Large-scale mapping of antigenic relationships in Sars-CoV-2

Peter C. Jentsch[1,4], Finlay Maguire[3,5], and Samira Mubareka[1,2]

[1] *Sunnybrook Research Institute, Toronto, Canada*
[2] *University of Toronto, Toronto, Canada*
[3] *Dalhousie University, Halifax, Canada*
[4] *Simon Fraser University, Burnaby, Canada*
[5] *Shared Hospital Laboratory, Toronto, Canada*

November 8, 2022

1 Background

The rapid and widespread adoption of immunization has saved millions of lives in the COVID-19 pandemic. As the world's population gains immune experience with Sars-CoV-2, either through previous infection or vaccination, antigenic drift gives rise to new variants that exhibit significant immune escape [19]. Similar to the annual reformulation of influenza vaccinations, public health planning has shifted resources towards increasing resilience against the emergence of SARS-CoV-2 variants. The Moderna and Pfizer mRNA vaccine boosters were updated accordingly to a bivalent version, including mRNA encoding both the ancestral spike protein, and the spike protein from the BA.1 lineage of the Omicron variant [4]. In the mean time, the BA.4 and BA.5 Omicron sublineages have become widespread, becoming the majority of COVID-19 cases by July 2022 [17].

Since these sublineages are also antigenically distinct from their BA.1 ancestors [3], development of bivalent mRNA boosters which cover BA.4 and BA.5 quickly followed [1]. The uptake of updated vaccine boosters will undoubtedly reduce the burden of new variants on health systems worldwide, but perhaps we can do better than a reactive approach. The level of genomic surveillance during this pandemic has been unprecedented, and many new methods have been developed to take advantage of the quantity of data collected. Understanding this immune landscape and the dynamics of viral evolution within it will be key to remaining in control of the resulting antigenic arms race.

Mapping antigenic relationships between related pathogens was pioneered with Influenza A [8]. The immune response of an antiserum to a different antigen can be quantified as a reduction in concentration of the newly introduced antigen, known as antigenic distance. These relationships can be measured

between many related strains of a pathogen to create a map of changing immune responses through the evolution of the pathogen. Antigenic space consists of a set of points between which antigenic distances can be defined. With manifold reduction techniques such as multidimensional scaling, antigenic relationships can often be well approximated in two or three dimensional cartesian maps, clearly showing the pathogen moving through antigenic space as it evolves. This technique was used to map the antigenic evolution of Influenza A, and argue that it can be understood as primarily occurring in only two dimensions [8, 13]. Wilks et al. have developed upon this technique to create an antigenic map of approximately 17 major variants of Sars-CoV-2 from serum neutralization assays, also showing that a two-dimensional map is an adequate approximation of the Sars-CoV-2 antigenic landscape [10, 16, 18].

This work on data-driven approximations of antigenic space motivates the development of models that explicitly incorporate data on antigenic relationships obtained from these neutralization assays. The dynamics of related pathogen strains evolving within a shared host population can rapidly become intractable, and therefore researchers have devised models which can manage this complexity. One method is to constrain the possible viral strains to points on a finite one or two dimensional lattice, thereby making the analysis of strain evolution tractable [5] . In two dimensional strain space, cross-immunity of a pathogen is specified by an arbitrary function of the euclidean distance between strains, and mutation is implemented as diffusion with at a fixed rate. To accurately parameterize these models, the map of antigenic space should use as much genomic information as is available. Fortunately, for Sars-CoV-2, the most sequenced biological entity, we have a huge amount of information. With millions of the samples in the global SCV2 tree, our combined knowledge of the antigenic space of SCV2 could be orders of magnitude more detailed. In this manuscript, we describe methods for constructing more detailed antigenic maps using a subset this genomic diversity, and provide some examples of these maps.

2 Methods

To add viral genomes to our map without obtaining explicit antigenic distances via laboratory assays, we need some method of estimating antigenic distances from genomic data. There are two additional sources of information that we use for this interpolation. Deep mutational scanning of the receptor binding domain (RBD) of the Sars-CoV-2 genome has been used to measure the polyclonal antibody binding affinity for a mutation at every nucleotide in the RBD [14]. Using this data, it is possible to approximate the polyclonal binding affinity for an arbitrary given RBD sequence [6]. This algorithm is gives the binding affinity for a given RBD is returned by the aformentioned tool as a fraction between zero and one, where a binding affinity of one represents perfect binding and zero represents no binding, or complete antibody escape. While antigenic distance is more complex than just antibody binding affinity to the viral RBD, we include it as one aspect of our antigenic distance approximation.

