Determining epitope specificity of T cell receptors with TCRGP

Emmi Jokinen [1,*], Jani Huuhtanen [2,3], Satu Mustjoki [2,3],
Markus Heinonen [1,4] and Harri Lähdesmäki [1,*]

[1]Department of Computer Science, Aalto University, Espoo, 02150, Finland.
[2]Translational Immunology Research program and Department of Clinical Chemistry and Hematology, University of Helsinki, Helsinki, 00014, Finland.
[3]Hematology Research Unit Helsinki, Helsinki University Hospital Comprehensive Cancer Center, Helsinki, 00014, Finland.
[4]Helsinki Institute for Information Technology, Espoo, 02150, Finland.

[*] To whom correspondence should be addressed.

T cell receptors (TCRs) can recognize various pathogens and consequently start immune responses. TCRs can be sequenced from individuals and methods analyzing the specificity of the TCRs can help us better understand individuals' immune status in different diseases. We have developed TCRGP, a novel Gaussian process method to predict if TCRs recognize certain epitopes. This method can utilize CDR sequences from TCRα and TCRβ chains and learn which CDRs are important in recognizing different epitopes. We have experimented with with epitope-specific data against 29 epitopes and performed a comprehensive evaluation with existing prediction methods. On this data, TCRGP outperforms other state-of-the-art methods in epitope-specificity predictions. We also propose a novel analysis approach for combined single-cell RNA and TCR$\alpha\beta$ (scRNA+TCR$\alpha\beta$) sequencing data by quantifying epitope-specific TCRs with TCRGP in phenotypes identified from scRNA-seq data. With this approach, we find HBV-epitope specific T cells and their transcriptomic states in hepatocellular carcinoma patients.

Introduction

The adaptive immune system implements various complex mechanisms for surveillance against both pathogens and pathological cells arising in our body. To initiate an adequate adaptive immune response, a peptide, called epitope must first be bound by the major histocompatibility complex (MHC) class I or II molecule expressed on the surface of a nucleated cell or a professional antigen-presenting cell. The peptide-MHC (pMHC) complex is then presented to T cells which can recognize the complex via the T cell receptor (TCR) protein, consequently leading to T cell activation and proliferation by clonal expansion[1]. During clonal expansion, a fraction of T cells gain a long-living memory phenotype and therefore a clonal population of T cells with identical TCR rearrangements remain for years against the recognized antigen[2], thus forming a potentially mappable immunological signature. Learning these signatures could have implications in broad range of clinical applications including infectious diseases, autoimmunity and tumor immunology.

T cells undergo non-homologous recombination during T cell development, which involves rearrangement of the germline TCR loci from a large collection of variable (V), diversity (D) and joining (J) gene segments as well as template-independent insertions and deletions at the V-D and D-J junctions[3,4]. TCRs are formed by a pair of α and β-chains (90-95% of T cells) or γ and δ-chains (5-10%) and V(D)J recombination happens in each locus independently. It is estimated that this rearrangement can result in the range of 10^{18} different TCR genes[5,6] which provides enormous diversity for epitope-specific T cell repertoire. Furthermore, due to the low affinity of TCR-pMHC-interaction, TCR recognition is degenerate and a single TCR can interact with more than 1 million different epitopes (cross-reactivity), and a given epitope can elicit response from millions of TCRs[7,8]. Given these three levels of diversity, predicting TCR's epitope specificity is notably challenging[9].

The complementarity determining regions (CDRs) of a TCR determine whether the TCR recognizes and binds to an antigen or not[10]. Of these regions, CDR3 is the most variable and primarily interacts with the peptide, while CDR1 and CDR2 primarily interact with the peptide binding groove of the MHC protein presenting the peptide, but can also be in contact with the peptide[11,12]. Dash et *al.*[13] noted that also a loop between CDR2 and CDR3 (IMGT®[14] positions 81-86), which they called CDR2.5, has sometimes been observed to make contact with pMHC in solved structures. Figure 1 shows these CDRs in interaction with a peptide-MHC-complex (pMHC).

It is well known that for example the CDR3β of a TCR is important in recognizing peptides presented to the T cell, but it is still unclear which specific physicochemical or structural features of the CDR3β or of other parts of the TCR determine the antigen recognition specificity of that T cell. Although high-throughput DNA sequencing has enabled large-scale characterization of TCR sequences[6], it still remains exhaustive to profile epitope-specific TCRs as they require sample-consuming experiments with distinct pMHC-multimers for each epitope of interest. Therefore, there is a great need for models that examine which epitopes a TCR can recognize or to which TCRs an epitope can bind to. Curated databases of experimentally verified TCR-peptide interactions have recently been launched, such as the VDJdb, IEDB, and McPAS[16,17,18]. Such data sources are enabling more comprehensive, data-driven analysis of TCR-peptide interactions, and make it possible to

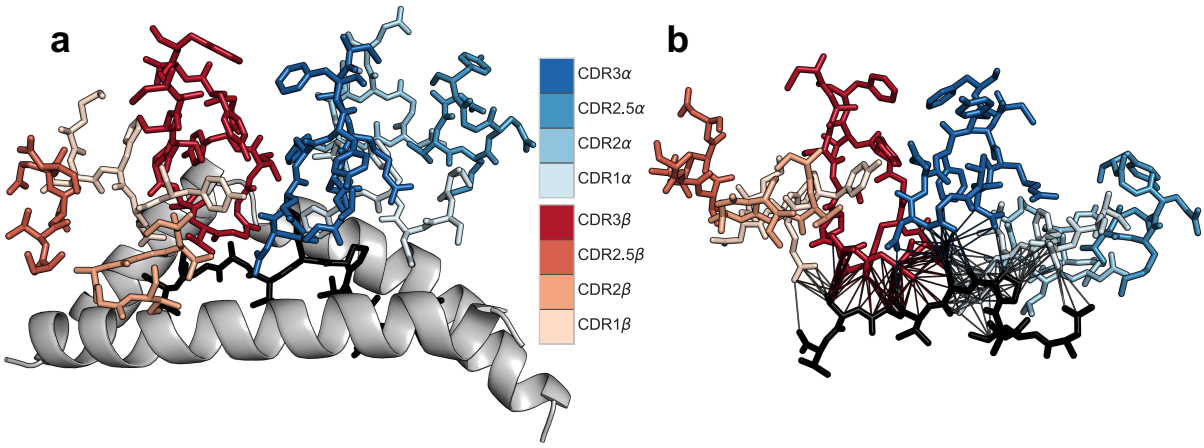

Figure 1: **Structure of a TCR-pMHC -complex. a)**CDRs of a TCR binding to CMV-epitope pp65$_{495\text{-}503}$ (shown in black), presented by MHC-protein, whose binding groove is shown as a white cartoon. **b)** Distances below 5 Å between the atoms of the epitope and the different CDRs are shown. The original structure was determined by Gras et *al.*[15].

use techniques from statistical machine learning for the aforementioned tasks. Yet only a few computational TCR specificity models have been proposed in the literature[12,13,19], some of which rely on heuristics, may be suboptimal for small datasets and have not been benchmarked against each other.

We propose a method called TCRGP which builds on non-parametric modelling using Gaussian process (GP) classification. Probabilistic formulation of GPs allows robust model inference also from small data sets, as is currently the case for TCR-peptide interaction information in curated databases. As the space of all TCRs that can recognize a certain epitope is potentially very large, it is important to avoid overfitting to the limited sample of TCRs that is available. Indeed, TCRGP clearly outperforms the current state-of-the art methods for predicting the epitope specificity of TCRs. At the same time, TCRGP also scales to extremely large data sets which we expect for the future epitope specific TCR-seq data sets. We also analyze the effects of utilizing different sections of the TCR amino acid sequence, and examine how the number of available TCRs for training affects the predictions. Finally, we demonstrate the usefulness of TCRGP in analyzing single-cell RNA+TCR$\alpha\beta$-sequencing data from hepatocellular carsinoma patients.

Results

Gaussian process classifier for TCRs

Gaussian processes (GP) are a flexible class of models that have become popular in machine learning and statistics with various applications in molecular biology, bioinformatics and other fields[20,21,22,23,24]. We have developed TCRGP, a Gaussian process based probabilistic classifier to predict TCRs' epitope specificity. GPs are nonparametric and differ from standard parametric models in that they define priors for entire nonlinear functions, instead of their parameters. GPs implement a Bayesian nonparametric kernel method for learning from data. Properties of GPs are defined by the kernel function, which is a function of objects that we want to classify. Our objects are amino acid sequences (strings) that have varying lengths. While kernel functions can be defined for strings, we use a feature representation that first aligns the amino acid sequences into a fixed length presentation using IMGT® definitions. BLOSUM based substitution matrices are then used to measure similarities between aligned amino acids via squared exponential kernel function, whose hyperparameters control the complexity and smoothness of the classifier.

TCRGP utilizes GPs' probabilistic nature and infers the classifier using variational inference. Probabilistic inference makes the method more robust in small data regime, where the current experimental data sets are, while sparse variational inference scales the method to extremely large (future) data sets. Since it is currently poorly understood that how different CDR regions and α/β chains (features) contribute to the epitope specificity (see Fig. 1), we extend TCRGP to use all these features by using multiple kernel learning and use experimental data to automatically calibrate the strength of each feature's contribution to the final classifier. See Methods Section for a detailed description of TCRGP method.

