Supplementary Note 1: detailed description and diagnosis of the tessa model

October 16, 2020

1 Overview

T cells play a central role in moderating host immune response. T cell receptors (TCRs) identify specific TCR clonotypes and mediate recognition of antigens and activation of T cell. Therefore, an informative characterization of TCRs is critical for studying T cell immunity and responsiveness to immunotherapies. Many experimental and in-silico approaches have been developed to characterize the TCR repertoire. However, most existing approaches are not able to elucidate the functional relevance of TCR sequences, which significantly hinders and could mislead interpretation of the roles of T cells in various biological processes. To address this inadequacy, we developed a Bayesian model named tessa (TCR Functional Landscape Estimation Supervised with ScRNA-Seq Analysis), to integrate TCR sequence profiling with the transcriptomes of T cells. Enabled by the recently developed single cell sequencing techniques, which provide both TCR sequences and RNA sequences of each T cell concurrently, tessa allows us to map the functional landscape of the TCR repertoire, and generates insights into understanding human immune response to diseases.

2 The tessa model

2.1 Input data

The aims of tessa include: (1) estimating TCR sequence similarities and building TCR networks, and (2) capturing the association between expression and TCR repertoire. The 10x Genomics Chromium single cell immune profiling technique, and other advanced sequencing techniques, are capable of providing both gene expression and T cell receptor sequencing for each T cell simultaneously. For each T cell, we have (1) gene expression levels, as one numerical vector. Each element in the vector represents the RNA expression level of one gene, and (2) one TCR sequence, which is converted into a numeric vector through a stacked auto-encoder. The TCR sequences considered in the tessa model include only the TCR-β chain CDR3 region sequences. Typically, one T cell has only one productive TCR-β chain detected. In very rare conditions when one T cell has two productive TCR-β chains detected from the sequencing data, we select the sequence with the higher expression level. The group of T cells from the same patient that share the same TCR sequence are referred to as the same clone, and they are of the same clonotype. No T cell appears in more than one clone, and there are some clones that have only one unique T cell. The stacked auto-encoder extracts sequence features from each unique TCR and converts it into a numerical vector. Let \( i = 1, ..., T \) be the indices for T cells. Each T cell has a vector of numeric gene expression values, denoted by \( e_i \), and a vector of numeric TCR sequence features, denoted by \( t_i \).

2.2 Model specification

As the first step, T cells are assigned into networks depending on their sequence similarity, where T cells in one network are likely to target the same antigen. Naturally, T cells sharing the same TCR should be assigned to the same network. Therefore, networks are built on the units of TCRs, or clones, rather than those of T cells. We use \( t \) to represent all the unique TCRs, in the 30-dim embedded space, in one dataset. However, the RNA expression values of T cells sharing the same TCR are usually not the same, so we use \( e \) to represent expression of all individual T cells rather than T cell clones in one dataset. We consider identifying \( K \) networks among all T cells with observed data \( t \) and \( e \). The number of networks \( K \) is a positive integer determined through a Dirichlet Process dynamically, which assigns the network membership \( m_i \in \{1, ..., K\} \) for the \( i \)-th clone.

In each TCR network, we will determine a network center T cell clonotype. For the \( k \)-th network consisting of T cell clones with \( \varphi_k \) different TCR sequences \( \{ t_{k,1}, ..., t_{k,\varphi_k} \} \), we identify \( t_{k,c_k} \), where \( c_k \in \{1, ..., \varphi_k\} \), as the center TCR
that is "closest" to the mean of all embedded TCR sequences in this network. The remaining TCRs are considered as non-center in network $k$. Due to the inner structure of the networks such that there are center clones and non-center T cell clones in each network, we used the terminology of networks rather than clusters in this work, to emphasize this internal hierarchy.

Next, we maximize the association between TCRs and gene expression. Based on our observation of the real data, if two T cell clones within the same network have more similar TCRs, their expression profiles are also more similar. Motivated by this observation, we specify the relationship between TCR distances and expression distances in network $k$ via a linear regression model,

$$d_{k,c,r}^r = a_k d_{k,c,r}^e + \epsilon_{k,r},$$

where $d_{k,c,r}^e$ are distances between the center TCR and the non-center TCRs, $d_{k,c,r}^r$ represents the average of the pairwise distances, in terms of gene expression, between all T cells in the center clone and all T cells in each non-center clone, for $r = 1, ..., \varphi_k$. The regression coefficient and the random error are denoted by $a_k$ and $\epsilon_{k,c,r}$, respectively. The error term $\epsilon_{k,c,r}$ is assumed to follow a normal distribution with mean 0 and a variance of $\sigma_r^2$. Specifically, the two sets of distances in (1) can be calculated as,

$$d_{k,c,r}^e = \frac{1}{2} \sum_{q=1}^{Q} \frac{(t_{k,r,q} - t_{k,c,q})^2}{b_q}, 0 < r, c_k \in 1, ..., \varphi_k,$$

and,

$$d_{k,c,r}^r = \frac{\sum_{r \in g_k,r} \sum_{j \in g_k,c} ||e_i - e_j||^2}{|g_k,r| \cdot |g_k,c|}, 0 < r, c_k \in 1, ..., \varphi_k,$$

where $g_k,r$ and $g_k,c$ refer to the sets of the cells that belong to non-center TCR clone $t_{k,r}$ and the center TCR clone $t_{k,c}$, respectively.

As we previously mentioned, the networks are built on TCR clones rather than T cells, and the transcriptomes of the T cells in the same clonotype are usually not the same. The expression distances in (3) are defined as the average of the expression distances between every cell in one clone and every cell in the other clone. Assuming that different embedding digits of TCRs have different importance in terms of describing the part of TCR that is more relevant for the function of T cells, we employ the weighting factor $b = \{b_q\}_{q=1}^{Q}$ in (2) to adjust the weights of each digit to maximize the association between expression and TCRs in (1). As a result, the component of TCRs that is more relevant to the function of T cells is more influential in the tessa model. $Q$ is the number of embedding digits. For instance, $Q = 30$ with our stacked auto-encoder.

For the network assignments $m_i$ ($m_i = 1, ..., K$) for clone $i$, we employ a Dirichlet Process with a very diffuse base measure, $G_0 : MVN((t_{mean,1}, ..., t_{mean,i}, ..., t_{mean,30})^T, \lambda I_Q)$, where $t_{mean,i}$ represents the average value of the $i$-th element of all the input embedded TCR sequences $t$, and $\lambda$ is a large positive number which is pre-defined to keep the base measure diffuse enough and cover the dynamic region of the embedded TCR sequences. We define $t_{k,0}$ as the mean of the multivariate normal distribution that can describe the group of (embedded) TCRs in network $k$. We also define the center TCRs $c_k$, which is the TCR clonotype that is closest to $t_{k,0}$, and the fixed scaling parameter $\xi$ for sampling $m_i$. We assume a multivariate normal distribution for $t$,

$$p(t_{k,1}, ..., t_{k,\varphi_k})|t_{k,0}, b, m_i) \propto \prod_{r=1, ..., \varphi_k} e^{-d_{k,0,r}^r} \left(\prod_{b_q}\right)^{b_q} 0.5 \cdot$$

Given the network assignment in (4), for each network $k$, we define the center as,

$$c_k = \arg\min_s |t_{k,0} - t_{k,s}|, s \in \{1, ..., \varphi_k\}.$$  

We then define priors and hyperpriors. A diffuse inverse gamma prior distribution $IG(a = 0.1, v = 0.1)$ is taken for $b_q$ in (2) and for $\sigma_r^2$ of the random error $\epsilon_{k,r}$ in (1). We assign the regression coefficient $a_k$ in (1) with a $g$–prior $N(a, g\sigma^2)$. For the hyperparameter $a$, we assume a truncated normal prior to enforce $a$ to have a positive value,

$$TN(a; 0, \tau, a > 0) \propto e^{-\frac{a^2}{\tau}} ,$$

where $\tau$ is a pre-defined large positive number to make the prior diffuse. The overall model structure is summarized in SN1 Fig 1.
2.3 Posterior computation

Let $\Theta = \{\{a_k\}_{k=1}^{K}, \{b_q\}_{q=1}^{Q}, \{m_i\}_{i=1}^{K}, \{t_{k,0}\}_{k=1}^{K}, a, \sigma^2\}$ denote the set of model parameters. The joint posterior distribution is,

$$p(\Theta | t, e) \propto \prod_{k=1}^{K} G_0(t_{k,0}; \lambda) \cdot \prod_{q=1}^{Q} IG(b_q; u, v) \cdot \prod_{k=1}^{K} P(\{t_{k,1}, ..., t_{k,\varphi_k}\}| t_{k,0}, \{b_q\}_{q=1}^{Q}, \{m_i\}_{i=1}^{K}) \cdot \mathcal{N}(a; 0, \tau, a > 0).$$

With the log-posterior in (8), we employ the MCMC sampling to draw samples for each parameter in $\Theta$.

$$\log(p(\Theta | t, e)) = -\frac{1}{2} \sum_{k=1}^{K} \sum_{q=1}^{Q} \log(\sigma^2_k) - \frac{1}{2} \sum_{k=1}^{K} \sum_{r=1}^{\varphi_k} \log(d_{k,r}^2) + \frac{v}{b_q} - \sum_{k=1}^{K} \sum_{r=1}^{\varphi_k} a_k d_{k,r}^2 + \frac{\sum_{q=1}^{Q} \log(b_q)}{2} - \frac{a^2}{\sigma^2} - [(u+1)\log(\sigma^2) + \frac{v}{\sigma^2}].$$

We update $\Theta$ according to the following steps.

