ARTICLES
https://doi.org/10.1038/s41590-022-01129-x

nature immunology

Repertoire analyses reveal T cell antigen receptor sequence features that influence T cell fate

Kaitlyn A. Lagattuta[1,2,3,4,5], Joyce B. Kang[1,2,3,4,5], Aparna Nathan[1,2,3,4,5], Kristen E. Pauken[6,7], Anna Helena Jonsson[3], Deepak A. Rao[3], Arlene H. Sharpe[5,6,7], Kazuyoshi Ishigaki[1,2,5,8,10] ✉ and Soumya Raychaudhuri[1,2,3,4,5,9,10] ✉

T cells acquire a regulatory phenotype when their T cell antigen receptors (TCRs) experience an intermediate- to high-affinity interaction with a self-peptide presented via the major histocompatibility complex (MHC). Using TCRβ sequences from flow-sorted human cells, we identified TCR features that promote regulatory T cell (T_{reg}) fate. From these results, we developed a scoring system to quantify TCR-intrinsic regulatory potential (TiRP). When applied to the tumor microenvironment, TiRP scoring helped to explain why only some T cell clones maintained the conventional T cell (T_{conv}) phenotype through expansion. To elucidate drivers of these predictive TCR features, we then examined the two elements of the T_{reg} TCR ligand separately: the self-peptide and the human MHC class II molecule. These analyses revealed that hydrophobicity in the third complementarity-determining region (CDR3β) of the TCR promotes reactivity to self-peptides, while TCR variable gene (*TRBV* gene) usage shapes the TCR's general propensity for human MHC class II-restricted activation.

During T cell development, T_{reg} cells acquire their suppressive phenotype when the affinity of their TCR to the peptide–MHC complex (pMHC) is intermediate to high. In most cases, randomly rearranged *V*, *D* and *J* genes produce a TCR with too low an affinity to pMHC, so most developing T cells do not survive positive selection in the thymus ('death by neglect'). However, TCRs with too strong of an affinity to pMHC result in negative selection by T cell apoptosis. For the T cells that survive both positive and negative selection, however, a divergence in phenotype emerges: those whose TCRs have lower affinity to pMHC tend to become T_{conv} cells, and those whose TCRs have higher affinity tend to gain the T_{reg} phenotype[1–8]. Following thymic selection, a crucial prerequisite for the peripheral induction of T_{reg} cells is suprathreshold affinity to pMHC, although other factors, such as co-stimulatory signals, exert additional influence[7,9].

The body of evidence that regulatory versus conventional T cell phenotypes are largely driven by TCR signal strength suggests that the developmental fate of $CD4^+$ T cells may be influenced by sequence features of the TCR. Indeed, the degree of overlap in TCR sequence between T_{reg} cells and T_{conv} cells is minimal compared to T cell samples of the same phenotype[10]. The distinguishing features of T_{reg} and T_{conv} TCRs could shed light on the determinants of TCR strength, but the majority of extant work has focused on exact sequence matching rather than generalizable TCR sequence features.

To identify all sequence features that influence TCR strength, we examined 5.7×10^7 TCRβ chain sequences from six published datasets. Using multiple mixed effects logistic regression models, we quantified the effect of each TCR feature on T_{reg} fate and aggregated these results into a TiRP score that can be applied to any TCR. Our work reveals that the TCR sequence consistently informs T cell fate and function across diverse biological contexts, including the fetal thymus and tumor microenvironment.

Results

Study design. We first derived a comprehensive collection of TCR features (Supplementary Table 1) by examining the mutual information (MI) structure of the TCR amino acid sequence. We then tested each sequence feature for differential abundance between T_{reg} cells and T_{conv} cells in two human cohorts of TCRβ chains from flow-sorted T cells[11,12] (Supplementary Table 2). From these results, we developed a T_{reg} propensity scoring system for the TCR (TiRP) (Fig. 1a). Upon confirming its accuracy in two datasets of thymic T cells[13,14], we applied TiRP to tumor-infiltrating T cells and found that clone plasticity (the presence of induced T_{reg} cells (iT_{regs}) or exT_{reg}[15] cells; Fig. 1b) corresponded to a significantly high TiRP score. Finally, to shed light on the etiology of the observed TCR sequence biases, we separately examined the two elements of the T_{reg} TCR ligand, the self-peptide and the human MHC class II molecule. For these analyses, we calculated human TiRP for (1) murine T_{reg} cells and (2) human memory T_{conv} cells, respectively (Fig. 1c). These results demonstrated two separable components of TiRP: CDR3β hydrophobicity promotes reactivity to self-peptides, while the *TRBV* gene shapes the TCR's general activatability in the context of human MHC class II restriction.

Defining features of the TCR sequence. The TCR is a membrane-anchored heterodimeric protein consisting of an α-

[1]Center for Data Sciences, Brigham and Women's Hospital, Boston, MA, USA. [2]Division of Genetics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA. [3]Division of Rheumatology, Inflammation, and Immunity, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA. [4]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA. [5]Broad Institute of MIT and Harvard, Cambridge, MA, USA. [6]Department of Immunology, Blavatnik Institute, Harvard Medical School, Boston, MA, USA. [7]Evergrande Center for Immunologic Diseases, Harvard Medical School and Brigham and Women's Hospital, Boston, MA, USA. [8]Laboratory for Human Immunogenetics, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan. [9]Centre for Genetics and Genomics Versus Arthritis, Manchester Academic Health Science Centre, University of Manchester, Manchester, UK. [10]These authors contributed equally: Kazuyoshi Ishigaki, Soumya Raychaudhuri. ✉e-mail: kazuyoshi.ishigaki@riken.jp; soumya@broadinstitute.org

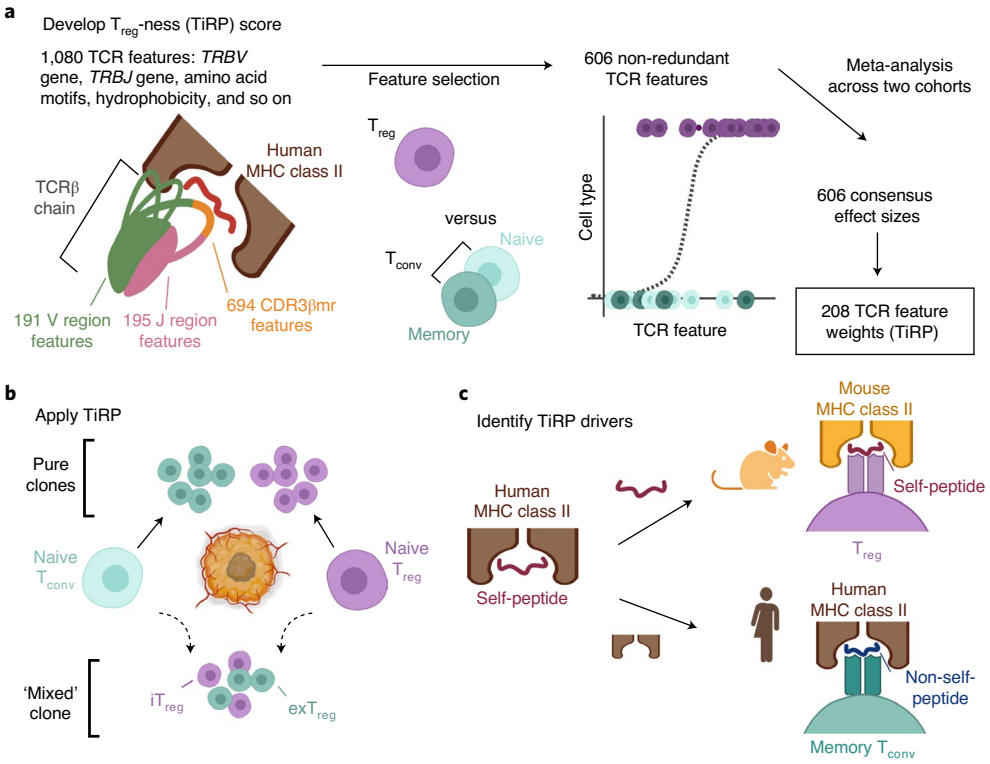


Fig. 1 | Study design. a, We first examined the structure of the TCR sequence to define 1,080 sequence features. Depicted is a TCR β chain in complex with antigenic peptide (red) and human MHC class II molecules (brown). The TCR is colored by region. The V region (including CDR1β and CDR2β loops) is in green, the CDR3β middle region (CDR3βmr) is in orange and the J region is in pink. We used MI analysis and mixed effects model comparisons to select 606 non-redundant TCR features that best explained variance in T cell state. We fit mixed effects logistic regression models for 70% of the data in the discovery and replication cohorts separately and combined the effect sizes for each TCR feature across the two cohorts by meta-analysis. TiRP was calibrated to include only 208 of the 606 TCR features that had Bonferroni-significant meta-analytic *P* values. **b**, We then applied TiRP to tumor-infiltrating CD4$^+$ cells to study mixed clones: groups of T_{reg} cells and T_{conv} cells with the same *TRB* and *TRA* nucleotide sequences observed in the same individual. These mixed clones likely represent lineages of T cells that have undergone a peripheral conversion between the regulatory and conventional phenotypes. Such clones may include iT_{reg} cells (T_{conv} cells that have acquired a regulatory phenotype), exT_{reg}[15] cells (T_{reg} cells that have lost the regulatory phenotype), or both. **c**, Finally, we investigated the drivers of TiRP by separately examining the two elements of the human T_{reg} TCR ligand, the self-peptide and the human MHC class II molecule. Figure created with BioRender.com.

and a β-chain. Each of the two chains includes three highly variable peptide loops that protrude toward the pMHC complex. The most variable of these loops is the CDR3β region in the β chain, which mediates recognition of specific antigens. Because *TRBV*, *TRBD* and *TRBJ* genes each encode a region of CDR3β, we anticipated that the CDR3β sequence would consist of blocks of strongly correlated residues. To determine the boundaries of these regions, we examined the MI structure of CDR3β peptides in a previously published cohort of targeted TCR sequencing in multiple tissues and peripheral blood mononuclear cells (PBMCs)[11] ('discovery cohort'; Supplementary Table 2). To assess generalizability of any findings, we held out data from six randomly selected donors (Methods).

MI calculations between CDR loop residues revealed three distinct regions of the TCR: the V region (International ImMunoGeneTics information system (IMGT) position 1 (p1)–p107), the CDR3β middle region (CDR3βmr, p108–p112) and J region (p113–p118) (Fig. 2a,b and Extended Data Fig. 1a–g). While random nucleotide insertions in the highly variable CDR3βmr obscured the identity of the *TRBD* gene, the germline-encoded V and J regions demonstrated sequence conservation and high inter-residue MI (Fig. 2a). MI was concentrated at the flanking ends of CDR3β such that 8 p104–p106 tripeptides ('Vmotifs') and 42 p114–p118 pentapeptides ('Jmotifs') accounted for >90% of observations. After observing minimal MI between the three regions, we elected to undertake a three-pronged modeling approach, in which we would examine the V, middle and J regions independently.

T_{reg} cells use specific amino acids in the CDR3βmr. We first examined the CDR3βmr of T_{reg} cells (CD4$^+$CD127$^-$CD25$^+$) and T_{conv} cells (CD4$^+$CD127$^+$) in the discovery cohort. Calculating the mean percentage of CDR3βmr residues occupied by each amino acid yielded strikingly consistent T_{reg}–T_{conv} differences across donors: phenylalanine, leucine, tryptophan and tyrosine were consistently enriched in T_{reg} cells, while aspartic acid and glutamic acid were consistently enriched in T_{conv} cells (Fig. 3a). Categorization of amino acids by physicochemical features showed that hydrophobic amino acids were enriched in T_{reg} cells, while negatively charged amino acids were enriched in T_{conv} cells (Extended Data Fig. 1h).

To quantify these effects, we used forward selection to build a statistical model that increased in complexity (degrees of freedom, d.f.) with the addition of each TCR feature. We observed that 15 amino acid features had an independent effect on T_{reg} fate, each affording an incremental gain in variance explained (Fig. 3b, middle, and Supplementary Table 3). At each step, we used nested conditional mixed effects logistic regression to account for interindividual differences, such as those driven by human leukocyte antigen (*HLA*) genotype and tissue source (Methods).

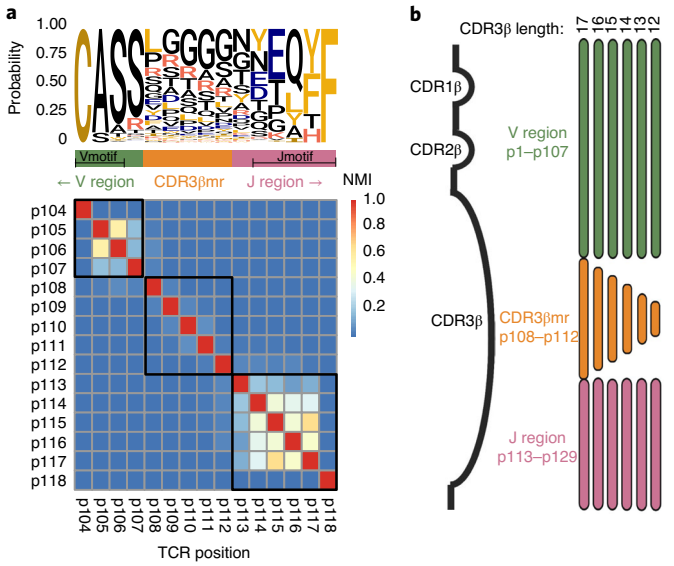


Fig. 2 | TCR sequence structure. a, Probability of each amino acid in each CDR3β position depicted by a sequence logo, with a heat map of normalized MI (NMI) between each pair of CDR3β residues for the most frequent CDR3β length, 15 amino acids. Based on this MI structure, we partitioned the CDR3β sequence into a Vmotif within a V region, a CDR3βmr and a Jmotif within a J region. **b**, Schematic showing TCRs of multiple lengths aligned to the TCR β-chain structure. Three CDRs within the TCR β-chain protrude as loops into the pMHC–TCR complex: CDR1β, CDR2β and CDR3β. CDR1β and CDR2β are encoded by the *TRBV* gene, while CDR3β spans *TRBV*-encoded residues, random nucleotide insertions (CDR3βmr) and *TRBJ*-encoded residues. Random nucleotide insertions from VDJ recombination occur at the V/D and D/J junctions, creating variation in CDR3βmr length. Regions suggested by MI structure are not drawn to scale.

To confirm that these effects were consistent across donors and clinical phenotypes, we estimated them in each of the 18 individuals and in the type 1 diabetes (T1D) and healthy subsets of the discovery cohort separately. We found consistent effect sizes in all contexts (Extended Data Fig. 2a,b, Supplementary Table 3 and Methods). We compared this model to an alternative approach in which CDR3βmr was scored by physicochemical features (hydrophobicity, isoelectric point (pI) and volume) rather than percentages of individual amino acid residues (Supplementary Table 4 and Methods). Physicochemical features did not capture as much information as amino acid percentages (Fig. 3b, middle); hence, we proceeded with an amino acid-based model of the CDR3βmr.

We then ran a separate mixed effects model for each CDR3βmr position (IMGT p108–p112), testing whether the amino acid at the given position explained variance in T cell fate beyond that accounted for by the CDR3βmr amino acid percentages (Methods). We found that each position indeed conveyed additional information regarding the likelihood of T_{reg} fate, but these position-specific effects all together did not explain as much variance as the general amino acid composition of the CDR3βmr (Fig. 3c and Supplementary Table 5).

CDR3β V and J regions explain variance in T cell state. We then examined the V region of the TCR. Previous studies have established that genetic variation in the MHC locus shapes the frequency with which *TRAV*/*TRBV* genes are used in the T cell repertoire[16]. MHC polymorphisms explain far more variance in *TRAV* gene usage than *TRBV*[16], consistent with protein structure data demonstrating that TRAV contacts MHC at polymorphic sites while TRBV contacts MHC at conserved sites[17]. We hypothesized that variation in *TRBV*-encoded residues may alter TCR affinity to these conserved MHC sites and thereby influence T cell fate.

To test this hypothesis, we extracted sequence features from the V region and tested their association with T_{reg} fate using mixed effects logistic regression (Methods). In consideration of multicollinearity, we computed all pairwise correlations between V region TCR features and avoided joint modeling of TCR features with any $|r| > 0.7$ (Extended Data Fig. 3 and Methods). Through model comparisons, we found that a joint model including *TRBV* gene identity and p107 best represented the region, because the 58 *TRBV* genes explained far more variance than the eight Vmotifs (Fig. 3b left and Methods). To account for interindividual variation in *TRBV* gene selection, we included a thymic selection parameter (V gene selection rate, VGSR) for each *TRBV* gene as a covariate (Supplementary Note and Extended Data Fig. 4). Despite adjusting for VGSR, *TRBV* gene usage continued to explain a significant amount of variance in T cell fate, with three *TRBV* genes reducing the odds of T_{reg} fate by more than 30% compared to the reference (most common) gene *TRBV05-01* ($P = 1.3 \times 10^{-804}$, likelihood ratio test (LRT); Supplementary Table 6). As in the CDR3βmr analysis, we confirmed that these associations replicated in models isolated to each individual and to both T1D and healthy cohort subsets (Extended Data Fig. 2c,d and Supplementary Table 6). The consistency in *TRBV* gene effects across individuals suggests that their influence on T_{reg} fate indeed occurs through interactions with conserved MHC residues and is largely independent of MHC variability between individuals.

We then examined the J region with the same approach. In contrast to the V region, wherein strong p104–p106 sequence conservation constrained multiple *TRBV* genes to the same Vmotif, variable nucleotide editing at the *D*/*J* junction resulted in multiple Jmotifs associated with each *TRBJ* gene. The 42 Jmotifs explained slightly more variance than the 13 *TRBJ* genes (Fig. 3b, right), so we proceeded with a joint model containing the Jmotif and p113 residue. Across six CDR3β lengths, the most important TCR features for T cell fate determination were the *TRBV* gene identity and the percent composition of amino acids in the CDR3βmr (Fig. 3c). Each TCR region played an important role, with the greatest variance explained per residue in the CDR3βmr. Relative gains in variance explained were proportional to fractional occupancy of the TCR, which was dependent on CDR3β length (Fig. 3d and Methods). To compare these results to a null model, we conducted 1,000 permutations of the cell-type labels and confirmed that the observed amount of variance explained far exceeded the distribution in the null model (Supplementary Table 7 and Methods). To assess whether these results were mediated by invariant TCRs such as those of invariant natural killer T (iNKT) cells, we excluded putative iNKT cell receptors from the data and observed minimal changes in TCR feature effect sizes (Supplementary Table 8 and Methods). Thus, our reported effects are statistically well calibrated and robust to niche or invariant TCRs.

