STAT 520A Project Report - A Bayesian method for SARS-CoV-2 variant risk analysis

Matthew Nguyen

April 2021

1 Introduction

The COVID-19 pandemic has to date caused by the SARS-CoV-2 virus has to date affected over 150 million individuals, with over 3 million deaths [1]. As the virus continues to spread, mutations such as single nucleotide polymorphisms (SNPs) accumulate in its genome. Studies have been performed to investigate the association between SARS-CoV-2 variants and severe health outcomes [2, 3]. However, these studies only determine associations through genome-wide association studies (GWAS) or by using a frequentist method such as logistic regression. Here, we introduce a Bayesian method for viral variant risk analysis, specifically applied to SARS-CoV-2. The method estimates the relative risk of variants given a number of independent samples (or sequences). We base our method off of Bayesian methods for variant risk analysis in humans [4, 5, 6, 7], which are meant to be applied to human GWAS results for associations between variants and human diseases (e.g. autism, schizophrenia). These methods require separated case and control cohorts to evaluate high risk variants, whereas we are able to predict this directly from the available data. Moreover, many variant risk analysis studies depend only on the annotated SNPs, which contain information on the predicted severity and risk of an independent SNP. However, using annotations as the only criteria can lead to both false positives and false negatives as the frequency of a variant may also be important in assessing its relative risk. Therefore, we introduce two separate methods, one which does not require annotations and one which incorporates annotations as a prior but both taking into consideration the frequency of a variant. All the code for this work is publicly available on Github: https://github.com/matnguyen/STAT520A-Bayesian-Variant-Risk-Project.

2 Methods

2.1 Data Acquisition and Preprocessing

All complete high coverage ($< 1\%$ N's and $< 0.05\%$ unique amino acid mutations) SARS-CoV-2 genome sequences that included patient status metadata was downloaded from GISAID, yielding 26 855 sequences. Using the patient status, samples were classified as being severe (terms: hospitalized, inpatient, deceased, severe) or mild (terms: outpatient, asymptomatic, mild, home, not hospitalized). Many samples were excluded because they contained vastly different patient status terms, so the final patient cohort consisted of 8921 samples.

2.2 Mutation Calling and Annotation

The 8921 SARS-CoV-2 sequences were aligned to the Wuhan-1 reference sequence using MAFFT [8]. SNPs were called from the multiple sequence alignment using SNP-sites [9] yielding 3989 SNPs. The SNPs were then annotated using SnpEff [10]. SnpEff annotates each variant with a potential risk (low, moderate, high and modifier for intergenic regions). We kept both coding and non-coding variants in the analysis, as studies have shown that both types of SNPs can have an effect on viral replication [11, 12].

2.3 Unannotated Model

The input for the unannotated model is a matrix X with rows i containing different variants, and columns j containing variables of interest. We denote the number of samples in the dataset by N (in our case $N = 8921$), X_i and $X_i^{(0)}$ as the number of samples containing the variant i and classified as having severe and mild health outcomes respectively, and $T_i = X_i + X_i^{(0)}$ as the total variant count for variant i. We also denote $F_i = \frac{T_i}{N}$ as the frequency of variant i in the dataset. Let Z_i be an indicator of whether variant i is a risk variant or not, and follows a Bernoulli distribution with

mean η. η can be interpreted as the fraction of variants in the dataset that are risk variants. We model η as a random variable following $\texttt{Beta}(\alpha, \beta)$ where α and β are shape parameters. For this method, we use $\alpha = 1$ and $\beta = 100$, as we expect the proportion of risk variants to non-risk variants is very small. Therefore, we obtain:

$$Z_i \sim \texttt{Bernoulli}(\eta) \qquad \eta \sim \texttt{Beta}(\alpha = 1, \beta = 100) \tag{1}$$

