Name: Edmond Brewer
Date: 4/30/2020
Class: BMMB 551

A Comparative Analysis of Common Coronaviruses

The year 2020 has so far been singularly defined by the COVID-19 pandemic that has ravaged populations on every populated continent. This respiratory illness is caused by the SARS-CoV-2 virus, a close relative of the original SARS virus that originated in 2002. Both viruses are part of the broad family of coronaviruses. When viewed under an electron microscope, coronaviruses have a spherical morphology with a "halo" of spike proteins that look similar to a solar corona viewed during an eclipse. The coronavirus family has 7 known members that can infect humans: HKU1, NL63, OC43, 229E, SARS, SARS-CoV-2, and MERS. The first 4 coronaviruses are mild, and continually circulate in the human population. They usually recur in the winter months to cause 10-15% of the world's common colds. The last 3 coronaviruses have potential to have much more severe symptoms, however, and have usually been found in acute outbreaks so far, although MERS has seemed to be circulating in the Middle East continually since 2012. As of April 30th, 2020, COVID-19 has 3.25 million reported cases worldwide, 231,415 deaths, and 1.05 million cases in the US alone[1]. By comparison, the SARS outbreak of 2002-2004 had only 8,029 worldwide cases, with only 774 deaths[2]. The MERS outbreak that is still ongoing has had 2,494 cases since it started in 2012, and 858 deaths[3]. Undoubtably, government responses to each outbreak have shaped the respective case and death counts, and an ineffective and delayed response can make an outbreak more devastating that it has to be, and vice versa. But, each virus also has epidemiological idiosyncrasies that have an effect on its spread, such as how long the disease takes to kill the patient, what percentage of patients die, the method and ease that the virus spreads, for example via bodily fluids, aerosols, and surviving for long periods of time on inanimate, high-touch objects such as doorknobs. These epidemiological traits have biological and genetic origins, and the questions that I am attempting to answer in my project are what gene and protein level differences exist between these select coronaviruses, and to what degree I can explain the epidemiological differentiation from my analyses.

The first part of my analysis utilizes the publicly available genomes for all 7 human coronaviruses, all of which are available at the NCBI Genome Database[4]. I used a bash script (see Code section) to download the genomes from the NCBI, perform a multiple sequence alignment using the *clustal omega*[5] package to align them to each other. I then loaded the .maf file that was created into SnapGene[6] to create a consensus sequence FASTA file. Unfortunately, the *clustal omega* package did not provide a built-in stats method, and I was unable to find a way to automate finding the composition of the multiple alignment, so I had to count the occurrences in a text editor. There were 25,423 total bases/ characters, and of those, 6363 were gaps, 14,703 were N, meaning any base, and 4357 were A, G, C, or T. Taking the consensus bases and dividing them by the total gives us a minimum sequence similarity of 17.1%. That doesn't tell the whole story, however, as a lot of those Ns are mutations with a 50%-50% split between two bases. Furthermore, this alignment technique found the consensus sequence among all genomes, whereas for my purposes, I am interested in how these other genomes differ from SARS-Cov-2. Nonetheless, it was clear that aligning each individual genome to SARS-Cov-2 is necessary to understand the genetic diversity between these viruses. I did that using *stretcher*, a global pairwise Needleman-Wunsch aligner (*Table 1*). The results are actually very interesting, in that the only coronavirus that SARS-Cov-2 has greater than 56% similarity with is SARS, hence the name SARS-Cov-2. It had 79.4% sequence identity and similarity, with only 1.2% gaps. I was surprised that the alignment with MERS was no better than the other, more mild coronaviruses. This could be due to the fact that MERS spreads through camels, and is recurring in the Middle East mainly, whereas SARS has similar zoonotic origins as SARS-Cov-2, as well as also originating in China. Once you account for the other relevant epidemiological information, some interesting and contradicting trends emerge.

SARS-Cov-2 and SARS are part of the sarbecovirus subgenus of the Betacoronavirus genus, but SARS-Cov-2 is no more related to other Betacoronaviruses than it is to the Alphacoronaviruses. Neither the original location of the human infection, nor the zoonotic origin also doesn't seem to have any consistent effect on how genetically similar SARS-Cov-2 is to other coronaviruses. Most surprisingly, the theorized entry receptor for viral entry and infection of cells doesn't have an effect, because NL63 uses the same receptor as SARS and SARS-Cov-2, yet like the rest of the coronaviruses besides SARS, it has around 53-55% similarity. Now, given the high mutation rate traditional to viruses due to reproduction

rate and selective pressures, the lack of similarity between viruses in a family makes sense. These findings correlate with the determination by the Coronaviridae Study Group of the International Committee on Taxonomy of Viruses that SARS-Cov-2 is in fact a strain of SARS, and not a separate species[12].

