Signatures of functional responses to acute and chronic COVID-19 infections

Rhea M. Grover
*University of California, Berkeley**
(Dated: August 2023)

Post-acute sequelae of COVID-19 (PASC), characterized by lingering symptoms of disease after the acute COVID-19 infection has passed, affects anywhere from 31-69% of individuals infected with COVID-19. It is unknown why some individuals develop PASC; we currently lack a quantitative description of the differences between individuals with and without PASC. In this project, we analyzed TCR sequencing information from a group of individuals with and without PASC. We detected TCR clonal expansion and contraction between time points during the infection and analyzed the probability of a clone being shared between patients, with the goal of finding predictors of PASC based on the dynamics of an individual's immune repertoire.

I. INTRODUCTION

Post-acute sequelae of COVID-19 (PASC), is an emerging problem following the recent SARS-CoV-2 pandemic. PASC is a complication of COVID-19, the disease caused by SARS-CoV-2, which can have severe and life changing effects on the lives of those who experience it. Much is still unknown about the causes of PASC; we do not know exactly how many people have PASC, who will develop it after COVID-19 infection, the mechanisms that lead to development of PASC, or even all of the manifestations of PASC.

In the future, we hope to be able to predict someone's chance of developing PASC based on their immune repertoire and other health markers. In this project, we analyze TCR sequencing data from a group of individuals infected with SARS-CoV-2, with varying levels of disease severity and PASC status and symptoms. By analyzing T cell repertoires, we hope to characterize the immune response to acute and chronic COVID-19 to understand more about PASC.

A. Post acute sequelae of COVID-19 (PASC)

Post acute sequelae of COVID-19 (PASC), commonly known as long COVID, is characterized by lingering symptoms of illness for 4 or more weeks after infection with SARS-CoV-2 [1]. Researchers have identified over 200 symptoms associated with PASC [2]. Symptoms include respiratory, gastrointestinal, and neurological difficulties, and anosmia/dysgeusia (loss/diminishment of smell/taste). Anywhere from 31-69 % of individuals who are infected with COVID-19 will go on to develop PASC. The reasons some individuals with COVID-19 go on to develop PASC, and the mechanisms behind PASC, are still not fully understood; with so many people living with PASC after the recent pandemic of COVID-19, understanding more about PASC is an important research question.

B. T cells

T cells are part of the adaptive immune system, and respond to pathogens to fight off infections. They identify infected cells through the presentation of peptides (protein fragments) by major histocompatibility complex (MHC) molecules. MHC molecules on the surface of cells display peptides being produced in the cell to T cells, specifically their T cell receptors (TCRs). If a TCR binds to the protein fragment and the cell is understood to be infected with a pathogen, the cell will either be destroyed by the T cell or flagged for destruction. There are several types of T cells, which serve various functions. Killer (cytotoxtic) T cells destroy infected cells. Helper T cells secrete signaling molecules that aid in, among other functions, activating B cells, another type of adaptive immune system cell which neutralizes pathogens and secretes antibodies. Regulatory T cells prevent the immune system from responding incorrectly [3].

Each T cell displays a single kind of TCR and is covered in thousands of them. TCRs have two main components, the α chain and the β chain. Each chain, and thus TCR, is produced as a result of combinatorics and stochastic processes. TCRβ chains are generated by V(D)J gene recombination in which alleles of the V, D, and J genes are combined and spliced together by deleting and inserting nucleotides randomly between the V and D genes and the D and J genes. [4]. The part of the chain on which recombination occurs is called the CDR3 region. A TCR can be identified uniquely by its V genes, J genes, and CDR3 nucleotide sequences.

* Research performed in the Nourmohammad Group, Physics Department, University of Washington

A group of T cells with a particular TCR is called a clone. It is estimated that V(D)J recombination alone results in anywhere between 10^{15} to 10^{61} distinct TCR clones [5].

Not all of the TCRs produced via recombination will effectively detect cells harmful to the body, i.e., non-self. The body attempts to make sure that T cells do not recognize the body's own cells as non-self, which would lead to autoimmune disease. T cells must therefore go through a round of selection after recombination to ensure that only those that detect infected cells make it into the body. A current theory is that TCRs are tested against self-antigen (proteins produced in the body) for binding strength. TCRs that bind too weakly are unlikely to effectively detect infected cells while those that bind too strongly to self-antigen may mistakenly detect healthy cells as non-self. Cells on either side of this spectrum undergo apoptosis [3] However, recent work has called this theory into question. Among other observations, the timescale on which T cells undergo selection—four to five days—is much too small for them to be tested against all self-antigens. Thus, selection is thought to be a leaky process, with some autoreactive T cells passing through successfully. [6].

After all is said and done, 10^{12} T cells survive functional selection [5]. Due to sampling constraints, it is unknown exactly how many distinct clonotypes there are [7]; however, recent estimates have put the number of distinct TCRs between 10^6 and 10^8 [4].

