

Letter to the Editor

Cysteine and hydrophobic residues in CDR3 serve as distinct T-cell self-reactivity indices

To the Editor:

The largely stochastic process of T-cell antigen receptor (TCR) gene assembly inevitably produces self-reactive T cells. Although most self-reactive T cells are deleted in the thymus, rare thymocytes survive strong self-antigen engagement and differentiate into type A precursors of CD8 α^+ intestinal intraepithelial lymphocytes (type A IELs) in the thymic cortex¹ or CD4⁺ forkhead box P3–positive regulatory T (Treg) cells in the thymic medulla.² Because TCR sequencing revealed amino acid motifs in complementarity-determining region 3 (CDR3) that are enriched in type A IELs or Treg cells,^{3,4} we reasoned that the expression of these self-reactive TCR motifs might serve as a biomarker to evaluate T-cell self-tolerance.

In human thymus samples we used programmed cell death protein 1 and CD10 to sort type A IELs⁵ and CD8⁺ conventional T (Tconv) cells, whereas CD127 and CD25 were used to distinguish Treg and CD4⁺ Tconv cells (Fig 1, A). Human type A IELs and Treg cells expressed more Helios than CD4⁺ and CD8⁺ Tconv cells (Fig 1, B), which is indicative of stronger TCR signaling. TCR-sequencing (for polymerase chain reaction primers, see Table E3 in this article's Online Repository at www.jacionline.org) revealed that, compared with other lineages, type A IELs used more proximal TCR α variable (*TRAV/Trav*) and joining (*TRAJ/Traj*) gene segments in human subjects (Fig 1, C) and mice (Fig 1, D),⁴ indicating that type A IELs tend to receive a TCR signal early after initiating TCR α recombination.⁵

Compared with CD4⁺ Tconv cells, Treg cells are enriched in hydrophobic amino acid doublets at positions 6 and 7 of the CDR3 in the TCR β chain (CDR3 β), a biomarker of self-reactivity.³ The frequency of sequences carrying these position 6 and 7 doublets (hydrophobic index) was greater in the TCR α and/or TCR β repertoires of type A IELs than in thymic and peripheral Treg cells and CD4⁺ and CD8⁺ Tconv cells in human subjects (Fig 1, E) and mice (Fig 1, F). These findings broaden the scope of the hydrophobic index to encompass both TCR α and TCR β repertoires and reveal heightened self-reactivity in the T-cell population that undergoes strong TCR signaling in the thymic cortex.^{1,5}

The frequency of TCR α and/or TCR β sequences with cysteine within 2 positions of the CDR3 apex (cysteine index) was also greater in type A IELs than in other thymic and peripheral T-cell subsets in human subjects (Fig 1, G) and mice (Fig 1, H).⁴ This was not due to preferential use of TCR gene segments because none of the cysteine residues detected by using the cysteine index in this study were encoded by germline codons; rather, they were encoded by codons spanning the junctions of TCR gene segments. Unlike the hydrophobic index, the cysteine index was not greater in Treg cells than in CD4⁺ Tconv cells, suggesting that T cells with cysteine within 2 positions of the CDR3 apex typically acquire tolerance in the thymic cortex at a stage preceding the onset of Treg cell selection.

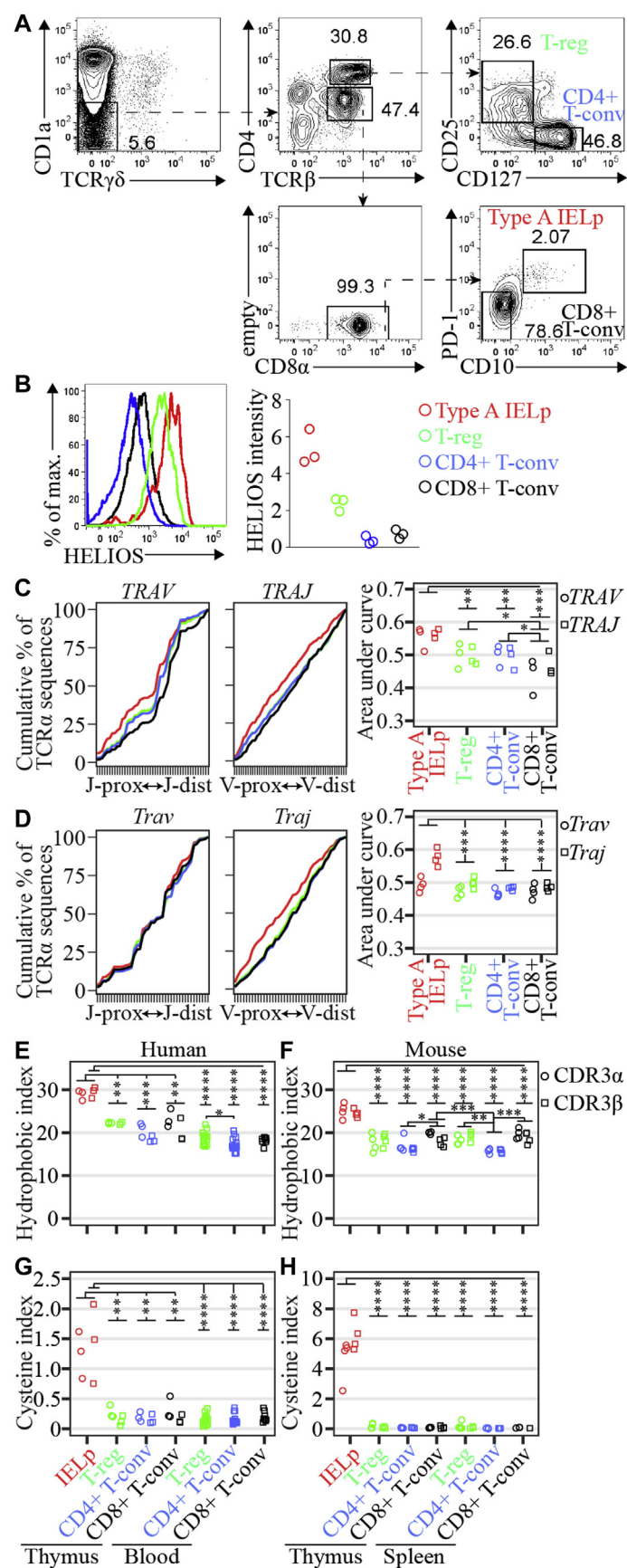
To investigate the effects of reduced thymopoiesis on T-cell self-reactivity indices, we examined patients with biallelic mutations in recombination-activating gene 1 (*RAG1*) or *RAG2*.