If variant i is a risk variant, we expect the number of severe cases to be higher than the number of mild cases. Let γ_i be the fold increase in frequency if variant i is a risk variant. We model γ_i as a random variable following $\texttt{Gamma}(\bar{\gamma}, \sigma)$, where $\bar{\gamma}$ is the mean of relative risk of variants and σ is the dispersion parameter. We use $\bar{\gamma} = 2$ and $\sigma = 1$ similar to in [4]. So we obtain:

$$\gamma_i \sim \texttt{Gamma}(\bar{\gamma} = 2, \sigma = 1) \tag{2}$$

Finally, we can model X_i as a random conditional variable following a Binomial distribution:

$$X_i|\gamma_i, T_i, F_i, Z_i = 0 \sim \texttt{Bin}(T_i, F_i) \qquad X_i|\gamma_i, T_i, F_i, Z_i = 1 \sim \texttt{Bin}(T_i, F_i \times \gamma_i) \tag{3}$$

The directed graphical model can be seen in Figure 1. For the model, we are most interested in inferring the posterior distribution of γ_i, as it can be interpreted as the relative risk of variant i compared to other variants in the dataset. Inference is performed using the non-reversible parallel tempering (PT) algorithm, implemented in Blang [13]. This is a parallel Markov-Chain Monte Carlo algorithm.

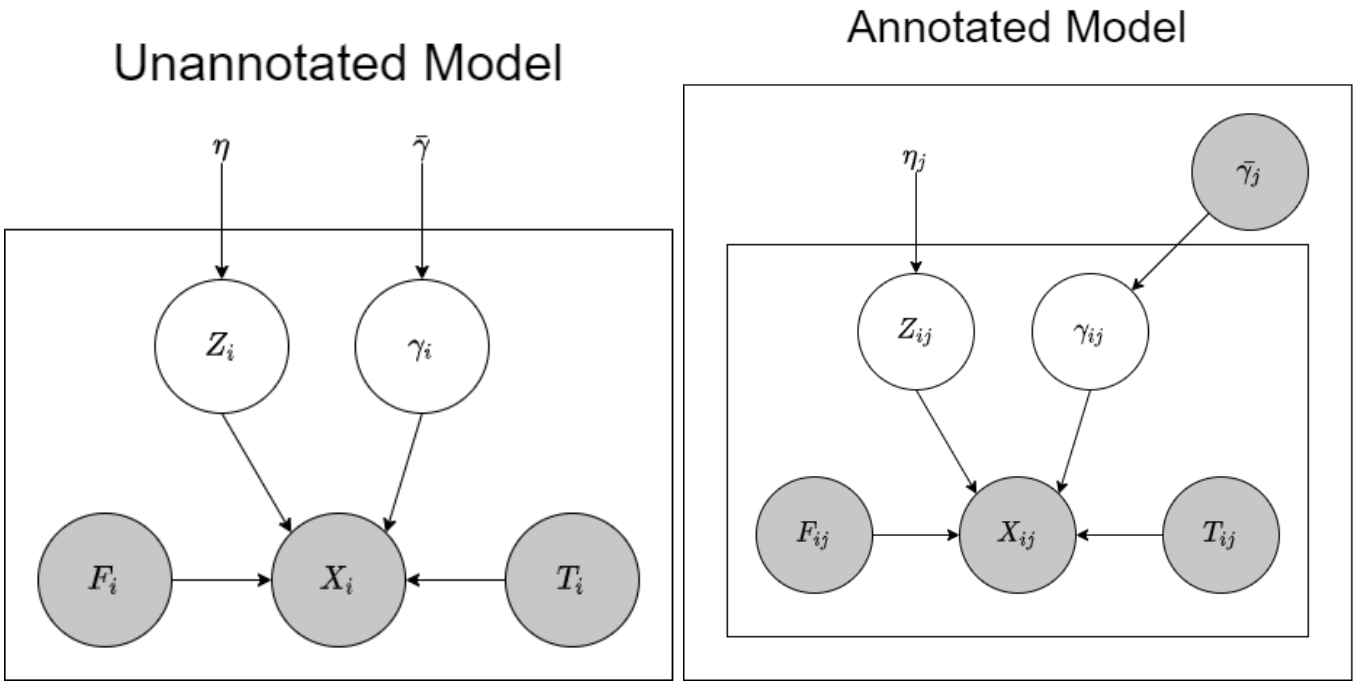

Figure 1: Directed graphical models for the unannotated and annotated models. The definitions of variables and parameters are in the text. Shaded circles refer to observed variables, whilst white circles refer to hidden variables. For the unannotated model, the box corresponds to a single variant in the dataset. For the annotated model, the outer box corresponds to a gene in the genome whilst the inner box corresponds to a single variant within the gene.

