Genomics and SARS-CoV-2

Introduction

The outbreak of the respiratory disease, SARS-CoV-2, which originated in Wuhan, China, has disseminated globally and influenced human life at a wide scale. The virus has exhibited a high transmissibility rate, causing over 500,000,000 positive cases and 6,200,000 deaths worldwide and growing in weekly cases by the millions and deaths by the tens of thousands, despite global vaccination and quarantining efforts (*WHO Coronavirus (COVID-19) Dashboard*, 2022). It has persisted through the human population possibly due to a mix of obstacles, including improper quarantining initiatives, the arrival of mutated strains within SARS-CoV-2 such as Delta and Omicron, and a lack of understanding on the complexity of the virus' genetic progression. Even after several years of extensive research on the virus, there still lied a need to understand a disease with vast implications.

This is where the role of genomics came in. A full investigation of the COVID-19 genome was required to better understand this virus, which in turn, uncovered findings with an array of purposes: for understanding genome structure (Naqvi et al., 2020), analyzing open-reading frames (ORFs) with unknown functions and protein-coding structures (Jungreis et al., 2021), analyzing mutations and conservation of nucleotide sequences and their corresponding codons for drug targeting (Chan et al., 2020; Rangan et al., 2020), and observing or comparing SARS-CoV-2 variants (Kandeel et al., 2022; Mlcochova et al., 2021). The use of genomics has been integral in researching not only SARS-CoV-2, but other similar viruses like SARS-CoV, MERS-CoV, and bat-related coronaviruses. Hence, this paper will integrate a brief and current overview of the landscape of genomics and SARS-CoV-2 via an analysis of 6 primary literature research articles; the goal for this analysis is to provide a better sense of how genomics has been used with relation to the pandemic, what kinds of findings have stood out as useful, what the implications of those findings were, and possibilities for the field in the future.

Review of the Literature

For any living organism, its genetic composition determined, among other things, the key proteins which function for its survival and reproduction. In the case of SARS-CoV-2, many of the encoded proteins played a vital role in structure and in turn, determined its ability to infect host cells and multiply. One such protein was the spike glycoprotein (S), which bound to the host cell through its receptor-binding domain (RBD); the SARS-CoV-2 S protein "[was] composed of 1273 amino acid residues containing three subunits, namely S1, S2, and S2'" (Naqvi et al., 2020). The S1 subunit was of particular interest to researchers, as it allowed the virus to interact with the human ACE2 receptor, thereby attaching virions to the host cell (Naqvi et al., 2020). Furthermore, the S2 subunit worked to fuse virions with mammalian cell membranes; the S2' subunit was a fusion peptide (Naqvi et al., 2020). Via a genomic comparison between four *Sarbecovirus* strains belonging to betaCoV lineage B – these were SARS-CoV-2, SARS-CoV, MERS-CoV, and animal CoVs from bats, pangolins, and civets – nucleotide conservation and variability was observed. The S protein sequence area was a point of consideration for researchers because it was comparable to other coronaviruses and therefore useful for drug targeting. Sequence similarities for the spike stalk S2 of SARS-CoV-2 showed a near 99% match with the sequences from "bat SARS-like CoV's and human SARS-CoV… indicating a broad-spectrum use of antiviral compounds… which may be useful in COVID-19 therapy" (Naqvi et al., 2020). However, compared with other functioning protein-coding sequences in the COVID-19 genome, the RBD of the S protein demonstrated the most variability across betaCoV lineage B strains. Specifically, a "~90 amino acid long receptor binding motif of the RBD… [showed] the least conservation, suggesting the involvement of multiple mechanisms in pathogenesis" (Naqvi et al., 2020). Although the S2 subunit sequence area was a potential target region, the RBD of the S protein displayed inter-strain variability, so the feature was not an obvious vulnerability of the virus and other sequences should be considered.

Although targeting the S protein may prove futile due to its variability, there were other more-conserved regions which elicited the interest of researchers: non-coding elements of the SARS-CoV-2 genome, ORFs with still-unknown functions, and protein-coding features (Chan et al., 2020; Jungreis et al., 2021; Naqvi et al., 2020). For example, Naqvi et al. highlighted areas of conservation for sequences which coded for spike proteins (S), envelope membrane proteins (E), membrane proteins (M), and nucleocapsid proteins (N) in comparison of four strains within betaCoV lineage B:

