bioRxiv preprint doi: https://doi.org/10.1101/2020.11.16.384743; this version posted November 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Recurrent mutations in SARS-CoV-2 genomes isolated from mink point to rapid host-adaptation

Lucy van Dorp[1], Cedric CS Tan[1], Su Datt Lam[2,3], Damien Richard[1,4], Christopher Owen[1], Dorothea Berchtold[1], Christine Orengo[3], François Balloux[1]

1) UCL Genetics Institute, University College London, United Kingdom.
2) Department of Applied Physics, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, Malaysia.
3) Institute of Structural and Molecular Biology, University College London, United Kingdom.
4) UCL Institute of Child Health, University College London, United Kingdom.

Correspondence: lucy.dorp.12@ucl.ac.uk (Lucy van Dorp), f.balloux@ucl.ac.uk (François Balloux)

Keywords:

Zoonotic disease, SARS-CoV-2, COVID-19, mink, host-adaptation, spike protein

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.16.384743; this version posted November 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Abstract [234 words]

Severe acute respiratory coronavirus 2 (SARS-CoV-2), the agent of the ongoing COVID-19 pandemic, jumped into humans from an unknown animal reservoir in late 2019. In line with other coronaviruses, SARS-CoV-2 has the potential to infect a broad range of hosts. SARS-CoV-2 genomes have now been isolated from cats, dogs, lions, tigers and minks. SARS-CoV-2 seems to transmit particularly well in mink farms with outbreaks reported in Spain, Sweden, the Netherlands, Italy, the USA and Denmark. Genomic data from SARS-CoV-2 isolated from infected minks provides a natural case study of a secondary host jump of the virus, in this case from humans to animals, and occasionally back again. We screened published SARS-CoV-2 genomes isolated from minks for the presence of recurrent mutations common in mink but infrequent in SARS-CoV-2 genomes isolated from human infections. We identify 23 recurrent mutations including three nonsynonymous mutations in the Receptor Binding Domain of the SARS-CoV-2 spike protein that independently emerged at least four times but are only rarely observed in human lineages. The repeat emergence of mutations across phylogenetically distinct lineages of the virus isolated from minks points to ongoing adaptation of SARS-CoV-2 to a new host. The rapid acquisition and spread of SARS-CoV-2 mutations in minks suggests that if a similar phenomenon of host adaptation had occurred upon its jump into humans, those human-specific mutations would likely have reached fixation already before the first SARS-CoV-2 genomes were generated.

Data Summary

All genome assemblies considered in this manuscript are openly available on registration with GISAID (https://www.gisaid.org). Information on the included assemblies, including the accessions used in the global analysis are provided in **Tables S1-S2**.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.16.384743; this version posted November 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Introduction

SARS-CoV-2, the agent of the COVID-19 pandemic, is part of the Coronaviridae family, whose members are widespread among wild populations of birds and mammals and tend to have fairly broad host ranges[1]. Seven coronavirus species are currently known to infect humans; SARS-CoV-1, SARS-CoV-2, MERS-CoV, and four endemic 'common cold' coronaviruses (HCoV-HKU1, HCoV-OC413, HCoV-229E, HCoV-NL63), all of which are of zoonotic origin. The original host of SARS-CoV-2 has not been identified to date, but bats are considered to be a plausible source[2,3]. Bats are also believed to represent the original host reservoir for NL63 and 229E[4–6], SARS-CoV-1[7–10] and MERS-CoV[11–14], whereas the zoonotic source for HCoV-HKU1 and HCoV-OC43 may be rodents, or possibly ruminants for the latter[15–17].

Upon an initial successful host jump, the ability of a pathogen to spread within its novel host population depends on its ability to transmit effectively between individuals. The four endemic HCoVs transmit readily in humans and are highly contagious particularly in children[18,19]. SARS-CoV-1 was probably less transmissible than SARS-CoV-2[20], which likely contributed to the 2003 SARS outbreak being comparably easy to control. MERS-CoV transmits poorly between humans and the repeat outbreaks it causes are primarily due to spillovers into humans from a reservoir it established in camelids in the Arabic Peninsula[21,22]. Finally, SARS-CoV-2 is highly transmissible in humans with an unusually high basic reproductive numbers (R_0) for a respiratory virus having been reported in several setting[23].

Transmissibility of a pathogen is expected to vary in different hosts and may often be limited in the early stages after the host jump. Such initial maladaptation may lead to a phase of intense natural selection when a pathogen acquires key mutations that maximize its transmission ability in the novel host. Efforts to identify mutations associated to SARS-CoV-2's transmissibility failed to identify obvious candidates for adaptation to its human host[24,25]. The only possible exception is the spike D614G mutation, whose role as a driver of transmission, or a neutral genetic marker of the otherwise successful SARS-CoV-2 B.1 lineage remains debated[24,26–28]. One plausible reason for the lack of mutations unambiguously associated to transmission in SARS-CoV-2 is that they may have emerged and spread before the virus was identified. Indeed, if by the time the earliest genomes were being sequenced, all key mutations associated to increased transmission in humans had already reached fixation, it would be impossible to identify those solely by analyzing SARS-CoV-2 genomes in circulation in humans.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.16.384743; this version posted November 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Beyond humans and non-human primates, carnivores are predicted to be particularly susceptible to SARS-CoV-2 infection[29–34]. Genomic data is currently available for SARS-CoV-2 isolated from dogs, domestic cats, tigers, lions and minks[31,35–43]. Transmissibility may vary between susceptible carnivore hosts. For example domestic cats (*Felis catus*) are both more susceptible to SARS-CoV-2 and more infectious than dogs (*Canis lupus familiaris*)[41]. SARS-CoV-2 has also been shown to transmit well in experimental mustelid systems[35,39,44,45] and, in real-world settings, minks (*Neovison vison* and *Mustela lutreola*) are highly susceptible to SARS-CoV-2 circulating in humans. Major mink farm outbreaks have been reported in multiple countries since April 2020[37,38,43].

The generation and release of genomic data from SARS-CoV-2 outbreaks in farmed European (*Mustela lutreola*) and American minks (*Neovison vison*) offer opportunities to identify mutations emerging and spreading in mink farms following the host jump(s). The conditions in intensive farming settings may provide highly suitable environments for small fitness selective differentials between strains to overcome the effect of genetic drift, allowing for natural selection to further increase transmissibility. Particularly promising candidates for host adaptation are those mutations occurring recurrently in different mink SARS-CoV-2 lineages, and that are uncommon in human viruses. The secondary host jump from humans into minks offers a glimpse into the window of early viral host adaptation of SARS-CoV-2 to a new host that has likely been missed at the start of the COVID-19 pandemic.

We screened publicly available SARS-CoV-2 genomes isolated from minks for the presence of recurrent mutations. We identified seven nonsynonymous mutations independently arising at least three times as plausible candidates for adaptation to transmission in minks. All seven candidate mutations are at low frequency in large repositories of SARS-CoV-2 strains circulating in humans. We note in particular three recurrent nonsynonymous mutations in the receptor-binding domain of the SARS-CoV-2 spike protein (S-protein), the region essential for binding to host Angiotensin-converting enzyme 2 (ACE2) receptors allowing cell entry[46,47]. We computationally predict the role of candidates in this region on modulating the binding stability of the human and mink spike protein ACE2 complex. Our results highlight the rapid and repeat emergence of mutations in the spike protein, across phylogenetically distinct lineages of the virus isolated from minks in Denmark and the Netherlands, and point to rapid adaptation of SARS-CoV-2 to a new host.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.16.384743; this version posted November 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Materials and Methods

Alignment of mink SARS-CoV-2

All publicly available genome assemblies of SARS-CoV-2 isolated from minks were downloaded from both GISAID[48,49] (https://www.gisaid.org) and NCBI (https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/) as of 14th October 2020, which included published genomic data associated to two studies of SARS-CoV-2 in minks in the Netherlands[37,38]. A full list of the 239 considered accessions is provided in **Table S1** which includes 227 genome assemblies sampled in the Netherlands and 12 genome assemblies sampled in Denmark covering the period of the 24th of April 2020 to 16th of September 2020. Some of the genomes were flagged as 'low quality' (commented in **Table S1**) but were included in the initial analaysis. All genome assemblies were profile aligned to the SARS-CoV-2 reference genome Wuhan-Hu-1 (NC_045512.2) using MAFFT v7.205[50]. SNPs flagged as putative sequencing errors were masked (a full list of 'masked' sites is available at https://github.com/W-L/ProblematicSites_SARS-CoV2/blob/master/problematic_sites_sarsCov2.vcf, accessed 14/10/2020)[51]. Wuhan-Hu-1 was retained to root the tree. A maximum likelihood phylogenetic tree was built across the alignment (**Figure S1**) using RaxML[52] run via the Augur (https://github.com/nextstrain/augur) tree implementation (**Figure 1**). We informally estimated the substitution rate over the mink SARS-CoV-2 alignment by computing the root-to-tip temporal regression implemented in BactDating v1.0.1[53] (**Figure S2**).

