CEI: A Clonal Expansion Identifier for T-cell receptor clones following SARS-CoV-2 vaccination

Yunbei Pan[a], Christian Hofmann[b], Barbara Banbury[c], Harsh Patel[c], Stephanie A. Bien[c], Tom Chou[a,d], Otto O. Yang[b]

[a]*Department of Computational Medicine, UCLA, 621 Charles E. Young Drive S., Los Angeles, 90095-1766, CA, USA*
[b]*Depts. of Medicine and Microbiology, Immunology, and Molecular Genetics, UCLA, 615 Charles E. Young Drive S., East, Los Angeles, 90095, CA, USA*
[c]*Adaptive Biotechnologies, 1165 Eastlake Ave E, Seattle, 98109, WA, USA*
[d]*Department of Mathematics, UCLA, 520 Portola Plaza, MS 6363, Los Angeles, 90095-1555, CA, USA*

Abstract

Each T cell typically carries a specific T-cell receptor (TCR) that determines its specificity against an epitope presented by the HLA complex on a target cell. Antigenic challenge triggers the expansion of reactive cells within a diverse pool of T cells with randomly generated receptors, a process that results in epitope-driven shifts of TCR frequencies over time. Here, we analyze the effects of SARS-CoV-2 vaccination on the TCR populations in peripheral blood drawn from seven COVID-naive individuals, before vaccines were widely available. To identify SARS-CoV-2 vaccine-associated TCR sequences among the $\sim 10^5 - 10^6$ TCR sequences sampled before and after vaccination, we develop statistical criteria to detect significant increases in abundance of positive TCR clones. Application of our statistical methods shows a robust identification of TCR sequences that respond to SARS-CoV-2 vaccination *in vivo*, illustrating the feasibility of quantifying the clone-specific dynamics of T-cell abundance changes following immunological perturbations.

Keywords: T-cell activation, COVID-19, vaccination, statistical identification

Introduction

mRNA vaccines that deliver SARS-CoV-2 spike protein (S protein) to the blood have been highly effective in reducing COVID-19 morbidity and mortality [1, 2, 3, 4, 5, 6, 7]. Vaccination promotes HLA-restricted T-cell responses against the S protein that play a critical role in this protective effect. These responses are mediated by TCRs recognizing epitopes from the S protein. TCRs are heterodimers [8, 9, 10, 11, 12] consisting of an α-chain (TCRA) and a β-chain (TCRB) [13]. In the generation of T cells, somatic recombination of the different domains [the variable (V) and joining (J) domains in TRA and TRB, and the diversity (D) segment in TRB] in the associated TCR sequence occurs. Along with nucleotide insertions and deletions, a complementarity-determining region 3 (CDR3) is generated with up to 10^{15} possible combinations of TCR sequence [13, 14, 15, 16].

Of this possible diversity, it has been estimated that $\sim 10^8$ unique TCR sequences are circulating in the peripheral blood of a person [17, 18, 19, 20, 21]. CTLs (cytotoxic T lymphocytes) are $CD8^+$ T lymphocytes that eliminate target cells through cytolysis and apoptosis. A second group of $CD4^+$ lymphocytes includes key "helper T cells" that direct immune responses by supporting CTLs and contributing to antigen-specific antibody production.

These T-cell responses have been shown to play a key role in mitigating the pathogenicity of SARS-CoV-2 infection [13, 22], and an early T-cell response has been associated with a mild outcome of COVID-19 [23, 24]. In contrast to rapidly waning antibody responses [4, 25], SARS-CoV-2-specific T cells persist over longer periods and provide an immediate immune response against COVID-19 re-infection [26, 27]. Furthermore, the evolution of SARS-CoV-2 into variants has been marked predominantly by mutations in and around the S protein receptor binding domain (RBD), leading to escape from neutralizing antibodies (which generally target the RBD), while T cells target epitopes distributed throughout the S protein and thus have a broader recognition of variants. The large size and abundance of the S protein make it a good target for the adaptive immune response [28]. To better understand the T-cell response against COVID-19, large libraries of TCR sequences have been analyzed using various bioengineering methods [29, 30, 31].

Quantifiable detection of vaccine-induced peptide-specific expansion of TCR clones is required in many applications. However, challenges include analyzing $\sim 10^6$ TCR clones without replicates and defining abundance changes among clones with highly dispersed abundances. In addition, different subjects may experience very different immune response intensities. These factors complicate the accurate assessment of T-cell clone responses during the post-vaccination visits and the capture of significant changes in the abundances of individual clones.

Several advanced Bayesian methods for detecting clonal expansions, such as NoisET [32] and edgeR [33], which are based on Bayesian models, and have been recently developed. The incomplete overlap in identified clones using these methods and the absence of a ground truth for evaluating their reliability motivates us to develop a new method grounded in fundamental statistical principles for a better comparative analysis.

In this study, we develop CEI, a Python package designed for the rapid, scalable detection of clonal expansion from single samples from repertoires pre- and post-vaccination (or in general, any perturbation). Applied to TCRB repertoires from blood drawn before and after mRNA/adenoviral-vectored vaccination, CEI identifies vaccine-associated TCRB sequences using a two-proportion test and analyses of differences and log-fold change of abundances. While developed for TCRB data in this work, the framework is applicable to other replicate-limited count comparisons (e.g., TCR/BCR repertoires or CRISPR screens) where feature-level proportions are contrasted across two distinct conditions.

Materials and Methods

Data preprocessing

Relevant to this study (comparing pre- and post-vaccination TCR abundances in COVID-naive subjects), samples were collected from seven individuals (labeled CLE0083, CLE0099, CLE0100, CLE0101, CLE0108, CLE0117, and CLE0136) between April 2020 and May 2021. Post-vaccination samples were taken from subjects in January-September 2021, very early in the pandemic before vaccines were widely available. These subjects received their first vaccinations between 12/18/2020 and 1/8/2021, and their second vaccinations between 01/06/2021 and 01/30/2021. Subject CLE0099 received the AstraZeneca chimpanzee adenovirus vectored vaccine while all others received the BNT162b2 mRNA vaccine. Additionally, all subjects had no prior history of COVID-19; baseline spike RBD antibodies were also confirmed negative by ELISA around the time of the first vaccination. This dataset represents a rare and scientifically valuable window into the earliest immune responses to SARS-CoV-2 vaccination, captured during a critical period at the onset of the global vaccine rollout. This temporal isolation from natural infection ensures that the observed TCR dynamics are most likely to be attributable to vaccination.

Overall, samples were taken at three time points: baseline (0-3 days before the first vaccination, labeled B), 14 to 20 days after the first vaccination but before (labeled P1), and 11 to 21 days after the second vaccination (labeled P2). Further details are provided in Table 1. The TCRs of both $CD4^+$ and $CD8^+$ T cells in these samples were sequenced by Adaptive Biotechnologies.