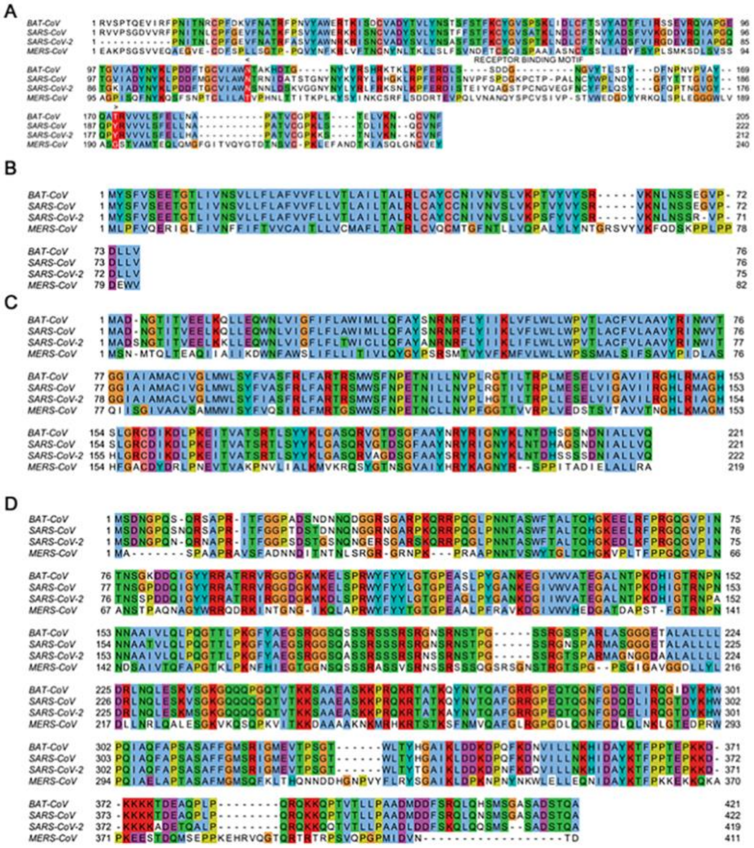


Figure 1. Multiple sequence alignment of A. S protein, B. E protein, C. M protein, D. N protein. Sourced from Naqvi et al., 2020.

The colors represented conservation or likeness across the 4 strains; from the alignment graph, it showed the S protein (A) with the whitest space and therefore the least conservation, while the three other proteins (B, C, D) displayed higher conservation (Figure 1). Likewise, the E protein (B) had the most conservation compared to the 3 other protein-coding sequences and was one of the most conserved regions in the SARS-related genomes (Chan et al., 2020). The envelope protein was the smallest of the structural CoV proteins consisting of 76-109 amino acids, but it played a crucial role in “release and insertion of the virus to host cell” (Aldaais et al., 2021). Among BAT-CoV, SARS-CoV, and SARS_CoV-2, significant sequence conservation was observed in the E protein; the MER-CoV strain sequence varied the most amongst the compared species. Despite their more conserved regions, the M and E proteins have garnered less attention than the Spike protein, which called for further research into these other vital structures and coding regions (Aldaais et al., 2021).

In addition to the observed conservation of the E, M, and N protein-coding sequences, there were other features, such as non-protein coding sequences, which were marked by low nucleotide variability amongst the compared strains of betaCoV lineage B (Chan et al., 2020). Chan et al. (2020) highlighted the degree of similarity among 28 annotated genomic features (“ORFs, processed peptides, and UTRs”) across 109 CoV family genomes. A BLAST search compared nucleotide and amino acid sequences of family genomes across betaCoV lineage B. These regions included “the 3’-UTR… ORF10, the 5’-UTR, and nsp10,” which were four of the five most conserved alignments (the fifth being the E gene).

First, the genomic terminals for betaCoV lineage B isolates were both very conserved, which suggested a functional significance and an area for more research on drug targeting (Chan et al., 2020). Additionally, a multiple sequence alignment (MSA) “on 620 near-full length betaCoV lineage B genomes” shows 94% sequence identity of 3’-UTR positions and 84% of 5’-UTR positions in 99% of the aligned genomes (Chan et al.). One useful aspect of high similarities was that it allowed for classification of distinct nucleotide profiles. High variability led to an abundance of profiles, which posed as cumbersome to discern for consistencies. On the other hand, low variability enabled researchers to detect unique variants across the observations with more clarity. Thus, nucleotide variability across specific positions resulted in classification profiles for distinct sub-clades of betaCoV genomes based on their UTR signatures, of which there were 15 UTR signatures clustered into two groups: SARS-CoV-2 and SARS-CoV (Chan et al., 2020). This explained why treatment used against SARS in the 2002 outbreak was not carried over to combat SARS-CoV-2 – the groups differed by 76% of non-identical nucleotides in their respective UTR signature positions (Chan et al., 2020).