T_{reg} cells are enriched for CDR1β charge and CDR3β hydrophobicity. We next aimed to localize physicochemical effects underlying CDR3βmr residue enrichments to specific TCR positions. At each CDR1β–CDR3β loop amino acid position, we estimated the effect of hydrophobicity, pI and volume on T_{reg} fate using a ridge regression model (Supplementary Table 9 and Methods). Intriguingly, these results provided a physicochemical basis for some of the *TRBV* gene differences observed. T_{reg} cells were enriched for positively charged amino acids at p37 of CDR1β (Fig. 4a). Seven *TRBV* genes assessed in our models harbor a negatively charged residue at p37; all seven of these were significantly depleted for T_{reg} cells compared to the reference gene *TRBV05-01*, which has a positively charged arginine at p37 (Fig. 4b). As expected from our earlier findings, CDR3βmr featured positive coefficients for hydrophobicity in every position (Fig. 4a). At each position, 1-s.d. increase in hydrophobicity

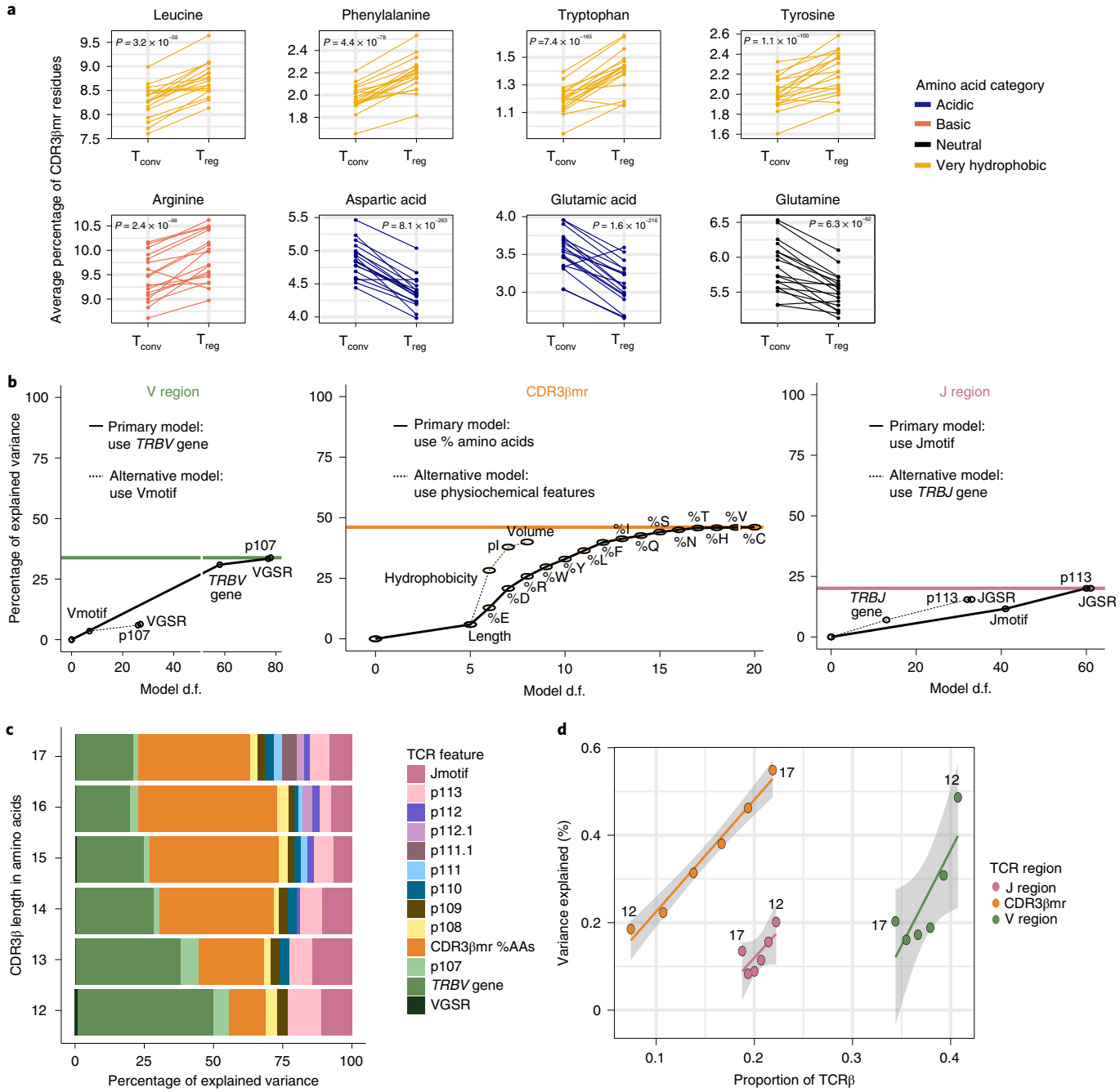


Fig. 3 | Broad differences exist between the TCRs of T_{reg} cells and T_{conv} cells. a, Percentage of select amino acids in the CDR3βmr, plotted as the mean for each donor sample in the discovery cohort, separated by cell type and colored by amino acid groups. *P* values are computed by a two-sided Wald test on the coefficient for each amino acid term in a mixed effects logistic regression model (Methods). **b**, Incremental variance explained by the addition of labeled TCR features to the V region (left), CDR3βmr (middle) and J region (right) mixed effects logistic regression models. The addition of each TCR feature increased model complexity by adding 1 d.f. for each quantitative feature and $k-1$ d.f. for each qualitative feature, where k is equal to the number of possible values for the qualitative feature ($k=58$ for 58 possible *TRBV* genes; $k=8$ for 8 possible Vmotifs). For each region, the primary modeling approach was compared to the alternative modeling approach, and the modeling approach that explained greater variance was selected. Colored horizontal lines depict the total percent of explained variance attributable to each TCR region, summing to 100%. **c**, Percentage of explained variance by each TCR feature type, summing to 100% for each length of CDR3β. **d**, Variance explained by each TCR region for different CDR3β lengths. As CDR3β length increases, CDR3βmr occupies a greater proportion of the TCR (fraction of amino acid residues) at the expense of V and J region proportions. Select CDR3β lengths (number of amino acids) are labeled to show the direction of these trends. The *x* axis corresponds to the proportion of TCR β-chain amino acids derived from the V, J and middle regions (summing to 100 for each CDR3β length; Methods), while the *y* axis corresponds to the absolute variance explained (scale, 0–100%). A line of best fit is drawn for each TCR region; the 95% confidence interval (95% CI) is shaded in gray VGSR, V gene selection rate (Supplementary Note); CDR3βmr %AAs, percent composition of amino acids in the CDR3βmr.

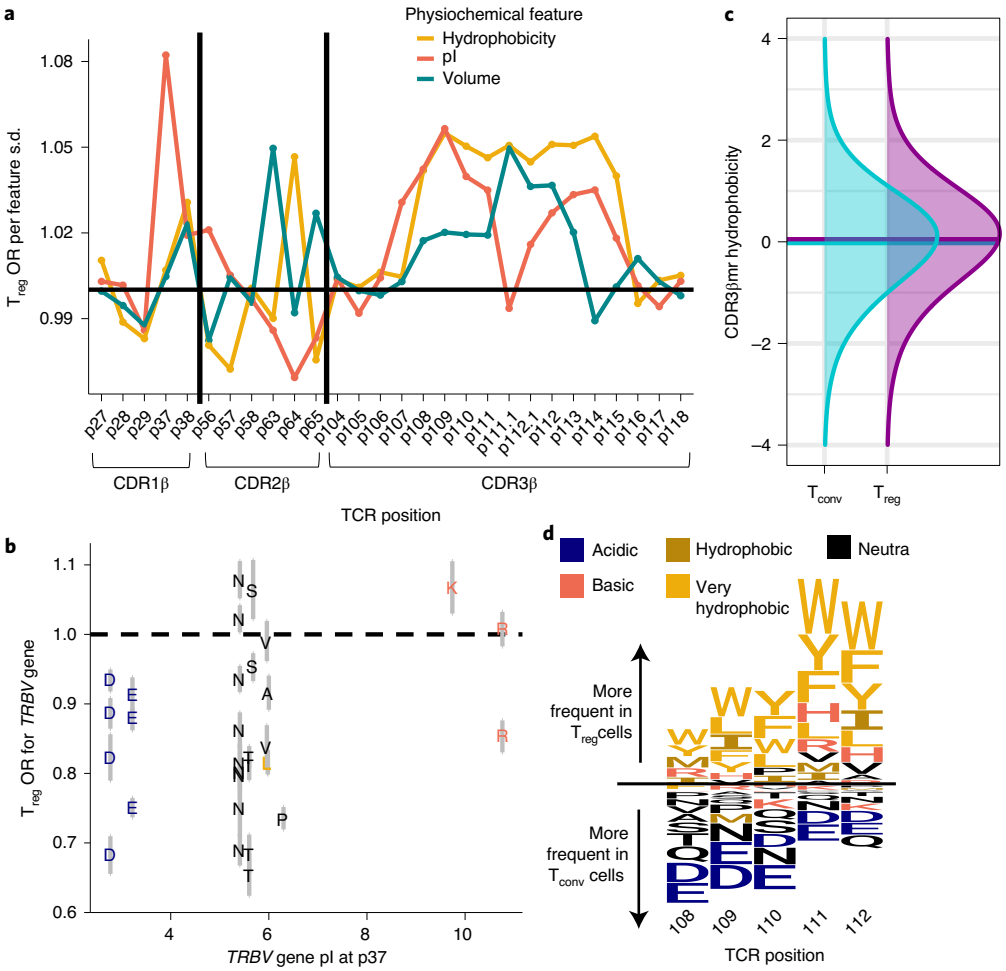


Fig. 4 | T_{reg} cells exhibit position-specific TCR sequence features. a, Estimated OR (per s.d.) for each physicochemical feature at each CDR1β–CDR3β loop position; features with an estimate >1 are positively associated with T_{reg} fate, while features with an estimate <1 are negatively associated. ORs denote the change in T_{reg} odds per s.d. increase in the given physicochemical feature at the given TCR position. Within each CDR3β length, all effects were estimated jointly via L2-regularized logistic regression with a penalty weight tuned via tenfold cross-validation (Methods). Shown are the OR estimates for each positional feature averaged across the six CDR3β lengths. Vertical lines denote the boundaries of each CDRβ loop. **b**, Correspondence between *TRBV* gene pI at p37 (apex of CDR1β) and *TRBV* gene OR for T_{reg} fate compared to the reference gene *TRBV05-01*. Each *TRBV* gene is labeled with its amino acid residue at p37 and the 95% CI for the OR. **c**, Distribution of CDR3βmr hydrophobicity in T_{conv} cells compared to T_{reg} cells in the discovery dataset. Hydrophobicity values are averaged over the CDR3βmr for each TCR and then scaled to have a mean of 0 and a variance of 1. Horizontal lines depict the mean for each population (T_{reg} mean CDR3βmr hydrophobicity = 0.05; T_{conv} mean hydrophobicity = −0.03; Wald test $P = 2.3 \times 10^{-523}$). **d**, Sequence logo depicting the effects of amino acids in the highly entropic CDR3βmr residues, sized proportionally to the associated change in T_{reg} odds, with amino acids more frequent in T_{reg} cells above the horizontal line and amino acids more frequent in T_{conv} cells below.

led to a 2.5% (L17, p113) to 6.3% (L12, p113) increase in odds of T_{reg} fate (odds ratio (OR) = 1.025, 95% CI = 1.011–1.039 and Wald test $P = 2.7 \times 10^{-4}$ for L17, p113; OR = 1.063, 95% CI = 1.051–1.074 and Wald test $P = 5.2 \times 10^{-28}$ for L12, p113; Extended Data Fig. 5 and Supplementary Table 9). Although highly consistent across samples, this effect was subtle: average CDR3βmr hydrophobicity was 0.08 s.d. higher in T_{reg} cells than in T_{conv} cells (Fig. 4c; OR = 1.08, 95% CI = 1.076–1.083, Wald test $P = 2.3 \times 10^{-523}$). Sensitivity analyses revealed that p37 charge and CDR3βmr hydrophobicity effects were relatively robust to the weight of the ridge penalty term (Supplementary Table 10). Interestingly, statistical interactions between physiochemical values at different TCR residues were largely insignificant except for a few relating to bulky adjacent amino acids (Methods and Supplementary Table 11).

To directly visualize the amino acids associated with T_{reg} fate, we generated a sequence logo representation of the CDR3βmr based on differential amino acid usage at each position (Fig. 4d and Methods). Our results are consistent with previous findings suggesting that hydrophobicity at p109 and p110 promotes the development of T cells that recognize self-antigens[18]. Importantly, we show that this principle extends beyond p109–p110 throughout the stretch of CDR3βmr residues. Thus, randomly recombined TCR amino acids play a parsimonious role in T cell fate acquisition; increasing hydrophobicity raises affinity to self-pMHC and thereby promotes T_{reg} development.

Reproducing TCR associations in an independent dataset. Having identified TCR features associated with T_{reg} identity, we next sought to validate them in a public dataset of TCRβ sequences from sorted T_{reg} ($CD4^+CD25^{high}CD127^{low}$) and T_{conv} ($CD4^+CD25^{low}CD27^+$) cells sampled from the peripheral blood of 16 donors[12] ('replication cohort'; Supplementary Table 2). Despite a different distribution of

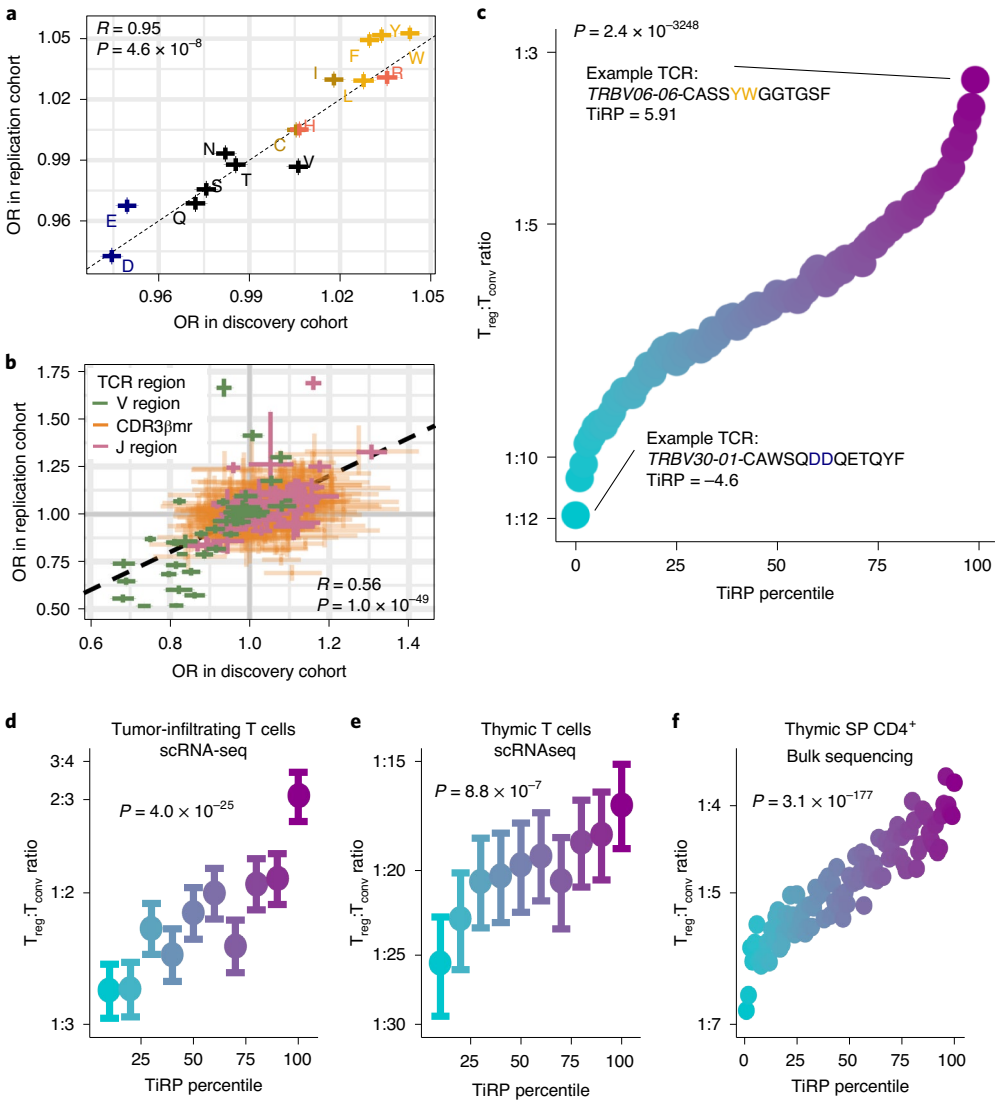


Fig. 5 | T_{reg} TCR sequence biases replicate in independent cohorts. a, Correspondence between the discovery and replication cohort ORs for CDR3βmr compositional amino acids; OR corresponds to the change in T_{reg} odds associated with 1-s.d. increase in CDR3βmr percentage for a given amino acid. Colors for amino acids correspond to Extended Data Fig. 1h. **b**, Comparison in **a** for all other TCR sequence features; OR corresponds to the change in T_{reg} odds associated with the presence of the given feature compared to the reference feature (Supplementary Table 1). For **a** and **b**, *R* represents the Pearson's correlation coefficient, and *P* values are computed by a two-sided *t*-test with Fischer transformation. **c**, Validation of the TiRP score in held-out donors of the discovery and replication datasets ($n = 3{,}277{,}036$ TCRs). Each s.d. increase in TiRP was associated with a 23% increase in the odds of T_{reg} status (OR = 1.231, 95% CI = 1.227–1.235, LRT $P = 2.4 \times 10^{-3{,}248}$). Percentile points are colored by T_{reg}:T_{conv} ratio ranging from blue (lowest) to purple (highest). **d**, Validation of TiRP in single-cell RNA sequencing (scRNAseq) of CD4$^+$ tumor microenvironment T cells[19,20] ($n = 27{,}721$ cells). Each unit increase in TiRP (corresponding to 1 s.d. for the scores in **c**) was associated with a 16% increase in the odds of T_{reg} status (OR = 1.16, 95% CI = 1.13–1.19, LRT $P = 4.0 \times 10^{-25}$). **e**, Validation of TiRP in human thymic T cells[13] ($n = 60{,}424$ cells). Among developing thymocytes, each unit increase in TiRP was associated with a 9% increase in the odds of T_{reg} fate (OR = 1.09, 95% CI = 1.05–1.13, LRT $P = 8.8 \times 10^{-7}$). For **d** and **e**, error bars outline 95% CIs for T_{reg}:T_{conv} odds in each TiRP score decile computed by bootstrap resampling (Methods). **f**, Validation of TiRP in TCR-targeted genomic DNA (gDNA) sequencing from grafted human thymi of humanized mice[14] ($n = 466{,}551$ TCRs). Each unit increase in TiRP was associated with a 12% increase in the odds of T_{reg} status (OR = 1.12, 95% CI = 1.11–1.12, LRT $P = 3.1 \times 10^{-177}$).

tissue sources in this dataset, the CDR3βmr amino acid percentage effects were nearly identical (Pearson $R = 0.95$, $P = 4.6 \times 10^{-8}$; Fig. 5a and Supplementary Table 3). Effects for individual *TRBV* genes and Jmotifs and position-specific amino acid effects were also consistent with discovery (Pearson $R = 0.56$, $P = 7.5 \times 10^{-57}$; Fig. 5b, Supplementary Tables 5 and 6 and Methods). In the replication cohort, *TRB* sequences were collected by reverse transcription and amplification of RNA rather than direct DNA sequencing. Thus, relative changes in T_{reg} likelihood induced by these TCR sequence features are robust not only to different tissue sources but also to technical differences in sorting and sequencing protocols.