(I) Updating network membership $m_i$ for each cell $i$, using a Dirichlet process for all the clones $g_{k,r} = \{i| t_i = t_k, m_i = k\}$. The updating rule of each clone is as follows,

(a) If this clone belongs to a network with $\varphi_k = 1$, we remove the network $k$, the associated $t_{k,0}$ and $a_k$, and update $\Lambda = \{1, ..., K/k\}$.

(b) If this clone belongs to a network with $\varphi_k > 1$, we update the network $k$ by changing $\varphi_k$ to $\varphi_k - 1$.

Then we estimate the probability for each clone to belong to any of the existing networks or a new network following these steps.

(a) We build a new network $K + 1$, the ‘putative center’ $t_{K+1,0}$ is drawn from $G_0(t_{K+1,0}; \lambda)$ and the linear coefficient $a_{K+1}$ is drawn from $\mathcal{N}(a_{K+1}|a, \sigma^2)$. The center of the new network is the clone $t_{k,r}$, which is denoted as $c_{K+1} = t_{k,r}$.

(b) We sample a network $k_s$ from current networks, which includes the new one we created in the last step. The probability of drawing $k_s$ from one of $\{\Lambda, K+1\}$ is proportional to

$$\left[I(k_s \in \Lambda)\varphi_k + I(k_s = K + 1) \times 10^5\right] \cdot MN(t_{k,r}| t_{k,0}, \{b_q\}_{q=1}^{Q}) \cdot N(d_{k,r}^2| q \cdot a_k d_{k,r}^2, \sigma^2_e).$$

where $N(d_{k,r}^2| q \cdot a_k d_{k,r}^2, \sigma^2_e)$ refers to the regression between the TCR sequence and the T cell expression for the sampled $k_s$.

(c) We assign $t_{k,r}$ to the sampled $k_s$ according to the last step. Update $K$ if needed, and rename the associated indices including $\varphi_k, m_i, d_{k,r}, d_{k,r}^2$, and $c_k$.

(II) We update $t_{k,0,q}$ by drawing each $t_{k,0,q}$ from,

$$N\left(\frac{\sum_{r=1}^{\varphi_k} t_{k,r,q}}{b_q \left[1 + \frac{\varphi_k}{b_q}\right]} \cdot \left[1 + \frac{\varphi_k}{b_q}\right]^{-1}\right).$$

SN1 Fig. 1: The tessa model structure.
We apply tessa to analyze T cells from multiple sources.

Let us create an indicator function $a_k$ to determine whether the potential assignment of $k$ to the same source. Otherwise, if the network contains clones from more than one source, we set $a_k = 0$.

The joint posterior distribution in (7) is adjusted to,

$$P(A|B) = \frac{\prod_{k=1}^{K} G_0(t_{k,0}; \lambda) \cdot \prod_{q=1}^{Q} IG(b_q; u, v) \cdot \prod_{k=1}^{K} P(t_{k,l} \mid \{t_{k,0}, \{b_q\}_{q=1}^Q, m_v\}, TN(a; 0, \tau, a > 0) \cdot IG(\sigma^2; u, v) \cdot \prod_{k=1}^{K} \prod_{r=1}^{\phi_k} N(a_k|a, \sigma^2) \prod_{k=1}^{K} \prod_{r=1}^{\phi_k} N(d_{k,c,r}^2|a_k d_{k,c,r}^2, \sigma^2) \cdot I(g)$$

The additional requirement does not change any posterior sampling procedure, except for the sampling of network memberships. The DP probability needs to be updated as $p(k_i | I(g^i))$, where $p(k_i)$ is the same as (9), and $I(g^i)$ tests whether the potential assignment of $k_i$ to $\{\Lambda, K + 1\}$ is in compliance with the prerequisite that networks should only comprise clones from the same subset of data.

### 2.5 Linear relationship between TCR and expression of T cells

In this work, we assumed a linear relationship between TCR distances and expression distances of T cells in each network. One might argue that the linear relationship is not very obvious, as we showed in Extended Data, Fig. 2. In fact, in Extended Data, Fig. 2, the distances are calculated for all TCR clones of one dataset in a pair-wise manner. Therefore, in the right segments of each figure, where the TCR distances are large, such distances are more likely to be from pairs of T cells from different networks. In the left segments, the distances are more likely to be from pairs of T cells from the same networks. We only assume the intra-network distances to follow the linear relationship in the
tessa model. As can be seen from Extended Data. Fig. 2, the relationship between TCR and expression in the left parts of the figures conform better to a linear relationship. For Extended Data. Fig. 2a, we subset the pairwise distances and kept only the top 10% pairs with the smallest TCR distances, likely more of which are intra-network pairs. The correlation coefficient increased from 0.493 (overall) to 0.784 (within this subset), as expected.

We also look further into the networks inferred by tessa and show that the intra-network pairs of the top three networks with the most TCR clones (SN1 Fig 2, dashed lines: regression line between expression distances and TCR distances, data from the healthy-CD8-3 dataset referred in the Supplementary Table.1). The intra-network expression and TCR distances confirm better to a linear relationship when viewed in this way.

On the other hand, assuming a linear model between TCRs and expression of T cells will have the advantage of being simple, which can avoid inadvertently forcing a very high and artificial correlation between TCRs and expression. But our future work can explore other possibilities of assuming non-linear relationships in the tessa model.

3 Simulation study

We conduct simulation studies to evaluate the performance of tessa under various settings. The simulation procedure is described as follows:

(a) We first define $K$ networks for one simulation experiment denoted as $\Phi = \{\phi_k\}$, for $k = 1, \ldots, K$, where $\phi_K$, drawn from $U(1, 2\phi_{ave})$, represents the number of clones of each network, and $\phi_{ave}$ is the average number of clones among $K$ networks. Depending on $K$ and $\Phi$, we generate the simulated network memberships $M = \{m_i\}$, where $m_i \in 1, \ldots, K$, for each of the clone $i$. To simplify the simulation scheme, we skip the step of combining data from different T cells in the same clone when handling the real data. Instead, we directly generate the clone-level data.

(b) Each $b_q$, for $q \in 1, \ldots, Q$, is randomly drawn from an inverse-gamma distribution, the shape and the scale of which are determined to match the real data, we let $Q = 30$, the shape of the inverse-gamma distribution equals to 5, and the scale equals to 2.

(c) For each network, each element of the ‘putative center’ $t_{k,0}$ is randomly drawn from a normal distribution $N(0, (\sum_{q=1}^{Q} \sqrt{b_q} \times t_{dist})^2)$, where $k \in 1, \ldots, K$, and $t_{dist}$ is a diffusion factor that controls the overall TCR distances between network centers.

(d) For each network, the embedding of each TCR clonotype sequence is simulated from a normal distribution $N(t_{k,0,q}, b_q)$, where $q = \{1, \ldots, Q\}$, and different TCR sequences are simulated for the network $k$. All simulated TCR sequences are denoted as a matrix of $t$.

(e) Next, we calculate $d^t_k, c_k, r$ for all $k = 1, \ldots, K$ with $t$ and $t_0$. To simplify the calculation, here we assume that $t_{k,0}$ and $c_k$ are the same, and both of them are real TCR centers for networks. Therefore, $d^t_k, c_k, r = \sum_{q=1}^{Q} \frac{(t_{k,q}-t_{0,q})^2}{b_q}$, where $k = 1, \ldots, K$.

(f) Next we simulate the expression data for each network. We define a three-element expression vector for each clone. To simplify the simulation, we assume there are only three genes. Actually, any number of genes can

SN1 Fig. 2: Scatter plots showing the linear relationships between the intra-network distances.
be assumed, which has limited effect on the simulation results. Again, since the expression distance in tessa is calculated between two TCR clones, here we generate simulated expression data at the clonal level directly. For each network, each of the three elements of a ‘putative expression center’ $c_{k,0}$ is randomly drawn from a normal distribution $N(0, 100)$, where $k = 1, ..., K$, and we define the variance according to our observation of the real data. Only the center clones’ expressions are simulated in this way. The other clones’ expression are not simulated directly, but rather the clone-to-clone expression distances are simulated as described below.

(g) Depending on the linear relationship we described in (1), for each of the network, $d_{k,ck,r}^e$ is simulated from $d_{k,ck,r}^e$. For each network $k$, the expression distance $d_{k,ck,r}^e$ between clone $r$ and center TCR $c_k$ (which is replaced by $l_{k,0}$) is randomly drawn from $N(c_{k,0}, \epsilon_{dist}^2 \times d_{k,ck,r}^e)$, where $k \in 1, ..., K$, and $\epsilon_{dist}$ is a diffusion factor that influences the approximate expression distances between network centers.

We consider the basic simulation setting with $K = 500$, $\varphi_{ave} = 10$, $t_{dist} = 1.5$, and $\epsilon_{dist} = 1$, to match the real data. Additional scenarios are considered in Section 3.3. For each scenario, we repeat the simulation study five times using different random seeds. The MCMC was iterated for 1,000 times, with the first 500 burn-ins.