Post processing of immunosequencing data for differential abundance analysis

Differential abundance analysis is a statistical framework used to determine whether a particular receptor rearrangement is more abundant in one sample than in another. It can be applied to nucleotide, amino acid, or bioidentity (V gene/amino acid/J gene) sequences depending on the desired resolution and biological context. While nucleotide-level matching provides the highest resolution, amino acid-level matching allows for grouping of clonotypes that may differ at the nucleotide level but encode the same protein sequence. We employed the IMSEQ [34] algorithm to collapse nucleotide sequences for identical clonotypes for downstream differential abundance analysis of bioidentities. Since the amino acid sequence determines protein function, we established criteria to determine whether two nucleotide sequences are identical:

- If the V subtype, CDR3 amino acids, and J subtype are all identical, we consider these two sequences to be identical;
- If the two sequences differ only in their V subtype, and one or both are ambiguous, we treat them as the same if the longer sequence contains the shorter;

Subjects	Demographics					Sample times versus vaccine doses (days)		
	Age	Sex	Race	Hisp	Vaccine	B to Vacc #1	P1 to Vacc #1	P2 to Vacc #2
CLE0083	49	M	A	N	BNT162b2	-3	18	18
CLE0099	51	M	W	N	AZD1222	0	15	14
CLE0100	55	M	A	N	BNT162b2	-2	16	14
CLE0101	36	F	A	N	BNT162b2	0	14	13
CLE0108	50	F	Iranian	N	BNT162b2	-1	20	11
CLE0117	64	M	W	N	BNT162b2	0	14	15
CLE0136	42	F	W	N	BNT162b2	0	18	21
Mean (Range)						-1 (-3 to 0)	16 (14 to 20)	15 (11 to 21)

Table 1: Vaccination records for all the individuals. BNT162b2 is an mRNA vaccine, while AZD1222 is a chimpanzee adenovirus-vectored vaccine. "Vacc #1" and "Vacc #2" refer to the first and second doses of the vaccine, respectively. The timing of sample collection in relation to vaccination is denoted by sample times versus vaccine doses. For example, in "B to Vacc #1," -3 represents a pre-vaccination baseline sample that was taken three days before the first "P1" vaccination.

- If the two sequences differ by only one base (A, C, T, G) in the nucleotide sequence of the CDR3 region, we consider them identical.

For identical sequences, we sum their template counts for the subsequent analysis of changes in abundance.

Quantification of clone abundance changes

Due to the specificity of TCR binding to assigned antigen peptides, we can statistically screen for sequences that expand after vaccination by comparing with the baseline sample. Our analysis identifies TCR sequences that are enriched in response to vaccine-associated antigens (including but not limited to S protein) as "positive" and those that do not target S proteins as "negative." We assume that all T cells and SARS-CoV-2 protein-expressing cells generated by mRNA vaccination are well-mixed, and neither the exposure nor the expansion capacity of reactive T cells is limiting. Although false-positive sequences (*i.e.*, non–vaccine-associated sequences that expand) may occur, their number is relatively small compared to the true-positive sequences. Likewise, the likelihood of false-negative sequences (*i.e.*, SARS-CoV-2–specific sequences not responding) can be neglected.

Intuitively, we need to assess the magnitude of the difference in the abundances of individual clones (here, a clone denotes the collection of cells with the same V gene, CDR3 amino acid, and J gene identity), measured in samples taken before (baseline) and after vaccination. In addition, we can evaluate the intensity of the response at each post-vaccination time point. One approach is to treat the clone abundances across clone identities as a distribution and then measure the dissimilarity between the distributions at different visit times within one individual. For instance, we can consider metrics/distances such as KL divergence [35], Wasserstein distance [36, 37], cosine similarity [38], Bhattacharyya distance [39], among others. However, these approaches face challenges in effectively addressing scenarios where clones were not initially sampled at the baseline but have grown significantly after vaccination. Moreover, because these metrics typically quantify changes in the global shape of the distributions, very high-abundance clones may dominate the metrics and thereby obscure relatively small clones that have nonetheless undergone substantial growth. It is also probable that extremely large clones are not SARS-CoV-2 specific if they did not appreciably expand upon vaccination.

Here, we will focus on the changes in the abundance of individual clones between samples taken from two repertoires. We assign a fixed index $1 \leq i \leq I$ to each clone, where I is the total number of different clones observed in the two samples drawn from the same subject before (B) and after (P) vaccination. Given a clone with index i, we denote its true *in vivo*, whole body sizes at time $t \in \{B, P\}$ as integers $n_i^{(t)}$. We also define the corresponding total number of T cells at these two time points as $N^{(B)}$ and $N^{(P)}$, respectively.

Assuming $N^{(B)} \approx N^{(P)}$, consider a scenario where the abundance of a sampled clone changes from $10,000$ to $10,100$. Although the absolute difference of 100 is appreciable, if this difference arises in an already-large clone, the

relative change in abundance could be small and attributed to the sampling variability or intrinsic fluctuations in a "homeostatic" large clone. On the other hand, a fold-change from 0 to 5 is considered large (infinite in this case), but an absolute difference of 5 might also arise from noise, sampling or otherwise. Therefore, we will use both the absolute difference and the fold-change as quantities for the identification of statistically significant changes in abundance.

We motivate and develop two clone-level quantities or "indicators" that can be used to effectively screen expanded clones and measure the strength of response following a perturbation, vaccination in our case.

1. **Difference in clone abundance** The normalized change in abundances from data is defined as

$$d_i \equiv \hat{p}_i^{(P)} - \hat{p}_i^{(B)} \tag{1}$$

where $\hat{p}_i^{(t)} = \frac{s_i^{(t)}}{S^{(t)}}$ represents the abundance fraction of clone i in the sample drawn at time $t \in \{B, P\}$.

2. **Regularized log-fold change** We define the log of a regularized abundance ratio by

$$r_i \equiv \log_2 \frac{s_i^{(P)} + \varepsilon}{S^{(P)}} - \log_2 \frac{s_i^{(B)} + \varepsilon}{S^{(B)}} \tag{2}$$

where ε is a Haldane-Anscombe correction to stabilize zero counts (the default value $\varepsilon = 0.5$ is typically used). Taking the logarithm defines r_i over the entire real line ($-\infty < r_i < \infty$).

Variance, Wald statistic, and p-values for d_i

Under a simple random sampling assumption (*i.e.,* each T cell has the same probability of being sampled) with total sample size $S^{(B)}$ and $S^{(P)}$, the abundances $\left\{s_i^{(t)}\right\}$ of sampled clones i at time $t \in \{B, P\}$ follow a multinomial distribution:

$$\left(s_1^{(t)}, s_2^{(t)}, \ldots, s_I^{(t)}\right) \sim \text{Multinomial}\left(S^{(t)};\ p_1^{(t)}, p_2^{(t)}, \ldots, p_I^{(t)}\right), \qquad t \in \{B, P\} \tag{3}$$

where $p_i^{(t)} = \frac{n_i^{(t)}}{N^{(t)}}$ is the probability of the clone i being sampled at time t. Note that because very few clones from the entire body are sampled, $S^{(t)} = \sum_{i=1}^{I} s_i^{(t)} \ll N^{(t)}$ for $t \in \{B, P\}$ and $\sum_{i=1}^{I} p_i^{(t)} \ll 1$.