Global context of mink SARS-CoV-2 infections

To place SARS-CoV-2 genomes isolated from minks into a global context we downloaded 56,803 high quality assemblies (high coverage, >29,700bp and with a fraction of 'N' nucleotides <5%) from the worldwide diversity of human SARS-CoV-2 available on GISAID[48,49] on 25/08/2020. All animal strains were removed as well as samples flagged by NextStrain as 'exclude' (https://github.com/nextstrain/ncov/blob/master/defaults/exclude.txt as of 25/08/2020). Following alignment and masking as described above, this left 54,793 assemblies for downstream analysis to which we added the mink SARS-CoV-2 alignment. A full metadata table, list of acknowledgements and exclusions is provided in **Table S2**. We constructured a maximum likelihood phylogenetic tree over the 55,030 included genomes using IQ-TREE v2.1.0 Covid release[54] specifying the fast mode. Following construction of the tree, 57 long branch phylogenetic outliers were detected by TreeShrink[55] and were subsequently removed, together with two low-quality mink SARS-CoV-2 assemblies which

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.16.384743; this version posted November 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

were also considered phylogenetic outliers (EPI_ISL_577816 and EPI_ISL_577819). Trees were queried and plotted using the R packages Ape v5.4[56] and ggtree v1.16.6[57] (see **Figure 1**).

Mutation analysis

All variable positions in the coding regions of the genome were identified and annotated for synonymous or nonsynonymous status (**Figure S3-S4**). This was done by retrieving the amino acid changes corresponding to all SNPs at these positions using a custom Biopython (v1.76) script (https://github.com/cednotsed/mink_homoplasies/blob/main/dnds/snp_to_sav_parser.py) with annotations reported in **Table S3**. The Orf coordinates used (including the Orf1ab ribosomal frameshift site) were obtained from the associated metadata for Wuhan-Hu-1 (NC_045512.2). Assemblies were also uploaded to CoVSurver (https://www.gisaid.org/epiflu-applications/covsurver-mutations-app/) to report the prevalence of mutations and indels relative to SARS-CoV-2 assemblies available on the GISAID database[48,49].

The frequencies of each type of mutation in the human and mink SARS-CoV-2 alignments were calculated using the base.freq function from the R package Ape v5.4[56] (**Figure S5**). We tested whether the mutational frequencies in the human and mink genomes differed using a Monte Carlo simulation of the $\chi 2$ statistic with fixed margins (2000 iterations)[58,59]. This was implemented using the chisq.test function in R with the simulate.p.value flag. CpG dinucleotide frequencies for both SARS-CoV-2 alignments of genomes isolated from human and mink were also calculated using a custom R script (https://github.com/cednotsed/mink_homoplasies/blob/main/CpG/plot_CpG.R) (**Figure S6**). A Wilcoxon rank sum test implemented using the wilcox.test function in R was used to test if the distribution of CpG dinucleotide frequencies differed between the two datasets.

Identification of recurrent mutations

We screened for the presence of recurrent mutations in the mink SARS-CoV-2 masked alignment using HomoplasyFinder v0.0.0.9[60], as described in our previous work[61,62]. HomoplasyFinder employs the method first described by Fitch[63], providing, for each site, the site specific consistency index and the minimum number of independent emergences in the phylogenetic tree. All nucleotide positions with a consistency index <0.5 are considered homoplasic (**Figure 2, Figure S7, Table S4**). Sites identified as homoplasic were screened against the global dataset of 54,793 human SARS-CoV-2 (**Figure S8**). In addition amino acid

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.16.384743; this version posted November 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

replacements were identified in human samples in the CoV-GLUE[64] repository (accessed 16th November 2020, last update from GISAID 9th November 2020), which provides frequently updated screens against all assemblies shared to GISAID[48,49] (**Figure 2, Table S4**).

Structural data used for the analysis

The structure of the SARS-CoV-2 spike protein, reference strain, bound to human ACE2 has been solved at 2.45Å resolution[65] (PDB ID 6M0J). We visualised this structure using PyMOL v2.4.1[66]. We used this as the template to model the structures of the American mink (*Neovision vision*) ACE2 bound to the SARS-CoV-2 reference (Wuhan-Hu-1) and the mink ACE2 protein bound to versions of SARS-CoV-2 mutated for three candidate sites in the RBD. We also built a model of the human ACE2 bound to the mutated SARS-CoV-2.

We generated query–template alignments using HH-suite[67] and predicted 3D models using MODELLER v.9.24[68]. We used the 'very_slow' schedule for model refinement to optimise the geometry of the complex and interface. We generated 10 models for each S-protein:ACE2 complex and selected the model with the lowest nDOPE[69] score, which reflects the quality of the model. Positive scores are likely to be poor models, while scores lower than -1 are likely to be native-like. The sequence similarity of the human ACE2 and the mink ACE2 is fairly high (83% amino-acid sequence identity) and all generated models were of high quality (nDOPE < -1).

Measuring changes in the stability of the S-protein:ACE2 complex following mutation

We calculated the change in stability of the S-protein:ACE2 complex using two independent methods. The first, HADDOCK[70], is one of the top-performing protein-protein docking servers in the CAPRI competition[71]. The HADDOCK scoring function uses linear combination of various energies, van der waals, electrostatics and desolvation. We employed HADDOCK (v2.4 web server) to score the complexes (**Figure S9**). We also calculated the change in the stability of the S-protein:ACE2 complex using mCSM-PPI2[72] (**Figure S10**). This assigns a graph-based signature vector to each mutation, which is then used within machine learning models to predict the binding energy. The signature vector is based upon atom-distance patterns in the wild-type protein, pharmacophore information and available experimental information, evolutionary information and energetic terms. We used the mCSM-PPI2 server (http://biosig.unimelb.edu.au/mcsm_ppi2/) for the simulations. These methods were used

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.16.384743; this version posted November 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

because we found in a previous study that they reported stability changes, following mutations in the S-protein:ACE2 complex, that correlated well with the available *in vivo* and *in vitro* experimental data on susceptibility to infection[29].

In particular we performed the following calculations:

1. Using the model of the mink ACE2:SARS-CoV-2 reference complex we mutated the target residue to those RBD candidates identified in mink SARS-CoV-2.
2. Using the model of the mink ACE2:SARS-CoV-2 mink complex we mutated the target residues to the human SARS-CoV-2 reference strain.
3. Using the structure of the human ACE2:SARS-CoV-2 reference we mutated the target residue to those RBD candidates identified in mink SARS-CoV-2.
4. Using the model of the human ACE2:SARS-CoV-2 mink complex we mutated the target residue to the human SARS-CoV-2 reference strain.

For HADDOCK, a value that is more negative than for the reference Wuhan-Hu-1 S-protein: ACE2 complex suggests stabilisation of the complex. Whilst for mCSM-PPI-2 negative ΔΔG values reflect destabilisation of the complex by the mutation and positive values reflect stabilisation of the complex.

We also evaluated structural changes for all combinations of RBD mutations and receptor complexes (**Figures S11-S13**).