### 3.1 Validation of the simulated data

To demonstrate that the data from our simulation experiments mimic real data, we first explored the simulated embedded TCRs with t-SNE plots (the same method as Fig. 1e in the main article) and one typical example (one simulated dataset with $K = 500$, $\varphi_{ave} = 10$, $t_{dist} = 1$, $\epsilon_{dist} = 1$) is shown in SN1 Fig 3. According to the simulation figure (SN1 Fig 3 below) and Fig. 1c of the main text, the ratios for within and between network TCR distances, in the simulated and real data, are on a comparable scale.

![SN1 Fig 3: T-SNE plots based on the TCR embeddings of the simulated TCR networks. TCRs in the same network were in the same color. Only the networks with more than 18 TCRs were shown to avoid the figures being overcrowded. The simulated TCR embeddings were adjusted by the simulated weighting factor $b$.](image)

We further calculated the pair-wise TCR distances and expression distances of the simulated TCRs, and analyzed the correlation between the two sets of distances using the same method as in Extended Data Fig. 2 in the main article. Typical examples (two simulated datasets with $K = 500$, $\varphi_{ave} = 10$, $t_{dist} = 1.5$, $\epsilon_{dist} = 1$, and different random seeds) were shown in SN1 Fig 4 (the $r$ represents the Pearson correlation coefficient and the shaded area denotes the 95% confidence intervals for linear regressions). As expected, we found that the similar simulated TCRs were more likely to share similar simulated expression profiles, and the Pearson correlation coefficients between the two sets of distances were similar to those in the real datasets.

### 3.2 Diagnosis for convergence

We employed trace plots and auto-correlation functions (ACFs) to investigate the convergence of the MCMC process in the tessa model. Key parameters including $\alpha$, and $\sigma$ were examined. SN1 Fig 5a and SN1 Fig 6a shows that four parallel chains, initialized from four different random starting points, were well-mixed within 1000 iterations,
indicating convergence of parameters $a$ and $\sigma$ in the tessa model. From SN1 Fig 5b and SN1 Fig 6b we found that the autocorrelations of the log-scaled parameters decreased to being not significantly different from zero at about lag 3, which indicates only small serial correlations and our trace plots are reliable diagnostics for convergence. Furthermore, we calculated the Gelman-Rubin statistic for all the key parameters, where $b$ achieved a point estimate = 1.05 with 95% upper limit being 1.07. For $a$ and $\sigma$, the point estimates were 1.07 and 1.11, and the 95% upper limit were 1.23 and 1.31, respectively. The Gelman-Rubin diagnostic results also suggested the MCMC procedure converged in tessa.

3.3 Model performance evaluation

First we examined the ‘accuracy’ of the tessa networks measured by the Adjusted Rand Index (ARI). ARI estimates the similarity between the simulated clonal memberships $M_{\text{sim}} = \{m_{1}, m_{2}, ... , m_{i}, ..., m_{N}\}$, where $m_{i} \in \{1, 2, ..., K\}$ and $N$ is the total number of clones, and the tessa predicted clonal memberships $M_{\text{pred}}$. An ARI of 0 represents random labeling, and 1 means that the $M_{\text{sim}}$ and $M_{\text{pred}}$ are identical. We vary the following parameters to create different
4 Hyper-parameter selection in real data

In the real data analysis, a standard diffuse prior was chosen for most hyper-parameters, and the inference is not sensitive to the choice of the value. There are two hyper-parameters that were tuned during the tessa model building, which are \( g \) and \( \xi \). \( g \) is used in the prior for \( a_k \) in the formula (11). \( g \) stayed the same for all real datasets and had

SN1 Fig. 6: Trace plots and auto-correlation function plots of \( a \) in four parallel chains.

SN1 Fig. 7a and SN1 Fig 7b showed that within the range of numbers used in the simulation, the number of the total TCR network in each dataset did not have much influence on the accuracy. However, building tessa networks with larger sizes achieved a higher modeling accuracy. We evaluated the real data and found that, mostly \( K \) is larger than 500 and smaller than 1000, and the average number of clones in the networks \( \varphi_{ave} = 10 \). Tessa has achieved high ARIs (\( ARI > 0.90, SD < 0.10 \)) in the simulation tests within the range of the real data.

Next we evaluated whether the \( t_{dist} \), the approximate TCR distances between networks, have influence on the accuracy. As we expected, increasing the \( t_{dist} \) resulted in a more dispersed network structure. Therefore, the results of tessa are more accurate as the networks become better separated by large distances (SN1 Fig 7c). In the real data, the estimated \( t_{dist} \) value is about 1.5. In SN1 Fig 7c we observed that tessa achieved \( ARI = 0.94, SD < 0.001 \) when \( t_{dist} = 1.5 \). Next we evaluated the \( e_{dist} \), which controls the clone-to-clone expression distances within networks. Increasing the \( e_{dist} \) resulted in higher ARIs, but the influence was limited (SN1 Fig 7d), which is expected as tessa does not primarily rely on the clones within the same network have similar or different expression profiles, but rather, emphasizes the regression between expression and TCRs within each network. In the real data, the estimated \( e_{dist} \) value ranges from 0.1 to 1. We found in SN1 Fig 7d that tessa achieved \( ARI \) between 0.91 and 0.94 (\( SD < 0.001 \)) within the real data range of the \( e_{dist} \).

Lastly, we evaluated the tessa estimations of the weighting factor \( b \). As we elaborated in the main article, the weighting factor \( b \) weights the importance of each digit of the TCR sequence embedding, which respect to the expression of T cells. We compared the \( b_{sim} \) vectors that are simulated and the \( b_{pred} \) vectors from the tessa estimation results. Under all simulation conditions, the Pearson correlation coefficients were close to 1, with typical examples shown in SN1 Fig 8.
SN1 Fig. 7: ARIs of different simulation experiments. Error bars plotted as SDs of 5 random repeats, hard to be observed because of small differences.

limited effect on the final estimation outcome. $g$ was chosen to be 0.001 to achieve the best convergence of the model. The hyper-parameter $\xi$ is from the formula (9). $\xi$ determines how likely TCRs are clustered into networks (larger number leads to fewer networks). $\xi$ was 25 (default) for all datasets, except for the four healthy-CD8 datasets (Supplementary Table. 1), where $\xi$ was 40. This choice was made to ensure the clustering rates of TCRs are comparable across the different datasets.

In summary, these simulation analyses have established the robustness of tessa, its good statistics characteristics, and its capability of accurately mapping the functional landscape of TCR repertoire.
SN1 Fig. 8: Correlation between $b_{\text{sim}}$ generated in the simulation and $b_{\text{pred}}$ estimated by tessa.
Supplementary Note 2: Further bioinformatics analyses and discussion of tessa

Functional differences between center and non-center T cell clones

The 10X Immune Profiling technique (the 4 10X datasets) is very powerful and has incorporated multiple modalities. One of the modalities is the measurement of expression of a number of cell surface protein markers (in addition to regular scRNA expression). The captured markers for these cells include cell type sorting markers such as CD3 and CD8, and there are two, HLA-DR and PD-1, which can characterize T cell functional status. We examined the top 1% (in terms of network size) TCR networks that also have at least 3 different TCR clones, and divided the clones into three segments (group 1, group 2, and group 3), according to their tessa-weighted distances from the center clones, where group 1 are the most close clones to centers and group 3 are the most distant clones. We examined the expression of HLA-DR and PD-1 on T cells falling into each segment. It is quite interesting that, in each of the four datasets, a decreasing gradient can be seen, which means that T cells with TCRs more similar to the center TCRs have higher expression of HLA-DR and PD-1. As HLA-DR and PD-1 are markers of T cell activation\textsuperscript{47-49}, the results suggest that there is a gradient over group 1 to group 2 to group 3 such that the T cells with TCRs more similar to the center TCRs are more activated, consistent with their higher clonal expansion (\textbf{Fig. 2c-e}). For all boxplots appearing here and below, box boundaries represent interquartile ranges, whiskers extend to the most extreme data point which is no more than 1.5 times the interquartile range, and the line in the middle of the box represents the median.
**SN2 Fig. 1** The expression of HLA-DR on the surface of the T cells in the TCR clones that were divided into group1, group2, and group3, across the 4 10X datasets. One-sided Jonckheere test was applied to test if there is a decreasing trend of HLA-DR expression from group 1 to group 3. The Fisher’s combined p-value for the four datasets is $1.80 \times 10^{-3}$. The number of clones in the three groups are: $N1=7$, $N2=6$, $N3=19$ (Donor 1), $N1=4$, $N2=4$, $N3=11$ (Donor 2), $N1=4$, $N2=4$, $N3=12$ (Donor 3), and $N1=7$, $N2=7$, $N3=20$ (Donor 4), respectively. Same for the figure below.
SN2 Fig. 2 The expression of PD-1 on the surface of the T cells in the TCR clones that were divided into group1, group2, and group3, across the 4 10X datasets. One-sided Jonckheere test was applied to test if there is a decreasing trend of PD-1 expression from group 1 to group 3. The Fisher’s combined p-value for the four datasets is $1.75 \times 10^{-3}$.

