IMMUNOLOGY

Self-supervised learning of T cell receptor sequences exposes core properties for T cell membership

Romi Goldner Kabeli, Sarit Zevin†, Avital Abargel, Alona Zilberberg, Sol Efroni*

The T cell receptor (TCR) repertoire is an extraordinarily diverse collection of TCRs essential for maintaining the body's homeostasis and response to threats. In this study, we compiled an extensive dataset of more than 4200 bulk TCR repertoire samples, encompassing 221,176,713 sequences, alongside 6,159,652 single-cell TCR sequences from over 400 samples. From this dataset, we then selected a representative subset of 5 million bulk sequences and 4.2 million single-cell sequences to train two specialized Transformer-based language models for bulk (CVC) and single-cell (scCVC) TCR repertoires, respectively. We show that these models successfully capture TCR core qualities, such as sharing, gene composition, and single-cell properties. These qualities are emergent in the encoded TCR latent space and enable classification into TCR-based qualities such as public sequences. These models demonstrate the potential of Transformer-based language models in TCR downstream applications.

INTRODUCTION

Many of the tasks of the immune system involve T cells (*1*, *2*). T cells kill infected host cells, detect foreign proteins, activate other immune cells, and regulate immunity. The required specific interaction with a wide variety of antigens leads for a need of a large number of T cells, each with its own pattern recognition means (*3*, *4*). This pattern recognition is mediated through the T cell receptor (TCR). The TCR is made of amino acids, and the collection of TCRs makes up the T cell repertoire (*1*, *5*). Most TCRs consist of α and β chains. Each TCR is antigen relevant, and the interaction is dominated by the third complementarity-determining region (CDR3) of the α and β chains. The CDR3 sequence itself, averaging 16 amino acids in length, is generated by the extensively studied V(D)J recombination, involving a semi-random rearrangement of multiple V, (in β) D, and J gene segments (*1*, *5*–*7*). The studied sequences are obtained from either RNA or DNA using either bulk sequencing or, more recently, single-cell sequencing technologies. While bulk RNA sequencing currently allows for the processing of a larger population of cells, single-cell technologies offer higher resolutions and the tandem exploration of α and β, enabling the exploration of cell-specific characteristics (*8*).

CDR3 sequences were long thought to be unique to each individual, referred to as "private" sequences, but over the past two decades, and especially since high-throughput sequencing has become available, it has been shown that many CDR3 sequences are shared between individuals. These sequences are called "public" sequences (*9*–*11*). The vast potential diversity of CDR3 sequences, estimated at approximately 10^{18} unique combinations (*12*), might intuitively suggest that the occurrence of public sequences would be statistically rare. However, closer examination reveals that such sequences are actually a predictable consequence of the mechanisms driving TCR diversity (*13*). The identification of public and private sequences within the CDR3 region may offer insights into the molecular underpinnings of TCR usage patterns and their distribution across individuals.

The Mina & Everard Goodman Faculty of Life Sciences, Bar-Ilan University, Ramat-Gan, Israel.
*Corresponding author. Email: sol.efroni@biu.ac.il
†Present address: LeddarTech Inc., Québec, Canada.

Progress in computing power and computational tools greatly improved the ability to analyze and research sequential data. This is evident in natural language processing (NLP) in general and, within the scope of this work, in the use of language models (specifically Transformers), to study sequential data such as DNA and proteins, leading to promising results (*3*, *14*–*18*). The language used to study these types of sequences is that of either nucleotides, with their 4-letter representation, or amino acids and their 20-letter representation. In this context, two types of Transformers—encoders and decoders—are of interest. Encoder-based models aim at producing meaningful embeddings out of their inputs, while decoder-based models are used mainly for generation. BERT is an encoder-based Transformer that has been shown to be effective with sequential data, such as DNA (*17*) and proteins (*19*). Regardless of the task it is used for, BERT trains unsupervised to learn the grammatical structure of large, unlabeled datasets.

Since CDR3 sequences are assembled from amino acids, with their function highly dependent on the specific order of these acids (*1*), we posit that a language model—a sequential model—might yield meaningful embeddings to analyze CDR3 sequence features (*20*). This study reveals that the prevalence of a sequence as public or private can be discerned through embeddings encoded by the Transformer, reflecting intrinsic sequence information. In addition, these embeddings facilitate the investigation of "sister" TCRβ sequences that couple with identical TCRα in single-cell data.

Language models have previously predicted TCR specificity (*21*), and various methods have classified private and public sequences (*13, 22*), with tools grouping sequences by editing distance (*23, 24*). Our encoder-based Transformer, CountVonCount (CVC), is trained on 5 million unique CDR3 TCRβ amino acid sequences—half public and half private. CVC stands out for its robust embeddings that enable unsupervised clustering, phenotypic feature delineation, and diverse classification tasks. Notably, when benchmarked against TCR-BERT and ESM-2, CVC's embeddings show superior performance in clustering and classification, highlighting its potential for advanced TCR sequence research.

While the CVC model provides unique insights into public versus private status, across thousands of samples, it lacks awareness of TCR α-β pairing at the single-cell level. To address this limitation, we leveraged a large dataset of over 2 million single T cells, which

together provided a total of 4.2 million TCRα and TCRβ sequences. Using these rich single-cell data, we developed scCVC—a model that enables inspection of TCR features within individual T cells. The scCVC Transformer models the co-occurrence patterns of TCRα and TCRβ chains, providing a more nuanced understanding of TCR presentation. Moreover, the model provides insight into mucosal-associated invariant T (MAIT) cell status and the role of CDR3 sequences in encoding cell type information.

RESULTS

CVC and scCVC are based on the BERT architecture. CVC was trained by processing CDR3 TCRβ amino acid sequences as input, while scCVC was trained on the combined CDR3 TCRα and TCRβ sequences, according to their linked single-cell association. scCVC's input is in the form of single cells, represented by their TCR (α and β) joined by a separator token, enabling a more comprehensive analysis of TCR behavior and features. Each amino acid would be a word in the original BERT architecture, while the CDR3 sequences is a would-be sentence. The model processes each input and outputs their embeddings: a 768-dimensional (768D) numerical vector. The data collection used for training CVC includes 1590 TCRβ samples that translate to 91,758,698 unique CDR3 sequences (see Materials and Methods). Of these, 5 million CDR3 TCRβ sequences were randomly selected for CVC's unsupervised training, with a subdivision of 2.5 million private and 2.5 million public sequences, to avoid bias. As for scCVC, a collection of single-cell data was used, including 2,120,565 single cells that total to 4,200,335 TCR sequences.

An unsupervised language model is trained by masking a certain percentage of the input, and it learns by predicting these masked items. In our case, 15% of each sequence's amino acids were masked, and the model predicted the missing information, with feedback. Once the Transformer is trained, we produce TCR embeddings for further analysis. The pipeline visualizations in Fig. 1 (A and B) illustrate how these embeddings are used. The trained model receives amino acid CDR3 sequences to create their embeddings. We visualized the embedding space in 2D using Uniform Manifold Approximation and Projection (UMAP) (*25*). Each point in CVC represents a sequence, while each point in scCVC represents a cell. In the different visualizations, point color is used for the specific feature analyzed.

CVC identifies public sequences in an unsupervised manner

To evaluate whether CVC encodes meaningful, latent, information about a sequence's biology in its embeddings, we fed the Transformer with 1,000,000 randomly sampled sequences to obtain their embeddings. Among the 1,000,000 sequences, 15% were public and 85% were private, keeping the original distribution of these labels across the entire dataset. We then visualized (UMAP) the 150,000 public and 850,000 private sequences. The results are shown in Fig. 2A, where each sequence (each point) is colored according to its public/private label. A sequence is labeled as public when it appears in more than one sample in the original database. Otherwise, it is labeled private. From the visualized embedding space, it is apparent that sequences are clustered into roughly dozen groups (unsupervised), with public sequences clustered at the tips of each group. Later, we discuss the dozen groups.

We examined the distinct behavior of public sequences by evaluating various thresholds used to tag a sequence as public or private. Our analysis with different criteria for classifying public sequences, based on their frequency across samples, provided consistent results, confirming the robustness of public sequence identification. Independently of the chosen threshold, we sought to determine whether the characteristic of publicity—the extent to which a sequence is common in the population of samples—is inherently captured by the Transformer's embeddings. We quantified the appearances of each sequence and analyzed the correlation between publicity and sequence length. Figure 2B (top right inset) illustrates that sequence length distribution aligns with being bell-shaped. We segmented these into percentiles (10, 25, 50, 75, and 90%), correlating to sequence lengths of 13, 14, 15, 16, and 18. Figure 2B displays the publicity distribution, colored by sequence length percentiles, with the *x* axis indicating the count of public appearances and the *y* axis showing the sequence count on a logarithmic scale. Consistent with previous findings (*26*), our findings indicate that sequences frequently found in public repertoires are generally shorter and have distinct characteristics that are less commonly observed in private sequences.

Using information from the distribution, we divided publicity values into 24 bins of different sizes. To demonstrate how the different sequences are encoded by the Transformer, we sampled sequences from each bin, maintaining the ratio of the complete dataset, leading to 1,037,748 sequences. CVC was used to create embeddings from the sequences, exclusively from the sequences, without considering samples or other features. In fig. S1, a UMAP of the embeddings is displayed using a color code showing the size-bin affiliation. The figure shows that the spectrum of publicity is associated with directionality in the embedded space. The more public a sequence is, the further it is from the private ones. Furthermore, in our analysis, we observe approximately a dozen prominent clusters, akin to those identified in previous observations. The clusters are not identical each time, which can be attributed to variations in the sampling process. Each iteration of sampling can introduce slight differences, leading to observable but not exact replications of cluster formations. As an interim summary, we showed that the embeddings created by CVC capture, in an unsupervised manner, biological features that are integral to the CDR3 sequence itself.

Sequence length, convergent recombination, and publicity

As previously shown, different CDR3 sequence lengths display varying degrees of publicity. Our analysis aimed to ascertain whether this variation in publicity is captured within the transformed embeddings' latent space. Figure 2C demonstrates the bell-shaped distribution of sequence lengths, which corresponds to the full dataset distribution. From this, we sampled 1,050,000 sequences, maintaining the proportion across different length percentiles for the embedding process using CVC. Panels D and E of Fig. 2 respectively depict sequence length percentiles and public/private status in the same UMAP space. These two figures show that the embeddings form roughly a dozen clusters, each containing sequences from all percentiles, suggesting a gradient from larger to smaller percentiles as noted in Fig. 2B. Further, when we compare this gradient with the public/private status in Fig. 2E, we find that public sequences predominantly reside within the lower to mid percentiles, whereas private sequences are more common in the higher percentiles. This pattern aligns with the correlation between sequence length and publicity observed in fig. S1, indicating the CVC-created embeddings' sensitivity to sequence length variations.