Table 1: Sequence Similarity and Biological Information

Virus aligned to SARS-Cov-2	Sequence Identity %	Sequence Similarity %	Gap %	Zoonotic origin[9]	Original Location	Genus[8]	Entry Receptor[9]
SARS-Cov-2	100%	100%	0%	Potentially Bat or Pangolin[11]	China	Beta Sarbecovirus	ACE2[10]
SARS	79.4%	79.4%	1.2%	Palm Civet and Bat	China	Beta Sarbecovirus	ACE2
MERS	55.4%	55.4%	8.0%	Bat and Camel	Middle East	Beta Merbecovirus	DPP4
HKU1	54.8%	54.8%	8.1%	Mice	Hong Kong	Beta Embecovirus	9-O-Acetylated sialic acid
OC43	54.2%	54.2%	9.5%	Cattle	Unknown	Beta Embecovirus	9-O-Acetylated sialic acid
NL63	53.1%	53.1%	11.2%	Palm Civet and Bat	Netherlands	Alpha Setracovirus	ACE2
229E	52.5%	52.5%	12.1%	Bat	Unknown	Alpha Duvinacovirus	CD13

A raw percentage of genetic similarity doesn't give the full depth of biological relationships between coronaviruses. A bigger determining factor of epidemiological outcomes is the protein composition in each viral genome. Thankfully, viral genomes are much smaller than plants or animals, and as such are much easier to work with and catalog their protein content. There isn't an annotated SARS genome in the UCSC Genome Browser database, and I could not get the blastp program to accurately find SARS-Cov-2 proteins in the SARS genome, so I had to locally align the SARS-Cov-2 cDNA sequences to the SARS genome, using *water* and report on alignment statistics as a proxy for whether there was a SARS ortholog for SARS-Cov-2 proteins. The proteins and regions that I chose were determined by a variety of factors. I looked at the MultiZ alignment of SARS to SARS-Cov-2 on the UCSC Genome Browser, and looked at proteins where the alignment suggested a lot of non-synonymous mutations (see Figure 1, select proteins are outlined in red). I also examined proteins and regions that were relevant to key viral properties, such as spike proteins, the ACE2-binding region, and the receptor-

binding domain. The overall similarity between SARS and SARS-Cov-2 is 79.4%, which means that regions that have less similarity than that could be candidates for positive selection.

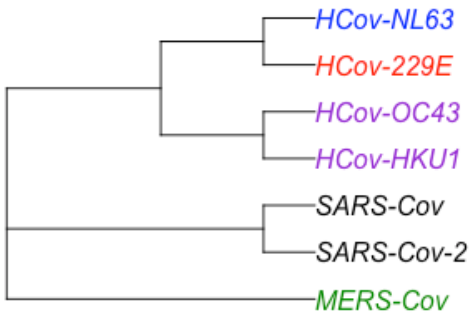
Table 2: Key SARS-Cov-2 Protein Sequence alignment to SARS Genome

SARS-Cov-2 Protein aligned to SARS Genome	Sequence Identity %	Sequence Similarity % (sorted ascending)	Gap %
ACE2-binding domain (in Receptor-binding domain)	63.8	63.8	12.7
Spike Protein Subunit 1	64.9	64.9	14.8
Non-structural protein 2	68.4	68.4	10.9
Receptor-binding domain	72.1	72.1	6.2
Non-structural protein 3	73.4	73.4	6.0
Replicase polyprotein 1A	75.9	75.9	5.1
Replicase polyprotein 1AB	79.8	79.8	3.4
Spike Protein Subunit 2	81.2	81.2	1.2
M-protein	85.8	85.8	1.3
N-protein	88.6	88.6	1.3
RNA-binding domain (of N-protein	92.0	92.0	0.0
E-protein	93.4	93.4	1.3

Of the 11 proteins and regions that I looked at in *Table 2*, 6 had a lower than average similarity. The region with the lowest similarity was the ACE2-binding domain, which is a part of the spike protein subunit 1, at 63.8% similarity. The second lowest was the whole spike protein subunit 1 region, at 64.9% sequence similarity. The research behind differential binding affinity between SARS-Cov-2 and SARS seems to confirm this hypothesis of positive selection, as Wrapp et al found that the binding affinity was 10-20 fold higher in SARS-Cov-2[14]. The authors offer that the greater ease at which SARS-Cov-2 binds to ACE2 could account for the greater ease at which it has spread in the population. Another region with high variation is the non-structural protein 2 region, which had 68.4% similarity. According to UniProt, this protein "May play a role in the modulation of host cell survival signaling pathway by interacting with host PHB and PHB2"[15]. A paper published in February in the *Journal of Medical Virology* by Angiletti, et al, suggests that there is significant positive selection in *nsp2* and *nsp3* (which I found to have 74.3%