In patients with *RAG* mutations, the hydrophobic index was not significantly altered in CDR3 α , but it was increased in CDR3 β (Fig 2, A, and see Table E2 in this article's Online Repository at www.jacionline.org). Because the cysteine index was also not significantly altered in CDR3 α but was increased in the CDR3 β repertoire of Treg cells from patients with *RAG* mutations, subsequent analyses were confined to CDR3 β (Fig 2, B). Significant correlations were detected when each index among unique sequences was plotted as a function of the same index among total sequences (Fig 2, C and D). Samples were on either side of the diagonal for the hydrophobic index (Fig 2, C), whereas those cysteine indices that deviated from the diagonal had a higher index among unique sequences than among total sequences (Fig 2, D), suggesting that clonal expansion is lower than average in cells carrying a cysteine near the CDR3 β apex.

For some mutations, the extent of recombination activity was determined by expressing the mutant *RAG* protein in Abelson virus–transformed *Rag1*^{−/−} or *Rag2*^{−/−} pro-B cells (see Table E1 in this article's Online Repository at www.jacionline.org).⁶ *RAG* function was not correlated with the hydrophobic index (Fig 2, E) but was inversely correlated with the cysteine index (Fig 2, F). Thymic tissue in patients with *RAG* mutations lacks segregated K5⁺ and K8⁺ thymic epithelial cell populations,⁷ indicating immaturity of thymic stroma. The expression level and diversity of self-peptide/MHC complexes presented by such immature thymic tissue might be insufficient to induce tolerance in a normal proportion of nascent self-reactive $\alpha\beta$ TCR⁺ thymocytes. Aberrant differentiation of these thymocytes would account for the high self-reactivity indices observed in peripheral TCR repertoires of most patients with *RAG* mutations.

To explore the self-reactivity of T cells differentiated *in vitro*, we analyzed TCR β repertoires of CD3⁺CD4⁺CD8[−] immature single-positive and CD3⁺CD4⁺CD8⁺ double-positive cells cultured from induced pluripotent stem cells. The cysteine indices of immature single-positive and double-positive samples were low (see Fig E1 in this article's Online Repository at www.jacionline.org), suggesting that deletion by the cysteine-dependent mechanism can occur *in vitro*.

We also sequenced the TCR β repertoires of Treg and CD4⁺ Tconv cells from mice with hypomorphic mutations in *Rag2* (R229Q) or *Rag1* (R972Q, F971L, and R972W).⁸ The 3 missense *Rag1* mutations support different capacities to accomplish V(D)J recombination and T- and B-cell development (R972Q > F971L > R972W), and T-cell infiltrates were detected in target organs of the most severe R972W model (see Fig E2 in this article's Online Repository at www.jacionline.org). The cysteine index was normal in all mouse strains examined (Fig 2, G), suggesting that thymocytes with cysteine in the CDR3 apex are rendered tolerant in the cortex of mice with hypomorphic *Rag1* or *Rag2* mutations. Unlike in human subjects, thymic architecture is partially intact in mice lacking *RAG* function completely, including segregated K8⁺K5[−] (cortical) and K8[−]K5⁺ (medullary) thymic epithelial cell populations.⁹ By contrast, hydrophobic indices were increased in the F971L and R972W mice, with a greater increase in the R972W strain (Fig 2, G). Together with small thymic medullary regions and reduced expression of autoimmune regulator,⁸ the simplest interpretation of these data is that defects in medullary tolerance



