RESEARCH

Within host and across host SARS-CoV-2 single nucleotide variants (SNVs) and structural variants (SVs) and their implications on SARS-CoV-2 detection, transmission, and evolution

Jane E Doe[1*†] and John RS Smith[1,2]

*Correspondence:
jane.e.doe@cambridge.co.uk
[1]Department of Zoology, Cambridge, Waterloo Road, London, UK
Full list of author information is available at the end of the article
[†]Equal contributor

Abstract

First part title: Text for this section.

Second part title: Text for this section.

Keywords: sample; article; author

Background

On March 11, 2020, the WHO determined that an outbreak of a novel coronavirus (family Coronaviridae, order Nidovirales; 2019-nCoV, or SARS-CoV-2; PMID: 31978945PMID: 31986257) that began in Wuhan, China in late December, 2019, had reached pandemic status. Deep meta-transcriptomic RNA sequencing of bronchoalveolar lavage fluid (BALF) samples from COVID-19 affected patients admitted to and hospitalized in Wuhan in late December 2019 revealed sequence similarity to a SARS-like coronaviruses (87.6-89.1% nucleotides) PMID: 31986257PMID: 32004165. This genus, Betacoronavirus, subgenus Sarbecovirus, was the viral etiologic agent of the previously caused the 2002-2003 SARS outbreak in humans of SARS (e.g., or SARS-CoV-1) PMID: 12690091

Both viruses share the same putative cell surface entry receptor for non-endosomal membrane fusion entry, angiotensin converting enzyme 2 (ACE2) [1, 2, 3], and can co-opt use of the same cell protease (TMPSS2) for spike protein cleavage to facilitate entry into the cell [3]. However, simultaneously blocking TMPRSS2 and the cysteine proteases CATHEPSIN B/L inhibits in vitro host cell entry of SARS-CoV-1 [4] but not SARS-CoV-2, suggesting that one difference rendering much higher rates of human-to-human transmission may be viral genome expansion of motifs which allow for recruitment of additional host proteases in priming SARS-CoV-2 for enhanced cell entry. One proposed example is FURIN, a known S protein cleavage protease involved in MERS-CoV but not SARS-CoV-1 co-facilitation of host cell entry [5, 3, 6]. Whether higher pathogenicity of SARS-CoV-2 might also be related to variation in viral strain expansion of these or other consensus elements related to spike protein cleavage and into the host-cells via higher affinity or efficiency of ACE2 binding is presently unknown, but an area of active investigation [6]. Conversely, it is unknown

whether these or other spike protein proteases and ACE2 are differentially regulated in their cellular expression by virtue of advancing age, sex, co-morbidities such as smoking and hypertension, or during pregnancy.

Despite these genomic and functional similarities between SARS-CoV-1 and CoV-2, SARS and COVID-19 disease are different diseases with disparate case fatality rates and R0 infectivity estimates (PMID: 12690091). In addition, the variability with respect to the range of clinical outcomes among SARS-CoV-2 infected patients is marked: the majority will have mildly symptomatic disease, 10-15% will suffer severe disease, and 3-5% will die (PMID: 32168464).

Diversity of a virus is often characterized by the occurrence of single nucleotide polymorphsisms (SNPs) in the assembled genomes. However, read data offers an additional perspective on the genomic variants present. In this work we focus our analysis on the single nucleotide variants (SNVs) that occur below the consensus level threshold. We compare the SNVs found in read datasets from Mason lab at Cornell University and Baylor College of Medicine to the consensus SNPs identified in the GISAID genomes. This approach allows us to study the viral diversity and variation at greater depth and provides a bridge between the SNP and SNV information contained in the genomes.

OUTLINE

1 Introduction and Background
   (a) Introduce ss positive sense RNA (what does this mean wrt biology?)
   (b) Introduce recent studies showing artifacts in sequencing and error prone positions etc
   (c) Introduce proofreading mechanism and discussion of within host polymorphisms in other viruses

2 Results
   (a) Variability: of SNVs and SV in SARS-CoV-2 (more variable than expected) z@
      i. SNPs from GISAID genomes
      ii. SNVs from public, masonlab, and baylor reads
      iii. SVs from public, masonlab, baylor reads
   (b) Biology: Not just artifacts, fixing/drift/selection/etc z@
      i. comparison of SNVs at AF $< 50\%$ to SNPs in GISAID genomes
      ii. show frequency/pattern of SNVs and SVs on the tree
      iii. show differential expression of genes/ORFs that contain SNVs
      iv. show plot of Ct/viral titer to coverage (Huw)
      v. show patterns over time, February, March, April, May
   (c) Comparison of SARS-CoV-1, MERS, Myl-CoV z@
      i. compare SNV AFs (adjust for GC content?)
      ii. compare SV AFs
      iii. compare synonymous to nonsynonymous mutations
      iv. run structural modeling of non-synonymous mutations?
      v. compare changes found in Spike protein?
   (d) Inferring transmission and bottleneck size z@
      i. Histogram of number of shared SNVs per pair of samples
      ii. Bar plots comparing source/recipient allele frequencies