similarity with SARS, so below average) that could explain the increase in observed contagiousness. On the other end of the spectrum, the N-protein, or nucleocapsid protein, which is the 4th of 4 structural protein that is responsible for containing the viral genome. It has had remarkably little mutation compared to SARS, at 88.6% similarity, which could make it a potential vaccine candidate since it is highly conserved and likely to remain unchanged if SARS-Cov-2 perennial circulates in the population. The RNA-binding domain (of the N-protein) has the second highest similarity, at 92.0%. This domain binds to the coronavirus RNA, creating a ribonucleoprotein complex, which is essential for viral replication. This makes it a potential target for RNA-binding inhibitors, as this paper by Sarma, et al, describes[18]. The domain with the highest sequence similarity is the E, or envelope protein. This protein has a 93.4% similarity, but I was unable to find any research on its applicability to a treatment or a vaccine. In total, there is significant variation of sequence similarity too SARS among proteins and domains in the SARS-Cov-2 genome, and it begins to tell a story about how this virus evolved. It certainly will be fantastic fodder for research, both for public health purposes as well as for academic research.

Figure 1: Phylogenetic tree of 7 human-infecting coronaviruses



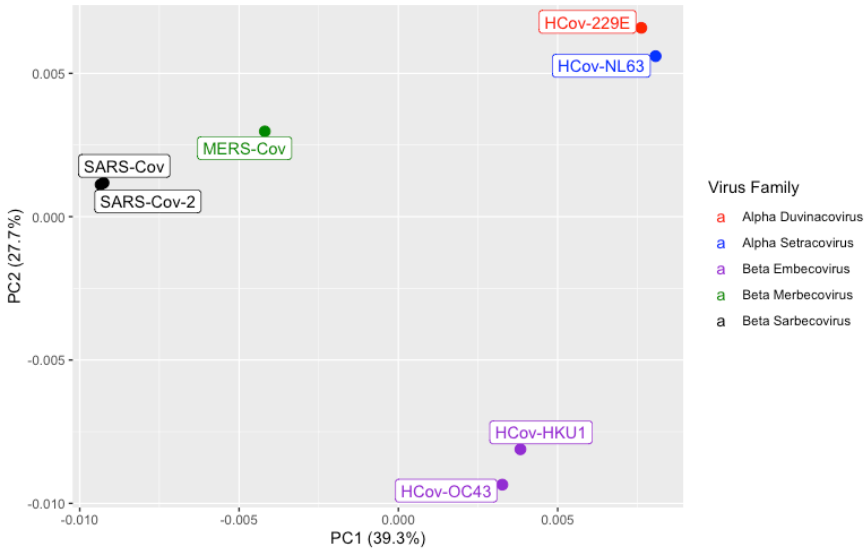
Having looked at the differences between SARS-Cov-2 and SARS in terms of protein content and homology, I thought it would be instructive to perform several cluster analyses on the genomes to get a broad summary of the differences and similarities between human-infecting coronaviruses. My first analysis was attempting to recreate a phylogeny of the 7 viruses that I looked at. I knew the general taxonomy of these viruses, but how they looked in a phylogeny that excluded all other viruses sounded intriguing. To get me started, I found two fantastic articles online, one by Drew Kerkhoff[22], and one by Dr. T.F. Khang[23], that provided me with code that I then mashed together for my purposes. I used the multiple alignment file (.maf) that I created earlier, using *clustal omega*, as the read-in file for the read.phyDat function in the *phangorn*[19] package, a phylogeny creation package in R[20]. This converted the sequence data into their proprietary phyDat format, which was then plugged into their genetic distance function, dist.ml. According to the documentation, this function computes “pairwise distances for an

object of class phyDat. dist.ml uses DNA / AA sequences to compute distances under different substitution models." I used their NJ function, which performs the neighbor-joining tree estimation, on my dist.ml object, and plotted the produced phylogeny with plot.phylo, from the *ape* package[21]. The results are in Figure 1. I colored them according to their actual genus and subgenus from *Table 1*, so for example Beta Sarbecovirus is black, and Beta Merbecovirus is green. It seems that my phylogeny reconfirmed the official taxonomy, and It was interesting to find that HCov-NL63 and HCov-229E were close enough related to be in the same fork, while MERS-Cov was distinct enough to be on its own branch. Another surprising finding was that even though the purple block of HCov-OC43 and HCov-HKU1 are part of the Beta Embecovirus genus and the blue/red block of HCov-NL63 and HCov-229E are part of the Alpha Setracovirus and Alpha Duvinacovirus genus', respectively, they split from a branch earlier than the SARS and MERS blocks do, even though those two are both Betacoronaviruses. Both Alpha- and Betacoronaviruses descend from bat viral lines, so maybe the differences between the two aren't as pronounced as other group comparisons. Another possibility is that since viral mutation occurs so rapidly, the resemblance of viruses in the same genus to each other isn't as high as in eukaryotes. This would mean that the grouping of inter-genus viruses together is born of a lack of similarity between intra-genus viruses, rather than extraordinary inter-genus similarity.