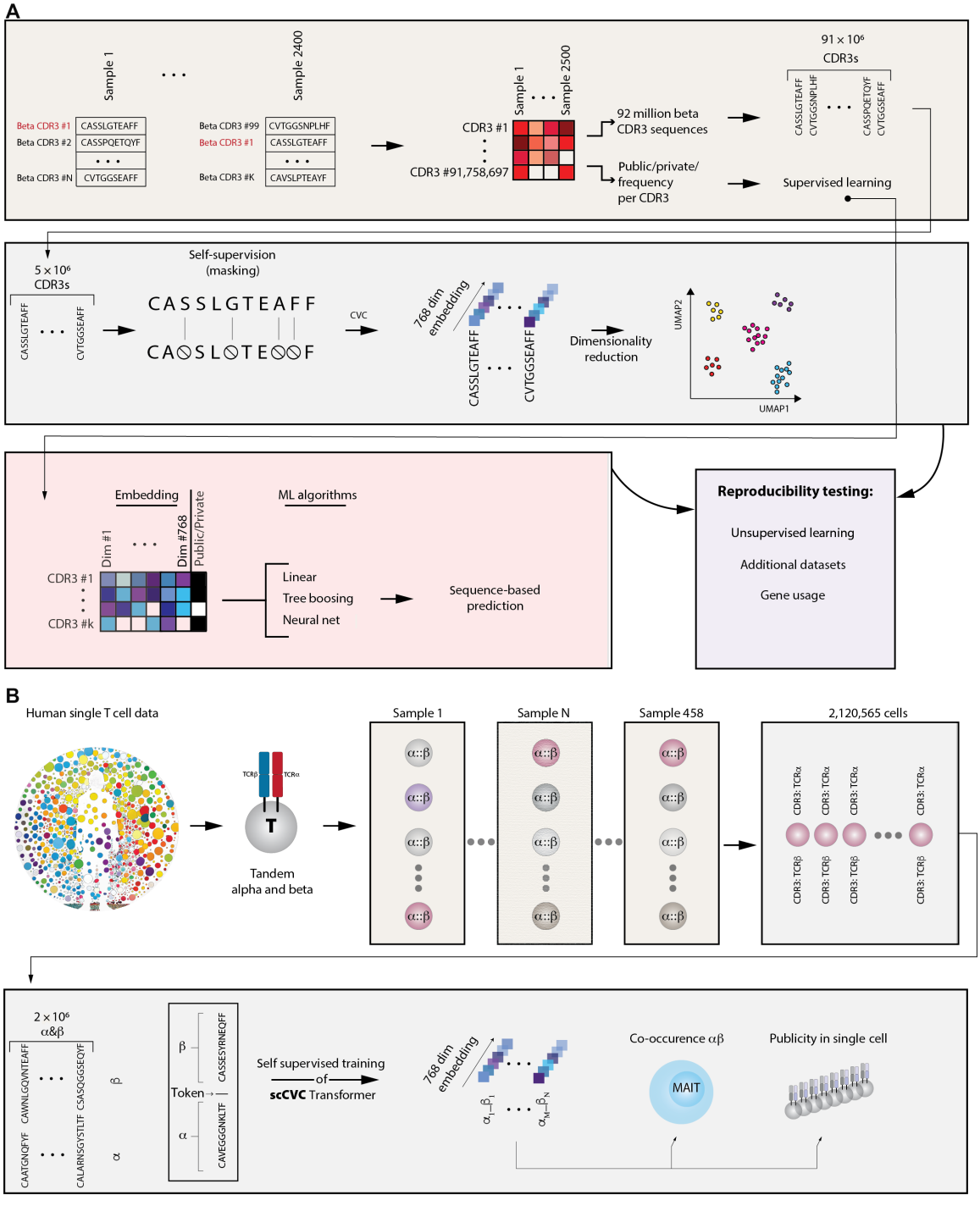

Fig. 1. Using CVC and scCVC to embed TCR sequences. A schematic representation of the processes used for constructing and applying the Transformers for bulk and single-cell TCR repertoires. (**A**) The bulk (CVC) model uses a representative subset of 5 million TCRβ sequences out of over 92 million available for self-supervised learning in the BERT framework, resulting in a 768D embedding for each sequence. (**B**) The single-cell (scCVC) model uses data from over 2 million single T cells, encompassing a curated subset of 4.2 million TCR sequences, with the higher sequence count reflecting instances of cells expressing multiple variants of TCRα and TCRβ chains. Sequences from the same cell are concatenated using a separator token "|," facilitating the Transformer to learn a joint representation, and subsequently producing a 768D embedding for each joint sequence (refer to Materials and Methods for details).

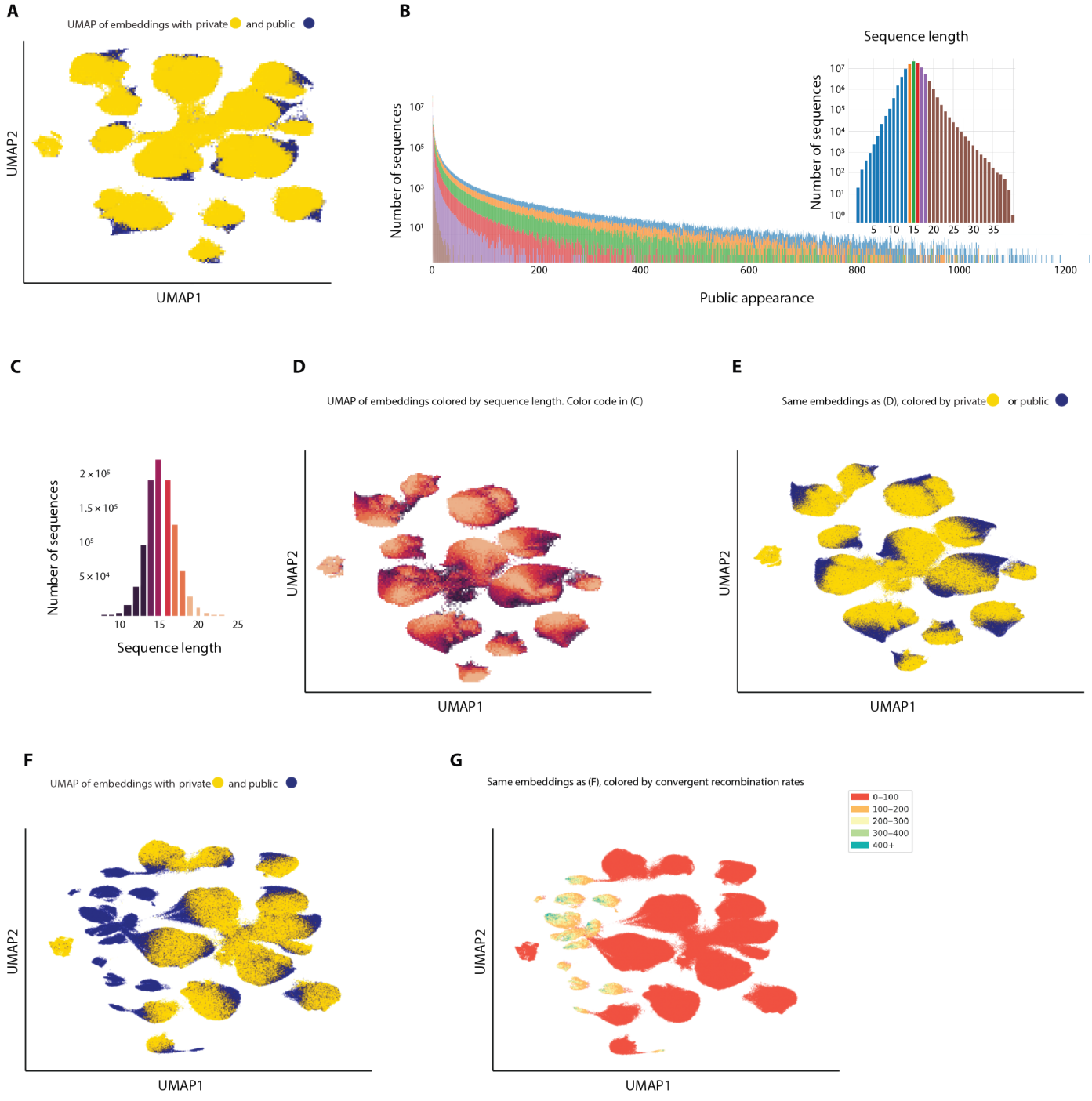

Fig. 2. TCR publicity in CVC space and its association with sequence length and convergent recombination. For all panels in this figure, CVC was used to create the embeddings of CDR3 TCRβ sequences, followed by dimensionality reduction for visualization (using UMAP). (**A**) UMAP of the embeddings of 1,000,000 TCRβ sequences colored according to their public/private label. Yellow points represent private sequences, while blue points represent public sequences. (**B**) public appearance distribution of the sequences in the dataset, colored according to sequence length percentiles, displayed in the upper right corner. The percentiles are 10, 25, 50, 75, and 90% corresponding to lengths 13, 14, 15, 16, and 18. (**C**) Sequence length distribution of 1,050,000 TCRβ sequences colored by sequence length percentiles: 10, 25, 50, 75, and 90%, which corresponded to amino acid length of 13, 14, 15, 16, and18, respectively. For (**D** and **E**), we created embeddings for the sequences used to generate (C). Both UMAP representations display the same latent space for the embeddings, colored initially (D) according to the sequence length percentiles and then (E) according to the private/public label of each sequence, showing the association between sequence length and sequences' sharing status. For (**F** and **G**), we created embeddings for 536,932 TCRβ sequences. Both UMAP representations display the same latent space for the embeddings, colored initially (F) according to public/private status and then (G) according to their convergent recombination ranges. We show five convergent recombination ranges. From each range, we included a set of sequences according to their distribution in the dataset: 0 to 100 with 500,000 sequences, 100 to 200 with 30,799 sequences, 200 to 300 with 4574 sequences, 300 to 400 with 1132 sequences, and 400 and above with 427 sequences. It is easy to see that the Transformer captures publicity and convergent recombination simultaneously in latent space.

Beyond sequence length, public sequences frequently exhibit convergent recombination (CR), where diverse nucleotide sequences encode identical amino acid sequences (*11, 27*), suggesting functional convergence among public sequences across individuals. We categorized sequences into five CR frequency groups and visualized a subsample of 536,932 sequences, revealing a clear correlation: Sequences with higher CR levels are predominantly public (Fig. 2, F and G). This finding supports our embeddings' ability to capture not only sequences' identity but also their immunological relevance. To further explore the relationship between CR and the TCR's structural components, we analyzed CR patterns across different J genes. Figure S5 (A and B) shows the percentage of sequences with CR levels above certain thresholds for each J gene, highlighting that some J genes may be associated with higher rates of CR. This subset correlates with those J genes documented as more prevalent in the general population (*27*). The implication of this correlation may suggest a selective advantage for these J genes in the immune repertoire, contributing to their higher representation and potential public nature in T cell responses.

The CVC produced embeddings space stratifies by J gene affiliation

2D dimensionality reduction of the embedded representation shows an intriguing partition into 12 to 13 large clusters. As a reminder, the embeddings were created unsupervised; that is, CDR3s were not tagged with any labels during self-supervision and were therefore not associated with their origin J gene.

As Fig. 3 (A and B) shows, the J gene region of the TCR gene lies within the CDR3 region and is of 13 types: J1:1 to J1:6 and J2:1 to J2:7 (*28*). To show a substantial amount of J gene tags on a UMAP, we used the ImmuneCODE database (*29*), which includes millions of TCR sequences from more than 1400 individuals, with high-quality information about the V and J gene sources of each CDR3 sequence. We randomly selected 7 million sequences. The distribution of the J genes is shown in Fig. 3C, with TCRBJ02 to TCRBJ04 and TCRBJ02 to TCRBJ06 showing the lowest frequency in the dataset, while the rest of the J genes differ slightly in their frequency. To level the representation, we downsampled to 9% of the sequences from each of the J genes except for TCRBJ02 to TCRBJ04 and TCRBJ02 to TCRBJ06, for which all available sequences have been used.

Given that J gene–associated clustering has been observed in the embedding space, we aimed to evaluate the reproducibility of this phenomenon using an additional dataset. We used the aforementioned dataset, which encompasses data not previously analyzed in our work. The UMAP visualizations, annotated by public/private labels, support our initial findings as demonstrated in Fig. 3D. This consistency validates the patterns we observed with our baseline dataset. To further explore whether the spatial stratification in the embedding space is related to specific J genes, we applied the CVC model to the sequences and reduced their dimensionality using UMAP, coloring each point to correspond with its J gene. Results are shown in Fig. 3E. The apparent color coding of the different clusters reveals that the embedding space stratifies CDR3 sequences according to their J genes. This influence likely stems from the fact that the J segment constitutes a substantial portion of the sequence, which could explain its notable presence within the clusters.

When contrasting with other related language models like TCR-BERT and ESM-2, distinct clustering patterns emerge. As demonstrated in fig. S7, the J gene–driven stratification is pronounced in CVC's visualization but is less discernible with TCR-BERT and ESM-2 embeddings. This distinction suggests that task-specific Transformer models like CVC are adept at capturing biologically pertinent features, potentially overshadowed in more generalized models. A comprehensive comparison is elaborated in the Supplementary Materials (fig. S7, A to F), reinforcing the specialized capabilities of CVC in TCR sequence analysis.

The fraction of the J segment within each cluster, indicating the extent to which the J segment is represented in the sequences, may teach us more of this behavior. Our fraction plot (fig. S6A) reveals consistent J segment proportions across different J gene types, suggesting uniformity in sequence length (fig. S6B). Supporting this are sequence logos (fig. S6, C to O) that visualize the prevalence of specific motifs at the CDR3 J segment junctions.

The clear importance of J genes in embeddings space led us to query the role of V genes. To do this, we again used the ImmuneCODE dataset, this time focusing on V gene available information. A total of 65 V genes, from TCRBV1 to TCRBV30, were represented in the data. Roughly 2% of sequences from each type were used; their embeddings were calculated and charted in fig. S2A. We created fig. S2B to see whether the V genes are associated with the public status of sequences. The red line in the figure is at the 50% mark, meaning that any bars over that threshold are for V genes with a greater than 50% chance of being public.

On the basis of the V genes of those bars, we generated fig. S2 (C and D), which displays the embedding space with the corresponding V gene and public/private labels. In fig. S2C, all clusters contain all types of V genes in which the sequences are grouped together by the different types. Regarding the publicity of these genes, we see that the same behavior occurs (fig. S2D), but with a larger presence of the public sequences. This combines to show that embeddings also link to show similarities between sequences with the same V gene, which has been demonstrated in similar related research (*30*).