We use two data sets to demonstrate TCRGP's accuracy in predicting TCR epitope specificity: a recently published data set of tetramer sorted TCR sequences for 10 epitopes, introduced by Dash et *al.*[13], and a new dataset of medium and high quality epitope-specific TCR sequences extracted from VDJdb database[16]. The Dash data provides the largest set of epitope-specific paired TCR$\alpha\beta$-data that we are aware of, and VDJdb provides a comprehensive selection of available epitope-specific TCRb-data currently available. We also considered using using TCRs from IEDB[17] and McPAS[18], but they had significant overlap with VDJdb and their collections of TCRbs were not as extensive. Both of the selected data sets are combined with a set of background TCRs, also presented by Dash et *al.*[13] that are not expected to recognize the epitopes in the two data sets. See Methods Section for details for the data sets. Our work is accompanied by an efficient software implementation that contains pre-existing models for predicting TCRs' specificity to epitopes involved in data sets used in this study as well as tools for building new epitope specificity models from new datasets.

Evaluating the significance of utilizing different CDRs

To evaluate the benefit of using different CDRs, we used the data set of Dash et *al.*[13] which includes 4635 pMHC-tetramer sorted single-cell sequenced TCR$\alpha\beta$ clonotypes from 10 epitope-specific repertoires (from hereon referred as the Dash data). We trained our TCRGP model using either only CDR3 or also with CDR1, CDR2, and CDR2.5 from TCRα, TCRβ, or both. We applied leave-one-subject-out cross-validation as described in Methods Section. Figure 2 a presents the cross-validation results for a single BMLF1$_{280\text{-}288}$-epitope from EBV and demonstrates how the classification results vary between different subjects likely due to the high variety of the TCRs.

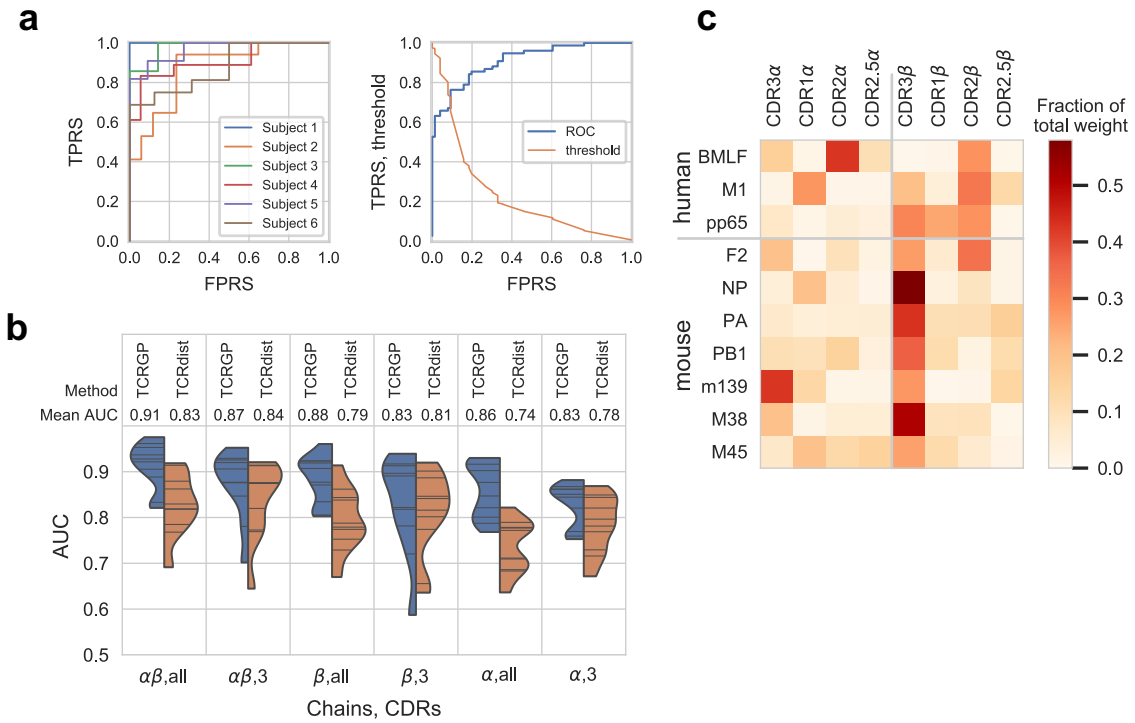

Figure 2: **Epitope-specificity prediction with Dash data. a)** The left panel shows the cross-validated ROC curves for each subject in the Dash data for $\text{BMLF1}_{280\text{-}288}$, when TCRGP has been trained using all CDRs from TCRα and TCRβ. The mean AUROC is 0.905. The right panel shows the ROC curves when the predictions have been combined and also the corresponding threshold values. From this figure we can determine which threshold values correspond to different true positive rates (TPRS) and false positive rates (FPRS). **b)** The blue parts of the violin plots illustrate the AUROC-scores of predictions made by TCRGP for all the epitopes. The orange sides illustrate the AUROC-scores obtained with TCRdist. A horizontal line within a violin plot presents the mean AUROC-score obtained for one epitope. The used chains (α and/or β) and CDRs (three or all) are indicated below each panel. **c)** Fractions of total weight given to kernels corresponding to different CDRs, when TCRGP has been trained to predict which TCRs are specific to the epitopes in the Dash data using all CDRs from both TCR chains.

AUROC-scores of the predictions for different combinations of CDRs and α/β chains are summarized in Fig. 2 b. For comparison, we also trained TCRdist in the same manner. Figure 2 b shows that both methods, TCRGP and TCRdist, perform on average better when using TCRβ than when using TCRα, although using both α and β chains generally provides the best results. There are few exceptions, as shown in Supplementary Fig. S1. For example, with peptides pp65 both models perform better when using CDR3α instead of CDR3β. Overall TCRGP is better than TCRdist in utilizing information from CDRs other than CDR3. TCRGP achieves higher AUROC-scores on average when trained using all CDRs instead of only CDR3, whereas with TCRdist the AUROC-scores seem to be similar or better when only CDR3 is utilized. Notably TCRGP outperforms TCRdist in prediction accuracy for 57 of the 60 comparisons (Supplementary Fig. S1, Fig. 2 a).

Figures 2 b and S1 also show that the AUROC-scores can have notable differences between different epitopes even when the same combinations of CDRs and α/β chains have been utilized. Some of these differences may be explained by the differences in the number of available training samples, for example for pp65 there were only 76 TCRs from 6 subjects in the Dash data, which may have contributed to a lower prediction accuracy. To address this, we evaluated the models also using leave-one-out cross-validation with only unique, private TCRs to see how the models perform when predictions are done only on new TCRs. We consider a TCR to be unique when it consists of a unique combination of CDR3 amino acid sequence and V-genes from both chains. With both TCRGP and TCRdist, the average AUROC-scores improve slightly (Supplementary Fig. 2 and Fig. 3), demonstrating that the models can predict the specificity of completely new sequences and that the larger number of TCRs used for training (due to the larger folds in leave-one-out cross validation) improve the model performances.

To better understand the significance of the different CDRs for TCR-pMHC recognition, we also examined more closely how TCRGP weighted the kernels created for the different CDRs, when all CDRs from both

chains were utilized. Figure 2 c illustrates which CDRs were found important for the different epitopes. As one might expect, with most of the epitopes most weight was given to the CDR3β, but utilizing several epitopes was found beneficial with all epitopes. This is in agreement with an alignment of 52 TCR sequences from TCR–pMHC PDB structure complexes, which demonstrates that all CDRs can be within 5Å of peptide[12]. For example with CMV-epitope pp65$_{495-503}$, experimental characterization of the structure observed contacts between the peptide and CDR3β, CDR1β, CDR3α and CDR1α (Fig. 1 b), and CDR2β is also in proximity of the peptide (within 5.8Å). Another TCR-pMHC-complex structure (PDBid 5D2L) for the same pp65$_{495-503}$-epitope suggests that CDR2β was also within 5 Å of the peptide. Indeed, the optimized weights for the pp65 epitope (Fig. 2 c) show some correspondence to the observed contacts (Fig. 1 b). However, with some epitopes CDR3β was not considered very important, as for example with mCMV-epitope m139$_{419\text{-}426}$ CDR3α is more important for the prediction, while with EBV-epitope BMLF1$_{280\text{-}288}$ most of the weight was given to CDR2α and CDR2β.

Comparisons to other methods

We also experimented with a data set we obtained from VDJdb, which gathers published epitope-specific TCR-sequencing results and is currently the largest collection of such data (from hereon referred as the VDJdb data). We again used leave-one-subject-out cross-validation as described in Methods Section. We trained TCRGP and TCRdist using only CDR3β and then also with the other CDRβs. Figure 3 a shows the ROC curves when TCRGP was trained with all CDRβs to predict which TCRs are specific to the HCV-epitope NS3$_{1436\text{-}1444}$.

We also trained a TCR-classifier as proposed by De Neuter et *al.*[19] [19] using the same data. Unfortunately, the background TCR data set from[13] did not contain information of the J-gene, which is requested by this TCR-classifier. However, according De Neuter et *al.*[19] themselves, not much weight was given to the J-gene at least in their experiments. This TCR-classifier does not utilize other CDRβs in addition to CDR3β, but the Vβ-gene from which the CDRβs can be derived from. Thus all these three methods get the same sequence information, when TCRGP and TCRdist use all CDRs, although in a slightly different form.

Figure 3 b shows the distributions of mean AUROC scores for each model trained for the 22 different epitopes. With this data set from VDJdb, we can see that TCRGP and TCRdist both perform better, when all CDRβs have been utilized. Remarkably, TCRGP outperforms the other methods when using all CDRs, but also when only the CDR3β is utilized. AUROC scores for the different epitopes are presented in Supplementary Fig. S4.