Since only a small number of T cell clones respond to vaccination, we consider no change in abundance, $p_i^{(B)} = p_i^{(P)} = p_i$, as the null hypothesis (H_0). We expect very few clones to violate this hypothesis.

Since the joint statistical properties for two signals (difference and log-fold change in abundance) under a true multinomial null are analytically complex, we adopt a resampling-based approach to assess clone abundance change significance. To reduce computational burden, we replace the full multinomial in Eq. 3 with a marginalized probability distribution, treating each clone independently. Specifically, considering the probability of each clone to be sampled is small and library sizes are extremely large, we use a Poisson model:

$$s_i^{(t)} \sim \text{Poisson}\left(\lambda_i^{(t)}\right), \qquad \text{where } \lambda_i^{(t)} = S^{(t)} p_i \text{ and } t \in \{B, P\}. \tag{4}$$

Since the true p_i is unknown, we approximate it using the pooled proportion estimator:

$$\hat{p}_i = \frac{s_i^{(B)} + s_i^{(P)}}{S^{(B)} + S^{(P)}} \tag{5}$$

The variance of the difference d_i for clone i is thus calculated by

$$\text{Var}(d_i) = \text{Var}\left(\frac{s_i^{(P)}}{S^{(P)}}\right) + \text{Var}\left(\frac{s_i^{(B)}}{S^{(B)}}\right) = \frac{\hat{p}_i}{S^{(P)}} + \frac{\hat{p}_i}{S^{(B)}} \tag{6}$$

To better balance the effects of the difference and log-fold change, we need to nondimensionalize the difference. Consider taking a sample from each clone as a single experiment. Instead of focusing on global Z scores, we should utilize Wald statistics to find asymptotic Z scores under H_0 for each clone separately. Therefore, for clone i, we have

$$Z_{d_i} = \frac{d_i}{\sqrt{\hat{p}_i\left(\frac{1}{S^{(P)}} + \frac{1}{S^{(B)}}\right)}} \tag{7}$$

The one-sided p-value $p_{d_i} = 1 - \Phi(Z_{d_i})$ for the clone i is used to detect expansion. Here $\Phi(x) = \frac{1}{2}[1 + \text{erf}(x/\sqrt{2})]$ is the cumulative density function of the standard normal distribution.

Variance, Wald statistic, and p-values for log-fold change r_i

Considering the empirical abundance fraction $s_i^{(t)}/S^{(P)}$ of clone i in sample $t \in \{B, P\}$. As assumed previously, the count of each clone follows a Poisson distribution. Given the definition of the log-fold change $r_i = \log_2 \tilde{p}_i^{(P)} - \log_2 \tilde{p}_i^{(B)}$, the null hypothesis for this case is supposed to satisfy $H_0 : r_i = 0$ (no change upon vaccination). By assuming the independence between visits, the multivariate delta method yields the following variance:

$$\mathrm{Var}(r_i) \approx \frac{1}{(\ln 2)^2} \left(\frac{1}{\tilde{p}_i^{(P)} S^{(P)}} + \frac{1}{\tilde{p}_i^{(B)} S^{(B)}} \right) \tag{8}$$

where $\tilde{p}_i^{(t)} = \frac{s_i^{(t)} + \varepsilon}{S^{(t)}}$ ($t \in \{B, P\}$) are empirical proportions smoothed by a Haldane-Anscombe correction ε (default $\varepsilon = 0.5$).

Applying the Wald test gives us the asymptotic Z scores:

$$Z_{r_i} = \frac{r_i}{\sqrt{\mathrm{Var}(r_i)}} = \frac{\ln 2 \left(\log_2 \tilde{p}_i^{(P)} - \log_2 \tilde{p}_i^{(B)} \right)}{\sqrt{\frac{1}{\tilde{p}_i^{(P)} S^{(P)}} + \frac{1}{\tilde{p}_i^{(B)} S^{(B)}}}} \tag{9}$$

which is approximately normally distributed when H_0 holds. In this case, the one-sided p-value is found from $p_{r_i} = 1 - \Phi(Z_{r_i})$.

CEI-Joint method

Since both differences in abundance and log-fold changes between two samples are considered important in identifying amplified clones, it is natural to consider their associated Z scores as components of a two-dimensional vector $Z_i = (Z_{d_i}, Z_{r_i})^\top$.

To include standardized statistics for both differences and log-fold changes into a single test statistic, we can consider a score based on a modified Mahalanobis distance for each clone i,

$$D_i^2 \equiv (Z_i - m)^\top \, \Sigma^{-1} \, (Z_i - m) \tag{10}$$

where Σ denotes a covariance matrix

$$\Sigma = \mathbb{E}\left[(Z_{d_i} - m_d)(Z_{r_i} - m_r) \right] \tag{11}$$

Here, we use the trimmed median vector m after removing the largest and smallest 10% of clones by $\|Z_i\|_\infty$. Vaccine stimulation leads to a very small number of clones experiencing significant expansion, induces modest contraction in several others, but leaves the majority of clones statistically unchanged. Consequently, the median serves as a more stable representation of the "no-change" baseline than the mean.

We control false positives by testing our CEI-Joint method with a multinomial permutation under the no-change null hypothesis. By applying the pooled clone frequencies $\hat{p}_i$, we redraw entire libraries as $s_{i,r}^{(B)} \sim \mathrm{Multinomial}(S^{(B)}, \hat{p}_i)$ and $s_{i,r}^{(P)} \sim \mathrm{Multinomial}(S^{(P)}, \hat{p})$ in each permutation $r = 1, \cdots, R$. For each permutation, we recompute the same per-clone statistics (*i.e.*, Z_{d_i}, Z_{r_i}) and the 2D Mahalanobis distance. We then pool all permuted distances across clones to obtain the null distribution. The empirical p-value for clone i is calculated as follows:

$$\hat{p}_i = \frac{1 + \#\{D^2_{\mathrm{null}} \geq D_i^2\}}{1 + IR} \qquad \text{(with } I \text{ clones, } R \text{ permutations)} \tag{12}$$

This formula avoids zero p-values and accounts for dependence between the two Z scores. Finally, we apply the Benjamini-Hochberg (BH) procedure to control the false discovery rate (FDR) [40, 41]. By setting an FDR threshold, denoted as α (*e.g.*, $\alpha = 10\%$). Given a total of I clones and their corresponding p-values, denoted as $p_1, p_2, \ldots, p_I$, we sort these p-values in ascending order to obtain $p_{r_1}, p_{r_2}, \ldots, p_{r_I}$. Next, we identify the largest index i such that $p_{r_i} \leq (i/I)\alpha$. The clones ranked from 1 to i are referred to as "expanded." This filtering step preferentially removes clones whose apparent changes are consistent with random sampling noise, given the fixed library sizes.