2.4 Annotated Model

The full annotated model is a hierarchical model that adds two components to the unannotated model. First, include gene information, as different genes will have different values of η (proportion of risk variants). Some genes such as the ORF1ab and Spike protein genes may be more conducive to worst prognosis [14]. Therefore, for a variant i in gene j, we have:

$$Z_{ij} \sim \texttt{Bernoulli}(\eta_j) \qquad \eta_j \sim \texttt{Beta}(\alpha = 1, \beta = 100) \tag{4}$$

Second, we add information from the variant annotations obtained by SnpEff. We encode this in the prior, by varying $\bar{\gamma}$ depending on the predicted effect. We set $\bar{\gamma} = 9$ for a high effect, $\bar{\gamma} = 5$ for a moderate effect and $\bar{\gamma} = 2$ for a low effect and for intergenic regions. Therefore, the higher effect will effectively shift the Gamma distribution such that the relative risk is higher. We include non-coding variants but set a smaller prior for them because although they may have an effect on disease severity, non-coding risk variants are much rarer than coding risk variants. The conditional distribution of X_{ij} is the same as for the unannotated model (Equation 3). The directed graphical model can be seen in Figure 1. Inference is again performed using the PT algorithm in Blang. Similarly to the unannotated model, we are interested in inferring the posterior distribution of γ_{ij}. Moreover, due to the hierarchical nature and independent modelling of η_j for each gene j, we can also infer the relative risk of each gene by inferring the posterior distribution of η_j.

2.5 Simulation and Calibration

Model calibration was performed using a simulated dataset. We only performed calibration on the unannotated model, as it was unclear how to simulate gene and annotation information and we assume that a calibrated simpler model with very similar priors would also indicate that the more complex model is likely calibrated as well. We assume a well-specified setup, and simulate 20 replicates of a 1000 variant dataset. For each dataset, we sample η from the $\texttt{Beta}(\alpha, \beta)$ with $alpha = 1$ and $\beta = 100$. We assume that risk variants are very rare. We let the cohort size of patients be 9000. For each variant in the dataset, we do the following steps:

1. We sample the risk variant indicator Z_i for variant i such that $Z_i \sim \texttt{Bernoulli}(\eta)$.
2. We sample the relative risk $\gamma_i \sim \texttt{Gamma}(\bar{\gamma}, \sigma)$, where $\bar{\gamma} = 2$ and $\sigma = 1$.
3. We sample the total variant count $T_i \sim \texttt{Poisson}(\lambda)$ where $\lambda = 32$, as the mean total variant count we observed in our dataset was 32. We add 1 to T_i to avoid counts of 0, which are possible when sampling.
4. We obtain the frequency by dividing the T_i with the cohort size of 9000.
5. We sample X_i depending on the value of Z_i. If $Z_i = 0$, then we sample $X_i \sim \texttt{Bin}(T_i, F_i)$, otherwise we sample $X_i \sim \texttt{Bin}(T_i, F_i \times \gamma_i)$

We then perform inference with our unannotated model using the PT algorithm in Blang, and compare the obtained γ_i values and its highest density interval to the true γ_i values form the simulated dataset.

3 Results

3.1 Top risk variants between both models were concordant

Potential risk variants can be obtained by extracting variants with the highest inferred median γ point estimate. The top 20 risk variants from both the unannotated and annotated models were quite concordant, with only 2 variants that differed, whilst the top 10 risk variants were the same for both models (but with some different orderings). Moreover, the highest risk variant is the same for both models: the G26144T substitution in the ORF3a gene. It is also interesting to note that there is only one non-coding mutation that is present in the top 20 risk variants from both models: a G29540T substitution upstream from the ORF10 gene. Table 1 shows the top 20 risk variants that are in common in both models. We observe that in these top variants, transitions (mutations that are interchanges between adenine and guanine or cytosine and thymine) are the most common, with C $\rightarrow$ T transitions being the most common. The only transversions observed in the top 20 are guanine to thymine substitutions, with the highest risk variant being such a transversion. These trends were also seen in [3] where they identified high risk genes through frequentist methods.