In the VDJdb data, there were also TCRs that appeared in samples collected from multiple subjects (see Table 3). We therefore trained the models also using leave-one-out cross-validation with only unique TCRs. In this case we considered a TCR to be unique if it had a unique combination of CDR3β amino acid sequence and Vβ-gene, as we only utilized the TCRβ. As with the Dash data above, our results (Supplementary Fig. S5 and Fig. S6) show that the models can predict the specificity of completely new sequences, thus demonstrating their use for epitope specificity prediction for previously unseen TCRs.

Significance of the number of training samples

To assess how the number of epitope-specific TCRs affects the performance of TCRGP classifier, we trained our model using different numbers of epitope-specific TCRs from the VDJdb data. We selected all unique TCRs for each epitope and took 100 random samples from them for each training set size, using always an equal number of randomly chosen control TCRs. Learning curves for four epitopes are shown in Fig. 3 c and learning curves for all 22 epitopes in Supplementary Fig. S7. In general, the predictive performance of the TCRGP classifiers improve when more training samples are available. However, the exact number of TCRs required to achieve a certain level of accuracy varies greatly between the different epitopes. This likely reflects the fact that different epitopes can be more selective in choosing their TCR interactions. In other words, TCRs that recognize one epitope can be more diverse than the TCRs that recognize another epitope[13], and if the TCRs are very heterogeneous, it requires more sampling to get a representative sample of these TCRs for the model training. Indeed, we observed a negative correlation between TCRs' diversity and prediction accuracy (Fig. 3 d).

These learning curves also further demonstrate the benefit of using multiple CDR sequences: With most of the epitopes using all CDRs produces better or comparable AUROC-scores with all sample sizes, although there are a few epitopes with which the AUROC-scores are higher when utilizing only the CDR3β if the sample sizes are very small (≤ 40). These results also suggest that with many epitopes it may be more beneficial to sequence a moderate amount of TCRs in such precision that in addition to the CDR3 also the V-gene and

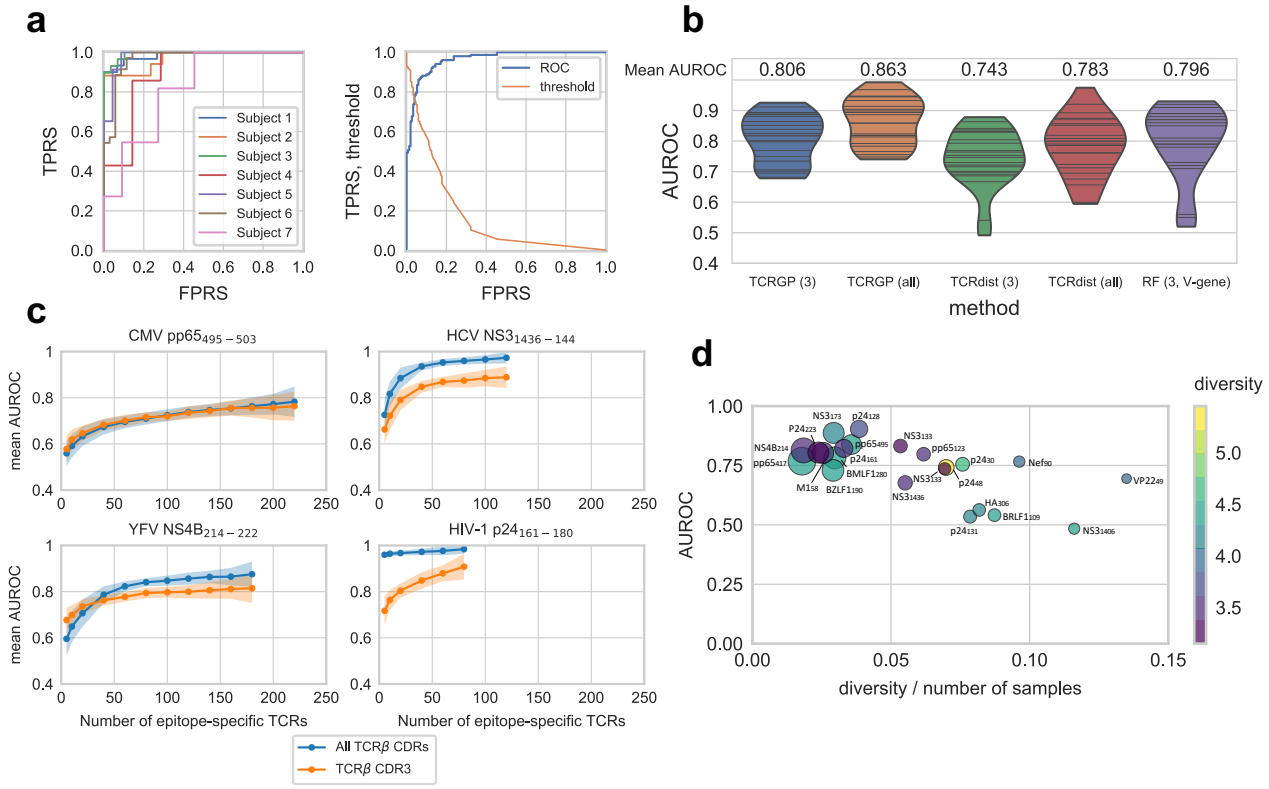

Figure 3: **Epitope specificity prediction with VDJdb data. a)** Left panel shows the cross-validated ROC curves for each subject in the VDJdb data for HCV $NS3_{1436\text{-}1444}$-epitope, when TCRGP has been trained using TCRα and TCRβ with all CDRs. The mean AUROC is 0.944. Right panel shows the ROC curves when all predictions have been combined and also the threshold values for classification are shown. From this figure we can determine which threshold values correspond to different true positive rates (TPRS) and false positive rates (FPRS). **b)** One violin plot presents the mean AUROC-scores obtained with one method for all epitopes in our VDJdb data. Below each violin plot there is the name of the method used and in the brackets which CDRs have been used (3 for CDR3, all for CDR1, CDR2, CDR2.5, and CDR3). A horizontal line within a violin plot presents the mean AUROC-score obtained for one epitope. RF refers to the Random Forest TCR-classifier of De Neuter et *al.*[19]. **c)** For each epitope from the VDJdb dataset, TCRGP models were trained using different numbers of unique epitope-specific TCRβs, always complemented with the same number of control TCRβs. For each point of the learning curve the model was trained with 100 random samples of the TCRβs, using either CDR1, CDR2, CDR2.5, and CDR3 (blue curves), or only CDR3 (orange curves). The darker lines show the mean of the predictions and the shaded areas $\pm$ the standard deviation for the 100 folds. The points indicate the tested sample sizes. Here learning curves for four peptides are shown. **d)** Leave-one-out cross-validated AUROC-scores correlate with the diversity and number of samples (Pearson correlation -0.66). The sizes of the circles indicate the number of unique TCRs used for training (see Table 3).

allele (and thus the CDR1, CDR2, and CDR2.5) can be determined, than to sequence large amounts of only CDR3s. These findings are inline with TCRGP-predicted weights for each CDR3 for individual epitope, as we can see in the case of CMV-epitope $pp65_{495\text{-}503}$, EBV-epitope $BMLF1_{280-288}$ and IAV-epitope $M1_{58-66}$. With $pp65_{495\text{-}503}$ most weight was given to CDR3β and thus information from other CDR3s are not as beneficial; with $BMLF1_{280-288}$ almost no weight was given to CDR3β and in the learning curves there is a clear improvement when all CDRβs are used; with $M1_{58-66}$ some weight was given to CDR3β, but most weight fell to CDR2β and correspondingly there is a small improvement in the learning curves, when all CDRβs are utilized. Overall, the learning curves show that TCRGP can learn an accurate predictor even from a small data set, thus making it applicable to the currently existing TCR-peptide interaction data sets. On the other hand, our results also show that TCRGP's prediction accuracy increases along with increasing number of training examples, enabling analysis of larger TCR-peptide interaction data sets in the future.

Leveraging TCRGP in single-cell RNA+TCR$\alpha\beta$-sequencing data analysis

We next demonstrate how TCRGP can be utilized to implement a novel analysis of combined single-cell RNA and TCR$\alpha\beta$ (scRNA+TCR$\alpha\beta$) sequencing data. Hepatocellular carcinoma (HCC) is one of the leading causes for cancer-related deaths worldwide[25]. Globally the predominant cause of HCC is considered to be Hepatitis B virus (HBV) as half of the HCC patients are estimated to be chronic HBV carriers[26]. During the course of natural infection, HBV integrates itself into the genome of the hepatocytes and thus a proportion of the HCC cells expresses HBV antigens[27]. Therefore, the malignant cells could be targeted by HBV-specific T-cell clonotypes and the high-dimensional characterization of these clonotypes could be crucial in understanding the viral control of HBV-infection and its association to HCC. To address this previously unanswered question we used TCRGP to analyze a published single-cell RNA and TCR$\alpha\beta$ dataset by Zheng et *al.*[28] of T cells from HBsAg-positive HCC-patients from blood, non-malignant liver tissue and tumour tissue (from hereon referred as the Zheng data).

Recently, Cheng et *al.*[29] mapped HBV-reactive T cell populations by exhaustively screening the whole HBV genome with an HLA-class I restricted multiplexed pMHC-tetramer strategy and characterized T cells against four interesting HBV-epitopes from two antigens with TCRβ-sequencing. We used this TCRβ data to train TCRGP classifiers (see Methods section for details) to enable prediction for the unselected TCR repertoire in the Zheng data against the four epitopes ($\text{HBV}_{\text{core169}}$, $\text{HBV}_{\text{core195}}$, $\text{HBV}_{\text{pol282}}$, $\text{HBV}_{\text{pol387}}$, where core refers to core protein and pol to polymerase protein) (Fig. 4 a).