Results

Genomic diversity in mink SARS-CoV-2

At the time of writing (6th November 2020) 239 SARS-CoV-2 genome assemblies isolated from minks were publicly available on GISAID[48,49] and NCBI (**Table S1**) spanning a sampling period of the 24th of April 2020 to the 16th of September 2020. An alignment across all mink SARS-CoV-2 comprised 1073 Single Nucleotide Polymorphisms (SNPs) (**Figure S1**) with a mean pairwise distance of 6.5 (range: 4.1 – 11.7) mutations between any two genomes. No genome deviated from the reference genome, Wuhan-Hu-1 (NC_045512.2), by more than 24 mutations and the alignment exhibited a highly significant temporal signal (**Figure S2**). Dynamic assignment of assemblies to lineages[73] (run 6th November 2020) placed mink SARS-CoV-2 into seven different SARS-CoV-2 lineages. All 12 genomes from mink farms in

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.16.384743; this version posted November 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Denmark fell into a single lineage: B.1.1. Both lineage assignments and phylogenetic placement of mink SARS-CoV-2 in a large global phylogeny of human isolates are consistent with multiple zoonotic jumps of SARS-CoV-2 from human to mink hosts seeding local farm-related outbreaks[37,38] (**Figure 1, Tables S1-S2**).

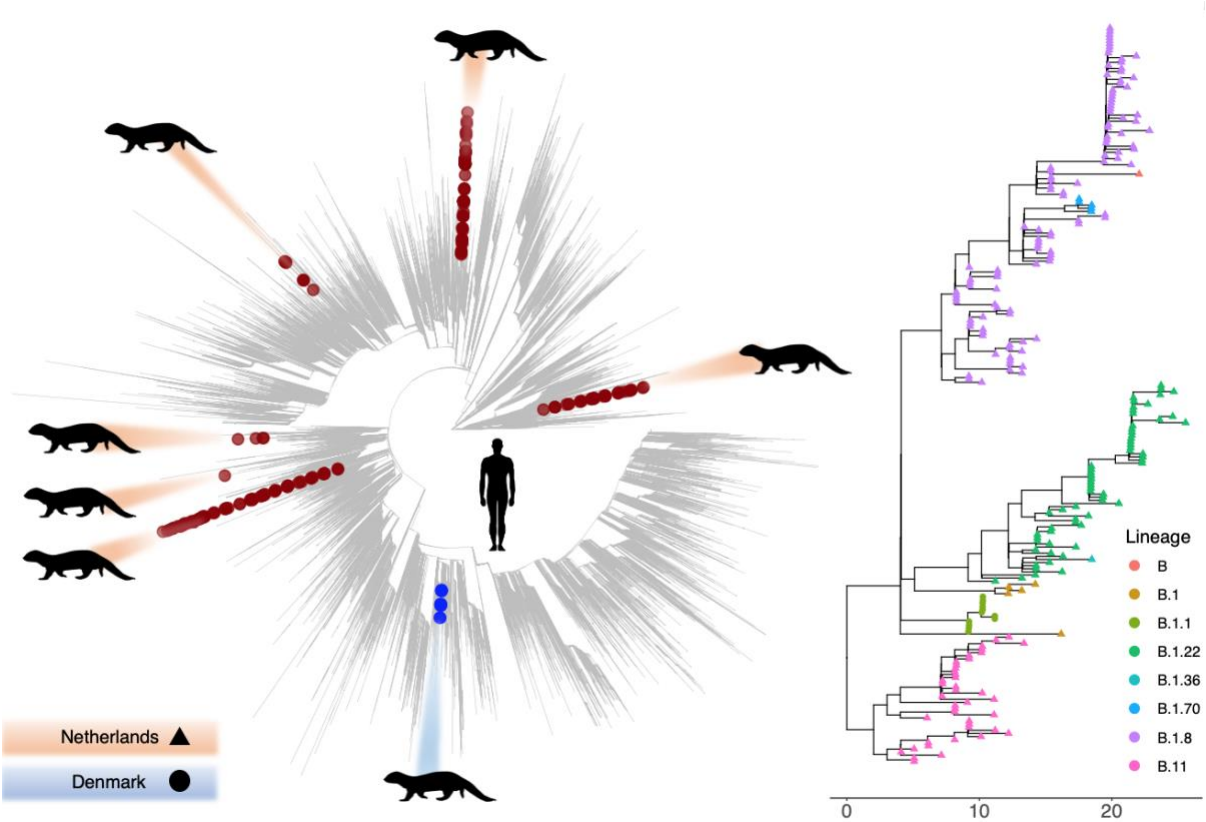

Figure 1: *(left) SARS-CoV-2 radial phylogenetic tree providing the relationships amongst SARS-CoV-2 isolated from humans and minks rooted on Wuhan-Hu-1. Genomes isolated from minks are shown with large tip symbols and coloured blue for those from Denmark and brown for those from the Netherlands. (right) Phylogenetic tree providing the relationships among SARS-CoV-2 genomes from minks with lineage assignments provided by tip colours and location by tip symbols (circle Denmark; triangle Netherlands). Full lineage assignments and a list of included genomes are provided in* ***Table S1*** *and* ***Table S2****.*

Across the mink SARS-CoV-2 alignment we observed a ratio of nonsynonymous to synonymous changes of 1.84, in line with the previous estimate of 1.88 for SARS-COV-2 lineages circulating in humans[24]. A full list of all identified mutations is provided in **Table S3** with nonsynonymous mutations displayed in **Figures S3-S4**. As previously reported for human SARS-CoV-2, we observe marked compositional asymmetries in mink SARS-CoV-2 likely deriving from the mutational action of host RNA-editing mechanisms[24,74,75]. For example, 37% of all nonsynonymous mutations represent C→U changes (**Figure S5**). This results in 87% of

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.16.384743; this version posted November 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

all possible nonsynonymous mutations involving a C being C→U changes, in line with C→U changes also representing by far the most common mutations for SARS-CoV-2 strains circulating in humans[24,74,75]. Though, the overall pattern of nucleotide substitutions is statistically significantly different between strains circulating in humans and minks ($p < 0.001$), with in particular a deficit of G→C and an excess of C→A mutations in minks **(Figure S5)**. SARS-CoV-2 genomes isolated from minks also showed a stronger depletion of CpG sites ($p = 2.2 \times 10^{-16}$) relative to their counterparts circulating in humans **(Figure S6)**.

Recurrent mutations in mink SARS-CoV-2

Across the masked alignment we detect 23 mutations that have appeared independently at least twice in SARS-CoV2 circulating in minks, corresponding to a consistency index (CI) <0.5[60]. Of those 23 mutations, 16 comprise nonsynonymous and seven synonymous changes (**Table S4**). Nonsynonymous mutations which have emerged independently at least three times include one within Orf1ab (12,795 - G4177E), three within Orf3a (25,936 – H182Y; 26,047 – L219V; 26078 – T229I) and three within the spike protein (22,920 – Y453F; 23,018 – F486L; 23,064 – N501T). Of note the latter three recurrent mutations in the spike protein correspond to residue changes in the SARS-CoV-2 spike RBD (**Figure 2, Figure S4, Figure S7**), which may be a particularly important region for adaptation of SARS-CoV-2 to host receptors[47]. In particular, we infer five independent emergences of Y453F across four lineages, five emergences of F486L across three lineages and four emergences of N501T across three lineages.

Mutations specifically increasing transmission in minks can be expected to have emerged only a limited number of times in SARS-CoV-2 strains circulating in humans and to have remained at low frequency in the human SARS-CoV-2 population. Using a phylogenetic dataset of >54,000 human SARS-CoV-2 genomes, we identify some overlap between sites homoplasic in minks and in humans (**Figure S8**). This suggests that mutations having emerged recurrently in viruses circulating in minks may not all reflect host adaptation but could also be the result of mutations induced by vertebrate APOBEC proteins[24,74,75]. This is in line with the large proportion of homoplasies involving C→T changes (12/23) consistent with know patterns of hypermutation in vertebrates[74] (**Figure 2**, **Table S4**). As such, we expect strong candidate mutations for SARS-CoV-2 host adaptation to minks to be at low frequency in SARS-CoV-2 circulating in humans.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.16.384743; this version posted November 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