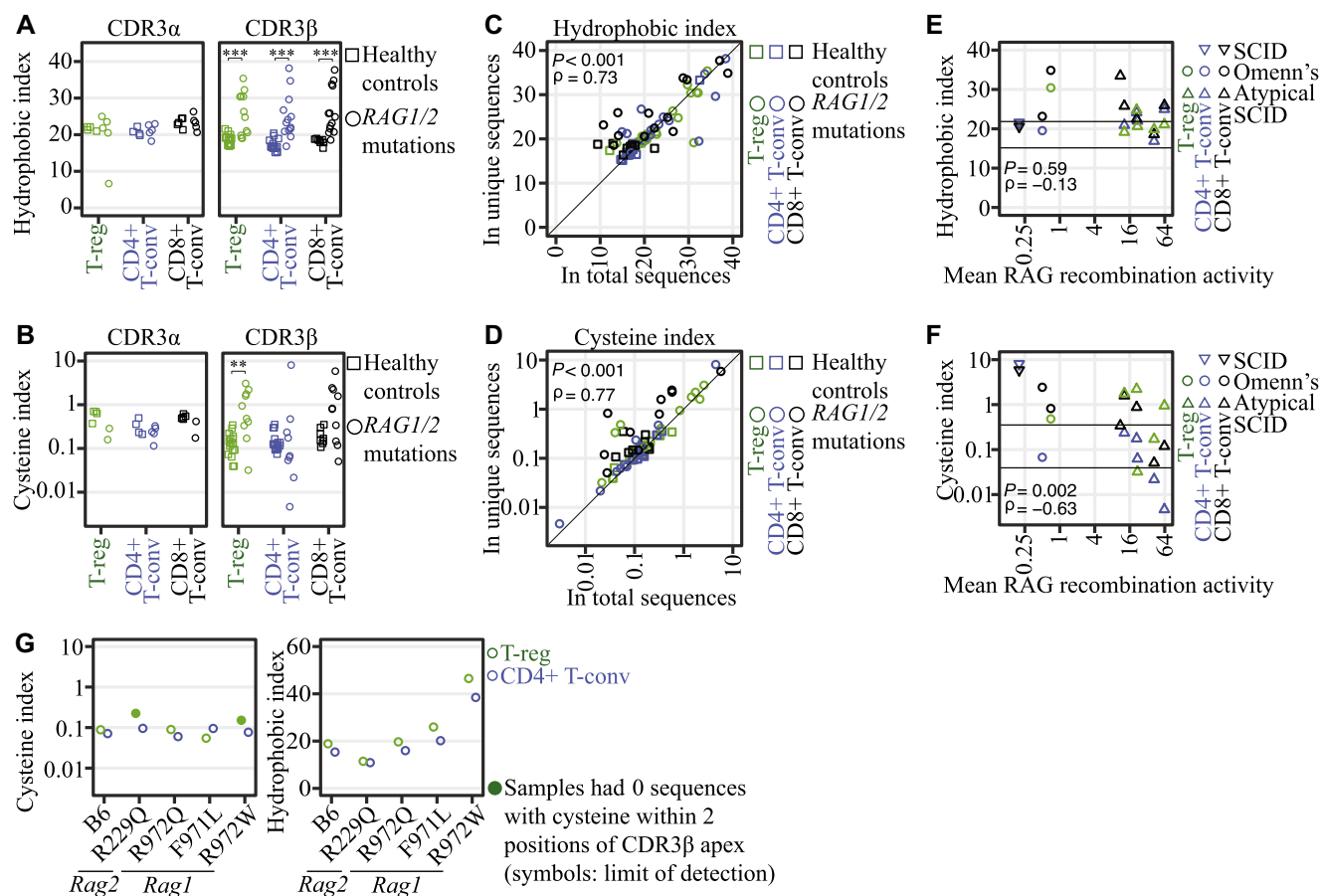


FIG 2. Aberrant self-reactivity indices of T-cell populations selected in the presence of *RAG* mutations. **A** and **B**, Hydrophobic index (Fig 2, A) and cysteine index (Fig 2, B) in CDR3α and CDR3β repertoires of T-cell subsets (*x*-axis) from healthy control subjects (squares) and patients with *RAG* mutations (circles). Fifteen samples from patients with *RAG* mutations were excluded from Fig 2, B, because no sequences were detected with cysteine within 2 positions of the CDR3 apex, and hence the cysteine index was undefined (see Table E2 in this article's Online Repository at www.jacionline.org). **C** and **D**, Hydrophobic index (Fig 2, C) and cysteine index (Fig 2, D) among unique sequences (*y*-axis) as a function of the same index among total sequences (*x*-axis). **E** and **F**, Hydrophobic index (Fig 2, E) and cysteine index (Fig 2, F) as a function of mean RAG recombination activity. **G**, Cysteine index (left) and hydrophobic index (right) in unique TCRβ sequences of CD4⁺ Tconv and Treg cells sorted from B6 control mice or mice with homozygous mutations in *Rag2* (R229Q) or *Rag1* (R972Q, F971L, or R972W). In Fig 2, A and B, 2-way ANOVA was used to compare healthy control subjects with patients with *RAG* mutations adjusted with the Sidak multiple comparisons test. ***P* < .01 and ****P* < .001. In Fig 2, C-F, *P* and *ρ* values were determined by using the Spearman test for correlation. Horizontal lines in Fig 2, E and F, show the ranges observed in healthy control subjects.

underlie the increased hydrophobic indices in F971L and R972W mice.

Both cysteine and hydrophobic indices are highly reproducible in healthy control subjects, revealing a surprising predictability in

normal T-cell selection. Hydrophobic doublets can promote self-reactivity by forming hydrophobic bonds that exclude water molecules from the TCR-peptide/MHC interface, whereas cysteine residues can form stronger disulfide bonds between the

FIG 1. Strong and early TCR activation in type A IELs. **A**, Phenotypes of sorted human thymocyte subsets. **B**, Electronically gated subsets (see color code at right) analyzed for Helios expression, with a summary of mean Helios labeling intensity normalized to CD4⁺CD8⁺ thymocytes in the same sample. **C**, Cumulative percentage of unique sequences that used each *TRAV* or *TRAJ* segment as a function of chromosomal location in the *TRAV* (left) or *TRAJ* (right) locus, with a summary of the area under the curve for each sample (right). **D**, *Trav* and *Traj* use in mouse T-cell subsets⁴ analyzed as in Fig 1, C. **E-H**, Percentage of unique sequences with a self-reactive CDR3 position 6 and 7 doublet (hydrophobic index; see Fig 1, E and F) or percentage of unique sequences with cysteine within 2 positions of the CDR3 apex (cysteine index; Fig 1, G and H; Fig 1, E and G, for human and Fig 1, F and H, for mouse T-cell subsets). Two-way ANOVA was used with repeated measures based on sample donor and TCR chain (α and β), whereas 1-way ANOVA was used to compare TCRβ repertoires. *ns*, *P* > .05. Mouse data in Fig 1, H, have been published⁴ and are included here to enable comparison with human data.