   iii. Bottleneck size estimates
   iv. Integration of info back on to phylogenetic tree
3 Discussion
   (a) Interpreting AF from RNA-seq variant calls (coverage means frequency and gene expression!)
   (b) Limitations of inferring transmission and bottlenecks with unknown metadata/inappropriate study design (no known transmission pairs, hard to fully evaluate)
   (c) Effect of SNVs on RT-PCR based detection methods
   (d) SVs and implication on biology (cite paper for 29bp deletion etc)
   (e) anything else?
4 Conclusion
   (a) Summarize main results
   (b) Request for future studies designed to evaluate utility of LFVs for transmission analyses
5 Methods
   (a) Read QC (trimming etc)
   (b) Read mapping (bwa-mem, parameters)
   (c) SNV calling (lofreq, parameters)
   (d) SNV annotation (SNPeff, parameters)
   (e) Count table generation
   (f) DE (limma-voom, parameters)
   (g) SV calling (manta, parameters)
   (h) MSA (parsnp, parameters)
   (i) Tree constricution (raxml, parameters)
   (j) Transmission and bottleneck calculuations (equation, paper to cite)
   (k) anything else?

Results

We have analyzed variants occurring in three datasets: GISAID public data, Baylor College of Medicine sequencing read sets for 11 samples collected in Houston, Texas, and Mason lab sequencing read sets for 140 samples collected in New York City. Variants detected in GISAID data are consensus SNPs present in the final assembled genomes. Variants detected in the read sets range from ultra low frequency variants (allele frequency below 5%) to common variants (allele frequency $> 50\%$). As can be seen from Figures 3, 4 the distribution of variants with respect to the observed allele frequency is bimodal with peaks occurring near the ends of available values spectrum. Thus, we can naturally divide SNVs into three sub-categories: ultra-low frequency (AF between 2% and 5%), low-frequency (AF between 5% and 30%), and high-frequency (AF $> 95\%$). We note that the later category mostly overlaps with the consensus level SNPs, and therefore does not present interest from the perspective of low-frequency variant analysis, as this variant data is already captures by SNPs in the assembled genomes. Thus, for most of our analyses we focus on the ultra-low frequency and low frequency SNVs.

Viral Transmission

In addition to phylogenetic inference, genomic analyses of pathogenic microbes are increasingly being performed to attempt to infer direct person to person transmission events [7, 8, 9]. While phylogenies and transmission trees, made up of links between all the inferred transmission events in a data set, are fundamentally linked to one another, inferring a phylogeny does not guarantee simultaneous inference of a transmission tree [10]. One key difference between phylogenetic inference and inference of transmission events is in the data used for inference. Phylogenetic inference is commonly performed on a set of consensus level genomes. Meanwhile, inferred transmission events typically rely on extra information such as metadata linking viral genomes together, for instance as having come from two individuals from the same household, and shared low frequency variation inside viral samples.

In the extreme case, definitively labelling viral samples involved in direct transmission events relies on intensive metadata collection. As discussed in [11], which includes a review of previous transmission analyses as well as related methodologies, one study on transmission dynamics involved barcoding influenza to be able to recover individual transmission events through animals [12]. In humans, reporting the ability to directly infer transmission events has relied on data such as hospitalization data [7] and shared household information [13].

Alongside metadata, sequencing data sets can provide an in depth look at viral transmission dynamics between transmission pairs. When viral transmissions occur, under the assumption of a single discrete transmission event, a founding population of virions will travel from the transmission donor to the transmission recipient. This will include a major bottle necking event, as this founder population will be much smaller than the size of the full population inside the donor. If this founder population is large enough, also referred to as a 'wide' or 'loose' transmission bottleneck, variants inside the donor population have the potential to be transmitted from the donor to the recipient. The link between these shared donor and recipient variants and the size of the transmission bottleneck, or founder population size, was thoroughly explored in [14]. Essentially, the wider the transmission bottleneck, or the larger the founder population size, the more shared variants we expect to see between the donor and recipient genomic data.

For SARS-CoV-2, the recent study of [15] reported on shared variants between 18 samples, and how these shared variants could be used to further refine a phylogenetic tree based on consensus level SARS-CoV-2 sequences. They also discuss the potential of these shared variants to reveal possible direct or indirect transmission events. In this study, we similarly investigate links between pairs of samples to examine the potential for revealing possible transmission events directly from sequencing data. Given a lack of confirmatory metadata for establishing definitive spacial and temporal links between samples, our analyses are strictly exploratory. Similarly, given previously mentioned factors that must be taken into account [HERE LOOK OVER WHATS IN THE PAPER THEN MENTION THE HUGE LIST OF CONFOUNDERS OR SAY WE HAVE IN DISCUSSION THEN PUT IN DISCUSSION], we cannot ignore the potential for technical artifacts, in fact we report on strong shared variant signal coming from site 29871 which is likely erroneous due to its location at the end of the genome.

