scientific reports

Check for updates

OPEN

Human thymic putative CD8αα precursors exhibit a biased TCR repertoire in single cell AIRR-seq

Marte Heimli[1], Siri Tennebø Flåm[1], Hanne Sagsveen Hjorthaug[1], Pål Marius Bjørnstad[1], Maria Chernigovskaya[2], Quy Khang Le[2], Xavier Tekpli[1], Victor Greiff[2] & Benedicte Alexandra Lie[1]✉

Thymic T cell development comprises T cell receptor (TCR) recombination and assessment of TCR avidity towards self-peptide-MHC complexes presented by antigen-presenting cells. Self-reactivity may lead to negative selection, or to agonist selection and differentiation into unconventional lineages such as regulatory T cells and CD8αα T cells. To explore the effect of the adaptive immune receptor repertoire on thymocyte developmental decisions, we performed single cell adaptive immune receptor repertoire sequencing (scAIRR-seq) of thymocytes from human young paediatric thymi and blood. Thymic *PDCD1*$^+$ cells, a putative CD8αα T cell precursor population, exhibited several TCR features previously associated with thymic and peripheral *ZNF683*$^+$ CD8αα T cells, including enrichment of large and positively charged complementarity-determining region 3 (CDR3) amino acids. Thus, the TCR repertoire may partially explain the decision between conventional vs. agonist selected thymocyte differentiation, an aspect of importance for the development of therapies for patients with immune-mediated diseases.

Abbreviations

aa	Amino acid
AIRR	Adaptive immune receptor repertoire
APC	Antigen presenting cell
BBKNN	Batch balanced K nearest neighbour
CDR3	Complementarity-determining region 3
CoNGA	Clonotype neighbour graph analysis
CTL	Cytotoxic T cell
DN(P)	Proliferating $CD4^-CD8^-$ double negative
DN(Q)	Quiescent $CD4^-CD8^-$ double negative
DP(P)	Proliferating $CD4^+CD8^+$ double positive
DP(Q)	Quiescent $CD4^+CD8^+$ double positive
GEX	Gene expression
IEL	Intraepithelial lymphocyte
imhc	Independent of peptide:MHC
mait	Mucosal-associated invariant T cell
MHC	Major histocompatibility complex
NKT	Natural killer T cell
P_{adj}	Adjusted P value
PBMC	Peripheral blood mononuclear cells
scAIRR-seq	Single cell adaptive immune receptor repertoire sequencing
SP	$CD4^+CD8^-$ or $CD4^-CD8^+$ single positive
TCM	T central memory
TCR	T cell receptor
TEM	T effector memory
T_{reg}	Regulatory T cell

[1]Department of Medical Genetics, University of Oslo and Oslo University Hospital, 0424 Oslo, Norway. [2]Department of Immunology, University of Oslo and Oslo University Hospital, 0372 Oslo, Norway. ✉email: b.a.lie@medisin.uio.no

UMAP	Uniform manifold approximation and projection
V(D)J recombination	Variable, (Diversity), and Joining gene segment recombination

Diversity of the adaptive immune cell receptor repertoire (AIRR) is ensured by genetic recombination of Variable (V), Joining (J), and Diversity (D) gene segments[1,2]. During thymic T cell development, V(D)J recombination of β, γ, and δ T cell receptor (TCR) chains occurs at the $CD4^{-}CD8^{-}$ double-negative (DN) stage, while TCRα chain recombination occurs later at the $CD4^{+}CD8^{+}$ double-positive (DP) stage[1,3]. TCR recombination is coupled to selection checkpoints in order to restrict maturation to thymocytes bearing in-frame, functionally recombined, TCRs without self-reactive capabilities[4]. Despite the selection processes, a fraction of self-reactive thymocytes escape to the periphery, conferring risk of autoimmune responses[5,6].

Escaped self-reactive T cells may be held in check by unconventional, agonist selected T cell populations such as regulatory T cells (T_{regs})[7]. Agonist selected populations are of high interest due to their immune-modulating functions and therapeutic potential[8–11], however, knowledge regarding thymic heterogeneity and development of agonist selected T cells remains incomplete[12–14]. In similarity to developing T_{regs}, thymic precursors of CD8$\alpha\alpha^{+}$ intraepithelial lymphocytes (IELs) have been suggested to undergo agonist selection[15,16].