Developing TiRP: a T_{reg} propensity score for the TCR. Having replicated the effect of a comprehensive set of TCR features in two independent cohorts, we next developed a method to quantify the TiRP of a T cell. Briefly, for a given TCR, TiRP is the sum of T_{reg} association effect sizes of independent sequence features in all three TCR regions (Methods). We used meta-analytic effect size

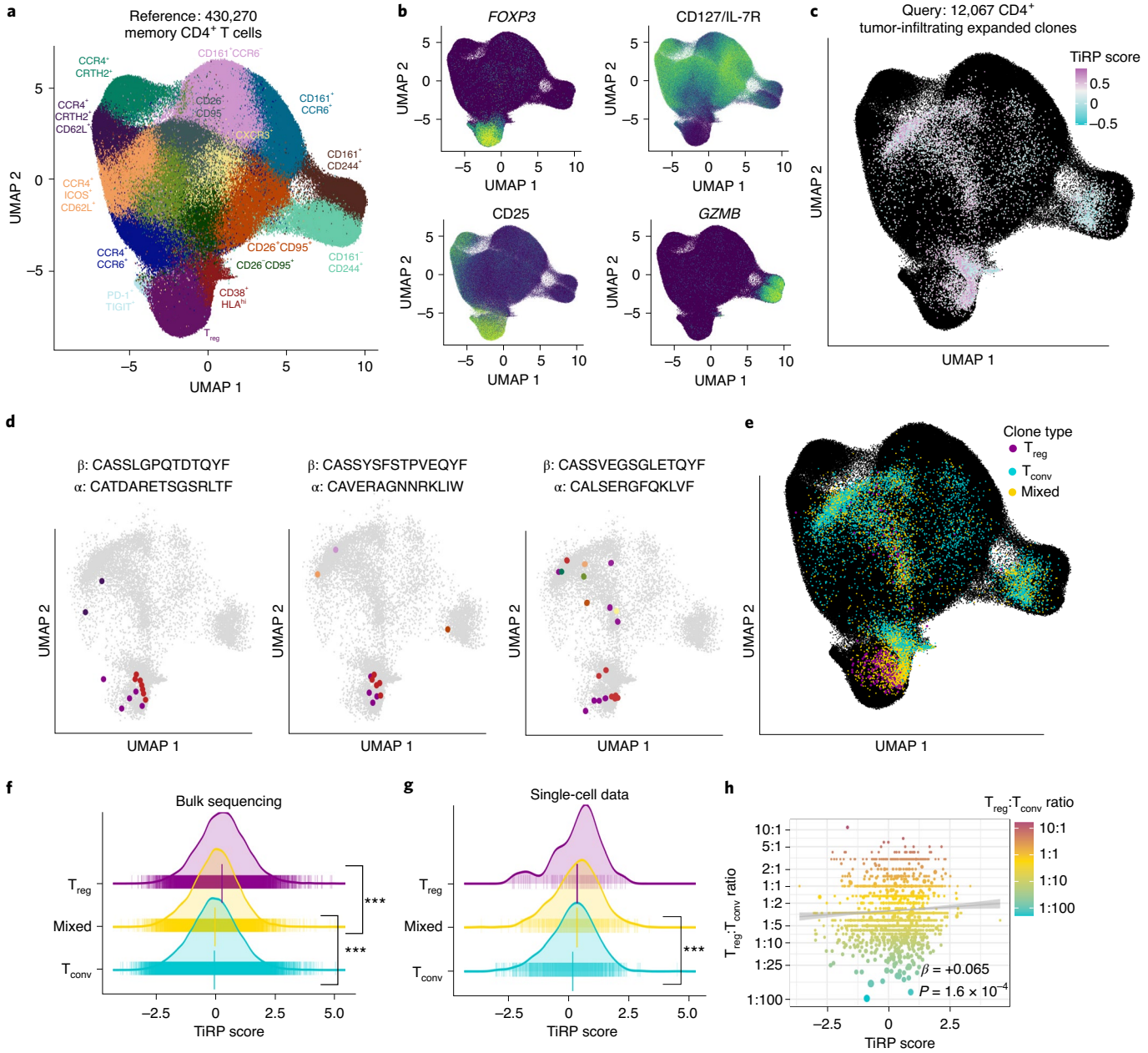


Fig. 6 | TiRP helps to explain clonal plasticity in the tumor microenvironment. a, Reference T cell dataset, colored by cell-type clusters according to transcriptional and surface marker variation depicted in Extended Data Fig. 7c,d. **b**, Select gene expression (*FOXP3* and *GZMB*) and surface marker abundance (CD25 and CD127) for cells in the reference T cell dataset; low, purple; high, light green). **c**, Tumor microenvironment T cells of expanded clones mapped into the reference embedding by Symphony. Each cell is colored by the TiRP score of its paired TRB chain, with *k* nearest neighbor smoothing for visualization (Methods). TiRP is scaled such that 0 corresponds to the mean score, and 1 unit corresponds to 1 s.d. of held-out bulk sequencing TCRs (Fig. 5c). **d**, Cell members of three example mixed clones are highlighted in color according to their cell-type classification by Symphony (colors as in **a**). Within a given plot, each cell expresses the same *CDR3β* DNA sequence and the same CDR3α amino acid sequence and was observed within the same donor (the CDR3β amino acid sequence is listed above the CDR3α amino acid sequence for each). **e**, Same as **c**, with each tumor-infiltrating T cell colored according to clone type: purple for clones containing only T_{reg} cells, blue for clones containing only T_{conv} cells and yellow for clones containing both T_{reg} and T_{conv} cells ('mixed' clones). **f**, TiRP scores of T_{conv}, T_{reg} and mixed expanded clones from held-out bulk sequencing data; $P = 2.0 \times 10^{-40}$ for the mixed–T_{conv} difference and $P = 9.1 \times 10^{-16}$ for the mixed–T_{reg} difference. **g**, Scores as in **f** for tumor-infiltrating scRNAseq data; $P = 3.0 \times 10^{-4}$ for mixed–T_{conv} difference and $P = 0.55$ for mixed–T_{reg} difference. For **f** and **g**, vertical bars denote mean and s.e.m. per clone type. **h**, Correspondence between TiRP score and the T_{reg}:T_{conv} ratio for each clone. The best fit line is shown in gray; clones are colored by T_{reg}:T_{conv} ratio and sized proportionally to the number of constituent cells. β corresponds to the slope of the regression line between the log transform of the T_{reg}:T_{conv} ratio and TiRP score. For **f**–**h**, *P* values are computed by the LRT between mixed effects logistic regression models (Methods).

estimates across the two cohorts and included only features with a significant effect on T cell fate based on a Bonferroni *P* value threshold (Methods). As a result, TiRP is the weighted sum of 25 *TRBV* genes, 23 Jmotifs, 4 CDR3β lengths, 14 CDR3βmr amino acid percentages and 142 positional amino acids (Supplementary Table 12).

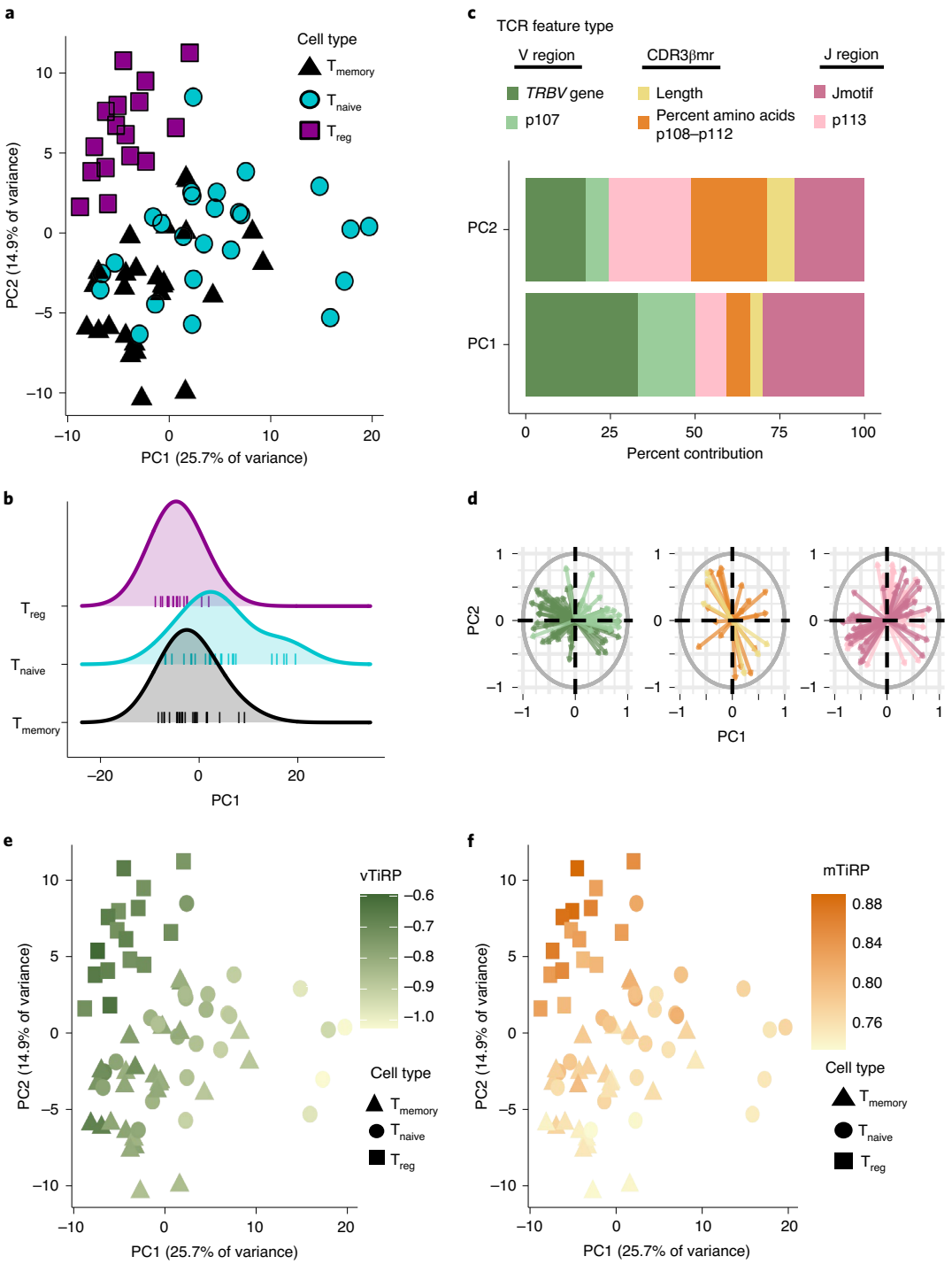


Fig. 7 | Two axes of TCR-driven cell states. a, Sixty-seven samples from the replication cohort colored by cell type and arranged in principal component (PC) space according to variation in TCR sequence feature frequencies (Methods). **b**, Distribution of PC1 embeddings for each cell type; each vertical line corresponds to one sample. Naive T_{conv} cells have the highest PC1 embedding in 15 of the 16 donors with all three cell types available. *P* value is computed by the binomial test with $n=16$ and $k=15$. **c**, Percent contribution of each type of TCR sequence feature to the first two PCs. **d**, Loadings of each of the TCR sequence features on PC1 and PC2, depicted by arrows, separated by TCR region and colored by the same scheme as in **c**. **e**, Samples arranged in PC space as in **a**, colored by mean TiRP in the V region of the TCR (vTiRP). **f**, Same as in **e**, colored by mean TiRP in the CDR3βmr (mTiRP). *P* values for **e**–**f** are calculated by a two-sided *t*-test with Fischer transformation on Pearson's *R*; jTiRP, TiRP of the J region of the TCR (IMGT p113–p118); mTiRP, TiRP of the middle region of the TCR (IMGT p108–p112); vTiRP, TiRP of the V region of the TCR (IMGT p1–p107).

We then tested our TiRP score on the four discovery cohort donors and two replication cohort donors whose repertoire data had been withheld from all former analyses. We observed that a 1-s.d. increase in TiRP in these held-out data resulted in a 23% increase in the odds of T_{reg} status (OR = 1.231, 95% CI = 1.227–1.235, LRT $P = 2.4 \times 10^{-3,248}$; Fig. 5c, Supplementary Table 13 and Methods). TCRs in the highest-scoring decile were more than twice as likely as TCRs in the lowest-scoring decile to belong to a T_{reg}, 1 in every

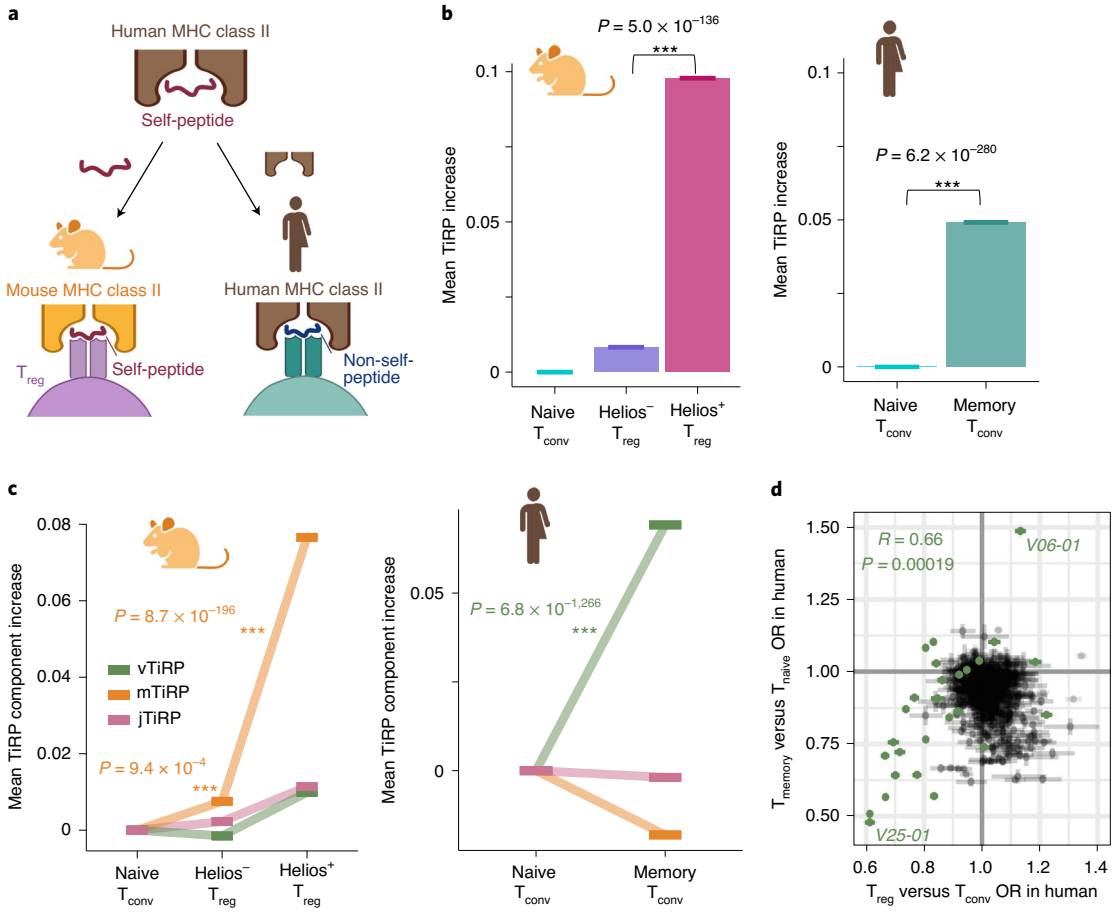


Fig. 8 | Isolating the drivers of TiRP. a, We investigated the drivers of TiRP by separately examining the two elements of the human T_{reg} TCR ligand: the self-peptide and the human MHC class II molecule. To do so, we scored (1) murine T_{reg} TCRs, which share an affinity to mammalian self-peptides but not to human MHC class II molecules, and (2) human memory T_{conv} TCRs, which share an affinity to human MHC class II molecules but not to self-peptides. **b**, Left, mean increase in TiRP score of *Helios*-sorted T_{reg} cells compared to naive T_{conv} cells in *Helios*-green fluorescent protein (GFP) *Foxp3*-red fluorescent protein (RFP) reporter mice[31]. Right, mean increase in TiRP score of memory T_{conv} cells compared to naive T_{conv} cells from held-out donors of the replication dataset. **c**, Left, TiRP score increases in *Helios*-sorted murine T_{reg} cells broken down into TiRP score components by TCR region. Right, TiRP score increase in human memory T_{conv} cells broken down into TiRP score components by TCR region. **d**, Correspondence between TCR feature odds ratios for T_{reg}–T_{conv} odds (*x* axis, meta-analytic odds between discovery and replication cohort) and memory-naive odds (*y* axis, replication cohort only) with their 95% CI values. *TRBV* genes are highlighted in green; *V06-01* indicates *TRBV06-1*; *V25-01* indicates *TRBV25-01*. Pearson's *R* and its corresponding *P* value pertain to *TRBV* gene ORs only. *P* values in **b** and **c** are calculated by the LRT between mixed effects models (Methods); the *P* value in **d** is calculated by a two-sided *t*-test with Fischer transformation on Pearson's *R*. Figure created with BioRender.com.

3.9 compared to 1 in every 9.1. To ensure that this TCR–T cell state covariation was contingent on the biology of surface-expressed TCRs, we repeated this analysis on the non-productive TCRs in the four held-out donors for which out-of-frame reads were available (Methods). This indeed abrogated the association between T_{reg}-ness score and T_{reg} fate (OR = 1.00, 95% CI = 0.97–1.04, LRT $P = 0.96$).

To externally validate our scoring system, we calculated TiRP in four published datasets[13,14,19,20] (Supplementary Table 2). We scored each TCR and assessed whether the TiRP explained variance in T cell phenotype, as defined by standard mRNA clustering for the three scRNAseq cohorts (Methods and Extended Data Figs. 6 and 7a,b) and by CD25 and CD127 flow sorting[14]. Consistent with our previous observations, there was a nearly twofold increase in T_{reg} likelihood in the top TiRP decile compared to the bottom TiRP decile in all cohorts (Fig. 5d–f), including the tumor microenvironment (OR = 1.16 per unit increase in TiRP, 95% CI = 1.13–1.19, LRT $P = 4.0 \times 10^{-25}$; Fig. 5d and Supplementary Table 13). TiRP elevation in thymic T_{reg} cells[13] confirmed the direct relevance of TiRP to the thymus (Fig. 5e; OR = 1.09, 95% CI = 1.05–1.13, LRT $P = 8.8 \times 10^{-7}$). Similar results in TCRs from flow-sorted single positive (SP) CD4$^+$ thymic T cells[14] (Fig. 5f; OR = 1.12, 95% CI = 1.11–1.12, LRT $P = 3.1 \times 10^{-177}$) pinpointed the stage of thymic development in which TiRP promotes T_{reg} fate. Importantly, these SP CD4$^+$ thymocytes include T cells observed before negative selection. Because the T_{reg} population represents a terminal differentiation state in the thymus, young T cells that will be negatively selected are more likely to be observed in the precursor non-regulatory population. Thus, the blunting in TiRP effect size that we observe in thymic data is consistent with high TiRP of T cells that are negatively selected for their affinity to self-pMHC. Evidently, our TCR scoring system describes T_{reg} TCR features in diverse biological contexts, including thymic selection.

TiRP explains T_{reg} plasticity in the tumor microenvironment. We next asked whether TiRP could help to explain T_{reg} plasticity. It is well recognized that naive T_{conv} thymic emigrants can be peripherally induced to adopt a regulatory phenotype[21,22]. Conversely, some T_{reg} cells have been observed to lose *FOXP3* expression and adopt a

proinflammatory phenotype[23–26] (exT_{reg}[15] cells; Fig. 1b). Expanded T cell clones (possessing the same TCR) observed as both T_{reg} cells and T_{conv} cells within the same donor (hereafter referred to as 'mixed clones') represent lineages of T cells that have undergone such peripheral conversions. We hypothesized that the TiRP of these T cells may be intermediate, rendering them most susceptible to peripheral conversion.