Exploratory analysis of potential transmission events

To begin our analyses, we calculated the number of sites where a low frequency variant in one sample also had a variant present in another sample. Fig. **??** shows the distribution of these pairs and the number of shared variants between them. Here, for each pair, we arbitrarily assign one sample as the donor and one as the recipient, including both such pairs, and we narrow down the donor alleles to only include those with allele frequency (AF) between 0.03 and 0.5. We consider a site to be shared if the recipient also has that same nucleotide variant present at any frequency. [IN METHODS NEED TO DISCUSS THE INTRICACIES OF THIS; WE EXCLUDE 0s EVEN THO LOW AFs COULD GO TO ZERO AND HERE WE DO INCLUDE THINGS LIKE 0.6 and 1.0 CUZ SOMETHING WITH LIKE 0.45 AF COULD LATER BECOME SOMETHING LIKE 0.6 etc] We show these results on the raw data from .vcf files [METHODS HAS DISCUSSION ON HOW VCFS WERE GENERATED??] as well as on the same data but after applying masking to sites near the ends of the genome. Before masking, we see in Fig. **??** that most pairs have 0 to 3 shared variants, with about 150 pairs having 4 or more shared variants. After masking, these numbers drop substantially by reducing likely noise from the variant calls, and we see most pairs sharing 0 to 2 variants, with about 100 pairs having 3 or more shared variant sites.

To begin looking at possible transmission dynamics, we looked at pairs of samples that either shared many variant sites in common, or that shared fewer sites in common while also having few variant sites in the donor that were not present in the recipient. For instance, in Fig. **??**A we show estimated bottleneck sizes (number of estimated virions in the founder population) for each pair of samples sharing 5 or more variants from the raw variants data. In this style of analysis, we run the bottleneck size estimation method from [11] while treating each pair of samples as a putative transmission pair. In Fig. **??**B we visualize the donor and recipient allele frequencies for the alleles from the donor between 0.03 and 0.5 allele frequency. Similarly, we show the same results in Fig. **??**A and **??**B for pairs of samples sharing 4 variant sites with less than 10 variant sites in the donor not being present in the recipient. Here we see much greater signal for shared variants that could have the potential to be generated by transmission pairs with a large bottleneck. But, upon examining the actual sites that seem to be shared in the pairs with the largest inferred bottlenecks, we that most pairs have large shared allele frequencies at site 29871.

In the recent report of (http://virological.org/t/issues-with-sars-cov-2-sequencing-data/473), many sites were examined that showed extensive homoplasy. It was found that most of these sites showed strong signals of being the result of technical artifacts, such as being present only at the ends of reads or occurring after poly-X nucleotide runs. Based on quality control recommendations, we subsequently examined pairs after masking sites 1-55 and 29804-29903. After masking, we examined pairs sharing 4 or more variant sites in Fig. **??**, pairs sharing 2 or 3 variants sites with less than 10 variant sites present only in the donor in Fig. **??**, and pairs sharing only 1 variant site with an allele frequency greater than 0.1 and with any non shared site having less than 0.1 allele frequency in Fig. [FIG REF??]. When examining these pairs, we see that the majority of pairs show no signal for an inferred large bottleneck. This is to be expected given that the majority of pairs in

a large batch of sequenced SARS-CoV-2 samples are not expected to have been direct or indirect transmissions. Despite this, we do still see striking outlier pairs with large inferred bottlenecks, where we observe significant conservation of high allele frequency within host diversity between sample pairs.

A first example pair with striking similarity between low frequency variants within hosts is the pair [CANT READ NAMES] that appears in both Fig. **??**B and Fig. **??**B. In Fig. **??**B we also see this pair reported with reversed donor and recipient samples. [SAY SOME STUFF ABOUT IT. pcr negs also?]

[THEN JUST A BIT MORE DISCUSSION ON THE FINAL FIGS AND INTERESTING LOOKING PAIRS. MENTION ANY THAT WERE MENTIONED IN THAT QC ARTICLE AS POTENTIAL TECHNICAL ARTIFACTS]

Finally, do we want a final fig grouping discussed maybe interesting pairs as well as manually pulling out a few pairs not included in the top ones that might look kind of interesting?

Discussion

Conclusion

Methods

Read QC and mapping

SNV calling and annotation

Multiple sequence alignment

Phylogenetic tree construction

Transmission Analyses

In this study, we applied *BB bottleneck* software to estimate SARS-COVID2 bottleneck sizes, that is, the founding viral population size in the recipient host[11]. To quantify the bottleneck sizes, *BB bottleneck* uses beta-binomial sampling technique to analyze shared variants in direct transmission pairs. Comparing to previous approaches, beta-binomial sampling takes variant calling thresholds and differences between variant frequencies at the founding time and frequencies at the sampling time into consideration.