Previously, we have studied gene expression (GEX) in cell populations in young paediatric thymi, in order to explore the influence of the local cellular milieu for thymic agonist selection[17]. Since the TCR could also play a role in determining divergence towards an agonist-selected T cell fate, we here performed single cell AIRR sequencing (scAIRR-seq) to assess the TCR repertories of the same cells. The current results indicated changes in the TCR repertoire during developmental progression of pre-selection DP thymocytes. In addition, we found biases in the TCR repertoire for a thymic *PDCD1*$^{+}$ putative CD8$\alpha\alpha$ T cell precursor population that have previously been described for *ZNF683*$^{+}$ CD8$\alpha\alpha$ T cells. Our results add insights regarding how the TCR influence decision checkpoints during thymocyte development, in particular the decision between agonist versus conventional selection.

Methods

Study population

Thymic tissues and EDTA blood samples were collected from young paediatric donors (3 male, 2 female, age span 7 days–13.5 months) undergoing corrective heart surgery at the Department of Cardiothoracic surgery, Oslo University Hospital (Supplementary table 1). The project was approved by the Regional Ethics committee of South East Norway (REC 31516) and conducted in compliance with the Declaration of Helsinki. Written informed consent was obtained from donor parents.

Sample processing

Sample collection and processing has been described previously[17]. Briefly, thymic tissue was dissociated by mechanical and enzymatic (Liberase TM) treatment. For each donor (N = 5), four samples were profiled: (1) Peripheral blood mononuclear cells (PBMC), (2) thymic cells without enrichment, (3) thymic cells enriched for APCs by density gradient centrifugation, and (4) CD45-depleted thymic cells.

Single cell AIRR sequencing

Single cell sequencing was performed according to the 10× Genomics Chromium Single Cell 5′ V(D)J Reagent Kit User guide with Feature Barcode Technology for Cell Surface Protein, v1 Chemistry (10× Genomics protocol CG000186, RevD). Amplification of full length cDNA prepared from mRNA was run for 13 cycles for PBMC and unenriched samples, and 15 cycles for APC-enriched and CD45-depleted samples, before single cell TCR V(D)J enriched libraries were constructed. In brief, 2 µl cDNA was used to enrich full-length V(D)J segments via PCR amplification with primers specific to the $\alpha\beta$TCR, before enzymatic fragmentation and size selection by beads. Indexing PCR was run for 9 cycles for all samples as per manufacturer's recommendation, and sequencing was performed on an Illumina NovaSeq S2 flow cell (read length 27 for R1, 92 for R2).

Pre-processing and filtering of single cell AIRR data

Single cell AIRR sequencing data was pre-processed using CellRanger v.3.1.0 with alignment to the refdata-cellranger-vdj-GRCh38-alts-emsebl-5.0.0 reference[18,19]. Most libraries reached a sequencing depth of > 5000 read pairs/cell, with two exceptions for the CD45-depleted TCR library for donor 3 (3300 read pairs/cell) and the APC-enriched TCR library for donor 2 (4353 read pairs/cell). Clonotypes were defined as by CellRanger v.3.1.0, according to exact match of the complementarity-determining region 3 (CDR3) nucleotide sequence. Alternatively, for analyses focusing on the CDR3 amino acid sequence, we defined clonotypes by exact match in CDR3 amino acid (CDR3aa) sequence.

Further analysis (except CoNGA) was performed in R v.4.1.3[20] using Immunarch v.0.8.0[21]. The filtered_contig_annotation.csv files from CellRanger were loaded as paired chain data by use of repLoad() with .mode = "paired", and barcodes annotated as thymocytes/T cells in the GEX dataset were retained. Next, CellRanger-derived clonotypes were filtered by chain combinations, permitting a) a single *TRB* chain or a pair of one *TRA* and one *TRB* for thymocytes, and b) a pair of one *TRA* and one *TRB* for peripheral blood T cells.

Downstream analyses in Immunarch were performed using single receptor chains, loaded by repLoad() with .mode = "single". The resulting dataset was filtered by keeping barcodes that existed in filtered, paired chain data, and grouped by GEX data cell type annotations.

For calculation of numbers and abundances of unique CDR3aa sequences in the filtered single chain data, identical CDR3aa sequences were combined and their counts summed up. Unique CDR3aa numbers were

calculated using repExplore() with .method = "volume" and .col = "aa". The abundance of unique CDR3aa sequences was calculated using repExplore() with .method = "count" and .col = "aa".