T cells which have passed selection are ready to start recognizing and responding to infections. Once a T cell has detected an infected its cognate antigen, it becomes activated. The TCR sends a signal to the nucleus of the T-cell, setting off a chemical cascade which leads to TCRs on the surface clustering and sending various signals based on the circumstance. The activated T-cell then proliferates rapidly, expanding its clonal population. After the infection is gone, the clone population contracts, except for some memory T cells which persist in the body and are easier to activate if the pathogen is encountered again [3].

C. About the data

This project used the INCOV cohort from the study described in [8]. The data used for this project was obtained from individuals infected with COVID-19 at Swedish Medical Center in Seattle, WA. This cohort contained 124 individuals who had been tested for PASC, consisting of 63 individuals without PASC (28 male and 35 female) and 61 with PASC (24 male and 37 female). The participants

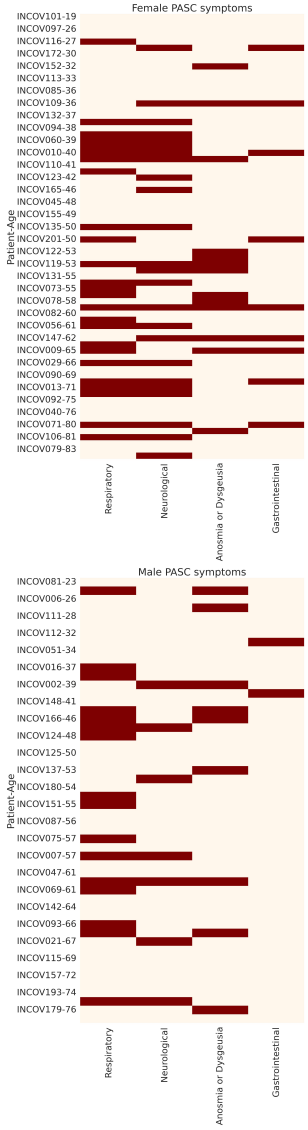

FIG. 1. Heatmaps showing PASC symptoms experienced by male and female individuals with PASC. Dark boxes correspond to experiencing that symptom.

were aged between 19 and 86 years old at the time of the study and had a wide variety of disease severities and preexisting conditions.

Blood samples were obtained at three time points during the infection: baseline (BL), taken at time of clinical diagnosis of COVID-19; acute (AC), taken at the peak of infection; and convalescent (CV), 2-3 months after first symptoms of disease [8]. The distributions of these sampling times are shown in Fig. 2.

Bulk TCR sequencing from collected blood samples was performed at Adaptive Biotechnologies in Seattle, WA. The sequencing data gave us information about the β chains of specific clones found in

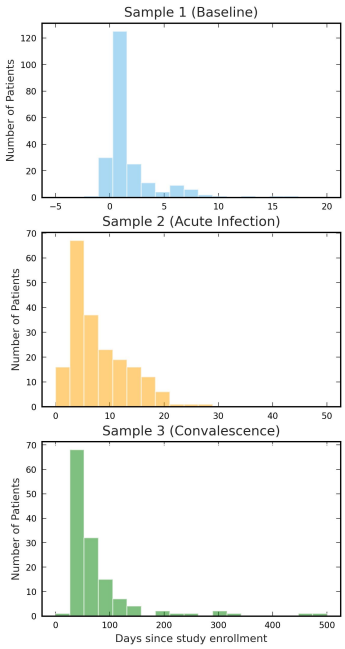

FIG. 2. Time each sample was taken. Most baseline samples were taken 0-5 days after study enrollment, most acute samples were taken 5-15 days after study enrollment, and most convalescent samples were taken 22-100 days after study enrollment.

the sample, which included their CDR3 sequence, their V-gene and J-gene, and the number of times the clone appears in the sample (the read count). On average, there were about 275,000 unique TCR clones in each sample.

Individuals were tested for PASC symptoms during the CV time point by looking for symptoms of PASC that fell into four categories: respiratory, gastrointestinal, and neurological difficulties, and anosmia and dysgeusia. We considered an individual to have PASC if they displayed at least one of these four symptoms. A more detailed look at the PASC symptoms experienced by individuals in this study can be seen in Fig. 1. Of the 124 individuals tested for PASC, some had samples taken at all three time points; others only had samples from the AC and CV time points. We only considered individuals who had a sample from the CV time point. Samples were taken from individuals experiencing a wide variety of disease severity and with a wide range of preexisting conditions.

More information about this data can be found in [8].

D. Goals

The goal of this project was to understand more about the immune response of patients with COVID-19. Using T cell repertoire data, we aimed to identify predictors of PASC from an individual's immune response. We aimed to describe clonal expansion and contraction between time points in our data sequence features of the expanded and contracted clones, and analysis of what phenomena were shared between demographic groups, to identify which clones were responding to COVID-19. We also looked at the immune responses in four groups, females with PASC, females without PASC, males with PASC, and males with PASC, to look for differences associated with sex and PASC status, because there are differences in immune repertoires between females and males [9]. In doing these analyses, we hoped to understand more about the individuals that went on to develop PASC after COVID-19 infection, in particular how their T-cell repertoires differed from those individuals without PASC.