CDR3 apex and the peptide. The available data suggest that an increased cysteine index is a specific biomarker of defective cortical tolerance mechanisms. The hydrophobic index appears more sensitive for detection of a self-tolerance defect but is not specific for either cortical or medullary tolerance mechanisms. Thus the cysteine and hydrophobic indices provide complementary information in the diagnosis and classification of T-cell self-tolerance defects.

Stephen R. Daley, PhD^a
Hui-Fern Koay, PhD^{b,c}
Kerry Dobbs, MSc^d
Marita Bosticardo, PhD^d
Rushika C. Wirasinha, PhD^a
Francesca Pala, PhD^d
Riccardo Castagnoli, MD^d
Jared H. Rowe, MD, PhD^e
Lisa M. Ott de Bruin, MD, PhD^f
Sevgi Keles, MD^g
Yu Nee Lee, PhD^h
Raz Somech, MD, PhD^h
Steven M. Holland, MD^d
Ottavia M. Delmonte, MD^d
Debbie Draper, RN, BSN^d
Sandra Maxwell, RN, BSN^d
Julie Niemela, MSⁱ
Jennifer Stoddard, BSⁱ
Sergio D. Rosenzweig, MD, PhD^j
Pietro Luigi Poliani, MD, PhD^j
Valentina Capo, PhD^k
Anna Villa, MD^{k,l}
Dale I. Godfrey, PhD^{b,c}
Luigi D. Notarangelo, MD^d

From ^athe Infection and Immunity Program, Monash Biomedicine Discovery Institute and Department of Biochemistry and Molecular Biology, Monash University, Melbourne, Australia; ^bthe Department of Microbiology and Immunology, Peter Doherty Institute for Infection and Immunity, and ^cthe Australian Research Council Centre of Excellence in Advanced Molecular Imaging, University of Melbourne, Melbourne, Australia; ^dthe Laboratory of Clinical Immunology and Microbiology, Division of Intramural Research, National Institute of Allergy and Infectious Diseases, and ^ethe Immunology Service, Department of Laboratory Medicine, Clinical Center, National Institutes of Health, Bethesda, Md; ^fthe Division of Hematology-Oncology, Boston Children's Hospital and Dana Farber Cancer Institute, Harvard Medical School, Boston, Mass; ^gthe Division of Immunology, Boston Children's Hospital, Harvard

Medical School, Boston, Mass; ^hthe Division of Pediatric Immunology and Allergy, Meram Medical Faculty, Necmettin Erbakan University, Konya, Turkey; ⁱthe Immunology Service, Edmond and Lily Safra Children's Hospital Sheba Medical Center, Tel Hashomer, Israel; ^jthe Department of Molecular and Translational Medicine, Pathology Unit, University of Brescia, Brescia, Italy; ^kSan Raffaele Telethon Institute for Gene Therapy, Istituto di Ricovero e Cura a Carattere Scientifico (IRCCS) San Raffaele Scientific Institute, Milan, Italy; and ^lIstituto di Ricerca Genetica e Biomedica (IRGB), Consiglio Nazionale delle Ricerche (CNR), Milan Unit, Milan, Italy. E-mail: stephen.daley@monash.edu. Or: luigi.notarangelo2@nih.gov.

Supported by the Monash Biomedicine Discovery Institute, the National Health and Medical Research Council (grant 1107464), and the Division of Intramural Research, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, Maryland.

Disclosure of potential conflict of interest: L. D. Notarangelo is supported by the National Institute of Allergy and Infectious Diseases; has board memberships with the *Journal of Clinical Immunology*, *Clinical Immunology*, and *Frontiers in Immunology*; and has received royalties from UpToDate. The rest of the authors declare that they have no relevant conflicts of interest.

REFERENCES

1. Ruscher R, Kummer RL, Lee YJ, Jameson SC, Hogquist KA. CD8 $\alpha\alpha$ intraepithelial lymphocytes arise from two main thymic precursors. *Nat Immunol* 2017;18:771-9.
2. Fontenot JD, Dooley JL, Farr AG, Rudensky AY. Developmental regulation of Foxp3 expression during ontogeny. *J Exp Med* 2005;202:901-6.
3. Stadinski BD, Shekhar K, Gomez-Tourino I, Jung J, Sasaki K, Sewell AK, et al. Hydrophobic CDR3 residues promote the development of self-reactive T cells. *Nat Immunol* 2016;17:946-55.
4. Wirasinha RC, Singh M, Archer SK, Chan A, Harrison PF, Goodnow CC, et al. $\alpha\beta$ T-cell receptors with a central CDR3 cysteine are enriched in CD8 $\alpha\alpha$ intraepithelial lymphocytes and their thymic precursors. *Immunol Cell Biol* 2018;96:553-61.
5. Verstichel G, Vermijlen D, Martens L, Goetgeluk G, Brouwer M, Thiault N, et al. The checkpoint for agonist selection precedes conventional selection in human thymus. *Sci Immunol* 2017;2.
6. Tirosh I, Yamazaki Y, Frugoni F, Ververs FA, Allenspach EJ, Zhang Y, et al. Recombination activity of human recombination-activating gene 2 (RAG2) mutations and correlation with clinical phenotype. *J Allergy Clin Immunol* 2019;143:726-35.
7. Poliani PL, Facchetti F, Ravanini M, Gennery AR, Villa A, Roifman CM, et al. Early defects in human T-cell development severely affect distribution and maturation of thymic stromal cells: possible implications for the pathophysiology of Omenn syndrome. *Blood* 2009;114:105-8.
8. Villa A, Notarangelo LD. RAG gene defects at the verge of immunodeficiency and immune dysregulation. *Immunol Rev* 2019;287:73-90.
9. Klug DB, Carter C, Crouch E, Roop D, Conti CJ, Richie ER. Interdependence of cortical thymic epithelial cell differentiation and T-lineage commitment. *Proc Natl Acad Sci U S A* 1998;95:11822-7.