The 3’-UTR signature was distinct for each group of SARS-CoV-2, SARS-CoV, and the bat CoVs isolates, however, “these three positions [were] overlapped with a conserved RNA motif S2m… previously identified in coronavirus and astrovirus” (Chan et al., 2020). Despite unique signatures across the three groups, the predicted RNA secondary structure, S2m was conserved and may have played a role in structure or function (Chan et al., 2020). S2m was highly conserved across similar groups, thus it demanded more attention considering a structure for drug therapy targets. For example, the human miRNA overexpression of hsa-miR-1307-3p was complementary to H1N1 nsp1 which inhibited wild-type H1N1 replication and reduced nsp1 expression.

> “Sequence matches to the human miRNAs hsa-miR-1307-3p and has-miR-1304-3p were located within the broader conserved RNA motif S2m. Interestingly, a recent study of IAV H1N1 provided supporting functional evidence of hsa-miR-1307-3p in mediating antiviral responses and inhibiting viral replication” (Chan et al., 2020).

Additionally, the conserved RNA motif S2m was found in other *Sarbecovirus* families including gammaCoVs and deltaCoVs (Chan et al., 2020). With consideration to the high variability of the S protein and historical success of variants of SARS-CoV-2, further pathways for investigation into the conserved regions and structures, such as the RNA motif, S2m, demanded further attention and opened research for the future.

The next paper continued the search for RNA gene conservation and found 79 identifying regions >15 nt were exactly conserved (Rangan et al., 2020). As an aside, an initial positive aspect of the Rangan et al. (2020) paper was its awareness of the COVID-19 timeline, as it noted that the sequences were taken from strains at the beginning of the pandemic. None of the other papers identified which dates their analyses presided over. Rangan et al. (2020) found conserved structured and unstructured regions which could be predicted and therefore were marked as potential therapeutic targets. Specifically, they used MSA comparing SARS, SARS-CoV-2, and SARS-related bat coronaviruses via three methods:

1. Alignment captured from sequences curated from another reference.
2. BLAST search of 100 genome sequences close to SARS-CoV-2 reference genome.
3. Complete genome betacoronavirus sequences obtained from NCBI database.

RNA secondary structures were investigated and the 5'-UTR region was marked by five stem-loop structures: SL1-SL5 (Rangan et al., 2020). SL1 was found to be thermodynamically unsable "to allow for the formation of long-range-interactions) (Rangan et al., 2020). SL2 demonstrated the highest sequence conservation across the 5'-UTR and was shown to be critically involved in subgenomic RNA synthesis (Rangan et al., 2020). SL3 was conserved only in betacoronaviruses, SL4 contained a short upstream ORF labeled as uORF, and SL5 may have played a role in viral packaging plus it contained the AUG start codon for ORF1ab (Rangan et al., 2020). Again, a more clear identification of the structures and functions of RNA secondary structures of SARS-CoV-2 initiated opportunities of further research for RNA-targeting interventions and treatment.

Open-reading frames (ORFs) referred to any portion of a genomic sequence that did not include a stop codon. A definition of the function or the protein-coding elements of an ORF was difficult, thus the use of multiple sequence alignment (MSA) was employed for detecting start and stop codons; comparative genomics adds on to this detection model, where gene prediction can only be done via a large training set of many genomes to classify ORF function (Galperin & Frishman, 1999). The SARS-CoV-2 genome's first two-thirds is comprised of two ORFs: 1a and 1b. The last third of the genome encoded for the S, E, M, and N proteins, plus several ORFs specific to the *Sarbecovirus* genus: 3a, 6, 7a, 7b, and 8 were previously identified in other species; these ORFs also included 2b, 9b, 9c, and 10 (Jungreis et al., 2021). Some, but not all, of each ORF's function or protein-coding status was known, hence Jungreis et al. compared the genes among 44 *Sarbecovirus* genomes to uncover the protein-coding signatures of each ORF based on the criteria:

> Selection of 44 genomes "at evolutionary distances well-suited for identifying protein-coding genes and non-coding purifying selection, spanning ~3 substitutions per 4-fold degenerate site on average (comparable to 29-mammals/12-flies projects[29,30]), and ranging from 1.2 (E) to 4.8 (O-MT/nsp16) and higher" (Jungreis et al., 2021).