Supervised classification using CVC

With clinical applications aiming to control specific TCR sequences in patients, the use of embeddings to expose sequence-based information that associates a TCR with its population-level quantities may greatly benefit clinical TCR uses. To determine whether these embeddings could be used to tag sequences as public or private, we randomly selected 200,000 sequences, 100,000 from each type (public/private), and produced embedding vectors (768D) through CVC. We then used these data (tabular, 200,000 × 768, label 0/1) for supervised binary classification. We tried multiple classification algorithms and eventually focused on three: LDA, xgBoost, and a deep neural network (DNN) (see details in Materials and Methods, in table S1—DOME Report, and in Fig. 4A for the DNN architecture), which showed areas under the curve (AUCs) (over test set) of 0.89, 0.89, and 0.9, respectively. The models provided an accuracy of 81.5, 80.635, and 81.7%.

To learn about the added information content provided by the transformed model, we used machine learning over a one-hot representation of the CDR3 sequences. In this approach, we represent each amino acid using a 20D binary vector. Each vector with 19 zeros and one is placed at the index of the specific amino acid. To maintain an equal length for all sequences in the dataset, we set all one-hot transformations to be the length of the longest sequence (LS), while shorter sequences were padded with zeros. This led to a

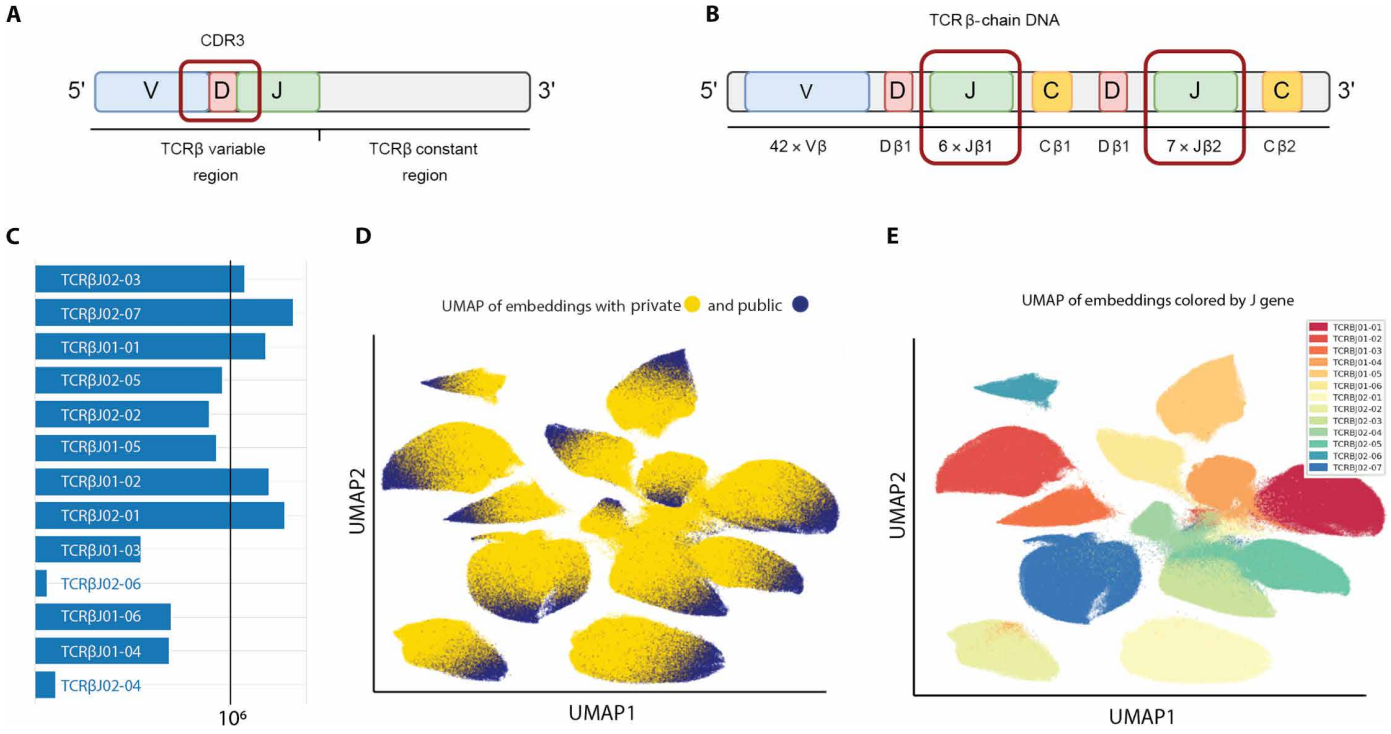

Fig. 3. J gene clustering in embedding space. (**A**) The structure of the CDR3 region of the RNA transcript of a TCRβ chain. (**B**) Structure of the DNA used to produce TCRβ chains before recombination, consisting of the variable (V), joining (J), constant (C), and diversity (D) regions. Segment from each region, together with deletion/addition/replacement of nucleotides, generates the TCR through the process of VDJ recombination. The red marked areas are the J genes, J1:1-6 and J2:1-7. (**C**) Bar plot representation of the number of CDR3 sequences, in our dataset, according to their use of J genes. All the sequences of TCRBJ02-04 and TCRBJ02-06 were taken and 9% of sequences from each of the remaining J gene types were randomly selected to create the represented embedding space and to provide meaningful representations for the visualization of all J genes. We colored the embedding space by the corresponding public/private label of each sequence (**D**) and by the different J gene types (**E**). We can see a near-perfect segmentation of the latent space according to J gene association.

200,000 × LS × 20 table as the algorithm's input. Using these data, we achieved an AUC of 0.76, 0.81, and 0.8, respectively. The accuracy of the models was 69.98, 73.75, and 72.7%. xgBoost did better here, but only by a small margin. The receiver operating characteristic (ROC) curve can be seen below in Fig. 4B. To place our model within the landscape of existing Transformer-related architectures, we performed the same binary classification task using embeddings from the models TCR-BERT and ESM-2. As fig. S8A indicates, CVC showed superior results compared with these two Transformers. These differences demonstrate the importance of the latent space for classifying the sequences as public or private, with a substantial increase in AUC and accuracy when using CVC.

To see whether the embeddings created by CVC could be used to classify a sequence's J gene without previous knowledge of the composition of the TCR sequences, and only the CDR3 representation in embedding space, we used the same set of algorithms used before: xgBoost, LDA, and a modified DNN (Fig. 4A), both on the embeddings and on the one-hot representation of the sequences. Figure 4C displays the accuracies for the DNN, while the other methods appear in fig. S3. All methods did well in predicting the J gene of a sequence when it is represented by the embeddings, but also quite well when the sequences are represented by one-hot encoding. In a similar manner to the comparison previously done with other Transformer models, CVC achieved the highest accuracy, as shown in fig. S8B, surpassing the results of TC-Bert and ESM-2.

Co-occurrence of TCRα and TCRβ and publicity in single-cell data

Single-cell immune profiling provides us with the knowledge of which TCRα and TCRβ chains are expressed in the same cell, allowing exploration of their co-occurrence and possible functional implications. To investigate this, we analyzed two distinct examples: (i) the study of MAIT cells and (ii) the analysis of TRB sister sequences. MAIT cells are a unique type of T cell identifiable by their α chain's specific J and V genes TRAV1-2 joined with TRAJ33/20/12. Using single-cell data, we tagged MAIT cells with this V/J information (available at the data source). Figure 5 (A and B) shows that MAIT cells do not cluster, neither in the single-cell embedding space (scCVC) nor in TCRβ space (CVC). This behavior indicates that unique transcriptional and functional characteristics of MAIT cells are driven primarily by their TCRα. To investigate the publicity of MAIT, we used the TCRβ embeddings at our disposal to classify MAIT cells as public or private according to their TCRβ sequences. We used a DNN classifier like the one described earlier, and as can be seen in Fig. 5C, roughly 60% of the MAIT cells were classified as public. Given the demonstrated success in classifying public sequences with CVC embeddings, and the fact that many MAIT cells were public, we explored whether MAIT cells could be classified as such, using only their TCRβ CDR3 (using CVC embeddings) or only their TCRα CDR3 (scCVC embeddings), without any information about V or J genes. As Fig. 5 (D and E) shows, we were able to

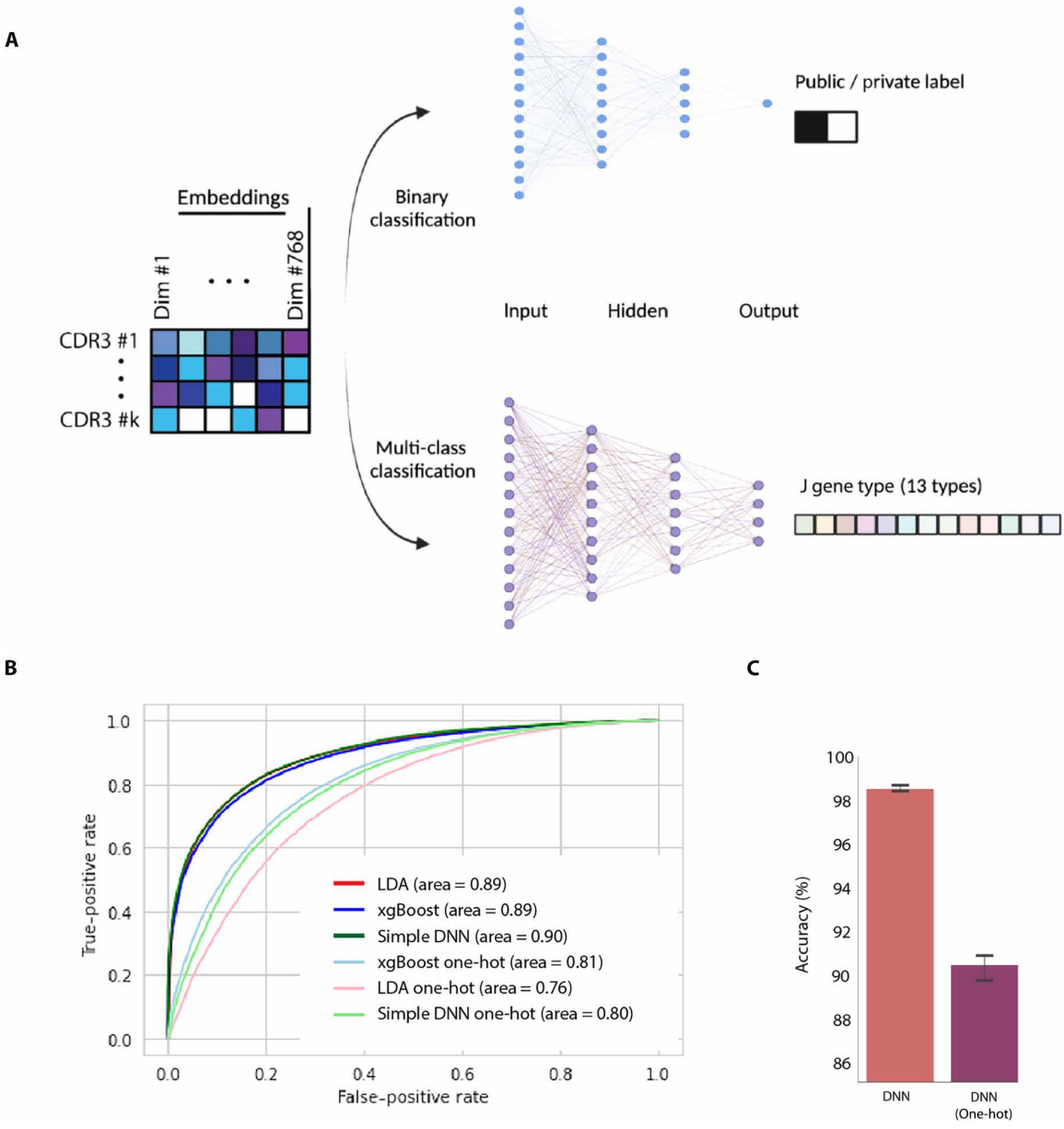

Fig. 4. CVC embeddings for supervised classification tasks. (**A**) We used DNNs, xgBoost, and LDA for the task of binary classification of sequences for their public/private status, and DNN alone for the task of multi-class classification of the J gene of each sequence. In all cases, input is the embeddings of each sequence, produced by CVC. (**B**) ROC of the LDA, xgBoost, and DNN classifiers trained over the task of binary classification of public and private sequences. Each algorithm was applied twice, using the embeddings created by CVC and using one-hot encoding. As shown in the figure, classifiers over embeddings achieved higher scores compared to the one-hot representation: AUC of 0.89, 0.89, and 0.9 compared to 0.76, 0.81, and 0.8, respectively. (**C**) Multi-class classification results of J gene type prediction using DNN on both the embeddings and one-hot vector representation of the sequences. The network was applied three times, and average result accuracies were 98.57% on the embeddings and 90.44% using one-hot encoding. All results are for the test set (previously unseen data); see code for details.

achieve an AUC of 0.71 for β-based classification and an AUC of 0.83 for α-based classification. These results demonstrate that information about the cell type is strongly encoded into the CDR3 sequence, and by translating this sequence into the Transformer-based embeddings, without any gene information, we can effectively classify MAIT cells. The differences in accuracy between the α-based and the β-based classifications are expected, as the tagging itself is α-based. It is unexpected to find that β sequences hold relevant information about the MAIT status of the cell.