BB bottleneck has two modes: APPROX mode and EXACT mode. APPROX mode takes the donor and recipient shared variants' allele frequencies (AFs) as input and assumes the read coverage is infinity. For EXACT mode, in addition to the variant frequency information, it also requires read number at recipient variant sites and the total number of recipient reads. Although the EXACT mode provides more accurate bottleneck size estimation, it is much more computationally intensive than the APPROX mode. Due to the time and resource constraint, we employ APPROX mode to approximate SARS-COVID2 the effective founding population sizes.

Commonly, to recover the viral transmission bottleneck sizes in real world, the first step is to identify direct transmission pairs through corresponding epidemiological or physiological data. However, we do not have any metadata related to direct transmissions. Thus, we infer transmission pairs using solely sequence variant calling results. We postulate that, comparing to any other random genome pairs, a direct transmission pair share relatively more number of variants. To find pairs with a large number of shared variants, we first count the number of genome-wide variants

shared between any two samples. We only exam variants with allele frequencies (AFs) ranging from 0.03 to 0.5, also referred as low frequency variants (LFVs), inside donor samples. We use 0.03 as the LFV calling cutoff in order to exclude potential artificial mutations inside the donor sequences. On the other hand, since the variant frequencies in recipient samples partially rely on stochastic pathogen replication process in the early infection, we take all variants (with any AFs) into account. If a recipient has a variant at position i which has the same nucleotide as its donor's LFV does at the same position, we say the donor and the recipient share a variant at site i and record their corresponding AFs as input for the *BB bottleneck* APPROX mode. If the recipient does not have the same base as the donor does or the recipient does not have any variants at position i while mapping to the reference sequence, we call this variant in the donor sequence as unique donor variant and assign the recipient a 0.0 AF at that position.

To narrow down the number of transmission pairs, we select the pairs on the top two levels (the most number of shared variants) as putative transmission pairs. While considering the whole viral genomes, pairs sharing at least 5 variants are classified as the first level pairs. For the second level, we include pairs with 4 shared variants. Since 148 pairs belong to the second level, we further narrow down the group by only considering pairs whose donors have no more than 9 unique variants.

In [http://virological.org/t/issues-with-sars-cov-2-sequencing-data/473] study, authors suspect that mutations at the beginning and the end of the SARS-Cov-2 genomes most likely come from sample preparation, sequencing or mapping errors instead of evolutionary events. Therefore, they propose to mask the ends of the viral genomes (position 1-55 and 29804-29903) before doing downstream analysis. We take their advice and identify transmission pairs using masked genomes. Specifically, for masked genome, we only look at the variants from position 56 to 29803, and utilize pairs with at least 3 shared variants in our bottleneck size approximations.

Competing interests
The authors declare that they have no competing interests.

Author's contributions
Text for this section ...

Acknowledgements
Text for this section ...

Author details
[1]Department of Zoology, Cambridge, Waterloo Road, London, UK. [2]Marine Ecology Department, Institute of Marine Sciences Kiel, Düsternbrooker Weg 20, 24105 Kiel, Germany.

References

1. Lu, R., Zhao, X., Li, J., Niu, P., Yang, B., Wu, H., Wang, W., Song, H., Huang, B., Zhu, N., *et al.*: Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. The Lancet **395**(10224), 565–574 (2020)
2. Yan, R., Zhang, Y., Li, Y., Xia, L., Guo, Y., Zhou, Q.: Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2. Science **367**(6485), 1444–1448 (2020)
3. Hoffmann, M., Kleine-Weber, H., Krüger, N., Mueller, M.A., Drosten, C., Pöhlmann, S.: The novel coronavirus 2019 (2019-nCoV) uses the SARS-coronavirus receptor ACE2 and the cellular protease TMPRSS2 for entry into target cells. BioRxiv (2020)
4. Kawase, M., Shirato, K., van der Hoek, L., Taguchi, F., Matsuyama, S.: Simultaneous treatment of human bronchial epithelial cells with serine and cysteine protease inhibitors prevents severe acute respiratory syndrome coronavirus entry. Journal of virology **86**(12), 6537–6545 (2012)
5. Burkard, C., Verheije, M.H., Haagmans, B.L., van Kuppeveld, F.J., Rottier, P.J., Bosch, B.-J., de Haan, C.A.: ATP1A1-mediated Src signaling inhibits coronavirus entry into host cells. Journal of virology **89**(8), 4434–4448 (2015)