An initial takeaway from this study was that the comparative genomics training programs were informative for future study. Jungreis et al. developed PhyloCSF32 to train on entire genomes for comparison of codon substitutions and frequencies “in alignments of related genomes to coding and noncoding models” (2021). CodAlignView was used for visual examination of alignments for substitutions, stop codons, insertions, and deletions (Jungreis et al., 2021). The figure on the right side of the page displayed genome alignment for ORF 3c across 44 *Sarbecovirus* strains. Synonymous constraint elements overlapped 38 of 41 codons in ORF 3a to highlight the sequence as “localized nearly perfectly on the dual-coding region” (Jungreis et al., 2021). Furthermore, ORF 3c demonstrated completely conserved start codons, excluding one near-cognate substitution, and entirely conserved stop codons, with exception of a single-codon extension (Figure 2). Many synonymous mutations were present in ORF 3c which “[indicated] ORF3c may be an equally strong driver of constraint in the dual-coding region” (Jungreis et al., 2021). On the other hand, ORF 3a’s genomic comparison exhibited a high concentration of non-synonymous mutations. Due to these reasons, Jungreis et al. (2021) concluded that ORF3c encoded for a conserved and functional protein. Since there was no one-method for using comparative genomics, the usage of PhyloCSF32 and CodAlignView enabled the study to distinguish protein-coding status of genes. Jungreis et al. (2021) highlighted the advantages of its methodology, which employed programs to “[focused] on the patterns of change characteristic of protein-coding constraint (specific codon substitution frequencies and reading frame conservation) rather than the overall number of substitutions… [and were] resilient to the recombination events that [characterized] coronavirus genomes” (2021). Therefore, the possible future usage for PhyloCSF32 and CodAlignView for full genome comparisons can expand the understanding of the still-unclear coding statuses of several ORFs. This study marked the first time these programs have been used for comparison of viral genomes; hence, further practice was required for future studies. However, Jungreis et al. have already classified “all existing and potential SNVs and known RNA modification sites into likely-functional vs like-neutral,” which forwarded the frontier of what is known about the genome and elucidated more important information in the arms race between host and virus (2021).

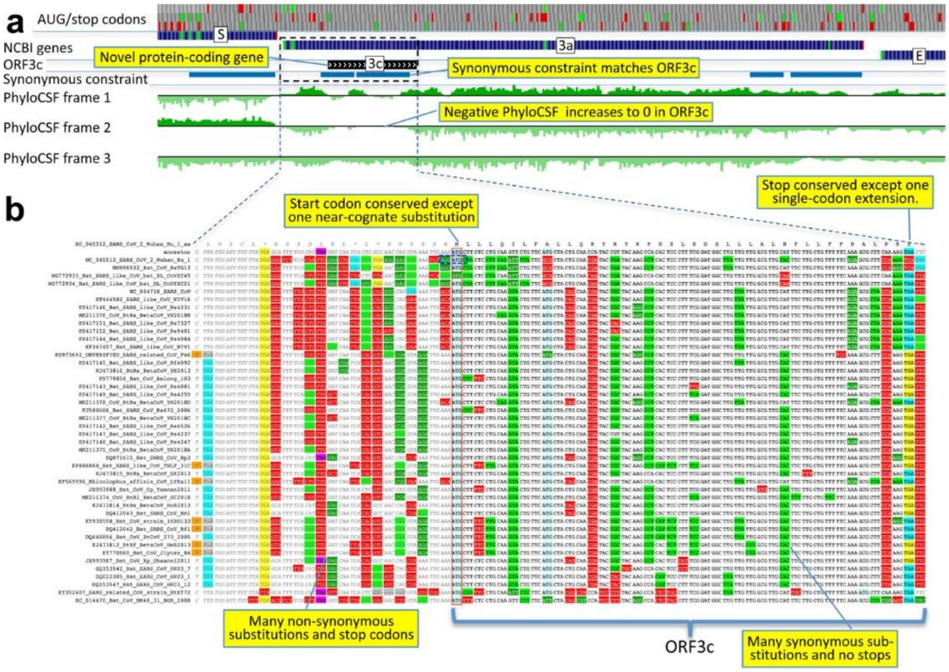
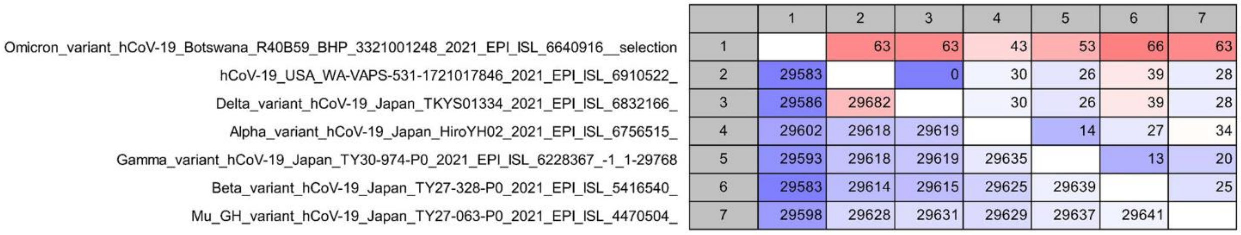