<https://doi.org/10.1016/j.jaci.2019.03.022>

Online Repository for the article:

Cysteine and Hydrophobic Residues in the CDR3 Serve as Distinct T Cell Self-Reactivity Indices

Authors: Stephen R. Daley, PhD, Hui-Fern Koay, PhD, Kerry Dobbs, MSc, Marita Bosticardo, PhD, Rushika C. Wirasinha, PhD, Francesca Pala, PhD, Riccardo Castagnoli, MD, Jared H. Rowe, MD, PhD, Lisa M. Ott de Bruin, MD, PhD, Sevgi Keles, MD, Yu Nee Lee, PhD, Raz Somech, MD, PhD, Steven M. Holland, MD, Ottavia M. Delmonte, MD, Debbie Draper RN, BSN, Sandra Maxwell RN, BSN, Julie Niemela, MS, Jennifer Stoddard, BS, Sergio D. Rosenzweig, MD, PhD, Pietro Luigi Poliani, MD, PhD, Valentina Capo, PhD, Anna Villa, MD, Dale I. Godfrey, PhD, and Luigi D. Notarangelo, MD

Contents:

1. **Figure E1.** T cell self-reactivity indices in T cells derived from induced pluripotent stem cells (iPSC)
2. **Figure E2.** T cell infiltrates in mice with homozygous R972W mutations in *Rag1*
3. **Materials and Methods**
4. **Acknowledgments**
5. **Supplementary references**
6. **Table E1.** Clinical phenotype and mutation details of RAG patients
7. **Table E2.** TCR sequencing dataset summarized by sample
8. **Table E3.** Primers used to amplify TCR α and TCR β transcripts from human thymocytes

Figure E1. T cell self-reactivity indices in T cells derived from induced pluripotent stem cells (iPSC). Cysteine index (left) and hydrophobic index (right) in unique TCR β sequences of iPSC-derived T cells from a healthy control and from 3 patients with different *RAG1* mutations (**Table E1**).⁶

Figure E2. T cell infiltrates in mice with homozygous R972W mutations in *Rag1*. Immunohistochemistry labelling of CD3 (brown) in the indicated organs from mice with homozygous R972W mutations in *Rag1*, representative of 12 mice examined.

Materials and Methods

Study Design

This study was motivated by the goal of developing a blood test to measure the effectiveness of self-tolerance induction in the human thymus. The primary hypothesis was that the TCR repertoires of healthy tolerant and non-tolerant thymocytes should be distinguishable based on differential expression of self-reactive amino acid motifs, which were defined in previous studies. TCR sequencing was performed on sorted thymocyte subsets from three subjects in an initial experiment, which was deemed sufficient based on the statistical significance of effects observed. We further hypothesised that self-reactive amino acid motifs would be aberrantly expressed in T cells either differentiated from iPSC *in vitro* or selected *in vivo* in humans or mice with thymic pathology and immunodeficiency associated with previously defined mutations in *RAG1/Rag1* or *RAG2/Rag2*. Analyses of human samples represent an observational study of the TCR sequences expressed by T cell subsets and the number of samples analysed reflects sample availability during the timeframe of the study. Analyses of mouse samples were performed as

controlled laboratory experiments. Investigators were aware of the experimental groups of the samples.

Human samples

Human thymus samples were obtained upon informed consent, in accordance with University of Melbourne Human Research and Ethics committee guidelines (reference numbers 1035100 and 1443540) and Royal Children's Hospital Human Research Ethics Committee Approval (reference number 24131 G). Human peripheral blood samples were obtained upon informed consent, according to NIH IRB-approved protocols 93-I-0119, 06-I-0015, 16-I-N139 and 18-I-0041.

Mice

Mice with homozygous hypomorphic mutations in *Rag1* or *Rag2* were bred with mice bearing a *Foxp3*^{EGFP-Cre} (*Foxp3*^{tm2.1(EGFP/cre)Shori}) allele and the F₁ offspring were backcrossed to the homozygous *Rag*-mutant lines to produce homozygous *Rag*-mutant *Foxp3*^{EGFP-Cre} male N₂ animals from which T cells were sorted. Animal work was conducted in specific pathogen-free conditions, in accordance with the US Public Health Service Policy on Humane Care and Use of Laboratory Animals, with protocols approved by Boston Children's Hospital (protocols 13-08-2472R and 16-05-3176R) and by the National Institute of Allergy and Infectious Diseases Animal Care and Use Committee (protocol LCIM 16E).

Cell sorting

Human thymus samples were cut into small pieces and passed through a 70-μm cell strainer into ice-cold RPMI-1640 medium containing 2 mM EDTA before being washed into PBS plus 2% fetal calf serum (flow cytometry buffer). Single-cell suspensions were incubated with

LIVE/DEAD™ Fixable Aqua (ThermoFisher Scientific, Waltham, MA) plus assortments of the following surface monoclonal antibodies: anti-CD1a Brilliant Violet 421 (clone HI149), anti- $\gamma\delta$ TCR PECy7 or anti- $\gamma\delta$ TCR FITC (clone 11F2 or UC7-13D5), anti-CD4 BUV496 (clone SK3), anti-CD8 α BUV805 (clone SK1), anti-TCR β Brilliant Violet 605 (clone JOVI.1), anti-PD-1 Brilliant Violet 785 (clone EH12.2H7), anti-CD10-FITC (clone HI10a), anti-CD25-APCCy7 (clone M-A251) and anti-CD127-PECy7 (clone A019D5). For each population, 50000 cells (maximum) were sorted into flow cytometry buffer using a BD FACSAria III cell sorter. Cell purity was checked after sorting and found to be 85–98%. Cell pellets were stored at -80°C until RNA isolation. Methods used to sort thymocytes and splenocytes from B6 mice to support the conclusions of Figures 1 and 2 were described previously.¹ T-reg, CD4+ and CD8+ T-conv cells were sorted from human peripheral blood mononuclear cells as described.² Due to lymphopaenia in mice with homozygous *Rag1*/2 mutations, splenocytes from multiple mice were pooled to enable sorting of CD4+ Foxp3^{GFP}– (CD4+ T-conv) and CD4+ Foxp3^{GFP}+ (T-reg) samples. The number of mice pooled per genotype was as follows: B6 controls, n = 4; *Rag2* (R229Q), n = 15; *Rag1* (R972Q) n = 4; *Rag1* (F971L), n = 7 for CD4+ T-conv and n = 15 for T-reg; and *Rag1* (R972W), n = 5 for CD4+ T-conv and n = 11 for T-reg.