CDR3 lengths, clonal overlap, and gene usage in Immunarch

CDR3aa lengths were calculated for thymic TCRs by use of single chain data, after combining data from all samples for each cell type. For thymocytes, cell types were further grouped as early stage (DN(P), DN(Q), DP(P), DP(Q)), late stage ($\alpha\beta$T(entry), $CD4^+$ SP, $CD8^+$ SP) or agonist (CD8$\alpha\alpha$(I), CD8$\alpha\alpha$(II), $T_{(agonist)}$, T_{reg}(diff), T_{reg}) thymocytes. Identical CDR3aa sequences were combined and summed up before running repExplore() with .method = "len" and .col = "aa". A two-tailed Welsh t-test was performed between early stage thymocytes and either late stage or agonist thymocytes, with a null hypothesis of no difference in group CDR3aa length means (H_0: $\mu_1 = \mu_2$), an alternative hypothesis of a difference in group CDR3aa length means (H_a: $\mu_1 \neq \mu_2$), and an alpha level of 0.05, followed by Bonferroni correction for two tests.

Clonal overlap was evaluated for singe chain data, after combining data from all samples by either cell type or donor, and grouping and summing up identical CDR3aa sequences. Overlap was determined by the Morisita-Horn overlap index, which accounts for both numbers and abundances of overlapping clones[22,23]. For this, repOverlap() was used with .method = "morisita" and .col = "aa".

Gene usage was determined based on single chain data combined by cell type, according to CellRanger-derived clonotype definitions. One clonotype lacking *TRAJ* gene information and one clonotype lacking *TRBJ* gene information were excluded. Weighted gene usage was calculated as proportions weighted by the abundance of each clonotype, using the Immunarch geneUsage() function with .quant = "count", .norm = T, and .ambig = "maj".

Overlap of the TCR dataset to McPAS-TCR

For identification of pathology-specific TCR sequences, we assessed our data for overlap to the McPAS-TCR database (accessed 02.02.2023), which contains pathology-associated *TRA* and *TRB* CDR3aa sequences with known antigen specificities[24]. Human sequences under the "Autoimmune" and "Pathogens" categories of McPAS-TCR were extracted, encompassing 26258 *TRB* sequences and 6999 *TRA* sequences. We then tested for overlap to the database among unique, single chain CDR3aa *TRA* or *TRB* sequences for early stage thymocytes, late stage thymocytes, agonist thymocytes, peripheral CD4 T cells (CD4 cytotoxic T cell (CTL), CD4 naive, CD4 Proliferating, CD4 T central memory (TCM), CD4 T effector memory (TEM)), and peripheral CD8 T cells (CD8 naive, CD8 TCM, CD8 TEM). Enrichment of pathology-specific sequences among cell groups relative to the database was assessed by a one-tailed Fisher's exact t-test (H_0: true odds ratio = 1, H_A: true odds ratio > 1) with Bonferroni correction for the number of tested diseases and the number of tested cell type groups, based on the approach by Amoriello et al.[25].

Joint TCR and transcriptomic analysis using CoNGA

We performed clonotype neighbour graph analysis (CoNGA)[26], a joint analysis pipeline for TCR and GEX data in order to identify correlations between the data types. The CoNGA pipeline includes creation of separate neighbourhood graphs for clonotypes in the TCR and GEX datasets, followed by identification of clonotypes that reside within overlapping neighbourhoods in both datasets. For CoNGA analysis (commit ID b572f73e5c90aaae59e11f187553a50324f26f02) of either thymocytes or peripheral T cells, filtered, paired chain TCR data was used, and analysed together with previously reported GEX data for either thymocytes or T cells.

After log normalization and then regression in order to account for the effect of number of counts and percentage of mitochondrial genes in the GEX data, data was reduced to one representative cell per CellRanger-derived clonotype. Neighbourhood graphs were created for the GEX and TCR datasets by use of BBKNN v. 1.5.1[27], before UMAP dimensionality reduction[28] and clustering by the Leiden algorithm[29] at resolution 1.0. Differential gene expression analysis was performed by the Wilcoxon Rank Sum test by the rank_genes_group() function from Scanpy v. 1.9.1, as implemented by the CoNGA find_gex_cluster_degs() function.

Next, the CoNGA graph-versus-graph analysis was used, which calculates a "CoNGA score" for identification of clonotypes residing within overlapping neighbourhoods in GEX and TCR datasets, and combines overlapping clonotypes into "CoNGA clusters". This was done by the CoNGA run_graph_vs_graph() function, with a minimum size for identified "CoNGA clusters" of 22 clonotypes for the thymocyte dataset, and 8 for the PBMC-derived T cell dataset, constituting at least 0.1% of clonotypes in the dataset.