Furthermore, using the RNA expression data of the T cells, we examined genes that showed a monotonous increasing or decreasing trend over group 1 to group 2 to group 3. We input them into the GOrilla server$^{50,51}$. We found that, in these datasets, the monotonously changing genes are enriched in immune activation (group 1>group 2>group 3) and metabolomic terms (group 1<group 2<group 3), which echoes the PD-1 and HLA-DR results above. Due to the issue of space, we only show the Gene Ontology results from the first dataset. The P values shown are from two-sided Hypergeometric tests. The original P values and False Discovery Rates (FDRs) of multiple comparisons are both shown.
SF2 Table 1 Up-regulated pathways in group 1 T cell clones

| GO Term | Description | P-value | FDR q-value | Enrichment N | B | n | b |
|--------|-------------|---------|-------------|---------------|---|---|---|
| GO:0001817 | regulation of cytokine production | 2.35E-07 | 3.64E-03 | 3.23 | 17281 | 689 | 194 | 25 |
| GO:0002376 | immune system process | 5.25E-07 | 4.07E-03 | 1.99 | 17281 | 2323 | 194 | 52 |
| GO:0050776 | regulation of immune response | 1.95E-06 | 1.01E-02 | 2.56 | 17281 | 1045 | 194 | 30 |
| GO:0006955 | immune response | 1.15E-05 | 4.44E-02 | 2.3 | 17281 | 1202 | 194 | 31 |
| GO:0018219 | positive regulation of cytokine production | 3.64E-05 | 1.13E-01 | 3.25 | 17281 | 438 | 194 | 16 |
| GO:0048518 | positive regulation of biological process | 1.13E-04 | 2.93E-01 | 1.30 | 17281 | 5778 | 194 | 90 |
| GO:0048584 | positive regulation of response to stimulus | 1.27E-04 | 2.81E-01 | 1.75 | 17281 | 2240 | 194 | 44 |
| GO:0009987 | cellular process | 1.33E-04 | 2.57E-01 | 1.15 | 17281 | 12992 | 194 | 167 |
| GO:0050778 | positive regulation of immune response | 1.52E-04 | 2.63E-01 | 2.44 | 17281 | 768 | 194 | 21 |
| GO:0045087 | innate immune response | 1.76E-04 | 2.72E-01 | 2.73 | 17281 | 554 | 194 | 17 |
| GO:002682 | regulation of immune system process | 1.77E-04 | 2.50E-01 | 1.91 | 17281 | 1587 | 194 | 34 |
| GO:0017755 | cell activation | 1.92E-04 | 2.48E-01 | 2.23 | 17281 | 958 | 194 | 24 |
| GO:0041131 | cell activation | 2.13E-04 | 2.54E-01 | 4.88 | 17281 | 143 | 194 | 8 |
| GO:0045321 | leukocyte activation | 2.22E-04 | 2.46E-01 | 2.31 | 17281 | 848 | 194 | 22 |
| GO:0006950 | response to stress | 2.51E-04 | 2.60E-01 | 1.59 | 17281 | 2964 | 194 | 53 |
| GO:0070936 | protein K48-linked ubiquitination | 2.87E-04 | 2.78E-01 | 8.57 | 17281 | 52 | 194 | 5 |
| GO:0043122 | regulation of I kappa B kinase/NF-kappaB signaling | 3.00E-04 | 2.73E-01 | 3.84 | 17281 | 232 | 194 | 10 |
| GO:0043170 | macromolecule metabolic process | 3.12E-04 | 2.68E-01 | 1.36 | 17281 | 5679 | 194 | 87 |
| GO:0050929 | positive regulation of NF-kappaB transcription factor activity | 3.38E-04 | 2.63E-01 | 4.69 | 17281 | 152 | 194 | 8 |
| GO:0016197 | endosomal transport | 3.79E-04 | 2.93E-01 | 4.09 | 17281 | 196 | 194 | 9 |

SF2 Table 2 Up-regulated pathways in group 3 T cell clones

| GO Term | Description | P-value | FDR q-value | Enrichment N | B | n | b |
|--------|-------------|---------|-------------|---------------|---|---|---|
| GO:0036222 | cellular protein-containing complex assembly | 3.8E-04 | 1.6E-01 | 2.75 | 17281 | 795 | 182 | 23 |
| GO:0009987 | cellular process | 3.99E-05 | 1.08E-01 | 1.17 | 17281 | 12992 | 182 | 160 |
| GO:0044237 | cellular metabolic process | 3.26E-05 | 1.68E-01 | 1.37 | 17281 | 6925 | 182 | 100 |
| GO:0043170 | macromolecule metabolic process | 3.56E-05 | 1.38E-01 | 1.44 | 17281 | 5679 | 182 | 86 |
| GO:0006413 | translational initiation | 4.24E-05 | 1.31E-01 | 6.28 | 17281 | 121 | 182 | 8 |
| GO:0045669 | regulation of virus genome replication | 5.28E-05 | 1.36E-01 | 7.22 | 17281 | 92 | 182 | 8 |
| GO:0016071 | mRNA metabolic process | 5.43E-05 | 1.20E-01 | 2.89 | 17281 | 591 | 182 | 18 |
| GO:0090304 | nucleic acid metabolic process | 8.03E-05 | 1.56E-01 | 1.85 | 17281 | 2055 | 182 | 40 |
| GO:0043983 | protein-cortactin complex subunit organization | 9.89E-05 | 1.70E-01 | 2.02 | 17281 | 1507 | 182 | 32 |
| GO:0044260 | cellular macromolecule metabolic process | 1.09E-04 | 1.68E-01 | 1.5 | 17281 | 4373 | 182 | 69 |
| GO:0034645 | cellular macromolecule biosynthetic process | 1.15E-04 | 1.67E-01 | 2.35 | 17281 | 928 | 182 | 23 |
| GO:0042733 | ribosomal large subunit biogenesis | 1.48E-04 | 1.92E-01 | 14.61 | 17281 | 26 | 182 | 4 |
| GO:0033441 | cellular nitrogen compound metabolic process | 1.50E-04 | 1.79E-01 | 1.63 | 17281 | 3031 | 182 | 52 |
| GO:0006139 | nucleoside-containing compound metabolic process | 1.51E-04 | 1.67E-01 | 1.7 | 17281 | 2566 | 182 | 46 |
| GO:0049900 | regulation of multi-organism process | 2.06E-04 | 2.12E-01 | 3.24 | 17281 | 381 | 182 | 33 |
| GO:0046003 | protein-cortactin complex assembly | 2.25E-04 | 2.18E-01 | 2.08 | 17281 | 1322 | 182 | 27 |
| GO:0008152 | metabolic process | 2.33E-04 | 2.12E-01 | 1.3 | 17281 | 7698 | 182 | 105 |
| GO:0050792 | regulation of viral process | 2.36E-04 | 2.04E-01 | 4.36 | 17281 | 196 | 182 | 9 |
| GO:0006725 | cellular aromatic compound metabolic process | 2.47E-04 | 2.02E-01 | 1.64 | 17281 | 2776 | 182 | 48 |
| GO:0002244 | hematopoietic progenitor cell differentiation | 2.59E-04 | 2.00E-01 | 6.78 | 17281 | 84 | 182 | 6 |

Further benchmarking tessa against GLIPH, for discovering novel biological insights

We carried out more systematic analyses to compare the results from tessa and also from GLIPH (purely TCR-based analyses) for the datasets used in Fig. 234, in order to demonstrate the capability of tessa in deriving novel biological insights in real data applications.

First we performed the analyses as we did in Fig. 2e, but with GLIPH. Here we show that GLIPH is not able to reveal the gradients of antigen binding affinity within the TCR clusters it identified.
SN2 Fig. 3 GLIPH cannot reveal the decreasing gradient of antigen binding strength for TCRs, along with increasing dissimilarity to the center TCRs of the TCR networks. The same analysis was done as in Fig. 2e, but with GLIPH instead of tessa. The TCR clonotypes from each dataset were divided into six groups of equal size (N(TCR)=198).

We also performed GLIPH clustering on the BCC datasets of Fig. 3. As in our main text, we defined the ‘pre-1, pre-2, post-1, post-2’ subsets based on the clustering results. That is, post-2 clones are from clusters including only post-treatment clones, pre-2 clones are from pre-treatment-only clusters, and pre-1 and post-1 are from mixed clusters. We calculated pathway activity scores for the four subsets. The exhausted pathway is attached below. The figures show that GLIPH clustering is unable to distinguish post-treatment clones into two subsets, because all post clones were clustered with pre clones, leaving no post-treatment only clusters.

SN2 Fig. 4 GLIPH clustering is unable to distinguish post-treatment cells into two subsets. The numbers of T cells in the six subgroups analyzed were: responders: N(pre-1)=1638, N(post-1)=830, N(pre-2)=859; non-responders: N(pre-1)=1023, N(post-1)=828, N(pre-2)=1029, respectively.

Next we applied GLIPH clustering on all the single cell datasets and again calculated the ‘unexplained variations’ as we did in Fig. 4d to test the level of constraint on transcriptomic variations by TCRs, in the GLIPH clusters. In Fig. 4d, we took the top 20%, 40%, 60%, and 80% tessa networks with the largest numbers of unique TCR clones. But GLIPH identified much
fewer clusters with each cluster being much bigger, in each dataset. Therefore, we cannot take this subsetting approach for GLIPH, and thus have showed all GLIPH clusters below:

\[ \text{SN2 Fig. 5} \] The unexplained variance in gene expression of all GLIPH clusters from twelve tumor samples of different cancer types and seven healthy donors.

Our results show that, unlike tessa, GLIPH cannot reveal the differences in unexplained variations between tumor and normal scRNA datasets.