In addition to MAIT cells, we used single-cell data to analyze single T cells to identify tenets of co-occurrences between different TCRβ chains and the same TCRα chain in different cells. That is, we studied TCRβ sequences appearing in different cells that share the same TCRα sequences. We refer to these β sequences as TCRβ

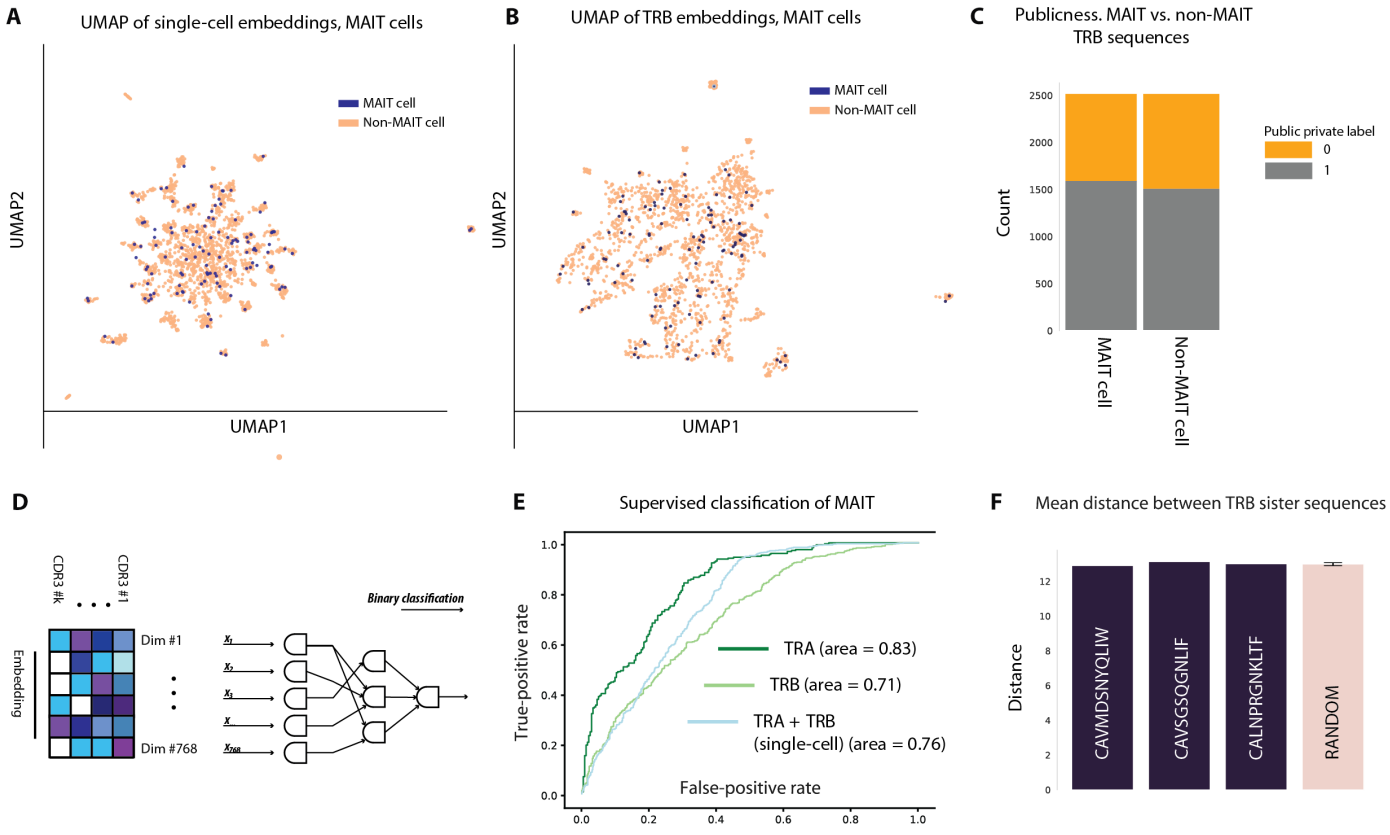

Fig. 5. MAIT cells and TCRβ sister sequences in CVC and scCVC embedding space. The 10x Genomics single-cell lung cancer dataset was used to examine the distribution of MAIT cells in the embedding space. MAIT cell barcodes were labeled according to their TRA J and V genes: TRAV1-2 combined with TRAJ33/20/12, enabling the labeling of corresponding TRB sequences by MAIT barcodes. (**A**) UMAP visualization of MAIT and non-MAIT single-cell embeddings generated using scCVC. (**B**) UMAP visualization of MAIT and non-MAIT TRB sequences produced by CVC. In both cases, we see that the MAIT cells did not cluster together. Eight 10x Genomics single-cell datasets were combined for a more comprehensive analysis. (**C**) Publicity distribution for 2508 MAIT and 2508 non-MAIT cells, revealing that over 60% of MAIT cells are public. (**D**) DNN architecture used for binary classification of MAIT cells, with embeddings as input. (**E**) Results were evaluated using three types of embeddings: TRA only, TRB only, and TRA combined with TRB. ROC AUC values were 0.83, 0.71, and 0.76, respectively. (**F**) Using 100,000 single-cell sequences from the single-cell database (see Data and materials availability), TRB sequences coexpressed with the same TRA sequence, i.e., TRB sister sequences (see text), were grouped together. UMAP visualization of the embedding space for these cells highlights TRB sister sequences belonging to TRA CAVMDSNYQLIW, CAVSGSQGNLIF, and CALNPRGNKLTF (fig. S4, A to C, respectively), showing that they do not cluster together. The mean distance between them was calculated and compared to the distance between them and the rest of the (random) sequences, revealing no difference in mean distance.

sisters. Using single-cell data (see Materials and Methods and Data and materials availability), TCRβ sisters were analyzed, and their embeddings were generated using CVC. To see whether these TCRβ sisters occupy a contained area in embedding space, we measured distances between sister TCRβs and compared these distances with the measured distances between sister TCRβs and random TCRβ sequences. We also projected the TCRβ sequences onto a 2D UMAP plot. As Fig. 5F indicates, the distances within and without the TRB sisters showed no substantial difference. The same phenomenon can be seen in fig. S4 (A to C), which shows TCRβ sisters scattered throughout the embedding space. These results indicate the diversity within sister TCRβ sequences.

DISCUSSION

Our understanding of the immune system's versatility and its ability to perform its multitude of responsibilities is intricately linked to the specificity of TCRs, particularly within the CDR3 region. While TCRBuilder (*31*) and other structural NLP-based modeling tools (*21*, *32*, *33*) clarify parts of TCR functionality, translating the sequence of amino acids directly into functional insights remains challenging. Here, we introduced Transformer language models—CVC, trained on bulk TCRβ sequences, and scCVC, trained on single-cell TCRα and TCRβ presentation in isolated T cells. These models unveil underlying patterns in CDR3 sequences, informing of previously hidden associations.

These models detect spatial separations in latent Transformer space between public and private TCR sequences and manifest self-organized clusters indicative of J gene usage. Intriguingly, we observed certain J genes exhibiting higher CR, hinting at selective pressures in the immune system's evolution that merit further investigation. Moreover, our model is able to classify public and private TCR sequences and to multi-class label J gene types. The utility of such classification tasks is also demonstrated in their ability to identify specialized T cell types, including MAIT cells, showcasing the potential of these tools to parse the complex parts of the T cell phenotype domain with precision.

An additional layer of complexity was addressed by analyzing TCRβ sister sequences using our scCVC model. By examining the co-occurrence of different TCRβ chains within the context of shared TCRα sequences across single T cells, this study follows the diversity and the potential functional interplay between sister chains. Despite the common TCRα linkage, TCRβ sisters demonstrate a high degree of diversity, occupying varied regions within the embedding space. This finding may assist in tracing the recombination mechanisms at the core of these unique cell subsets.

Other language models have been used to explore the TCR. Comparing the results of CVC against one such model, TCR-BERT, which is similarly trained on CDR3 sequences, or ESM-2, which is more generalized, we see that our task-specific Transformer models demonstrated superior capabilities in clustering and in classifying. This shows the contrast between models fine-tuned for specific biological and general models. The clear distinction in performance is a testament to the benefit of CVC's architecture, which is adept at capturing the subtle complexities of TCR specificity.

Future work and future uses build on our ability to scale these models. We believe that such work could increase both parameter numbers and the number of sequences the model would be trained on (*34*). Improvement in single-cell technology may provide computational tools with the data needed to not only better our understanding of immune cell biology but also catalyze the development of innovative T cell–based therapies (*35*–*37*). Insights into public and private TCR distinctions may pave innovative pathways for cancer immunotherapies by identifying public TCRs that target common tumor antigens across patients. This transition from measured biology to latent mathematical space, and back into clinical implementation, has the potential to improve health.

MATERIALS AND METHODS

Bulk sequencing data

The dataset we collected for model training (CVC) includes information from 34 published papers. All included papers report T cell repertoire sequencing from bulk RNA (in contrast with single-cell data). The library preparation has been done using multiple methods, as well as the sequencing itself. All samples are human samples. While some of these papers reported α-chain sequencing, as well as β-chain sequencing, we only included β-chain sequencing in the training dataset. Further, as we were only interested in the sequences themselves, we did not refer in our analyses to any metadata (such as tissue type). These metadata may be the subject of future work.

Out of each of the papers, tables, FASTQ files, or any form of collection was stripped down to two items: TCR sequence and sample identification. These were aggregated to a larger table. The collection finally included 4217 samples that held 221,176,713 sequences.

Single-cell sequencing data

The single-cell dataset we collected for model training (scCVC) and analysis includes information from 31 published experiments. All included papers report T cell repertoire from single-cell RNA sequencing. The library preparation has been done using multiple methods, as well as the sequencing itself. All samples are human samples. As we were mainly interested in the sequences themselves, of both α- and β-chain types, we did not refer in our analyses to additional metadata.

Out of each experiment, tables, FASTQ files, or any form of collection was stripped down to three items: TCR sequence, sample identification, and unique cell identification. These were aggregated to a larger table. The collection finally included 458 samples that held 6,159,652 sequences.

Language models: CVC and scCVC

Both CVC and scCVC are language models based on the BERT model architecture (*38*), a language model that has been shown to have state-of-the-art results on different NLP-related tasks. They were implemented in Python using PyTorch (*39*) and the Transformer libraries (*40*). The models use a mechanism called attention to learn complex interactions within the input sequence, and in our case the interactions and correlations between the amino acids. This allows, after some training, to understand the grammar of the amino acid language in an unsupervised manner.

The difference between the two models is mainly in the number of training samples each model was trained on, the input sequences themselves, and how they were presented during training:

1) CVC was trained on 5 million CDR3 TCRβ sequences, with an internal split of 2.5 private and 2.5 public sequences. The input was individual CDR3 TCRβ sequences taken from the bulk-sequencing data mentioned above. The training was achieved by using the masking technique: 15% of each sequences' amino acids were masked, and the model had to predict them.

2) scCVC was trained on 2,120,565 single cells (consisting of 4,200,335 TCRα and TCRβ sequences) from the previously mentioned single-cell sequencing data. The input consisted of single cells represented by a concatenated representation of the CDR3 that belong to them, joined by a separator token. The training process was achieved by first generating a random permutation of the sequences that constitute the single cells and then using the masking technique: 15% of each sequences' amino acids were masked, and the model had to predict them. The randomization of sequence order was used to ensure that the model did not assign any importance to a particular order.

The hyperparameters that were used were the following, having most kept equal to the default BERT values: hidden representation dimensionality: 768; intermediate representation dimensionality: 1536; number of attention heads: 12; number of Transformer layers: 12; batch size: 1024; training epochs: 50; learning rate: 5×10^{-5}; maximum positional embedding: 64; optimizer: Adam; loss: NLL (negative log likelihood); approximately 86 million parameters.

Because of the large computational needs, the models were trained (separately) on the Google Cloud Platform with the NVIDIA Tesla A100 GPU and 120 GB of memory. With this hardware, it took about 6 days to train. Adding parallelization of eight GPUs decreased the training time to about 2 days.

Once the training was complete, each model was ready to be used for embedding creation. The inputs of CVC were CDR3 TCRβ sequences, and the inputs of scCVC were either individual CDR3 sequences of both chain types or single cells, in the format explained above. The lengths (L) differed. Each sequence was then padded with a prefix token, C, and a suffix token, S. The padded input gets passed to an embedding layer that transposed each amino acid token into a 768D vector. Along with position embeddings, all the embedded tokens were passed into a set of 12 layers that created the whole sequence embedding matrix with dimensions of $(L + 2) \times 768$. This matrix was then reduced to be of dimension 1×768 by calculating the mean of its embeddings. The method for dimensionality reduction could be changed, but the mean was set as the default method.