6. Lukassen, S., Chua, R.L., Trefzer, T., Kahn, N.C., Schneider, M.A., Muley, T., Winter, H., Meister, M., Veith, C., Boots, A.W., et al.: SARS-CoV-2 receptor ACE2 and TMPRSS2 are primarily expressed in bronchial transient secretory cells. The EMBO journal (2020)
7. Snitkin, E.S., Zelazny, A.M., Thomas, P.J., Stock, F., Henderson, D.K., Palmore, T.N., Segre, J.A., Program, N.C.S., *et al.*: Tracking a hospital outbreak of carbapenem-resistant klebsiella pneumoniae with whole-genome sequencing. Science translational medicine **4**(148), 148–116148116 (2012)
8. Hall, M., Woolhouse, M., Rambaut, A.: Using genomics data to reconstruct transmission trees during disease outbreaks. Revue scientifique et technique (International Office of Epizootics) **35**(1), 287 (2016)
9. Hall, M.D., Colijn, C.: Transmission trees on a known pathogen phylogeny: enumeration and sampling. Molecular biology and evolution **36**(6), 1333–1343 (2019)
10. Sánchez-Pacheco, S.J., Kong, S., Pulido-Santacruz, P., Murphy, R.W., Kubatko, L.: Median-joining network analysis of sars-cov-2 genomes is neither phylogenetic nor evolutionary. Proceedings of the National Academy of Sciences (2020). doi:10.1073/pnas.2007062117. https://www.pnas.org/content/early/2020/05/06/2007062117.full.pdf
11. Leonard, A.S., Weissman, D.B., Greenbaum, B., Ghedin, E., Koelle, K.: Transmission bottleneck size estimation from pathogen deep-sequencing data, with an application to human influenza a virus. Journal of virology **91**(14), 00171–17 (2017)
12. Varble, A., Albrecht, R.A., Backes, S., Crumiller, M., Bouvier, N.M., Sachs, D., García-Sastre, A., *et al.*: Influenza a virus transmission bottlenecks are defined by infection route and recipient host. Cell host & microbe **16**(5), 691–700 (2014)
13. Poon, L.L., Song, T., Rosenfeld, R., Lin, X., Rogers, M.B., Zhou, B., Sebra, R., Halpin, R.A., Guan, Y., Twaddle, A., *et al.*: Quantifying influenza virus diversity and transmission in humans. Nature genetics **48**(2), 195 (2016)
14. Worby, C.J., Lipsitch, M., Hanage, W.P.: Shared genomic variants: identification of transmission routes using pathogen deep-sequence data. American journal of epidemiology **186**(10), 1209–1216 (2017)
15. Ramazzotti, D., Angaroni, F., Maspero, D., Gambacorti-Passerini, C., Antoniotti, M., Graudenzi, A., Piazza, R.: Characterization of intra-host sars-cov-2 variants improves phylogenomic reconstruction and may reveal functionally convergent mutations. bioRxiv (2020)

Figures

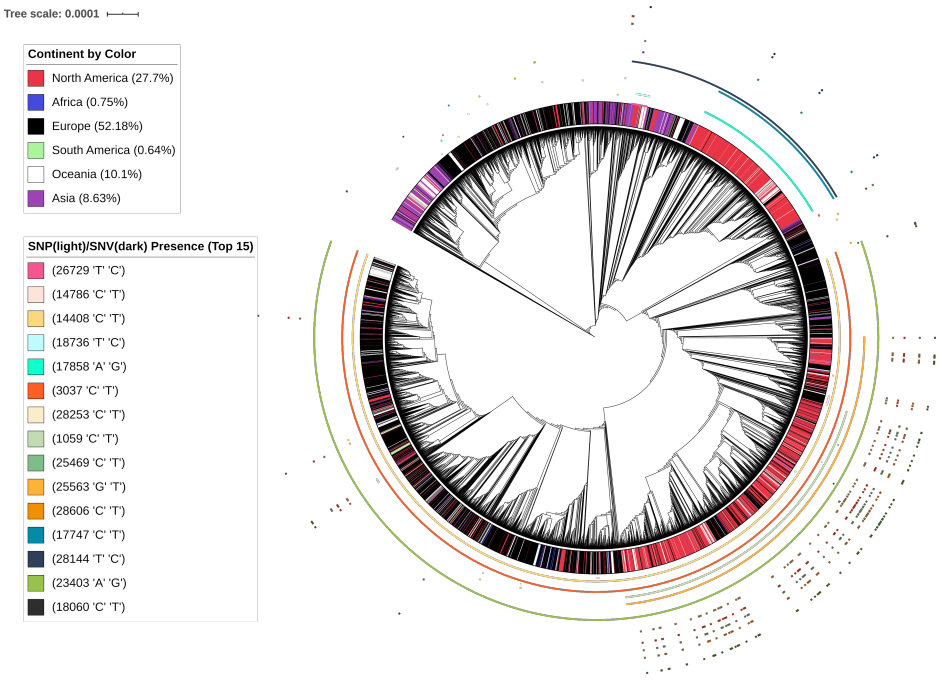

Figure 1 This figure shows the phylogenetic tree obtained from 6,928 GISAID genomes with addition of simulated genomes that include SNVs from the read sets. The inner most circle captures the continent corresponding to the GISAID metadata for the given genome. The outer circles represent SNPs (in lighter colors) and SNVs (in darker colors) for the assembled or simulated variant genomes respectively. We note formation of two clades corresponding to the SNP co-occurrence pattern. We also note that we do not currently have read data from Europe, and as a result most of the simulated variant genomes cluster with North America based data.