The next analysis that I did was a principle components analysis, or PCA. According to the prcomp function from the *stats*[20] package that I used, "[t]he calculation is done by a singular value decomposition of the (centered and possibly scaled) data matrix, not by using eigen on the covariance matrix. This is generally the preferred method for numerical accuracy." My linear algebra knowledge is not sufficient to explain how that works, but prcomp is a standard PCA function and it produced intuitive results that actually mirror the official taxonomy more closely than my phylogeny. The prcomp function gave 7 principal components, but since the 6th and 7th components only accounted for 2.3% and 0% of the variance, respectively, I only looked at pairwise comparison of the first 5 components (see Table S1 for the summary statistics for the PCA). In the pairwise comparison of PC1 through PC5, PC1 vs PC2 seemed to have the best representation of the expected relationship between the data, and combined, PC1 and PC2 accounted for 67% of variation. This gave me confidence that the first two components best

represented the data. The other pairwise comparisons can be seen in Figure S2 for reference. Looking at the PC1 vs PC2 data, several trends are clear. First, the SARS and SARS-Cov-2 genomes are so closely related in this analysis as to be indistinguishable on the graph. This shows how, relative to the other human-infecting coronaviruses, they are incredibly similar. The second takeaway is that the severe-disease causing viruses are segregated from the more mild viruses, which seems to be a function of PC1 on the x-axis, as their Y-axis positions don't show a clear distinction between severe and mild. The third takeaway is that the Beta Embecoviruses, in purple, are very distinct from the Alphacoronaviruses, in red and blue. This seems to be a function of PC2 on the y-axis, as on the x-axis they are close together. The final takeaway is that the PCA plot differs from both my recreated phylogeny, and the official taxonomy in important ways. Since the SARS/SARS-Cov-2, MERS, and HKU1/OC43 subgenus' are all Betacoronaviruses, it would make sense that they would be closer together than the Alphacoronaviruses. In looking at the plot, however, it is clear they are not. Whether this is a measure of the vast genetic diversity of the viruses from each other or a limitation of my analyses is unclear. When looking at the pairwise comparison of the first 5 principal components (Figure S2), it is hard to find a plot that approximates a divide between Alpha- and Betacoronaviruses, with the best one possibly being PC1 vs PC5, but the graph also divides the two Alphacoronaviruses on two opposite sides of the graph. However, PC5 only accounts for 5% of the variation, diminishing its importance. Taken all together, the PCA plots suggest a much more idiosyncratic relationship between these select viruses than the taxonomy shows.

Figure 2: Normalized PCA of Human-infecting coronaviruses



The X-Y graphing of PC1 vs PC2 is very informative, and could lend some credence to the idea that there is a correlation between the severity of the resulting disease and how similar the viruses are genetically, as well as pointing out a potential limitation in the traditional taxonomical approach. Strictly

examining alignment statistics, as I did in the first section of my report, doesn't bear this result out, though. A PCA is a blunt tool, that when used correctly can be excellent for exploratory data analysis and giving a zoomed-out view of the subject at hand. More in-depth research on specific genetic differences is needed, and is certainly being done as we speak. I think that my analysis could be used as a springboard for several steps. I think performing a global alignment, but on the first 2/3 of each virus' genome, excluding the structural M, E, N, and S proteins, would be informative and show how the viruses differ from each other in the portion of their sequence that is more apt to mutate and much less conserved than the structural proteins. It would be interesting to see what the phylogeny of the non-structural viral genome looks like, and the various *nsp*'s as well, because outside of the S-protein subunit 1, all the structural proteins were highly conserved between SARS and SARS-Cov-2.

The questions that I was asking at the beginning of this project were what gene- and protein-level differences existed between these select coronaviruses, and to what degree I could explain the epidemiological differentiation from my analyses. The answer to the first question is clear in that there are vast differences in gene and protein expression between the select viruses. This is to be expected because viruses evolve so fast. It also was striking how a virus that could have 55% similarity with SARS-Cov-2, could also have 55% similarity with HCov-HKU1, for example. The distinctiveness of these viral genomes surprised me, and I think it speaks volumes on how strong the selective pressure for viruses is against genomic waste. They have such small genomes, and depend on their biological efficiency for their survival much more than eukaryotes do. That creates incentives for rapid mutations and a plethora of biological diversity, and I hadn't fully appreciated this fact before. Given the similarity between the two, the SARS-Cov virus and 2002-2004 outbreak seem to be the most instructive in comparison to the current pandemic, especially considering the divergent worldwide outcomes.