The second metric we use is the number of SNPs between two genomes, because identical viral genomes should receive very similar immune reactions, and genomes that have many mutations between them will likely have provoke different immune reactions. However, the degree to which a given SNP will affect the antigenic escape of the mutated virus depends significantly on the location and base of the SNP in the genome. For instance, since the spike protein is the main antigen of the virus, mutations in the spike protein are far more relevant than mutations in other proteins [7]. Although there are certainly exceptions, we think that mutations which are sampled more frequently in the global phylogenetic tree of Sars-CoV-2 might be more associated in some way with immune escape. To that end, we weight SNPs by the frequency with which they independently reoccur within in the global tree [2]. This is done by analyzing homoplasic sites in the global SCV2 tree created with USHER [15], using the software tool matUtils [9].

For each pair of genomes g_i and g_j $(i \neq j)$ aligned with the reference, first we find the closest lineages in the existing antigenic map to g_i, and g_j using the Pango lineage assignments of g_i and g_j. Precisely, given the set of lineages examined by [18], we first approximate the position of new genomes by mapping them to the most recent ancestor lineage present in the map using their Pango lineage assignment, computed with the Pangolin software tool [11]. Using these closest mapped lineages, we use their mapped coordinates x_i and x_j as a first approximation for the antigenic distances. Assume the polyclonal binding affinity is given by $B(g)$ (computed by [6]), where g is some viral genome sequence. Further let $g_i^{(k)}$ be the kth nucleotide base in g_i and $\chi(g_i^{(k)}, g_j^{(k)})$ the two argument indicator function given by equation 1.

$$\chi(a,b) = \begin{cases} 1 & \text{if } a = b \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

Assume $h^{(k)}$ is a vector containing the number of homoplasic mutations at site k in the global tree, and $W > 0$ is a real number denoting the weighting of the extra distances. Then we compute the distance between g_i and g_j as given by equation 2.

$$d(g_i, g_j) = \|x_i - x_j\| + W \frac{B(g_i) + B(g_j)}{2} \frac{\sum_k \chi\left(g_i^{(k)}, g_j^{(k)}\right) h^{(k)}}{\sum_k h^{(k)}} \tag{2}$$

Computing $d(g_i, g_j)$ for every pair of Sars-CoV-2 genomes g_i and g_j gives a distance matrix between all genomes in the sample. We use multidimensional scaling (MDS) as in [8,18] to place the genomes in 2 dimensional euclidean space in a way that closely approximates their distances with respect to d.

The dataset we use for this is the set of public genomes made available by the NCBI [12]. To avoid sampling bias we take only the genomes sequenced as a result of random sampling, for a total of 1,895,794 unique Sars-CoV-2 genomes. However, we only consider the SNPs in the RBD, plus the top 100 most frequent

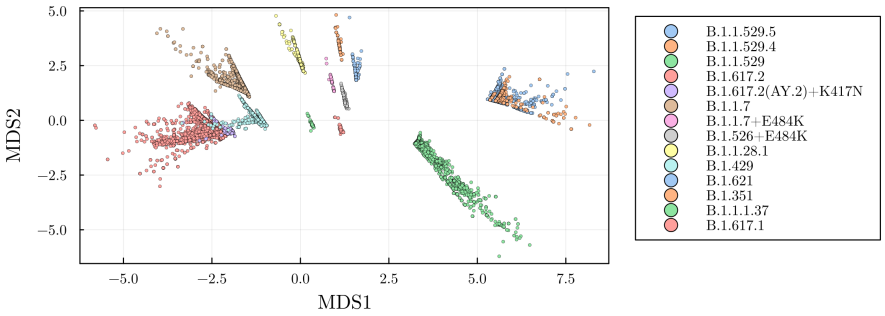

Figure 1: 2D scaling map of 18533 unique Sars-CoV-2 genomes sampled in the USA at the time of writing, accounting for only the 100 most reoccurring SNPs in the global tree.

SNPs with respect to h, because the weight assigned to other SNPs is largely negligible. This simplification reduces the total number of unique genomes to 18533, which is much more tractable for processing with commonly available MDS software packages.