Before testing our hypothesis, we used Symphony[27] to standardize cell-type definitions across the two tumor microenvironment datasets[19,20] by mapping cells of expanded clones from both datasets (12,067 cells) into a common reference atlas[28] of T cell states based on joint transcriptional and proteomic profiling (Fig. 6a–c, Supplementary Table 2, Extended Data Figs. 7c,d and 8a–d and Methods). On average, 19.2% of expanded clones from the same donor were observed in both the T_{reg} and T_{conv} states, including a few large clones with a relatively even balance (Fig. 6d,e and Supplementary Table 14).

We next tested whether the TiRP score of mixed clones was in between that of purely T_{conv} and T_{reg} clones (Methods). In the previously held-out bulk sequencing data, the TiRP scores of mixed clones were significantly greater than those of expanded T_{conv} clones and less than those of expanded T_{reg} clones (Fig. 6f, mixed–T_{conv} difference = 0.03, $P = 2.0 \times 10^{-40}$; mixed–T_{reg} difference = –0.29, $P = 9.1 \times 10^{-16}$, LRT; Methods). These single-cell data confirmed that T_{reg} cells of mixed clones indeed exhibited greater *FOXP3* expression than T_{conv} cells within the same clonal expansion (Extended Data Fig. 8e and Methods). As in the previously held-out bulk sequencing data, mixed clones in single-cell data had intermediate TiRP scores, which were significantly greater than the scores of expanded, pure T_{conv} clones (Fig. 6g, mixed–T_{conv} mean TiRP difference = 0.182, $P = 3.0 \times 10^{-4}$, LRT; Methods). With the limited extent of T_{reg} expansion, we were underpowered to detect significant differences between mixed and T_{reg} clones in these data (mixed–T_{reg} mean TiRP difference = –0.005, $P = 0.57$, LRT). When we quantified clone phenotypes by the proportion of T_{reg} cells and T_{conv} cells within each clone, however, increasing TiRP corresponded to more T_{reg}-skewed clonal expansions (LRT $P = 0.003$; Fig. 6h and Methods). To our knowledge, TiRP is the first metric to identify TCR-intrinsic, rather than TCR-extrinsic, factors relevant to peripheral phenotypic conversion.

Separable drivers of TiRP: self-peptide and human MHC. We next asked whether TiRP captured the major sources of TCR sequence variation between sorted T cell samples from diverse individuals. For this, we conducted a principal components analysis (PCA) of TCR feature frequencies in the sorted samples of the replication dataset, in which all T cell states of interest were available (Methods). We observed that the major axes of TCR sequence variation corresponded to T cell state rather than donor *HLA* genotype or clinical phenotype (Fig. 7a and Extended Data Fig. 9a,b). While our previous supervised modeling was designed to focus on T_{reg}–T_{conv} differences, this approach recovered the importance of T cell state in an unsupervised manner.

PCA delineated two axes of TCR-driven cell states: antigen-experienced (T_{reg} and memory T_{conv}) versus naive (PC1) and regulatory versus conventional (PC2) (Fig. 7a,b). The axis dividing antigen-experienced from inexperienced samples (PC1) was most reliant on *TRBV* gene frequencies, while the axis dividing regulatory versus conventional samples (PC2) was most reliant on mean percent composition of amino acids in CDR3βmr and the CDR3βmr-adjacent residue p113 (Fig. 7c,d). Because TiRP is a weighted sum of TCR features from the V, J and middle regions, the score can be divided into three score components corresponding to these three regions. TiRP scoring by TCR region revealed that V region-specific TiRP (vTiRP) and CDR3βmr-specific TiRP (mTiRP) indeed captured PC1 and PC2, respectively (Fig. 7e,f; vTiRP and PC1, $R = -0.86$, $P = 1.5 \times 10^{-20}$; mTiRP and PC2, $R = 0.85$, $P = 2.6 \times 10^{-20}$).

We next investigated possible biological drivers for vTiRP and mTiRP. The biological structure of the pMHC–TCR complex suggests that different regions of the TCR may promote T_{reg} fate via particular affinities; MHC class II mostly contacts the V region of the TCR, while the self-peptide is in closest contact with CDR3βmr[17,29,30] (Fig. 1a). Thus, we hypothesized that vTiRP enhanced affinity to human MHC class II, while mTiRP facilitated recognition of self-antigens. To test this idea, we examined TiRP in two complementary datasets: (1) murine T_{reg} TCRs[31], which recognize self-antigens but are not human MHC restricted, and (2) human memory T_{conv} TCRs[12,32], which are human MHC restricted but do not recognize self-antigens (Fig. 8a and Supplementary Table 2).

To apply TiRP to murine data, we first translated murine *TRBV* genes to their human homologs (Methods). We observed that human TiRP was significantly elevated in murine T_{reg} cells compared to T_{conv} cells (Fig. 8b, left; $P = 5.0 \times 10^{-136}$ for Helios$^+$ T_{reg} cells and $P = 0.003$ for Helios$^-$ T_{reg} cells, LRT; Methods). Thus, TiRP facilitates recognition of self, even in the context of an entirely different species' MHC restriction. A parsimonious explanation for this finding, among several possible explanations, is that TiRP enhances affinity to self-peptides. Consistent with this explanation, TiRP is significantly elevated in the 361 CD4$^+$ autoreactive TCRs currently documented in McPAS-TCR[33] and VDJdb[34] (Extended Data Fig. 10; $P = 1.5 \times 10^{-9}$, Wald test). Across 11 studies, these 361 autoreactive TCRs were identified by their reactivity to tetramers or antigen-presenting cells (APCs) presenting peptides known to be targeted in four autoimmune diseases (T1D, Celiac disease, multiple sclerosis and inflammatory bowel disease).

TiRP was dramatically elevated in murine T_{reg} cells that expressed *Helios*, a marker of thymic T_{reg} fate acquisition (Fig. 8b, left). Consistent with our TCR region hypothesis, the TiRP component with the greatest increase between murine T_{conv} cells and T_{reg} cells was mTiRP (Fig. 8c, left). CDR3βmr amino acid percentage effect sizes replicated strongly between murine and human data (Extended Data Fig. 9c, Pearson's $R = 0.85$, $P = 0.00013$, while other TCR features did not (Extended Data Fig. 9d, Supplementary Table 15 and Methods). These results strongly suggest that CDR3βmr features such as hydrophobicity promote T_{reg} fate via enhanced recognition of self. Interestingly, mTiRP also accounted for the increased TiRP of mixed clones of the human tumor microenvironment (Extended Data Fig. 9e; $P = 2.9 \times 10^{-4}$, Wald test). Taken together, these results suggest self-peptide recognition by exT_{reg} cells in the tumor microenvironment and underline the role of interactions between CDR3βmr and the antigenic peptide in T_{reg} fate acquisition.

To understand the role of human MHC, we next compared TiRP in naive and memory T_{conv} TCRs[12], which do not strongly recognize self-peptides[6] (Fig. 8a, Supplementary Table 2 and Methods). TiRP was significantly elevated in human memory T_{conv} cells compared to human naive T_{conv} cells (Fig. 8b, right), indicating that affinity to human MHC class II also contributes to TiRP. Consistent with the hypothesis of V region-based affinity to human MHC class II molecules, vTiRP was the only TiRP component to increase in human memory T_{conv} cells (Fig. 8c, right). As expected, large-effect-size TCR features between memory T_{conv} cells and naive T_{conv} cells were predominantly *TRBV* genes (Fig. 8d and Extended Data Fig. 9f), and the extent of each gene's enrichment in memory T_{conv} cells correlated with the extent of its enrichment in T_{reg} cells (Fig. 8d; Pearson's $R = 0.702$ and $P = 4.5 \times 10^{-5}$). These effects further replicated in an entirely independent cohort of sorted memory and naive T cells from five healthy donors[32] (Supplementary Tables 2 and 16 and Extended Data Fig. 9g). Thus, as structural interactions in the pMHC–TCR complex would suggest, V region features modulate

affinity to MHC, thereby shaping the T cell's general disposition for activation.

Discussion

Because the TCR sequence arises from a random process before T cell fate determination, associations between the TCR and T cell fate indicate causal effects of the TCR. The majority of T_{reg} research to date has focused on TCR-extrinsic determinants of T cell fate, such as the effect of co-stimulatory receptors, antigenic peptides and cytokines[35]. Although each of these elements certainly plays an essential role in T cell fate, the contribution of the TCR sequence itself has not yet been comprehensively investigated. TCR-intrinsic factors are relevant to nearly all immunological contexts, including the engineering of TCRs for immune therapies.

In this work, we leveraged the affinity-based partition of the repertoire into T_{reg} cells and T_{conv} cells to uncover determinants of TCR avidity toward the self-pMHC class II complex. We identified TCR sequence features that are predictive of T_{reg} cell fate across seven independent cohorts, encompassing diverse genetic, clinical and tissue contexts as well as sequencing protocols. Donor TCR samples were excluded due to incomplete cell sorting in only two of these seven cohorts. Using mixed effects logistic regression, we developed a scoring system that captures the TiRP of a given TCR. We validated this scoring system in three external datasets, including TCRs from the human thymus. We observed that TiRP largely reflects centrally derived T_{reg} TCRs but is also moderately elevated in peripherally derived T_{reg} cells. Excitingly, TiRP helped to explain the variable tendency of T cell clones to exhibit a regulatory phenotype in the tumor microenvironment. The application of TiRP scoring to murine data demonstrated that these TCR differences persist even with limited pathogen exposure. As evidenced by these diverse contexts, TiRP quantifies the extent to which a T cell is fated to be a T_{reg}, purely due to its TCR.

It is important to recognize several limitations to our approach. First, the amount of variance in T cell state explained by the TCR is significant but modest considering the full diversity of the repertoire. For any given TCR, specific antigenic contacts and co-stimulatory signals are likely the major determinants of T cell phenotype. Our results show, however, that TCR features, such as hydrophobicity, consistently predispose the T cell to adopt a regulatory phenotype. Second, our analyses focused on the β-chain of the TCR. The β-chain is more variable than the α-chain and is largely considered to mediate antigen specificity. However, the α-chain may also play a role in determining T cell phenotype, which remains to be explored. Third, although we found preliminary evidence that TiRP is elevated in $CD4^+$ autoreactive TCRs, the current data represent only four of many diseases that have been described as autoimmune. This finding will need to be reassessed as efforts progress to identify a comprehensive set of autoreactive TCRs for these diseases and for others.

The broadest takeaway from our work is the hydrophobic bias of T_{reg} TCRs, present at each of the peptide contact residues of CDR3β. This observation extends previous work[18,36] regarding p109 and p110 of T_{reg} TCRs and demonstrates that the hydrophobic bias is in fact not specific to these positions. As a group, hydrophobic amino acids are among the strongest interacting[37]. The concept that the strength of amino acid interactions may influence the thymic fate of a TCR was first predicted by Kosmrlj et al[38]. In this computational model of thymic selection, TCRs with 'weakly interacting amino acids' (QNSTAG) best evaded negative selection. Antigen specificity then followed; for TCRs with only weak amino acid interactions, any change in peptide sequence abrogates TCR recognition. If the T_{reg} population is thought of as 'partially' negatively selected (that is, precisely the TCRs for which pMHC recognition in the thymus is high but not to a fatal extent), their TCRs should be enriched in strongly interacting amino acids (IVYWREL). Our analyses confirm this enrichment in T_{reg} cells and suggest that the phenomena also applies to fully negatively selected TCRs. If strongly interacting residues make TCR recognition relatively robust to changes in peptide sequence, antigen specificity may be reduced in T_{reg} cells compared to T_{conv} cells. Perhaps such degenerate 'stickiness' allows the T_{reg} to generalize from the self-peptide encountered in the thymus to a larger pool of protected self-antigens.

Importantly, however, CDR3βmr hydrophobicity is not the full picture. *TRBV* gene usage explained nearly as much variance in T cell fate, and *TRBV* gene effects were not related to hydrophobicity. Our work suggested instead that the pI of CDR1β p37 encoded by the *TRBV* gene shapes affinity to conserved sites of MHC class II (ref. [17]). While the T_{reg}-promoting effect of hydrophobic CDR3βmr amino acids did not translate to the development of memory T_{conv} cells, memory T_{conv} cells and T_{reg} cells exhibited strikingly similar *TRBV* gene biases compared to the naive repertoire. These results suggest that hydrophobic residues in the CDR3βmr may only be 'sticky' toward self-peptides, while T_{reg}-promoting *TRBV* genes enhance affinity to MHC class II and thereby predispose $CD4^+$ T cells to recognize both self and non-self.

These phenomena offer a new lens on the T cell immune response; although each TCR tends to recognize a specific cognate antigen, all TCRs are subject to common processes that shape T cell activation. Due to these common processes, not all TCRs are created equal; those with a higher baseline for general reactivity may require a less 'perfect' cognate antigen for activation. Existing tools provide rough annotations for 'TCR strength', but these are based on frequently interacting residues in general protein structures[38]. TiRP sharpens our understanding of high-affinity amino acids in the context of the pMHC–TCR complex, providing a crucial functional annotation for this immune receptor.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41590-022-01129-x.

Received: 14 June 2021; Accepted: 5 January 2022;
Published online: 17 February 2022

References

1. Jordan, M. S. et al. Thymic selection of $CD4^+CD25^+$ regulatory T cells induced by an agonist self-peptide. *Nat. Immunol.* **2**, 301–306 (2001).
2. Yun, T. J. & Bevan, M. J. The Goldilocks conditions applied to T cell development. *Nat. Immunol.* **2**, 13–14 (2001).
3. Sakaguchi, S., Yamaguchi, T., Nomura, T. & Ono, M. Regulatory T cells and immune tolerance. *Cell* **133**, 775–787 (2008).
4. Klein, L., Hinterberger, M., Wirnsberger, G. & Kyewski, B. Antigen presentation in the thymus for positive selection and central tolerance induction. *Nat. Rev. Immunol.* **9**, 833–844 (2009).
5. Romagnoli, P. & van Meerwijk, J. P. M. Thymic selection and lineage commitment of $CD4^+Foxp3^+$ regulatory T lymphocytes. *Prog. Mol. Biol. Transl. Sci.* **92**, 251–277 (2010).
6. Moran, A. E. et al. T cell receptor signal strength in T_{reg} and iNKT cell development demonstrated by a novel fluorescent reporter mouse. *J. Exp. Med.* **208**, 1279–1289 (2011).
7. Ohkura, N. et al. T cell receptor stimulation-induced epigenetic changes and Foxp3 expression are independent and complementary events required for T_{reg} cell development. *Immunity* **37**, 785–799 (2012).
8. Li, M. O. & Rudensky, A. Y. T cell receptor signalling in the control of regulatory T cell differentiation and function. *Nat. Rev. Immunol.* **16**, 220–233 (2016).
9. Sidwell, T. et al. Attenuation of TCR-induced transcription by Bach2 controls regulatory T cell differentiation and homeostasis. *Nat. Commun.* **11**, 252 (2020).
10. Bolotin, D. A. et al. Antigen receptor repertoire profiling from RNA-seq data. *Nat. Biotechnol.* **35**, 908–911 (2017).

11. Seay, H. R. et al. Tissue distribution and clonal diversity of the T and B cell repertoire in type 1 diabetes. *JCI Insight* **1**, e88242 (2016).
12. Gomez-Tourino, I., Kamra, Y., Baptista, R., Lorenc, A. & Peakman, M. T cell receptor β-chains display abnormal shortening and repertoire sharing in type 1 diabetes. *Nat. Commun.* **8**, 1792 (2017).
13. Park, J.-E. et al. A cell atlas of human thymic development defines T cell repertoire formation. *Science* **367**, eaay3224 (2020).
14. Khosravi-Maharlooei, M. et al. Cross-reactive public TCR sequences undergo positive selection in the human thymic repertoire. *J. Clin. Invest.* **129**, 2446–2462 (2019).
15. Joller, N. & Kuchroo, V. Good guys gone bad: exT_{reg} cells promote autoimmune arthritis. *Nat. Med.* **20**, 15–17 (2014).
16. Sharon, E. et al. Genetic variation in MHC proteins is associated with T cell receptor expression biases. *Nat. Genet.* **48**, 995–1002 (2016).
17. Reche, P. A. & Reinherz, E. L. Sequence variability analysis of human class I and class II MHC molecules: functional and structural correlates of amino acid polymorphisms. *J. Mol. Biol.* **331**, 623–641 (2003).
18. Stadinski, B. D. et al. Hydrophobic CDR3 residues promote the development of self-reactive T cells. *Nat. Immunol.* **17**, 946–955 (2016).
19. Azizi, E. et al. Single-cell map of diverse immune phenotypes in the breast tumor microenvironment. *Cell* **174**, 1293–1308 (2018).
20. Yost, K. E. et al. Clonal replacement of tumor-specific T cells following PD-1 blockade. *Nat. Med.* **25**, 1251–1259 (2019).
21. Samstein, R. M., Josefowicz, S. Z., Arvey, A., Treuting, P. M. & Rudensky, A. Y. Extrathymic generation of regulatory T cells in placental mammals mitigates maternal-fetal conflict. *Cell* **150**, 29–38 (2012).
22. Cebula, A. et al. Thymus-derived regulatory T cells contribute to tolerance to commensal microbiota. *Nature* **497**, 258–262 (2013).
23. Zhou, X. et al. Instability of the transcription factor Foxp3 leads to the generation of pathogenic memory T cells in vivo. *Nat. Immunol.* **10**, 1000–1007 (2009).
24. Setoguchi, R., Hori, S., Takahashi, T. & Sakaguchi, S. Homeostatic maintenance of natural $Foxp3^+CD25^+CD4^+$ regulatory T cells by interleukin (IL)-2 and induction of autoimmune disease by IL-2 neutralization. *J. Exp. Med.* **201**, 723–735 (2005).
25. Komatsu, N. et al. Pathogenic conversion of $Foxp3^+$ T cells into T_H17 cells in autoimmune arthritis. *Nat. Med.* **20**, 62–68 (2014).
26. Zemmour, D. et al. Single-cell gene expression reveals a landscape of regulatory T cell phenotypes shaped by the TCR. *Nat. Immunol.* **19**, 291–301 (2018).
27. Kang, J. B. et al. Efficient and precise single-cell reference atlas mapping with Symphony. *Nat. Commun.* **12**, 5890 (2021).
28. Nathan, A. et al. Multimodally profiling memory T cells from a tuberculosis cohort identifies cell state associations with demographics, environment and disease. *Nat. Immunol.* **22**, 781–793 (2021).
29. Jorgensen, J. L., Esser, U., Fazekas de St Groth, B., Reay, P. A. & Davis, M. M. Mapping T-cell receptor–peptide contacts by variant peptide immunization of single-chain transgenics. *Nature* **355**, 224–230 (1992).
30. Garcia, K. C. et al. An αβ T cell receptor structure at 2.5 Å and its orientation in the TCR–MHC complex. *Science* **274**, 209–219 (1996).
31. Thornton, A. M. et al. $Helios^+$ and $Helios^-$ T_{reg} subpopulations are phenotypically and functionally distinct and express dissimilar TCR repertoires. *Eur. J. Immunol.* **49**, 398–412 (2019).
32. Soto, C. et al. High frequency of shared clonotypes in human T cell receptor repertoires. *Cell Rep.* **32**, 107882 (2020).
33. 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).
34. Shugay, M. et al. VDJdb: a curated database of T-cell receptor sequences with known antigen specificity. *Nucleic Acids Res.* **46**, D419–D427 (2018).
35. Lee, Y. K., Mukasa, R., Hatton, R. D. & Weaver, C. T. Developmental plasticity of T_H17 and T_{reg} cells. *Curr. Opin. Immunol.* **21**, 274–280 (2009).
36. Daley, S. R. et al. Cysteine and hydrophobic residues in CDR3 serve as distinct T-cell self-reactivity indices. *J. Allergy Clin. Immunol.* **144**, 333–336 (2019).
37. Košmrlj, A., Jha, A. K., Huseby, E. S., Kardar, M. & Chakraborty, A. K. How the thymus designs antigen-specific and self-tolerant T cell receptor sequences. *Proc. Natl Acad. Sci. USA* **105**, 16671–16676 (2008).
38. Miyazawa, S. & Jernigan, R. L. Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules* **18**, 534–552 (1985).