Graph-versus-feature analysis was used for identification of TCR features exhibiting variation across the GEX dataset, by running the CoNGA run_graph_vs_feature() function. Reported adjusted P-values are from assessment of variation in TCR features across a GEX cluster graph, where clonotypes residing in the same GEX cluster are connected. As implemented in CoNGA in order to increase computational speed, features were first subjected to a preliminary t-test, and p-values were adjusted by multiplication for the number of performed tests. Features with multiplied p-values ≤ 10 were then subjected to a two-tailed Mann–Whitney U test with Bonferroni correction, at an alpha level of 0.05.

Visualisations

Visualisations were prepared as implemented in Immunarch or CoNGA, in addition to ggplot2 v.3.3.5[30].

Results

Sample preparation and data filtering

To study TCR repertoires of human thymus and blood, scAIRR-seq was performed. Briefly, one PBMC sample and three thymic samples were prepared for each young paediatric donor (N = 5), with thymic samples consisting

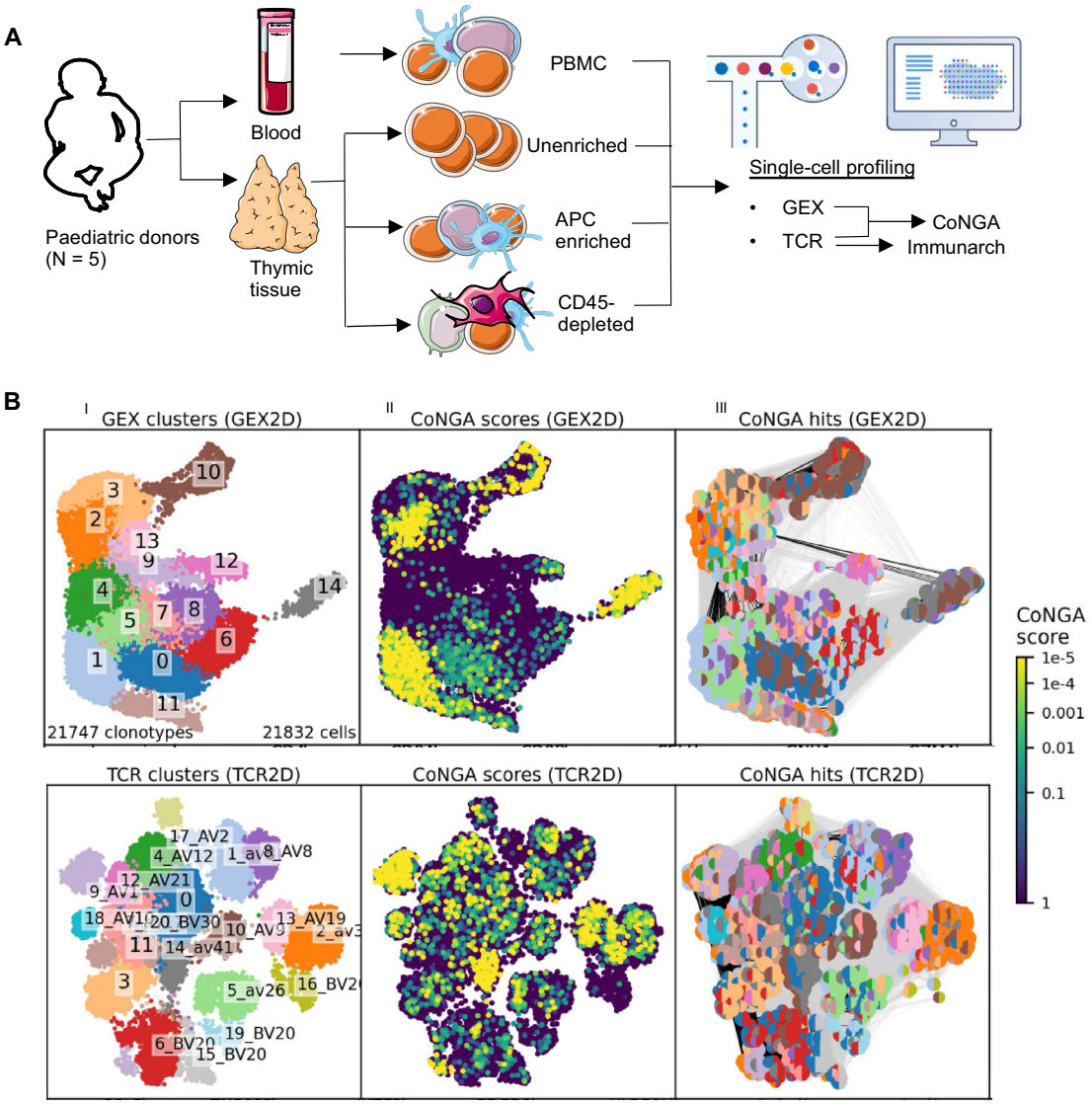

Figure 1. Single cell gene expression (GEX) and adaptive immune receptor repertoire (AIRR) sequencing of young paediatric (N = 5) thymi and blood. (**A**) Experimental set-up. For each donor, single cell AIRR sequencing was performed for one PBMC sample and three thymic samples. Created using content from Servier Medical Art, provided by Servier, licensed under a Creative Commons Attribution 3.0 unported license, and graphics adapted from 10 × Genomics. (**B**) Clonotype Neighbour Graph Analysis (CoNGA) for identification of correlations between thymocyte GEX and T cell receptor (TCR) datasets. (I) Clonotypes that are highly similar in either the GEX (top row) or TCR (bottom row) datasets, based on neighbourhood graphs, are clustered together. (II) Groups of clonotypes with similar GEX and TCR profiles are identified as having low "CoNGA scores". The "CoNGA scores" reflect the likelihood of seeing equal or greater overlap between the GEX and TCR dataset by chance, when overlap is assessed as shared neighbours between the two datasets for each clonotype. (III) Clonotypes with low "CoNGA scores" ("CoNGA hits") are grouped into "CoNGA clusters", bi-coloured discs indicate their GEX (left half) and TCR (right half) cluster assignment.

of one unenriched sample, one sample enriched for APCs, and one CD45-depleted sample (Fig. 1A). Single cell TCR sequencing was performed for all 20 samples.