My second question was how much could I explain the epidemiological variation with my gene- and protein-level analysis. I think that my results give a mixed answer, mainly because I do not have the means to experiment on these viruses and their hosts. But, it is clear from my domain-specific genetic similarity analysis that the answer is yes. Some regions are much more differentiated than others, and the types of regions that are heavy with mutations have seemingly direct effects on the epidemiological profile of the viruses. My analyses give a surface-level answer to my question, but more importantly, give

fodder for research that goes in two directions. The first direction is toward a vaccine, and its targets should be regions that are highly conserved, for the ideal vaccine will protect the patient from SARS-Cov-2 for as long as possible. The M, N, and E proteins seem to be suitable targets, although biological considerations and mechanisms that I am not knowledgeable enough about may dictate them to be unsuitable. The second direction is the research and palliative care route. This takes us to the highly mutated proteins, which are also likely the ones responsible for the increased contagiousness and activity of SARS-Cov-2, though research should not exclude other proteins and regions. It is worth noting that at the current moment, that SARS-Cov-2 has a slightly lower official death rate (official deaths / official cases from the numbers that I mentioned in the beginning) than SARS, though the difference is minimal and both rates almost certainly undercount the true case number. Figuring out a medicine and method that disrupts the ability of SA RS-Cov-2 to spread and reproduce in the body would certainly help lessen the treatment time of COVID-19, and lessen the societal impact of the pandemic. Discovering the causes of why SARS-Cov-2 isn't SARS certainly involves examining public policy responses, but scientists should be looking at proteins that have changed drastically from SARS to SARS-Cov-2 as a first step towards that goal.

Supplementary Figures

Table S1: PCA Summary Statistics

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Standard deviation	0.007518	0.006317	0.004773	0.003701	0.002738	0.001843	9.88E-18
Proportion of Variance	0.393070	0.277510	0.158440	0.095230	0.052130	0.023620	0E+00
Cumulative Proportion	0.393070	0.670580	0.829020	0.924250	0.976380	1.000000	1E+00

Figure S1: UCSC Genome Browser Screenshot with select domains highlighted

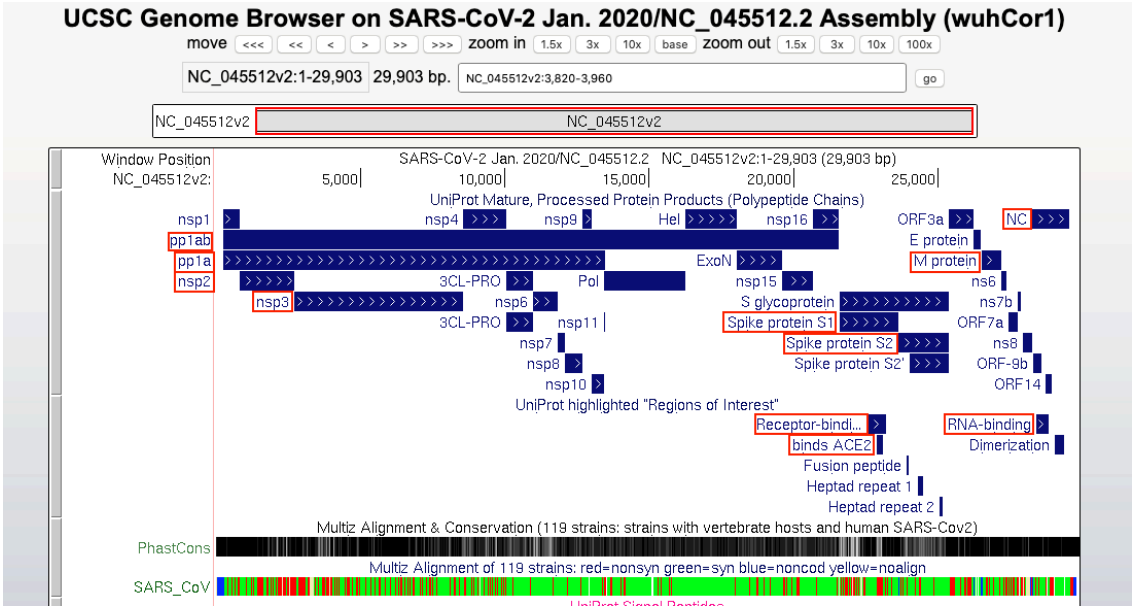
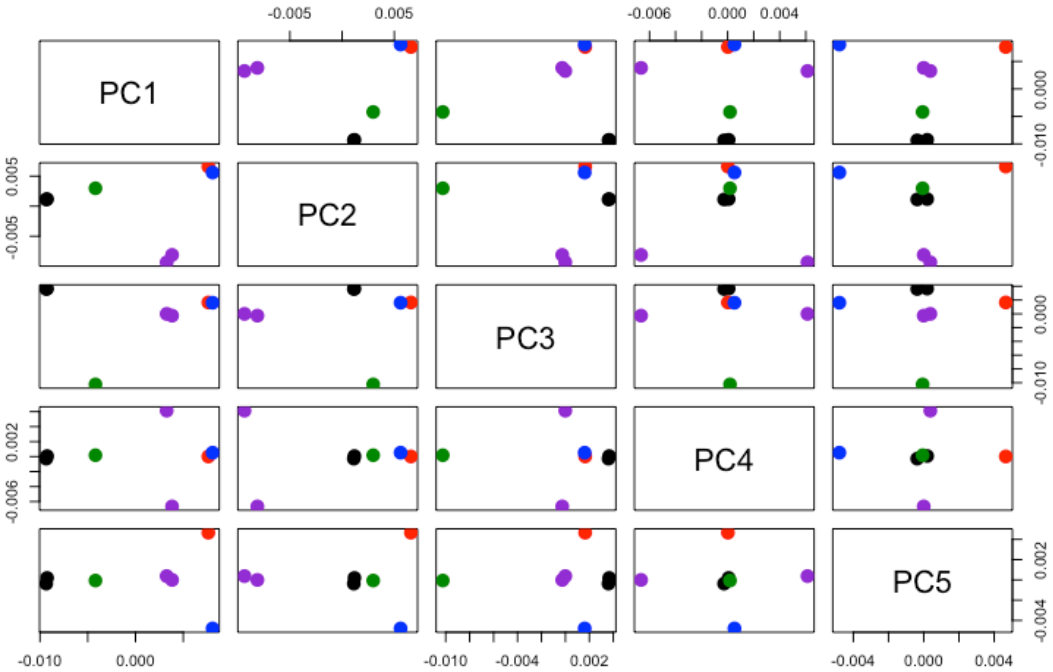