3 Results

The 2-dimensional MDS map showing all 18533 unique genomes up to the 100 most reoccurring SNPs is shown in Figure 1. Clusters corresponding to the points on Wilks' map can be seen clearly, with additional diversity arranged around each point. The size of the clusters generally correspond to the parameter W, here chosen to be 10000. To assess the fit of the MDS map, we can plot the sum of the squared differences between the actual distances and the projected euclidean distances in a given output dimension (Figure 2). We see that for values of $W < 10000$, an output of dimension of 2 provides the lowest output stress, but beyond that the weight of the additional distance is large enough that 3 dimensions minimize the stress. The smallest stress is given by $W = 0$, since the map points x_i are already embedded in a 2 dimensional plane.

4 Discussion

References

[1] Pfizer and biontech granted fda emergency use authorization of omicron ba.4/ba.5-adapted bivalent covid-19 vaccine booster for ages 12 years and older. `https://www.pfizer.com/news/press-release/press-release-detail/`

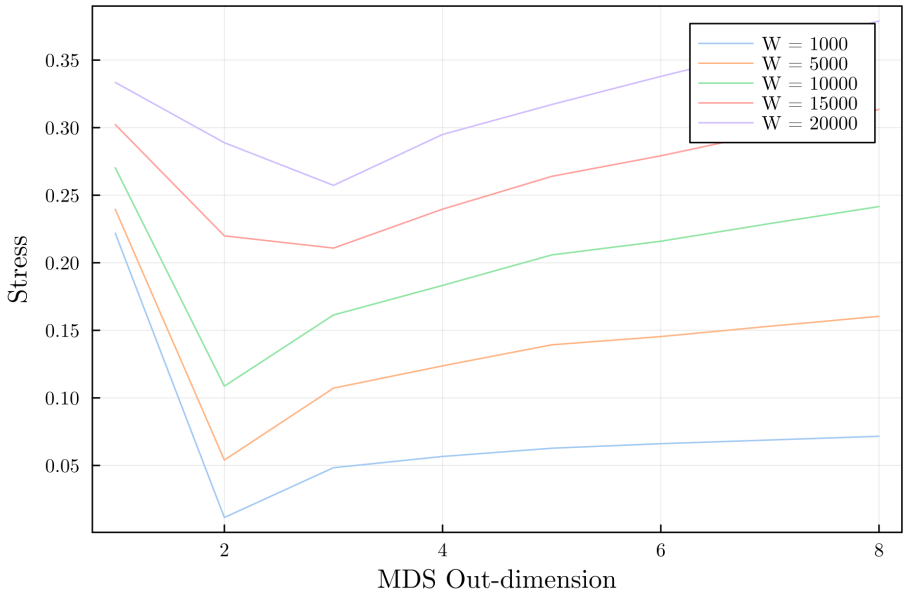

Figure 2: An output dimension of 2 minimizes the stress in the multidimensional scaling dimension reduction.

`pfizer-and-biontech-granted-fda-emergency-use-authorization`, 2022.

[2] Stephen W Attwood, Sarah C Hill, David M Aanensen, Thomas R Connor, and Oliver G Pybus. Phylogenetic and phylodynamic approaches to understanding and combating the early sars-cov-2 pandemic. *Nature Reviews Genetics*, pages 1–16, 2022.

[3] Yunlong Cao, Ayijiang Yisimayi, Fanchong Jian, Weiliang Song, Tianhe Xiao, Lei Wang, Shuo Du, Jing Wang, Qianqian Li, Xiaosu Chen, et al. Ba. 2.12. 1, ba. 4 and ba. 5 escape antibodies elicited by omicron infection. *Nature*, 608(7923):593–602, 2022.

[4] Spyros Chalkias, Charles Harper, Keith Vrbicky, Stephen R Walsh, Brandon Essink, Adam Brosz, Nichole McGhee, Joanne E Tomassini, Xing Chen, Ying Chang, et al. A bivalent omicron-containing booster vaccine against covid-19. *New England Journal of Medicine*, 2022.

[5] J. R. Gog and B. T. Grenfell. Dynamics and selection of many-strain pathogens. *Proceedings of the National Academy of Sciences*, 99(26):17209–17214, December 2002.

[6] Allison J Greaney, Tyler N Starr, and Jesse D Bloom. An antibody-escape estimator for mutations to the sars-cov-2 receptor-binding domain. *Virus evolution*, 8(1):veac021, 2022.