**Controlling for T cell stages when interpreting the TCR-transcriptome associations**

In *Fig. 4*, we derived the observation that the unexplained variations by TCRs are much larger for the tumor datasets than the normal datasets. As we discussed in our paper, TCRs and intrinsic/extrinsic cues both impact the functions of T cells, which is also the conclusions of Tubo *et al*\(^1\) and Buchholz *et al*\(^2\). Therefore, it seems possible that the different levels of TCR/transcriptome association are strongly influenced by the differential stimulations in the tumor and normal contexts. To investigate this possibility, we divided the T cells into different stages, and refined our analyses in *Fig. 4d*.

We first categorized the T cells in each tumor/normal dataset into different functional states. As in Yost *et al*\(^3\), we observed that the T cells formed a V-shaped distribution. At the joint of the two branches, are the naive cells and memory cells. The pseudotime analysis cannot distinguish the naive and memory cells, which is also the same as in Yost *et al*\(^3\). One of the branches has the activated cells and the other branch has the exhausted cells. We distinguished the activated and the memory cells by expression of marker genes. Specifically, we compared the average expression levels of five marker genes between the cells from the two branches. The cells in the activated branch should have higher expression levels of *IFNG*, *TNF* and *FOS*, and the cells in the exhausted branch should highly express *ENTPD1* and *HAVCR2*\(^3\). One example dataset was shown below (the Breast-1 dataset of *Supplementary Table 1*)
SN2 Fig. 6 Pseudotime of T cells. The pseudotime of 2,403 T cells was inferred and overlaid onto the diffusion map.

Next we repeated the analyses in Fig. 4d in each of the naive/memory, activated, and exhausted stages. In Fig. 4d, we carried out a percentage-wise sub-sampling approach at several cutoffs and observed the same phenomenon. Here, as we limited ourselves to a subset of T cells each time (e.g. naive/memory), this sub-sampling approach does not work, as we have lost sample sizes in this sub-sampling process. Therefore, the analyses below included all available cells.

SN2 Fig. 7 Unexplained variances by TCRs in each of the naive/memory, activated and exhausted stages in the twelve tumor datasets and seven healthy datasets.

From the figure above, we make two interesting observations:

(1) The unexplained variances in the activated and exhausted cells in the tumor datasets are still larger than the normal datasets (P-values of 0.117 and 0.077 from one-sided Student’s T-tests,
respectively), but the differences are smaller than in Fig. 4d. The unexplained variances in the naive/memory cells serve as a control (P-value of 0.186 from one-sided Student’s T-test), which shows an opposite trend. These differences suggest that indeed the different functional statuses of T cells in tumor and normal contexts, as a consequence of differential stimulation, strongly influence how much TCRs constrain the functions of the T cells. Furthermore, the fact that activated and exhausted T cells still show more unexplained variances in tumor contexts than normal contexts suggest that there could be additional non-TCR-dependent cues from the tumor micro-environment that affected the functions of T cells that are supposed to be executing cytotoxic roles in tumor cells.

(2) On the other hand, in the normal contexts, T cells in the activated and exhausted stages have less unexplained variances by the TCRs than the naive/memory cells. But this phenomenon is reversed in the tumor context. This is also interesting, as in the normal contexts, the environmental cues are relatively quiescent, and TCRs may play more important roles in influencing T cell transcriptomics, when activated in a TCR-dependent manner. In comparison, the tumor microenvironment likely confers more complicated cues upon the T cells, especially the activated and exhausted subsets, so that TCRs of T cells could play less influential roles.

Re-coloring Fig. 3a

To make it easier to discern the different groups of patients (e.g. responder pre-1) in Fig. 3a, we have used 8 different colors to re-label these 8 different groups and shown the figure below.
**SN2 Fig. 8** t-SNE plot of the pre- and post-treatment T cells from all the BCC patients (dataset BCC, Supplementary Table 1). The colors represent clonal level labels.

This coloring scheme will make it easier to discern the 8 different groups of cells. But the coloring scheme in **Fig. 3a** makes it easier to be compared against **Fig. 3b**, which has a similar coloring scheme. We encourage the readers to consider these three figures together to visualize the changes between clone level labels and network level labels. The group of T cells that we want the readers to focus on is the “responder post-1” T cells, which is colored in orange in **SN2 Fig. 8**.

**Quantifying the mixing effects of the clones in the CTCL patient and the healthy donor**

For **Fig. 4ab**, we objectively quantified the mixing effects of the clones with Kullback-Leibler (KL) divergence between the original clonal distributions and randomly mixed distributions. For each dataset, we fixed the positions of cells on the t-SNE plot (**Fig. 4ab**) and gridded the t-SNE space into n x n (we selected n=10) squares. For cells from each clone, we calculated their frequency in the n² bins according to their t-SNE coordinates. Then we randomly permuted the clone labels of the cells and calculated their random frequency in these bins again. Then for each clone, a KL divergence statistic between the two sets of frequencies was calculated and regarded as ‘the relative difficulty for one clone becoming randomly distributed around the t-SNE space’.
The KL divergence calculation was repeated 1,000 times for each clone and the maximum S.D. across all the clones was 2.8. The 1,000 repeats were averaged to generate the KL divergence for each clone. As we show below, the healthy control clones have generally larger KL divergences. Therefore, the clone frequency distributions are less random or less ‘mixed’ than those from the CTCL patients. One-tailed student’s t-test was performed between the two groups of KL divergence statistics, and the P-value was 0.011.

SN2 Fig. 9 KL divergence statistics of the CTCL T cell clones and the healthy donor T cell clones for quantifying the mixing of the clones in the t-SNE space. N(CTCL)=416 and N(Healthy)=433.

This KL divergence calculation used the true sample size and clone size information, which were the same for every round of simulation. Therefore, sample sizes and clone sizes were naturally controlled.

Additional discussion on the significance of tessa

Here we provide additional discussions of the significance of our work. They are adapted from the comments of one of our reviewers, whom we acknowledge deeply for providing such a nice summary of our work.

Tessa is built upon the landmark studies of GLIPH and TCRdist, which allowed for an analysis of ‘likeness’ or ‘distance’ between TCRs. Prior to this, TCRs were considered either clonally distinct or clonally identical, and therefore analyses of TCR repertoires were limited in that there was no provision for appreciating the ‘effective diversity’ of any epitope specific population. Our current study builds significantly on those by overlaying an analogous TCR distance algorithm with an analysis of transcriptional distance, which is of huge significance to the field especially in light of the numerous platforms now that support simultaneous resolution of both TCR sequences and transcriptional profiling. The bottleneck now really is in the analysis and
interpretation of such parallel datasets and an algorithm, such as tessa, which aligns the two, would provide a substantial advance than what is currently available.

This work is powerful as it will provide not just critical fundamental information about the extent, to which TCR groupings are associated with distinct transcriptional profiles (and thus the extent to which the TCR is responsible for transcriptional phenotypes), but will enable the identification and selection of functionally potent TCR groupings when determined in the context of epitope specific populations. This in turn could facilitate the selection of TCRs for engineering/optimization/expression in clinical scenarios for the treatment of chronic viral infections or cancer immunotherapies. As an extension, one can imagine tessa facilitating the structural resolution of TCR recognition of pMHC class I proteins that drives a transcriptionally optimal phenotype. That is, by linking TCR clonotypes with transcriptional profiles, it will facilitate a deeper understanding of the key recognition characteristics that underpin this association.

References

47. Saraiva, D. P., Jacinto, A., Borralho, P., Braga, S. & Cabral, M. G. HLA-DR in Cytotoxic T Lymphocytes Predicts Breast Cancer Patients’ Response to Neoadjuvant Chemotherapy. Front. Immunol. 9, 2605 (2018).

48. Hong, J. J., Amancha, P. K., Rogers, K., Ansari, A. A. & Villinger, F. Re-evaluation of PD-1 expression by T cells as a marker for immune exhaustion during SIV infection. PLoS ONE 8, e60186 (2013).

49. Petrovas, C. et al. PD-1 is a regulator of virus-specific CD8+ T cell survival in HIV infection. J. Exp. Med. 203, 2281–2292 (2006).

50. Eden, E., Navon, R., Steinfeld, I., Lipson, D. & Yakhini, Z. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics 10, 48 (2009).

51. Eden, E., Lipson, D., Yoge, S. & Yakhini, Z. Discovering motifs in ranked lists of DNA sequences. PLoS Comput. Biol. 3, e39 (2007).
### Supplementary Table 1 Data cohorts and details.