Of the 789 CD8+ cells from Zheng data analyzed with TCRGP, 108 cells (13.688%), were predicted to be reactive against HBV with at least a probability of 85%, most of which against HBVcore195-epitope (59 cells) (Fig. 4 a, b and c). On the contrary, 176 cells were predicted to be reactive against common viruses (CMV=22, EBV=88 and Influenza A=66 cells) (Fig. 4 b), showing that HBV was the most common target for antigen-specific T cells in HCC patients.

After unsupervised clustering of the CD8+ cells' scRNA-seq data, we received 6 different phenotypes that were similar to the phenotypes described by Zheng et *al.*[28], but had the exhausted cells divided into 3 different clusters instead of one (naïve, effector, memory, exhausted 1, exhausted 2 and exhausted 3) (Fig. 4 c). Interestingly, cells in exhausted 3 cluster showed the highest enrichment of the clonotypes targeting $\text{HBV}_{\text{core195}}$-epitope (Fisher's exact test p=2.913e-06, Benjamini-Hochberg corrected for multiple testing p.adj=0.00103), but not to any other epitope-specific clonotypes (Fig. 4 d, e). By calculating exhaustion score for each T cell, we found that exhausted 3 cluster was the most exhausted (against exhausted 2 p=0.0032, against exhausted 1 p=0.0021) and the least cytotoxic cluster (p=0.019 and p=1.3e-05). Further, gene-level analysis showed high expression of *TIGIT* and *HAVCR2* (encoding TIM-3), which have been associated with late-stage exhaustion after long antigen exposure. Upregulated pathways for exhaustion cluster 3 were IL2-STAT5 signaling pathway (exh3 vs exh1 q=0.022 and exh3 vs exh2 q=0.000) and myogenesis pathway (q =0.016 and q=0.001).

In summary, TCRGP was able to identify a T cell cluster that was enriched with HBV-targeting clonotypes, which was the most exhausted and least functional. These differences in transcriptomes of the exhausted clonotypes could explain the role of viral control in the development of HCC in HBV-carriers, which is further complicated by differential expression of HBV antigens that can elicit either more effector or exhaustion-prone immune response.

Discussion

In this paper we have demonstrated that we can accurately predict if a previously unseen TCR can recognize an epitope, for which we have had sufficient amount of experimentally produced epitope-specific TCR-sequencing data for training. The performances of the models for different epitopes depend greatly on the size, quality, and heterogeneity of the repertoire of the epitope-specific T cells available for training, and not all the epitopes elicit oligoclonal responses that can be interpreted with machine learning models. We have also shown that the other CDRs in addition to CDR3β can provide useful information for the classification task and that it can depend on the epitope in question which of the CDRs are important.

As the amount of epitope-specific T cell sequences has expanded recently and the computational methods available are fairly new, no comparative effort has thus far emerged to gain understanding of the already available toolbox. In this work we provide the most thorough analysis of the current epitope-specificity prediction algorithms on the biggest data sets publicly available and thus provide important information to the community. The currently available sequence data of epitope-specific TCRs has allowed us to come this far, but it will be interesting to see what can be achieved when more data becomes available with modern high-throughput techniques presented recently[30,31]. Because of the limited data, we have not yet been able to consider the

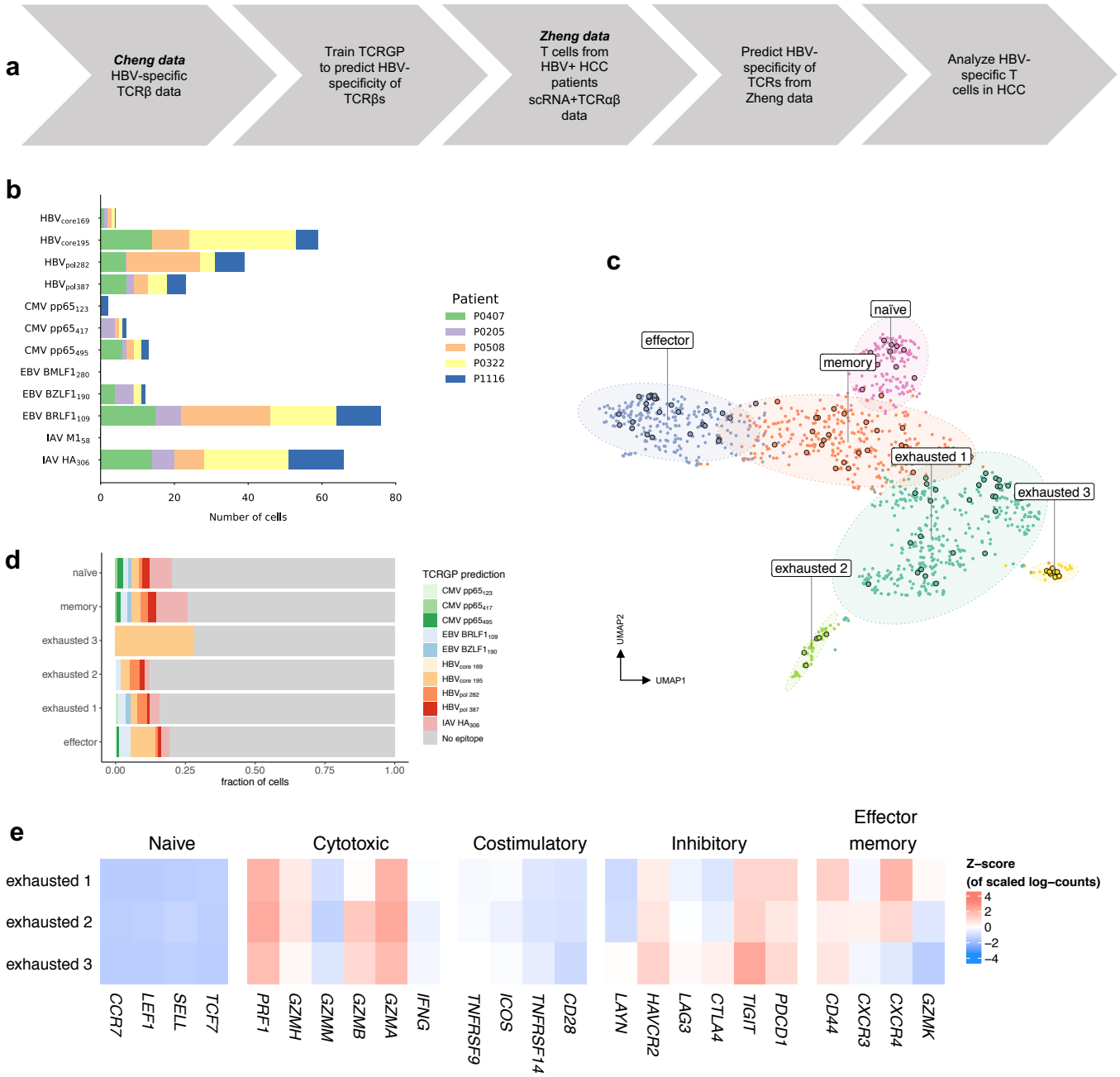

Figure 4: **Analysis of HBV-specific T cells in HCC patients. a)** Schematics for the analysis of single-cell RNA and TCR$\alpha\beta$ sequencing data using TCRGP and multimer-sorted data. **b)** Numbers of cells predicted to recognize different epitopes by TCRGP with probability of at least 85%. HBV-reactivity was assessed by four different TCRGP classifiers trained against four different HBV-epitopes ($\text{HBV}_{\text{core169}}$, $\text{HBV}_{\text{core195}}$, $\text{HBV}_{\text{pol282}}$, $\text{HBV}_{\text{pol387}}$). Other predictions were made using the models trained with the VDJdb data. **c)** Dimensionality reduced representation (UMAP) of the 1189 CD8+ T cells from HBV+ HCC-patients from peripheral blood, normal adjacent tissue and tumour tissue. Encircled dots represent the T cells predicted to be HBV-reactive by TCRGP.**d)** The frequencies of T cells predicted to recognize different HBV-epitopes in each cluster. **e)** The frequencies of T cells predicted to recognize different viral epitopes in each cluster. **f)** Z-score normalized mean expressions of known canonical markers to assess CD8+ cell phenotypes (naïve, cytotoxic, costimulatory inhibitory, and effector memory markers) in the three different exhausted cell clusters. Exhausted 3 was predicted to be enriched for HBV-targeting T cells (p=2.913e-06, p.adj=0.00103).

similarities of the epitopes and the significance of the different HLA-types of the MHC proteins presenting the epitopes. Having a larger variety of epitopes and TCRs that recognize them, would allow us to also better model the cross-reactivity of the TCRs. Eventually, when larger dataset become available, we may be able to model the similarity of the epitopes and consider the HLA-types in addition to the similarity of the TCRs and even predict if a TCR can recognize a previously unseen epitope. Furthermore, the proposed Gaussian process formalism has been shown to scale to very large datasets up to billion data points[32] by optimizing a small number of landmark sequences.