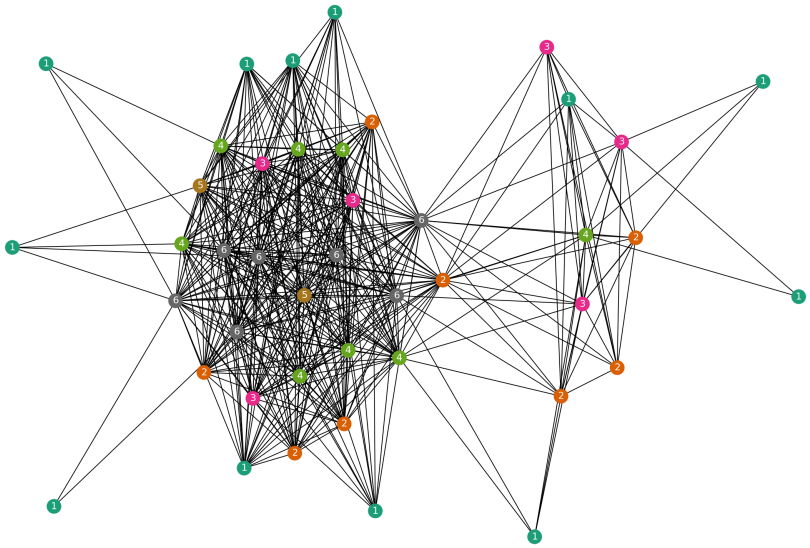

Figure 2 This figure shows the overlap between different variant configurations encountered in GISAID, Mason and Baylor samples. Each nodes corresponds to a unique combination of variants, with the number representing the total number of co-occuring variants. Edges indicate an overlap by at least one variant between the two configurations. We can see that while the entire graph is connected, there is a small subgraph on the right that is only loosely connected to the main set of variant configurations.

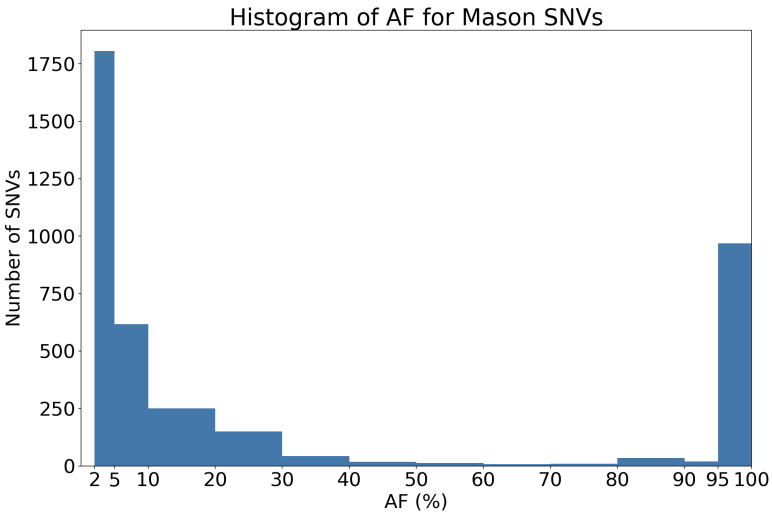

Figure 3 This histogram captures the distribution of allele frequencies of variants in the Mason lab samples. Note that we have filtered out the variants present below 2% allele frequency threshold. The distribution of allele frequencies is bimodal with peaks near the left (2-5%) and right (95-100%) ends.