| Cohort       | Reference          | Accession methods                                                                 | #T cells | #Unique TRB sequences | Role                        | Data type                                                                 |
|--------------|--------------------|-----------------------------------------------------------------------------------|----------|-----------------------|-----------------------------|---------------------------------------------------------------------------|
| TCGA         | Citation 23        | https://gdc.cancer.gov/about-data/publications/painnnucm                          | 204,881  | 181,787               | Training auto-encoder       | TCRs called from tumor RNA-Seq                                             |
| Kidney-bulkRNA | Citation 24        | https://github.com/jcạo9757/TESSA/tree/master/Tessa_released_data                | 80,801   | 75,157                | Training auto-encoder       | TCRs called from tumor RNA-Seq                                             |
| IEDB         | https://www.iedb.org/ | https://www.iedb.org/database_export_v3.php                                        | 93       | 58                    | Training auto-encoder       | TCR database export                                                       |
| McPAS        | Citation 26        | http://friedmanlab.weizmann.ac.il/McPAS-TCR/                                     | 702      | 521                   | Training auto-encoder       | TCR database export                                                       |
| NSCLC-1      | Chromium Single Cell Immune Profiling Solution on the 10x website                 | https://support.10xgenomics.com/single-cell-vdj/datasets/2.2.0/vdj_v1_hs_nsccl_5ge | 1,741    | 1,401                 | Training auto-encoder & Tessa                                             | ScRNA-Seq multiplexed with TCR-Seq                                       |
| NSCLC-2      | Citation 26        | EGAS00001002430                                                                  | 3,628    | 845                   | Training auto-encoder & Tessa                                             | ScRNA-Seq multiplexed with TCR-Seq                                       |
| CRC          | Citation 27        | EGAS00001002791                                                                  | 3,463    | 891                   | Training auto-encoder & Tessa                                             | ScRNA-Seq multiplexed with TCR-Seq                                       |
| HCC          | Citation 28        | EGAS00001002072                                                                  | 3,536    | 2,506                 | Testing auto-encoder & Tessa                                              | ScRNA-Seq multiplexed with TCR-Seq                                       |
| Breast-1     | Citation 29        | GSE114727, GSE114724                                                             | 6,376    | 4,145                 | Training auto-encoder & Tessa                                             | ScRNA-Seq multiplexed with TCR-Seq                                       |
| Breast-2     | Citation 29        | Same as above                                                                    | 6,289    | 4,028                 | Training auto-encoder & Tessa                                             | ScRNA-Seq multiplexed with TCR-Seq                                       |
| Breast-3     | Citation 29        | Same as above                                                                    | 4,455    | 2,246                 | Training auto-encoder & Tessa                                             | ScRNA-Seq multiplexed with TCR-Seq                                       |
| Tissue | Citation | Sample | Training Method | Seq Configuration |
|--------|----------|--------|-----------------|-------------------|
| Breast-4 | Citation 29 | Same as above | 4,840 | 2,452 | Training auto-encoder & Tessa | ScRNA-Seq multiplexed with TCR-Seq |
| Breast-5 | Citation 29 | Same as above | 4,189 | 2,243 | Training auto-encoder & Tessa | ScRNA-Seq multiplexed with TCR-Seq |
| Healthy-CD8-1 | Chromium Single Cell Profiling Solution on the 10x website | Same as above | 12,784 | 9,502 | Tessa | ScRNA-Seq multiplexed with TCR-Seq |
| Healthy-CD8-2 | Chromium Single Cell Immune Profiling Solution on the 10x website | Same as above | 16,246 | 4,846 | Tessa | ScRNA-Seq multiplexed with TCR-Seq |
| Healthy-CD8-3 | Chromium Single Cell Immune Profiling Solution on the 10x website | Same as above | 7,839 | 4,389 | Tessa | ScRNA-Seq multiplexed with TCR-Seq |
| Healthy-CD8-4 | Chromium Single Cell Immune Profiling Solution on the 10x website | Same as above | 10,873 | 6,817 | Tessa | ScRNA-Seq multiplexed with TCR-Seq |
| Healthy-PBMC-1 | Chromium Single Cell Immune Profiling Solution on the 10x website | | 709 | 589 | Tessa | ScRNA-Seq multiplexed with TCR-Seq |
| Healthy-PBMC-2 | Chromium Single Cell Immune Profiling Solution on the 10x website | | 540 | 450 | Tessa | ScRNA-Seq multiplexed with TCR-Seq |
| Melanoma | Citation 30 | GSE123139 | 3,615 | 1,690 | Tessa | ScRNA-Seq multiplexed with TCR-Seq |
| BCC | Citation 31 | GSE113590 | 6,207 | 2,482 | Tessa | ScRNA-Seq multiplexed with TCR-Seq |
|                  | Citation | ExpId       | Seq       | Seq Description                                      |
|------------------|----------|-------------|-----------|------------------------------------------------------|
| ECCITE-CTCL      | Citation 16 | GSE126310   | Tessa     | ScRNA-Seq multiplexed with TCR-Seq                   |
| ECCITE-ctrl      | Citation 16 | GSE126310   | Tessa     | ScRNA-Seq multiplexed with TCR-Seq                   |
| Glanville        | Citation 10 | https://www.nature.com/articles/nature22976#MOESM1 | Test antigen-specificity | TCR-antigen specificity data                         |
| Dash             | Citation 11 | SRP101659   | Test antigen-specificity | TCR-antigen specificity data                         |
**Supplementary Table 2** The genes in the T cell functional pathways related to **Fig. 3e-g** and **Extended Data Fig. 6.**

| Naive   | Memory | Activated | Exhausted | Inhibition |
|---------|--------|-----------|-----------|------------|
| IL7R    | GZMK   | IFNG      | GZMB      | TGFB1      |
| PABPC1  | CST7   | JUNB      | GNLY      | LDHA       |
| CCR7    | DUSP2  | SLC2A3    | ENTPD1    | FTH1       |
| RPS12   | SH2D1A | CD69      | PRF1      | SRGN       |
| FTH1    | TRAT1  | NFkBIA    | KRT86     | MTRNR2L8   |
| LTB     | TC2N   | PPP1R15A  | ACP5      | SRSF2      |
| ANXA1   | CMC1   | TNF       | GZMH      | CREM       |
| TCF7    | LITAF  | NR4A2     | GAPDH     | RPS26      |
| S1PR1   | EOMES  | FOS       | AC092580.4| SH2D2A     |
| RPL13   | GIMAP4 | CCL4      | GALNT2    | PABPC1     |
| MT-ND2  | CNN2   | NR4A1     | ITGAE     | HNRNPUL1   |
| EEF1B2  | COTL1  | DUSP1     | LAYN      | ATP1B3     |
| CD55    | RPS15A | HSPA1B    | CTSW      | ZFP36      |
| CXCR3   | IER2   | CXCL13    |           |            |
| LDHA    | GADD45B| TNFRSF18  |           |            |
| RPS26   | JUN    | AHI1      |           |            |
| TRMO    | CSRNP1 | CXCR6     |           |            |
| FAM102A | CDKN1A | VCAM1     |           |            |
| APOBEC3G| TSC22D3| JAML      |           |            |
| MTRNR2L8| ALOX5AP|           |           |            |
| IL9R    |        | LAG3      |           |            |
| TGBF1   |        | LSP1      |           |            |
| HLA-B   |        | PTMS      |           |            |
| PLP2    |        | HLA-DRB1  |           |            |
| CCR7    |        | HLA-DPB1  |           |            |
|         |        | CCL5      |           |            |
|         |        | CD74      |           |            |
|         |        | HLA-DRA   |           |            |
|         |        | HLA-DPA1  |           |            |
**Supplementary Table 2** The genes in the T cell functional pathways related to Fig. 4e.