CEI-ACAT Method

Alternatively, we can combine the two statistics into a single p-value measurement using the Aggregated Cauchy Association Test (ACAT) [42]. ACAT merges multiple p-values by mapping each one to a Cauchy distribution and then returning a single p-value. We use this method to consolidate the two one-sided p-values for Z_{d_i} and Z_{r_i} into a single score. ACAT remains valid even if dependency between the two indicators exists. The ACAT method constructs the quantity

$$T_i = \frac{1}{2}\tan\left[\pi\left(0.5 - p_{d_i}\right)\right] + \frac{1}{2}\tan\left[\pi\left(0.5 - p_{r_i}\right)\right] \tag{13}$$

from the one-sided p-values $p_{d_i} = 1 - \Phi(Z_{d_i})$ and $p_{r_i} = 1 - \Phi(Z_{r_i})$, respectively. The corresponding effective p-value for clone i is then calculated from

$$p_i^{\mathrm{ACAT}} = 0.5 - \frac{1}{\pi}\arctan(T_i). \tag{14}$$

Again, BH-FDR control is utilized to filter out false positive clones.

Results

We now apply the methods described above to SARS-CoV-2 vaccination data to identify vaccine-associated TCR sequences. The FDR will be controlled by the BH procedure, and we set $\alpha = 1\% - 10\%$ as a possible range for selection.

Comparison of positive-associated clones identified by different methods

When applying a 10% FDR to both the CEI-Joint and the CEI-ACAT methods, the table shows strong inter-individual and inter-visit heterogeneity in the number of positive clones. To provide a clear overview of this variability, we list the number of positive-associated clones in Table 2. The CEI-Joint method typically identifies many more clones than the CEI-ACAT method (for example, it reports thousands of clones for CLE0108 and CLE0117). In contrast, CEI-ACAT tends to be more conservative, although it occasionally calls more positive clones than CEI-Joint in specific cases (such as for CLE0101 B–P1, CLE0100 B–P2, and CLE0136 B-P1). Additionally, the results patterns differ depending on the visit pairs: some individuals show their strongest signals in P2 compared to P1 (indicating a later response), while others have the largest signal in P2 compared to B (indicating an overall strong post-vaccination expansion). Several participants, however, exhibited very few positive clones across all comparisons (*e.g.*, CLE0100 and CLE0136). Since subject CLE0099 received the chimpanzee adenovirus-vectored vaccine, the set of activated TCR clones in this individual consisted of clones responding not only to SARS-CoV-2 antigens but also to antigens derived from the adenoviral vector itself.

We compared the positively associated clones identified by our CEI-Joint and CEI-ACAT methods, as illustrated in Fig. 1 and in Figs. S1-S6 in the Supplementary Material. The positive clones selected at different times were analyzed by comparing sample P1 with sample B, sample P2 with sample B, and sample P2 with sample P1. Results are presented in Fig. 2 and Figs. S7-S12 in the Supplementary Material. For better visualization, both figures display selections made by the CEI-Joint method at a 10% FDR and selections made by the CEI-ACAT method at a 1% FDR.

Comparison with standard heuristic selection criteria (Adap) and edgeR

Our TCR identification methods were further evaluated by comparing their predictions with those from methods previously developed to identify differentially abundant clones [40, 43] in different contexts.

The standard heuristic selection method (Adap) previously used by Adaptive Biotechnologies designates a clone as activated if its sampled baseline and post-vaccination counts exhibit at least a 2-fold increase *and* that the abundances pass a two-sided binomial test with a p-value below 0.01. In this analysis, the two-sided binomial test p-values are adjusted using the BH procedure to effectively control the FDR. Furthermore, to ensure the reliability of the detection, only clones with a minimum count abundance of five in the sample are considered. edgeR, which tests negative binomial or generalized linear models, was also applied to our single sample (no replicates) data.

Figure 3 compares the clones identified using two heuristic criteria (Adap and edgeR) with those identified using our CEI-Joint and CEI-ACAT methods for the subjects CLE0083 and CLE0101. All methods were controlled at a 1% FDR. The panels highlight significant subject-specific differences and distinct profiles for each method.

Individuals	R_B	R_{P1}	R_{P2}	B vs P1		B vs P2		P1 vs P2	
				Joint	ACAT	Joint	ACAT	Joint	ACAT
CLE0083	184, 988	315, 615	322, 569	359	5	373	3	2	2
CLE0099	223, 828	85, 703	231, 395	451	216	44	27	2584	12
CLE0100	245, 481	219, 434	267, 635	1	1	4	21	17	16
CLE0101	252, 373	271, 931	177, 847	24	89	101	47	202	33
CLE0108	136, 808	591, 274	208, 447	5127	5	46	2	2001	60
CLE0117	133, 666	89, 962	369, 441	117	5	4248	3	5119	11
CLE0136	370, 013	386, 065	297, 501	2	4	5	6	17	7

Table 2: Table of the total number of positive-associated clones across multiple participants. R_B, R_{P1}, and R_{P2} are the number of clones (richness) detected for each individual in the sample B, P1, and P2, respectively. In the B versus P1, B versus P2, and P1 versus P2 columns, we list the number of SARS-CoV-2 clones identified through their sufficient increases in abundance as determined by our CEI-Joint and CEI-ACAT approaches. We set the FDR $\alpha = 10\%$ for selection.

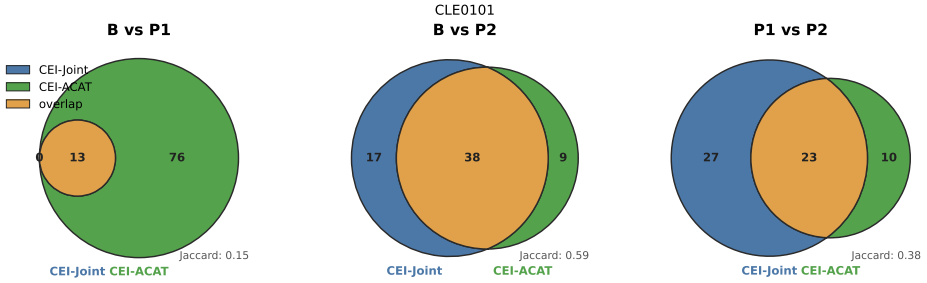

Figure 1: Comparisons between the positive sequences called by each method for individual CLE0101. The blue and green circles represent sets of clones identified by the CEI-Joint and CEI-ACAT methods, respectively. The size of each circle indicates the number of positive clones detected. The yellow portion illustrates the overlap of positive clones identified by both the CEI-Joint and CEI-ACAT methods. Left: Clones identified through comparison of sample P1 to sample B. Middle: Clones identified through comparison of sample P2 to sample B. Right: Clones identified through comparison of sample P2 to sample P1.