The previous supervised algorithms developed are presented in the case of epitope-specific data, but we believe that to answer clinically relevant questions we need to address the unselected repertoire data which is far more numerous in size and more easily produced. Therefore we presented a novel workflow for analysis of scRNA+TCRab data in a clinically relevant question, showing the power of determining the epitope-specifity *in silico* to reveal underlying transcriptomic heterogeneity of the epitope-specific T cells, which to our knowledge has not been tackled before with single-cell RNA-sequencing in tumor infiltrating lymphocytes in any tumor. As the number of scRNA+TCRab and conventional TCRb sequencing data in clinical settings is increasing[33,34,35,36,37,38,39,40], we expect that models like ours can be applied to a variety of research questions where exhaustive *ex vivo* pMHC-multimer assays are not feasible. In conclusion, we propose that TCRGP could be useful in the diagnosis and follow-up of infectious diseases, in autoimmune disorders and cancer immunotherapy.

Methods

Data

Our experiments focus on TCRs formed by a pair of α and β chains, as those are the most common type of TCRs[41]. The CDR3 sequence is formed by V(D)J recombination, but CDR1, CDR2, and CDR2.5 sequences are determined completely by the V-gene and allele[3]. Dash et *al.*[13] provide a table of all V-gene and allele combinations and the corresponding CDR1, CDR2, and CDR2.5 amino acid sequences aligned according to IMGT® definitions[14]. Our method can utilize the aligned amino acid sequences of all these CDRs from either one or both of the α- and β-chains of the TCR. Table 1 shows a few examples of TCR sequences and their alignment.

Epitope gene	CDR3α	CDR1α	CDR2α	CDR2.5α	CDR3β	CDR1β	CDR2β	CDR2.5β
BMLF1$_{280\text{-}288}$	CAASDGAGGTSYGKLTF	NSM------FDY	ISSI---KDK	NKSAKH	CASSLWT---GSHEQYF	SGH-------TS	YDE----GEE	F-PNYS
BMLF1$_{280\text{-}288}$	CAESL-------DMLTF	DSS------STY	IFSN---MDM	NKKDKH	CASSVVG----GNEQFF	SGD-------LS	YYN----GEE	F-PDLH
BMLF1$_{280\text{-}288}$	CAMREVMD--SNYQLIW	TSDP-----SYG	QGSY--DQQN	QKARKS	CASSVAQLAGGTDTQYF	SGD-------LS	YYN----GEE	F-PDLH
pp65$_{495\text{-}503}$	CAGQAS----QGNLIF	SIF-------NT	LYKA---GEL	GITRKD	CASSIQ--------ALLTF	SGH-------DY	FNN----NVP	P-NASF
pp65$_{495\text{-}503}$	CAVRDNSITGGFKTIF	TSG------FYG	NAL----DGL	SRSDSY	CASSYF-------DEKLFF	DFQ------ATT	SNEG---CKA	A-SLTL
pp65$_{495\text{-}503}$	CILSNN-----NDMRF	TISG-----TDY	GLT-----SN	AEDRKS	CSARDPSGLAGGLAETQYF	DFQ------ATT	SNEG---SKA	A-SLTL

Table 1: An example of aligned TCR sequences (from the Dash data) for two peptides. Each CDR type has been aligned separately according to IMGT definitions. CDR1, CDR2 and CDR2.5 sequences for both the α- and β-chains are defined by germline Vα- and Vβ-genes. The alignments for all possible CDR1, CDR2 and CDR2.5 sequences have been determined by Lefranc[14] and we use these alignments with all epitopes. CDR3s are aligned by adding a gap at the top of the loop[14]. The length of the alignment can then be determined by the length of the longest CDR3 available and can vary between different models.

In our experiments, we use a data set collected by Dash et *al.*[13], which contains epitope-specific paired TCRα and TCRβ chains for three epitopes from humans and for seven epitopes from mice, see Table 2 for details.

We also gather a new data set from VDJdb (`https://vdjdb.cdr3.net`), which is a database that contains TCR sequences with known antigen specificity[16]. Every entry in VDJdb has been given a confidence score between 0 and 3 (0: critical information missing, 1: medium confidence, 2: high confidence, 3: very high confidence). We constructed our data set so that we selected all epitopes that have at least 50 TCRβ sequences with a confidence score at least 1 and found 22 such epitopes, see Table 3 for details. VDJdb also contains TCRα sequences, but since these are not in general paired with corresponding TCRβ sequences, we chose to only experiment with the TCRβ sequences.

For the training and testing of the models, we also required some background TCRs that we do not expect to recognize the epitopes in our data sets. For this purpose we used a set of background TCRs constructed in Dash et *al.*[13].

The data sets we have used can be found from `github.com/emmijokinen/TCRGP`.

Species	Epitope species	Epitope gene	Epitope	Subjects	Samples	Unique TCR$\alpha\beta$s
Human	EBV	BMLF1$_{280\text{-}288}$	GLCTLVAML	6	76	69
	CMV	pp65$_{495\text{-}503}$	NLVPMVATV	10	61	60
	IAV	M1$_{58\text{-}66}$	GILGFVFTL	15	275	237
Mouse	IAV	PB1-F2$_{62\text{-}70}$	LSLRNPILV	9	117	117
	IAV	NP$_{366\text{-}374}$	ASNENMETM	24	305	263
	IAV	PA$_{224\text{-}233}$	SSLENFRAYV	15	324	293
	IAV	PB1$_{703\text{-}711}$	SSYRRPVGI	34	642	584
	mCMV	m139$_{419\text{-}426}$	TVYGFCLL	8	87	87
	mCMV	M38$_{316\text{-}323}$	SSPPMFRV	14	158	143
	mCMV	M45$_{985\text{-}993}$	HGIRNASFI	13	291	271

Table 2: The Dash data contains epitope-spcefic TCRs for Epstein-Barr virus (EBV), human Cytomegalovirus (CMV), Influenza A virus (IAV) and mouse Cytomegalovirus (mCMV).

Epitope species	Epitope gene	Epitope	Subjects	Samples	Unique TCRβs
CMV	pp65$_{123\text{-}131}$	IPSINVHHY	17	65	58
CMV	pp65$_{417\text{-}426}$	TPRVTGGGAM	29	184	122
CMV	pp65$_{495\text{-}503}$	NLVPMVATV	103	413	242
EBV	BMLF1$_{280\text{-}288}$	GLCTLVAML	54	299	152
EBV	BZLF1$_{190\text{-}197}$	RAKFKQLL	17	225	149
EBV	BRLF1$_{109\text{-}117}$	YVLDHLIVV	6	66	51
IAV	M1$_{58\text{-}66}$	GILGFVFTL	50	239	138
IAV	HA$_{306\text{-}318}$	PKYVKQNTLKLAT	11	56	50
HCV	NS3$_{1073\text{-}1081}$	CINGVCWTV	7	76	39
HCV	NS3$_{1406\text{-}1415}$	KLVALGINAV	4	65	65
HCV	NS3$_{1436\text{-}144}$	ATDALMTGY	7	152	139
HSV-2	VP22$_{49\text{-}57}$	RPRGEVRFL	5	68	29
YFV	NS4B$_{214\text{-}222}$	LLWNGPMAV	5	223	198
DENV1	NS3$_{133\text{-}142}$	GTSGSPIVNR	11	65	59
DENV3-4	NS3$_{133\text{-}142}$	GTSGSPIINR	8	51	46
HIV-1	p24$_{30\text{-}40}$	KAFSPEVIPMF	44	134	104
HIV-1	p24$_{48\text{-}56}$	TPQDLNTML	21	52	40
HIV-1	p24$_{128\text{-}135}$	EIYKRWII	12	81	60
HIV-1	p24$_{131\text{-}140}$	KRWIILGLNK	27	212	141
HIV-1	p24$_{161\text{-}180}$	FRDYVDRFYKTLRAEQASQE	17	141	95
HIV-1	p24$_{223\text{-}231}$	GPGHKARVL	1	62	53
HIV-1	Nef$_{90\text{-}97}$	FLKEKGGL	21	104	78

Table 3: Data set gathered from VDJdb contains epitope-specific TCRs for Cytomegalovirus (CMV), Epstein-Barr virus (EBV), Influenza A virus (IAV), Hepatitis C virus (HCV), Herpes Simplex virus type 2 (HSV-2), Yellow Fever virus (YFV), Dengue virus type 1 (DENV1), Dengue virus type 3 (DENV3-4), and Human immunodeficiency virus type 1 (HIV-1).

Leave-one-subject-out cross-validations

For the evaluation of the methods developed by us and others, we needed to divide our data sets for training and testing. Both of the data sets we use determine the subjects from whom each TCR in the data has been obtained from. We therefore chose to use leave-one-subject-out cross-validation, where we leave out all TCRs from one subject, train the model with all the other TCRs, test it with the TCRs left out, and repeat this for all subjects. This way the average number of TCRs per fold in the Dash data set varied between 6 (for $\text{pp65}_{\text{123-131}}$) and 22 (for $\text{M45}_{\text{985-993}}$), and in the VDJdb dataset between 3 (for $\text{p24}_{\text{30-40}}$) and 45 (for $\text{NS4B}_{\text{214-222}}$). The number of subjects and samples for each epitope can be found from Tables 2 and 3.

We also randomly selected a set of background TCRs, so that there was always an equal number of epitope-specific and background TCRs in both training and test sets. We thought this would be the most realistic procedure for the evaluation, as this is likely how these kinds of models will be applied to new data: A model is trained with some set of TCRs and then predictions should be made for TCRs sequenced from an individual from who we have not seen any TCR sequences beforehand.