This final embedding representation could then be used in various downstream tasks like the ones we present below.

Benchmarking—implementation details of protein language models

For the benchmarking process, we used two prominent protein language models: TCR-BERT and ESM-2. The TCR-BERT model is tailored for TCR sequences, while ESM-2 is a generic protein model that has demonstrated broad capabilities in sequence representation.

TCR-BERT: The TCR-BERT model, derived from the original BERT architecture, is designed specifically for TCR sequences. Using the HuggingFace Transformer library, the "wukevin/tcr-bert-mlm-only" version of TCR-BERT includes a base architecture of 12 Transformer layers, each with 12 self-attention heads, and produces embeddings of 768D. The model was trained for 50 epochs with a learning rate of 5×10^{-5} and comprises approximately 58 million parameters, making it particularly suited for analyzing TCRβ sequences and ensuring its relevance for comparison with our CVC model.

ESM-2: We chose the 150 million parameter variants of ESM-2 models for its scale compatibility with other models in parameter count and embedding dimensions. The "facebook/esm2_t30_150M_UR50D" model, also implemented via the HuggingFace Transformer library, features 30 Transformer layers and 20 attention heads, and creates 640D embeddings. This structure provides a balance between efficiency and the ability to capture complex sequence information. It was trained over 500,000 epochs with a learning rate of 4×10^{-4} and a weight decay of 0.01.

A benchmarking set comprising 400,000 TCRβ sequences was extracted from our extensive database to evaluate the performance of these models. The sequences were processed through TCR-BERT and ESM-2 to generate embeddings, which were then benchmarked against those produced by our CVC model.

The benchmarking analysis was conducted under uniform computational settings to ensure equitable conditions, and both models followed identical preprocessing protocols to maintain consistency in the evaluation. Results are presented in figs. S7 and S8.

Downstream clustering (UMAP)

CVC outputs embeddings with dimension of 768. To view these high-dimensional embeddings on a 2D plot, dimensionality reduction had to take place. UMAP (*25*) was used in this case, after the application of principal components analysis (*41*). The scanpy package (*42*) was used to apply this technique with the receiving of an AnnData object consisting of the embeddings and dimensionality reduction coordinates. We also tried t-distributed Stochastic Neighbor Embedding (t-SNE), but results were clearer and faster using UMAP.

Classification models

According to the guidelines of the recent DOME standard (*43*) for reporting results of supervised machine learning, we included, in addition to the description here in Materials and Methods, a supplementary table that carefully follows the DOME standard. This can be found in table S1—DOME Report.

Input data presentation

For each of the models described below, the input was either the embeddings created by CVC or the one-hot encoding representation of the sequences. The one-hot encoding representation transformed each amino acid to a 1×20D one-hot vector.

LDA

The LDA algorithm is a supervised dimensionality reduction technique that was used here to classify both the public/private label and the J gene of a given sequence. The python package that was used to apply this algorithm was sklearn (*44*). It was used with its default hyperparameters. Hyperparameter tuning did not improve results.

xgBoost

The xgBoost algorithm is a well-known classification algorithm that gives high-accuracy results when applied on tabular data. Here, we used it in a supervised manner to classify both the public/private label and the J gene of a given sequence. The sklearn package was again used for this algorithm with default hyperparameters. Changing the parameters did not give better results.

Deep neural network

The classification tasks the DNN was applied on were to predict the public/private label, MAIT cell label, and the J gene of a given sequence.

Predicting public/private label and MAIT cells

For this task, the best results were achieved by using a simple three-layer network with dimensions of 128, 32, and 1. The nonlinear function was ReLU for the first two layers and sigmoid for the last, with a learning rate of 1×10^{-5}, the Adam optimizer, and binary cross entropy loss. For public/private classification, a batch size of 1024 and 150 epochs was used, as opposed to a batch size of 256 and 80 epochs for classifying MAIT cells.

Predicting J gene

For this task, the best results were achieved by using a simple three-layer network with dimensions of 64, 32, and 13 (13 types of J genes). The nonlinear function was ReLU, with a learning rate of 1×10^{-5}, the Adam optimizer, batch size of 1024, cross entropy loss, and 80 epochs. Adding dropout and batch normalization did not improve the results.

Supplementary Materials

This PDF file includes:

Figs. S1 to S8
Tables S1 to S3
References

REFERENCES AND NOTES

1. C.-Y. Wang, Y.-X. Fang, G.-H. Chen, H.-J. Jia, S. Zeng, X.-B. He, Y. Feng, S.-J. Li, Q.-W. Jin, W.-Y. Cheng, Z.-Z. Jing, Analysis of the CDR3 length repertoire and the diversity of T cell receptor α and β chains in swine CD4+ and CD8+ T lymphocytes. *Mol. Med. Rep.* **16**, 75–86 (2017).
2. I. Papadopoulou, A.-P. Nguyen, A. Weber, M. R. Martínez, DECODE: A computational pipeline to discover T cell receptor binding rules. *Bioinformatics* **38**, i246–i254 (2022).
3. J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, D. Hassabis, Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
4. J. Reading, K. Foster, K. Joshi, B. Chain, Tracking down tumor-specific T cells. *Cancer Cell* **40**, 351–353 (2022).
5. M. Ou, F. Zheng, X. Zhang, S. Liu, D. Tang, P. Zhu, J. Qiu, Y. Dai, Integrated analysis of B-cell and T-cell receptors by high-throughput sequencing reveals conserved repertoires in IgA nephropathy. *Mol. Med. Rep.* **17**, 7027–7036 (2018).
6. X. Hou, M. Wang, C. Lu, Q. Xie, G. Cui, J. Chen, Y. Du, Y. Dai, H. Diao, Analysis of the repertoire features of TCR beta chain CDR3 in human by high-throughput sequencing. *Cell. Physiol. Biochem.* **39**, 651–667 (2016).

7. R. Arnaout, W. Lee, P. Cahill, T. Honan, T. Sparrow, M. Weiand, C. Nusbaum, K. Rajewsky, S. B. Koralov, High-resolution description of antibody heavy-chain repertoires in humans. *PLOS ONE* **6**, e22365 (2011).
8. R. Bacher, C. Kendziorski, Design and computational analysis of single-cell RNA-sequencing experiments. *Genome Biol.* **17**, 63 (2016).
9. F. Serana, A. Sottini, L. Caimi, B. Palermo, P. G. Natali, P. Nisticò, L. Imberti, Identification of a public CDR3 motif and a biased utilization of T-cell receptor V beta and J beta chains in HLA-A2/Melan-A-specific T-cell clonotypes of melanoma patients. *J. Transl. Med.* **7**, 21 (2009).
10. W. Huisman, L. Hageman, D. A. T. Leboux, A. Khmelevskaya, G. A. Efimov, M. C. J. Roex, D. Amsen, J. H. F. Falkenburg, I. Jedema, Public T-cell receptors (TCRs) revisited by analysis of the magnitude of identical and highly-similar TCRs in virus-specific T-cell repertoires of healthy individuals. *Front. Immunol.* **13**, 851868 (2022).
11. V. Venturi, D. A. Price, D. C. Douek, M. P. Davenport, The molecular basis for public T-cell responses? *Nat. Rev. Immunol.* **8**, 231–238 (2008).
12. V. Greiff, P. Bhat, S. C. Cook, U. Menzel, W. Kang, S. T. Reddy, A bioinformatic framework for immune repertoire diversity profiling enables detection of immunological status. *Genome Med.* **7**, 49 (2015).
13. Y. Elhanati, Z. Sethna, C. G. Callan Jr., T. Mora, A. M. Walczak, Predicting the spectrum of TCR repertoire sharing with a data-driven model of recombination. *Immunol. Rev.* **284**, 167–179 (2018).
14. Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, A. dos Santos Costa, M. Fazel-Zarandi, T. Sercu, S. Candido, A. Rives, Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
15. M. Lennox, N. Robertson, B. Devereux, Deep learning proteins using a triplet-BERT network. *Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.* **2021**, 4341–4347 (2021).