Flow cytometry

After surface staining as above, cells were fixed and permeabilized with the eBioscience Foxp3 Fixation/Permeabilization kit according to the manufacturer's instructions to enable intracellular staining with anti-Helios PE (clone 22F6) and anti-Foxp3 Alexa647 (clone 259D). Data were acquired with an LSR II flow cytometer (Becton Dickinson, Franklin Lakes, New Jersey) and analyzed using FlowJo software (FlowJo LLC, Ashland, Oregon).

TCR sequencing

Sorted cells from human peripheral blood and from mouse spleen were sequenced using immunoSEQ (Adaptive Biotechnologies, Seattle, WA). For human thymocyte samples, RNA was isolated using Qiagen RNeasy Mini kits with an elution volume of 22 μL , from which 12 μL were used to synthesise cDNA (Quantitect RT, QIAgen). Using 5 μL of cDNA template per reaction, high-fidelity Q5 polymerase (NEB) was used to PCR amplify TCR β transcripts with a mix of 27 *TRBV*-specific forward primers and a single *TRBC*-specific reverse primer and TCR α transcripts using a mix of 40 *TRAV*-specific forward primers and a single *TRAC*-specific reverse primer (Table E3).³ Forward and reverse primers had distinct 5' overhang adapter sequences that enabled addition of sample-specific indices and P5/P7 sequencing adapters in a second PCR using the Illumina Nextera XT DNA library preparation kit (Illumina Inc., San Diego, CA). Before the second PCR, Agencourt AMPure XP magnetic beads (Beckman Coulter, Brea, CA) were used to enrich amplicons >100 bp. Conditions for the first PCR were 98°C for 5 min, 20 cycles of 98°C for 10 s, 60°C for 30 s and 72°C for 30 s, followed by 72°C for 2 min and for the second PCR were 72°C for 3 min and 98°C for 30 s, 20 cycles of 98°C for 10 s, 63°C for 30 s and 72°C for 30 s, followed by 72°C for 1 min. After determining amplicon concentrations using a QIAxcel capillary electrophoresis machine (Qiagen Pty Ltd, Venlo, Netherlands), equimolar amounts of amplicons were pooled into a single tube, concentrated using Agencourt AMPure XP magnetic beads and then 400–600 bp amplicons were gel-purified (Monarch DNA Gel Extraction Kit, New England Biolabs, Ipswich, MA) before sequencing on an Illumina MiSeq machine, with a short read 1 of 25 bases followed by a read 2 of 275 bases. A summary of the TCR dataset is shown in Table E2.

TCR sequence processing and analysis

For human thymus samples, TCR sequences were aligned to V (D) and J gene segments, and CDR3 nucleotide and amino acid sequences were ascertained, using molecular identifier groups-based error correction (MIGEC) software (version 1.2.6).⁴ To exclude PCR or sequencing errors, low-abundance sequences were excluded based on read count. To determine a “threshold read count” on a per-sample basis, sequences were ranked from most abundant to least abundant based on read count and a cumulative percentage of reads was assigned to each sequence within each sample. For each sample, the “threshold read count” was defined as the read count of the sequence that spanned the 80th percentile of cumulative reads. Sequences with a read count equal to or lower than the “threshold read count” were excluded from further analysis. Subsequent analyses of human thymocyte samples, and the mouse TCR sequence dataset¹ presented in Figure 1, were based solely on unique sequences (ignoring the read count) because read count does not necessarily correspond to the number of cells that carried a given TCR sequence in the original sample. By contrast, values obtained from the Adaptive Biotechnologies immunoSEQ assay are reflective of the number of cells that carried a given TCR sequence in the original sample, enabling parallel analyses of unique and total sequences. All data were filtered to exclude non-functional CDR3 sequences that were either out-of-frame, contained a stop codon or were < 8 amino acids long; the last filter was imposed because a conserved Phe or Try is present at position 6 or 7 of CDR3 sequences that are 6 or 7 amino acids long. We also excluded a total of 9 mouse *Trbv1* sequences with a CDR3 < 10 amino acids long because the germline-encoded Cys at CDR3 position 3 is within 2 positions of the CDR3 apex in these sequences. Self-reactive CDR3 β P6-P7 doublet expression was determined by calculating the percentage of sequences using any of the 175 amino acid doublets identified as promoting self-reactivity.⁵ The same filtered datasets were examined for sequences with a cysteine within 2 positions of the CDR3

apex as described.¹ Briefly, for a CDR3 sequence of n amino acids, the amino acid at the largest position not greater than $(n/2 + 1)$ was designated as the CDR3 apex. For the chromosomal distribution of TCR α gene segment analysis, sequences that mapped to > 1 *TRAV/Trav* segment were excluded and the area under the curve was determined for each sample by dividing the sum of the cumulative percentages by 100, then by the number of distinct *TRAV/Trav* or *TRAJ/Traj* segments in the filtered dataset.