| IL-2 pathway #1 | IL-2 pathway #2 | IFN-a/b | IL-12 pathway #1 | IL-12 pathway #2 | Control pathway #1 |
|-----------------|-----------------|---------|------------------|------------------|--------------------|
| IRF4            | NUP62           | APBA2   | DCAF17           | TPTE             | IPMK               |
| CD69            | ESM1            | ZNF574  | PAPPA2           | CUL2             | INPP5D             |
| BATF            | NPLGC4          | ZNHit3  | C2orf88          | CCDC70           | ALAS2              |
| CCR7            | TMPO            | ZDHHC9  | AVEN             | FANCM            | YWHAE              |
| IL2RA           | CD81            | XCR1    | MB21D1           | CEP192           | ATG4B              |
| PRDM1           | APIP            | WNT10B  | GCAT             | CCDC122          | HADHB              |
| MTOR            | ATP1A2          | VAPA    | PTPRM            | HAUS6            | EFTUD2             |
| HRAS            | FARp1           | UFD1L   | TRAT1            | GPR34            | CCNE2              |
| KRAS            | MAPRE2          | TUSC2   | ATP1B3           | PTGER2           | SRSF2              |
| PRF1            | RAB6A           | TULP3   | GUCY2C           | PCGF6            | GSTA1              |
| GZMB            | SH2D2A          | TTF1    | CAPN9            | UHRF1BP1L        | GNS                |
| IFNG            | DNA2            | TSFM    | NEURL            | RIBC1            | FMO5               |
| HMGCR           | CIT             | TRHR    | LOC389493        | USP46            | PHKA1              |
| PDSS1           | TPBG            | DISP1   | MDM2             | SLC27A2          |                    |
| VASH1           | TMEM85          | GTF2H1  | PPP1R15B         | GUCY1A2          |                    |
| CBFB            | TMEM30B         | RAB1A   | TNPO3            | PIDD             |                    |
| CARHSP1         | TMEM159         | TIPRL   | ORC2             | ALG9             |                    |
| UBR7            | TIGD5           | ZCCHC2  | NPEPPS           | UGT2A3           |                    |
| TPBG            | THOC7           | RPN1    | GRINA            | CYP2C9           |                    |
| CCDC41          | THBS3           | C9orf41 | SAMSN1           | GM2A             |                    |
| ANGPLT2         | THAP7           | MAPRE1  | VNN1             | PPP1CA           |                    |
| G2E3            | TFEB            | CEP57L1 | LARP4            | GRN2A            |                    |
| RPA2            | TEKT2           | TMEM161B| MIS12            | PRKACG           |                    |
| STIL            | TBPL1           | CLP1    | ASZ1             | LDLRAP1          |                    |
| RBM44           | TBC1D23         | EIF2B3  | MPHOSPH10        | RAF1             |                    |
| STK39           | TALDO1          | SSPN    | PLS3             | FLT1             |                    |
| BMP2K           | SYT7            | MLF1IP  | RAD50            | PRPF38A          |                    |
| WDR62           | SYNGR2          | BTD     | UBE2Q2           | GTF2H3           |                    |
| SOD1            | SWAP70          | ST18    | GNL2             | PLCD1            |                    |
| ICA1            | STX2            | ACTC1   | GOLGA7           | AOC3             |                    |
| Gene  | Gene  | Gene  | Gene  | Gene  |
|-------|-------|-------|-------|-------|
| CA13  | SRPK1 | MORC2 | SCN7A | NEU1  |
| PMCH  | SRMS  | EEPD1 | KPTN  | PMAIP1|
| RPA1  | SPATA5| IMP4  | CEP170| ADCY9 |
| CLIC4 | SON   | KCNE3 | FOLR2 | SYNJ1 |
| CASP7 | SNX2  | TMEM181| CX3CR1| DCXR  |
| ACTG1 | SLC6A6| WDR45L| PDE4DIP| ABCA1 |
| C6orf168| SLC38A4| UMODL1| DNM1L | BRAF  |
| PRICKLE3| SLC35A2| PGM5   | HELQ  | GGT1  |
| DUSP14| SLC30A1| TRIM66 | SESTD1| PPAP2B|
| POLE  | SLC25A28| ZNF18  | GSPT2 | GYS2  |
| PMVK  | SLC22A12| MAB21L3| EEF2K | ADSSL1|
| FUT4  | SLC20A1| GNPDA2 | FAF2  | FTMT  |
| EXPH5 | SIRT2 | SEC24D | SLC33A1| UGT1A5|
| GPD2  | SFTPC | SGTA  | RNF115| ACO1  |
| LGALS1| 4-Sep | MED7  | PTPLB | SDS   |
| LSM2  | 10-Sep| IL1RL1 | CYSLTR2| ENTPD1|
| DAP   | SENP6 | FTSJD2 | TERF2IP| Treh  |
| LCLAT1| SEMA6C| NUDCD1 | HINFP | CDA   |
| IL10  | SEC23A| LRRC59 | PTMA  | GNS   |
| MYBL1 | SEC22B| AGPS  | UTP14A| ABL1  |
| PSMD14| SAR1A | MDH2  | USP1  | HADH  |
| AMDHD2| SAA4  | SBDS  | NEK4  | ADH1A |
| EIF4H | RYK   | C15orf57| TIAL1 | ITPKB |
| PIF1  | ROBO3 | LCOR  | HOOK1 | NT5C2 |
| ACER3 | RNASE3| SHCBP1L| CASP3 | PSMD3 |
| TPI1  | RGS4  | TMEM97 | ATP9B | CYP1A1|
| MMD   | RERE  | CHCHD2| NR4A1 | PRIM2 |
| RDBP  | RENBP | RGAG4 | ACACA | CRLF2 |
| BCAT1 | REG3G | PCCB  | MRFAP1| RPS6KA2|
| HIF1A | RAX   | C7orf42| SBF2  | MCEE  |
| GNB4  | RASA4 | CKAP2L | HNRNPD| NRG2  |
| FBXO5 | RAB5A | RBL1  | DARS  | PRPS2 |
| KDM2B | PTPN23 | RGMA | SCNN1G | HMGCS1 |
|-------|--------|------|--------|--------|
| FAM81A | PSTPIP2 | SDR42E1 | PPP1R3B | B4GALT1 |
| ERBB3 | PSMC1 | ERP27 | CBFA2T2 | DRD4 |
| SERPINF1 | PRSS58 | C21orf91 | ZC3H15 | ALDH3A1 |
| INTS7 | PRPF8 | FSCN3 | CLPX | ADRB2 |
| RAD51AP1 | PRPF39 | RBM25 | TBC1D7 | PTH2R |
| CENPN | PRAF2 | GALNTL5 | NAV2 | LYN |
| SRGN | POU3F3 | NUDT3 | HNRPLL | MARS |
| TRIP12 | PON1 | MRAS | PRDM5 | BMPR2 |
| GTSF1 | POLR3D | SERPINB11 | SHROOM2 | CASP3 |
| GINS2 | POLR3A | SGK1 | ODF2L | CCL26 |
| GLRX | PLAU | CDK9 | IDI1 | PGK1 |
| STAU2 | PIP4K2C | MRPS30 | MRPL32 | UGT2B10 |
| SLC25A20 | PHLDA2 | PCDHB3 | VPS33B | GGT7 |
| SLMO2 | PGAM2 | BDP1 | LY96 | ACSS3 |
| FAM111A | PEMT | BDKRB2 | RFX5 | GADD45B |
| SHC4 | PCDH7 | RPAP3 | SRGAP2 | PPAP2C |
| DSP | P2RX6 | APPBP2 | GNA13 | AK7 |
| SLC25A10 | OR6A2 | PLAG1 | CDC14A | ALG12 |
| TUBB3 | ORPD1 | DOPEY1 | MRGBPF | ABL2 |
| IL1R2 | OMD | NUCB1 | TMX1 | EGFR |
| THOC6 | ODZ1 | SYT12 | PVT1 | CCR6 |
| DERL2 | NUP62 | B4GALT7 | CPEB2 | PRIM2 |
| OSBPL9 | NRCAM | CHIC2 | APAF1 | WARS |
| ORC6 | NKAIN1 | BICD2 | DNAJB2 | PIP4K2B |
| PRELID2 | NCOA1 | C20orf94 | DNAJC3 | BUB1 |
| ANXA7 | NCK2 | PRKACA | SNX16 | CALM2 |
| CHEK1 | NAT6 | IL12RB1 | NFYA | SETD1A |
| INCENP | NAB1 | ST13 | CENPF | CMPK2 |
| LITAF | NAA11 | TOX3 | ACER3 | PLA2G3 |
| UBASH3B | MYT1 | XPNPEP3 | TMBIM1 | JMJD7-PLA2G4B |
| CPD | KAT7 | DCTN5 | TRO | CACNA1S |
| RORC  | MYO10 | GCNT2 | LAGE3 | GPAT2 |
|-------|-------|-------|-------|-------|
| UBE2T | MXI1  | APIP  | DTX3  | ODC1  |
| MTCH1 | MTHFR | POLR3A| KDEL2C| XCR1  |
| HMGN3 | MSH3  | GTF2F2| POMT1 | POLR2H|
| DEPDC1B| MRPS7 | GABARAPL1| TBL1X| UGT2A1|
| MELK  | MRPS18B| XIST | ABCA7 | DCP1A |
| TNFRSF9| MRPL3 | DDX20 | DHX29 | RFC2  |
| HMGB3 | MORC4 | KIF5B | MFSD6 | ALOX12|
| CYP11A1| MLLT1 | CYP4A11| POLK  | ABCD2 |
| SPRED1| MINA  | CYP2C8| CTSD  | STX19 |
| KLHDC2| MIF4GD| RGS13 | ZHX3  | PTPLB |
| TSSC1 | MFNG  | VWC2  | EIF2AK4| PPP3CC|
| H2AFX | METTL7A| RHBDL2| USP37 | TP53AIP1|
| CDC25C| MEST  | C19orf28| PPAP2A| POLD4 |