II. TCR POPULATION DYNAMICS ANALYSIS

A. Read Counts and Undersampling

The sequencing data contains information about the clone identities in each sample and the number of times the clone was sequenced. So given one sample from each time point, why not just compare the read counts for each clonotype and see which ones are larger or smaller? Indeed, PCA on the read count trajectories (Fig. 3) shows many read counts get much larger or smaller between time points; there is quite a bit of fluctuation in read counts between samples. However, the process of identifying expanding or contracting clone populations is not so simple.

There are about 10^{12} T cells in the body [10], but only 10^6 are sampled and sequenced. Clone populations follow a power law distribution, $\rho(f) = Cf^{-2}$, for some constant C (Fig. 4), which means many clones have low frequency in the immune repertoire. Because T cells are highly undersampled, there is a lot of variation in clone reads between samples, unrelated to clonal expansion or contraction. Fig. 5 illustrates this variation.

This variation in read counts means that we must distinguish between clones that have a higher or lower read count due to experimental noise from undersampling, and clones that have a higher or lower read count because the clone population has expanded or contracted. Dealing with this experimental noise is a key challenge in clonal population

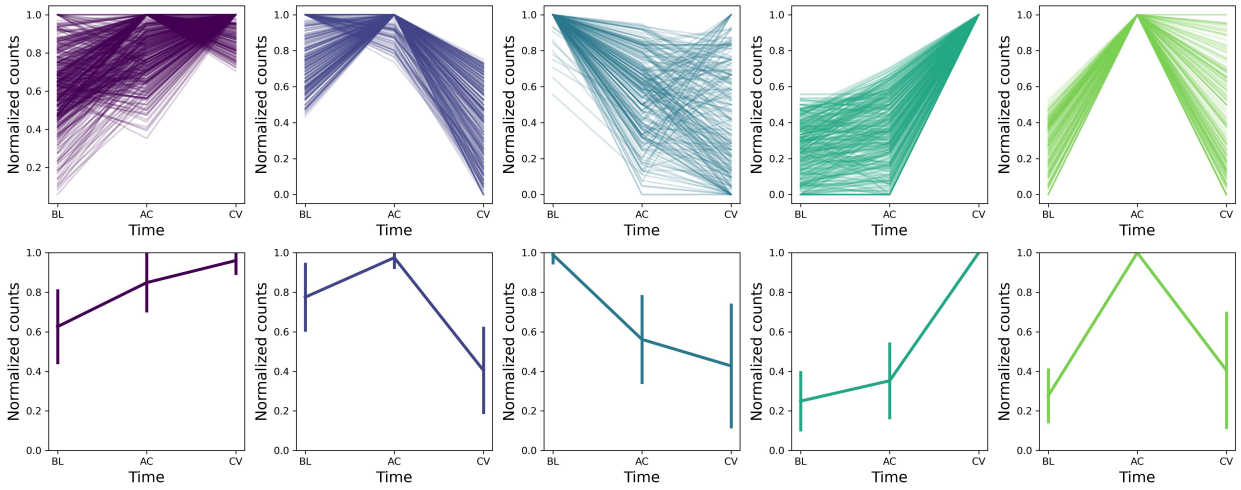

FIG. 3. Trajectories of clone read counts for top 1000 most abundant clonotypes across all patients. Clustered into groups showing roughly the same behavior over time using principal component analysis.

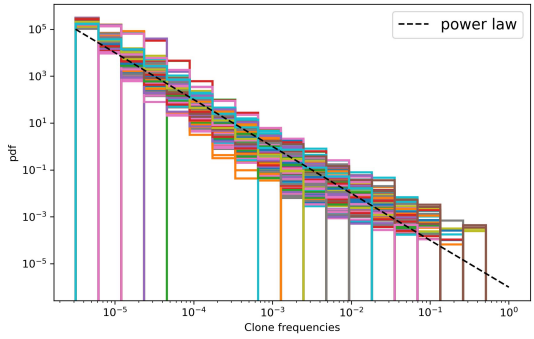

FIG. 4. Clone size distribution follows a power law, $\rho(f) = Cf^{-2}$

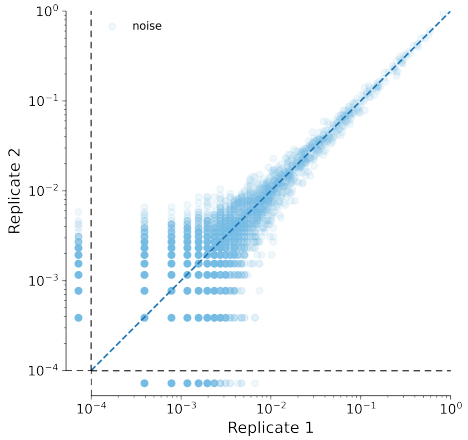

FIG. 5. Variation in read counts between two biological replicates (samples taken at the same time from the same person). Because the samples are taken at the same time, clone populations have not expanded or contracted; therefore, any variation in the number of reads is due to noise.

dynamics analysis.