Methods

Data preparation. *Bulk sequencing data.* We downloaded the discovery cohort[11], replication cohort[12], murine cohort[31] and memory cohort[32] sequencing data from the Adaptive Biotechnologies immuneACCESS site (URLs), and we downloaded the thymic bulk sequencing cohort[14] from GitHub (URLs). For all data, we defined CDR3 amino acid sequences with stop codons or frameshifts to be non-productive amino acid sequences. We restricted all analyses to CDR3 sequences of a length within 12 and 17 amino acids, representing 91.8% of observations in the discovery cohort. We aligned CDR3 amino acids to positions defined by IMGT (URLs), wherein sequences less than 15 amino acids have CDR3βmr gaps, and sequences longer than 15 amino acids have extra CDR3βmr positions. We examined only one copy of each CDR3β sequence within each individual. Unless explicitly noted, we excluded CDR3β reads that were observed in both the T_{reg} and T_{conv} samples of any individual (0.63% of observations in the discovery cohort and 1.9% of observations in the replication cohort). For the discovery cohort, we restricted our analysis to the 24 donors with both T_{reg} and T_{conv} TCRs available. For the replication cohort, we restricted our analysis to the 16 donors with both T_{reg} and T_{conv} TCRs available.

Single-cell sequencing data. We downloaded scRNAseq tumor microenvironment data[19,20] from the Gene Expression Omnibus (GEO) through accession numbers GSE114727, GSE114724 and GSE123814. For the scRNAseq thymic data, we downloaded fastq files from ArrayExpress under accession number E-MTAB-8581 and metadata from Zenodo (https://doi.org/10.5281/zenodo.3711134). For quality control, we included only cells for which (1) more than 1,000 genes were expressed, (2) less than 25% of detected unique molecular identifiers were of mitochondrial origin and (3) exactly one productive TCR β-chain was detected. We followed the quality control process of the original authors for the multimodal memory T cell dataset[28], which is available for download from the GEO through accession number GSE158769.

Statistical analyses. All mixed effects models were fit with the R package lme4. All model comparisons were computed with the R package stats. All significance tests on Pearson's *R* were *t*-tests with the Fischer transformation. All analyses were performed with R version ≥3.6.1.

Holding out observations for calibration and testing. To leverage both the discovery[11] and replication[12] cohorts in the development of TiRP, we used approximately 70% of the TCR clones from each cohort for training, 10% for calibration and 20% for testing. To preserve the novelty of held-out data, we kept all TCR clone observations from the same individual together in this process, holding out entire repertoire samples. In the discovery cohort, we held out two individuals for TiRP calibration (donor IDs 6279 and 6196, accounting for 8.4% of TCR clones in the discovery cohort) and four individuals (donor IDs 6161, 6193, 6207 and 6287, accounting for 20.3% of clones in the discovery cohort) for TiRP testing. In the replication cohort, we held out one individual for TiRP calibration (T1D3) and three individuals (HD1, HD2 and T1D6) for validation. TCR sequence feature effect sizes were estimated in a separate mixed effects model for each cohort for each independent region of the TCR.

MI structure of the CDR3β sequence. We first calculated the conditional MI for all possible trios of CDR3β positions: the normalized MI of positions A and B given position C. For all trios, we normalized conditional MI by dividing by the mean conditional entropy of positions A and B given position C, such that the normalized MI was ultimately equivalent to 'symmetric uncertainty'[39] or the harmonic mean of the uncertainty coefficients. We used the R package 'infotheo' to compute all conditional MI and conditional entropy values.

We then calculated the Shannon entropy[40] of each CDR3β position and the MI[41] between all pairs of CDR3β positions with the R package DescTools. Again, to normalize MI, we divided MI for a given pair of positions by the mean entropy of those two positions.

Selection of random effects and model comparisons. In the discovery cohort[11], T cells were sampled from four tissues: peripheral blood (PBMCs), spleen, pancreatic lymph node and inguinal/irrelevant lymph node. We reasoned that there were three sensible ways to model tissue as a source of variation in T cell state.

(1) Model as a fixed effect:

$$\log(\frac{p}{1-p}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + b_{0i}$$

where p is the probability that the CD4$^+$ sorted CDR3β sequence belongs to a T_{reg}, β_0 is an intercept, X_1 is an indicator variable set to 1 if the sequence is from a PBMC sample, X_2 is an indicator variable for spleen origin, X_3 is an indicator variable for inguinal/irrelevant lymph node origin (pancreatic lymph node as a reference), and b_{0i} is a modification to the intercept fit to each individual i, normally and independently distributed (NID) with mean 0 and variance σ_0^2.

(2) Model as a random intercept effect independent from the random intercept effect per individual, wherein matched tissues across donors have the same (zero-centered) intercept effect:

$$\log(\frac{p}{1-p}) = \beta_0 + b_{0i} + b_{1j}$$

where b_{1j} is a modification to the intercept fit to each tissue j, NID with mean 0 and variance ${\sigma_1}^2$ and all other variables maintain previous definitions.

(3) Model as a nested random intercept effect, wherein each tissue–donor pair is modeled as a unique batch of correlated observations within the individual-level and tissue-level variances:

$$\log(\frac{p}{1-p}) = \beta_0 + b_{0i} + b_{1j} + b_{2i,j}$$

where $b_{2i,j}$ is a modification to the intercept fit to each individual i, tissue j pair, NID with mean 0 and variance ${\sigma_2}^2$ and all other variables maintain previous definitions. For stable numerical results, we included the marginal random effects for donor and tissue in this nested random intercept model.To determine which of these models was most appropriate, we calculated the pseudo-R^2 by the conventional McFadden[42] approach (range 0–1) and multiplied the result by 100 (variance explained range of 0–100). All measures of variance explained in this study were computed with this approach. For this analysis, we compared models 1–3 to a baseline model that fit the log odds of T_{reg} status only to a random intercept for each individual:

$$\log(\frac{p}{1-p}) = \beta_0 + b_{0i}$$

These model comparisons revealed that tissue explained 1.90% of variance as a fixed effect and 1.15% of variance as a random effect ($P = 1.15 \times 10^{-11,211}$ fixed and $P = 4.68 \times 10^{-10,229}$ random, LRT). However, tissue as a random effect nested within individual explained 6.27% of variance ($P = 1.32 \times 10^{-55,291}$, LRT). We therefore concluded that nesting a random tissue effect within the donor random effect was the most appropriate model for the batch structure of these data and proceeded with three random intercepts for each mixed effects model: the nested donor–tissue effect, the marginal donor effect and the marginal tissue effect.

CDR3βmr mixed effects logistic regression. For each amino acid, we calculated the percentage of CDR3βmr positions occupied by this residue; a percentage of 0 means that the residue is missing for a given TCR, while a percentage of 100 means that the residue is present at every CDR3βmr position. We scaled this percentage to have a mean of 0 and variance of 1, and tested the scaled percentage in a separate mixed effects logistic regression for each amino acid with random intercepts as described above. We controlled for CDR3β sequence length by including it as a categorical covariate, reasoning that conformational differences in the MHC–TCR complex may not scale linearly with additional residues. To collect the relevant amino acid proportions, we did a forward search in which we iteratively added to the mixed effects model the amino acid proportion that provided the greatest improvement in model fit. On the first round, the percentage of CDR3βmr positions occupied by glutamic acid in each TCR explained the most variance, with a 9.7% fall in odds of T_{reg} fate per additional glutamic acid residue for CDR3βs of length 15 (pseudo-$R^2 = 0.036\%$, LRT $P = 8.37 \times 10^{-196}$, OR = 0.954, 95% CI = 0.951–0.957). Conditioning on this feature revealed that the next amino acid with the greatest independent effect was aspartic acid (pseudo-$R^2 = 0.042\%$, LRT $P = 1.01 \times 10^{-225}$, OR = 0.95, 95% CI = 0.947–0.953). We repeated this process until the remaining amino acid percentages no longer passed the Bonferroni-corrected significance threshold ($P = 0.05/20$ for 20 amino acids; Fig. 3b, middle). We confirmed that this threshold kept the type I error rate below 0.05 by repeating this analysis 1,000 times, with T_{conv} and T_{reg} labels for each TCR randomly shuffled within the data for each donor on each run.

Position-specific mixed effects logistic regressions. To parse the *TRBV*-encoded region, we asked if the 5′-flanking CDR3β residues could be represented by a handful of motifs. Indeed, the eight p104–p106 sequences ('Vmotifs') present in each donor with a frequency of >0.001 in every donor accounted for 96.2% of TCRs. We labeled the remaining 3.8% of TCRs with a Vmotif of 'other'.

To avoid multicollinearity in our selection of covariates, we calculated all correlation coefficients for each pair of TCR features in the discovery dataset. This computation for *TRBV* gene and Vmotif, for example, yields 57 non-reference *TRBV* genes × 7 non-reference Vmotifs = 399 correlation coefficients. Visualized in Extended Data Fig. 3a–c is the correlation coefficient with the maximum absolute value for each TCR feature pair. All pairs of features derived from the V region exhibited $|r| > 0.7$, except for pairings with p107 (Extended Data Fig. 3b).

P107 featured moderate correlation coefficients with other V region features, suggesting two viable models for comparison: (1) joint modeling of the *TRBV* gene identity with the p107 amino acid and (2) joint modeling of Vmotif with p107. By comparing the pseudo-R^2 of these two models (Fig. 3b, left), we concluded that the V region was best modeled by joint estimation of *TRBV* gene and p107 residue effect sizes. To account for donor-individualized *TRBV* gene thymic selection, we included VGSR as a fixed covariate in this final model (Supplementary Note).

Similarly, to parse the *TRBJ*-encoded region, we asked if the 3′-flanking CDR3β residues could be represented by a handful of motifs. Indeed, the 42 p114–p118 sequences (Jmotifs) present in each donor with a frequency of >0.001 in every donor accounted for 91.5% of TCRs. Computation of all pairwise correlation coefficients for TCR features in the J region (Extended Data Fig. 3c) suggested two

possible non-multicollinear models: (1) joint modeling of the *TRBJ* gene identity with the p113 amino acid and (2) joint modeling of Jmotif with p113. In contrast to the V region, here it appeared that the motif afforded a greater pseudo-R^2 than the gene (Fig. 3b, right), and so we proceeded with joint estimation of Jmotif and p113 for the J region.

To confirm the absence of multicollinearity in these models, we computed the inflations in variance for coefficient estimates and found that avoiding pairs with any $|r|>0.7$ successfully corrected variance inflation (Extended Data Fig. 3d,e). To make the variance inflation comparable across multiple d.f., we used the generalized variance inflation factor[43] $GVIF^{\frac{1}{2\times d.f.}}$, computed with the R package 'car'.

To protect against numerically unstable estimates, we report only the effect sizes of TCR features with a frequency greater than 0.005 in the training data for both the discovery and replication cohorts.

Calculating TCR proportions. To approximate the proportion of the TCR occupied by each TCR region in Fig. 3d, we divided the number of amino acids in a given TCR region by the estimated total number of TCR β-chain amino acids protruding into the MHC–TCR complex (Fig. 2b). To estimate the total number of amino acids protruding into the MHC–TCR complex, we added 11 to the observed CDR3β length because over 70% of TCR clones in the discovery training data express a *TRBV* gene with exactly 11 amino acids in the CDR1β and CDR2β loops. Thus, we estimated the absolute size of the V region to be 15 amino acids (11 + 4 CDR3β amino acids), the size of the J region to be 6 amino acids, and the size of the CDR3βmr to vary with CDR3β length (Fig. 2b).

Null model comparisons for variance explained by TCR features. To generate a suitable null model for variance explained by TCR features, we conducted permutation analyses. Within each donor and tissue sample of the discovery cohort used for training, we permuted the cell type labels (T_{reg} versus T_{conv}) for each TCR 1,000 times. On each permutation, we fit mixed effects logistic regression models for the CDR3βmr and J region as described above (Supplementary Table 7).

Estimating the effects of physicochemical features. To estimate the effects of physicochemical features, we represented each CDRβ loop residue as a vector of length 3, corresponding to the amino acid's hydrophobicity, pI and volume. For consistency with the closely related work by Stadinksi et al.[18], we used the whole-residue interfacial hydrophobicity scale[44]. We used pI values from the CRC Handbook of Chemistry and Physics[45] and volume estimates from IMGT's conversion of Zamyatnin's[46] measurements to cubed angstroms (URLs). Each value was scaled to have a mean 0 and variance 1 for regression analysis.

To localize the importance of these physicochemical features within the TCR, we represented each residue belonging to a CDRβ loop as a vector of length 3 corresponding to the amino acid's hydrophobicity, pI and volume and modeled T_{reg} fate as an outcome of these features using multiple logistic regression. We followed IMGT positioning, wherein the human CDR1β loop consists of positions 27, 28, 29, 37 and 38, while the human CDR2β loop consists of positions 56, 57, 58, 63, 64 and 65. We used only TCR reads with a resolved *TRBV* gene (78.5% of observations) and imputed CDR loop amino acids based on *TRBV* gene identity using IMGT (URLs). To enable TCR alignment, we discarded 3.6% of observations with a resolved *TRBV* gene for which there were not exactly five CDR1β amino acids and six CDR2β amino acids or for which CDR1–CDR2 amino acids were not available via IMGT.

To handle the densely correlated TCR features within the CDR1β and CDR2β loops, we applied a ridge penalty to the logistic regression using the R package 'glmnet'. This regularization served as a penalization strategy alternative to random effects, and so we included batch (donor and tissue source of the TCR) as a fixed and penalized covariate. As in the *TRBV* gene analysis, we used VGSR as a covariate to partial out genetic variation in TCR–MHC affinity (Supplementary Note). All predictors were scaled to a have mean 0 and variance 1. We did not assume that position-wise physicochemical effects would translate across different CDR3β lengths, and so we fit a separate logistic regression for each length. For each regression, we tuned the λ penalty by testing the 100 values generated by the glmnet package and selecting the one that gave the minimum mean cross-validated error across 10-folds of the training data in the discovery cohort. Sensitivity analyses confirmed that $\lambda=0.01$ was an appropriate choice for the data (Supplementary Table 10).

In a separate analysis isolated to the CDR3βmr, we fit a separate mixed effects logistic regression for each length–position combination in the discovery cohort training data (Extended Data Fig. 5b). We included all three physicochemical features as fixed covariates for each position and modeled donor and tissue sources as random effects as described above. Each physicochemical feature was scaled to have a mean 0 and variance 1 for each length–position combination.

For the Fig. 4d visualization, we included only TCRs with a CDR3β length of 15 amino acids in the discovery cohort training data and fit a separate mixed effects logistic regression for each position. Each regression included random intercepts as described above and one fixed covariate corresponding to the amino acid identity at the given position. We cast the most common amino acid as the reference: leucine for p108 and glycine for all other positions.

Assessing TCR residue interactive effects on T cell fate. Because the physicochemical features of hydrophobicity, pI and volume captured most of the variance explained by the CDR3βmr (Fig. 3b), we used these three features to test for TCR residue interactions with respect to T_{reg} fate. For each pair of TCR positions a and b, we fit nine mixed effects logistic regression models, one for each of the nine possible pairs of the three physicochemical features:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_{1a}X_{1a} + \beta_{1b}X_{1b} + \beta_{2a}X_{2a} + \beta_{2b}X_{2b} + \beta_{3a}X_{3a} + \beta_{3b}X_{3b} + \beta_4X_{1a}X_{1b} + b_{0i} + b_{1j} + b_{2i,j} \quad (1)$$

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_{1a}X_{1a} + \beta_{1b}X_{1b} + \beta_{2a}X_{2a} + \beta_{2b}X_{2b} + \beta_{3a}X_{3a} + \beta_{3b}X_{3b} + \beta_4X_{2a}X_{2b} + b_{0i} + b_{1j} + b_{2i,j} \quad (2)$$

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_{1a}X_{1a} + \beta_{1b}X_{1b} + \beta_{2a}X_{2a} + \beta_{2b}X_{2b} + \beta_{3a}X_{3a} + \beta_{3b}X_{3b} + \beta_4X_{3a}X_{3b} + b_{0i} + b_{1j} + b_{2i,j} \quad (3)$$

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_{1a}X_{1a} + \beta_{1b}X_{1b} + \beta_{2a}X_{2a} + \beta_{2b}X_{2b} + \beta_{3a}X_{3a} + \beta_{3b}X_{3b} + \beta_4X_{1a}X_{2b} + b_{0i} + b_{1j} + b_{2i,j} \quad (4)$$

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_{1a}X_{1a} + \beta_{1b}X_{1b} + \beta_{2a}X_{2a} + \beta_{2b}X_{2b} + \beta_{3a}X_{3a} + \beta_{3b}X_{3b} + \beta_4X_{2a}X_{1b} + b_{0i} + b_{1j} + b_{2i,j} \quad (5)$$

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_{1a}X_{1a} + \beta_{1b}X_{1b} + \beta_{2a}X_{2a} + \beta_{2b}X_{2b} + \beta_{3a}X_{3a} + \beta_{3b}X_{3b} + \beta_4X_{2a}X_{3b} + b_{0i} + b_{1j} + b_{2i,j} \quad (6)$$

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_{1a}X_{1a} + \beta_{1b}X_{1b} + \beta_{2a}X_{2a} + \beta_{2b}X_{2b} + \beta_{3a}X_{3a} + \beta_{3b}X_{3b} + \beta_4X_{3a}X_{2b} + b_{0i} + b_{1j} + b_{2i,j} \quad (7)$$

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_{1a}X_{1a} + \beta_{1b}X_{1b} + \beta_{2a}X_{2a} + \beta_{2b}X_{2b} + \beta_{3a}X_{3a} + \beta_{3b}X_{3b} + \beta_4X_{1a}X_{3b} + b_{0i} + b_{1j} + b_{2i,j} \quad (8)$$

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_{1a}X_{1a} + \beta_{1b}X_{1b} + \beta_{2a}X_{2a} + \beta_{2b}X_{2b} + \beta_{3a}X_{3a} + \beta_{3b}X_{3b} + \beta_4X_{3a}X_{1b} + b_{0i} + b_{1j} + b_{2i,j} \quad (9)$$

Here, p is the probability that the CDR3β sequence belongs to a T_{reg}, X_{1a} is the hydrophobicity of residue a, X_{2a} is the pI of residue a and X_{3a} is the volume of residue a (with analogous values X_{1b}, X_{2b} and X_{3b} for the physicochemical features of residue b), and intercept terms β_0, b_{0i}, b_{1j} and $b_{2i,j}$ are as defined previously. To test for interactive effects, we compared each of these models to a baseline model in which $\beta_4 = 0$:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_{1a}X_{1a} + \beta_{1b}X_{1b} + \beta_{2a}X_{2a} + \beta_{2b}X_{2b} + \beta_{3a}X_{3a} + \beta_{3b}X_{3b} + b_{0i} + b_{1j} + b_{2i,j}$$

All model comparisons were computed by the LRT. As depicted in Fig. 2b, the CDR3βmr is of variable length, ranging from 2 amino acids in CDR3βs of length 12 to 7 amino acids in CDR3βs of length 17; $\binom{2}{2}$ pairs of CDR3βmr residues in length 12 + $\binom{3}{2}$ pairs of CDR3βmr residues in length 13 + $\binom{4}{2}$ pairs of CDR3βmr residues in length 14 and so forth to $\binom{7}{2}$ pairs of CDR3βmr residues in length 17 totals to 56 total pairs of CDR3βmr residues. We fit the nine mixed effects logistic regression models enumerated above for each of these 56 pairs in both the discovery and replication cohorts and integrated the results via meta-analysis as described for other TCR features. With 606 non-interactive TCR features (Supplementary Table 1) and 56×9 interactive effects, the Bonferroni significance threshold for these meta-analytic P values was $0.05/((9\times56)+606)=4.5\times10^{-5}$.