The data set we gathered from VDJdb contains TCR sequences from multiple studies, many of which have used same conventions for naming their subjects. Therefore we used the combination of the PMID of the publication and the subject id as the subject identifier. For two epitopes, $\text{p24}_{\text{223-231}}$ and $\text{NS3}_{\text{1406-1415}}$ there were very few separate subjects, only one and four, respectively. With these epitopes we used 5-fold cross-validation instead of the leave-one-subject-out cross-validation.

Sequence representation

Computational methods require the data to have some presentation, that they can utilize. Character sequences with variable lengths often provide some challenges as many methods rely on numerical inputs of fixed sizes. One solution is to compare subsequences of same length instead of the complete sequences, which is what for example Generic String kernel (GSkernel) does[42].

However, by aligning the sequences more approaches become applicable. According to the IMGT definitions[14] CDR3s can be aligned by introducing a gap in the middle of the sequence (i.e. top of the loop). Alignments for CDR1s, CDR2s, and CDR2.5s can be found from `www.imgt.org`. When the sequences are aligned, all the sequences within a CDR class (1, 2, 2.5 or 3) have the same length (see Table 1).

We observe sequences $a_1 a_2 \cdots a_L$ of amino acids $a_j \in \mathcal{A} = \{A, R, N, \ldots, -\}$ at aligned positions $j = 1, \ldots, L$. The alignment guarantees that all sequences have the same possibly padded length L. We encode the amino acids a with global feature vectors $\boldsymbol{\phi}(a) \in \mathbb{R}^D$ that associate a D-length real-valued code with each of the 21 amino acids including the gap symbol. The sequences are then encoded as data vectors $\mathbf{x}$ by concatenating the L feature vectors into a $D \cdot L$ length column vectors $\mathbf{x} = (\boldsymbol{\phi}(a_1)^T, \ldots, \boldsymbol{\phi}(a_L)^T)^T \in \mathcal{X}$. We collect a dataset of N sequences into a matrix $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)^T \in \mathbb{R}^{N \times DL}$ with rows as sequences and columns as individual amino acid features in aligned order. Each sequence is associated with a class label $y_i \in \{0, 1\}$ that indicates whether the sequence was epitope-specific or not. We collect the class labels into an output vector $\mathbf{y} = (y_1, \ldots, y_N)^T \in \{0, 1\}^N$.

We can observe amino acid sequences for both the α- and β-chains and the four complementarity determining regions (CDR) $\{1, 2, 2.5, 3\}$ from a single TCR. Sequence data for each chain and CDR combination has individual alignments and sequence lengths. We denote the data as $(\mathbf{X}_{\alpha,1}, \mathbf{X}_{\alpha,2}, \mathbf{X}_{\alpha,2.5}, \mathbf{X}_{\alpha,3}, \mathbf{X}_{\beta,1}, \mathbf{X}_{\beta,2}, \mathbf{X}_{\beta,2.5}, \mathbf{X}_{\beta,3}, \mathbf{y})$.

Substitution matrices such as BLOSUM62[43] describe the similarity of each amino acid. We modified the BLOSUM62 to include also the gap used in alignments and scaled the matrix values between 0 and 1. The resulting matrix $\mathbf{B} \in \mathbb{R}^{21 \times 21}$ is then positive semidefinite. We apply eigendecomposition $\mathbf{B} = \mathbf{V}\mathbf{S}\mathbf{V}^T$, where the column vectors of $\mathbf{V}$ encode orthogonal projections of the amino acids on the rows. We use the row vectors of $\mathbf{V}$, indexed by the amino acids a from the modified BLOSUM62, as our descriptions $\boldsymbol{\phi}(a) = \mathbf{V}_{a,:}$ with a feature representation $\boldsymbol{\phi}(a)^T \mathbf{S} \boldsymbol{\phi}(b) = [\mathbf{B}]_{ab}$ for any two amino acids $a, b \in \mathcal{A}$. It is possible to use also different substitution models and feature vectors obtained from different sources, or even use the so-called one-hot-encoding, but here we relied only on the eigenvectors of the (gap-extended) BLOSUM62. For our model, we utilized all the 21 components, but in Fig. S8 we show how the amino acids locate on the first two components.

Gaussian process classification

We use Gaussian process (GP) classification[44] to predict if a TCR recognizes a certain epitope or not. Gaussian processes model Gaussian distributions of non-parametric and non-linear functions. We apply a link function to squash the function values to a range $[0, 1]$ suitable for classification. GPs have a clear advantage of

characterizing the prediction uncertainty with class probabilities instead of point predictions. GPs naturally model sequences through kernel functions focusing on sequence similarity as the explaining factor for class predictions.

We use a GP function f to predict the latent epitope-specificity *score* $f(\mathbf{x}) \in \mathbb{R}$ of a sequence $\mathbf{x}$. A zero-mean GP prior

$$f(\mathbf{x}) \sim \mathcal{GP}(0, k(\mathbf{x}, \mathbf{x}')),$$

defines a distribution over functions $f(\mathbf{x})$ whose mean and covariance are

$$\mathbb{E}[f(\mathbf{x})] = 0$$
$$\mathbf{cov}[f(\mathbf{x}), f(\mathbf{x}')] = k(\mathbf{x}, \mathbf{x}'),$$

where $k(\cdot, \cdot)$ is the kernel function. We use the standard squared exponential kernel on the vectorized feature representation,

$$k(\mathbf{x}, \mathbf{x}'|\theta) = \sigma^2 \exp\left(-\frac{(\mathbf{x}-\mathbf{x}')^T(\mathbf{x}-\mathbf{x}')}{2\ell^2}\right), \tag{1}$$

where ℓ is the length-scale parameter, σ^2 is the magnitude parameter and $\theta = (\ell, \sigma^2)$. For any collection of TCR sequences $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$, the function values follow a multivariate normal distribution

$$p(\mathbf{f}) = \mathcal{N}(\mathbf{f}|\mathbf{0}, \mathbf{K_{XX}}), \tag{2}$$

where $\mathbf{f} = (f(\mathbf{x}_1), \ldots, f(\mathbf{x}_N))^T \in \mathbb{R}^N$ collects all function predictions of the sequences, and $\mathbf{K_{XX}} \in \mathbb{R}^{N \times N}$ is the sequence similarity matrix with $[\mathbf{K_{XX}}]_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$. The key property of Gaussian processes is that they couple all predictions to be dependent. The Gaussian process predicts similar epitope values $f(\mathbf{x}), f(\mathbf{x}')$ for sequences $\mathbf{x}, \mathbf{x}'$ if they are similar according to the kernel $k(\mathbf{x}, \mathbf{x}')$.

The latent function $f(\mathbf{x})$ represents an unbounded real-valued classification score, which we turn into a classification likelihood by the probit link function $\Phi : \mathbb{R} \mapsto [0, 1]$,

$$\Phi(f) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{f} \exp\left(-\frac{1}{2}\tau^2\right) d\tau. \tag{3}$$

The joint model then decomposes into a factorized Bernoulli likelihood and Gaussian prior,

$$p(\mathbf{y}, \mathbf{f}) = p(\mathbf{y}|\mathbf{f})p(\mathbf{f}) \tag{4}$$
$$= \left[\prod_{i=1}^{N} \mathrm{Ber}(y_i|\Phi(f_i))\right] \cdot \mathcal{N}(\mathbf{f}|\mathbf{0}, \mathbf{K_{XX}}), \tag{5}$$

where f_i is a shorthand for $f(\mathbf{x}_i)$. The objective of Gaussian process modelling is to infer the posterior distribution $p(\mathbf{f}|\mathbf{y})$, which is intractable for many non-Gaussian likelihoods. Additionally inferring the kernel hyper-parameters θ entails computing the marginalized *evidence*

$$p(\mathbf{y}; \theta) = \mathbb{E}_{p(\mathbf{f};\theta)}[p(\mathbf{y}|\mathbf{f})], \tag{6}$$

which is also intractable in general and has a limiting cubic complexity $\mathcal{O}(N^3)$[44]. We tackle the scalability with sparse Gaussian processes[45] and the intractability with stochastic variational inference[46].

Variational inference for low-rank GP approximation

We consider low-rank sparse Gaussian processes by augmenting the system with M inducing *landmark* pseudo-sequences $\mathbf{z}_j \in \mathcal{X}$ with associated (label) function values $u_j = f(\mathbf{z}_j) \in \mathbb{R}$. We collect all inducing points into structures $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_M)^T$ and $\mathbf{u} = (u_1, \ldots, u_M)^T$. By conditioning the GP with these values we obtain the augmented Gaussian process joint model

$$p(\mathbf{y}, \mathbf{f}, \mathbf{u}) = p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{u})p(\mathbf{u}) \tag{7}$$
$$p(\mathbf{f}|\mathbf{u}) = \mathcal{N}(\mathbf{f}|\mathbf{A}\mathbf{u}, \mathbf{K_{XX}} - \mathbf{A}\mathbf{K_{ZZ}}\mathbf{A}^T) \tag{8}$$
$$p(\mathbf{u}) = \mathcal{N}(\mathbf{u}|\mathbf{0}, \mathbf{K_{ZZ}}) \tag{9}$$
$$\mathbf{A} = \mathbf{K_{XZ}}\mathbf{K_{ZZ}}^{-1}, \tag{10}$$

where $\mathbf{K_{XX}} \in \mathbb{R}^{N \times N}$ is the kernel between observed sequences, $\mathbf{K_{XZ}}$ is between observed and induced sequences and $\mathbf{K_{ZZ}}$ is between induced sequences. The matrix $\mathbf{A}$ projects the M inducing points to the full observation space of N sequences.