B. NoisET

To detect clonal expansion and contraction, this project used the NoisET software [11]. NoisET is a Python package which uses Bayesian inference to learn experimental noise and predict which clones are most likely to have undergone expansion or contraction between two time points. NoisET does this in two steps. First, NoisET learns the null (noise) model from two biological replicates, modeling the immune repertoire with one of three probability distributions chosen by the user. After learning null model parameters, NoisET then performs the expansion/contraction calculations. NoisET calculates a p-value that the read counts for a particular clone at each time point are due to experimental noise (i.e., that the clone population did not expand or contract). The user then selects the significant p-values. More information about NoisET can be found in [11].

NoisET allows the user to choose between three probability distributions to model the experimental noise between two replicates: a Poisson distribution, a negative binomial distribution, and a distribution that combines the two and attempts to account for the number of reads of a clone in a sample vs. the actual number of the clone in the sample (m), since

one cell may contribute multiple reads to the data. These distributions are as follows [11]:

$$P(\hat{n}|f) = \text{Pois}(fN_r)$$

for N_r the total number of reads in the sample

$$P(\hat{n}|f) = \text{NegBin}(\hat{n}; N_r f, N_r f + a(N_r f)^b)$$

for a and b parameters dictating the variance, learned when calculating noise parameters

$$P(\hat{n}|f) = \sum_{m_i}^{\infty} P(\hat{n}|m_i)P(m_i|f)$$

where M is the number of T-cells in a sample (as opposed to the number of reads), fM is the average population of a clone with frequency f in a populaton of M cells, $P(m_i|f) = \text{NegBin}(m_i; fM, fM + a(fM)^b)$ and $P(\hat{n}|m_i) = \text{Pois}(m_i N_r/M)$.

After NoisET has learned the parameters of the experimental noise, it goes on to calculate which clones are most likely to have expanded or contracted given the null model. By modeling the frequency of a clone at two points in time as $f(t_1) = f$ and $f(t_2) = fe^s$, NoisET calculates the probability of the observed read counts given the fraction of responding clones γ and the average selection factor $\bar{s}$ as follows [11]:

$$P((\hat{n_i}(t_1) = \hat{n_1}, \hat{n_i}(t_2) = \hat{n_2})|\gamma, \bar{s}) =$$
$$\iint df\rho(f)dsP(s|\gamma, \bar{s})P(\hat{n_1}|f)P(\hat{n_2}|fe^s)$$

where $P(n|f)$ are the noise distributions learned in the earlier step, $\rho(f)$ is the power-law frequency distribution, and $P(s|\gamma, bars$ is the prior on s. NoisET then calculates the probability of seeing the observed s given the two read counts at each time, using Bayes' Rule:

$$P(s|\hat{n_1}, \hat{n_2}) = \frac{P(\hat{n_1}, \hat{n_2}|s, \gamma, \bar{s})P(s|\gamma, \bar{s})}{P(\hat{n_1}, \hat{n_2})}$$

This results in a p-value for each clone which represents the probability that the clone population has expanded or contracted, given the null model. The user then chooses which p-values are significant.

C. Using NoisET on our data

One of the challenges in using NoisET on our data came in the first step, learning the null model. NoisET is designed to learn the null model from two biological replicates. However, our data did not include biological replicates; we only had one sample per time point for each individual. In order to learn the null model, we turned to subsampling. We initially started with hypergeometric sampling based on the number of reads of each clone, with 20,000 reads total in each in silico replicate. However, we realized that this was too small a sample to obtain adequate results, as these "replicates" did not contain clone populations as small as would be seen in an actual sample.

We also considered using unproductive sequences to train the noise model. Unproductive sequences are those sequences that do not produce a functional TCR. This usually occurs because the sequences are out of frame for transcription or because they contain a premature stop codon. These sequences are found in cells that have functional TCRs, because the unproductive sequence is encoded by one chromosome and the productive sequence is encoded by the other. [4]. We knew that unproductive sequences travel with productive sequences, but we thought we might try using them to train on noise, since they do not generate functional TCRs that would respond to infection. However, after further analysis, in which we performed PCA on the population trajectories over time for both productive and unproductive sequences, we found that the behavior of the unproductive sequences was too similar to that of productive sequences, reflecting the fact that they "hitchhike" along with productive sequences, which expand and contract in response to disease.

After considering these two ideas, we turned to the idea of data thinning [12, 13]. Data thinning is a method to separate a random variable X into multiple random variables follow the same distribution as the original follow the same distribution as the larger sample.