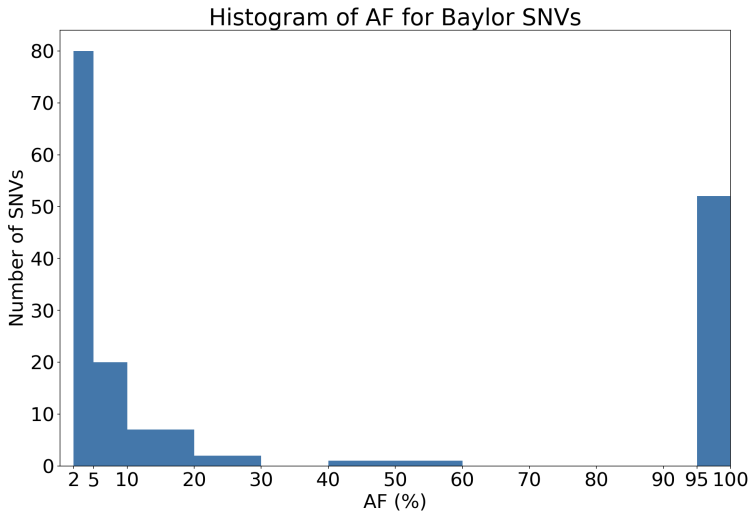

Figure 4 This histogram captures the distribution of allele frequencies of variants in the Baylor samples. Note that we have filtered out the variants present below 2% allele frequency threshold. The distribution of allele frequencies is bimodal with peaks near the left (2-5%) and right (95-100%) ends. We see sharper peaks with lighter tails in this figure as compared to the Mason data. However, this can be explained by the lower number of samples and total number of variants in the Baylor data.

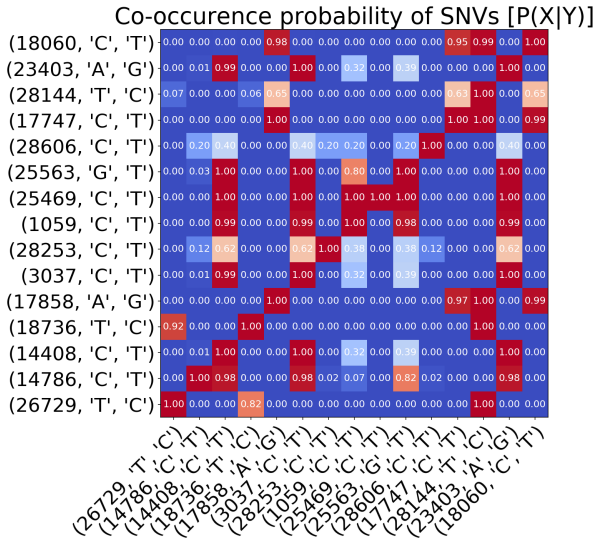

Figure 5 The figure above captures the SNV co-occurrences in the available genomic data. We distinguish two groups of variants that tend to happen simultaneously, as shown by their empirical conditional occurrence probabilities.

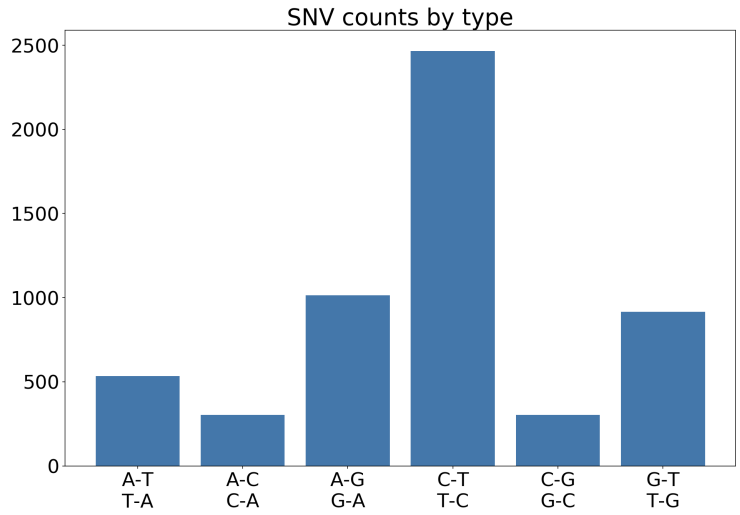

Figure 6 This figure shows the distribution of SNPs and SNVs by the type of the nucleic acid change. We note that predominantly C>T and T>C changes occur with G>C and C>G changes being the least common.

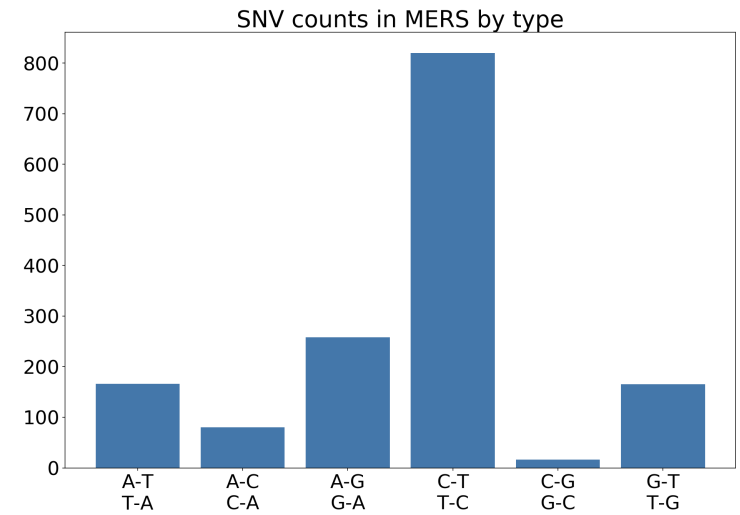

Figure 7 This figure shows the distribution of SNPs and SNVs in the MERS virus by the type of the nucleic acid change. We note that predominantly C>T and T>C changes occur with G>C and C>G changes being the least common. This pattern is analogous to the one we observe for the SARS-CoV-2.