Next, we define a variational approximation for the inducing points,

$$q(\mathbf{u}) = \mathcal{N}(\mathbf{u}|\mathbf{m}, \mathbf{S}) \tag{11}$$

$$q(\mathbf{f}) = \int p(\mathbf{f}|\mathbf{u})q(\mathbf{u})d\mathbf{u} \tag{12}$$

$$= \mathcal{N}(\mathbf{f}|\mathbf{A}\mathbf{m}, \mathbf{K_{XX}} + \mathbf{A}(\mathbf{S} - \mathbf{K_{ZZ}})\mathbf{A}^T), \tag{13}$$

where $\mathbf{m} \in \mathbb{R}^M$ and $\mathbf{S} \succeq \mathbf{0} \in \mathbb{R}^{M \times M}$ are free variational parameters to be optimized. It can be shown that minimizing the Kullback-Leibler divergence $\mathrm{KL}[q(\mathbf{u})||p(\mathbf{u}|\mathbf{y})]$ between the approximative posterior $q(\mathbf{u})$ and the true low-rank posterior $p(\mathbf{u}|\mathbf{y})$ is equivalent to maximizing the evidence lower bound (ELBO)[47]

$$p(\mathbf{y}) \geq \sum_{i=1}^{n} \mathbb{E}_{q(f_i)}[\log p(y_i|f_i)] - \mathrm{KL}[q(\mathbf{u})||p(\mathbf{u})]. \tag{14}$$

The log expectation is tractable for Probit likelihoods[48], while the KL term similarly has a closed form for two Gaussian densities.

Due to the small data regime we choose the optimal assignment of selecting $\mathbf{Z} = \mathbf{X}$ and $\mathbf{u} = \mathbf{y}$, which corresponds to the full Gaussian variational approximation of Nickish et *al.*[49], while for larger datasets the inducing landmark points can also be optimised[46]. We then optimize the ELBO (14) with respect to the variational parameters $\mathbf{m}$ and $\mathbf{S}$ as well as the kernel hyperparameters θ, that is, the lengthscales ℓ_{cr} and weights w_{cr}.

Finally, predictions $\mathbf{f}_*$ of new test sequences $\mathbf{X}_* \subset \mathcal{X}$ follow a variational predictive posterior

$$p(\mathbf{f}_*|\mathbf{y}) = \int p(\mathbf{f}_*|\mathbf{u})p(\mathbf{u}|\mathbf{y})d\mathbf{u} \tag{15}$$

$$\approx \int p(\mathbf{f}_*|\mathbf{u})q(\mathbf{u})d\mathbf{u} \tag{16}$$

$$= \mathcal{N}(\mathbf{f}_*|\mathbf{A}_*\mathbf{m}, \mathbf{K}_{\mathbf{X}_*\mathbf{X}_*} + \mathbf{A}_*(\mathbf{S} - \mathbf{K_{ZZ}})\mathbf{A}_*^T), \tag{17}$$

where $\mathbf{A}_*$ indicates projection from the landmark points $\mathbf{Z}$ to the new sequences $\mathbf{X}_*$. The predictive distribution is a Gaussian distribution for the latent test values $\mathbf{f}_*$, from which the distributions of the test labels can be retrieved through the link function.

We have implemented our model using GPflows VGP-model[50,51]. Our code, data sets, and some examples can be found from `github.com/emmijokinen/TCRGP`.

Multiple kernel learning

As mentioned in Section 4.1, when a TCR binds to a pMHC, its CDR3β is presumably always in contact with the peptide while the other CDRs may contact the peptide, but mainly contact the peptide binding groove of the MHC presenting the peptide.[13] took this into account by giving fixed weights for the distances between amino acids within different CDRs, giving more weight to the CDR3. As it can vary which CDRs can be in contact with different peptides, we did not want to determine the importance of these different CDRs beforehand, but instead created separate kernels for each CDR and let our model decide which of them are important. We define the kernel as a convex combination of the four CDR regions r and the two chains c,

$$k(\mathbf{x}, \mathbf{x}') = \sum_{\substack{r \in \\ \{1,2,2.5,3\}}} \sum_{\substack{c \in \\ \{\alpha,\beta\}}} w_{cr} k_{cr}(\mathbf{x}, \mathbf{x}'; \theta_{cr}), \tag{18}$$

where the weights $w_{cr} \geq 0$ are non-negative.

TCR repertoire diversity

To estimate the diversity of the epitope-specific TCRs for each epitope, we developed a diversity measure following the example of Dash et *al.*[13]. The Simpson's diversity index was then generalized to account for the

similarity of TCRs by utilizing the Gaussian kernel function as follows:

$$\text{diversity} = \left(\frac{\sum_{i=0}^{N-1} \sum_{j=i+1}^{N} \sigma^2 \exp\left(-\frac{||\mathbf{x}_i - \mathbf{x}_j||^2}{2l^2} \right)}{\frac{1}{2}(N-1)N} \right)^{-1} . \tag{19}$$

Here σ^2 is the kernel variance and l is the lengthscale of the Gaussian kernel used by TCRGP, and $\mathbf{x}_i$ and $\mathbf{x}_j$ are feature vectors for the TCRs $i, j \in [1, N]$. The kernel variance and lengthscale were set to the average values used for the 22 epitopes in the VDJdata ($\sigma^2 = 5.52, l = 2.50$).

TCRGP classifiers for HBV-epitopes

Utilizing the TCRs specific to HBV epitopes $\text{HBV}_{\text{core169}}$, $\text{HBV}_{\text{core195}}$, $\text{HBV}_{\text{pol282}}$, and $\text{HBV}_{\text{pol387}}$ from Cheng et *al.*[29] and control sequences from Dash et *al.*[13] we trained a TCRGP classifier for each epitope. We utilized all epitope-specific TCRs from which we could determine also CDR1β, CDR2β, and CDR2.5β in addition to CDR3β and complemented these epitope-specific TCRs with the same amount of control TCRs. We considered TCRs which were predicted to recognize the epitopes with 85 % probability as epitope-specific. The amounts of epitope-specific TCRs and AUROC-scores obtained from leave-one-subject-out cross-validations for each epitope are shown in Table 4. We used TCRGP and VGP with all epitopes except for $\text{HBV}_{\text{pol387}}$, with which we used SVGP with 700 inducing points due to the high number of samples.

Epitope	Samples	Subjects	AUROC
$\text{HBV}_{\text{core169}}$	699	9	0.756
$\text{HBV}_{\text{core195}}$	588	12	0.847
$\text{HBV}_{\text{pol282}}$	459	12	0.880
$\text{HBV}_{\text{pol387}}$	1348	12	0.760

Table 4: HBV-epitopes for which we trained TCRGP classifiers, the numbers of epitope-specific TCRs and subjects and mean AUROC-scores from leave-one-subject-out cross-validations.

Single-cell RNA-sequencing analysis

The unnormalized expression count data of T cells passing the quality control in the Zheng data were fetched from GEO (GSE98638) along with the TCR$\alpha\beta$-sequences inferred from the full-transcript single-cell RNA-sequencing data and inferred phenotypic states as described by Zheng et *al.*[28]. As the TCRβ-sequenced training data for HBV-specific epitopes was HLA-A -restricted, we focused our analysis only on T cells capable of peptide recognition in HLA-A restricted manner, namely clusters CD8-LEF1, CD8-CX3CR1, CD8-LAYN and CD8-GZMK. The data was log-normalized to 10 000 counts per cell and scaled accordingly with the Seurat 3.0.2.[52] package for R 3.5.2. The highly variable genes (HVGs) were chosen as the genes showing the highest mean to variance ratio (min expression = 0.5, max expression 3, min variance 0.5) with the FindVariableFeatures-function. The linear dimensionality reduction was calculated with PCA for the scaled expression matrix containing only HVGs. Non-linear dimensionality reduction was performed with UMAP for principal components that had standard deviation > 2 using standard parameters with the RunUMAP-function. To receive a better grouping for the selected cells, we used a graph-based clustering approach implemented in the Seurat tool. To find the shared nearest neighbor graph, the function FindNeighbors was used with the same amount of PCs as with UMAP. To determine optimal clustering, FindClusters-functions was used with several parameter values for the resolution parameter, ranging from $[0.1, 3]$. The optimal clustering was decided by agreement of grouping in the UMAP-embedding and the labels from clustering by visual interpretation. The cytotoxic and exhaustion signatures for the clusters were calculated as cell-wise mean expression of cytotoxic (*NKG7*, *CCL4*, *CST7*, *PRF1*, *GZMA*, *GZMB*, *IFNG*, *CCL3*) and exhaustion genes (*CTLA4*, *PDCD1*, *HAVCR2*, *TIGIT*, *LAG3*). The one-sided Fisher's test for enrichment of epitope-specific T cells to different phenotypes was calculated independently for individual and pooled patients, epitopes and tissues which were then adjusted with Benjamini-Hochberg for false-discovery.

Data and code availability

TCRGP software tool and the data sets used for the evaluation of the method are available at `https://github.com/emmijokinen/TCRGP`. Software and data for the single-cell RNA-sequencing analysis of HCC-patients are available at `https://github.com/janihuuh/tcrgp_manu_hcc`.

Acknowledgements

We would like to acknowledge the computational resources provided by the Aalto Science-IT. This work has been supported by The European Research Council (M-IMM project), Academy of Finland (project numbers: 287224, 299915, 313271, 314442 and 314445), Finnish special governmental subsidy for health sciences, research and training, the Sigrid Juselius Foundation and the Finnish Cancer Societies

Author Contributions

All authors contributed to designing the study. E.J., M.H. and H.L. co-developed the method. E.J. implemented TCRGP and carried out the experiments with support from other authors. J.H. developed the analysis approach for scRNA+TCR$\alpha\beta$-sequencing data from HCC-patients and performed the analysis with support from other authors. S.M. helped with analyzing the results. All authors contributed to the manuscript writing.