By treating each clone read count as a random variable X, we sampled without replacement to obtain two "replicates" which summed to the original [14]:

$$X^{(1)} + X^{(2)} = X$$

We chose to use this method for creating in silico replicates over other methods for several reasons. First, the larger sample included clone counts with low frequencies, which were left in our earlier method of only sampling 20,000 reads. Second, this method ensured that the smaller samples followed the same clone frequency distribution as the original. Third, the number of significant clonal expansions and contractions detected with this method were closer to what we would expect to see, compared to the earlier subsampling method, which detected many more expansions and contractions than expected.

After developing our data thinning technique, we had to choose which noise model to use to describe our data. Unfortunately, we were not able to per-

form the calculations for the mixed negative binomial/Poisson model because the program took too long; we then focused our attention on the other two models. We selected a random sample of individuals from each of four groups based on the individual's sex and PASC status; we then produced 100 sets of in silico replicates and calculated the Poisson and negative binomial noise parameters for each sample. We found that parameters for both noise models across each of the 100 sets of replicates were generally quite similar. Because the likelihood values and parameters between the two models were similar, we then looked at how the parameters affected the calculation of the number of expansions and contractions. For one patient from the four groups, I calculated the p-values of expansion and contraction for each of the clones for each of the 100 parameters, between the BL and AC timepoints, for each noise model. We considered a clone to have significantly expanded or contracted if it had a p-value ¡ 0.05. We found that the negative binomial noise model was more sensitive to the parameters than the Poisson noise model; there was far more variation in the numbers of significant clonal expansions/contractions detected and in which clones were detected. There were patients for which no one clone appeared in all 100 sets of calculations. This was in contrast to the Poisson model parameters, which consistently detected the same number or a close number of significant contractions and expansions, and detected the same clones in most, if not all, 100 calculations (Fig. 6). We think this consistency may be because of our data thinning method, which may be more consistent with a Poisson distribution than a negative binomial one. Because of the consistency of the detections with the Poisson noise parameters, we chose to use the Poisson noise model for the rest of of our calculations.

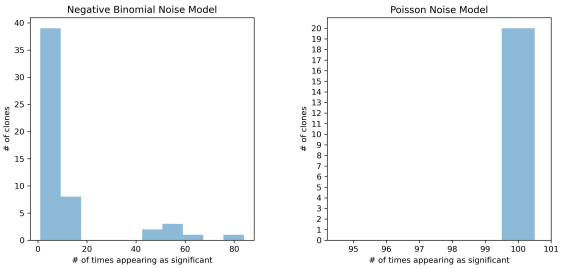

FIG. 6. Clonal expansions and contractions detected with the Poisson noise parameters were more consistent across all 100 sets of in-silico replicates than those detected with negative binomial noise parameters. For this particular individual, no clonotype was detected as contracted or expanding in all 100 calculations; however, all 100 calculations using Poisson noise parameters detected the same clones.

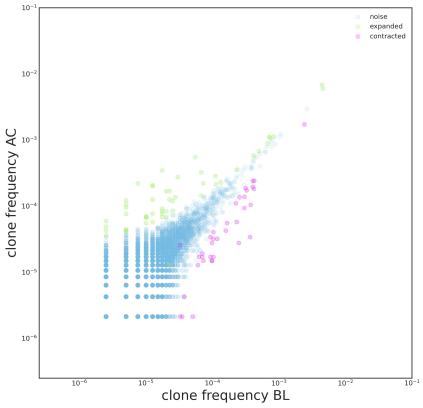

FIG. 7. Example of detection of expanding and contracting clone populations for one individual between the baseline and acute time points.

D. Final Algorithm

After refining our in silico replicate procedure and choosing a noise model, we settled on a final procedure for detecting expansions and contractions, as follows:

1. Performed data thinning 100 times to obtain 100 sets of in silico replicates for each sample.

2. Calculated 100 sets of noise parameters for each set of replicates.

3. With each of the 100 sets of noise parameters, detected expansions and contractions between two time points.

4. Intersected the set of significant clones obtained from each the 100 sets of noise parameters; considered clones to have significantly expanded or contracted only if they were in this set intersection.

We chose to perform data thinning 100 times to increase confidence in the results, as sampling is a random process. We then chose to intersect all 100 sets of results for the same reason, and as a correction for multiple testing.

E. Results

We ran our algorithm between the BL and AC time points and AC and CV time points for each

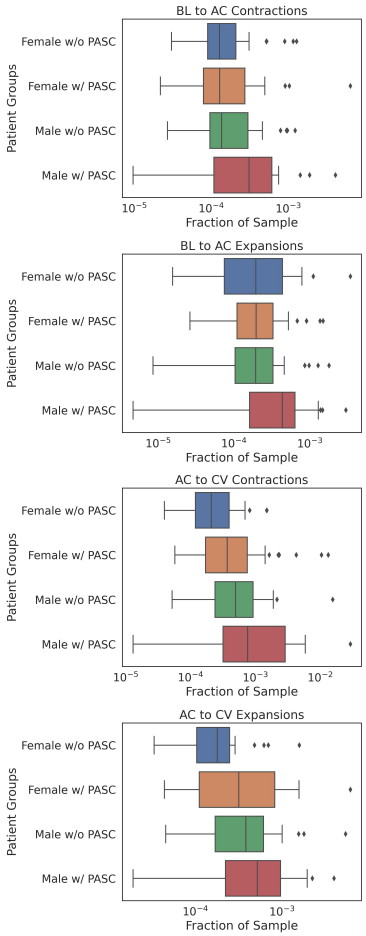

FIG. 8. Number of contractions and expansions detected between time points per demographic group