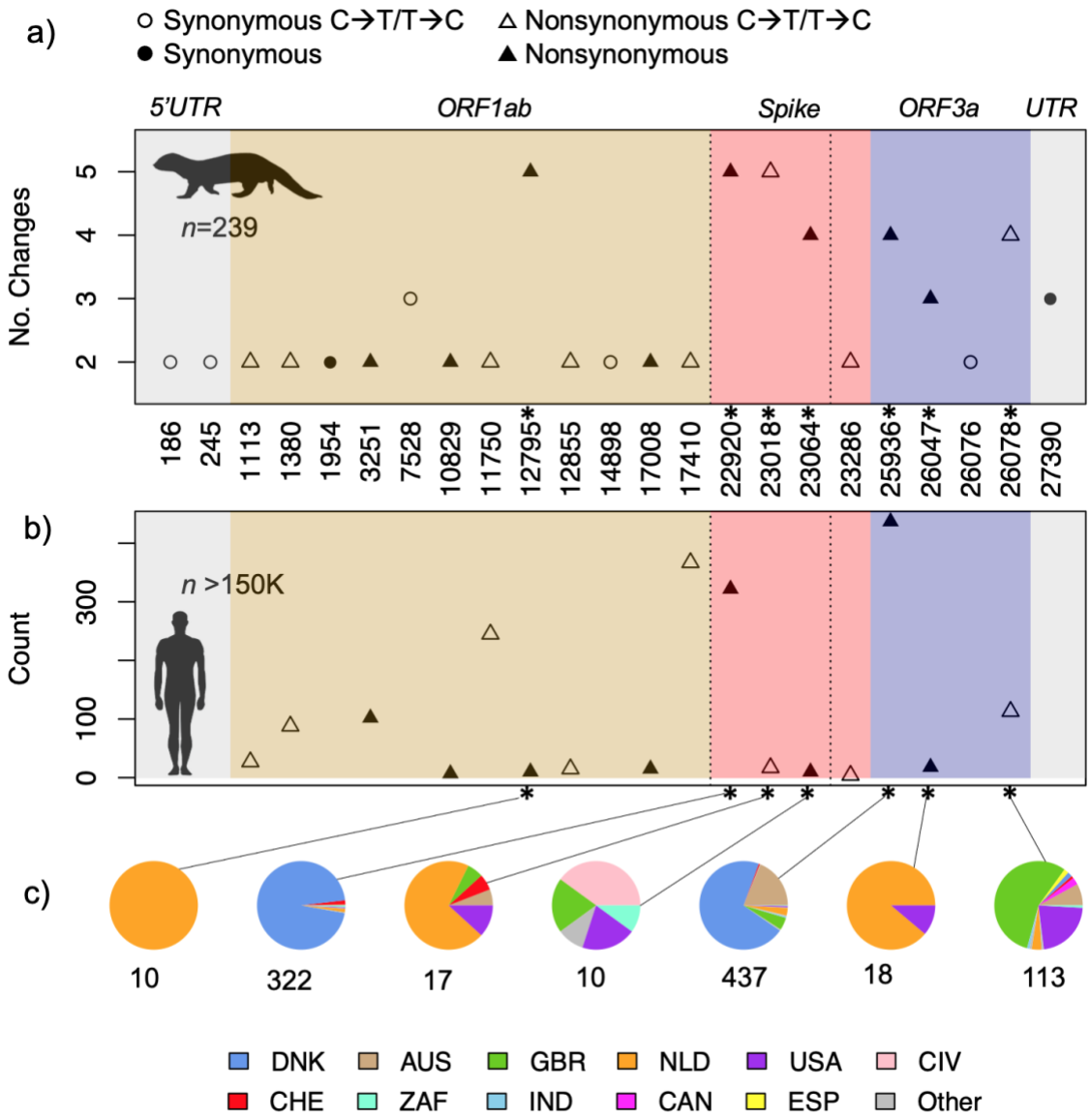

Figure 2: *a) 23 recurrent mutations identified in the mink SARS-CoV-2 alignment. Y-axis provides the minimum number of change points detected by HomoplasyFinder. Colours provide the genome annotations as given at top with different categories of mutation denoted by the symbols. Recurrent mutations in the spike RBD are bounded by vertical dashed lines. b) Number (y-axis) of human associated SARS-CoV-2 genomes annotated as carrying nonsynonymous mutations recurrent in mink associated SARS-CoV-2 as of 9th November 2020. c) Pies providing the geographic distribution of human SARS-CoV-2 assemblies carrying each of the seven candidate changes (corresponding to nonsynonymous mutations with at least three emergences in minks). Legend at bottom provides the country assignments: AUS – Australia, CHE – Switzerland, CAN – Canada, CIV – Ivory Coast, DEN – Denmark, ESP – Spain, GBR – Great Britain, IND – India, NLD – Netherlands, RUS – Russia, USA – United States of America, ZAF – South Africa. A full list is provided in* ***Table S4****.*

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.16.384743; this version posted November 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Following a screen of amino acid replacements in CoV-GLUE[64] (accessed 16th November 2020) we found that all seven nonsynonymous mutations that have independently emerged at least three times in strains circulating in minks are carried by fewer than ~0.3% of the SARS-CoV-2 genomes isolated from humans available on large publicly available sequencing repositories (**Figure 2, Table S4**). Of the mink homoplasies located in the spike protein, two derive from A→T (nucleotide position 22920 – Y453F) and A→C (nucleotide position 23018 – F486L) changes while one corresponds to a T→C change (nucleotide position 23064 – N501T). The recurrent emergence of these mutations in phylogenetically distant lineages, and their relative scarcity in SARS-CoV-2 samples circulating in humans, further supports these as strong candidates for ongoing host-adaptation of the virus to transmission in minks. Indeed, many of the cases identified in humans seem putatively linked to mink farm outbreaks in the Netherlands and Denmark[37,38,76] (**Figure 2c**).

Predicted impact of recurrent mutations in the spike protein

Having observed the presence of three candidate mutations falling within the SARS-CoV-2 spike RBD we considered their role in receptor binding affinity as putative sites of adaptation to a mink host. The spike RBD (codon positions 319-541[77]) provides a critical region for SARS-CoV-2 to attach to host cells via docking to ACE2 receptors, thereby allowing subsequent SARS-CoV-2 entry into host cells and eventual replication[46,47]. This also makes the RBD the target of neutralising antibodies[78–81], i.e. antibodies preventing cellular infection upon prior exposure to the pathogen or a vaccine. Specific residues within the RBD have been identified as critical for receptor binding[29,82,83], with potential to modulate both infectivity and antigenicity[77]. All three recurrent spike RBD mutations (Y453F, F486L and N501T) suggested by the phylogenetic analyses are in residues directly involved in contacts in the S-protein:ACE2 interface and are therefore relevant to the binding and stability of the complex (**Figure 3**).

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.16.384743; this version posted November 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

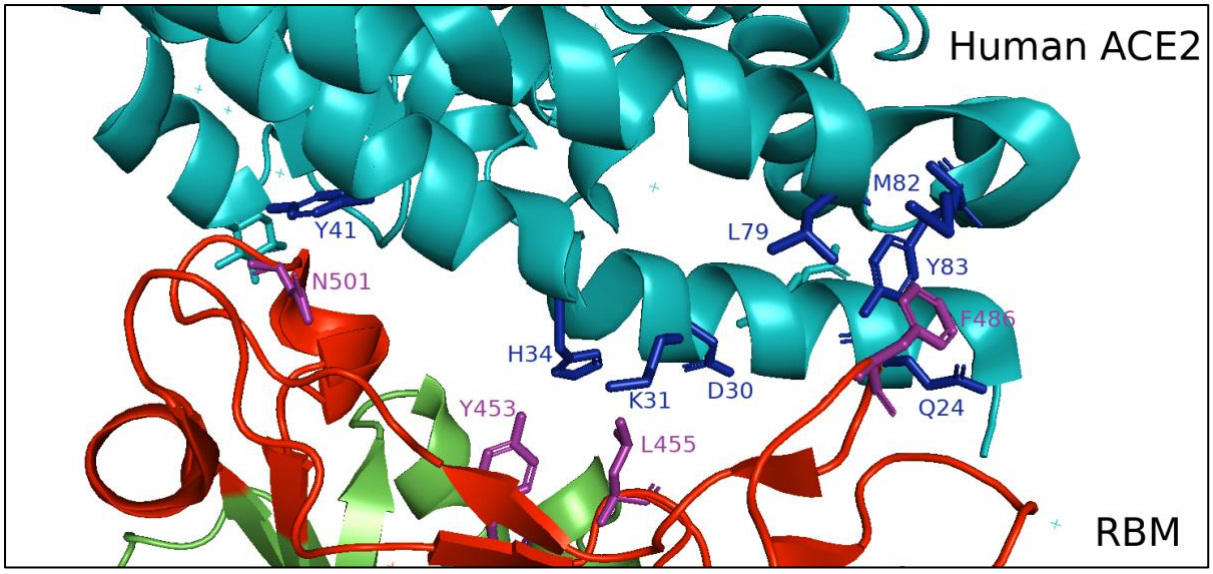

Figure 3: *Protein structure of the receptor binding motif (RBM) of the receptor binding domain (RBD) in complex with human ACE2. Human ACE2, the RBM and the rest of the receptor binding domain are shown in teal, red and green respectively. The wild-type amino acid residues for the recurrent nonsynonymous mutations Y453F, F486L, N501T are shown in purple. The ACE2 residues reported previously to be interacting with the RBM residues L455 (in close proximity to F453), F486 and N501 are shown in blue. The Protein Data Bank (PDB) code for the SARS-CoV-2 RBM–ACE2 complex is 6M0J. This figure was rendered using PyMOL (v2.4.1).*