References

1. Davis, M. M. & Bjorkman, P. J. A model for T cell receptor and MHC/peptide interaction. In *Mechanisms of Lymphocyte Activation and Immune Regulation II*, 13–16 (Springer, 1989).
2. Miles, J. J. *et al.* T-cell grit: large clonal expansions of virus-specific cd8+ t cells can dominate in the peripheral circulation for at least 18 years. *Blood* **106**, 4412–4413 (2005).
3. Bassing, C. H., Swat, W. & Alt, F. W. The mechanism and regulation of chromosomal V(D)J recombination. *Cell* **109**, S45–S55 (2002).
4. Cabaniols, J.-P., Fazilleau, N., Casrouge, A., Kourilsky, P. & Kanellopoulos, J. M. Most α/β T cell receptor diversity is due to terminal deoxynucleotidyl transferase. *Journal of Experimental Medicine* **194**, 1385–1390 (2001).
5. Arstila, T. P. *et al.* A direct estimate of the human $\alpha\beta$ T cell receptor diversity. *Science* **286**, 958–961 (1999).
6. Robins, H. S. *et al.* Comprehensive assessment of T-cell receptor β-chain diversity in $\alpha\beta$ T cells. *Blood* **114**, 4099–4107 (2009).
7. Wooldridge, L. *et al.* A single autoimmune T cell receptor recognizes more than a million different peptides. *Journal of Biological Chemistry* **287**, 1168–1177 (2012).
8. Sewell, A. K. Why must T cells be cross-reactive? *Nature Reviews Immunology* **12**, 669 (2012).
9. Miho, E. *et al.* Computational strategies for dissecting the high-dimensional complexity of adaptive immune repertoires. *Frontiers in immunology* **9**, 224 (2018).
10. Lefranc, M.-P. & Lefranc, G. *The T cell receptor FactsBook* (Elsevier, 2001).
11. Rudolph, M. G., Stanfield, R. L. & Wilson, I. A. How TCRs bind MHCs, peptides, and coreceptors. *Annu. Rev. Immunol.* **24**, 419–466 (2006).
12. Glanville, J. *et al.* Identifying specificity groups in the T cell receptor repertoire. *Nature* **547**, 94 (2017).
13. Dash, P. *et al.* Quantifiable predictive features define epitope-specific T cell receptor repertoires. *Nature* **547**, 89 (2017).

14. Lefranc, M. The IMGT unique numbering for immunoglobulins, T-cell receptors, and Ig-like domains. *Immunologist* **7**, 132–136 (1999).

15. Gras, S. *et al.* Structural bases for the affinity-driven selection of a public TCR against a dominant human cytomegalovirus epitope. *The Journal of Immunology* **183**, 430–437 (2009).

16. Shugay, M. *et al.* VDJdb: a curated database of T-cell receptor sequences with known antigen specificity. *Nucleic acids research* **46**, D419–D427 (2017).

17. Vita, R. *et al.* The immune epitope database (iedb): 2018 update. *Nucleic acids research* **47**, D339–D343 (2018).

18. Tickotsky, N., Sagiv, T., Prilusky, J., Shifrut, E. & Friedman, N. McPAS-TCR: a manually curated catalogue of pathology-associated T cell receptor sequences. *Bioinformatics* **33**, 2924–2929 (2017).

19. De Neuter, N. *et al.* On the feasibility of mining CD8+ T cell receptor patterns underlying immunogenic peptide recognition. *Immunogenetics* **70**, 159–168 (2018).

20. Cheng, L. *et al.* An additive gaussian process regression model for interpretable non-parametric analysis of longitudinal data. *Nature communications* **10**, 1798 (2019).

21. Jokinen, E., Heinonen, M. & Lähdesmäki, H. mgpfusion: predicting protein stability changes with gaussian process kernel learning and data fusion. *Bioinformatics* **34**, i274–i283 (2018).

22. Romero, P. A., Krause, A. & Arnold, F. H. Navigating the protein fitness landscape with gaussian processes. *Proceedings of the National Academy of Sciences* **110**, E193–E201 (2013).

23. Clifton, L., Clifton, D. A., Pimentel, M. A., Watkinson, P. J. & Tarassenko, L. Gaussian processes for personalized e-health monitoring with wearable sensors. *IEEE Transactions on Biomedical Engineering* **60**, 193–197 (2012).

24. Chu, W., Ghahramani, Z., Falciani, F. & Wild, D. L. Biomarker discovery in microarray gene expression data with gaussian processes. *Bioinformatics* **21**, 3385–3393 (2005).

25. Bosetti, C., Turati, F. & La Vecchia, C. Hepatocellular carcinoma epidemiology. *Best Pract. Res. Clin. Gastroenterol.* **28**, 753–770 (2014).

26. Hassan, M. M. *et al.* Risk factors for hepatocellular carcinoma: synergism of alcohol with viral hepatitis and diabetes mellitus. *Hepatology* **36**, 1206–1213 (2002).

27. Wang, Y. *et al.* Different expression of hepatitis B surface antigen between hepatocellular carcinoma and its surrounding liver tissue, studied using a tissue microarray. *J. Pathol.* **197**, 610–616 (2002).

28. Zheng, C. *et al.* Landscape of infiltrating T cells in liver cancer revealed by single-cell sequencing. *Cell* **169**, 1342–1356 (2017).

29. Cheng, Y. *et al.* Multifactorial heterogeneity of virus-specific t cells and association with the progression of human chronic hepatitis b infection. *Science immunology* **4**, eaau6905 (2019).

30. Bentzen, A. K. *et al.* T cell receptor fingerprinting enables in-depth characterization of the interactions governing recognition of peptide–MHC complexes. *Nature biotechnology* **36**, 1191 (2018).

31. Zhang, S.-Q. *et al.* High-throughput determination of the antigen specificities of T cell receptors in single cells. *Nature biotechnology* **36**, 1156 (2018).

32. Salimbeni, H. & Deisenroth, M. Doubly stochastic variational inference for deep Gaussian processes. In *Advances in Neural Information Processing Systems*, 4588–4599 (2017).

33. Li, H. *et al.* Dysfunctional cd8 t cells form a proliferative, dynamically regulated compartment within human melanoma. *Cell* **176**, 775–789 (2019).

34. Azizi, E. *et al.* Single-cell map of diverse immune phenotypes in the breast tumor microenvironment. *Cell* **174**, 1293–1308 (2018).

35. Guo, X. *et al.* Global characterization of t cells in non-small-cell lung cancer by single-cell sequencing. *Nature Medicine* **24**, 978 (2018).

36. Zhang, L. *et al.* Lineage tracking reveals dynamic relationships of t cells in colorectal cancer. *Nature* **564**, 268 (2018).

37. Sade-Feldman, M. *et al.* Defining t cell states associated with response to checkpoint immunotherapy in melanoma. *Cell* **175**, 998–1013 (2018).

38. Emerson, R. O. *et al.* Immunosequencing identifies signatures of cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire. *Nature genetics* **49**, 659 (2017).

39. Savola, P. *et al.* Somatic mutations in clonally expanded cytotoxic T lymphocytes in patients with newly diagnosed rheumatoid arthritis. *Nature Communications* **8**, 15869 (2017).

40. Tumeh, P. C. *et al.* PD-1 blockade induces responses by inhibiting adaptive immune resistance. *Nature* **515**, 568 (2014).

41. Allison, T. J., Winter, C. C., Fournié, J.-J., Bonneville, M. & Garboczi, D. N. Structure of a human $\gamma\delta$ T-cell antigen receptor. *Nature* **411**, 820 (2001).

42. Giguere, S., Marchand, M., Laviolette, F., Drouin, A. & Corbeil, J. Learning a peptide-protein binding affinity predictor with kernel ridge regression. *BMC bioinformatics* **14**, 82 (2013).

43. Henikoff, S. & Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences* **89**, 10915–10919 (1992).

44. Rasmussen, C. E. & Williams, C. K. I. *Gaussian processes for machine learning* (The MIT Press, 2006).

45. Snelson, E. & Ghahramani, Z. Sparse Gaussian processes using pseudo-inputs. In *Advances in neural information processing systems*, 1257–1264 (2006).

46. Hensman, J., Matthews, A. G. d. G. & Ghahramani, Z. Scalable variational Gaussian process classification. In *Artificial Intelligence and Statistics* (2015).

47. Blei, D. M., Kucukelbir, A. & McAuliffe, J. D. Variational inference: A review for statisticians. *Journal of the American Statistical Association* **112**, 859–877 (2017).

48. Hegde, P., Heinonen, M. & Kaski, S. Variational zero-inflated Gaussian processes with sparse kernels. In *Uncertainty in Artificial Intelligence* (2018).

49. Nickisch, H. & Rasmussen, C. Approximations for binary Gaussian process classification. *Journal of Machine Learning Research* **9**, 2035–2078 (2008).

50. Matthews, A. G. d. G. *et al.* GPflow: A Gaussian process library using TensorFlow. *Journal of Machine Learning Research* **18**, 1–6 (2017). URL `http://jmlr.org/papers/v18/16-537.html`.

51. Opper, M. & Archambeau, C. The variational Gaussian approximation revisited. *Neural computation* **21**, 786–792 (2009).

52. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature biotechnology* **36**, 411 (2018).