individual who had samples from those time points (see Fig. 7 for an example). The number of significant expansions and contractions found between time points for each group of individuals is depicted in Fig. 8. While each of the boxplots has significant overlap with the others, a couple of features stand out: the variance in clonal expansions and contractions in males with PASC is higher at all times, and the median for that group is also higher than for any of the other groups.

While the numbers of expansions and contractions do not conclusively suggest differences between the groups, the detection of expansion and contraction gives us information that we can continue to use to characterize immune response to COVID-19.

III. CLONE SHARING AND P_{post} ANALYSIS

Every TCR has a probability of occurring in the body, which relates to the probability of being produced through V(D)J recombination and surviving functional and self-antigen testing. In analyzing the probability of a clone occurring in the body, we can find those clones that have a rare chance of occurring, but also occur in many individuals in our data. There are a few reasons that clones might be found in multiple individuals. The first is convergent recombination; since all human produce clones through V(D)J recombination, which has specific biases, certain clones are more likely to be found in many people [15]. Other reasons for clone sharing include experimental biases from the sequencing method, and similar exposure histories; if multiple individuals have been exposed to a certain pathogen, they will have a larger population of responding clones than those who have not been exposed to the pathogen. Therefore, clone sharing, particularly of rare clones, is an indication that those clones could be responding to infection [16].

To look for these rare clones, we used two softwares. V(D)J recombination is modeled by the software IGoR [17], which predicts the probability of generation via V(D)J recombination, P_{gen}. The selection process occurring after V(D)J recombination is modeled by the software SONIA [18], which predicts the probability of surviving selection, P_{post}.

A. Results

Before analyzing any probabilities, we first looked at the numbers of clones shared between individuals in each of our four demographic groups (Fig. 9 (a)). We combined samples from all three time points and only looked at productive sequences. Here we found a curious result. The distributions of number of clones shared between certain numbers of individuals is largely the same between the two groups of female patients (the drop off in the group with PASC can be attributed to the smaller sample size. However, the males without PASC had more sharing than the males with PASC, at all points in the distribution. Male patients also showed less sharing than female patients; however, prior research has shown that males tend to have more sharing than females as a result of producing a reduced diversity of TCRs compared to female individuals– in other words, clones in male individuals tend to have a higher P_{post} than those in female patients [9]. The result in our data does not track with this finding. However, more work must be done to examine how much of our result is due to sample sizes.

After looking at the numbers of clones shared, we looked at the distributions of P_{post} for the clones in each group Fig. 10. There were not significant differences in the P_{post} distributions between groups.

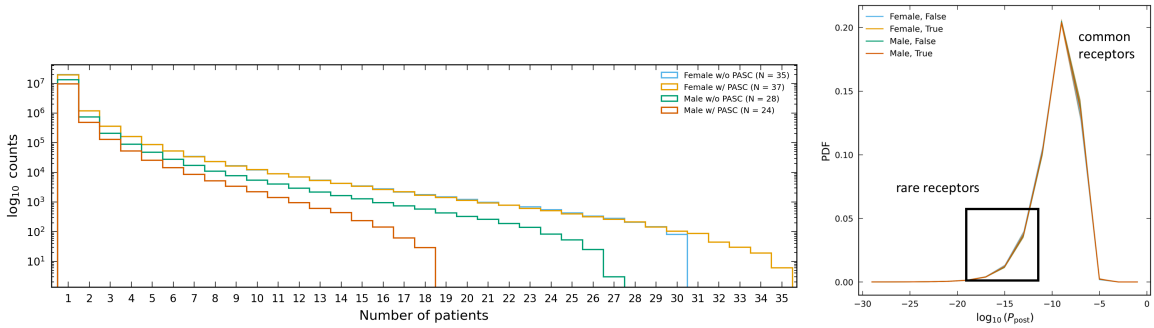

FIG. 9. (a) Distribution of clones shared between individuals in each demographic group. (b) Distribution of P_{post} for patients in each of four demographic groups. False indicates a individual without PASC; True indicates an individual with PASC. We are particularly interested in the rare receptors highlighted in the figure.