We used the HADDOCK[70] and mCSM-PPI2[72] protocols to analyse the change in stability of the mink SARS-CoV-2 S-protein:ACE2 complex for the three recurrent spike RBD mutations. To analyse the changes in stability we used 3D-models of the mink ACE2 bound to SARS-CoV-2 mink strain and then mutated the target residue to that found in human-associated reference SARS-CoV-2. We used this approach as previous work showed that it gave results that correlated well with experimental data on susceptibility to infection[29]. However, we got similar results when we performed the calculation by using the complex of the mink ACE2 bound to SARS-CoV-2 Wuhan-Hu-1 and mutated the target SARS-CoV-2 residue to those observed in minks (**Figure S9-S10**). The results indicate marginal changes in the stability of the complex for these mutations and are somewhat conflicting between the methods, but none of the values reported is expected to significantly stabilise or destabilise the complex.

Structural analyses

In light of the small estimated changes in binding energy associated with complex stabilities, additional structural analyses were performed to gain further insights into whether the mutations were likely to affect complex stability. The first of the candidate residues 453

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.16.384743; this version posted November 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

(nucleotide position 22920) lies close to the centre of the S-protein:ACE2 interface. In Wuhan-Hu-1 the tyrosine at this position interacts with H34 in human ACE2 enabling the formation of a polar hydrogen bond which would stabilise the complex. In the American mink ACE2, position 34 has a tyrosine residue which would lead to a clash with the polar tyrosine 453 for Wuhan-Hu-1 SARS-CoV-2. Therefore the emergence of mutation of Y453 to a phenyalanine, observed in 47 mink associated SARS-CoV-2 (**Table S3**), could be favourable as it reduces the clash of polar groups (see **Figure S11**).

The SARS-CoV-2 spike residue 486 (nt position 23018) has been identified as one of the most important locations – 'the receptor binding motif' - for human ACE2 binding to the spike protein[82,83] and lies within a hydrophobic pocket (**Figure S12**). It has been shown that SARS-CoV-2 exploits this pocket better than SARS-CoV-1, as the phenylalanine residue allows for pi stacking of the aromatic rings with Y83 in human ACE2 which enhances affinity[65,82,83]. Our previous analysis[29] showed the Y83 residue to be highly conserved across human and other animals including the American mink. While we note that this residue change arises from a T→C transition (**Figure 2**), this polymorphism has appeared at least five times and was observed in 96 mink associated SARS-CoV-2. Mutation of phenylalanine to leucine replaces the large aromatic ring with an aliphatic leucine which is not involved in any substantive interactions with the Y83 (see **Figure S12**).

Our final candidate, N501T (nt position 23064) is only observed in five mink SARS-CoV-2 but has independently emerged four times (**Table S4**). This residue change has been observed in only ten human SARS-CoV-2 assemblies to date (**Figure 2**). The recurrent mutation leads to the replacement of the asparagine by a threonine residue in the mink associated strain (**Figure S13**). This gives rise to a hydrogen bond to lysine 353 which may contribute to the small increase in stability reported by mCSM-PPI2 (**Figure S10**).

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.16.384743; this version posted November 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Discussion

The COVID-19 pandemic is understood to have been caused by a unique host jump into humans from a single yet-undescribed zoonotic source in the latter half of 2019 most likely in China[2,3,61]. This created a situation where the entire genetic diversity of the SARS-CoV-2 population was initially negligible despite the virus having rapidly swept across the world. Genetic diversity has been building up since the start of the pandemic through the acquisition of new mutations, but is still limited relative to other RNA viruses with broad distributions. Conversely, the secondary host jump into minks involved multiple independent secondary transmissions into mink farms in different countries. Whilst the data we analysed represents only a subset of infected mink farms from just two countries, the Netherlands and Denmark, we already identify a minimum of seven independent host jumps from humans to minks, involving strains representative of essentially the entire SARS-CoV-2 lineage diversity in circulation (**Figure 1**).

The secondary transmissions of SARS-CoV-2 from humans into minks provides a set of 'natural experiments' to identify mutations involved in the adaptation of the virus to a novel host. By analysing SARS-CoV-2 isolated from minks, we recovered 23 mutations having independently emerged multiple times. By restricting this set to the nonsynonymous mutations that have appeared at least three times in mink, we identify seven variants that are strong candidates for host adaptation to minks (**Figure 2, Table S4**). These seven candidates comprise a recurrent change in nsp9 of Orf1ab, a region involved in mediating viral replication[84], as well as the repeat emergence of three mutations in Orf3a, a protein thought to play an important role in triggering host inflammatory response[85,86].

We particularly focus on the three candidates that have emerged independently in minks at least four times in the RBD of the SARS-CoV-2 spike protein, a region vital for determining host-range[47] (**Figure 2, Figure S4**). We detect at least five independent emergences of Y453F across four phylogenetic lineages, five emergences of F486L across three lineages and four emergences of N501T across three lineages in minks. It is noteworthy that all three mutations are not, or only marginally homoplasic, in humans and all three have also been observed at low frequency in strains circulating in human COVID-19 infections which have been intensively sampled; ~150,000 high quality genome assemblies are available on GISAID at the time of writing (**Table S4, Figure 2**).

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.16.384743; this version posted November 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Quantifying the consequences of these candidate RBD mutations on the binding affinity of the SARS-CoV-2 spike protein to human and mink ACE2 receptors we predict only subtle changes in S-protein:ACE2 complex stability for Y453F and F486L (**Figure S9-S10**). Consistently, characterisation of mutated variants in the RBD region suggest the region has a high tolerance to mutation in the context of receptor binding[87] and we note that previous studies have shown that animals are susceptible to infection by SARS-CoV2 despite much higher destabilisation of the complex than reported for these mutants[29,30,88]. A published approach based on fluorescent detection of ACE2 binding reports negative binding constants for F486L though suggests a positive effect for Y453F and N501T to human ACE2[87] (https://jbloomlab.github.io/SARS-CoV-2-RBD_DMS/ accessed 12th November 2020).

We also identify structural support for a marginal increase in stability of the mink complex with the N501T mutation relative to wild type (**Figure S10, Figure S13**). In our dataset we detect the presence of this mutation in five mink SARS-CoV-2 though this corresponds to at least four independent emergences, with N501T being exceptionally rare in human infections (**Figure 2**). It may also be relevant that a different residue change in this position (N501Y) has been proposed as a mechanism of host adaption in mice infected with SARS-CoV-2[89] suggesting the broader role of residue changes at 501 as relevant affinity enhancing mutations[87].