Developing the TiRP scoring system. We defined TiRP as the sum of the TCR sequence features present in a given TCR, reasoning that the effects of TCR

features were additive provided that they were fit jointly or derived from independent regions of the TCR. To reach a consensus effect size for each TCR feature across the two cohorts, we used inverse variance-weighted meta-analysis. Due to the inconsistent effect size directions for the usage of Valine (V) in the CDR3βmr (Fig. 5a and Extended Data Fig. 2b), we included only 14 amino acid percent covariates in our final CDR3βmr models (Supplementary Table 1). To exclude potentially unreliable effect size estimates from the score computation, we calibrated a meta-*P* value significance threshold above which TCR features were excluded from the score. For this, we used a single mixed effects logistic regression for each threshold over a range of thresholds on the pooled discovery and replication TCRs held out for calibration (discovery cohort: 6279 and 6196; replication cohort: T1D3). Each mixed effects logistic regression estimated the fixed effect of TiRP on T cell fate, with random intercepts for donor source, tissue source and each donor–tissue source pair (see Selection of random effects and model comparisons). We found that no threshold led to significantly greater variance explained than the Bonferroni-corrected threshold (0.05/612 TCR features), resulting in 25 *TRBV* genes, 23 Jmotifs, 4 CDR3β lengths, 14 CDR3βmr amino acid percentages and 142 position-specific features relevant to TiRP computation (Supplementary Table 12).

Testing TiRP in held-out donors from bulk sequencing cohorts. To test TiRP in bulk sequencing data, we scored each unique productive TCR in donors held out from both TiRP training and calibration (discovery cohort donors 6161, 6193, 6207 and 6287 and replication cohort donors HD1, HD2 and T1D6). We then tested the association between TiRP and T cell state by comparing the additional variance explained by a mixed effects logistic regression model including TiRP as a fixed covariate to a baseline model containing only donor ID, tissue source and donor–tissue interaction as random intercepts (LRT). We conducted the same process for non-productive TCRs in held-out donors and restricted this analysis to the discovery cohort, in which TCR gDNA was sequenced, and therefore out-of-frame reads were available (Supplementary Table 2). To ascertain the difference between high-scoring and low-scoring TCRs in these held-out data, we collected the top and bottom decile of TCRs per donor and compared the ratio of T_{reg} cells to T_{conv} cells between the group of all top decile TCRs and the group of all bottom decile TCRs.

Validating TiRP in single-cell data. In single-cell data analyses, TCR clones were defined by a barcode consisting of their donor ID and CDR3β DNA sequence. As in bulk sequencing analyses, CDR3β chains with a length shorter than 12 amino acids or longer than 17 amino acids were discarded. Only cells with exactly one productive CDR3β detected were included in analyses.

We computed the TiRP score for each clone based on its CDR3β amino acid sequence and *TRBV* gene. So that TiRP scores would be comparable, percent amino acid values were scaled by the mean and s.d. of the TCRs held out for testing from the discovery cohort (transformation provided in Supplementary Table 12). *TRBV* gene usage was determined by MiXCR alignments for the Azizi et al.[19] cohort and Park et al.[13] cohort and by RNA expression in the Yost et al.[20] cohort. To determine *TRBV* gene usage based on RNA expression in the Yost et al. cohort, read counts were log normalized per cell and then scaled so that each *TRBV* gene had mean 0 and variance 1 within cells that had non-zero read counts for the given gene. Each cell was then assigned the *TRBV* gene with the highest normalized and scaled expression. Cells without any *TRBV* gene expression detected were given a *TRBV* gene value 'unresolved'.

To validate the TiRP score in these data, we tested the association between TiRP score and regulatory or conventional cell phenotype. For the Yost et al. cohort, cell phenotypes based on the original authors' clustering were available. We labeled all cells in the 'T_{regs}' and 'T_{reg}' cluster as T_{reg} and all cells in the 'T_{FH}', 'T_H17', 'CD4_T_cells' and 'Naive' to be $CD4^+$ T_{conv}. For the Azizi et al. cohort, we applied a standard scRNAseq pipeline to infer cell phenotypes; we excluded all cells with read counts from 1,000 genes or less or at least 25% of read counts from mitochondrial genes and then used the R package 'Seurat' with default parameters to (1) normalize the read counts per cell, (2) take the variance-stabilizing transform, (3) scale and center gene expression and (4) compute the first 20 PCs based on the 500 most variable genes. We then used Harmony[47] to batch correct the PC embeddings by sample (donor_batch ID) and constructed a shared-nearest-neighbor (SNN) graph based on these harmonized embeddings with $k=30$. Finally, we conducted Louvain clustering on the SNN graph with resolution 0.8 and ran uniform maniform approximation and projection (UMAP) on the first 10 harmonized PCs. After aligning fastq reads from the Park et al. cohort to GRCh38-3.0.0 with CellRanger version 6.1.1, we applied this same pipeline, including only the 29 samples from 11 donors (7 prenatal, 2 pediatric and 2 adult) with paired TCR sequences available, taking the top 1,000 variable genes per sample, harmonizing over donor ID, sample and enzyme used (collagenase or liberase) and using $k=10$ for the SNN graph. After clustering all cells with resolution 2.0, we distinguished T cells from other major lineages by expression of *CD3G*, *CD3D*, *NKG7*, *CD59*, *MS4A1*, *CD34* and *CD14*. We then filtered our analysis to T cells, retransformed expression, recomputed and harmonized the PCA, reconstructed the SNN graph and reclustered the cells at resolution 3.0 to identify T_{reg} thymocytes (Extended Data Fig. 6).

To create 95% CIs for T_{reg} odds per TiRP decile (Fig. 5d,e), we conducted bootstrapping with 10,000 iterations via the R package 'boot'.

Creating a $CD4^+$ memory T cell single-cell reference. To construct a reference of cellular phenotypes for $CD4^+$ memory T cells, we used a published dataset[28] of scRNAseq and CITE-seq for 500,000 memory T cells from 259 donors (Supplementary Table 2). From these quality-controlled data, we used cellular indexing of transcriptomes and epitopes by sequencing (CITE-seq) values to select 430,270 $CD4^+$ cells (normalized $CD4^+ > 1.5$ and normalized $CD8^+ < 1$, consistent with the original authors' procedure). We followed the method developed by Nathan et al.[28] to cluster the cells based on integrated mRNA and protein expression. First, we used the R package Seurat to normalize the read counts per cell, take the variance-stabilizing transform and then scale gene expression to have a mean of 0 and variance of 1. We selected the union of the 1,500 most variable genes (by mRNA expression) in each donor, resulting in 4,707 variable genes.

To integrate surface protein information, we used canonical correlation analysis (CCA). Following the original authors' procedure[28], we used the 'cc' function from R package CCA to resolve the coefficients that maximized the correlation between linear combinations of the 4,707 genes and the 31 manually curated surface proteins[28] in the CITE-seq panel. We then projected the cells into the 31 canonical dimensions in mRNA space and used Harmony[47] with default parameters to harmonize the embeddings of these canonical dimensions by donor. For visualization, we used the R package 'uwot' to conduct UMAP on the first 10 canonical dimensions using the cosine metric, a local neighborhood size of 30 and a minimum distance of 0.3 between embeddings. To identify cell types, we constructed an SNN graph ($k=10$) from the harmonized embeddings of the first 10 canonical dimensions and conducted Louvain clustering on the SNN graph with resolution 0.8, revealing one cluster (cluster 6) with markedly elevated *FOXP3* and CD25 expression and reduced CD127 expression. We labeled cells belonging to this cluster as T_{reg} cells and manually annotated the phenotypes of the other clusters based on surface expression of the 31 manually curated, immunologically relevant surface proteins as well as mRNA expression of *CCR7*, *IFNG*, *GZMK* and *CTLA4* (Extended Data Fig. 7c,d).

Mapping tumor-infiltrating T cells with Symphony. Before ascertaining mixed clones in tumor-infiltrating cells, we standardized T_{reg} and T_{conv} definitions between the two cohorts by projecting cells from both cohorts into the annotated low-dimensional space of the reference single-cell dataset. To accomplish this projection and simultaneously harmonize the tumor-infiltrating cells by cohort, donor and sample, we utilized Symphony[27]. Because the reference dataset consisted of only memory T cells and our hypothesis focused on expanded clones, we mapped only the tumor-infiltrating cells for which their paired *CDR3β* DNA sequence was detected on more than one cell within their sample (56.1% of cells in the Azizi et al. cohort, 60.6% of cells in the Yost et al.[20] basal cell carcinoma cohort and 73.7% of cells in the Yost et al. squamous cell carcinoma cohort). For each cohort separately, we used Symphony to map the query cells into the harmonized reference canonical variate embedding space while integrating over unwanted sources of technical variation tagged by donor and sample in the query. We used the resultant canonical variate embeddings to (1) impute cluster membership for query cells via *k* nearest neighbors in the reference cohort (R package 'class', $k=5$) and (2) project the query cells into the reference UMAP embedding. To visualize TiRP trends, we colored each cell by the average TiRP of its 100 nearest query neighbors in the 31 canonical dimensions (Fig. 6c).

Mixed clone analysis with bulk sequencing data. We conducted our mixed clone analysis with bulk sequencing data in the donors from the discovery and replication cohort that were held out from the estimation of TCR feature effect sizes and TiRP score calibration. Clones were defined by the 'barcode' consisting of their *CDR3β* nucleotide sequence, *TRBV* gene ID and donor ID. Because clonal expansion is a prerequisite to mixed clone status, we compared mixed clone TiRP scores to those of expanded T_{conv} and T_{reg} clones. For the discovery cohort, *CDR3β* chains were sequenced from gDNA, so clonal expansion could be derived from the number of 'templates' for each clone (number of biological molecules before PCR amplification, inferred by immunoSEQ via internal bias control). Because *TRB* chains were sequenced from cDNA in the replication cohort, we cannot be sure whether identical reads within the same sample represent *CDR3β* transcripts from one or multiple cells. However, we can deduce that identical reads across multiple flow-sorted samples from the same individual arose from multiple cells and therefore an expanded clone. Therefore, for the replication cohort, we collected a sample of the expanded clones from each donor by aggregating all *CDR3β* nucleotide sequences that arose in multiple flow-sorted samples from the same individual (T_{reg}, naive T_{conv}, central memory T_{conv} and stem cell-like memory T_{conv}). Because there was only one T_{reg} sorted sample for each individual, we could only detect pure T_{conv} or mixed clones in the replication cohort by this approach. We tested the effect of TiRP score on clone phenotype with mixed effects models as designed in the single-cell analyses.

Mixed clone analysis with single-cell data. To detect mixed clones in single-cell data, we aggregated cells into clones based on matching clonal barcodes consisting

of individual ID, *CDR3β* DNA sequence, *TRBV* gene and CDR3α amino acid sequence. To protect against contamination by doublets (droplets encapsulating two cells rather than one), we excluded cells with more than one unique *CDR3β* chain detected. Because the expression of multiple alpha chains, however, is a common biological phenomenon[48], we did not exclude multi-alpha chain cells. To assign a clonal barcode CDR3α for these cells, we selected the CDR3α sequence that was most often expressed by cells with a matching *CDR3β* DNA sequence in the given individual.

To model the effect of TiRP score on clone phenotype (T_{conv}, T_{reg} or mixed), we used mixed effects logistic regression with random intercept for the clone's source individual and the clone's source cohort (BRCA, squamous cell carcinoma or basal cell carcinoma). Because clonal expansion is a prerequisite to mixed clone status, only clones of size >1 were included. We used the LRT to compare the model including TiRP to a baseline model containing only the random covariates. We conducted this process twice, first to compare mixed clones to purely T_{conv} clones and second to compare mixed clones to purely T_{reg} clones.

We then quantified the clone phenotype by taking the natural log transform of the within-clone T_{reg}:T_{conv} ratio, with one 'hallucinated' T_{reg} and one 'hallucinated' T_{conv} per clone to protect against numerically unstable estimates. We tested the effect of TiRP score on this quantitative clone phenotype using mixed effects linear regression with random intercepts as described above and found a 0.065 increase in ln (T_{reg}:T_{conv} ratio) per s.d. increase in TiRP score (Fig. 6h; $P = 1.6 \times 10^{-4}$, LRT).

To check that *FOXP3* expression was significantly different between T_{reg} cells and T_{conv} cells within mixed clones, we conducted a Student's paired *t*-test and confirmed that this was indeed true (Extended Data Fig. 8e).

Analysis of murine TCRs. T cell clones were defined by the barcode consisting of CDR3β amino acid sequence, *TRBV* gene identity and donor ID. Due to ambiguity, clones observed in both T_{reg} and T_{conv} samples from the same donor or in both the *Helios*$^+$ and *Helios*$^-$ T_{reg} samples from the same donor were excluded from the following analyses. Clones with member cells in both the naive T_{conv} and memory T_{conv} samples from the same donor were labeled with the memory T_{conv} phenotype.

To compute the *TRBV* gene component of the TiRP score in murine data, we assigned each murine *TRBV* gene the TiRP coefficient of its human homolog according to human–mouse *TRBV* correspondences listed in IMGT (URLs). Murine and human *TRBV* genes were aligned for comparison in Extended Data Fig. 9d by this same correspondence scheme. Murine *TRBV* genes with multiple human *TRBV* gene homologs were assigned the average of their human homolog coefficients. Because the reference *TRBV* gene in human data, *TRBV05-01*, does not have a murine homolog, comparing *TRBV* gene effect sizes in mouse and human required a change to a common reference. We encoded *TRBV19-01* as the reference for murine mixed effects logistic regression models and translated human *TRBV* gene effect sizes to those that would be obtained from *TRBV19-01* as the reference by subtracting the meta-analytic effect size for *TRBV19-01* from all *TRBV* gene effect sizes (including *TRBV05-01*, originally at 0).

TCR feature PCA. To contextualize the amount of T cell phenotypic variation explained by TCR features identified in our work, we performed a PCA on the matrix of samples by TCR feature means for the replication cohort, in which sorted samples for all T cell phenotypes of interest were available (Supplementary Table 2 and Fig. 7a). For categorical TCR features such as *TRBV* gene or Jmotif, we one-hot-encoded the variable into a binary vector equal to the length of possible values and took the mean of each of the positions. As this process rapidly expands the dimensionality of each sample, we summarized the TCR features in the CDR3βmr by percent composition of each amino acid only. We used the function 'prcomp' from the R package stats to conduct singular-value decomposition of the centered and scaled matrix of samples by mean TCR features.

Analyzing the TiRP of autoreactive TCRs. To survey the TiRP of known autoreactive TCRs, we collected all CD4$^+$ β-chain TCRs currently documented in McPAS-TCR[33] and VDJdb[34] with an association to autoimmune disease. For TiRP scoring, we included only TCRs with a CDR3β length of 12–17 amino acids. For these 375 unique TCRs, we manually inspected their source publications and included only the 361 TCRs whose autoreactivity was confirmed by tetramers or APCs pulsed with a known pathogenic peptide. For reference, we compared these TiRP scores to repertoire memory CD4$^+$ T_{conv} cells from donors held out from TiRP training and calibration ($n = 3$ donors). Specifically, we fit a linear model of TiRP score as a function of TCR category (T_{conv} memory or autoimmune) and used the Wald test to assess whether TCR category is associated with a significant TiRP difference.

Memory-naive TCR comparisons. T cell clones were defined by the barcode consisting of CDR3β amino acid sequence, *TRBV* gene identity and donor ID. Due to ambiguity, clones observed in both T_{reg} and T_{conv} samples from the same donor were excluded from the following analyses. Clones with member cells in both the naive T_{conv} and memory T_{conv} samples from the same donor were labeled with the memory T_{conv} phenotype.

For the replication of T_{conv} memory-naive *TRBV* effects in the Soto et al. cohort[32], two additional steps were necessary to accommodate the deeper TCR sequencing within these individuals. First, only TCRs with a cysteine at p104 and phenylalanine at p118 were included. Although there does exist some minor physiologic variation at these conserved sites, such outlier sequences are not relevant to TiRP score computation. Second, although the donor source of each TCR was modeled as a random effect in other cohorts, we modeled it here as a fixed covariate, reducing computational burden and allowing the maximum likelihood estimation to converge.

URLs. ImmuneAccess:
https://clients.adaptivebiotech.com/immuneaccess
Thymic TCR bulk sequencing:
https://github.com/Aleksobrad/Humanized-Mouse-Data
Amino acids encoded by *TRBV* genes:
http://www.imgt.org/IMGTrepertoire/Proteins/proteinDisplays.php?species=human&latin=Homo%20sapiens&group=TRBV
Amino acid volumes:
http://www.imgt.org/IMGTeducation/Aide-memoire/_UK/aminoacids/abbreviation.html

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Data analyzed in this study were previously deposited in the following locations: immuneACCESS: https://doi.org/10.21417/B73S3K, https://doi.org/10.21417/B7C88S, https://doi.org/10.21417/AMT2019EJI, https://doi.org/10.21417/CS2020CR and https://doi.org/10.21417/B7001Z; GEO: GSE158769, GSE123813 and GSE114724; GitHub: https://github.com/aleksobrad/humanized-mouse-data; Zenodo: https://doi.org/10.5281/zenodo.3711134; ArrayExpress: E-MTAB-8581; 10x Genomics: https://cf.10xgenomics.com/supp/cell-exp/refdata-gex-GRCh38-2020-A.tar.gz; McPAS-TCR: http://friedmanlab.weizmann.ac.il/McPAS-TCR and VDJdb: https://vdjdb.cdr3.net.

Code availability

Custom analysis scripts are available on GitHub (https://github.com/immunogenomics/TiRP).

References

39. Witten, I. H. & Frank, E. *Data Mining: Practical Machine Learning Tools and Techniques* 2nd edn (Morgan Kaufmann, 2005).
40. Shannon, C. E. & Weaver, W. *The Mathematical Theory of Communication* (University of Illinois Press, 1998).
41. Ihara, S. *Information Theory for Continuous Systems* (World Scientific, 1993).
42. Zarembka, P. & Harcourt Brace & Company (1993–1999). *Frontiers in Econometrics* (Academic Press, 1974).
43. Fox, J. & Monette, G. Generalized collinearity diagnostics. *J. Am. Stat. Assoc.* **87**, 178–183 (1992).
44. Wimley, W. C. & White, S. H. Experimentally determined hydrophobicity scale for proteins at membrane interfaces. *Nat. Struct. Biol.* **3**, 842–848 (1996).
45. Lide, D. R. *CRC Handbook of Chemistry & Physics* 72nd edn (CRC Press, 1991).
46. Zamyatnin, A. A. Protein volume in solution. *Prog. Biophys. Mol. Biol.* **24**, 107–123 (1972).
47. Korsunsky, I. et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).
48. Schuldt, N. J. & Binstadt, B. A. Dual TCR T cells: identity crisis or multitaskers? *J. Immunol.* **202**, 637–644 (2019).