[7] William T Harvey, Alessandro M Carabelli, Ben Jackson, Ravindra K Gupta, Emma C Thomson, Ewan M Harrison, Catherine Ludden, Richard Reeve, Andrew Rambaut, Sharon J Peacock, et al. Sars-cov-2 variants, spike mutations and immune escape. *Nature Reviews Microbiology*, 19(7):409–424, 2021.

[8] Alan Lapedes and Robert Farber. The Geometry of Shape Space: Application to Influenza. *Journal of Theoretical Biology*, 212(1):57–69, September 2001.

[9] Jakob McBroome, Bryan Thornlow, Angie S Hinrichs, Nicola De Maio, Nick Goldman, David Haussler, Russell Corbett-Detig, and Yatish Turakhia. matutils: Tools to interpret and manipulate mutation annotated trees. *bioRxiv*, 2021.

[10] Nathaniel L. Miller, Thomas Clark, Rahul Raman, and Ram Sasisekharan. An Antigenic Space Framework for Understanding Antibody Escape of SARS-CoV-2 Variants. *Viruses*, 13(10):2009, October 2021.

[11] Áine O'Toole, Emily Scher, Anthony Underwood, Ben Jackson, Verity Hill, John T McCrone, Rachel Colquhoun, Chris Ruis, Khalil Abu-Dahab, Ben Taylor, et al. Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool. *Virus evolution*, 7(2):veab064, 2021.

[12] Eric W Sayers, Richa Agarwala, Evan E Bolton, J Rodney Brister, Kathi Canese, Karen Clark, Ryan Connor, Nicolas Fiorini, Kathryn Funk, Timothy Hefferon, et al. Database resources of the national center for biotechnology information. *Nucleic acids research*, 47(Database issue):D23, 2019.

[13] Derek J. Smith, Alan S. Lapedes, Jan C. de Jong, Theo M. Bestebroer, Guus F. Rimmelzwaan, Albert D. M. E. Osterhaus, and Ron A. M. Fouchier. Mapping the Antigenic and Genetic Evolution of Influenza Virus. *Science*, 305(5682):371–376, July 2004.

[14] Tyler N Starr, Allison J Greaney, Sarah K Hilton, Daniel Ellis, Katharine HD Crawford, Adam S Dingens, Mary Jane Navarro, John E Bowen, M Alejandra Tortorici, Alexandra C Walls, et al. Deep mutational scanning of sars-cov-2 receptor binding domain reveals constraints on folding and ace2 binding. *Cell*, 182(5):1295–1310, 2020.

[15] Yatish Turakhia, Bryan Thornlow, Angie S Hinrichs, Nicola De Maio, Landen Gozashti, Robert Lanfear, David Haussler, and Russell Corbett-Detig. Ultrafast sample placement on existing trees (usher) enables real-time phylogenetics for the sars-cov-2 pandemic. *Nature Genetics*, 53(6):809–816, 2021.

[16] Karlijn van der Straten, Denise Guerra, Marit van Gils, Ilja Bontjer, Tom G Caniels, Hugo D van Willigen, Elke Wynberg, Meliawati Poniman, Judith A Burger, Joey H Bouhuijs, et al. Mapping the antigenic diversification of sars-cov-2. *medRxiv*, 2022.

[17] Alexander Wilhelm, Shelesh Agrawal, Jens Schoth, Christina Meinert-Berning, Daniel Bastian, Laura Orschler, Sandra Ciesek, Burkhard Teichgräber, Thomas Wintgens, Susanne Lackner, et al. Early detection of sars-cov-2 omicron ba. 4 and ba. 5 in german wastewater. *Viruses*, 14(9):1876, 2022.

[18] Samuel H. Wilks, Barbara Mühlemann, Xiaoying Shen, Sina Türeli, Eric B. LeGresley, Antonia Netzl, Miguela A. Caniza, Jesus N. Chacaltana-Huarcaya, Xiaoju Daniell, Michael B. Datto, Thomas N. Denny, Christian Drosten, Ron A. M. Fouchier, Patricia J. Garcia, Peter J. Halfmann, Agatha Jassem, Terry C. Jones, Yoshihiro Kawaoka, Florian Krammer, Charlene McDanal, Rolando Pajon, Viviana Simon, Melissa Stockwell, Haili Tang, Harm van Bakel, Richard Webby, David C. Montefiori, and Derek J. Smith. Mapping SARS-CoV-2 antigenic relationships and serological responses. Preprint, Immunology, January 2022.

[19] Jonathan W. Yewdell. Antigenic drift: Understanding COVID-19. *Immunity*, 54(12):2681–2687, December 2021.