In addition to adaptation to a novel host through acquisition of new variants, RNA viruses are also under pressure of their hosts' immune system which comprises antiviral mechanisms such as APOBEC proteins inducing hypermutation at specific sites in their genome[74,75]. While C→U changes, most likely induced by proteins from the APOBEC family, represent the majority of mutations in SARS-CoV-2 circulating both in minks and humans, we also detected subtle shifts in mutational patterns likely due to differences in the innate immune system of humans and minks **(Figure S5-S6)**. The widespread depletion of CpG sites in RNA viruses can also be explained by another defense mechanism in the vertebrate innate immune system, in the form of Zinc-Finger Antiviral Proteins (ZAPs), which target viral RNA for exosome-mediated degradation[90,91]. There have been earlier suggestions that ZAPs from different vertebrate host taxa may exert varying levels of CpG depletion pressure[92,93]. In this context, it is an intriguing observation that SARS-CoV-2 genomes isolated from minks displayed a moderate but statistically highly significant depletion of CpG sites (p = 2.2 x 10^{-16}) relative to their counterparts circulating in humans **(Figure S6)**.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.16.384743; this version posted November 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

A combination of biological and epidemiological factors make mink farms highly susceptible to SARS-CoV-2 outbreaks with the risk of subsequent transmission back into humans[37,38,94]. As such, mink farms represent reservoirs for the virus which can greatly complicate the containment of COVID-19. At the time of writing a further 214 mink associated human COVID-19 infections were reported in the North Jutland region of Denmark, highlighting the risk of spill-overs from animal reservoirs (announced 5th November 2020, working paper available https://files.ssi.dk/Mink-cluster-5-short-report_AFO2 accessed 12th November 2020)[76]. A phylogenetic grouping of the virus termed 'Cluster 5' circulating in North Jutland has attracted particular attention. This cluster is reported to carry our Y453F candidate mutation, together with three other mutations in the spike protein falling outside the RBD (del69-70, I692V, M1229I), none of which we identify as mink-adapted. The cluster 5 lineage is known to have infected up to 12 people and its spread led to the Danish Government considering closing all mink farms in the country. The putative impact on antigen-escape of the Y453F mutation remains under study, though appears moderate[76].

Our work points to the frequent emergence of adaptive mutations in minks which may be retained during anthroponotic spillovers, at least transiently so. To date, the vast majority of mutations that have been observed in SARS-CoV-2 appear selectively neutral, or even deleterious in humans[24,95] and those which we identify as strong candidates for host-adaptation to minks have remained at low frequency globally in human SARS-CoV-2 to date (**Figure 2**). There is no *a priori* reason to expect that mutations adaptive in minks will lead to any marked change in the dynamic of human COVID-19 infections. Indeed, SARS-CoV-2 lineages carrying candidate mutations adaptive to transmission in minks have been sequenced since the first outbreaks of SARS-CoV-2 on mink farms in April 2020, with the first mink SARS-CoV-2 sample carrying Y453F dating back to the 24th April 2020.

The low prevalence of mink-adapted candidate mutations in strains in circulation in humans to date (November 2020) suggests they are not expected to increase transmission of the virus in humans. However, mutations located within the RBD of the SARS-CoV-2 spike protein do warrant careful monitoring. Genetic variation in the RBD region is common and studies of related Sarbecoviruses have previously identified signals consistent with complex selection histories in this region[2,10,96,97]. The RBD is the most immunodominant region of the SARS-CoV-2 genome and any mutation in this region therefore has potential implications for

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.16.384743; this version posted November 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

antigenic response[78–81]. The spike RBD is also the target of antibody based therapies and vaccines[98–101]. In this context, it may seem concerning that Y453F has been flagged as as a possible vaccine-escape mutation[102]. Though our work suggests this mutation is an adaptive change specifically enhancing transmission in minks. Other RBD variants, already found at higher frequency in human SARS-CoV-2, represent stronger candidates for vaccine-escape mutations[103,104]. Once vaccines will be deployed, careful characterisation and tracking of the frequency of such mutations will be essential to inform if and when vaccines will need to be redesigned. This critical effort will be greatly enabled by open genomic data sharing platforms[48,49,64,105].

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.16.384743; this version posted November 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

References

1. Shaw, L. P. *et al.* The phylogenetic range of bacterial and viral pathogens of vertebrates. *Mol. Ecol.* mec.15463 (2020). doi:10.1111/mec.15463
2. Boni, M. F. *et al.* Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. *Nat. Microbiol.* **5**, 1408–1417 (2020).
3. Zhou, P. *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579**, 270–273 (2020).
4. Corman, V. M. *et al.* Evidence for an Ancestral Association of Human Coronavirus 229E with Bats. *J. Virol.* **89**, 11858–70 (2015).
5. Huynh, J. *et al.* Evidence supporting a zoonotic origin of human coronavirus strain NL63. *J. Virol.* **86**, 12816–25 (2012).
6. Tao, Y. *et al.* Surveillance of Bat Coronaviruses in Kenya Identifies Relatives of Human Coronaviruses NL63 and 229E and Their Recombination History. *J. Virol.* **91**, (2017).
7. Ge, X.-Y. *et al.* Isolation and characterization of a bat SARS-like coronavirus that uses the ACE2 receptor. *Nature* **503**, 535–538 (2013).
8. Li, W. *et al.* Bats Are Natural Reservoirs of SARS-Like Coronaviruses. *Science (80-.).* **310**, 676–679 (2005).
9. Lau, S. K. P. *et al.* Severe acute respiratory syndrome coronavirus-like virus in Chinese horseshoe bats. *Proc. Natl. Acad. Sci.* **102**, 14040–14045 (2005).
10. Hu, B. *et al.* Discovery of a rich gene pool of bat SARS-related coronaviruses provides new insights into the origin of SARS coronavirus. *PLOS Pathog.* **13**, e1006698 (2017).
11. Corman, V. M. *et al.* Rooting the phylogenetic tree of middle East respiratory syndrome coronavirus by characterization of a conspecific virus from an African bat. *J. Virol.* **88**, 11297–303 (2014).
12. Anthony, S. J. *et al.* Further Evidence for Bats as the Evolutionary Source of Middle East Respiratory Syndrome Coronavirus. *MBio* **8**, (2017).
13. Lau, S. K. P. *et al.* Receptor Usage of a Novel Bat Lineage C Betacoronavirus Reveals Evolution of Middle East Respiratory Syndrome-Related Coronavirus Spike Proteins for Human Dipeptidyl Peptidase 4 Binding. *J. Infect. Dis.* **218**, 197–207 (2018).
14. Wang, Q. *et al.* Bat Origins of MERS-CoV Supported by Bat Coronavirus HKU4 Usage of Human Receptor CD26. *Cell Host Microbe* **16**, 328–337 (2014).
15. Su, S. *et al.* Epidemiology, Genetic Recombination, and Pathogenesis of

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.16.384743; this version posted November 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Coronaviruses. *Trends Microbiol.* **24**, 490 (2016).

16. Vijgen, L. *et al.* Complete genomic sequence of human coronavirus OC43: molecular clock analysis suggests a relatively recent zoonotic coronavirus transmission event. *J. Virol.* **79**, 1595–604 (2005).
17. Hogue, B. G., King, B. & Brian, D. A. Antigenic relationships among proteins of bovine coronavirus, human respiratory coronavirus OC43, and mouse hepatitis coronavirus A59. *J. Virol.* **51**, 384–8 (1984).
18. Nickbakhsh, S. *et al.* Epidemiology of Seasonal Coronaviruses: Establishing the Context for the Emergence of Coronavirus Disease 2019. *J. Infect. Dis.* **222**, 17–25 (2020).
19. Dijkman, R. *et al.* Human coronavirus NL63 and 229E seroconversion in children. *J. Clin. Microbiol.* **46**, 2368–73 (2008).
20. Cevik, M., Kuppalli, K., Kindrachuk, J. & Peiris, M. Virology, transmission, and pathogenesis of SARS-CoV-2. *BMJ* **371**, m3862 (2020).
21. Reusken, C. B., Raj, V. S., Koopmans, M. P. & Haagmans, B. L. Cross host transmission in the emergence of MERS coronavirus. *Curr. Opin. Virol.* **16**, 55–62 (2016).
22. Alagaili, A. N. *et al.* Middle East respiratory syndrome coronavirus infection in dromedary camels in Saudi Arabia. *MBio* **5**, e00884-14 (2014).
23. Liu, Y., Gayle, A. A., Wilder-Smith, A. & Rocklöv, J. The reproductive number of COVID-19 is higher compared to SARS coronavirus. *J. Travel Med.* **27**, (2020).
24. van Dorp, L. *et al.* No evidence for increased transmissibility from recurrent mutations in SARS-CoV-2. *bioRxiv* 2020.05.21.108506 (2020). doi:10.1101/2020.05.21.108506
25. MacLean, O. A. *et al.* Natural selection in the evolution of SARS-CoV-2 in bats, not humans, created a highly capable human pathogen. *bioRxiv* 2020.05.28.122366 (2020). doi:10.1101/2020.05.28.122366
26. Korber, B. *et al.* Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. *Cell* **182**, 812-827.e19 (2020).
27. Volz, E. M. *et al.* Evaluating the effects of SARS-CoV-2 Spike mutation D614G on transmissibility and pathogenicity. *medRxiv* 2020.07.31.20166082 (2020). doi:10.1101/2020.07.31.20166082
28. Grubaugh, N. D., Hanage, W. P. & Rasmussen, A. L. Making Sense of Mutation: What