16. M. H. Vu, R. Akbar, P. A. Robert, B. Swiatczak, V. Greiff, G. K. Sandve, D. T. T. Haug, Advancing protein language models with linguistics: A roadmap for improved interpretability. arXiv:2207.00982 [q-bio.QM] (2022).
17. Y. Ji, Z. Zhou, H. Liu, R. V. Davuluri, DNABERT: Pre-trained bidirectional encoder representations from transformers model for DNA-language in genome. *Bioinformatics* **37**, 2112–2120 (2021).
18. A. Weber, A. Pélissier, M. R. Martínez, T cell receptor binding prediction: A machine learning revolution. arXiv:2312.16594 [q-bio.QM] (2023).
19. S. Unsal, H. Atas, M. Albayrak, K. Turhan, A. C. Acar, T. Doğan, Learning functional properties of proteins with language models. *Nat. Mach. Intell.* **4**, 227–245 (2022).
20. K. Davidsen, B. J. Olson, W. S. DeWitt III, J. Feng, E. Harkins, P. Bradley, F. A. Matsen IV, Deep generative models for T cell receptor protein sequences. *eLife* **8**, e46935 (2019).
21. K. Wu, K. E. Yost, B. Daniel, J. A. Belk, Y. Xia, T. Egawa, A. Satpathy, H. Y. Chang, J. Zou, TCR-BERT: Learning the grammar of T-cell receptors for flexible antigen-xbinding analyses. bioRxiv 2021.11.18.469186 [Preprint] (2021). https://doi.org/10.1101/2021.11.18.469186.
22. V. Greiff, C. R. Weber, J. Palme, U. Bodenhofer, E. Miho, U. Menzel, S. T. Reddy, Learning the high-dimensional immunogenomic features that predict public and private antibody repertoires. *J. Immunol.* **199**, 2985–2997 (2017).
23. S. Valkiers, M. Van Houcke, K. Laukens, P. Meysman, ClusTCR: A python interface for rapid clustering of large sets of CDR3 sequences with unknown antigen specificity. *Bioinformatics* **37**, 4865–4867 (2021).
24. W. S. DeWitt, A. Smith, G. Schoch, J. A. Hansen, F. A. Matsen IV, P. Bradley, Human T cell receptor occurrence patterns encode immune history, genetic background, and receptor specificity. *eLife* **7**, e38358 (2018).
25. L. McInnes, J. Healy, J. Melville, UMAP: Uniform manifold approximation and projection for dimension reduction. arXiv:1802.03426 [stat.ML] (2018).
26. X. Hou, P. Zeng, X. Zhang, J. Chen, Y. Liang, J. Yang, Y. Yang, X. Liu, H. Diao, Shorter TCR β-chains are highly enriched during thymic selection and antigen-driven selection. *Front. Immunol.* **10**, 299 (2019).
27. J. D. Freeman, R. L. Warren, J. R. Webb, B. H. Nelson, R. A. Holt, Profiling the T-cell receptor beta-chain repertoire by massively parallel sequencing. *Genome Res.* **19**, 1817–1824 (2009).
28. L. M.-P. Lefranc, G. Lefranc, *The T cell receptor FactsBook,* (Academic Press, London, 2001), pp. 398, IMGT/LIGMDB: IMGT000021 (582960 bp), human (*Homo sapiens*) TRB locus.
29. S. Nolan, M. Vignali, M. Klinger, J. N. Dines, I. M. Kaplan, E. Svejnoha, T. Craft, K. Boland, M. Pesesky, R. M. Gittelman, T. M. Snyder, C. J. Gooley, S. Semprini, C. Cerchione, M. Mazza, O. M. Delmonte, K. Dobbs, G. Carreño-Tarragona, S. Barrio, V. Sambri, G. Martinelli, J. D. Goldman, J. R. Heath, L. D. Notarangelo, J. M. Carlson, J. Martinez-Lopez, H. S. Robins, A large-scale database of T-cell receptor beta (TCRβ) sequences and binding associations from natural and synthetic exposure to SARS-CoV-2. *Res. Sq.*, (2020).
30. N. Deutchmann, A. Pelissier, A. Weber, S. Gao, J. Bogojeska, M. R. Martínez, Do domain-specific protein language models outperform general models on immunology-related tasks? bioRxiv 2023.10.17.562795 [Preprint] (2023). https://doi.org/10.1101/2023.10.17.562795.
31. W. K. Wong, C. Marks, J. Leem, A. P. Lewis, J. Shi, C. M. Deane, TCRBuilder: Multi-state T-cell receptor structure prediction. *Bioinformatics* **36**, 3580–3581 (2020).
32. J.-W. Sidhom, H. B. Larman, D. M. Pardoll, A. S. Baras, DeepTCR is a deep learning framework for revealing sequence concepts within T-cell repertoires. *Nat. Commun.* **12**, 1605 (2021).
33. M. Ostrovsky-Berman, B. Frankel, P. Polak, G. Yaari, Immune2vec: Embedding B/T cell receptor sequences in $\mathbb{R}^N$ using natural language processing. *Front. Immunol.* **12**, 680687 (2021).
34. J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. de Las Casas, L. A. Hendricks, J. Welbl, A. Clark, T. Hennigan, E. Noland, K. Millican, G. van den Driessche, B. Damoc, A. Guy, S. Osindero, K. Simonyan, E. Elsen, J. W. Rae, O. Vinyals, L. Sifre. Training compute-optimal large language models. arXiv:2203.15556 [cs.CL] (2022).
35. C. Raffin, L. T. Vo, J. A. Bluestone, T_{reg} cell-based therapies: Challenges and perspectives. *Nat. Rev. Immunol.* **20**, 158–172 (2020).
36. M. Romano, G. Fanelli, C. J. Albany, G. Giganti, G. Lombardi, Past, present, and future of regulatory T cell therapy in transplantation and autoimmunity. *Front. Immunol.* **10**, 43 (2019).
37. C. Loretelli, E. Assi, A. J. Seelam, M. B. Nasr, P. Fiorina, Cell therapy for type 1 diabetes. *Expert Opin. Biol. Ther.* **20**, 887–897 (2020).
38. J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805 [cs.CL] (2018).
39. A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, PyTorch: An imperative style, high-performance deep learning library, in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. Buc, E. Fox, R. Garnett, Eds. (Curran Associates Inc., 2019), pp. 8024–8035.
40. T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, A. Rush, Transformers: State-of-the-art natural language processing, in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (Association for Computational Linguistics, Online); https://aclanthology.org/2020.emnlp-demos.6, pp. 38–45.
41. I. T. Jolliffe, J. Cadima, Principal component analysis: A review and recent developments. *Philos. Trans. A Math. Phys. Eng. Sci.* **374**, 20150202 (2016).
42. F. A. Wolf, P. Angerer, F. J. Theis, SCANPY: Large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
43. I. Walsh, D. Fishman, D. Garcia-Gasulla, T. Titma, G. Pollastri; ELIXER Machine Learning Focus Group, J. Harrow, F. E. Psomopoulos, S. C. E. Tosatto, DOME: Recommendations for supervised machine learning validation in biology. *Nat. Methods* **18**, 1122–1127 (2021).
44. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
45. Q. Jia, W. Wu, Y. Wang, P. B. Alexander, C. Sun, Z. Gong, J.-N. Cheng, H. Sun, Y. Guan, X. Xia, L. Yang, X. Yi, Y. Y. Wan, H. Wang, J. He, P. A. Futreal, Q.-J. Li, B. Zhu, Local mutational diversity drives intratumoral immune heterogeneity in non-small cell lung cancer. *Nat. Commun.* **9**, 5361 (2018).
46. G. Napolitani, P. Kurupati, K. W. W. Teng, M. M. Gibani, M. Rei, A. Aulicino, L. Preciado-Llanes, M. T. Wong, E. Becht, L. Howson, P. de Haas, M. Salio, C. J. Blohmke, L. R. Olsen, D. M. S. Pinto, L. Scifo, C. Jones, H. Dobinson, D. Campbell, H. B. Juel, H. Thomaides-Brears, D. Pickard, D. Bumann, S. Baker, G. Dougan, A. Simmons, M. A. Gordon, E. W. Newell, A. J. Pollard, V. Cerundolo, Clonal analysis of Salmonella-specific effector T cells reveals serovar-specific and cross-reactive T cell responses. *Nat. Immunol.* **19**, 742–754 (2018).
47. V. Giudice, X. Feng, Z. Lin, W. Hu, F. Zhang, W. Qiao, M. Del Pilar Fernandez Ibanez, O. Rios, N. S. Young, Deep sequencing and flow cytometric characterization of expanded effector memory $CD8^+CD57^+$ T cells frequently reveals T-cell receptor Vβ oligoclonality and CDR3 homology in acquired aplastic anemia. *Haematologica* **103**, 759–769 (2018).
48. C. S. Seet, C. He, M. T. Bethune, S. Li, B. Chick, E. H. Gschweng, Y. Zhu, K. Kim, D. B. Kohn, D. Baltimore, G. M. Crooks, A. Montel-Hagen, Generation of mature T cells from human hematopoietic stem and progenitor cells in artificial thymic organoids. *Nat. Methods* **14**, 521–530 (2017).
49. J. S. Sims, B. Grinshpun, Y. Feng, T. H. Ung, J. A. Neira, J. L. Samanamud, P. Canoll, Y. Shen, P. A. Sims, J. N. Bruce, Diversity and divergence of the glioma-infiltrating T-cell receptor repertoire. *Proc. Natl. Acad. Sci. U.S.A.* **113**, E3529–E3537 (2016).
50. R. Genolet, B. J. Stevenson, L. Farinelli, M. Osterås, I. F. Luescher, Highly diverse TCRα chain repertoire of pre-immune $CD8^+$ T cells reveals new insights in gene recombination. *EMBO J.* **31**, 4247–4248 (2012).
51. J. T. Neal, X. Li, J. Zhu, V. Giangarra, C. L. Grzeskowiak, J. Ju, I. H. Liu, S.-H. Chiou, A. A. Salahudeen, A. R. Smith, B. C. Deutsch, L. Liao, A. J. Zemek, F. Zhao, K. Karlsson,