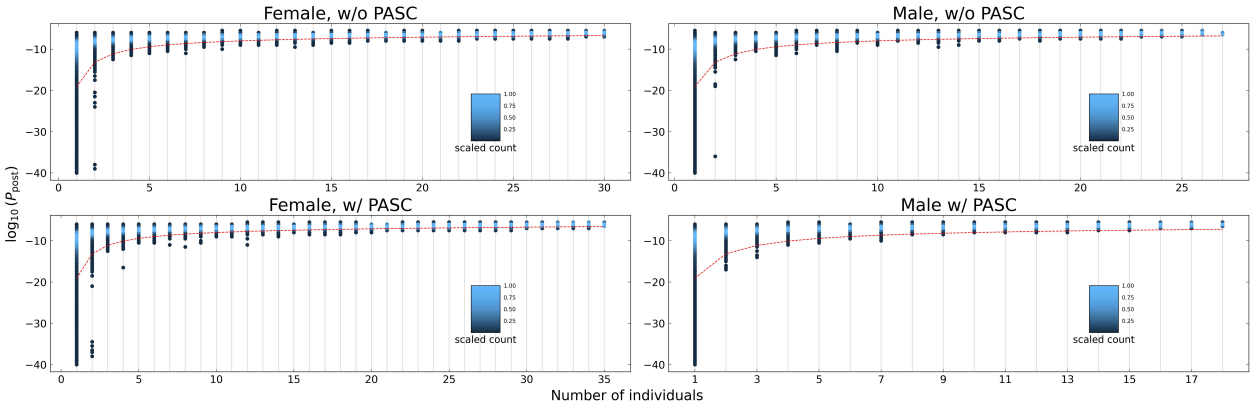

FIG. 10. Distribution of Pposts of clones shared by number of individuals in each group

We then plotted the P_{post} distributions for clones appearing in a certain number of patients. We calculated the probability of seeing a clone in a sample of N clones as follows:

$$\rho(\sigma; N) = 1 - (1 - P_{\text{post}}(\sigma))^N = 1 - e^{NP_{\text{post}}}$$

The distribution of P_{post} expected to see for clones seen in m individuals, out of a cohort of M with sample size N is a binomial distribution as follows:

$$P_{share}(\sigma; m, M, N) = \binom{M}{m}[\rho(\sigma; N)]^m[1 - \rho(\sigma; N)]^{M-m}$$

[16]. We chose the 0.05 quantile of clones as a cutoff, i.e., the rarest 5% of clones shared between each number of people. The clones below the this bound (the red line in Fig 11) are those clones that could be responding to disease; more work is needed to identify those clone sequences and connect them to our population dynamics analysis to see if they could be expanding or contracting as a response to infection.

IV. CONCLUSION AND FUTURE OUTLOOK

This project laid important groundwork towards our goal of characterizing the immune response to COVID-19 and PASC. First, we developed a data thinning method to perform population dynamics analysis in the absence of biological replicates. This method will allow that analysis to be performed on a wider array of datasets than previously possible. Second, this project identified expanding and contracting clones in all individuals in our study between all three time points. Third, we identified rare, shared clones. These identifications are an important first step in analyzing the characteristics of the immune repertoires in our cohort, and eventually towards identifying clones responding to COVID-19.

In the future, we will continue to refine our data thinning method to create in silico replicates which resemble biological replicates as closely as possible. This will involve testing analyses done with the in silico procedure against analyses done with biological replicate data. Additionally, we will test various

sampling methods to determine which is most suitable for our data and its underlying clone population distribution.

More work is needed to identify which clones may be responding to infection with SARS-CoV-2. This will involve comparing which clones appear in multiple patients and at what time points. We will also examine clones that have expanded or contracted and have a low P_{post}. Furthermore, we plan to do a network analysis with ALICE [19] on clone amino acid sequences to identify responding clones.

Additionally, we plan to integrate electronic health records into our analysis. These records include information about the symptoms and severity of COVID-19 experienced by the individuals in our study, any preexisting conditions, and other health markers. We plan to look for associations between that information and immune response.

V. ACKNOWLEDGMENTS

Thank you to my mentors, Zach Montague and Dr. Armita Nourmohammad, the Nourmohammad lab, and the University of Washington Physics REU program for this opportunity.

This work was supported by the National Science Foundation.

[1] Z. A. Sherif, C. R. Gomez, T. J. Connors, T. J. Henrich, and W. B. Reeves, Pathogenic mechanisms of post-acute sequelae of sars-cov-2 infection (pasc), Elife **12**, e86002 (2023).

[2] T. Thaweethai, S. E. Jolley, E. W. Karlson, E. B. Levitan, B. Levy, G. A. McComsey, L. McCorkell, G. N. Nadkarni, S. Parthasarathy, U. Singh, T. A. Walker, C. A. Selvaggi, D. J. Shinnick, C. C. M. Schulte, R. Atchley-Challenner, L. I. Horwitz, A. S. Foulkes, RECOVER Consortium Authors, and RECOVER Consortium, Development of a Definition of Postacute Sequelae of SARS-CoV-2 Infection, JAMA **329**, 1934 (2023).

[3] L. M. Sompayrac, *How the immune system works* (John Wiley & Sons, 2022).

[4] G. Altan-Bonnet, T. Mora, and A. M. Walczak, Quantitative immunology for physicists, Physics Reports **849**, 1 (2020).