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.16.384743; this version posted November 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

D614G Means for the COVID-19 Pandemic Remains Unclear. (2020). doi:10.1016/j.cell.2020.06.040

29. Lam, S. D. D. *et al.* SARS-CoV-2 spike protein predicted to form complexes with host receptor protein orthologues from a broad range of mammals. *Sci. Rep.* **10**, 16471 (2020).
30. Damas, J. *et al.* Broad host range of SARS-CoV-2 predicted by comparative and structural analysis of ACE2 in vertebrates. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 22311–22322 (2020).
31. Santini, J. M. & Edwards, S. J. L. Host range of SARS-CoV-2 and implications for public health. *The Lancet Microbe* **1**, e141–e142 (2020).
32. Melin, A. D., Janiak, M. C., Marrone, F., Arora, P. S. & Higham, J. P. Comparative ACE2 variation and primate COVID-19 risk. *Commun. Biol.* **3**, 641 (2020).
33. Infection with Novel Coronavirus (SARS-CoV-2) Causes Pneumonia in the Rhesus Macaques. (2020). doi:10.21203/RS.2.25200/V1
34. Rockx, B. *et al.* Comparative pathogenesis of COVID-19, MERS, and SARS in a nonhuman primate model. *Science* **368**, 1012–1015 (2020).
35. Shi, J. *et al.* Susceptibility of ferrets, cats, dogs, and other domesticated animals to SARS-coronavirus 2. *Science* **368**, 1016–1020 (2020).
36. McAloose, D. *et al.* From People to Panthera: Natural SARS-CoV-2 Infection in Tigers and Lions at the Bronx Zoo. *MBio* **11**, (2020).
37. Munnink, B. B. O. *et al.* Transmission of SARS-CoV-2 on mink farms between humans and mink and back to humans. *Science (80-.).* (2020). doi:10.1126/SCIENCE.ABE5901
38. Oreshkova, N. *et al.* SARS-CoV-2 infection in farmed minks, the Netherlands, April and May 2020. *Eurosurveillance* **25**, 2001005 (2020).
39. Kim, Y.-I. *et al.* Infection and Rapid Transmission of SARS-CoV-2 in Ferrets. *Cell Host Microbe* **27**, 704-709.e2 (2020).
40. Sia, S. F. *et al.* Pathogenesis and transmission of SARS-CoV-2 in golden hamsters. *Nature* **583**, 834–838 (2020).
41. Bosco-Lauth, A. M. *et al.* Experimental infection of domestic dogs and cats with SARS-CoV-2: Pathogenesis, transmission, and response to reexposure in cats. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 26382–26388 (2020).
42. Sit, T. H. C. *et al.* Infection of dogs with SARS-CoV-2. *Nature* **586**, 776–778 (2020).

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.16.384743; this version posted November 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

43. Molenaar, R. J. *et al.* Clinical and Pathological Findings in SARS-CoV-2 Disease Outbreaks in Farmed Mink (Neovison vison). *Vet. Pathol.* **57**, 653–657 (2020).

44. Richard, M. *et al.* SARS-CoV-2 is transmitted via contact and via the air between ferrets. *Nat. Commun.* **11**, 3496 (2020).

45. Belser, J. A. *et al.* Ferrets as Models for Influenza Virus Transmission Studies and Pandemic Risk Assessments. *Emerg. Infect. Dis.* **24**, 965–971 (2018).

46. Hoffmann, M. *et al.* SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. *Cell* **181**, 271-280.e8 (2020).

47. Letko, M., Marzi, A. & Munster, V. Functional assessment of cell entry and receptor usage for SARS-CoV-2 and other lineage B betacoronaviruses. *Nat. Microbiol.* **5**, 562–569 (2020).

48. Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data – from vision to reality. *Eurosurveillance* **22**, 30494 (2017).

49. Elbe, S. & Buckland-Merrett, G. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Glob. Challenges* **1**, 33–46 (2017).

50. Katoh, K. & Standley, D. M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).

51. De Maio, N. *et al.* Issues with SARS-CoV-2 sequencing data. *Virological* **5**, (2020).

52. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).

53. Didelot, X., Croucher, N. J., Bentley, S. D., Harris, S. R. & Wilson, D. J. Bayesian inference of ancestral dates on bacterial phylogenetic trees. *Nucleic Acids Res.* **46**, e134–e134 (2018).

54. Minh, B. Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).

55. Mai, U. & Mirarab, S. TreeShrink: fast and accurate detection of outlier long branches in collections of phylogenetic trees. *BMC Genomics* **19**, 272 (2018).

56. Paradis, E. & Schliep, K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526–528 (2019).

57. Yu, G., Smith, D. K., Zhu, H., Guan, Y. & Lam, T. T.-Y. ggtree : an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.* **8**, 28–36 (2017).

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.16.384743; this version posted November 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

58. Hope, A. C. A. A Simplified Monte Carlo Significance Test Procedure. *J. R. Stat. Soc. Ser. B* **30**, 582–598 (1968).

59. Patefield, W. M. Algorithm AS 159: An Efficient Method of Generating Random R × C Tables with Given Row and Column Totals. *Appl. Stat.* **30**, 91 (1981).

60. Crispell, J., Balaz, D. & Gordon, S. V. HomoplasyFinder: a simple tool to identify homoplasies on a phylogeny. *Microb. genomics* **5**, (2019).

61. van Dorp, L. *et al.* Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. *Infect. Genet. Evol.* **83**, 104351 (2020).

62. Richard, D. *et al.* No evidence for increased transmissibility from recurrent mutations in SARS-CoV-2. *Nat. Commun.* doi:10.5281/zenodo.4147272

63. Fitch, W. M. Toward Defining the Course of Evolution: Minimum Change for a Specific Tree Topology. *Syst. Biol.* **20**, 406–416 (1971).

64. Singer, J., Gifford, R., Cotten, M. & Robertson, D. CoV-GLUE: A Web Application for Tracking SARS-CoV-2 Genomic Variation. (2020). doi:10.20944/PREPRINTS202006.0225.V1

65. Lan, J. *et al.* Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. *Nature* **581**, 215–220 (2020).

66. Version 2.4.1 Schrödinger, L. The PyMOL Molecular Graphics System.

67. Steinegger, M. *et al.* HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics* **20**, 473 (2019).

68. Webb, B. & Sali, A. Comparative Protein Structure Modeling Using MODELLER. *Curr. Protoc. Bioinforma.* **54**, 5.6.1-5.6.37 (2016).

69. Shen, M.-Y. & Sali, A. Statistical potential for assessment and prediction of protein structures. *Protein Sci.* **15**, 2507–24 (2006).

70. van Zundert, G. C. P. *et al.* The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes. *J. Mol. Biol.* **428**, 720–725 (2016).

71. Koukos, P. I. *et al.* An overview of data-driven HADDOCK strategies in CAPRI rounds 38-45. *Proteins Struct. Funct. Bioinforma.* **88**, 1029–1036 (2020).

72. Rodrigues, C. H. M., Myung, Y., Pires, D. E. V & Ascher, D. B. mCSM-PPI2: predicting the effects of mutations on protein–protein interactions. *Nucleic Acids Res.* **47**, W338–W344 (2019).