3.2 Putative high risk genes were obtained from the annotated model

By looking at the η_i values for each gene i inferred from the annotated model, we can determine which genes can be considered risk variants. Figure 2 shows the posterior density of η_i for each gene. We see that ORF1ab has the highest posterior probability ($\sim 0.93\%$), with the next highest being the Spike protein gene S ($\sim 0.78\%$) followed by the nucleocapsid protein gene N ($\sim 0.63\%$). All other genes, except for ORF3a have a posterior probability less than 0.5, with the upper end of the highest density interval also being below 0.5. Interestingly, of the top 20 risk variants, 2 are within the M structural protein gene, whereas none are within the N gene. However, the posterior distribution takes into account all variants and not just the top 20 high risk ones. Therefore, we can say that the according to our model, only ORF1ab, S and N are putative high risk genes in the SARS-CoV-2 genome.

Variants	Gene	Variant Type	Effect	Number of severe cases	Number of mild cases	Frequency
G26144T	ORF3a	missense_variant	MODERATE	104	8	0.012564505
C11916T	ORF1ab	missense_variant	MODERATE	61	5	0.0074040830000000005
A26530G	M	missense_variant	MODERATE	63	6	0.007740633
C25731T	ORF3a	synonymous_variant	LOW	83	13	0.010769576000000001
C5055T	ORF1ab	missense_variant	MODERATE	56	3	0.006618801999999999
G17427T	ORF1ab	synonymous_variant	LOW	61	2	0.007067534
C2416T	ORF1ab	synonymous_variant	LOW	61	7	0.00762845
G24812T	S	missense_variant	MODERATE	53	3	0.006282253000000001
C14786T	ORF1ab	missense_variant	MODERATE	52	2	0.006057886
C18744T	ORF1ab	synonymous_variant	LOW	56	6	0.006955350999999999
C26895T	M	missense_variant	MODERATE	55	13	0.00762845
C23185T	S	synonymous_variant	LOW	171	1	0.01929549
C7765T	ORF1ab	synonymous_variant	LOW	63	20	0.009311195999999999
C17690T	ORF1ab	missense_variant	MODERATE	56	19	0.008413730999999999
C14724T	ORF1ab	synonymous_variant	LOW	46	5	0.005721337
G29540T	ORF10	upstream_gene_variant	MODIFIER	43	0	0.004823873
C23731T	S	synonymous_variant	LOW	57	22	0.008862464
G4006T	ORF1ab	missense_variant	MODERATE	46	13	0.006618801999999999

Table 1: The top 20 risk variants that are common in both unannotated and annotated models.

3.3 Calibration

After performing 20 inference replicates for simulated datasets with 1000 variants, we were able to observe how calibrated the model is. We performed calibration on the unannotated model, and compared the true γ_i value to inferred value of $gamma_i$ and its highest density interval (HDI). We found that the true γ_i value was within the inferred HDI only 77% of the time. This means that the model is not as highly calibrated as desired. Ideally we would like to see the (HDI) to be calibrated at least 90% of the time. However, this is still not a terrible calibration, and means the HDI is still somewhat calibrated, and our model can still be considered to be valid.

4 Discussion

In this work, we present a method for viral variant risk analysis applied to SARS-CoV-2 data. Using our method, we are able to model the heterogeneity of variant effects, such as the sparsity of risk variants and the varying risk of different genes. We provide two different models, one which only takes into account the frequency and the number of severe cases (the unannotated model) and one which also encodes the potential effect of a mutation as a prior (the annotated model).

Both models were able to predict high risk variants associated with the severity of COVID-19 health outcomes. The top variants were validated by looking at current literature on SARS-CoV-2 variant analysis. We identified the G26144T substitution in the ORF3a gene as the highest risk variant from both models. This is a missense variant that changes a glycine into a valine in the protein sequence. This mutation has also been shown as significantly associated with disease severity [3, 15, 16, 17]. ORF3a has been linked to disease pathogenesis [18, 19] and has been stipulated that the protein has an association with the host immune response through the JAK-STAT, chemokine and cytokine-related pathways. Therefore, it may be related in elevated disease severity because of the higher potential of cytokine storm. The specific G251V substitution in the protein sequence resulting from this variant has also been shown to drastically change the structure of the ORF3a protein, and thus may alter protein-protein interactions. The second highest risk variant, C11916T in the ORF1ab gene has also been shown to have a strong association with increased genome-wide mutation load [20], which may lead to severe health outcomes. Other inferred top risk variants have also been demonstrated as significant in other studies using frequentist methods. However, this is the first method that has employed a Bayesian method for this type of analysis.