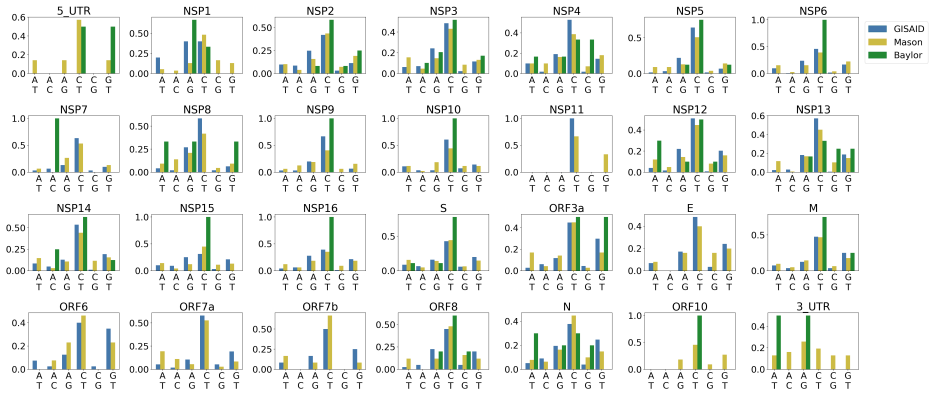

Figure 8 This figure shows the distribution of SNV types within each gene of SARS-CoV-2 colored by the dataset. We note that the overall patterns are consistent between all three datasets. The GISAID data is missing variants in the genes located at the ends of the genome, which can be explained with the assembly issues. We also note that in some cases we only observe variants in Mason and GISAID data, e.g. ORF6, ORF7a-b, NSP11, E genes which can be in part due to the smaller sample size of the Baylor dataset.

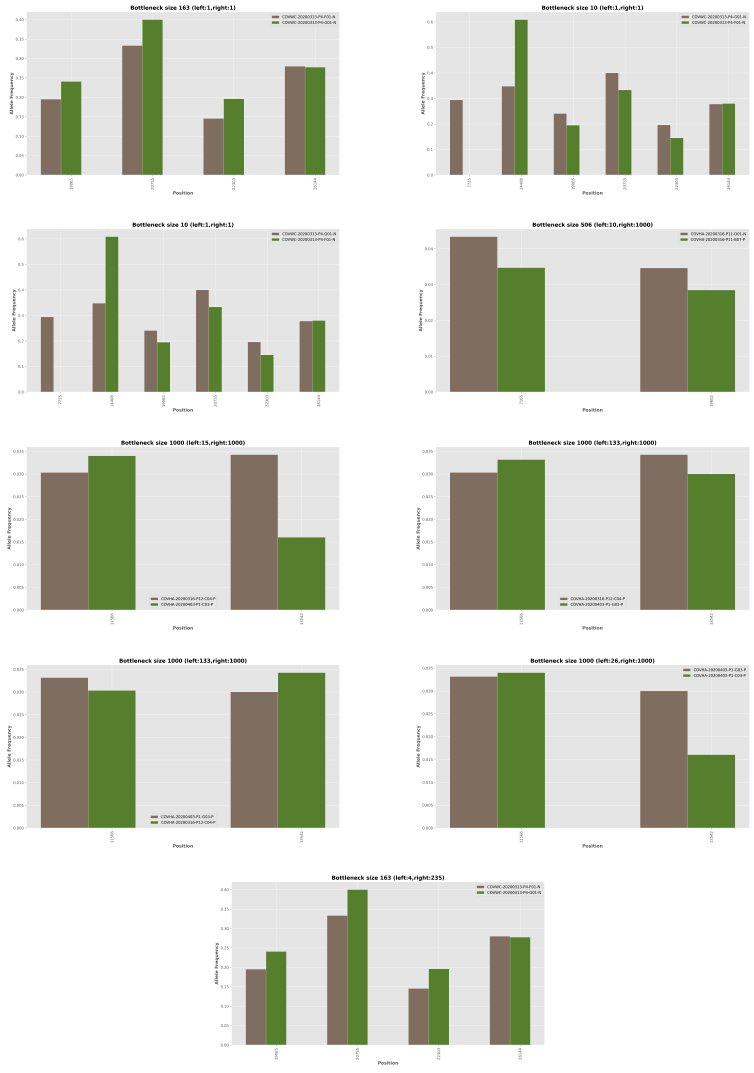

Figure 9 Interesting pairs to specifically discuss?