Acknowledgements

We thank M.B. Brenner for helpful scientific conversations regarding this work. K.A.L. and J.B.K. are each supported by award number T32GM007753 from the National Institute of General Medical Sciences. A.N. is supported by award number T32AR007530 from the National Institute of Arthritis and Musculoskeletal and Skin Diseases. D.A.R. is supported by National Institutes of Health (NIH) NIAMS K08 AR072791 and a Career Award for Medical Sciences from the Burroughs Wellcome Fund. A.H.S. is supported by NIH P01 AI039671, P01 CA236749 and P01 AI108545. S.R. is supported by NIH grants U19-AI111224-01, P01AI148102-01A1, U01-HG009379-04S1, 1R01AR063759 and UH2-AR067677.

Author contributions

K.A.L., K.I. and S.R. conceived the study. K.A.L. performed computational analyses with support from J.B.K. and A.N. K.A.L., K.I., S.R., J.B.K., A.N, K.E.P., A.H.J., A.H.S. and D.A.R. contributed to data interpretation. K.A.L., K.E.P., K.I. and S.R. contributed to writing the manuscript. All authors reviewed the manuscript. K.I. and S.R. supervised the study.

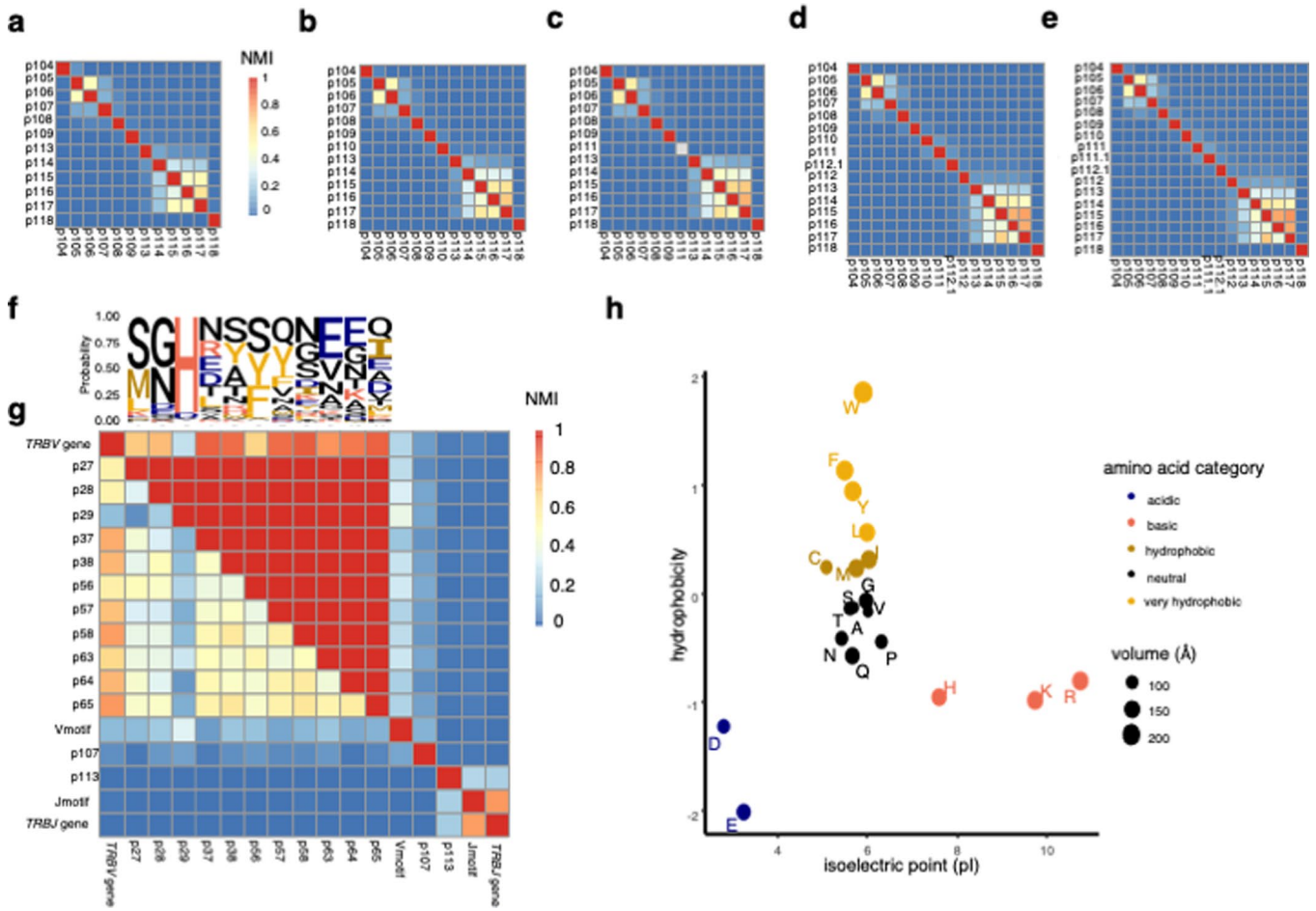
Competing interests
The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41590-022-01129-x.

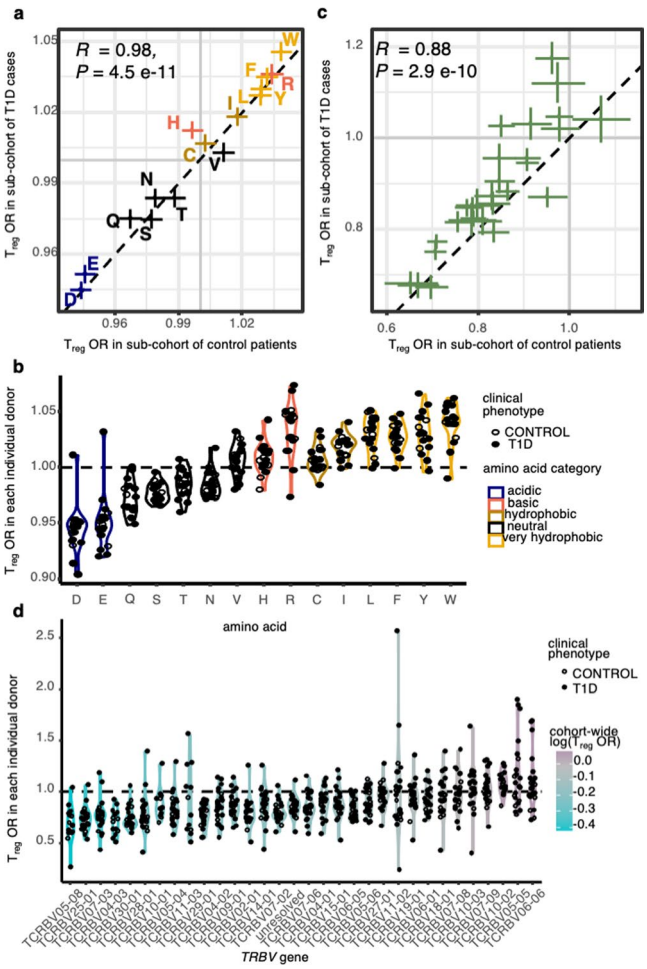
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41590-022-01129-x.

Correspondence and requests for materials should be addressed to Kazuyoshi Ishigaki or Soumya Raychaudhuri.

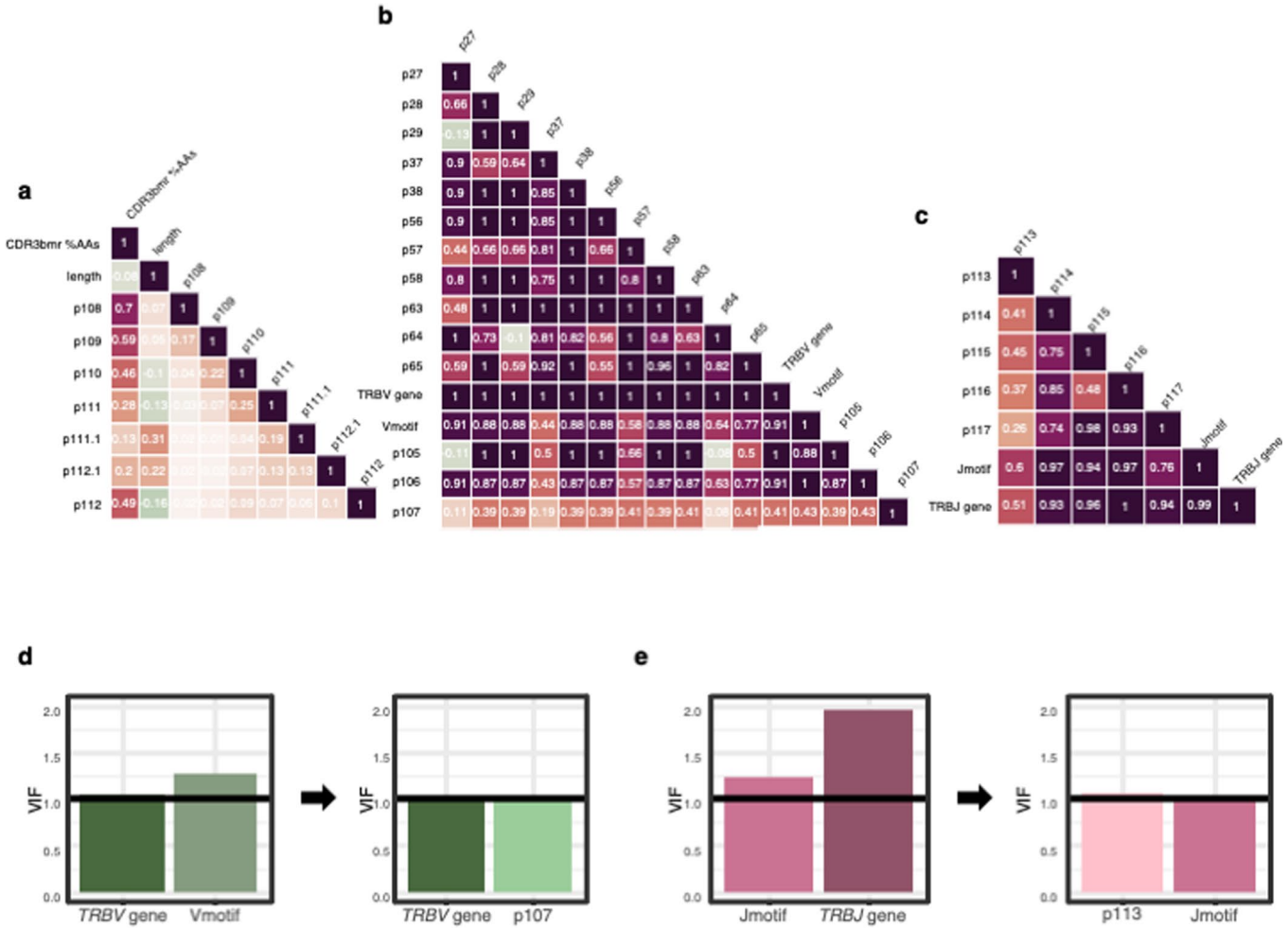
Peer review information *Nature Immunology* thanks the anonymous reviewers for their contribution to the peer review of this work. Zoltan Fehervari was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.



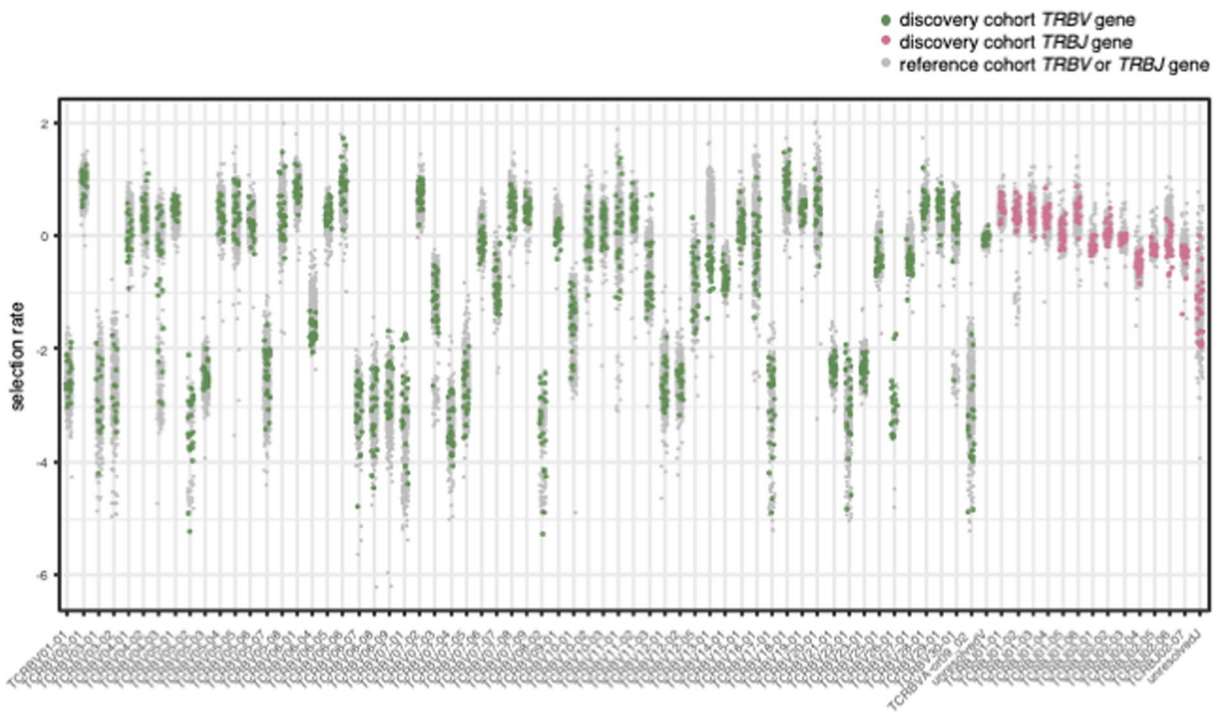
Extended Data Fig. 1 | Mutual information structure of the TCRβ sequence. (a) – (e) Heatmap depicting the mutual information structure of the CDR3β amino acid sequence for CDR3βs of length 12 (a), 13 (b), 14 (c), 16 (d), and 17(e) in the discovery dataset. The lower diagonal features normalized mutual information (NMI) between each pair of TCR positions, while the upper diagonal features the maximum mutual information achieved by conditioning on any other TCR position. NMI color scale for (a)-(e) is provided in (a). **(f)** Probability of each amino acid in each TCR position depicted by a sequence logo. **(g)** Heatmap as in (a) – (e) for CDR1β and CDR2β loop positions as well as TCR features derived from the flanking regions of CDR3β (Methods). **(h)** Categorization of amino acids by isoelectric point and interfacial hydrophobicity (Methods).



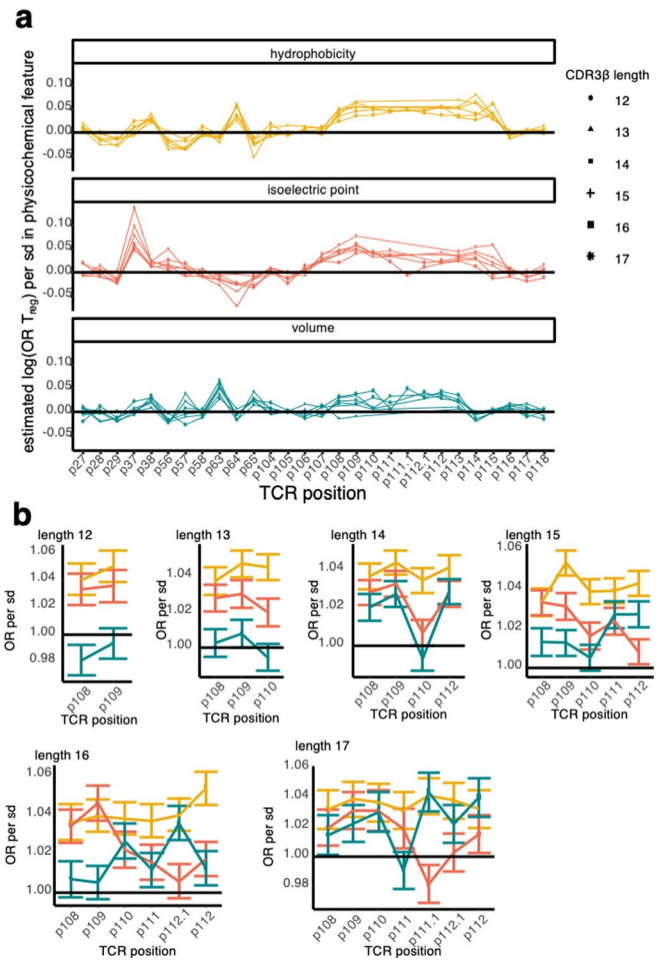
Extended Data Fig. 2 | Consistency of TCR feature effects across individuals and clinical phenotypes. (a) T_{reg} odds ratio per standard deviation increase in CDR3βmr occupancy by each of the 14 relevant amino acids, estimated separately for the T1D cases in the discovery cohort (y axis) and the controls (x axis) **(b)** T_{reg} odds ratio per standard deviation increase in CDR3βmr occupancy by each of the 15 relevant amino acids, estimated separately in each donor. **(c)** T_{reg} odds ratio for the usage of each *TRBV* gene relative to the reference gene *TRBV05-01*, estimated separately for the T1D cases in the discovery cohort (y axis) and the controls (x axis) **(d)** T_{reg} odds ratio for the usage of each *TRBV* gene relative to the reference gene *TRBV05-01*, estimated separately in each donor. *P* values in (a) and (c) are calculated by a two-sided t-test with Fischer transformation on Pearson's *R*.