a
AUG/stop codons
S
NCBI genes
3a
ORF3c
Novel protein-coding gene
3c
E
Synonymous constraint
Synonymous constraint matches ORF3c
PhyloCSF frame 1
PhyloCSF frame 2
Negative PhyloCSF increases to 0 in ORF3c
PhyloCSF frame 3
Stop conserved except one single-codon extension.
b
Start codon conserved except one near-cognate substitution
Many non-synonymous substitutions and stop codons
ORF3c
Many synonymous sub-stitutions and no stops

Figure 2. Alignment of Sarbecovirus genes at ORF 3c shows that this gene is protein-coding. From Jungreis et al., 2021

Gene mutations caused the rise of variants such as “Delta” and “Omicron” and were a key issue which required better understanding of the SARS-CoV-2 genome. So far, the research uncovered conserved therapeutic target areas, nucleotide variability among different strains, and function of formerly uncertain protein coding ORFs. Yet, there was still a need to investigate the two mutated strains which have succeeded in infecting vaccinated individuals and inducing 2nd and 3rd waves of the worldwide virus (Kandeel et al., 2022; Mlcochova et al., 2021). The final portion of this review will look at the genomes of B.1.617.2 "Delta" and B.1.1.529 "Omicron" variants, their mutated sequence area, and changed function contrasted to other SARS-CoV-2 variants.

The Delta variant displayed more resilience towards antibodies in recovering and vaccinated individuals compared to wild-type Wuhan-1 SARS-CoV-2, arising in India at the end of 2020 (Mlcochova et al., 2021). Meanwhile, on November, 2021, the Omicron variant was discovered in Botswana and was characterized by high transmissibility but milder symptoms compared to other variants (CDC, 2022). Both variants exhibited mutations on the S protein – the Delta variant bore the L452R spike receptor-binding motif (RBM) substitution while the Omicron variant demonstrated >=32 mutations in the spike protein. Delta showed a higher replication rate over the B.1.1.7 Alpha variant alongside its spike protein, which completely neutralized domain monoclonal antibodies and four out of nine non-RBM monoclonal antibodies (Mlcochova et al., 2021). Additionally, the S proteins exhibited evasion from an adenovirus vector vaccine (ChAdOx1) and an mRNA based (BNT162b2) vaccine; adaptations to the vaccine for improved efficacy against these variants were needed (Mlcochova et al., 2021).



		1	2	3	4	5	6	7
Omicron_variant_hCoV-19_Botswana_R40B59_BHP_3321001248_2021_EPI_ISL_6640916__selection	1		63	63	43	53	66	63
hCoV-19_USA_WA-VAPS-531-1721017846_2021_EPI_ISL_6910522_	2	29583		0	30	26	39	28
Delta_variant_hCoV-19_Japan_TKYS01334_2021_EPI_ISL_6832166_	3	29586	29682		30	26	39	28
Alpha_variant_hCoV-19_Japan_HiroYH02_2021_EPI_ISL_6756515_	4	29602	29618	29619		14	27	34
Gamma_variant_hCoV-19_Japan_TY30-974-P0_2021_EPI_ISL_6228367_-1_1-29768	5	29593	29618	29619	29635		13	20
Beta_variant_hCoV-19_Japan_TY27-328-P0_2021_EPI_ISL_5416540_	6	29583	29614	29615	29625	29639		25
Mu_GH_variant_hCoV-19_Japan_TY27-063-P0_2021_EPI_ISL_4470504_	7	29598	29628	29631	29629	29637	29641	

Figure 3. "Pairwise comparative matrix of Omicron with SARS-CoV-2 variants." Upper right diagonal shows gaps; the bottom left diagonal shows quantity of identical nucleotides. Figure from Kandeel et al., 2022.