Figure S2: Pairwise comparison of principle components 1-5



Multiple Sequence Alignment and Local Alignment Code in bash

```
#Name: Edmond Brewer
#Date: 4/27/20
#Class: Genomics BMMB 551
#Assignment: Final Project

# Make references folder and setup genome/gff downloads
set -uex
mkdir refs
cd refs

#COVID-19
mkdir covid
cd covid
COV=NC_045512.2
#Get genbank genome and convert to FASTA and GFF, then index genome
efetch -db nucleotide -format gb -id $COV > $COV.gb
cat $COV.gb | seqret -filter -feature -osformat fasta -ofname2 $COV.gff > $COV.fa
rm -R *.gb
samtools faidx $COV.fa
cd ..

#Severe Acute Respiratory Syndrome 1
mkdir sars
cd sars
SAR=NC_004718.3
#Get genbank genome and convert to FASTA and GFF, then index genome
efetch -db nucleotide -format gb -id $SAR > $SAR.gb
cat $SAR.gb | seqret -filter -feature -osformat fasta -ofname2 $SAR.gff > $SAR.fa
rm -R *.gb
samtools faidx $SAR.fa
cd ..

#Middle East Respiratory Syndrome
mkdir mers
cd mers
MER=NC_019843.3
# Get genbank genome and convert to FASTA and GFF, then index genome
efetch -db nucleotide -format gb -id $MER > $MER.gb
cat $MER.gb | seqret -filter -feature -osformat fasta -ofname2 $MER.gff > $MER.fa
rm -R *.gb
samtools faidx $MER.fa
cd ..

#Human Coronavirus HKU1
mkdir hku1
cd hku1
HKU=NC_006577.2
# Get genbank genome and convert to FASTA and GFF, then index genome
efetch -db nucleotide -format gb -id $HKU > $HKU.gb
cat $HKU.gb | seqret -filter -feature -osformat fasta -ofname2 $HKU.gff > $HKU.fa
rm -R *.gb
samtools faidx $HKU.fa
cd ..
#Human Coronavirus 229E
mkdir 229e
cd 229e
two29E=NC_002645.1
# Get genbank genome and convert to FASTA and GFF, then index genome
efetch -db nucleotide -format gb -id $two29E > $two29E.gb
cat $two29E.gb | seqret -filter -feature -osformat fasta -ofname2 $two29E.gff > $two29E.fa
rm -R *.gb
samtools faidx $two29E.fa
cd ..

#Human Coronavirus OC43
mkdir oc43
cd oc43
OC43=NC_006213.1
# Get genbank genome and convert to FASTA and GFF, then index genome
efetch -db nucleotide -format gb -id $OC43 > $OC43.gb
cat $OC43.gb | seqret -filter -feature -osformat fasta -ofname2 $OC43.gff > $OC43.fa
rm -R *.gb
samtools faidx $OC43.fa
cd ..

#Human Coronavirus NL63
mkdir nl63
cd nl63
NL63=NC_005831.2
# Get genbank genome and convert to FASTA and GFF, then index genome
efetch -db nucleotide -format gb -id $NL63 > $NL63.gb
cat $NL63.gb | seqret -filter -feature -osformat fasta -ofname2 $NL63.gff > $NL63.fa
rm -R *.gb
samtools faidx $NL63.fa
cd ..
cd ..

# Multiple Genome Alignment
cat refs/covid/*.fa refs/sars/*.fa refs/mers/*.fa refs/hku1/*.fa refs/229e/*.fa refs/oc43/*.fa refs/nl63/*.fa >
combined_genomes.fa
clustalo -i combined_genomes.fa -o clustalo_msa.fa -v

#Double Genome Alignment
stretcher refs/covid/*.fa -bsequence refs/sars/*.fa -outfile covid_vs_sars.txt
stretcher refs/covid/*.fa -bsequence refs/mers/*.fa -outfile covid_vs_mers.txt
stretcher refs/covid/*.fa -bsequence refs/hku1/*.fa -outfile covid_vs_hku1.txt
stretcher refs/covid/*.fa -bsequence refs/229e/*.fa -outfile covid_vs_229e.txt
stretcher refs/covid/*.fa -bsequence refs/oc43/*.fa -outfile covid_vs_oc43.txt
stretcher refs/covid/*.fa -bsequence refs/nl63/*.fa -outfile covid_vs_nl63.txt

#Align select SARS-Cov-2 Proteins to SARS Genome to find alignment statistics
water receptor_binding_domain.txt refs/sars/*.fa -gapopen 10.0 -gapextend 0.5 -outfile rbd_sars.txt
water spike_protein_S1.txt refs/sars/*.fa -gapopen 10.0 -gapextend 0.5 -outfile S1_sars.txt
water spike_protein_S2.txt refs/sars/*.fa -gapopen 10.0 -gapextend 0.5 -outfile S2_sars.txt
water RNA_binding.txt refs/sars/*.fa -gapopen 10.0 -gapextend 0.5 -outfile RNA_binding_sars.txt
water ACE2_binding.txt refs/sars/*.fa -gapopen 10.0 -gapextend 0.5 -outfile ACE2_binding_sars.txt
water N_protein.txt refs/sars/*.fa -gapopen 10.0 -gapextend 0.5 -outfile N_protein_sars.txt
water pp1ab.txt refs/sars/*.fa -gapopen 10.0 -gapextend 0.5 -outfile pp1ab_sars.txt
water nsp2.txt refs/sars/*.fa -gapopen 10.0 -gapextend 0.5 -outfile nsp2_sars.txt
water nsp3.txt refs/sars/*.fa -gapopen 10.0 -gapextend 0.5 -outfile nsp3_sars.txt
water pp1a.txt refs/sars/*.fa -gapopen 10.0 -gapextend 0.5 -outfile pp1a_sars.txt
water M_protein.txt refs/sars/*.fa -gapopen 10.0 -gapextend 0.5 -outfile M_protein_sars.txt
water E_protein.txt refs/sars/*.fa -gapopen 10.0 -gapextend 0.5 -outfile E_protein_sars.txt
```