Induced pluripotent stem cell cultures

Methods used to culture, sort and TCR sequence induced pluripotent stem cell-derived T cells have been described.⁶

Immunohistochemistry

Formalin-fixed, paraffin-embedded sections were subjected to hematoxylin and eosin (H&E) staining and immunohistochemistry, as previously described.⁷ In brief, sections were dewaxed and rehydrated, endogenous peroxidase activity was blocked by 0.3% H₂O₂/methanol for 20 min, and heat-induced antigen retrieval was obtained by thermostatic bath treatment using EDTA buffer, pH 8.0, followed by 1-h incubation with primary antibody rabbit anti-CD3 (1:100; Dako). Signal was revealed by using Real EnVision Rabbit-HRP (Dako) detection system followed by diaminobenzidine, and nuclei were counterstained with hematoxylin. Digital images were acquired by a camera (DP70; Olympus) mounted on a microscope (Bx60; Olympus) using Cell^F Imaging software (Soft Imaging System GmbH).

Statistical analysis

The “tidyverse”, “stringr” and “reshape2” packages were used in RStudio software for apex-out alignment and categorization of CDR3 sequences, to produce graphs and to perform

nonparametric Spearman tests for correlation. Other statistical analyses were performed using GraphPad Prism version 7.0a (GraphPad Software, La Jolla California USA) with multiple comparisons tests recommended in the GraphPad Prism software. Figures were made using Adobe Illustrator CC (Adobe Systems Inc., San Jose, CA).

Acknowledgments

We thank the staff of Monash University Micromon and the University of Melbourne DMI and MBC flow cytometry facilities. **Author contributions:** HK, KD, DG, RW, SD, JR, MB and LOdB performed cell sorting; RW and SD prepared TCR sequencing libraries; KD, MB, FP, RC, OD, and PLP performed experiments; YNL, RS, JN, JS and SR performed genetic studies; SK, RS, SH, DD, SM and AV provided samples; RW, HK and DG edited the manuscript; SD performed bioinformatic analyses, prepared the figures and wrote the manuscript with advice and editing by LN. **Data availability:** Thymocyte TCR sequencing datasets are available from the Monash University data repository using the Digital Object Identifier 10.26180/5c484681591c0 (human) or 10.4225/03/5a5376b442289 (mouse).

Supplementary references:

1. Wirasinha RC, Singh M, Archer SK *et al.* $\alpha\beta$ T-cell receptors with a central CDR3 cysteine are enriched in CD8 $\alpha\alpha$ intraepithelial lymphocytes and their thymic precursors. *Immunol Cell Biol.* 2018;96(6):553-61.
2. Rowe JH, Stadinski BD, Henderson LA *et al.* Abnormalities of T-cell receptor repertoire in CD4⁺ regulatory and conventional T cells in patients with RAG mutations: Implications for autoimmunity. *J Allergy Clin Immunol.* 2017;140(6):1739-43 e7.
3. Wang GC, Dash P, McCullers JA *et al.* T cell receptor $\alpha\beta$ diversity inversely correlates with pathogen-specific antibody levels in human cytomegalovirus infection. *Sci Transl Med.* 2012;4(128):128ra42.
4. Shugay M, Britanova OV, Merzlyak EM *et al.* Towards error-free profiling of immune repertoires. *Nat Methods.* 2014;11(6):653-5.
5. Stadinski BD, Shekhar K, Gomez-Tourino I *et al.* Hydrophobic CDR3 residues promote the development of self-reactive T cells. *Nat Immunol.* 2016;17(8):946-55.
6. Brauer PM, Pessach IM, Clarke E *et al.* Modeling altered T-cell development with induced pluripotent stem cells from patients with RAG1-dependent immune deficiencies. *Blood.* 2016;128(6):783-93.
7. Rigoni R, Fontana E, Guglielmetti S *et al.* Intestinal microbiota sustains inflammation and autoimmunity induced by hypomorphic RAG defects. *J Exp Med.* 2016;213(3):355-75.

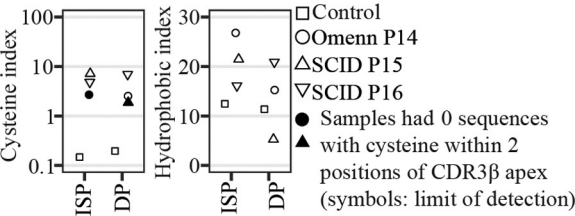
Table E1. Clinical phenotype and mutation details of RAG patients

Patient	Phenotype	Age	Gene and mutation	Mean recombination activity
P1	Atypical SCID	3 y	RAG1: p.H612R;p.A857V	NA
P2	Atypical SCID	9 y	RAG1: p.R474C;p.L506F	63
P3	Atypical SCID	11 y	RAG1: p.L514R;p.L514R	12.9
P4	Atypical SCID	10 y	RAG1: p.R410W;p.H375D	21
P5	Atypical SCID	4 mo	RAG1: p.H375D;p.H375D	41.9
P6	Omenn syndrome	4 mo	RAG1: p.R332X;p.L1025fs*39	NA
P7	Omenn syndrome	1 mo	RAG1: p.S259Afs*5;p.S259Afs*5	0.7
P8	SCID	6 mo	RAG2: p.C41W;p.C41W	0.2
P9	Atypical SCID	39 y	RAG1: p.R108*;p.W522C	21.7
P10	Atypical SCID	3 mo	RAG1: p.C328Y;p.P619L	NA
P11	Omenn syndrome	3 mo	RAG1: p.R396H;p.R396H	0.5
P12	Atypical SCID	2 y	RAG2: p.R73H;p.R73H	11
P13	Atypical SCID	4 y	RAG1: p.Q145Kfs*18;p.N855S	NA
P14	Omenn syndrome	2 mo	RAG1: p.K86Vfs*33;p.E722K	1.32
P15	SCID	2 mo	RAG1: p.N476Kfs*15;p.N476Kfs*15	0.07
P16	SCID	9 mo	RAG1: p.R394W;p.R394W	0.09