Using the annotated model, we were also able to identify putative high risk genes that may be involved in more severe health outcomes. The 4 genes with posterior probability greater than 0.5 were ORF1ab, S, N and ORF3a in order of descending probability. ORF1ab has been shown to be involved in cellular signaling and viral replication, but has also been speculated to be involved in viral pathogenesis through currently known methods [21]. The spike protein is involved in cell entry and is a key mechanism for immune evasion, cell infectivity and spread of the virus [22]. The

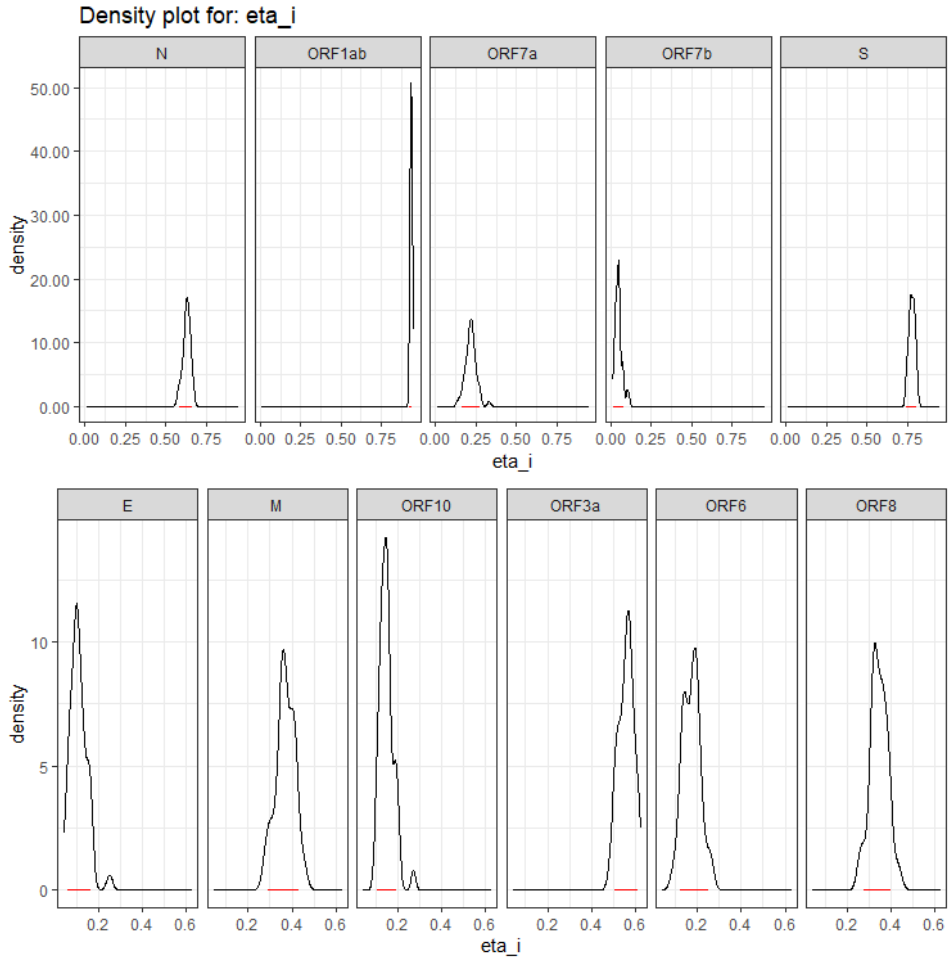

Figure 2: Posterior densities for η_i for each gene i, inferred from the hierarchical annotated model. This can be interpreted as the probability that a gene is considered a risk gene, where variants in the gene are more likely to lead to a more severe health outcome. Red lines indicate the highest density interval.

nucleocapside protein (N) plays a role in viral genome replication and cell-signaling pathways, and also plays a role in host immune responses [23]. The relevance of ORF3a has been described above. We see that the inference of these genes as potential risk genes are also validate by current research.