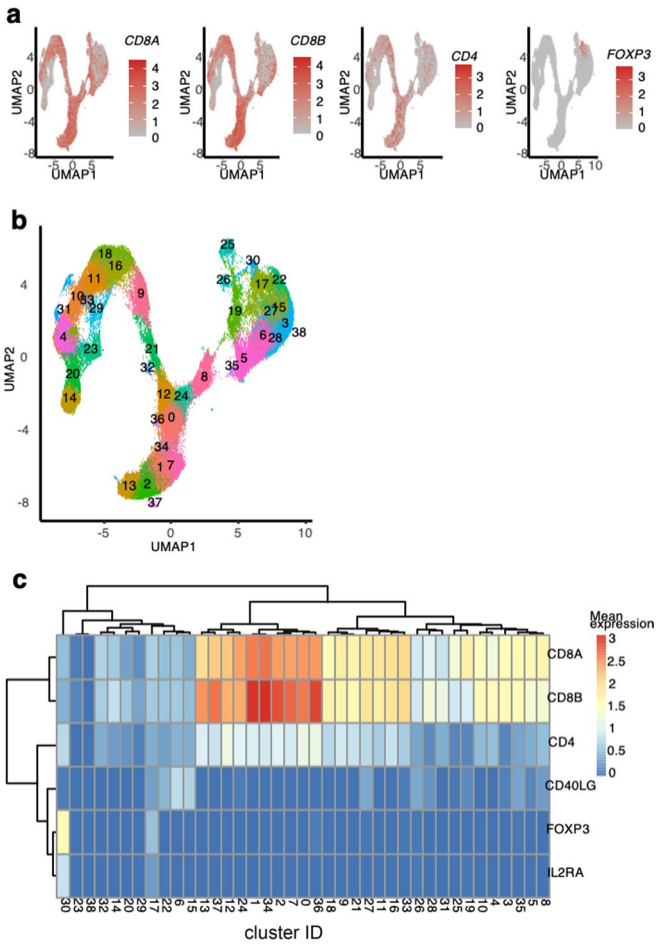
Extended Data Fig. 3 | Multicollinearity analysis. (a)-(c) Maximum Pearson's correlation observed between each pair of TCR features in the discovery dataset, for all possible combinations of amino acid-based TCR feature values (Methods). Heatmaps are separated by TCR region: (a) CDR3βmr, (b) *TRBV*-encoded (CDR1β loop, CDR2β loop, and the V-region of CDR3β) and, (c) *TRBJ*-encoded. **(d)** Feature selection for the V-region model based on variance inflation in estimated regression coefficients (Methods); each plot represents a candidate mixed effects logistic regression model jointly modeling the effects of TCR features on the x-axis. Black arrow denotes improvement from the first model to the second model via reduction of the variance inflation factor (VIF). Black horizontal line denotes the ideal VIF: zero inflation compared to a model with uncorrelated features. **(e)** Same as (d), for candidate J-region models.



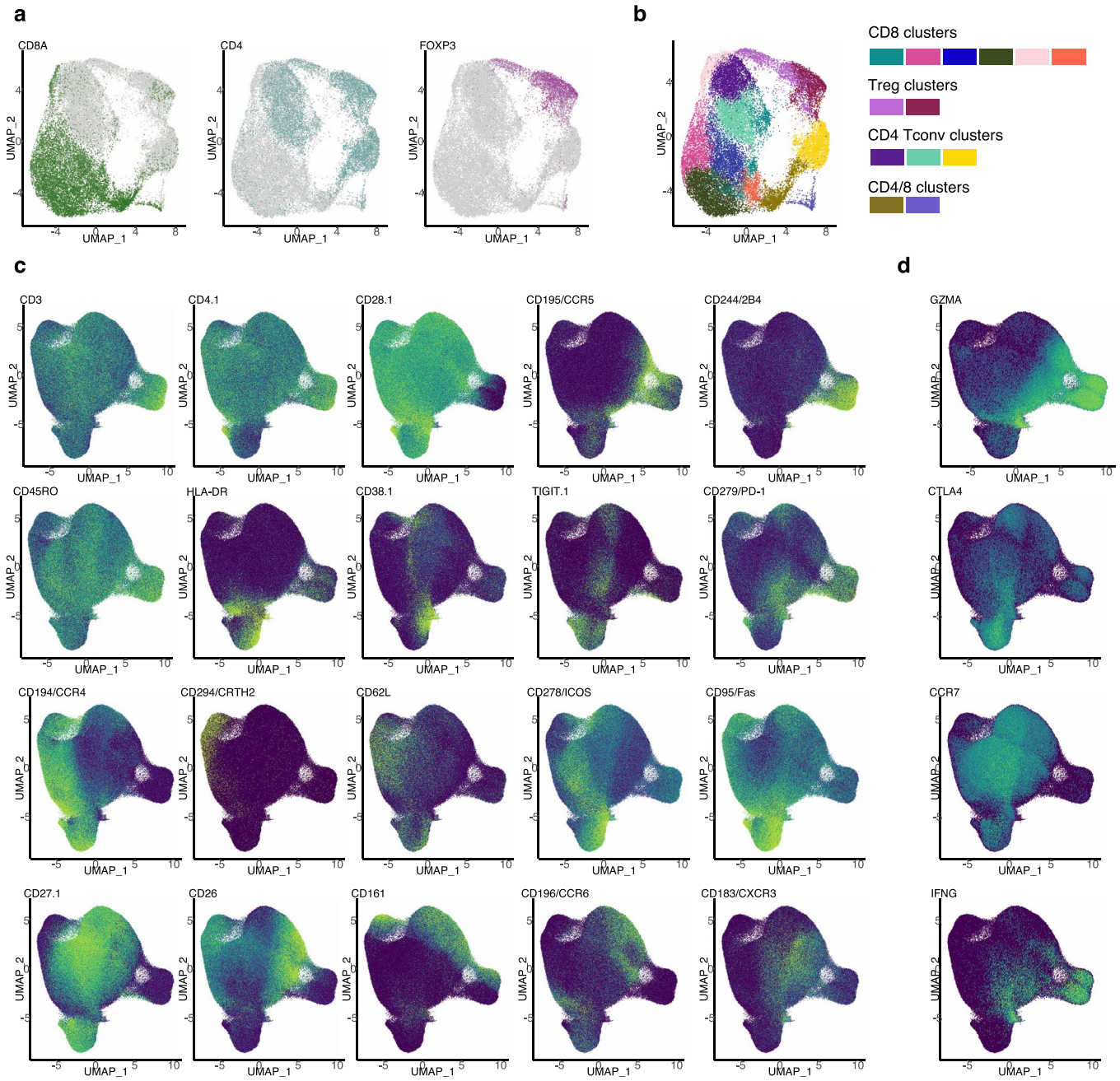
Extended Data Fig. 4 | Thymic selection rates for *TRBV* and *TRBJ* genes. Thymic selection rates for each *TRBV* and *TRBJ* gene in each donor in the discovery cohort and in a reference cohort of 666 healthy donors, inferred by relative gene usage in productive reads versus nonproductive reads (Supplementary Note).



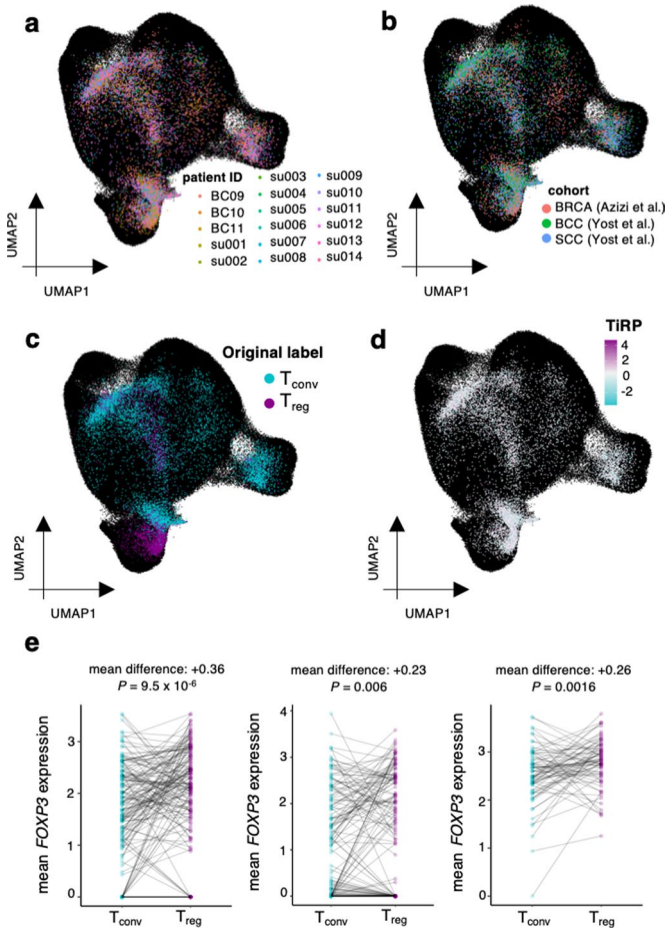
Extended Data Fig. 5 | Estimated effects of physicochemical features at each TCRβ position, stratified by CDR3β length. (a) Estimated log odds ratio for T_{reg} fate per standard deviation of each physicochemical feature at each CDRβ(1-3) loop position in each CDR3β length; features with an estimate > 0 are positively associated with T_{reg} fate while features with an estimate < 0 are negatively associated. For each CDR3β length, all effects were estimated jointly in an L2-regularized logistic regression with a penalty weight tuned via 10-fold cross-validation (Methods). **(b)** T_{reg} odds ratio per standard deviation increase in each physicochemical feature at each CDR3βmr position for each CDR3 length (Methods, Supplementary Table 9). Error bars denote 95% confidence interval for the estimated odds ratio.



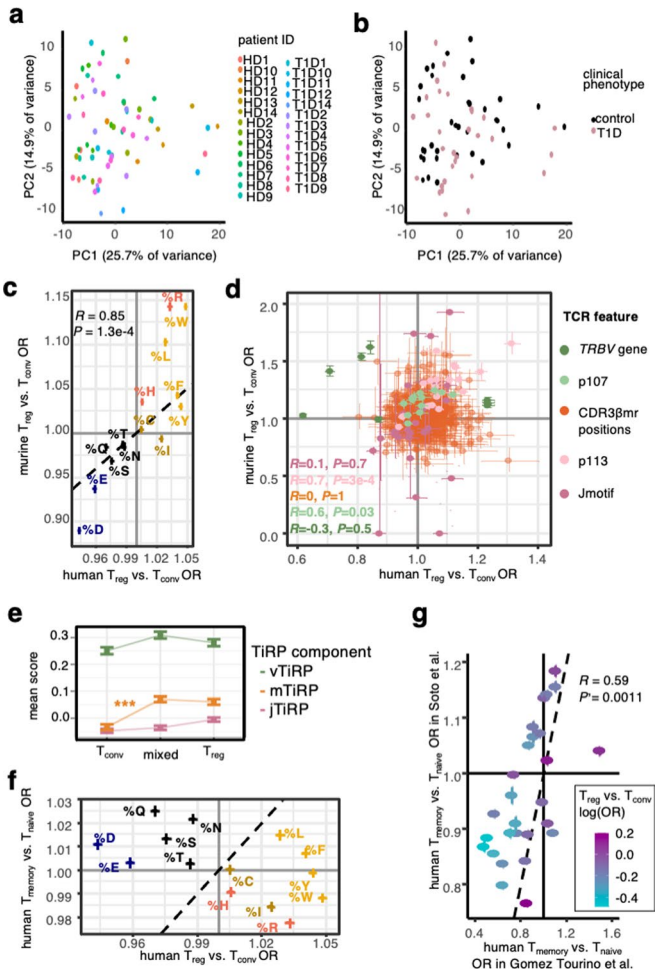
Extended Data Fig. 6 | Cell type identification for thymic T cells. (a) scRNAseq thymic dataset[13] cells arranged in a 2-dimensional embedding by UMAP and colored by normalized expression level of select transcripts; gray (low) to red (high). **(b)** Transcriptional cluster assignments **(c)** Average normalized expression of cell-type-relevant transcripts per cluster.



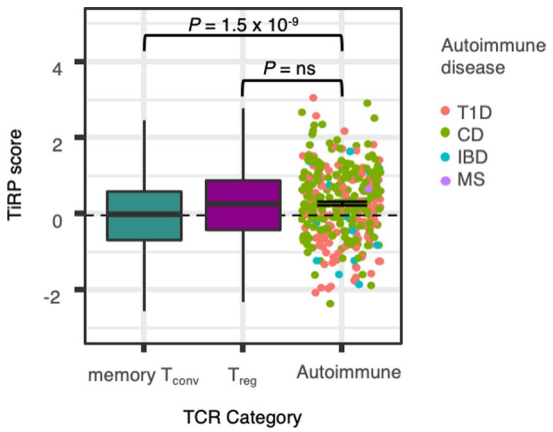
Extended Data Fig. 7 | Cell type identification for tumor microenvironment T cells and reference T cells. (a) Log-normalized *CD8A*, *CD4* and *FOXP3* mRNA expression in T cells from breast tumor biopsies in Azizi et al. 2018, organized into a 2-dimensional embedding by Uniform Maniform Approximation and Projection (UMAP). **(b)** Louvain clustering of breast tumor microenvironment T cells. Broad cell type labels are indicated for each cluster in the surrounding legend. **(c)** Levels of key surface proteins measured by CITE-seq in the CD4 + reference single cell dataset[26] (low = purple, high = light green). Protein levels are normalized by the centered log-ratio (CLR) transformation (Methods). **(d)** LogCP10K-normalized expression levels of key mRNA transcripts in the CD4 + reference single cell dataset[26] (low = purple, high = light green).



Extended Data Fig. 8 | Symphony mapping details. (a) Tumor microenvironment T cells mapped into the reference embedding by Symphony, colored by donor to reveal successful integration of donors. **(b)** same as (a), colored by cancer type to reveal successful integration of cohorts. **(c)** Tumor microenvironment T cells mapped into the reference embedding by Symphony, colored by cell types derived from internal clustering (by Yost et al. for the SCC and BCC samples, and as depicted in Extended Data Fig. 7a-b for the BRCA samples) to show the extent of concordance with Symphony's cell type solutions. **(d)** same as (a), colored by the TiRP score of their TCR. TiRP is scaled such that 0 corresponds to the mean score and one unit corresponds to one standard deviation of held-out bulk sequencing TCRs (Fig. 5c). **(e)** *FOXP3* expression differences between T_{regs} and T_{convs} within mixed clones of three representative donor samples. Each mixed clone is represented by a line connecting the average *FOXP3* expression of Tregs within the clone to the average *FOXP3* expression of T_{convs} within the clone. Each *P* value is computed by a two-sided paired t-test comparing the mean *FOXP3* expression in Tregs to that in T_{convs} within each mixed clone.



Extended Data Fig. 9 | Further analysis of principal components, murine Tregs, and human memory Tconv. (a) 67 samples from the replication cohort colored by donor ID and arranged by principal component space according to variation in TCR sequence feature frequencies. **(b)** Same as (a), colored by donor clinical phenotype. **(c)** Replication of CDR3βmr percent composition of amino acid effects in mice. Error bars correspond to 95% confidence intervals for ORs. Amino acids are colored by physicochemical categories defined in Extended Data Fig. 1h. **(d)** Lack of mouse-human correspondence for position-specific TCR feature effects. TCR features are colored by type; error bars denote OR 95% confidence intervals. Murine *TRBV* genes were mapped to their human homologs for comparison, only those with a human homolog are shown (Methods). **(e)** Mean TiRP component scores for $CD4^+$ expanded pure T_{conv}, pure T_{reg}, and mixed clones in the tumor microenvironment[16,17]. Error bars denote standard error of the mean. T_{conv} mTiRP compared to mixed clone mTiRP two-sided Wald test $P = 2.9 \times 10^{-4}$, all other comparisons nonsignificant. **(f)** Overall lack of correspondence between Treg-Tconv OR and memory-naïve OR for CDR3βmr percent composition of amino acids. Error bars correspond to 95% confidence intervals, and amino acids are colored by the scheme in (c). **(g)** Replication of memory T_{conv} – naive T_{conv} *TRBV* gene odds ratios in an independent dataset of sorted memory and naïve T cells from 4 healthy donors[32]. *TRBV* genes are colored by their T_{reg}-T_{conv} odds ratios. For (c), (d), (f), and (h), R = Pearson's correlation coefficient and P values are computed by a two-sided t-test with Fischer transformation. For (e)-(g), human T_{reg}-T_{conv} ORs result from fixed-effect meta-analysis across the discovery and replication cohorts.



Extended Data Fig. 10 | TiRP scoring of autoreactive T cell receptors. TiRP scores of McPAS and VDJdb autoimmune TCRs (points) compared to memory T_{convs} and T_{regs} from the replication dataset held out for testing (boxplots). Each point in the autoimmune category represents one TCR from McPAS or VDJdb, colored by disease Error bar denotes standard error of the mean TiRP for autoreactive TCRs, which is higher than reference memory T_{convs} ($P = 1.5 \times 10^{-9}$, two-sided Wald test), but not significantly different from reference T_{regs} ($P = 0.43$, two-sided Wald test). Within each boxplot, the horizontal lines reflect the median, the top and bottom of each box reflect the interquartile range (IQR), and the whiskers reflect the maximum and minimum values within each grouping no further than 1.5 x IQR from the hinge. T1D = Type 1 Diabetes. CD = Celiac Disease. IBD = Inflammatory Bowel Disease. MS = Multiple Sclerosis.

nature portfolio

Corresponding author(s): Soumya Raychaudhuri and Kazuyoshi Ishigaki

Last updated by author(s): Dec 22, 2021

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used for data collection.
Data analysis	All data were analyzed with open-source software available online and detailed in Methods: cellranger (version 5.0.1, GRCh38-3.0.0), R (version >= 3.6.1). R packages: stats, lme4, broom, broom.mixed, glmnet, CCA, uwot, car, VIF, boot, dplyr, DescTools, infotheo, singlecellmethods, harmony, symphony, Seurat, class, ggseqlogo, ggplot2, ggrastr, ggpubr, ggrepel, pheatmap, RColorBrewer, pals, patchwork, corrplot2, officer, rvg.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Data analyzed in this study were previously deposited in the following locations:

immuneACCESS

DOI: https://doi.org/10.21417/B73S3K

DOI: https://doi.org/10.21417/B7C88S

DOI: https://doi.org/10.21417/AMT2019EJI

DOI: https://doi.org/10.21417/CS2020CR

DOI: https://doi.org/10.21417/B7001Z

Gene Expression Omnibus (GEO)

GSE158769

GSE123813

GSE114724

Github

URL: https://github.com/aleksobrad/humanized-mouse-data

Zenodo

DOI: https://doi.org/10.5281/zenodo.3711134

ArrayExpress

E-MTAB-8581

10X Genomics

URL: https://cf.10xgenomics.com/supp/cell-exp/refdata-gex-GRCh38-2020-A.tar.gz

McPAS-TCR

URL: http://friedmanlab.weizmann.ac.il/McPAS-TCR

VDJdb

URL: https://vdjdb.cdr3.net

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	To allow for ample statistical power, we conducted our main analyses with bulk sequencing data. Aggregating several publicly available datasets of TCR sequences from human peripheral blood flow-sorted into Treg and Tconv populations amounted to more than 2.4E7 TCR observations, affording >0.99 power to detect a mean difference as small as 0.005 standard deviations.
Data exclusions	For bulk sequencing data, we considered each unique TCR sequence within each donor to be a single observation and ignored duplicates. To define a tractable number of TCR positions for study, we excluded CDR3b sequences shorter than 12 amino acids or longer than 17 amino acids. For training our Treg vs Tconv statistical model, we included only data from donors in which both T cell phenotypes of interest were

	collected. To focus on Treg-Tconv distinctions, we excluded TCR sequences that were observed as both Treg and Tconv within the same donor.
Replication	To replicate our findings, we fit mixed-effects logistic regression models for the V-, J-, and middle regions of the TCR with the same covariates in an independent cohort of bulk TCR sequencing from flow-sorted Tregs and Tconvs. We further replicated our findings by applying the TiRP scoring system to two more independent cohorts of bulk TCR sequencing as well as three independent cohorts of scRNA sequencing. The findings replicated in six out of six cohorts examined.
Randomization	We considered each TCR sequence to be randomly assigned by the process of V(D)J recombination in the thymus. To control for possible confounds in this pseudo-randomization, we model donor and batch ID as random effects and a donor-individualized thymic selection rate parameter (Supplementary Note) as a fixed effect.
Blinding	Data from each individual were analyzed in the same manner. For our outcome of interest (Treg vs Tconv), the level of observation is the T cell, for which blinding is not meaningful.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging