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

Kaitlyn A. Lagattuta1,2,3,4,5, Joyce B. Kang1,2,3,4,5, Aparna Nathan1,2,3,4,5, Kristen E. Pauken6,7, Anna Helena Jonsson3, Deepak A. Rao3, Arlene H. Sharpe5,6,7, Kazuyoshi Ishigaki1,2,5,8,10, and Soumya Raychaudhuri1,2,3,4,5,9,10

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 (Treg) 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 (Tconv) phenotype1–8. Following thymic selection, Treg cells and Tconv cells tend to become Tconv cells, and those whose TCRs have higher affinity to pMHC+ cell surface molecules tend to gain the Treg phenotype1–8. Following thymic selection, the development of CD4+ T cells into either Treg or Tconv cells is largely driven by TCR signal strength, but the majority of extant work has focused on co-stimulatory signals, exert additional influence9,10.

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 Treg cells and Tconv cells is minimal compared to T cell samples of the same phenotype10. The distinguishing features of Treg and Tconv 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 5.7×106 TCRβ chain sequences from six published datasets. Using multiple mixed effects logistic regression models, we quantified the effect of each TCR feature on Treg 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 Treg cells and Tconv cells in two human cohorts of TCRβ chains from flow-sorted T cells1,2 (Supplementary Table 2). From these results, we developed a Treg propensity scoring system for the TCR (TiRP) (Fig. 1a). Upon confirming its accuracy in two datasets of thymic T cells1,2,4, we applied TiRP to tumor-infiltrating T cells and found that clone plasticity (the presence of induced Treg cells (iTregs) or exTreg 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 Treg TCR ligand, the self-peptide and the human MHC class II molecule. For these analyses, we calculated human TiRP for (1) murine Treg 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 adaptability 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- and β-chain.

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

Nature Immunology

Such clones may include iTreg cells (Tconv cells that have acquired a regulatory phenotype), exTreg cells. These mixed clones likely represent lineages of T cells that have undergone a peripheral conversion between the regulatory and conventional phenotypes.

The gene, the germline-encoded V and J regions demonstrated sequence conservation and high inter-residue MI (Fig. 2a). While random nucleotide sequences observed in the same individual. We used MI analysis and mixed effects model comparisons to select 606 non-redundant TCR features that had Bonferroni-significant meta-analytic P values. We then applied TiRP to tumor-infiltrating CD4+ cells to study mixed clones: groups of Treg cells and Tconv cells with the same TCR sequences as Treg cells (CD4+CD127−CD25+) and Tconv cells (CD4+CD127+) in the discovery cohort. Calculating the mean percentage of CDR3mr residues occupied by each amino acid yielded strikingly consistent Treg–Tconv differences across donors: phenylalanine, leucine, tryptophan and tyrosine were consistently enriched in Treg cells, while aspartic acid and glutamic acid were consistently enriched in Tconv cells. Categorization of amino acids by physicochemical features showed that hydrophobic amino acids were enriched in Treg cells, while negatively charged amino acids were enriched in Tconv cells (Extended Data Fig. 1a).

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 Treg 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).

Treg cells use specific amino acids in the CDR3mr. We first examined the CDR3mr of Treg cells (CD4+CD127−CD25+) and Tconv cells (CD4+CD127+) in the discovery cohort. Calculating the mean percentage of CDR3mr residues occupied by each amino acid yielded strikingly consistent Treg–Tconv differences across donors: phenylalanine, leucine, tryptophan and tyrosine were consistently enriched in Treg cells, while aspartic acid and glutamic acid were consistently enriched in Tconv cells (Fig. 3a). Categorization of amino acids by physicochemical features showed that hydrophobic amino acids were enriched in Treg cells, while negatively charged amino acids were enriched in Tconv cells (Extended Data Fig. 1a).

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 Treg 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).
To confirm that these effects were consistent across donors and clinical phenotypes, we estimated them in 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 CDR3\text{\text{\textbar}}mr 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 CDR3\text{\text{\textbar}}mr.

We then ran a separate mixed effects model for each CDR3\text{\text{\textbar}}mr 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 CDR3\text{\text{\textbar}}mr amino acid percentages (Methods). We found that each position indeed conveyed additional information regarding the likelihood of T\text{\text{\textbar}}reg fate, but these position-specific effects all together did not explain as much variance as the general amino acid composition of the CDR3\text{\text{\textbar}}mr (Fig. 3c and Supplementary Table 5).

CDR3\text{\text{\textbar}}β 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\text{\text{\textbar}}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\text{\text{\textbar}}reg fate by more than 30% compared to the reference (most common) gene TRBV05-01 (P = 1.3 × 10\[^{−04}\], likelihood ratio test (LRT); Supplementary Table 6). As in the CDR3\text{\text{\textbar}}mr 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\text{\text{\textbar}}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 CDR3\text{\text{\textbar}}β 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 CDR3\text{\text{\textbar}}mr (Fig. 3c). Each TCR region played an important role, with the greatest variance explained per residue in the CDR3\text{\text{\textbar}}mr. Relative gains in variance explained were proportional to fractional occupancy of the TCR, which was dependent on CDR3\text{\text{\textbar}}β 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\text{\text{\textbar}}reg cells are enriched for CDR1\text{\textbar}β charge and CDR3\text{\textbar}β hydrophobicity. We next aimed to localize physicochemical effects underlying CDR3\text{\text{\textbar}}mr residue enrichments to specific TCR positions. At each CDR1\text{\textbar}β–CDR3\text{\text{\textbar}} loop amino acid position, we estimated the effect of hydrophobicity, pI and volume on T\text{\text{\textbar}}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\text{\text{\textbar}}reg cells were enriched for positively charged amino acids at p37 of CDR1\text{\textbar}β (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\text{\text{\textbar}}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, CDR3\text{\text{\textbar}}mr 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<sub>conv</sub> cells and T<sub>reg</sub> cells. a, Percentage of select amino acids in the CDR3<sub>mr</sub>, 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), CDR3<sub>mr</sub> (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 = 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 CDR3<sub>mr</sub>. As CDR3<sub>mr</sub> length increases, CDR3<sub>mr</sub> occupies a greater proportion of the TCR (fraction of amino acid residues) at the expense of V and J region proportions. Select CDR3<sub>mr</sub> lengths (number of amino acids) are labeled to show the direction of these trends. The x axis corresponds to the number of possible values for the qualitative feature (TRBV<sub>gene</sub> = 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%. d, Variance explained by each TCR region for different CDR3<sub>mr</sub> lengths. As CDR3<sub>mr</sub> length increases, CDR3<sub>mr</sub> occupies a greater proportion of the TCR (fraction of amino acid residues) at the expense of V and J region proportions. Select CDR3<sub>mr</sub> lengths (number of amino acids) are labeled to show the direction of these trends. The x axis corresponds to the number of possible values for the qualitative feature (TRBV<sub>gene</sub> = 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%.
led to a 2.5% (L17, p113) to 6.3% (L12, p113) increase in odds of T<sub>reg</sub> fate (odds ratio (OR) = 1.025, 95% CI = 1.011–1.039 and Wald test P = 2.7 × 10<sup>-4</sup> for L17, p113; OR = 1.063, 95% CI = 1.051–1.074 and Wald test P = 5.2 × 10<sup>-28</sup> for L12, p113; Extended Data Fig. 5 and Supplementary Table 9). Although highly consistent across samples, this effect was subtle: average CDR3<sub>mr</sub> hydrophobicity was 0.08 s.d. higher in T<sub>reg</sub> cells than in T<sub>conv</sub> cells (Fig. 4c; OR = 1.08, 95% CI = 1.076–1.083; Wald test P = 2.3 × 10<sup>-23</sup>). Sensitivity analyses revealed that p37 charge and CDR3<sub>mr</sub> 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<sub>reg</sub> fate, we generated a sequence logo representation of the CDR3<sub>mr</sub> β<sub>j</sub>-CDR3<sub>β</sub> loop position; features with an estimate >1 are positively associated with T<sub>reg</sub> fate, while features with an estimate <1 are negatively associated. ORs denote the change in T<sub>reg</sub> odds per s.d. increase in the given physiochemical feature at the given TCR position. Within each CDR3<sub>β</sub> 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 CDR3<sub>β</sub> lengths. Vertical lines denote the boundaries of each CDR3<sub>β</sub> loop. Features with an estimate >1 are positively associated with T<sub>reg</sub> fate, while features with an estimate <1 are negatively associated. ORs 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<sub>reg</sub> cells that recognize self-antigens<sup>16</sup>. Importantly, we show that this principle extends beyond p109–p110 throughout the stretch of CDR3<sub>mr</sub> 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<sub>reg</sub> development.

**Reproducing TCR associations in an independent dataset.** Having identified TCR features associated with T<sub>reg</sub> identity, we next sought to validate them in a public dataset of TCR<sub>β</sub> sequences from the peripheral blood of 16 donors<sup>12</sup> (‘replication cohort’; Supplementary Table 2). Though a different distribution of
Fig. 5 | TiRP: Treg 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 Treg propensity (TiRP) 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 Treg status (OR = 1.231, 95% CI = 1.237–1.235, LRT P = 4.0 × 10^{−25}). d, Validation of TiRP in single-cell RNA sequencing (scRNA-seq) of CD4+ tumor microenvironment T cells^{19,20} (n = 27,721 cells). Each unit increase in TiRP (corresponding to 1-s.d. increase in CDR3β mr percentage for a given amino acid). Colors for amino acids correspond to Extended Data Fig. 1h. e, Validation of TiRP in scRNA-seq of Thymic T cells (n = 60,424 cells). Among developing thymocytes, each unit increase in TiRP was associated with a 9% increase in the odds of Treg fate (OR = 1.195, 95% CI = 1.195–1.235, LRT P = 4.0 × 10^{−25}). For d and e, error bars outline 95% CIs for TiRP, 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 Treg status (OR = 1.12, 95% CI = 1.11–1.13, LRT P = 3.1 × 10^{−177}).

Developing TiRP: a Treg 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 TiRP association effect sizes of independent sequence features in all three TCR regions (Methods). We used meta-analytic effect size features are robust not only to different tissue sources but also to technical differences in sorting and sequencing protocols.
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. 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 Treg cells, blue for clones containing only Tconv cells and yellow for clones containing both Treg and Tconv cells (‘mixed’ clones). f. TiRP scores of Treg, Tconv and mixed expanded clones from held-out bulk sequencing data; P = 2.0 × 10^{-10} for the mixed-Treg difference and P = 9.1 × 10^{-10} for the mixed-Tconv difference. g. Scores as in f for tumor-infiltrating scRNAseq data; P = 3.0 × 10^{-4} for mixed-Treg difference and P = 0.55 for mixed-Tconv difference. For f and g, vertical bars denote mean and s.e.m. per clone type. h. Correspondence between TiRP score and the Treg-Tconv ratio for each clone. The best fit line is shown in gray; clones are colored by Treg-Tconv 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 Treg:Tconv ratio and TiRP score. For f–h, P values are computed by the LRT between mixed effects logistic regression models (Methods).
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 Treg, 1 in every

**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 naïve 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 CDR3mr (mTiRP). $P$ values for e–f are calculated by a two-sided t-test with Fischer transformation on Pearson’s $R$; mTiRP, 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).
previous observations, there was a nearly twofold increase in Treg
7a,b) and by CD25 and CD127 flow sorting14. Consistent with our
three scRNAseq cohorts (Methods and Extended Data Figs. 6 and
T cell phenotype, as defined by standard mRNA clustering for the
each TCR and assessed whether the TiRP explained variance in
decile in all cohorts (Fig. 5d–f), including the tumor microenviron-
likelihood in the top TiRP decile compared to the bottom TiRP
score and Treg fate (OR

in thymic Treg cells13 confirmed the direct relevance of TiRP to the

for which out-of-frame reads were available (Methods). This indeed abrogated the association between Treg–ness score and Treg 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 datasets13,14,19,20 (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 sorting11. Consistent with our previous observations, there was a nearly twofold increase in Treg 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 × 10⁻²⁵; Fig. 5d and Supplementary Table 13). TiRP elevation in thymic Treg cells13 confirmed the direct relevance of TiRP to the thymus (Fig. 5e; OR = 1.09, 95% CI = 1.05–1.13, LRT P = 8.8 × 10⁻⁷).

Similar results in TCRs from flow-sorted single positive (SP) CD4⁺ thymic T cells14 (Fig. 5f; OR = 1.12, 95% CI = 1.11–1.12, LRT P = 3.1 × 10⁻¹⁷) pinpointed the stage of thymic development in which TiRP promotes Treg fate. Importantly, these SP CD4⁺ thymocytes include T cells observed before negative selection. Because the Treg 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 Treg TCR features in diverse biological contexts, including thymic selection.

TiRP explains Treg plasticity in the tumor microenvironment. We next asked whether TiRP could help to explain Treg plasticity. It is well recognized that naïve Treg, thymic emigrants can be peripherally induced to adopt a regulatory phenotype12,21. Conversely, some Treg cells have been observed to lose FOXP3 expression and adopt a
proinflammatory phenotype (exTreg cells; Fig. 1b). Expanded T cell clones (possessing the same TCR) observed as both Treg cells and Tconv 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+ to standardize cell-type definitions across the two tumor microenvironment datasets by mapping cells of expanded clones from both datasets (12,067 cells) into a common reference atlas of T cell subsets 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 Treg and Tconv 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 Tconv and Treg clones (Methods). In the previously held-out bulk sequencing data, the TiRP scores of mixed clones were significantly greater than those of expanded Tconv clones and less than those of expanded Treg clones (Fig. 6f, mixed–Tconv difference = 0.03, \( P = 2.0 \times 10^{-48} \); mixed–Treg difference = -0.29, \( P = 9.1 \times 10^{-46} \), LRT; Methods). These single-cell data confirmed that Treg cells of mixed clones indeed exhibited greater FOXP3 expression than Tconv 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 Tconv clones (Fig. 6f, mixed–Tconv mean TiRP difference = 0.182, \( P = 3.0 \times 10^{-4} \), LRT; Methods). With the limited extent of Treg expansion, we were underpowered to detect significant differences between mixed and Treg clones in these data (mixed–Treg mean TiRP difference = -0.005, \( P = 0.57 \), LRT). When we quantified clone phenotypes by the proportion of Treg cells and Tconv cells within each clone, however, increasing TiRP corresponded to more Treg-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 Treg–Tconv interactions between that of purely Tconv and Treg clones (Methods). In the pre-existing TCR repertoire dataset, in which all T cell states of interest were available (Methods), we observed a few large clones with a relatively even balance (Fig. 6d,e and Supplementary Table 14).

We tested whether the TiRP score of mixed clones was in between that of purely Tconv and Treg clones (Methods). In the previously held-out bulk sequencing data, the TiRP scores of mixed clones were significantly greater than those of expanded Tconv clones and less than those of expanded Treg clones (Fig. 6f, mixed–Tconv difference = 0.03, \( P = 2.0 \times 10^{-48} \); mixed–Treg difference = -0.29, \( P = 9.1 \times 10^{-46} \), LRT; Methods). These single-cell data confirmed that Treg cells of mixed clones indeed exhibited greater FOXP3 expression than Tconv 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 Tconv clones (Fig. 6f, mixed–Tconv mean TiRP difference = 0.182, \( P = 3.0 \times 10^{-4} \), LRT; Methods). With the limited extent of Treg expansion, we were underpowered to detect significant differences between mixed and Treg clones in these data (mixed–Treg mean TiRP difference = -0.005, \( P = 0.57 \), LRT). When we quantified clone phenotypes by the proportion of Treg cells and Tconv cells within each clone, however, increasing TiRP corresponded to more Treg-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.

To test this idea, we examined TiRP in two complementary datasets: (1) murine Treg TCRs, which recognize self-antigens but are not human MHC restricted, and (2) human memory Tconv TCRs, 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 Tconv cells compared to Treg cells (Fig. 8b, left; \( P = 5.0 \times 10^{-106} \) for Helios* Treg cells and \( P = 0.003 \) for Helios* Treg 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–TCR25 and VDJdb+ (Extended Data Fig. 10; \( P = 1.5 \times 10^{-17} \), 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 Treg cells that expressed Helios, a marker of thymic T cell fate acquisition (Fig. 8b, left). Consistent with our TCR region hypothesis, the TiRP component with the greatest increase between murine Tconv cells and Treg cells was mTiRP (Fig. 8c, left). CDR3\βmr 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 CDR3\βmr features such as hydrophobicity promote Treg cell 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 exTreg cells in the tumor microenvironment and underline the role of interactions between CDR3\βmr and the antigenic peptide in Treg cell fate acquisition.

To understand the role of human MHC, we next compared TiRP in naive and memory Tconv TCRs, which do not strongly recognize self-peptides (Fig. 8a, Supplementary Table 2 and Methods). TiRP was significantly elevated in human memory Tconv cells compared to human naive Tconv 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 Tconv cells (Fig. 8c, right). As expected, large-effect-size TCR features between memory Tconv cells and naive Tconv cells were predominantly TRBV genes (Fig. 8d and Extended Data Fig. 9f), and the extent of each gene’s enrichment in memory Tconv cells correlated with the extent of its enrichment in Treg 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 (Supplementary Tables 2 and 16 and Extended Data Fig. 9g). Thus, as structural interactions in the MHC–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 Treg 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\(^{36}\). 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 Treg cells and Tconv cells to uncover determinants of TCR avidity toward the self-pMHC class II complex. We identified TCR sequence features that are predictive of Treg 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 Treg TCRs but is also moderately elevated in peripherally derived Treg cells. Excitingly, TiRP helped to explain the variable tendency of T cell clones to exhibit a regulatory phenotype. Second, our analyses focused on the TiRP scoring approach, 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 TCRs with ‘weakly interacting amino acids’ (IVYWREL) reflects centrally derived Treg TCRs but is also moderately elevated in peripheral Tconv cells. Perhaps such degenerate ‘stickiness’ allows the Treg to generalize from the self-peptide encountered in the thymus to a larger pool of protected self-antigens.

Importantly, however, CDR3\(\beta\)mr 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 Pf of CDR1p37 encoded by the TRBV gene shapes affinity to conserved sites of MHC class II (ref. \(^{17}\)). While the Treg-promoting effect of hydrophobic CDR3\(\beta\)mr amino acids did not translate to the development of memory Tconv cells, memory Tconv cells and Treg cells exhibited strikingly similar TRBV gene biases compared to the naive repertoire. These results suggest that hydrophobic residues in the CDR3\(\beta\)mr may only be ‘sticky’ toward self-peptides, while Treg-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\(^{35}\). 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., Wirnberger, 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 Treg and inNKT 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 Treg cell development. Immunology 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. 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).

456 NATURE IMMUNOLOGY | VOL 23 | MARCH 2022 | 446–457 | www.nature.com/natureimmunology
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. & Peake, 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: exTreg 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., Josepowicz, S. Z., Arvey, A., Treuting, P. M. & Rudensky, A. Y. Extrathyminic 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 Th17 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− Treg 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 T117 and Treg cells. *Curr. Opin. Immunol.* **21**, 274–280 (2009).

36. Daley, S. R. et al. Cysteine and hydrophobic residues in CDR3 serve as distinct markers of TCR plasticity of TH17 and Treg cells. *Curr. Opin. Immunol.* **21**, 274–280 (2009).

37. Košmrlj, A., Jha, A. K., Huseby, E. S., Kardar, M. & Chakraborty, A. K. How does the thymus design antigen-specific and self-tolerant T cell receptor repertoire? *J. Allergy Clin. Immunol.* **144**, 333–336 (2019).

38. Kochs, A., Jha, A. K., Huseby, E. S., Kardar, M. & Chakraborty, A. K. How does the thymus design antigen-specific and self-tolerant T cell receptor repertoire? *J. Allergy Clin. Immunol.* **144**, 333–336 (2019).

39. Miyazawa, S. & Jernigan, R. L. Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules* **18**, 534–552 (1985).

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2022
Methods

Data preparation. Bulk sequencing data. We downloaded the discovery cohort\(^1\), replication cohort\(^2\), murine cohort\(^3\) and memory cohort\(^4\) sequencing data from the Adaptive Biotechnologies immuneACCESS site (URLs), and we downloaded the thymic bulk sequencing cohort\(^5\) 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 CDR3\(_{mr}\) gaps, and sequences longer than 15 amino acids have extra CDR3\(_{mr}\) positions. We examined only one copy of each CDR3\(_j\) sequence within each individual. Unless explicitly noted, we excluded CDR3\(_j\) reads that were observed in both the \(T_{inj}\) and \(T_{imj}\) 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 \(T_{inj}\) and \(T_{imj}\) TCRs available. For the replication cohort, we restricted our analysis to the 16 donors with both \(T_{inj}\) and \(T_{imj}\) TCRs available.

Single-cell sequencing data. We downloaded scRNAsq tumor microenvironment data\(^6\) from the GTEx Expression Omnibus (GEO) through accession numbers GSE114727, GSE114724 and GSE123814. For the scRNAsq 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\(^7\), 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 \(r\)-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\(^1\) and replication\(^2\) 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 (HDL2, 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 CDR3\(_j\) sequence. We first calculated the conditional MI for all possible trios of CDR3\(_j\) 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’ 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\(^4\) of each CDR3\(_j\) position and the MI\(^5\) between all pairs of CDR3\(_j\) 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\(^1\), 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:

\[
\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + b_0 \]

where \(p\) is the probability that the CDA4\(^\ast\) sorted CDR3\(_{j}\) sequence belongs to a \(T_{out}\), \(b_0\) 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 set to 1 if the sequence is from a PBMC sample, \(X_3\) is an indicator variable for spleen origin, \(X_4\) is an indicator variable for inguinal/irrelevant lymph node origin (pancreatic lymph node as a reference), and \(b_0\) is a modification to the intercept fit to each individual \(i\), normally and independently distributed (NID) with mean 0 and variance \(\sigma^2_0\).

(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:

\[
\log\left(\frac{p}{1-p}\right) = \beta_0 + b_0 + b_{ij} \]

where \(b_0\) is a modification to the intercept fit to each individual \(i\), tissue \(j\) pair, NID with mean 0 and variance \(\sigma^2_i\) 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:

\[
\log\left(\frac{p}{1-p}\right) = \beta_0 + b_0 + b_i + b_{ij} \]

These model comparisons revealed that tissue explained 1.90% of variance as a fixed effect and 1.15% of variance as a random effect (\(P = 1.15 \times 10^{-11}\) fixed and \(P = 4.68 \times 10^{-10}\) random, LRT). However, tissue as a random effect nested within individual explained 6.27% of variance (\(P = 1.32 \times 10^{-18}\), 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.

CDR3\(_{mr}\) mixed effects logistic regression. For each amino acid, we calculated the percentage of CDR3\(_{mr}\) positions occupied by this residue; a percentage of means that the residue is missing for a given TCR, while a percentage of 100 means that the residue is present at every CDR3\(_{mr}\) 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 donor and TCR, and CDR3\(_{mr}\) sequence length 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 CDR3\(_{mr}\) positions occupied by glutamic acid in each TCR explained the most variance, with a 9.7% fall in odds of \(T_{inj}\) fate per additional glutamic acid residue for CDR3\(_{mr}\)s of length 15 (pseudo-\(R^2 = 0.0368\), LRT \(P = 8.37 \times 10^{-24}\), 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-\(R^2 = 0.0422\), LRT \(P = 1.01 \times 10^{-22}\), OR = 0.95, 95% CI = 0.947–0.953). We repeated this process until the remaining amino acid coefficients 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 \(T_{inj}\) and \(T_{imj}\) 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 CDR3\(_j\) 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\(_{x}\) non-reference Vmotifs = 599 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 computing 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\(_{x}\) gene and p107 residue effect sizes. To account for donor-individualized TRBV\(_{x}\) gene thymic selection, we included VGSR as a fixed covariate in this final model (Supplementary Note).

Similarly, to parse the TRBV-encoded region, we asked if the 3′-flanking CDR3\(_j\) region could be represented by a handful of motifs. Indeed, the p2114–p1118 sequences (‘Jmotifs’) present in each donor with a frequency of >0.001 in every donor accounted for 91.5% of TCRs. Computations 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 TRB1 gene identity with the p113 amino acid and (2) joint modeling of Jmol 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 Jmol 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). To make the variance inflation comparable across multiple d.f., we used the generalized variance inflation factor \( GVR \), 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 \( \beta \)-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 CDR\( 3 \mr \) length because over 70% of TCR clones in the discovery training data express a TRBV gene with exactly 11 amino acids in the CDR1\( \beta \) and CDR2\( \beta \) loops. Thus, we estimated the absolute size of the V region to be 15 amino acids (11 + 4 CDR3\( \beta \) amino acids), the size of the J region to be 6 amino acids, and the size of the CDR3\( \beta \)mr to vary with CDR3\( \beta \)mr 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 (Treg versus Tconv) for each TCR 1,000 times. On each permutation, we fit mixed logistic regression models for the CDR3\( \beta \)mr 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 CDR1\( \beta \) 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 Stadiniki et al.,\textsuperscript{16} we used the whole-residue interfacial hydrophobicity scale\textsuperscript{16}. We used pl values from the CRC Handbook of Chemistry and Physics\textsuperscript{42} and volume estimates from IMGT’s conversion of Zamyatin’s\textsuperscript{43} 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 CDR1\( \beta \) loop as a vector of length 3 corresponding to the amino acid’s hydrophobicity, pl and volume. For the Fig. 4d visualization, we included only TCRs with a CDR3\( \beta \)mr of variable length, ranging from 2 amino acids in CDR3\( \beta \)mr to 7 amino acids in CDR3\( \beta \)mr, with a mean 0 and variance 1 for each length–position combination.

All model comparisons were computed by the LRT. As depicted in Fig. 2b, the CDR3\( \beta \)mr is of variable length, ranging from 2 amino acids in CDR3\( \beta \)mr of length 12 to 7 amino acids in CDR3\( \beta \)mr of length 17; \( \frac{\beta}{2} \) pairs of CDR3\( \beta \)mr residues in length 12 + \( \frac{\beta}{2} \) pairs of CDR3\( \beta \)mr residues in length 13 + \( \frac{\beta}{2} \) pairs of CDR3\( \beta \)mr residues in length 14 and so forth to \( \frac{\beta}{2} \) pairs of CDR3\( \beta \)mr residues in length 17 totals to 56 total pairs of CDR3\( \beta \)mr 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 x 9 interactive effects, the Bonferroni significance threshold for these meta-analytic \( P \) values was 0.05/\((9 \times 56) + 606\) = 4.5 x 10\(-5\).

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 directionality of the usage of Valine V3.14 TRBV in the CDR3jmr (Fig. 5a and Extended Data Fig. 2b), we included only 14 amino acid percent covariates in our final CDR3jmr 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 TRIP 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) (24) and that no threshold could explain the variance explained than the Bonferroni-corrected threshold (0.05/612 TCR features), resulting in 25 TRBV genes, 23 motifs, 4 CDR3β lengths, 14 CDR3jmr amino acid percentages and 142 position-specific features relevant to TRIP computation (Supplementary Table 12).

Testing TRIP in held-out donors from bulk sequencing cohorts. To test TRIP in bulk sequencing data, we scored each unique productive TCR in donors held out from both TRIP training and calibration (discovery cohort donors 6161, 6193, 6207 and 6287 and replication cohort donors HD1, HDJ2 and T1D6). We then tested the association between TRIP and T cell state by comparing the additive variance explained by a mixed effects logistic regression model including TRIP as a fixed covariate to a baseline model containing only donor ID, tissue source and donor–tissue source interaction as random intercepts (LRT). We conducted the same procedure 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 not expressed (Supplementary Table 2). To visualize the resultant 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 Treg cells to Tconv cells between the group of all top decile TCRs and the group of all bottom decile TCRs.

Validating TRIP 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 TRIP score for each clone based on its CDR3β amino acid sequence and TRBV gene. So that TRIP 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. cohort and Park et al. cohort and by RNA expression in the Yost et al. 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 TRIP score in these data, we tested the association between TRIP score and regulatory or conventional cell phenotype. For the Yost et al. basal cell carcinoma cohort, 60.6% of cells in the Yost et al. 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 single-cell data. To accomplish this, we created a CD4+ memory T cell single-cell reference. To construct a reference of tumor-infiltrating T cells with Symphion. Before ascertaining mixed clones in tumor-infiltrating T cells, we standardized Treg and Tconv 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 constructing a reference of tumor-infiltrating cells for each cohort, donor and sample, we utilized Symphion. Because the reference dataset consisted of only memory T cells and our hypothesis was that low-abundance clones were 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 Yost et al. cohort), we used canonical variate embeddings to (1) impute cluster membership for query cells as the nearest neighbors in the reference UMAP embedding (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 TRIP 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 TRIP 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 TRIP scores to those of expanded Treg and Tconv 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 TRBV chains were sequenced from cDNA in the replication cohort, we cannot be sure whether the 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 (Treg naïve Tconv central memory Tconv and stem cell-like memory Tconv). Because there was only one Treg sorted sample for each individual, we could only detect pure Tconv or mixed clones in the replication cohort by this approach. We tested the effect of TRIP score on clone phenotype with mixed effects models as designed in the single-cell analyses.
of individual ID, CDR3β DNA sequence, TRBV gene and CDR3β 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, 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 (Tconv, Treg, 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 Tconv clones and second to compare mixed clones to purely Treg clones.

We then quantified the clone phenotype by taking the natural log transform of the within-clone Tconv/Treg ratio, with one ‘hallucinated’ Tconv and one ‘hallucinated’ Treg 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(Tconv/Treg ratio) per s.d. increase in TiRP score (Fig. 6h; P = 1.6 × 10⁻⁴, LRT).

To check that FOXP3 expression was significantly different between Treg cells and Tconv 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 Tconv and Treg samples from the same donor or in both the Helios+ and Helios− Tconv samples from the same donor were excluded from the following analyses. Clones with member cells in both the naïve Tconv and memory Treg samples from the same donor were labeled with the memory Tconv 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–murine 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 TRBV05-01 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). We used the function one-hot-encoded the variable into a binary vector equal to the length of possible TRBV gene component of the TiRP score in murine data.

To analyze the effect of TiRP score on clone phenotype (Tconv, Treg, 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 Tconv clones and second to compare mixed clones to purely Treg clones.

To model the effect of TiRP score on clone phenotype (Tconv, Treg, 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 Tconv clones and second to compare mixed clones to purely Treg clones.

To model the effect of TiRP score on clone phenotype (Tconv, Treg, 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 Tconv clones and second to compare mixed clones to purely Treg clones.

To model the effect of TiRP score on clone phenotype (Tconv, Treg, 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 Tconv clones and second to compare mixed clones to purely Treg clones.

To model the effect of TiRP score on clone phenotype (Tconv, Treg, 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 Tconv clones and second to compare mixed clones to purely Treg clones.

To model the effect of TiRP score on clone phenotype (Tconv, Treg, 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 Tconv clones and second to compare mixed clones to purely Treg clones.

To model the effect of TiRP score on clone phenotype (Tconv, Treg, 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 Tconv clones and second to compare mixed clones to purely Treg clones.

To model the effect of TiRP score on clone phenotype (Tconv, Treg, 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 Tconv clones and second to compare mixed clones to purely Treg clones.

To model the effect of TiRP score on clone phenotype (Tconv, Treg, 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 Tconv clones and second to compare mixed clones to purely Treg clones.

To model the effect of TiRP score on clone phenotype (Tconv, Treg, 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 Tconv clones and second to compare mixed clones to purely Treg clones.

To model the effect of TiRP score on clone phenotype (Tconv, Treg, 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 Tconv clones and second to compare mixed clones to purely Treg clones.

To model the effect of TiRP score on clone phenotype (Tconv, Treg, 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 Tconv clones and second to compare mixed clones to purely Treg clones.

To model the effect of TiRP score on clone phenotype (Tconv, Treg, 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 Tconv clones and second to compare mixed clones to purely Treg clones.

To model the effect of TiRP score on clone phenotype (Tconv, Treg, 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 Tconv clones and second to compare mixed clones to purely Treg clones.

To model the effect of TiRP score on clone phenotype (Tconv, Treg, 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 Tconv clones and second to compare mixed clones to purely Treg clones.
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.

Reprints and permissions information is available at www.nature.com/reprints.
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 Treg 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 Treg 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) Treg odds ratio per standard deviation increase in each physicochemical feature at each CDR3βmrm 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\textsuperscript{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 Manifold 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 dataset26 (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 dataset26 (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 Tregs within the clone to the average FOXP3 expression of Tconv 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 Tconv within each mixed clone.
Extended Data Fig. 9 | Further analysis of principal components, murine T regs, and human memory T conv. (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 Tconv, pure Treg, and mixed clones in the tumor microenvironment16,17. Error bars denote standard error of the mean. Tconv 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 Tconv – naïve Tconv TRBV gene odds ratios in an independent dataset of sorted memory and naïve T cells from 4 healthy donors18. TRBV genes are colored by their $T_{\text{naïve}}$-$T_{\text{conv}}$ 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 $T_{\text{naïve}}$-$T_{\text{conv}}$ 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_{conv} and T_{reg} 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_{conv} ($P = 1.5 \times 10^{-9}$, two-sided Wald test), but not significantly different from reference T_{reg} ($P = 0.43$, 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.
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.

- 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 any 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 5.0.3, GRCh38-3.0.0), R (version >= 3.6.1). R packages: stats, lme4, broom, broom.mixed, gimnet, CCA, uwot, car, VIF, boot, dplyr, DescTools, infotheo, singlecellmethods, harmony, symphony, Seurat, class, ggseqlogo, ggplot2, ggpr, ggpubr, ggrepel, pheatmap, RCColorBrewer, pals, patchwork, corplot2, 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:
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.

- [x] 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-fist.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 Tconv. 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 | Methods |
|----------------------------------|---------|
| n/a | n/a |
| ☒ | ☒ |
| Antibodies | ChiP-seq |
| Eukaryotic cell lines | Flow cytometry |
| Palaeontology and archaeology | MRI-based neuroimaging |
| Animals and other organisms | |
| Human research participants | |
| Clinical data | |
| Dual use research of concern | |