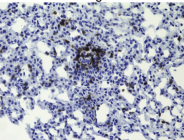
Table E3. Primers used to amplify TCR α and TCR β transcripts from sorted thymocytes

Primer target	Primer sequence
TRAV1	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGCCACCCACATTCTCTTAC
TRAV10	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGAAAGAACTGCACTCTTCAATG
TRAV12-1,12-2,12-3	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAAGATGGAAGGTTTACAGCAC
TRAV13-1	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTCAGACAGTGCCTCAAACCTAC
TRAV13-2	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCAGTGAAACATCTCTCTGTC
TRAV14	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAGGCTGTGACTCTGGACTG
TRAV16	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGTCCAGTACTCCAGACAACG
TRAV17	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCCACCATGAAGTGCAGTTAC
TRAV18	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTGACAGTTCCTCCACCTG
TRAV19	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTGTGACCTTGGACTGTGTG
TRAV2	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCACTCTGTGTCCAATGCTTAC
TRAV20	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTCTGGTATAGGCAAGATCCTG
TRAV21	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAACTTGGTTCTCAACTGCAG
TRAV22	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCTGACTCTGTGAACAATTTGC
TRAV23	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTGCATTATTGATAGCCATACG
TRAV24	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTGCCTTACACTGGTACAGATG
TRAV25	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTATAAGCAAAGGCCTGGTG
TRAV26-1	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCGACAGATTCACCTCCAG
TRAV26-2	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTTCACCTTGCCTTGTAAACAC
TRAV27	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTCACTGTGTACTGCAACTCC
TRAV29	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTGTGAAGGTCCTACATTC
TRAV3	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGATGCACCTATTCAGTCTCTGG
TRAV30	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAGAAGCATGGTGAAGCAC
TRAV34	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGATCTACCATAAACTGCACG
TRAV35	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGACCTGGCTATGGTACAAGC
TRAV36	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGATCTCTGGTTGCCACGAG
TRAV38-1,38-2	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCAGCAGGCAGATGATTCTC
TRAV39	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTCAACCACTTCAGACAGACTG
TRAV4	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGATTATATCACGTGGTACCAACAG
TRAV40	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGGAGGCGGAAATATTAAAGAC
TRAV5	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTACACAGACAGCTCCTCCAC
TRAV6	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTGGTACCGACAAGATCCAG
TRAV7	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGACAATTGCAAGTGGTACAGG
TRAV8-1	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGTCAACACCTTCAGCTTCTC
TRAV8-2,8-4	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAGAGTGAACCTCCTTCCAC
TRAV8-3	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTTTGAGGCTGAATTTAAGAGG
TRAV8-6	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAAACCAAGGACTCCAGCTTC
TRAV8-7	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGATCAGAGGTTTTGAGGCTG
TRAV9-1,9-2	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGAAACCACTTCTTCCACTTG
TRAV41	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTGTGTTATGCTGAGCTCAGG
TRBV2	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTTCACCTCTGAAGATCCGGTC
TRBV10-1,10-2,10-3	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTCCYCCTCACTCTGGAGTC
TRBV11-1,11-2,11-3	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGACTCCACTCTCAAGATCCA
TRBV12-3,12-4,12-5	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCYACTCTGARGATCCAGCC
TRBV13	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCATTCTGAAGTGAACATGAGC
TRBV14	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGATTCTACTCTGAAGGTGCAGC
TRBV15	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGATAACTTCCAATCCAGGAGG
TRBV16	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGAAAGATTTTCACTAAGTGCC
TRBV17	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTGTCTACTGGTACCGACAG
TRBV18	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCGATTTCTGCTGAATTTCC
TRBV19	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTTCCTCTCACTGTGACATCG
TRBV20-1	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGACTCTGACAGTGACCACTGC
TRBV23-1	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGCAATCCTGTCTCAGAAC
TRBV24-1	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGATGGATACAGTGTCTCTCGA
TRBV25-1	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCAGAGAAGGGAGATCTTTCC
TRBV27,28	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTTCYCCCTGATYCTGGAGTC
TRBV29-1	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTCTGACTGTGAGCAACATGAG
TRBV3-1	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAATCTTCACATCAATTCCTTG
TRBV30	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAGAATCTCTCAGCCTCCAGAC
TRBV4-1,4-2,4-3	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCCTGCAGCCAGAAGACTC
TRBV5-1,5-3,5-4	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCTTGGAGCTGGRSGACTC
TRBV5-5,5-6,5-7,5-8	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTCTGAGCTGAATGTGAACG
TRBV6-1,6-2,6-3,6-5,6-6,6-7,6-8,6-9	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGTGTRCCAGGATATGAACC
TRBV6-4	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTGGTTATAGTGTCTCCAGAGC
TRBV7-1,7-2,7-3	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTCTYACTCTGAMGWTCAGCG
TRBV7-4,7-6,7-7,7-8,7-9	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTGRMGATYACGCGACA
TRBV9	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGTACCAACAGAGCCTGGAC
TRAC	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGNNNNNNNTGTTGCTCTTGAAGTCCATAG
TRBC	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGNNNNNNNTTCTGATGGCTCAAACACAG

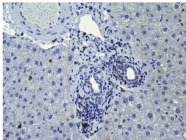
iPSC-derived T cells



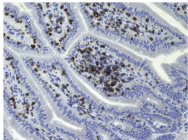
Lung



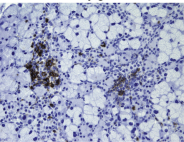
Liver



Small intestine



Salivary gland



Pancreas