L. M. Schultz, T. J. Metzner, L. D. Nadauld, Y.-Y. Tseng, S. Alkhairy, C. Oh, P. Keskula, D. Mendoza-Villanueva, F. M. De La Vega, P. L. Kunz, J. C. Liao, J. T. Leppert, J. B. Sunwoo, C. Sabatti, J. S. Boehm, W. C. Hahn, G. X. Y. Zheng, M. M. Davis, C. J. Kuo, Organoid modeling of the tumor immune microenvironment. *Cell* **175**, 1972–1988.e16 (2018).

52. E. Azizi, A. J. Carr, G. Plitas, A. E. Cornish, C. Konopacki, S. Prabhakaran, J. Nainys, K. Wu, V. Kiseliovas, M. Setty, K. Choi, R. M. Fromme, P. Dao, P. T. McKenney, R. C. Wasti, K. Kadaveru, L. Mazutis, A. Y. Rudensky, D. Pe'er, Single-cell map of diverse immune phenotypes in the breast tumor microenvironment. *Cell* **174**, 1293–1308.e36 (2018).
53. H. van den Heuvel, K. M. Heutinck, E. M. W. van der Meer-Prins, S. L. Yong, P. P. M. C. van Miert, J. D. H. Anholts, M. E. I. F. Dijk, X. Q. Zhang, D. L. Roelen, R. J. M. T. Berge, F. H. J. Claas, Allo-HLA cross-reactivities of Cytomegalovirus-, influenza-, and varicella zoster virus-specific memory T cells are shared by different healthy individuals. *Am. J. Transplant.* **17**, 2033–2044 (2017).
54. V. Béziat, J. Li, J.-X. Lin, C. S. Ma, P. Li, A. Bousfiha, I. Pellier, S. Zoghi, S. Baris, S. Keles, P. Gray, N. Du, Y. Wang, Y. Zerbib, R. Lévy, T. Leclercq, F. About, A. I. Lim, G. Rao, K. Payne, S. J. Pelham, D. T. Avery, E. K. Deenick, B. Pillay, J. Chou, R. Guery, A. Belkadi, A. Guérin, M. Migaud, V. Rattina, F. Ailal, I. Benhsaien, M. Bouaziz, T. Habib, D. Chaussabel, N. Marr, J. El-Benna, B. Grimbacher, O. Wargon, J. Bustamante, B. Boisson, I. Müller-Fleckenstein, B. Fleckenstein, M.-O. Chandesris, M. Titeux, S. Fraitag, M.-A. Alyanakian, M. Leruez-Ville, C. Picard, I. Meyts, J. P. D. Santo, A. Hovnanian, A. Somer, A. Ozen, N. Rezaei, T. A. Chatila, L. Abel, W. J. Leonard, S. G. Tangye, A. Puel, J.-L. Casanova, A recessive form of hyper-IgE syndrome by disruption of ZNF341-dependent STAT3 transcription and activity. *Sci. Immunol.* **3**, eaat4956 (2018).
55. E. P. Mimitou, A. Cheng, A. Montalbano, S. Hao, M. Stoeckius, M. Legut, T. Roush, A. Herrera, E. Papalexi, Z. Ouyang, R. Satija, N. E. Sanjana, S. B. Koralov, P. Smibert, Multiplexed detection of proteins, transcriptomes, clonotypes and CRISPR perturbations in single cells. *Nat. Methods* **16**, 409–412 (2019).
56. M. Buggert, S. Nguyen, G. S.-M. de Oca, B. Bengsch, S. Darko, A. Ransier, E. R. Roberts, D. D. Alcazar, I. B. Brody, L. A. Vella, L. Beura, S. Wijeyesinghe, R. S. Herati, P. M. D. R. Estrada, Y. Ablanedo-Terrazas, L. Kuri-Cervantes, A. S. Japp, S. Manne, S. Vartanian, A. Huffman, J. K. Sandberg, E. Gostick, G. Nadolski, G. Silvestri, D. H. Canaday, D. A. Price, C. Petrovas, L. F. Su, G. Vahedi, Y. Dori, I. Frank, M. G. Itkin, E. J. Wherry, S. G. Deeks, A. Naji, G. Reyes-Terán, D. Masopust, D. C. Douek, M. R. Betts, Identification and characterization of HIV-specific resident memory $CD8^+$ T cells in human lymphoid tissue. *Sci. Immunol.* **3**, eaar4526 (2018).
57. A. de Paula Alves Sousa, K. R. Johnson, J. Ohayon, J. Zhu, P. A. Muraro, S. Jacobson, Comprehensive analysis of TCR-β repertoire in patients with neurological immune-mediated disorders. *Sci. Rep.* **9**, 344 (2019).
58. T. F. Cloughesy, A. Y. Mochizuki, J. R. Orpilla, W. Hugo, A. H. Lee, T. B. Davidson, A. C. Wang, B. M. Ellingson, J. A. Rytlewski, C. M. Sanders, E. S. Kawaguchi, L. Du, G. Li, W. H. Yong, S. C. Gaffey, A. L. Cohen, I. K. Mellinghoff, E. Q. Lee, D. A. Reardon, B. J. O'Brien, N. A. Butowski, P. L. Nghiemphu, J. L. Clarke, I. C. Arrillaga-Romany, H. Colman, T. J. Kaley, J. F. de Groot, L. M. Liau, P. Y. Wen, R. M. Prins, Neoadjuvant anti-PD-1 immunotherapy promotes a survival benefit with intratumoral and systemic immune responses in recurrent glioblastoma. *Nat. Med.* **25**, 477–486 (2019).
59. D. Martino, M. Neeland, T. Dang, J. Cobb, J. Ellis, A. Barnett, M. Tang, P. Vuillermin, K. Allen, R. Saffery, Epigenetic dysregulation of naive CD4+ T-cell activation genes in childhood food allergy. *Nat. Commun.* **9**, 3308 (2018).
60. Y. Kagoya, M. Nakatsugawa, K. Saso, T. Guo, M. Anczurowski, C.-H. Wang, M. O. Butler, C. H. Arrowsmith, N. Hirano, DOT1L inhibition attenuates graft-versus-host disease by allogeneic T cells in adoptive immunotherapy models. *Nat. Commun.* **9**, 1915 (2018).
61. J. Wu, S. Jia, C. Wang, W. Zhang, S. Liu, X. Zeng, H. Mai, X. Yuan, Y. Du, X. Wang, X. Hong, X. Li, F. Wen, X. Xu, J. Pan, C. Li, X. Liu, Minimal residual disease detection and evolved *IGH* clones analysis in acute B lymphoblastic leukemia using *IGH* deep sequencing. *Front. Immunol.* **7**, 403 (2016).
62. W. Zhang, Y. Du, Z. Su, C. Wang, X. Zeng, R. Zhang, X. Hong, C. Nie, J. Wu, H. Cao, X. Xu, X. Liu, IMonitor: A robust pipeline for TCR and BCR repertoire analysis. *Genetics* **201**, 459–472 (2015).
63. K. E. Yost, A. T. Satpathy, D. K. Wells, Y. Qi, C. Wang, R. Kageyama, K. L. McNamara, J. M. Granja, K. Y. Sarin, R. A. Brown, R. K. Gupta, C. Curtis, S. L. Bucktrout, M. M. Davis, A. L. S. Chang, H. Y. Chang, Clonal replacement of tumor-specific T cells following PD-1 blockade. *Nat. Med.* **25**, 1251–1259 (2019).
64. I. Song, A. Gil, R. Mishra, D. Ghersi, L. K. Selin, L. J. Stern, Broad TCR repertoire and diverse structural solutions for recognition of an immunodominant $CD8^+$ T cell epitope. *Nat. Struct. Mol. Biol.* **24**, 395–406 (2017).
65. I. M. Stromnes, A. Hulbert, R. H. Pierce, P. D. Greenberg, S. R. Hingorani, T-cell localization, activation, and clonal expansion in human pancreatic ductal adenocarcinoma. *Cancer Immunol. Res.* **5**, 978–991 (2017).
66. R. Spreafico, M. Rossetti, J. van Loosdregt, C. A. Wallace, M. Massa, S. Magni-Manzoni, M. Gattorno, A. Martini, D. J. Lovell, S. Albani, A circulating reservoir of pathogenic-like $CD4^+$ T cells shares a genetic and phenotypic signature with the inflamed synovial micro-environment. *Ann. Rheum. Dis.* **75**, 459–465 (2016).
67. A. J. Carey, J. L. Hope, Y. M. Mueller, A. J. Fike, O. K. Kumova, D. B. H. van Zessen, E. A. P. Steegers, M. van der Burg, P. D. Katsikis, Public clonotypes and convergent recombination characterize the naïve $CD8^+$ T-cell receptor repertoire of extremely preterm neonates. *Front. Immunol.* **8**, 1859 (2017).
68. M. S. Abdel-Hakeem, M. Boisvert, J. Bruneau, H. Soudeyns, N. H. Shoukry, Selective expansion of high functional avidity memory CD8 T cell clonotypes during hepatitis C virus reinfection and clearance. *PLOS Pathog.* **13**, e1006191 (2017).
69. M. Rossetti, R. Spreafico, A. Consolaro, J. Y. Leong, C. Chua, M. Massa, S. Saidin, S. Magni-Manzoni, T. Arkachaisri, C. A. Wallace, M. Gattorno, A. Martini, D. J. Lovell, S. Albani, TCR repertoire sequencing identifies synovial Treg cell clonotypes in the bloodstream during active inflammation in human arthritis. *Ann. Rheum. Dis.* **76**, 435–441 (2017).
70. Y. Suessmuth, R. Mukherjee, B. Watkins, D. T. Koura, K. Finstermeier, C. Desmarais, L. Stempora, J. T. Horan, A. Langston, M. Qayed, H. J. Khoury, A. Grizzle, J. A. Cheeseman, J. A. Conger, J. Robertson, A. Garrett, A. D. Kirk, E. K. Waller, B. R. Blazar, A. K. Mehta, H. S. Robins, L. S. Kean, CMV reactivation drives posttransplant T-cell reconstitution and results in defects in the underlying TCRβ repertoire. *Blood* **125**, 3835–3850 (2015).
71. M. Hsu, S. Sedighim, T. Wang, J. P. Antonios, R. G. Everson, A. M. Tucker, L. Du, R. Emerson, E. Yusko, C. Sanders, H. S. Robins, W. H. Yong, T. B. Davidson, G. Li, L. M. Liau, R. M. Prins, TCR sequencing can identify and track glioma-infiltrating T cells after DC vaccination. *Cancer Immunol. Res.* **4**, 412–418 (2016).
72. J. F. Beausang, A. J. Wheeler, N. H. Chan, V. R. Hanft, F. M. Dirbas, S. S. Jeffrey, S. R. Quake, T cell receptor sequencing of early-stage breast cancer tumors identifies altered clonal structure of the T cell repertoire. *Proc. Natl. Acad. Sci. U.S.A.* **114**, E10409–E10417 (2017).
73. I. Gomez-Tourino, Y. Kamra, R. Baptista, A. Lorenc, M. Peakman, T cell receptor β-chains display abnormal shortening and repertoire sharing in type 1 diabetes. *Nat. Commun.* **8**, 1792 (2017).
74. C. Keane, C. Gould, K. Jones, D. Hamm, D. Talaulikar, J. Ellis, F. Vari, S. Birch, E. Han, P. Wood, K.-A. Le-Cao, M. R. Green, P. Crooks, S. Jain, J. Tobin, R. J. Steptoe, M. K. Gandhi, The T-cell receptor repertoire influences the tumor microenvironment and is associated with survival in aggressive B-cell lymphoma. *Clin. Cancer Res.* **23**, 1820–1828 (2017).
75. D. B. Page, J. Yuan, D. Redmond, Y. H. Wen, J. C. Durack, R. Emerson, S. Solomon, Z. Dong, P. Wong, C. Comstock, A. Diab, J. Sung, M. Maybody, E. Morris, E. Brogi, M. Morrow, V. Sacchini, O. Elemento, H. Robins, S. Patil, J. P. Allison, J. D. Wolchok, C. Hudis, L. Norton, H. L. McArthur, Deep sequencing of T-cell receptor DNA as a biomarker of clonally expanded TILs in breast cancer after immunotherapy. *Cancer Immunol. Res.* **4**, 835–844 (2016).
76. D. Wu, A. Sherwood, J. R. Fromm, S. S. Winter, K. P. Dunsmore, M. L. Loh, H. A. Greisman, D. E. Sabath, B. L. Wood, H. Robins, High-throughput sequencing detects minimal residual disease in acute T lymphoblastic leukemia. *Sci. Transl. Med.* **4**, 134ra63 (2012).
77. H. R. Seay, E. Yusko, S. J. Rothweiler, L. Zhang, A. L. Posgai, M. Campbell-Thompson, M. Vignali, R. O. Emerson, J. S. Kaddis, D. Ko, M. Nakayama, M. J. Smith, J. C. Cambier, A. Pugliese, M. A. Atkinson, H. S. Robins, T. M. Brusko, Tissue distribution and clonal diversity of the T and B cell repertoire in type 1 diabetes. *JCI Insight* **1**, e88242 (2016).
78. R. O. Emerson, W. S. DeWitt, M. Vignali, J. Gravley, J. K. Hu, E. J. Osborne, C. Desmarais, M. Klinger, C. S. Carlson, J. A. Hansen, M. Rieder, H. S. Robins, Immunosequencing identifies signatures of cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire. *Nat. Genet.* **49**, 659–665 (2017).
79. A. M. Leader, J. A. Grout, B. B. Maier, B. Y. Nabet, M. D. Park, A. Tabachnikova, C. Chang, L. Walker, A. Lansky, J. L. Berichel, L. Troncoso, N. Malissen, M. Davila, J. C. Martin, G. Magri, K. Tuballes, Z. Zhao, F. Petralia, R. Samstein, N. R. D'Amore, G. Thurston, A. O. Kamphorst, A. Wolf, R. Flores, P. Wang, S. Müller, I. Mellman, M. B. Beasley, H. Salmon, A. H. Rahman, T. U. Marron, E. Kenigsberg, M. Merad, Single-cell analysis of human non-small cell lung cancer lesions refines tumor classification and patient stratification. *Cancer Cell* **39**, 1594–1609.e12 (2021).
80. Y. Zhao, C. Kilian, J.-E. Turner, L. Bosurgi, K. Roedl, P. Bartsch, A.-C. Gnirck, F. Cortesi, C. Schultheiß, M. Hellmig, L. U. B. Enk, F. Hausmann, A. Borchers, M. N. Wong, H.-J. Paust, F. Siracusa, N. Scheibel, M. Herrmann, E. Rosati, P. Bacher, D. Kylies, D. Jarczak, M. Lütgehetmann, S. Pfefferle, S. Steurer, J. S. Zur-Wiesch, V. G. Puelles, J.-P. Sperhake, M. M. Addo, A. W. Lohse, M. Binder, S. Huber, T. B. Huber, S. Kluge, S. Bonn, U. Panzer, N. Gagliani, C. F. Krebs, Clonal expansion and activation of tissue-resident memory-like TH17 cells expressing GM-CSF in the lungs of patients with severe COVID-19 patients. *Sci. Immunol.* **6**, eabf6692 (2021).
81. Y. Tang, D. J. Kwiatkowski, E. P. Henske, Midkine expression by stem-like tumor cells drives persistence to mTOR inhibition and an immune-suppressive microenvironment. *Nat. Commun.* **13**, 5018 (2022).
82. K. L. Banta, X. Xu, A. S. Chitre, A. Au-Yeung, C. Takahashi, W. E. O'Gorman, T. D. Wu, S. Mittman, R. Cubas, L. Comps-Agrar, A. Fulzele, E. J. Bennett, J. L. Grogan, E. Hui, E. Y. Chiang, I. Mellman, Mechanistic convergence of the TIGIT and PD-1 inhibitory pathways necessitates co-blockade to optimize anti-tumor $CD8^+$ T cell responses. *Immunity* **55**, 512–526.e9 (2022).