Furthermore, genome alignment of different SARS-CoV-2 variants showed the Omicron variant with more gaps than other viruses while it had the highest shared nucleotide sequencing with the Alpha variant (Figure 3). Finally, a phylogeny of COVID-19 variants "using neighbor-joing method and Tamura substitution model" showcased Omicron as an outgroup (Kandeel et al., 2022). Depending on the model-usage, however, the resulting phylogenetic tree sometimes placed the Omicron variant within the same clade as the Alpha variant; in a larger clade, the Omicron shared a clade with the Delta variant and "hCov-19" (Kandeel et al., 2022). Kandel et al. (2022) suggested that the Omicron variant's genetic closeness to the Alpha variant possibly explained how it seemed to spread so quickly – the Omicron was already in circulation and was only discovered late after its initial spread. This is an interesting point to consider because it may indicate the spread of alternate, but undiscovered, variants in the population. Although, it is difficult to predict when, where, and what the next variant will be. For the tie being, alignments of genomes and modeled phylogenies could find mutated areas, like the S protein, along with the traced evolutionary history of these variants. However, there is a need to push genomics further for quicker detection and analysis of such variants.

In summation, the state of genomics and the COVID-19 virus remains a field with a fruitful bearing of research to be built upon. Comparative genomics played a key role in viral research as it moved the frontier forward. Understanding the SARS-CoV-2 whole genome was a tall task, but it began with a comparison of the genome against other *Sarbecovirus* strains to detect conserved, mutated, and variable nucleotides and codons. Using tools like BLAST, PhyloCSF32, and CodAlignView, the visualization of alignment comparison provided useful insights for answering questions about genetic regions of interest. Plus, the usage of PhyloCSF32 was at its beginning in terms of investigating viral genomes. With such tools that supported their research, studies paid particular attention to unknown gene regions as well as highly conserved regions, including RNA secondary structures, the E gene, the 3'-UTR, and the 5'-UTR, and considered these areas as therapeutic targets for future investigation. The Delta and Omicron variants demonstrated increased resilience to vaccines and higher transmissibility than other variants, which presented a challenge for a world still facing the pandemic. Thus, comparative genomics were required to find the mutated regions in the variants for preparedness against the possibility of future mutations.

Current State

As of 2022, there have been 161 genomics and COVID-19 studies published in the National Library of Medicine. Based on the titles, abstracts, and keywords from these studies, genomics of recent has been centered on the mapped genome of the Omicron variant and the continued expansion of vaccine research. Cases surged to the highest 7-day moving average in the COVID-19 timeline from January to March of 2022 (*WHO Coronavirus (COVID-19) Dashboard*, 2022). Plus, the more recent discovery of the variant, "Omicron XE," has sprung up in India and in the United Kingdom in hundreds of cases as a recombinant virus of Omicron BA.1 and BA.2 (Ma & Chen, 2022). The arms race between host and virus bore on as the number of cases, and opportunities to mutate, increased earlier this year. As COVID-19 variants continued to pop up around the globe, genomics research located the structural differences among the variants and pinpointed mutated regions. One advantage to the already-mapped genome of SARS-CoV-2 was the ability to compare sequences among new variants and detect mutations in pertinent regions. In this case, the XE Omicron variant demonstrated mutations nsp3 C3214T, nsp12 C14599T (synonymous mutations), and nsp3 V1069I (amino acid mutation) (Ma & Chen, 2022). As stated earlier, the rise of variants within the population called for further investigation of mutation areas and conserved regions for therapeutic targeting. While there was still no known method for predicting when the next variant will arise, the hope was that viral genomics would continue to lay a groundwork for mutation tracking.

An important aspect when considering the current state of genomics is access to sequenced data and genomics-associated technologies. Studies utilized gene sequencing technologies by researchers from highly funded universities, such as Stanford University (Rangan et al., 2020), the Translational Genomics Research Institute in Phoenix (Chan et al., 2020), and the Massachusetts Institute of Technology (Jungreis et al., 2021). The adequate funding enabled such studies to use genomic technologies with expensive up-front costs (Office, n.d.). However, sequencing technologies have been adapted since the beginning of the pandemic and have grown more available due to lowered costs – a whole SARS-CoV-2 genome could be sequenced for as little as $33.80 (Park et al., 2021). Ease of access has allowed for "67,000 genomes per day" to be deposited into public viral genome data repositories, which contained over 6 million SARS-CoV-2 genomes by April, 2022 (Knyazev et al., 2022). There were still many genomes to be sequenced as the virus persisted, yet with the collaborative global effort to contribute SARS-CoV-2 genomes to databanks, the data set has become more enriched.

An evolving field because of the pandemic has been genomic surveillance, which monitored the emergence of variants of concern via aggregated databases. Novel genomic technologies have promoted the sharing gene sequencing of SARS-CoV-2 globally. For example, CoV-Seq, a web server that enabled simple and rapid analysis of SARS-CoV-2 genomes, was developed in response to the pandemic (Liu et al., 2020). Through automatic analysis of gene boundaries via machine learning programs, CoV-Seq could detect genetic variants across a multitude of SARS-CoV-2 genome inputs. This was a highly accessible program and purposed for clinicians with "limited bioinformatics or programming knowledge" (Liu et al., 2020). Furthermore, "COVseq could be used to sequence thousands of samples at less than 15 USD per sample, including library preparation and sequencing costs" (Simonetti et al., 2021). Thus, the current state of genomic technology has attempted to bridge the gap between genetic specialists and clinicians who applied treatment for COVID-19; plus, an emphasis on shared knowledge amongst researchers has bolstered genomic surveillance as a reliable field of study today.

Future of the Field

Billions of dollars have already been put into vaccine research and COVID-19, so possibilities for the future of the field rely on the foundation of current studies which have already contributed to an outpour of research. One such foundational research item is genomic comparison and sequencing, as its continued has been and will be a key tool in the human-virus arms race. Within the six studies looked at in this paper, they used BLAST and introduced PhyloCSF as genome alignment tools useful for visualization and immediate analysis of the included genomes, based on the widely available data coming from sources like the NCBI as well as each study's respective main academic institution data base. As the virus persists, with a certain degree of uncertainty for when the next variant will pop up, the use of genomic comparison tools can drive research further to find those highly conserved or protein coding regions of importance to the virus. Therefore, a call for further research and a concerted investment into these kinds of tools and their development is needed. A possible direction for the current and future state of such tools is improvement in accuracy, refinement, readability, and usability. The PhyloCSF and CodAlignView analyses were interpreted by Jungreis et al., who pioneered this tool used for viral genomics, and therefore more usage across different researchers could provide consensus for best-use and application. This was not said to question the interpretation of Jungreis et al., but rather posited that the tool's functionality and ease of use would inevitably grow as it became adopted into the viral-genomics field. For example, although the use of PhyloCSF was a novel research method for viral genomics, codon-frequency-comparison proved to be useful for discerning ORF function and the study uncovered many questions about the SARS-CoV-2 genome. There are more possibilities that can be accessed via PhyloCSF as more strains of the virus become sequenced and discovered, which ultimately provides better input for sound comparisons and detection of conserved regions.

With consideration to COVID-19 treatment and resilience, an obvious future direction for the field was to then be more holistically connected with vaccine research. For example, COVID-19 mRNA vaccines (mRNA-1273 by Moderna and BNT162b2 by BioNTech-Pfizer) both targeted the virus' Spike 2P protein. The S 2P protein adds "two consecutive proline residues (2P) at the beginning of the [central helix]," which was proven to be effective against a virus' S protein that was subject to quick changes in mechanism (Huang et al., 2021). Despite this change, both the Delta and Omicron variants have been able to infect and transmit to and from vaccinated hosts. The use of comparative genomics highlighted the conserved regions across variants, which can then be areas of targeting. Ultimately, there was a need for more comparative genomics within vaccine research to identify weak areas of the virus; hypothetically, if a vaccine which could target the most conserved region of the SARS-CoV-2 genome was found, then the antigens induced via the vaccine could prevent COVID-19 as well as other strains with the same conserved regions. The current outlook on COVID-19 indicated another seasonal disease, much like H1N1 (the flu), however, an ideal outlook for future vaccine research was for the development of a single vaccine that could prevent infection across all SARS-CoV-2 strains.

Bibliography

Aldaais, E. A., Yegnaswamy, S., Albahrani, F., Alsowaiket, F., & Alramadan, S. (2021). Sequence and structural analysis of COVID-19 E and M proteins with MERS virus E and M proteins—A comparative study. *Biochemistry and Biophysics Reports*, *26*, 101023. https://doi.org/10.1016/j.bbrep.2021.101023

CDC. (2022, March 29). *Omicron Variant: What You Need to Know*. Centers for Disease Control and Prevention. https://www.cdc.gov/coronavirus/2019-ncov/variants/omicron-variant.html

Chan, Agnes. P., Choi, Y., & Schork, N. J. (2020). CONSERVED GENOMIC TERMINALS OF SARS-COV-2 AS CO-EVOLVING FUNCTIONAL ELEMENTS AND POTENTIAL THERAPEUTIC TARGETS. *BioRxiv*, 2020.07.06.190207. https://doi.org/10.1101/2020.07.06.190207

Galperin, M., & Frishman, D. (1999). *Open Reading Frame—An overview | ScienceDirect Topics*. https://www.sciencedirect.com/topics/neuroscience/open-reading-frame

Huang, Q., Zeng, J., & Yan, J. (2021). COVID-19 mRNA vaccines. *Journal of Genetics and Genomics*, *48*(2), 107–114. https://doi.org/10.1016/j.jgg.2021.02.006

Jungreis, I., Sealfon, R., & Kellis, M. (2021). SARS-CoV-2 gene content and COVID-19 mutation impact by comparing 44 Sarbecovirus genomes. *Nature Communications*, *12*(1), 2642. https://doi.org/10.1038/s41467-021-22905-7

Kandeel, M., Mohamed, M. E. M., Abd El-Lateef, H. M., Venugopala, K. N., & El-Beltagi, H. S. (2022). Omicron variant genome evolution and phylogenetics. *Journal of Medical Virology*, *94*(4), 1627–1632. https://doi.org/10.1002/jmv.27515

Knyazev, S., Chhugani, K., Sarwal, V., Ayyala, R., Singh, H., Karthikeyan, S., Deshpande, D., Baykal, P. I., Comarova, Z., Lu, A., Porozov, Y., Vasylyeva, T. I., Wertheim, J. O., Tierney, B. T., Chiu, C. Y., Sun, R., Wu, A., Abedalthagafi, M. S., Pak, V. M., … Mangul, S. (2022). Unlocking capacities of genomics for the COVID-19 response and future pandemics. *Nature Methods*, *19*(4), 374–380. https://doi.org/10.1038/s41592-022-01444-z

Liu, B., Liu, K., Zhang, H., Zhang, L., Bian, Y., & Huang, L. (2020). CoV-Seq, a New Tool for SARS-CoV-2 Genome Analysis and Visualization: Development and Usability Study. *Journal of Medical Internet Research*, *22*(10), e22299. https://doi.org/10.2196/22299

Ma, K., & Chen, J. (2022). Omicron XE emerges as SARS-CoV-2 keeps evolving. *The Innovation*, *3*(3), 100248. https://doi.org/10.1016/j.xinn.2022.100248

Mlcochova, P., Kemp, S. A., Dhar, M. S., Papa, G., Meng, B., Ferreira, I. A. T. M., Datir, R., Collier, D. A., Albecka, A., Singh, S., Pandey, R., Brown, J., Zhou, J., Goonawardane, N., Mishra, S., Whittaker, C., Mellan, T., Marwal, R., Datta, M., … Gupta, R. K. (2021). SARS-CoV-2 B.1.617.2 Delta variant replication and immune evasion. *Nature*, *599*(7883), 114–119. https://doi.org/10.1038/s41586-021-03944-y

Naqvi, A. A. T., Fatima, K., Mohammad, T., Fatima, U., Singh, I. K., Singh, A., Atif, S. M., Hariprasad, G., Hasan, G. M., & Hassan, Md. I. (2020). Insights into SARS-CoV-2 genome, structure, evolution, pathogenesis and therapies: Structural genomics approach. *Biochimica et Biophysica Acta. Molecular Basis of Disease*, *1866*(10), 165878. https://doi.org/10.1016/j.bbadis.2020.165878

Office, U. S. G. A. (n.d.). *Gene Sequencing Can Track COVID Variants, But High Costs and Security and Privacy Concerns Present Challenges*. Retrieved May 3, 2022, from https://www.gao.gov/blog/gene-sequencing-can-track-covid-variants%2C-high-costs-and-security-and-privacy-concerns-present-challenges

Park, S. Y., Faraci, G., Ward, P. M., Emerson, J. F., & Lee, H. Y. (2021). High-precision and cost-efficient sequencing for real-time COVID-19 surveillance. *Scientific Reports*, *11*(1), 13669. https://doi.org/10.1038/s41598-021-93145-4

Rangan, R., Zheludev, I. N., Hagey, R. J., Pham, E. A., Wayment-Steele, H. K., Glenn, J. S., & Das, R. (2020). RNA genome conservation and secondary structure in SARS-CoV-2 and SARS-related viruses: A first look. *RNA*, *26*(8), 937–959. https://doi.org/10.1261/rna.076141.120

Simonetti, M., Zhang, N., Harbers, L., Milia, M. G., Brossa, S., Huong Nguyen, T. T., Cerutti, F., Berrino, E., Sapino, A., Bienko, M., Sottile, A., Ghisetti, V., & Crosetto, N. (2021). COVseq is a cost-effective workflow for mass-scale SARS-CoV-2 genomic surveillance. *Nature Communications*, *12*(1), 3903. https://doi.org/10.1038/s41467-021-24078-9

WHO Coronavirus (COVID-19) Dashboard. (2022, April 27). https://covid19.who.int