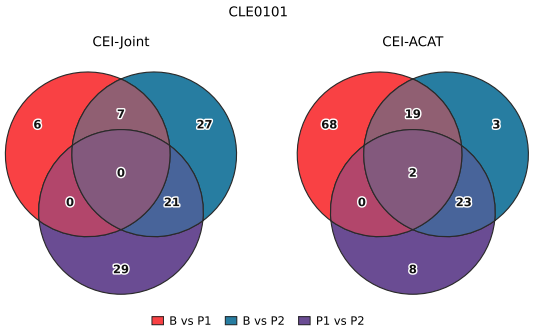

Figure 2: Comparisons between the positive sequences sampled at different times for subject CLE0101. The subplots, left and right, represent comparisons using the CEI-Joint and CEI-ACAT methods, respectively. Red, blue, and purple circles correspond to positive clones identified by comparing sample P1 with B, P2 with B, and P2 with P1. The size of each circle reflects the number of positive clones identified. Left: Clones identified via the CEI-Joint method. Right: Clones identified by the CEI-ACAT method.

For CLE0083 (a), our CEI-Joint test identifies a considerable number of expansions ($n = 87$), far exceeding the counts for CEI-ACAT ($n = 5$), Adap ($n = 7$), and edgeR ($n = 2$). Some of the CEI-Joint calls lie just above the B=P1 diagonal, indicating a broad but modest shift following vaccination. In contrast, for CLE0101 (b), the rankings shift.

CEI-ACAT (n = 70) and Adap (n = 52) identify the largest sets along the rising trend, while edgeR falls in between with n = 28. The CEI-Joint method is the most stringent, identifying only n = 13 clones.

Across both subjects, selected clones predominantly sit above the diagonal (indicating expansion), but the methods emphasize different regions of the plot. The standard heuristic method focuses solely on larger log-fold changes, while edgeR tends to favor high log-fold change signals with low clone abundance at baseline. Our CEI-ACAT and CEI-Joint methods strike a balance between absolute differences and log-fold changes. This results in many calls when there is a diffuse, concordant shift (as seen in CLE0083), but fewer calls when changes cluster along a tight trajectory (as seen in CLE0101).

Our collection of statistical techniques provides a more systematic framework for fine-tuning FDR, which allows us to control the stringency in identifying activated clones. These flexible tools can be customized to detect population expansions in different biological contexts, ensuring that the identification process is both rigorous and adaptable to varying experimental conditions.

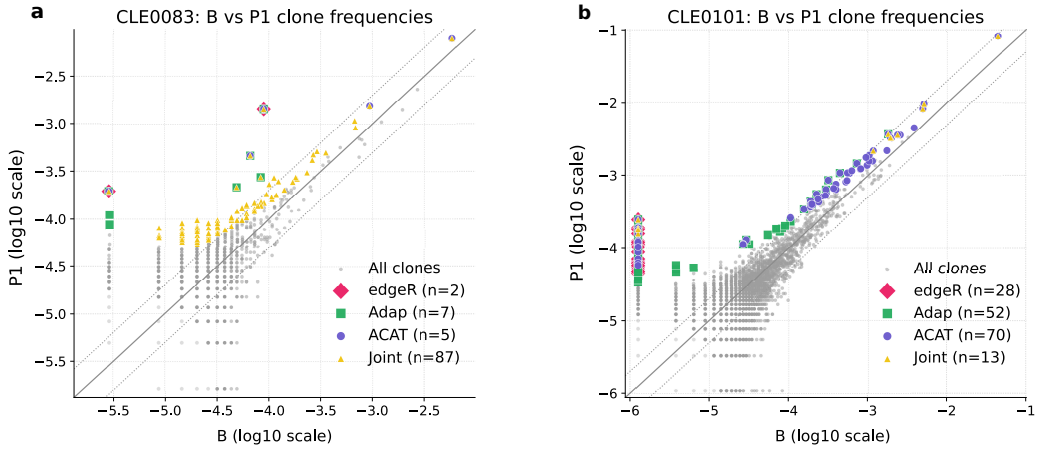

Figure 3: Abundances of individual TCR clones before vaccination (B) are plotted along the x-axis, while their abundances after vaccination (P1) are plotted along the y-axis. (a) and (b) correspond to subjects CLE0083 and CLE0101, respectively. Clones that are statistically significantly expanded (blue) are identified by a corrected binomial model [41] and additionally, a threshold of at least a 2-fold increase or decrease from B to P1. Each dot corresponds to a unique TCR clone: grey circles represent robust clones with an abundance greater than 5 templates, while clones with fewer than 5 templates are shown in white to indicate they fall below the threshold for reliable detection. Purple squares represent the expanded clones called by our ACAT method using a FDR at 1%, while green circles denote the expanded clones identified by the standard heuristic approach. Similarly, clones that have been detected by edgeR are indicated by the red triangles. These results illustrate the overlap between the clones identified by the ACAT method, the clones detected using the standard heuristic criteria, and those called by edgeR for differential abundance analysis.

Discussion and Conclusions

Our study builds on the efforts to map TCRs to SARS-CoV-2 antigens by offering a high-resolution, longitudinal view of both $CD4^+$ and $CD8^+$ T cell responses at three critical timepoints: pre-vaccination, post-dose 1, and post-dose 2. This design allows for precise tracking of TCR clonal expansion and repertoire reshaping in response to antigenic stimulation. By focusing on COVID-naive individuals, this study avoids confounding from prior infection and provides a clean baseline for understanding vaccine-induced T cell immunity—an essential component of long-term protection and immune memory.

To determine whether changes in TCR abundances are deemed vaccination-sensitive (*e.g.*, whether associated T cells proliferate following vaccination) requires us to carefully define statistical thresholds. We presented several measures to detect clones whose abundances increased significantly after vaccination. Our framework quantifies each clone with two complementary statistics—the rescaled abundance change of a clone d_i and its log-fold change in abundance r_i.

Human samples make collection of replicates challenging and make it difficult to filter out experimental noise during the selection process and sequencing procedure. To validate our methods, we compared our selections with the

simpler rule "$r_i \geq 2$" (plus a minimum of five template counts), NoisET and edgeR. NoisET and edgeR are well-known packages designed for clonal expansion detection. NoisET explicitly learns an experimental noise model and then applies a Dirichlet-Multinomial Bayesian test for expansion under stimulation. However, our dataset lacks replicates, which prevents us from using NoisET to learn about the noise and detect expanded clones. On the other hand, edgeR uses empirical Bayes moderation techniques for scenarios with very few or no replicates. However, it is too conservative and highly sensitive to the abundance of clones in the baseline sample, such that it can only detect expanded clones if their abundance in the baseline sample is nearly zero. If the size of these clones is even slightly larger in the baseline sample, edgeR may fail to identify the expansion. We provide a case example for reference in Fig. 4.

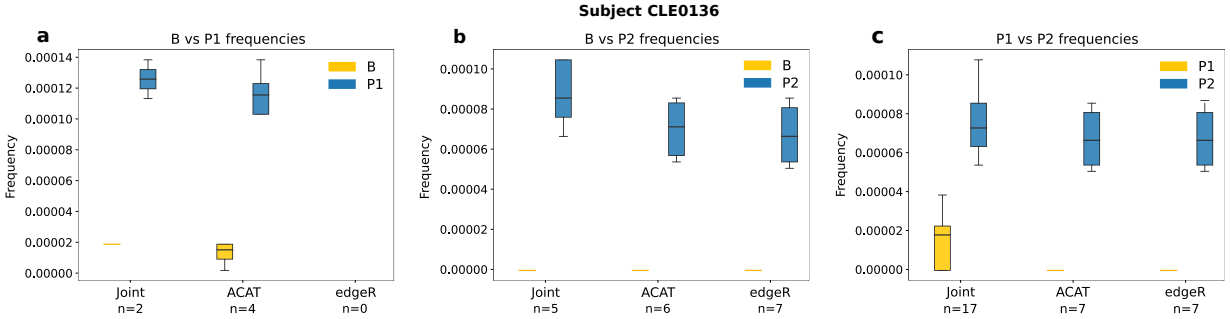

Figure 4: For subject CLE0136, the boxplots compare clone frequencies between time points for the clones called by each method (n listed below panels): (a) B vs P1, (b) B vs P2, and (c) P1 vs P2. For each contrast, boxplots display the frequencies of the identified clones by comparing two samples using three different methods: CEI-Joint, CEI-ACAT, and edgeR (with the number of calls, n, under each method). The CEI-Joint method yields significantly higher frequencies at both baseline and post-vaccination samples. Note that edgeR calls no positive sequences any significant expansion in the B vs P1 comparison, even when we relaxed the FDR.

We performed sensitivity analysis of the number of positive clones selected by scanning the FDR α from 0.001 to 0.5. Since most clones are of small abundance with relatively small changes, the number of detected clones increases sharply once α exceeds a certain level, shown in the Fig. 5.

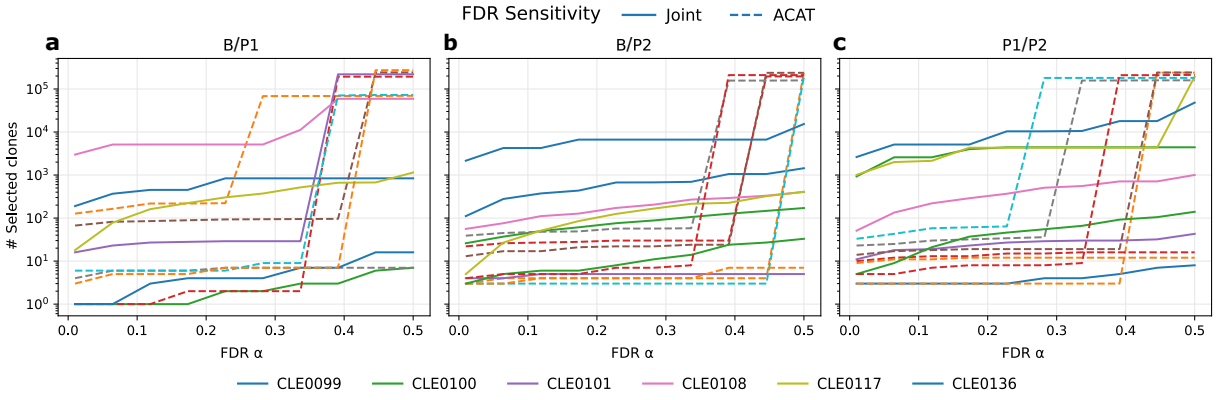

Figure 5: Sensitivity of the number of positive clones selected with respect to the FDR threshold α, where $\alpha \in (0, 0.5]$. The dashed lines correspond to the CEI-ACAT method, while the solid lines correspond to the CEI-Joint method. (a) Sensitivity analysis conducted by comparing sample P1 with B. (b) Sensitivity analysis conducted by comparing sample P2 with B. (c) sensitivity analysis conducted by comparing sample P2 with P1.

Our general methods can be applied to any multispecies system subject to an external perturbation that yields before and after snapshots of abundances. Using the same metrics and thresholds also enable direct comparisons across different induction protocols and experimental contexts, allowing one to quantify overlaps in results from different experimental/clinical methods. For example, using changes in the *in vivo* T-cell abundance as an indicator of vaccination response led to significantly fewer identified TCRs than using cellular immune assays using antigen libraries [44]. The substantial difference between clones that were measured to have expanded *in vivo* and those identified through antigen recognition in cell culture suggests a more complex pathway from antigen recognition to T

cell amplification and carries important clinical implications. Besides overcounting low-affinity TCRs, discrepancies between the *in vivo* and *ex vivo* methods may arise from other causes such as finite sampling from individuals, different timescales between *ex vivo* assays and *in vivo* clone abundance dynamics, and dynamical fluctuations of interacting T-cell populations in an individual. How these effects influence changes in clonal populations subjected to different stimulation protocols will be quantitatively compared in future investigations that use the methods described here.

Declaration of Interest Statement

HP, BB, SAB are employees of Adaptive Biotechnologies and hold stock or stock options in the company. YP, TC, and OY do not have any commercial or financial relationships with Adaptive Biotechnologies.

CRediT author contribution statement

YP: Writing – review & editing, Writing – original draft, Investigation, Conceptualization, Data curation, Methodology, Visualization, Formal Analysis; CH: Writing – review & editing, Investigation, Data curation, Resources; BB, HP, and SB: Data curation, Visualization, Resources, Formal Analysis, Review & Editing; TC: Writing – Original Draft, Writing – Review & Editing, Visualization, Supervision, Methodology, Conceptualization; OY: Writing – Review & Editing, Conceptualization, Methodology, Supervision.

Acknowledgements

The authors are grateful for support from the UCLA-CDU CFAR grant no. AI152501. Additional funding was provided by private philanthropic donors (including William Moses, Mari Edelman, Beth Friedman, Dana and Matt Walden, Kathleen Poncher, Scott Z. Burns, Gwyneth Paltrow and Brad Falchuk), with additional infrastructure support from the UCLA-Charles Drew University Center for AIDS Research (NIH grant AI152501), James B. Pendleton Trust, and McCarthy Foundation. Support for TCR immunosequencing was provided by Adaptive Biotechnologies, Seattle, WA, USA.

Supplementary materials

Additional comparisons of identified clones using different methods, across all other subjects are provided in the Supplementary materials available in the online version.

Data availability

Post-processed immunosequencing data (as described here) are freely available at the ImmuneACCESS database, `https://clients.adaptivebiotech.com/immuneaccess`, and will also be provided upon request.

Code availability

The software package is freely available under the MIT license from the GitHub website, `https://github.com/ybpan16/CEI`.

References

[1] L. R. Baden, H. M. El Sahly, B. Essink, K. Kotloff, S. Frey, R. Novak, D. Diemert, S. A. Spector, N. Rouphael, C. B. Creech, J. McGettigan, S. Khetan, N. Segall, J. Solis, A. Brosz, C. Fierro, H. Schwartz, K. Neuzil, L. Corey, P. Gilbert, H. Janes, D. Follmann, M. Marovich, J. Mascola, L. Polakowski, J. Ledgerwood, B. S. Graham, H. Bennett, R. Pajon, C. Knightly, B. Leav, W. Deng, H. Zhou, S. Han, M. Ivarsson, J. Miller, T. Zaks, Efficacy and safety of the mRNA-1273 SARS-CoV-2 vaccine, New England Journal of Medicine 384 (5) (2021) 403–416, pMID: 33378609. arXiv:https://doi.org/10.1056/NEJMoa2035389, doi:10.1056/NEJMoa2035389.
URL https://doi.org/10.1056/NEJMoa2035389

[2] W. F. Garcia-Beltran, E. C. Lam, K. S. Denis, A. D. Nitido, Z. H. Garcia, B. M. Hauser, J. Feldman, M. N. Pavlovic, D. J. Gregory, M. C. Poznansky, et al., Multiple SARS-CoV-2 variants escape neutralization by vaccine-induced humoral immunity, Cell 184 (9) (2021) 2372–2383.

[3] R. R. Goel, M. M. Painter, S. A. Apostolidis, D. Mathew, W. Meng, A. M. Rosenfeld, K. A. Lundgreen, A. Reynaldi, D. S. Khoury, A. Pattekar, et al., mRNA vaccines induce durable immune memory to SARS-CoV-2 and variants of concern, Science 374 (6572) (2021) abm0829.

[4] F. J. Ibarrondo, C. Hofmann, J. A. Fulcher, D. Goodman-Meza, W. Mu, M. A. Hausner, A. Ali, A. Balamurugan, E. Taus, J. Elliott, et al., Primary, recall, and decay kinetics of SARS-CoV-2 vaccine antibody responses, ACS Nano 15 (7) (2021) 11180–11191.

[5] R. Link-Gelles, M. E. Levy, K. Natarajan, S. E. Reese, A. L. Naleway, S. J. Grannis, N. P. Klein, M. B. DeSilva, T. C. Ong, M. Gaglani, et al., Estimation of COVID-19 mRNA vaccine effectiveness and COVID-19 illness and severity by vaccination status during Omicron BA. 4 and BA. 5 sublineage periods, JAMA network open 6 (3) (2023) e232598–e232598.

[6] F. P. Polack, S. J. Thomas, N. Kitchin, J. Absalon, A. Gurtman, S. Lockhart, J. L. Perez, G. Pérez Marc, E. D. Moreira, C. Zerbini, et al., Safety and efficacy of the BNT162b2 mRNA Covid-19 vaccine, New England Journal of Medicine 383 (27) (2020) 2603–2615.

[7] Z. Zhang, J. Mateus, C. H. Coelho, J. M. Dan, C. R. Moderbacher, R. I. Gálvez, F. H. Cortes, A. Grifoni, A. Tarke, J. Chang, et al., Humoral and cellular immune memory to four COVID-19 vaccines, Cell 185 (14) (2022) 2434–2451.

[8] D. G. Augusto, L. D. Murdolo, D. S. M. Chatzileontiadou, J. J. Sabatino Jr., T. Yusufali, N. D. Peyser, X. Butcher, K. Kizer, K. Guthrie, V. W. Murray, et al., A common allele of HLA is associated with asymptomatic SARS-CoV-2 infection, Nature 620 (7972) (2023) 128–136.

[9] P. R. Buckley, C. H. Lee, M. Pereira Pinho, R. Ottakandathil Babu, J. Woo, A. Antanaviciute, A. Simmons, G. Ogg, H. Koohy, HLA-dependent variation in SARS-CoV-2 CD8+ T cell cross-reactivity with human coronaviruses, Immunology 166 (1) (2022) 78–103.

[10] A. Sette, J. Sidney, S. Crotty, T cell responses to SARS-CoV-2, Annual Review of Immunology 41 (2023) 343–373.

[11] E. Taus, C. Hofmann, F. J. Ibarrondo, J. A. Fulcher, S. G. Kitchen, N. H. Tobin, O. O. Yang, Persistent memory despite rapid contraction of circulating T cell responses to SARS-CoV-2 mRNA vaccination, Frontiers in Immunology 14 (2023) 1100594.

[12] E. Taus, C. Hofmann, F. J. Ibarrondo, M. A. Hausner, J. A. Fulcher, P. Krogstad, K. G. Ferbas, N. H. Tobin, A. W. Rimoin, G. M. Aldrovandi, et al., Dominant CD8+ T cell nucleocapsid targeting in SARS-CoV-2 infection and broad spike targeting from vaccination, Frontiers in Immunology 13 (2022) 835830.

[13] K. Murphy, C. Weaver, Janeway's Immunobiology, Garland Science, 2016.

[14] J. Nikolich-Žugich, M. K. Slifka, I. Messaoudi, The many important facets of T-cell repertoire diversity, Nature Reviews Immunology 4 (2) (2004) 123–132.

[15] A. K. Sewell, Why must T cells be cross-reactive?, Nature Reviews Immunology 12 (9) (2012) 669–677.

[16] T. P. Arstila, A. Casrouge, V. Baron, J. Even, J. Kanellopoulos, P. Kourilsky, A direct estimate of the human $\alpha\beta$ T cell receptor diversity, Science 286 (5441) (1999) 958–961.

[17] G. Lythe, R. E. Callard, R. L. Hoare, C. Molina-París, How many TCR clonotypes does a body maintain?, Journal of Theoretical Biology 389 (2016) 214–224.

[18] T. Mora, A. M. Walczak, How many different clonotypes do immune repertoires contain?, Current Opinion in Systems Biology 18 (2019) 104–110.

[19] Q. Qi, Y. Liu, Y. Cheng, J. Glanville, D. Zhang, J.-Y. Lee, R. A. Olshen, C. M. Weyand, S. D. Boyd, J. J. Goronzy, Diversity and clonal selection in the human T-cell repertoire, Proceedings of the National Academy of Sciences 111 (36) (2014) 13139–13144.

[20] R. L. Warren, J. D. Freeman, T. Zeng, G. Choe, S. Munro, R. Moore, J. R. Webb, R. A. Holt, Exhaustive T-cell repertoire sequencing of human peripheral blood samples reveals signatures of antigen selection and a directly measured repertoire size of at least 1 million clonotypes, Genome Research 21 (5) (2011) 790–797.

[21] M. K. Jenkins, H. H. Chu, J. B. McLachlan, J. J. Moon, On the composition of the preimmune repertoire of T cells specific for peptide–major histocompatibility complex ligands, Annual Review of Immunology 28 (2009) 275–294.

[22] H. Aoki, M. Kitabatake, H. Abe, P. Xu, M. Tsunoda, S. Shichino, A. Hara, N. Ouji-Sageshima, C. Motozono, T. Ito, et al., CD8+ T cell memory induced by successive SARS-CoV-2 mRNA vaccinations is characterized by shifts in clonal dominance, Cell Reports (2024).

[23] V. Mallajosyula, C. Ganjavi, S. Chakraborty, A. M. McSween, A. J. Pavlovitch-Bedzyk, J. Wilhelmy, A. Nau, M. Manohar, K. C. Nadeau, M. M. Davis, CD8+ T cells specific for conserved coronavirus epitopes correlate with milder disease in patients with COVID-19, Science Immunology 6 (61) (2021) eabg5669.

[24] L. Swadling, M. O. Diniz, N. M. Schmidt, O. E. Amin, A. Chandran, E. Shaw, C. Pade, J. M. Gibbons, N. Le Bert, A. T. Tan, et al., Pre-existing polymerase-specific T cells expand in abortive seronegative SARS-CoV-2, Nature 601 (7891) (2022) 110–117.

[25] F. J. Ibarrondo, J. A. Fulcher, D. Goodman-Meza, J. Elliott, C. Hofmann, M. A. Hausner, K. G. Ferbas, N. H. Tobin, G. M. Aldrovandi, O. O. Yang, Rapid decay of anti–SARS-CoV-2 antibodies in persons with mild COVID-19, New England Journal of Medicine 383 (11) (2020) 1085–1087.

[26] L. B. Rodda, J. Netland, L. Shehata, K. B. Pruner, P. A. Morawski, C. D. Thouvenel, K. K. Takehara, J. Eggenberger, E. A. Hemann, H. R. Waterman, et al., Functional SARS-CoV-2-specific immune memory persists after mild COVID-19, Cell 184 (1) (2021) 169–183.

[27] A. K. Wheatley, J. A. Juno, J. J. Wang, K. J. Selva, A. Reynaldi, H.-X. Tan, W. Shi Lee, K. M. Wragg, H. G. Kelly, R. Esterbauer, et al., Evolution of immunity to SARS-CoV-2 in mild-moderate COVID-19, Nature Communications 12 (2021) 116.

[28] A. Grifoni, D. Weiskopf, S. I. Ramirez, J. Mateus, J. M. Dan, C. R. Moderbacher, S. A. Rawlings, A. Sutherland, L. Premkumar, R. S. Jadi, et al., Targets of T cell responses to SARS-CoV-2 coronavirus in humans with COVID-19 disease and unexposed individuals, Cell 181 (7) (2020) 1489–1501.

[29] E. Jokinen, A. Dumitrescu, J. Huuhtanen, V. Gligorijević, S. Mustjoki, R. Bonneau, M. Heinonen, H. Lähdesmäki, Tcrconv: predicting recognition between T cell receptors and epitopes using contextualized motifs, Bioinformatics 39 (1) (2023) btac788.
[30] A. P. Ferretti, T. Kula, Y. Wang, D. M. V. Nguyen, A. Weinheimer, G. S. Dunlap, Q. Xu, N. Nabilsi, C. R. Perullo, A. W. Cristofaro, et al., Unbiased screens show CD8+ T cells of COVID-19 patients recognize shared epitopes in SARS-CoV-2 that largely reside outside the spike protein, Immunity 53 (5) (2020) 1095–1107.
[31] A. Grifoni, J. Sidney, R. Vita, B. Peters, S. Crotty, D. Weiskopf, A. Sette, SARS-CoV-2 human T cell epitopes: Adaptive immune response against COVID-19, Cell Host & Microbe 29 (7) (2021) 1076–1092.
[32] M. B. Koraichi, M. P. Touzel, A. Mazzolini, T. Mora, A. M. Walczak, NoisET: noise learning and expansion detection of T-cell receptors, The Journal of Physical Chemistry A 126 (40) (2022) 7407–7414.
[33] Y. Chen, B. Pal, J. E. Visvader, G. K. Smyth, Differential methylation analysis of reduced representation bisulfite sequencing experiments using edgeR, F1000Research 6 (2018) 2055.
[34] L. Kuchenbecker, M. Nienen, J. Hecht, A. U. Neumann, N. Babel, K. Reinert, P. N. Robinson, IMSEQ—a fast and error aware approach to immunogenetic sequence analysis, Bioinformatics 31 (18) (2015) 2963–2971. arXiv:http://bioinformatics.oxfordjournals.org/content/31/18/2963.full.pdf+html, doi:10.1093/bioinformatics/btv309.
URL http://bioinformatics.oxfordjournals.org/content/31/18/2963.abstract
[35] S. Kullback, R. A. Leibler, On Information and Sufficiency, The Annals of Mathematical Statistics 22 (1) (1951) 79 – 86. doi:10.1214/aoms/1177729694.
URL https://doi.org/10.1214/aoms/1177729694
[36] L. V. Kantorovich, Mathematical methods of organizing and planning production, Management Science 6 (4) (1960) 366–422.
[37] L. N. Vaserstein, Markov processes over denumerable products of spaces, describing large systems of automata, Problemy Peredachi Informatsii 5 (3) (1969) 64–72.
[38] A. Singhal, et al., Modern information retrieval: A brief overview, IEEE Data Eng. Bull. 24 (4) (2001) 35–43.
[39] A. Bhattacharyya, On a measure of divergence between two multinomial populations, Sankhyā: The Indian Journal of Statistics (1946) 401–406.
[40] W. S. DeWitt, R. O. Emerson, P. Lindau, et al., Dynamics of the cytotoxic T cell response to a model of acute viral infection, Journal of Virology 89 (2015) 4517–4526.
[41] Y. Benjamini, Y. Gavrilov, A simple forward selection procedure based on false discovery rate control, Annals of Applied Statistics 3 (1) (2009) 179–198.
[42] Y. Liu, S. Chen, Z. Li, A. C. Morrison, E. Boerwinkle, X. Lin, ACAT: A fast and powerful p-value combination method for rare-variant analysis in sequencing studies, The American Journal of Human Genetics 104 (3) (2019) 410–421. doi:https://doi.org/10.1016/j.ajhg.2019.01.002.
URL https://www.sciencedirect.com/science/article/pii/S0002929719300023
[43] L. A. Rojas, Z. Sethna, K. C. Soares, C. Olcese, N. Pang, E. Patterson, J. Lihm, N. Ceglia, P. Guasp, A. Chu, et al., Personalized RNA neoantigen vaccines stimulate T cells in pancreatic cancer, Nature 618 (7963) (2023) 144–150.
[44] M. Klinger, F. Pepin, J. Wilkins, T. Asbury, T. Wittkop, J. Zheng, M. Moorhead, M. Faham, Multiplex Identification of Antigen-Specific T Cell Receptors Using a Combination of Immune Assays and Immune Receptor Sequencing, PLoS ONE 10 (10) (2015) e0141561.