| SLC12A8| MCM5  | ATXN1 | CENPC1| DGAT1 |
| ZBTB32| MBD4  | TAF9  | SH3BGRL| ASNS  |
| IL12RB2| MAST2 | TOP3B | UHMK1 | SIAH1 |
| IFIH1  | MAPKAPK5| GNGT1 | AIFM2 | PIK3CB|
| DUSP4 | LOC100506612| FAM120C| SERPINB12| HSD17B2|
| C9orf23| LITAF | ACAD8 | RNF11 | LOC652346|
| C9orf41| LGALS9B| CCNO  | ARHGAP5| CCL14 |
| LRR1   | LEPREL2| CERCAM| NSMCE2| CPS1  |
| ABCC4  | LCT   | PDE4D | ADSS  | CDC5L |
| MAP2K3 | LAMA5 | TNFAIP6| CCDC41| STX10 |
| RAF1   | LAMA4 | FAM53B| BCAP29| SCP2  |
| TYMS   | KTN1  | C17orf108| TBC1D5| FPGS  |
| GCNT1  | KRT6A | PAFAH1B1| PHACTR4| GADD45G|
| GPR25  | KPNA6 | ALG14 | SLC30A6| LYPLA2|
| CCDC18 | KLF4  | DACH1 | LGR4  | PIGX  |
| FHL2   | KLF12 | MYO3B | DEK   | SEPHS1|
| SMC6   | KCNK3 | OAZ2  | FBXW11| LDHD  |
| PCNA   | KCMF1 | WNT5A | HNRNPA3| LIPA  |
| Gene 1 | Gene 2 | Gene 3 | Gene 4 | Gene 5 |
|-------|-------|-------|-------|-------|
| ERMN  | ITGAV | DNAJB4 | MMD   | ADI1  |
| C3orf37 | IQGAP2 | CLIC4 | ARHGAP12 | NDUFB6 |
| CYB5B | INMT | PIGH | SLAMF6 | ZNRD1 |
| PASK | IL1R1 | FAM65B | RNF25 | SLC8A1 |
| CYTH2 | IL1A | ACER2 | A1CF | STMN1 |
| PSMD13 | IFIT3 | CLASP1 | TPMT | CCNE2 |
| KIAA1522 | IFIT2 | U2SURP | RAF1 | MAP2K2 |
| ADAP1 | IER3 | EIF4A1 | AHRR | LTA |
| HMOX1 | IBSP | SPG20 | DCP1B | |
| KLC3 | HTR3A | GRPEL1 | RAC2 | |
| CKAP2L | HSPA5 | PPP2R5E | ADRA2A | |
| EHBPl1 | HPD | SLC1A4 | PRKACB | |
| RTCD1 | HOXB7 | HNRNPM | TCIRG1 | |
| BRIP1 | GRPEL2 | BCAS1 | AARS | |
| GSR | GNPTG | C17orf89 | PLA2G3 | |
| FAM156A | GJA4 | TRIO | ADH1A | |
| GSTM5 | GFRA1 | ELN | DLAT | |
| BCL3 | GABRB1 | COQ3 | GOT1 | |
| SEMA4C | G0S2 | KIRREL3 | GALR2 | |
| BLM | FZD9 | CYP7A1 | GPX5 | |
| EPDR1 | FRAT1 | ALX1 | UGT2A3 | |
| ANKRD50 | FOSB | DNAJC24 | COX5A | |
| RAPGEF5 | FOLR2 | BCAP29 | RPN2 | |
| TREX1 | FKBp5 | AMMECR1L | CCR2 | |
| C11orf31 | FKBp10 | NEU2 | LPAR3 | |
| FKBp2 | FGD1 | CSMD1 | RPL34 | |
| FOXM1 | FCER2 | EGLN3 | BDH1 | |
| C11orf51 | FAM110A | PSEN1 | MGAT3 | |
| SMTN | EPHB2 | CEP128 | PDE6H | |
| SNRNP25 | EOMES | ASNS | GAD2 | |
| SH3RF1 | ENO1 | ATF3 | PSMC6 | |
| KIF20B | EIF2B4 | FLJ37453 | ACADM | |
| Gene  | Gene  | Gene  | Gene  |
|-------|-------|-------|-------|
| STOM  | EFNA5 | CD80  | CARS  |
| CDKN1A| E2F5  | NDUFA12| DPYD  |
| TOX2  | DVL2  | VWA3A | RPL27A|
| NUP93 | DUT   | XYLBJ| ARSA  |
| EHD4  | DPF3  | DNAJC12| JMJD7-PLA2G4B|
| PTGIR | DNAJB13| C3orf22| MDH1  |
| POC1A | DKK3  | KIAA1161| CERK  |
| UBL4A | DEK   | CLTC  | TAF9  |
| NUCB1 | DEGS1 | DRGX  | CD70  |
| FMR1  | DDX10 | ALS2  | AHCY  |
| FIGNL1| DCAF13| NUP43 | AWAT2 |
| ELF4  | CYR61 | KRT4  | GFPT1 |
| GARS  | CYP2A6| QRICH1| TGFB1 |
| TIMP2 | CTSC  | C9orf93| P4HA1 |
| TUFT1 | CSRPl | DDX39 A| LTB4R2|
| SRSF12| CRX   | HDHD3 | LIF   |
| KIF18A| CREBBP| SIPA1L1| CACNA2D4|
| BEND4 | COX7A1| TMED5 | RPA3  |
| AGPAT9| COL8A1| RALYL | HAGH  |
| C16orf61| COASY| PKD2L2| POLR2B|
| NSMAF | CMA1  | ANKRD34A| S1PR4 |
| ARHGAP19| CLPTM1L| CUL3 | SEC11A|
| GCG   | CLP1  | TAF1B | GRID1 |
| EXO1  | CIDE C| IL10  | GPDP1L|
| HMHR | CFI   | LINC00301| AGTR1 |
| TUBE1 | CELF1 | EXOSC8| GNPDA2|
| BLMH  | CDV3  | NAA50 | FABP7 |
| E2F8  | CDKN2D| GHSR  | UGT1A10|
| PRICKLE1| CD70 | EIF3B | AMDHD2|
| UBXN8 | CCND1 | EIF2B5| LSM3  |
| CDC45 | CCL4  | USP37 | YWHAH |
| BUB1  | CCKAR | YTHDF2| CCNB2 |
|    | FRMD4B | CCDC6 | OR4E2 | HGSNAT |
|----|--------|-------|-------|--------|
| IQGAP3 | CCDC28B | TAL1 | PPAP2B |
| PYCARD | CBFA2T3 | EIF2C3 | LAPT4A |
| KIF18B | CAT | SRRM1 | RAD50 |
| RCC1 | CAPZA2 | LYPLA2 | CCND2 |
| UBR5 | CAMK2D | P2RY14 | ADH5 |
| POLH | CA3 | MMP3 | PPCDC |
| C15orf23 | C9orf78 | CLVS2 | LAMTOR3 |
| CDK2AP1 | C20orf30 | PAK1IP1 | GCAT |
| MMP16 | C12orf41 | ABCC3 | PSMB11 |
| Control pathway #2          |
|----------------------------|
| B4GALT1                    |
| LAPTM4B                    |
| POLL                      |
| TAF2                      |
| DHRS9                     |
| CAMK2G                    |
| ALDH7A1                   |
| PLCB3                     |
| ALDH9A1                   |
| EHHADH                    |
| ABCA13                    |
| C1GALT1                   |
| LCMT2                     |
| CCKAR                     |
| ALDH9A1                   |
| PPP3CA                    |
| B3GNT1                    |
| UGT2B4                    |
| HTR4                      |
| RBX1                      |
| PNLIPRP2                  |
| GALC                      |
| CNTF                      |
| ATP2B2                    |
| FDFT1                     |
| DPYS                      |
| ABCB8                     |
| PSME2                     |
| ADORA2B                   |
| TRIM37                    |
| Gene      |
|-----------|
| FGF6      |
| SUOX      |
| HNRNPU    |
| SULT2B1   |
| GALT      |
| QDPR      |
| NAT6      |
| HRH1      |
| DNMT3B    |
| PIK3R1    |
| XAB2      |
| MAP2K7    |
| CLTB      |
| TAB1      |
| DGKQ      |
| RELT      |
| FPGS      |
| MAN2B1    |
| ATP6V0A2  |
| IL15RA    |
| GFPT2     |
| GSTT1     |
| RPA4      |
| IDUA      |
| RNASEH2A  |
| RAC1      |
| BCAT2     |
| CYP26C1   |
| TYMP      |
| BUB3      |
| LOC650621 |
| CDC16     |
| Gene   |
|--------|
| PIK3CD |
| ATG12  |
| POLR1D |
| ACOX2  |
| CYP2E1 |
| ACER2  |
| RBKS   |
| RPS9   |
| CCNE1  |
| ALDH2  |
| ATP8   |
| EHHADH |
| GNG7   |
| EDAR   |
| AP3S2  |
| UGT2B7 |
| COX7A2L|
| CACNA2D1|
| IL29   |
| NT5C3  |
| PNP    |
| MGST3  |
| CYP3A5 |
| NT5C1A |
| CYP1A1 |
| GSTA4  |
| SOAT2  |
| GRB2   |
| ALDH6A1|
| PRUNE  |
| ATP12A |
| CDKN2A |
| Gene       |
|------------|
| IFNGR1     |
| PTH2R      |
| CLTA       |
| GMPR       |
| GUCY2D     |
| LYN        |
| ADCY6      |
| ALDH9A1    |
| ADCY1      |
| CXCL10     |
| MAP2K3     |
| FADS2      |
| GOT1       |
| PGAM4      |
| GSTK1      |
| RPL3       |
| FTCD       |
| UGT1A10    |
| PTAFR      |
| BAAT       |
| SPTLC2     |
| PRKCA      |
| FPR1       |
| UGT2A3     |
| PPAP2A     |
| LPAR2      |
| HTR1A      |
| SYNJ2      |
| ALDH5A1    |
| RYR2       |
| PRIM2      |
| SORD       |
| Gene      |
|-----------|
| CHMP4A    |
| GLB1      |
| LTC4S     |
| SFN       |
| RAD50     |
| GK        |
| GDF5      |
| TRPV1     |
| ETNK2     |
| ABCC8     |
| CYSLTR1   |
| MKNK2     |
| CACNA1I   |
| ATP6V1H   |
| RDH11     |
| POLR3D    |
| POLR1A    |
| GGT1      |
| HSD17B12  |
| ALDH2     |
| IDS       |
| AKR1B10   |
| GSTM3     |
| AKT2      |
| DUT       |
| INS       |
| EDC4      |
| OXCT1     |
| ZAK       |
| M6PR      |
| GNG12     |
| UGT1A9    |
| Gene  |
|-------|
| PLD1  |
| TSHR  |
| GLS2  |
| METTL2B|
| CCNE2B|
| PIGZ  |
| B4GALT4|
| GNG5  |
| ALDH9A1|
| ACO2  |
| BCAT1 |
| CES1  |
| UGT1A10|
| TBXAS1|
| GRM5  |
| GSTM2 |
| TAF11 |
| VAMP3 |
| IL5RA |
| ALDH2 |
| PIGG  |
| IFNG  |
| FADS2 |
| IFNA7 |
| HAGH  |
| AMH   |
| TAF4B |
| ADCY8 |
| GOT1  |
| CHST4 |
| POLR2G|
| IL4I1 |
| Gene   |
|--------|
| ULK2   |
| PPAP2C |
| HRH1   |
| TIAM1  |
| CDC42  |
| TREH   |
| UGP2   |
| ITCH   |
| POLR3B |
| TCIRG1 |