83. K. M. Mahuron, J. M. Moreau, J. E. Glasgow, D. P. Boda, M. L. Pauli, V. Gouirand, L. Panjabi, R. Grewal, J. M. Luber, A. N. Mathur, R. M. Feldman, E. Shifrut, P. Mehta, M. M. Lowe, M. D. Alvarado, A. Marson, M. Singer, J. Wells, R. Jupp, A. I. Daud, M. D. Rosenblum, Layilin augments integrin activation to promote antitumor immunity. *J. Exp. Med.* **217**, e20192080 (2020).
84. M. Liao, Y. Liu, J. Yuan, Y. Wen, G. Xu, J. Zhao, L. Cheng, J. Li, X. Wang, F. Wang, L. Liu, I. Amit, S. Zhang, Z. Zhang, Single-cell landscape of bronchoalveolar immune cells in patients with COVID-19. *Nat. Med.* **26**, 842–844 (2020).
85. J. Biermann, J. C. Melms, A. D. Amin, Y. Wang, L. A. Caprio, A. Karz, S. Tagore, I. Barrera, M. A. Ibarra-Arellano, M. Andreatta, B. T. Fullerton, K. H. Gretarsson, V. Sahu, V. S. Mangipudy, T. T. T. Nguyen, A. Nair, M. Rogava, P. Ho, P. D. Koch, M. Banu, N. Humala, A. Mahajan, Z. H. Walsh, S. B. Shah, D. H. Vaccaro, B. Caldwell, M. Mu, F. Wünnemann, M. Chazotte, S. Berhe, A. M. Luoma, J. Driver, M. Ingham, S. A. Khan, S. Rapisuwon, C. L. Slingluff, T. Eigentler, M. Röcken, R. Carvajal, M. B. Atkins, M. A. Davies, A. Agustinus, S. F. Bakhoum, E. Azizi, M. Siegelin, C. Lu, S. J. Carmona, H. Hibshoosh, A. Ribas, P. Canoll, J. N. Bruce, W. L. Bi, P. Agrawal, D. Schapiro, E. Hernando, E. Z. Macosko, F. Chen, G. K. Schwartz, B. Izar, Dissecting the treatment-naive ecosystem of human melanoma brain metastasis. *Cell* **185**, 2591–2608.e30 (2022).
86. J. X. Caushi, J. Zhang, Z. Ji, A. Vaghasia, B. Zhang, E. H.-C. Hsiue, B. J. Mog, W. Hou, S. Justesen, R. Blosser, A. Tam, V. Anagnostou, T. R. Cottrell, H. Guo, H. Y. Chan, D. Singh, S. Thapa, A. G. Dykema, P. Burman, B. Choudhury, L. Aparicio, L. S. Cheung, M. Lanis, Z. Belcaid, M. E. Asmar, P. B. Illei, R. Wang, J. Meyers, K. Schuebel, A. Gupta, A. Skaist, S. Wheelan, J. Naidoo, K. A. Marrone, M. Brock, J. Ha, E. L. Bush, B. J. Park, M. Bott, D. R. Jones, J. E. Reuss, V. E. Velculescu, J. E. Chaft, K. W. Kinzler, S. Zhou, B. Vogelstein, J. M. Taube, M. D. Hellmann, J. R. Brahmer, T. Merghoub, P. M. Forde, S. Yegnasubramanian, H. Ji, D. M. Pardoll, K. N. Smith, Transcriptional programs of neoantigen-specific TIL in anti-PD-1-treated lung cancers. *Nature* **596**, 126–132 (2021).
87. S. S. Chandran, J. Ma, M. G. Klatt, F. Dündar, C. Bandlamudi, P. Razavi, H. Y. Wen, B. Weigelt, P. Zumbo, S. N. Fu, L. B. Banks, F. Yi, E. Vercher, I. Etxeberria, W. D. Bestman, A. D. C. Paula, I. S. Aricescu, A. Drilon, D. Betel, D. A. Scheinberg, B. M. Baker, C. A. Klebanoff, Immunogenicity and therapeutic targeting of a public neoantigen derived from mutated PIK3CA. *Nat. Med.* **28**, 946–957 (2022).
88. A. M. Luoma, S. Suo, Y. Wang, L. Gunasti, C. B. M. Porter, N. Nabilsi, J. Tadros, A. P. Ferretti, S. Liao, C. Gurer, Y.-H. Chen, S. Criscitiello, C. A. Ricker, D. Dionne, O. Rozenblatt-Rosen, R. Uppaluri, R. I. Haddad, O. Ashenberg, A. Regev, E. M. Van Allen, G. MacBeath, J. D. Schoenfeld, K. W. Wucherpfennig, Tissue-resident memory and circulating T cells are early responders to pre-surgical cancer immunotherapy. *Cell* **185**, 2918–2935.e29 (2022).
89. Y. Zheng, Z. Chen, Y. Han, L. Han, X. Zou, B. Zhou, R. Hu, J. Hao, S. Bai, H. Xiao, W. V. Li, A. Bueker, Y. Ma, G. Xie, J. Yang, S. Chen, H. Li, J. Cao, L. Shen, Immune suppressive landscape in the human esophageal squamous cell carcinoma microenvironment. *Nat. Commun.* **11**, 6268 (2020).
90. L. Han, S. Chen, Z. Chen, B. Zhou, Y. Zheng, L. Shen, Interleukin 32 promotes Foxp3$^+$ Treg cell development and CD8$^+$ T cell function in human esophageal squamous cell carcinoma microenvironment. *Front. Cell Dev. Biol.* **9**, 704853 (2021).
91. C. M. Anadon, X. Yu, K. Hänggi, S. Biswas, R. A. Chaurio, A. Martin, K. K. Payne, G. Mandal, P. Innamarato, C. M. Harro, J. A. Mine, K. B. Sprenger, C. Cortina, J. J. Powers, T. L. Costich, B. A. Perez, C. D. Gatenbee, S. Prabhakaran, D. Marchion, M. H. M. Heemskerk, T. J. Curiel, A. R. Anderson, R. M. Wenham, P. C. Rodriguez, J. R. Conejo-Garcia, Ovarian cancer immunogenicity is governed by a narrow subset of progenitor tissue-resident memory T cells. *Cancer Cell* **40**, 545–557.e13 (2022).
92. C. M. Anadon, C. Zhang, X. Wang, L. Cen, J. R. Conejo-Garcia, X. Yu, Protocol for the isolation of CD8$^+$ tumor-infiltrating lymphocytes from human tumors and their characterization by single-cell immune profiling and multiome. *Star Protoc.* **3**, 101649 (2022).
93. M. Heming, X. Li, S. Räuber, A. K. Mausberg, A.-L. Börsch, M. Hartlehnert, A. Singhal, I.-N. Lu, M. Fleischer, F. Szepanowski, O. Witzke, T. Brenner, U. Dittmer, N. Yosef, C. Kleinschnitz, H. Wiendl, M. Stettner, G. M. Zu Hörste, Neurological manifestations of COVID-19 feature T cell exhaustion and dedifferentiated monocytes in cerebrospinal fluid. *Immunity* **54**, 164–175.e6 (2021).
94. P. Gueguen, C. Metoikidou, T. Dupic, M. Lawand, C. Goudot, S. Baulande, S. Lameiras, O. Lantz, N. Girard, A. Seguin-Givelet, M. Lefevre, T. Mora, A. M. Walczak, J. J. Waterfall, S. Amigorena, Contribution of resident and circulating precursors to tumor-infiltrating CD8$^+$ T cell populations in lung cancer. *Sci. Immunol.* **6**, eabd5778 (2021).
95. N. Kourtis, Q. Wang, B. Wang, E. Oswald, C. Adler, S. Cherravuru, E. Malahias, L. Zhang, J. Golubov, Q. Wei, S. Lemus, M. Ni, Y. Ding, Y. Wei, G. S. Atwal, G. Thurston, L. E. Macdonald, A. J. Murphy, A. Dhanik, M. A. Sleeman, S. S. Tykodi, D. Skokos, A single-cell map of dynamic chromatin landscapes of immune cells in renal cell carcinoma. *Nat Cancer* **3**, 885–898 (2022).
96. Z. Wang, L. Xie, G. Ding, S. Song, L. Chen, G. Li, M. Xia, D. Han, Y. Zheng, J. Liu, T. Xiao, H. Zhang, Y. Huang, Y. Li, M. Huang, Single-cell RNA sequencing of peripheral blood mononuclear cells from acute Kawasaki disease patients. *Nat. Commun.* **12**, 5444 (2021).
97. X. Shi, Z. Li, R. Yao, Q. Cheng, W. Li, R. Wu, Z. Xie, Y. Zhu, X. Qiu, S. Yang, T. Zhou, J. Hu, Y. Zhang, T. Wu, Y. Zhao, Y. Zhang, J. Wu, H. Wang, X. Jiang, L. Chen, Single-cell atlas of diverse immune populations in the advanced biliary tract cancer microenvironment. *Npj Precis. Oncol.* **6**, 58 (2022).
98. M. Ferreira-Gomes, A. Kruglov, P. Durek, F. Heinrich, C. Tizian, G. A. Heinz, A. Pascual-Reguant, W. Du, R. Mothes, C. Fan, S. Frischbutter, K. Habenicht, L. Budzinski, J. Ninnemann, P. K. Jani, G. M. Guerra, K. Lehmann, M. Matz, L. Ostendorf, L. Heiberger, H.-D. Chang, S. Bauherr, M. Maurer, G. Schönrich, M. Raftery, T. Kallinich, M. A. Mall, S. Angermair, S. Treskatsch, T. Dörner, V. M. Corman, A. Diefenbach, H.-D. Volk, S. Elezkurtaj, T. H. Winkler, J. Dong, A. E. Hauser, H. Radbruch, M. Witkowski, F. Melchers, A. Radbruch, M.-F. Mashreghi, SARS-CoV-2 in severe COVID-19 induces a TGF-β-dominated chronic immune response that does not target itself. *Nat. Commun.* **12**, 1961 (2021).
99. A. Ramaswamy, N. N. Brodsky, T. S. Sumida, M. Comi, H. Asashima, K. B. Hoehn, N. Li, Y. Liu, A. Shah, N. G. Ravindra, J. Bishai, A. Khan, W. Lau, B. Sellers, N. Bansal, P. Guerrerio, A. Unterman, V. Habet, A. J. Rice, J. Catanzaro, H. Chandnani, M. Lopez, N. Kaminski, C. S. D. Cruz, J. S. Tsang, Z. Wang, X. Yan, S. H. Kleinstein, D. van Dijk, R. W. Pierce, D. A. Hafler, C. L. Lucas, Immune dysregulation and autoreactivity correlate with disease severity in SARS-CoV-2-associated multisystem inflammatory syndrome in children. *Immunity* **54**, 1083–1095.e7 (2021).
100. A. M. Gaydosik, C. J. Stonesifer, A. E. Khaleel, L. J. Geskin, P. Fuschiotti, Single-cell RNA sequencing unveils the clonal and transcriptional landscape of cutaneous T-cell lymphomas. *Clin. Cancer Res.* **28**, 2610–2622 (2022).
101. C. S. Eberhardt, H. T. Kissick, M. R. Patel, M. A. Cardenas, N. Prokhnevska, R. C. Obeng, T. H. Nasti, C. C. Griffith, S. J. Im, X. Wang, D. M. Shin, M. Carrington, Z. G. Chen, J. Sidney, A. Sette, N. F. Saba, A. Wieland, R. Ahmed, Functional HPV-specific PD-1$^+$ stem-like CD8 T cells in head and neck cancer. *Nature* **597**, 279–284 (2021).
102. D. Corridoni, A. Antanaviciute, T. Gupta, D. Fawkner-Corbett, A. Aulicino, M. Jagielowicz, K. Parikh, E. Repapi, S. Taylor, D. Ishikawa, R. Hatano, T. Yamada, W. Xin, H. Slawinski, R. Bowden, G. Napolitani, O. Brain, C. Morimoto, H. Koohy, A. Simmons, Single-cell atlas of colonic CD8$^+$ T cells in ulcerative colitis. *Nat. Med.* **26**, 1480–1490 (2020).
103. S. Gao, Z. Wu, B. Arnold, C. Diamond, S. Batchu, V. Giudice, L. Alemu, D. Q. Raffo, X. Feng, S. Kajigaya, J. Barrett, S. Ito, N. S. Young, Single-cell RNA sequencing coupled to TCR profiling of large granular lymphocyte leukemia T cells. *Nat. Commun.* **13**, 1982 (2022).
104. S. Saluzzo, R. V. Pandey, L. M. Gail, R. Dingelmaier-Hovorka, L. Kleissl, L. Shaw, B. Reininger, D. Atzmüller, J. Strobl, V. Touzeau-Römer, A. Beer, C. Staud, A. Rieger, M. Farlik, W. Weninger, G. Stingl, G. Stary, Delayed antiretroviral therapy in HIV-infected individuals leads to irreversible depletion of skin- and mucosa-resident memory T cells. *Immunity* **54**, 2842–2858.e5 (2021).
105. Y. Hu, G. Cao, X. Chen, X. Huang, N. Asby, N. Ankenbruck, A. Rahman, A. Thusu, Y. He, P. A. Riedell, M. R. Bishop, H. Schreiber, J. P. Kline, J. Huang, Antigen multimers: Specific, sensitive, precise, and multifunctional high-avidity CAR-staining reagents. *Matter* **4**, 3917–3940 (2021).
106. N. Borcherding, A. Vishwakarma, A. P. Voigt, A. Bellizzi, J. Kaplan, K. Nepple, A. K. Salem, R. W. Jenkins, Y. Zakharia, W. Zhang, Mapping the immune environment in clear cell renal carcinoma by single-cell genomics. *Commun. Biol.* **4**, 122 (2021).
107. I. S. Cheon, C. Li, Y. M. Son, N. P. Goplen, Y. Wu, T. Cassmann, Z. Wang, X. Wei, J. Tang, Y. Li, H. Marlow, S. Hughes, L. Hammel, T. M. Cox, E. Goddery, K. Ayasoufi, D. Weiskopf, J. Boonyaratanakornkit, H. Dong, H. Li, R. Chakraborty, A. J. Johnson, E. Edell, J. J. Taylor, M. H. Kaplan, A. Sette, B. J. Bartholmai, R. Kern, R. Vassallo, J. Sun, Immune signatures underlying post-acute COVID-19 lung sequelae. *Sci. Immunol.* **6**, eabk1741 (2021).

Acknowledgments: We thank T. Mora and G. Yaari for extremely useful comments. **Funding:** This work was supported by an ISF funded project 682/19, an ISF-SFC project 3382/20, an ICRF project 829965, a BSF project 20199090, and by the VATAT Data Science Program. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of ISF, ICRF, BSF, and VATAT. **Author contributions:** Conceptualization: R.G.K. and S.E. Methodology: R.G.K. and S.E. Investigation: R.G.K. A.A., S.Z., and A.Z. Visualization: R.G.K. and S.E. Funding acquisition: A.Z. and S.E. Project administration: R.G.K. Supervision: S.E. Writing—original draft: R.G.K. Writing—review and editing: R.G.K. and S.E. **Competing interests:** A.Z. and S.E. are founders of Clonal Company. S.E. and R.G.K. are inventors on patent application no. IL2023/050758 submitted by Bar-Ilan University that covers the method for TCR sequence identification and classification. The other authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. Bulk sequencing database: This database was created in our laboratory. It is a collection of public data of 4219 samples that correspond to 221,176,713 rows. The list of the PMIDs of the samples that make up this database can be found in table S2. For this project, we used only the data of TCRβ sequences, which translate to 91,758,697 unique sequences. Single-cell sequencing database: This database was created in our laboratory. It is a collection of public data of 458 samples that correspond to 6,159,652 rows. The list of the PMIDs of the samples that make up this database can be found in table S3. For this project, we filtered out duplicate and low-quality sequences,

which left us with 4,200,335 TCRα and TCRβ sequences. These translate to 2,120,565 cells. ImmuneCODE database: This database includes millions of TCR sequences that come from patients who were exposed to or infected with SARS-CoV-2. It includes over 1400 different subjects. In this research, it was specifically used to distinguish the embedding space with the V and J genes. To do so, 17 million sequences were randomly extracted from it and used for both tasks. This database is freely available (*29*) and is planned to be used in further research. 10x Genomics dataset: 10x Genomics offers many different single-cell datasets that can be used for different research investigations. Overall, we used six datasets in this research: NSCLC tumor dataset, 20,000 bone marrow mononuclear cells, PBMCs of a healthy donor, 10,000 human PBMCs (https://www.10xgenomics.com/resources/datasets/10-k-human-pbm-cs-5-v-2-0-chromium-x-2-standard-6-1-0), $CD8^+$ T cells of healthy donor 1, and $CD8^+$ of healthy donor 2. These were chosen on the basis of the number of cells they contained and not for any specific reason. The NSCLC tumor datasets, which were used for immune profiling, consist of about 3643 cells. This and the rest of the datasets, 20,000 bone marrow mononuclear cells, PBMCs, and $CD8^+$ T cells, were used for MAIT cell classification. These datasets contain 19,737, 6037, 14,632, 123,862, and 191,643 cells, respectively. More information and the datasets themselves can be found in the 10x Genomics website.

Submitted 23 August 2023
Accepted 26 March 2024
Published 26 April 2024
10.1126/sciadv.adk4670