Phylogeny and Principle Component Analysis Code in R

```
#Phylogram and Principle Components Analysis Code
#Code Template provided by Dr. T.F. Khang and Drew Kerkhoff. See references for citations

#library and setup
library(seqinr)
library(Biostrings)
library(ape)
library(textmineR)
library(phangorn)
library(ggplot2)
library(ggrepel)
setwd("/Users/Ned/BMMB_852/Genomics_Final/viral_genomes")

#read into phyDat, create taxonomy, create color scheme
corona.aln.phydat=read.phyDat("/Users/Ned/BMMB_852/Genomics_Final/clustalo_msa.maf", format="fasta", type = "DNA")
summary(corona.aln.phydat)[1:5,]
corona.names <- c("SARS-Cov-2", "SARS-Cov", "MERS-Cov", "HCov-HKU1", "HCov-229E", "HCov-OC43", "HCov-NL63")
names(corona.aln.phydat)=corona.names
names(corona.aln.phydat)
taxonomy <- data.frame(names(corona.aln.phydat),
                       c("Beta Sarbecovirus", "Beta Sarbecovirus", "Beta Merbecovirus", "Beta Embecovirus",
                         "Alpha Duvinacovirus","Beta Embecovirus", "Alpha Setracovirus"))
colnames(taxonomy) <- c("Virus","Family")
tipcolor <- c("red","blue","darkviolet","green4","black")[unclass(taxonomy$Family)]

#perform distance analysis, neighbor joining, and plot the phylogram
corona.phydat.dist=dist.ml(corona.aln.phydat)
corona.NJ=NJ(corona.phydat.dist)
plot.phylo(corona.NJ, x.lim = 15, y.lim = NULL, use.edge.length=FALSE, cex=1, tip.color = tipcolor)

#normalize phyDat data, perform principle components analysis, and plot pairs and final graph
M <- do.call(rbind, corona.aln.phydat)
M_norm <- t(apply(M, 1, function(k) k/sum(k)))
pca <- prcomp(M_norm)
summary(pca)
pairs(pca$x[,1:5], pch=16, cex=2, col=tipcolor)
pc1_v_2 <- data.frame(pca$x[,1], pca$x[,2])
colnames(pc1_v_2) <- c("PC1", "PC2")
ggplot(data = pc1_v_2, aes(x = PC1, y = PC2, color = taxonomy$Family)) +
  geom_point(color = tipcolor, size = 3, aes(color = taxonomy$Family)) +
  geom_label_repel(aes(label = rownames(pc1_v_2))) +
  labs(x = "PC1 (39.3%)", y = "PC2 (27.7%)", title = "Figure 2: Normalized PCA of Human-infecting coronaviruses") +
  scale_color_manual(name = "Virus Family",
                     breaks = c("Alpha Duvinacovirus", "Alpha Setracovirus", "Beta Embecovirus",
                                "Beta Merbecovirus", "Beta Sarbecovirus"),
                     values = c("Alpha Duvinacovirus" = "red", "Alpha Setracovirus" = "blue",
                                "Beta Embecovirus" = "darkviolet", "Beta Merbecovirus" = "green4",
                                "Beta Sarbecovirus" = "black"))
```