The model presented in this work can be a stepping stone for the application of Bayesian methods to viral variant risk analysis. Although the model is not ideally highly calibrated (if we aim for a calibrated 90% credible interval), it appears to perform fairly well and is able to infer possible high risk variants and genes. The model can be further fine-tuned given more time and resources (as run-time was an issue with over 3000 variants). The models also have limitations that may be interesting to address in future work. Firstly, the sample size of patients was quite small, and many variants were incredibly rare, with many occurring only once in 8000 patients. It is possible that the current models are mis-specified, but this can be mitigated by increasing the sample size. Secondly, the models do not incorporate further information about the impact of the variant, such as its effect on the protein sequence, or whether it is a transition or a transversion. This type of information may be crucial in determining a variant's impact on pathogenesis, as downstream products from a gene may be important to annotate and include in the model as priors. Finally, our method merely extracts potential risk variants, but does not predict the probability of severity given the variants from a given sample. Therefore, this method can only be used in an exploratory manner, rather than as a diagnostic or clinical tool. However, the basis of this method may be extended to formulate a Bayesian regression problem, where a direct outcome can be inferred. Nevertheless, our method demonstrates the power of a Bayesian hierarchical model for viral variant risk analysis, and can be extended beyond SARS-CoV-2 and applied to other viral pathogens.

References

[1] Ensheng Dong, Hongru Du, and Lauren Gardner. An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infectious Diseases*, 20(5):533–534, May 2020.

[2] Pakorn Aiewsakun, Patompon Wongtrakoongate, Yuttapong Thawornwattana, Suradej Hongeng, and Arunee Thitithanyanont. Sars-cov-2 genetic variations associated with covid-19 severity. *medRxiv*, 2020.

[3] Jameson D. Voss, Martin Skarzynski, Erin M. McAuley, Ezekiel J. Maier, Thomas Gibbons, Anthony C. Fries, and Richard R. Chapleau. Variants in sars-cov-2 associated with mild or severe outcome. *medRxiv*, 2020.

[4] Shengtong Han, Nicholas Knoblauch, Gao Wang, Siming Zhao, Yuwen Liu, Yubin Xie, Wenhui Sheng, Hoang T. Nguyen, and Xin He. A bayesian method for rare variant analysis using functional annotations and its application to autism. *bioRxiv*, 2019.

[5] Xin He, Stephan J. Sanders, Li Liu, Silvia De Rubeis, Elaine T. Lim, James S. Sutcliffe, Gerard D. Schellenberg, Richard A. Gibbs, Mark J. Daly, Joseph D. Buxbaum, Matthew W. State, Bernie Devlin, and Kathryn Roeder. Integrated model of de novo and inherited genetic variants yields greater power to identify risk genes. *PLoS Genetics*, 9(8):e1003671, August 2013.

[6] Nengjun Yi and Degui Zhi. Bayesian analysis of rare variants in genetic association studies. *Genetic epidemiology*, 35(1):57–69, Jan 2011. 21181897[pmid].

[7] Hoang T. Nguyen, Julien Bryois, April Kim, Amanda Dobbyn, Laura M. Huckins, Ana B. Munoz-Manchado, Douglas M. Ruderfer, Giulio Genovese, Menachem Fromer, Xinyi Xu, Dalila Pinto, Sten Linnarsson, Matthijs Verhage, August B. Smit, Jens Hjerling-Leffler, Joseph D. Buxbaum, Christina Hultman, Pamela Sklar, Shaun M. Purcell, Kasper Lage, Xin He, Patrick F. Sullivan, and Eli A. Stahl. Integrated bayesian analysis of rare exonic variants to identify risk genes for schizophrenia and neurodevelopmental disorders. *Genome Medicine*, 9(1):114, Dec 2017.

[8] Tsukasa Nakamura, Kazunori D Yamada, Kentaro Tomii, and Kazutaka Katoh. Parallelization of MAFFT for large-scale multiple sequence alignments. *Bioinformatics*, 34(14):2490–2492, March 2018.