73. Rambaut, A. *et al.* A dynamic nomenclature proposal for SARS-CoV-2 lineages to

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.16.384743; this version posted November 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

assist genomic epidemiology. *Nat. Microbiol.* 1–5 (2020). doi:10.1038/s41564-020-0770-5

74. Simmonds, P. Rampant C→U Hypermutation in the Genomes of SARS-CoV-2 and Other Coronaviruses: Causes and Consequences for Their Short- and Long-Term Evolutionary Trajectories. *mSphere* **5**, (2020).
75. Giorgio, S. Di, Martignano, F., Torcia, M. G., Mattiuz, G. & Conticello, S. G. Evidence for host-dependent RNA editing in the transcriptome of SARS-CoV-2. *Sci. Adv.* **6**, eabb5813 (2020).
76. Ecdc. *Detection of new SARS-CoV-2 variants related to mink*. (2020).
77. Li, Q. *et al.* The Impact of Mutations in SARS-CoV-2 Spike on Viral Infectivity and Antigenicity. (2020). doi:10.1016/j.cell.2020.07.012
78. Jiang, S., Hillyer, C. & Du, L. Neutralizing Antibodies against SARS-CoV-2 and Other Human Coronaviruses. *Trends Immunol.* **41**, 355–359 (2020).
79. Cao, Y. *et al.* Potent Neutralizing Antibodies against SARS-CoV-2 Identified by High-Throughput Single-Cell Sequencing of Convalescent Patients' B Cells. *Cell* **182**, 73-84.e16 (2020).
80. Ju, B. *et al.* Human neutralizing antibodies elicited by SARS-CoV-2 infection. *Nature* **584**, 115–119 (2020).
81. Shi, R. *et al.* A human neutralizing antibody targets the receptor-binding site of SARS-CoV-2. *Nature* **584**, 120–124 (2020).
82. Shang, J. *et al.* Structural basis of receptor recognition by SARS-CoV-2. *Nature* **581**, 221–224 (2020).
83. Brielle, E. S., Schneidman-Duhovny, D. & Linial, M. The SARS-CoV-2 exerts a distinctive strategy for interacting with the ACE2 human receptor. *bioRxiv* 2020.03.10.986398 (2020). doi:10.1101/2020.03.10.986398
84. Littler, D. R., Gully, B. S., Colson, R. N. & Rossjohn, J. Crystal Structure of the SARS-CoV-2 Non-structural Protein 9, Nsp9. *iScience* **23**, 101258 (2020).
85. Siu, K. *et al.* Severe acute respiratory syndrome Coronavirus ORF3a protein activates the NLRP3 inflammasome by promoting TRAF3-dependent ubiquitination of ASC. *FASEB J.* **33**, 8865–8877 (2019).
86. Issa, E., Merhi, G., Panossian, B., Salloum, T. & Tokajian, S. SARS-CoV-2 and ORF3a: Nonsynonymous Mutations, Functional Domains, and Viral Pathogenesis. *mSystems*

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.16.384743; this version posted November 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

5, (2020).

87. Starr, T. N. *et al.* Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding. *Cell* **182**, 1295-1310.e20 (2020).
88. Rodrigues, J. P. *et al.* Insights on cross-species transmission of SARS-CoV-2 from structural modeling. *bioRxiv* 2020.06.05.136861 (2020). doi:10.1101/2020.06.05.136861
89. Gu, H. *et al.* Adaptation of SARS-CoV-2 in BALB/c mice for testing vaccine efficacy. *Science* **369**, 1603–1607 (2020).
90. Zhu, Y. & Gao, G. ZAP-mediated mRNA degradation. *RNA Biol.* **5**, 65–67 (2008).
91. Guo, X., Ma, J., Sun, J. & Gao, G. The zinc-finger antiviral protein recruits the RNA processing exosome to degrade the target mRNA. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 151–6 (2007).
92. Digard, P., Lee, H. M., Sharp, C., Grey, F. & Gaunt, E. Intra-genome variability in the dinucleotide composition of SARS-CoV-2. *Virus Evol.* **6**, (2020).
93. Xia, X. Extreme Genomic CpG Deficiency in SARS-CoV-2 and Evasion of Host Antiviral Defense. *Mol. Biol. Evol.* **37**, 2699–2705 (2020).
94. Edwards, S. J. L. & Santini, J. M. Anthroponotic risk of SARS-CoV-2, precautionary mitigation, and outbreak management. *The Lancet. Microbe* **1**, e187–e188 (2020).
95. Wang, H., Pipes, L. & Nielsen, R. Synonymous mutations and the molecular evolution of SARS-Cov-2 origins. *bioRxiv* 2020.04.20.052019 (2020). doi:10.1101/2020.04.20.052019
96. Frank, H. K., Enard, D. & Boyd, S. D. Exceptional diversity and selection pressure on SARS-CoV and SARS-CoV-2 host receptor in bats compared to other mammals. *bioRxiv* 2020.04.20.051656 (2020). doi:10.1101/2020.04.20.051656
97. MacLean, O. A. *et al.* Natural selection in the evolution of SARS-CoV-2 in bats, not humans, created a highly capable human pathogen. *bioRxiv* 2020.05.28.122366 (2020). doi:10.1101/2020.05.28.122366
98. Chen, W.-H. *et al.* Yeast-Expressed SARS-CoV Recombinant Receptor-Binding Domain (RBD219-N1) Formulated with Alum Induces Protective Immunity and Reduces Immune Enhancement. *bioRxiv Prepr. Serv. Biol.* (2020). doi:10.1101/2020.05.15.098079
99. Quinlan, B. D. *et al.* The SARS-CoV-2 Receptor-Binding Domain Elicits a Potent

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.16.384743; this version posted November 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Neutralizing Response Without Antibody-Dependent Enhancement. *SSRN Electron. J.* (2020). doi:10.2139/ssrn.3575134

100. Yang, J. *et al.* A vaccine targeting the RBD of the S protein of SARS-CoV-2 induces protective immunity. *Nature* **586**, 572–577 (2020).
101. Krammer, F. SARS-CoV-2 vaccines in development. *Nature* **586**, 516–527 (2020).
102. Baum, A. *et al.* Antibody cocktail to SARS-CoV-2 spike protein prevents rapid mutational escape seen with individual antibodies. *Science* **369**, 1014–1018 (2020).
103. Thomson, E. C. *et al.* The circulating SARS-CoV-2 spike variant N439K maintains fitness while evading antibody-mediated immunity. *bioRxiv* 2020.11.04.355842 (2020). doi:10.1101/2020.11.04.355842
104. Ortuso, F., Mercatelli, D., Guzzi, P. H. & Giorgi, F. M. Structural Genetics of circulating variants affecting the SARS-CoV-2 Spike / human ACE2 complex. *bioRxiv* 2020.09.09.289074 (2020). doi:10.1101/2020.09.09.289074
105. Hadfield, J. *et al.* Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121–4123 (2018).

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.16.384743; this version posted November 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Data Availability

All analysed data (as of 6th November 2020) is available on registration to GISAID with the IDs provided in **Table S1** (mink associated) and **Table S2** (human associated) together with full acknowledgement of contributing and submitting laboratories. In addition, 12 genomes were included from the NCBI nucleotide archive, as also provided in **Table S1** which were also released to GISAID by the time of writing.

Code Availability

Alignment annotation scripts together with code to obtain genome-wide dinucleotide frequencies are available at GitHub repository https://github.com/cednotsed/mink_homoplasies/.

Competing Interests

The authors have no competing interests to declare.

Author Contributions

L.v.D and F.B designed the study. L.v.D, C.CS.T, D.R, C.Ow performed the genomics analysis. S.D.L and C.Or performed structural analyses. T.B. assisted in literature review. L.v.D and F.B. wrote the manuscript with contributions from all authors.

Acknowledgements and Funding

L.v.D and F.B. acknowledge financial support from the Newton Fund UK-China NSFC initiative (grant MR/P007597/1) and the BBSRC (equipment grant BB/R01356X/1). L.v.D. is supported by a UCL Excellence Fellowship. D.R. is supported by a NIHR Precision AMR award. C.Ow. is funded by a NERC-DTP studentship. SDL is funded by a Fundamental Research Grant Scheme from the Ministry Of Higher Education Malaysia (FRGS/1/2020/STG01/UKM/02/3). We wish to thank Paul Ashford for generously sharing his insight in structural biology and the SARS-CoV-2 Twitter community for alerting us to literature characterising the role of Y453F. Finally, we acknowledge the large number of originating and submitting laboratories openly sharing SARS-CoV-2 genome assemblies with the research community.