References

1. https://coronavirus.jhu.edu/map.html
2. https://www.who.int/csr/sars/country/table2004_04_21/en/
3. https://www.who.int/emergencies/mers-cov/en/
4. https://www.ncbi.nlm.nih.gov/home/genomes/
5. http://www.clustal.org/omega/
6. https://www.snapgene.com/
7. Myers E.W. and Miller W. (1988) Optimal alignments in linear space. Comput. Appl. Biosci. 4(1):11-7. PubMed: 3382986
8. https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Tree&id=11118&lvl=3&lin=f&keep=1&srchmode=1&unlock
9. Lim YX, Ng YL, Tam JP, Liu DX. Human Coronaviruses: A Review of Virus-Host Interactions. *Diseases*. 2016;4(3):26. Published 2016 Jul 25. doi:10.3390/diseases4030026
10. Xu X, Chen P, Wang J, et al. Evolution of the novel coronavirus from the ongoing Wuhan outbreak and modeling of its spike protein for risk of human transmission. *Sci China Life Sci*. 2020;63(3):457–460. doi:10.1007/s11427-020-1637-5

11. Lu R, Zhao X, Li J, et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *Lancet*. 2020;395(10224):565−574. doi:10.1016/S0140-6736(20)30251-8
12. Coronaviridae Study Group of the International Committee on Taxonomy of Viruses. The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. *Nat Microbiol*. 2020;5(4):536−544. doi:10.1038/s41564-020-0695-z
13. Smith TF, Waterman MS (1981) J. Mol. Biol 147(1);195-7
14. Wrapp D, Wang N, Corbett KS, et al. Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science*. 2020;367(6483):1260−1263. doi:10.1126/science.abb2507
15. https://covid-19.uniprot.org/uniprotkb/P0DTC1
16. Angeletti S, Benvenuto D, Bianchi M, Giovanetti M, Pascarella S, Ciccozzi M. COVID−2019: The role of the nsp2 and nsp3 in its pathogenesis. *Journal of Medical Virology*. 2020;92(6):584-588. doi:10.1002/jmv.25719.
17. Wu C, Liu Y, Yang Y, et al. Analysis of therapeutic targets for SARS-CoV-2 and discovery of potential drugs by computational methods [published online ahead of print, 2020 Feb 27]. *Acta Pharm Sin B*. 2020;10.1016/j.apsb.2020.02.008. doi:10.1016/j.apsb.2020.02.008
18. Sarma P, Sekhar N, Prajapat M, et al. In-silico homology assisted identification of inhibitor of RNA binding against 2019-nCoV N-protein (N terminal domain) [published online ahead of print, 2020 Apr 8]. *J Biomol Struct Dyn*. 2020;1−11. doi:10.1080/07391102.2020.1753580
19. Schliep K.P. (2011) phangorn: Phylogenetic analysis in R.Â *Bioinformatics*, 27(4) 592-593
20. R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
21. Paradis E. & Schliep K. 2018. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics 35: 526-528.
22. Kerkhoff D. Basics of Sequence Data and Phylogenies in R. https://rstudio-pubs-static.s3.amazonaws.com/144108_49ef6d7992684fe68a39e53d82f950d0.html. Published January 18, 2016.
23. Khang TF. Whole-genome viral phylogeny estimation without sequence alignment. BioinfoRmatics R Tutorial: Whole-genome viral phylogeny estimation without sequence alignment. https://bioinformaticshome.com/bioinformatics_tutorials/R/phylogeny_estimation.html.