[9] Andrew J. Page, Ben Taylor, Aidan J. Delaney, Jorge Soares, Torsten Seemann, Jacqueline A. Keane, and Simon R. Harris. SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microbial Genomics*, 2(4), April 2016.

[10] P. Cingolani, A. Platts, M. Coon, T. Nguyen, L. Wang, S.J. Land, X. Lu, and D.M. Ruden. A program for annotating and predicting the effects of single nucleotide polymorphisms, snpeff: Snps in the genome of drosophila melanogaster strain w1118; iso-2; iso-3. *Fly*, 6(2):80–92, 2012.

[11] Irwin Jungreis, Rachel Sealfon, and Manolis Kellis. Sars-cov-2 gene content and covid-19 mutation impact by comparing 44 sarbecovirus genomes. *bioRxiv*, 2020.

[12] Eric C. Rouchka, Julia H. Chariker, and Donghoon Chung. Variant analysis of 1,040 sars-cov-2 genomes. *PLOS ONE*, 15(11):1–18, 11 2020.

[13] Alexandre Bouchard-Côté, Kevin Chern, Davor Cubranic, Sahand Hosseini, Justin Hume, Matteo Lepur, Zihui Ouyang, and Giorgio Sgarbi. Blang: Bayesian declarative modelling of arbitrary data structures, 2019.

[14] Rozhgar A. Khailany, Muhamad Safdar, and Mehmet Ozaslan. Genomic characterization of a novel SARS-CoV-2. *Gene Reports*, 19:100682, June 2020.

[15] Alejandro Lopez-Rincon, Alberto Tonda, Lucero Mendoza-Maldonado, Eric Claassen, Johan Garssen, and Aletta D. Kraneveld. A missense mutation in SARS-CoV-2 potentially differentiates between asymptomatic and symptomatic cases. April 2020.

[16] Anastasis Oulas, Maria Zanti, Marios Tomazou, Margarita Zachariou, George Minadakis, Marilena M. Bourdakou, Pavlos Pavlidis, and George M. Spyrou. Generalized linear models provide a measure of virulence for specific mutations in SARS-CoV-2 strains. *PLOS ONE*, 16(1):e0238665, January 2021.

[17] Shuvam Banerjee, Shrinjana Dhar, Sandip Bhattacharjee, and Pritha Bhattacharjee. Decoding the lethal effect of SARS-CoV-2 (novel coronavirus) strains from global perspective: molecular pathogenesis and evolutionary divergence. April 2020.

[18] Elio Issa, Georgi Merhi, Balig Panossian, Tamara Salloum, and Sima Tokajian. SARS-CoV-2 and ORF3a: Nonsynonymous mutations, functional domains, and viral pathogenesis. *mSystems*, 5(3), May 2020.

[19] Parinita Majumdar and Sougata Niyogi. ORF3a mutation associated with higher mortality rate in SARS-CoV-2 infection. *Epidemiology and Infection*, 148, 2020.

[20] Doğa Eskier, Aslı Suner, Yavuz Oktay, and Gökhan Karakülah. Mutations of SARS-CoV-2 nsp14 exhibit strong association with increased genome-wide mutation load. *PeerJ*, 8:e10181, October 2020.

[21] Rachel L. Graham, Jennifer S. Sparks, Lance D. Eckerle, Amy C. Sims, and Mark R. Denison. SARS coronavirus replicase proteins in pathogenesis. *Virus Research*, 133(1):88–100, April 2008.

[22] Jian Shang, Yushun Wan, Chuming Luo, Gang Ye, Qibin Geng, Ashley Auerbach, and Fang Li. Cell entry mechanisms of SARS-CoV-2. *Proceedings of the National Academy of Sciences*, 117(21):11727–11734, May 2020.

[23] Chris R. Triggle, Devendra Bansal, Hong Ding, Md Mazharul Islam, Elmoubashar Abu Baker Abd Farag, Hamad Abdel Hadi, and Ali A. Sultan. A comprehensive review of viral characteristics, transmission, pathophysiology, immune response, and management of SARS-CoV-2 and COVID-19 as a basis for controlling the pandemic. *Frontiers in Immunology*, 12, February 2021.