[5] M. U. Gaimann, M. Nguyen, J. Desponds, and A. Mayer, Early life imprints the hierarchy of t cell clone sizes, Elife **9**, e61639 (2020).

[6] T. Mora and A. M. Walczak, Towards a quantitative theory of tolerance, Trends in Immunology (2023).

[7] T. Mora and A. M. Walczak, How many different clonotypes do immune repertoires contain?, Current Opinion in Systems Biology **18**, 104 (2019).

[8] Y. Su, D. Yuan, D. G. Chen, R. H. Ng, K. Wang, J. Choi, S. Li, S. Hong, R. Zhang, J. Xie, S. A. Kornilov, K. Scherler, A. J. Pavlovitch-Bedzyk, S. Dong, C. Lausted, I. Lee, S. Fallen, C. L. Dai, P. Baloni, B. Smith, V. R. Duvvuri, K. G. Anderson, J. Li, F. Yang, C. J. Duncombe, D. J. McCulloch, C. Rostomily, P. Troisch, J. Zhou, S. Mackay, Q. DeGottardi, D. H. May, R. Taniguchi, R. M. Gittelman, M. Klinger, T. M. Snyder, R. Roper, G. Wojciechowska, K. Murray, R. Edmark, S. Evans, L. Jones, Y. Zhou, L. Rowen, R. Liu, W. Chour, H. A. Algren, W. R. Berrington, J. A. Wallick, R. A. Cochran, M. E. Micikas, ISB-Swedish COVID-19 Biobanking Unit, T. Wrin, C. J. Petropoulos, H. R. Cole, T. D. Fischer, W. Wei, D. S. B. Hoon, N. D. Price, N. Subramanian, J. A. Hill, J. Hadlock, A. T. Magis, A. Ribas, L. L. Lanier, S. D. Boyd, J. A. Bluestone, H. Chu, L. Hood, R. Gottardo, P. D. Greenberg, M. M. Davis, J. D. Goldman, and J. R. Heath, engMultiple early factors anticipate post-acute COVID-19 sequelae, Cell **185**, 881 (2022).

[9] A. Trofimov, P. Brouillard, J.-D. Larouche, J. Séguin, J.-P. Laverdure, A. Brasey, G. Ehx, D.-C. Roy, L. Busque, S. Lachance, *et al.*, Two types of human tcr differentially regulate reactivity to self and non-self antigens, Iscience **25** (2022).

[10] J. Charles A Janeway, P. Travers, M. Walport, and M. J. Shlomchik, enT-cell receptor gene rearrangement, in en*Immunobiology: The Immune System in Health and Disease. 5th edition* (Garland Science, 2001).

[11] M. B. Koraichi, M. P. Touzel, A. Mazzolini, T. Mora, and A. M. Walczak, NoisET: Noise Learning and Expansion Detection of T-Cell Receptors, The Journal of Physical Chemistry A **126**, 7407 (2022), publisher: American Chemical Society.

[12] A. Neufeld, A. Dharamshi, L. L. Gao, and D. Witten, Data thinning for convolution-closed distributions (2023), arXiv:2301.07276 [stat].

[13] A. Dharamshi, A. Neufeld, K. Motwani, L. L. Gao, D. Witten, and J. Bien, Generalized Data Thinning Using Sufficient Statistics (2023), arXiv:2303.12931 [math, stat].

[14] A. Neufeld, J. Popp, L. L. Gao, A. Battle, and D. Witten, Negative binomial count splitting for single-cell rna sequencing data, arXiv preprint arXiv:2307.12985 (2023).

[15] M. V. Pogorelyy, A. A. Minervina, D. M. Chudakov, I. Z. Mamedov, Y. B. Lebedev, T. Mora, and A. M. Walczak, Method for identification of condition-associated public antigen receptor sequences, Elife **7**, e33050 (2018).

[16] Z. Montague, H. Lv, J. Otwinowski, W. S. DeWitt, G. Isacchini, G. K. Yip, W. W. Ng, O. T.-Y. Tsang, M. Yuan, H. Liu, I. A. Wilson, J. M. Peiris, N. C. Wu, A. Nourmohammad, and C. K. P. Mok, enDynamics of B cell repertoires and emergence of cross-

reactive responses in patients with different severities of COVID-19, Cell Reports **35**, 109173 (2021).

[17] Q. Marcou, T. Mora, and A. M. Walczak, High-throughput immune repertoire analysis with igor, Nature communications **9**, 561 (2018).

[18] Z. Sethna, G. Isacchini, T. Dupic, T. Mora, A. M. Walczak, and Y. Elhanati, Population variability in the generation and selection of t-cell repertoires, PLOS Computational Biology **16**, e1008394 (2020).

[19] M. V. Pogorelyy, A. A. Minervina, M. Shugay, D. M. Chudakov, Y. B. Lebedev, T. Mora, and A. M. Walczak, Detecting t cell receptors involved in immune responses from single repertoire snapshots, PLoS Biology **17**, e3000314 (2019).