AIRR data was filtered to retain immune receptor chains from thymocytes/T cells based on previous annotations[17], and presenting in expected chain combinations (*TRB* or *TRA* + *TRB* for thymocytes, *TRA* + *TRB* for peripheral T cells). This resulted in 46086 unique cell barcodes in the TCR data. The number of unique CDR3aa sequences for each chain was proportional to cell numbers and cell composition of samples, with most abundant cell types having higher numbers of unique CDR3aa sequences (Supplementary Fig. 1A). For each chain, the clonal abundance was low, with most unique CDR3aa sequences being associated with one single cell per sample (Supplementary Fig. 1B).

Joint TCR and transcriptomic analysis confirms previously identified thymocyte populations

In order to determine whether changes in TCR repertoires during the course of thymocyte developmental progression could be observed, we used two approaches for data analysis: One based on Immunarch using our previously reported annotations, and one based on the CoNGA pipeline[26]. The CoNGA pipeline performs a joint analysis of both GEX and AIRR data for identification of correlations between both data types to enable new discoveries not revealed by the GEX data alone.

In the CoNGA-generated GEX UMAP (Figs. 1B, 2A, B Supplementary Fig. 2A), populations previously reported by us and others were identified[14], including proliferating (P) DP thymocytes (GEX cluster 10), quiescent (Q) DP thymocytes (GEX clusters 3 and 2), $CD8^+$ SP thymocytes (GEX clusters 1 and 11), and $CD4^+$ SP thymocytes (GEX cluster 0 among others). GEX cluster 11 contained two distinguishable subsets, where one expressed genes associated with an innate-like phenotype, in resemblance to the population termed NKT cells in our previous report. We also observed clusters resembling *ZNF683*$^+$ CD8$\alpha\alpha$ cells (GEX cluster 12) and *FOXP3*$^+$ T_{regs} (GEX cluster 14), and a cluster (GEX cluster 9) expressing *PDCD1* in similarity to the thymic Type A CD8$\alpha\alpha$ precursor[15,31].

A shift from proximal to distal TRA gene usage occurs at the DP(Q) thymocyte stage

We next investigated potential differences in TCR gene usage between thymocyte populations. In the CoNGA pipeline, clonotypes that reside within overlapping neighbourhoods in the GEX and TCR datasets are combined into "CoNGA clusters" (GEX:TCR) (Fig. 1B, Supplementary Fig. 3). One "CoNGA cluster" (10:14) contained

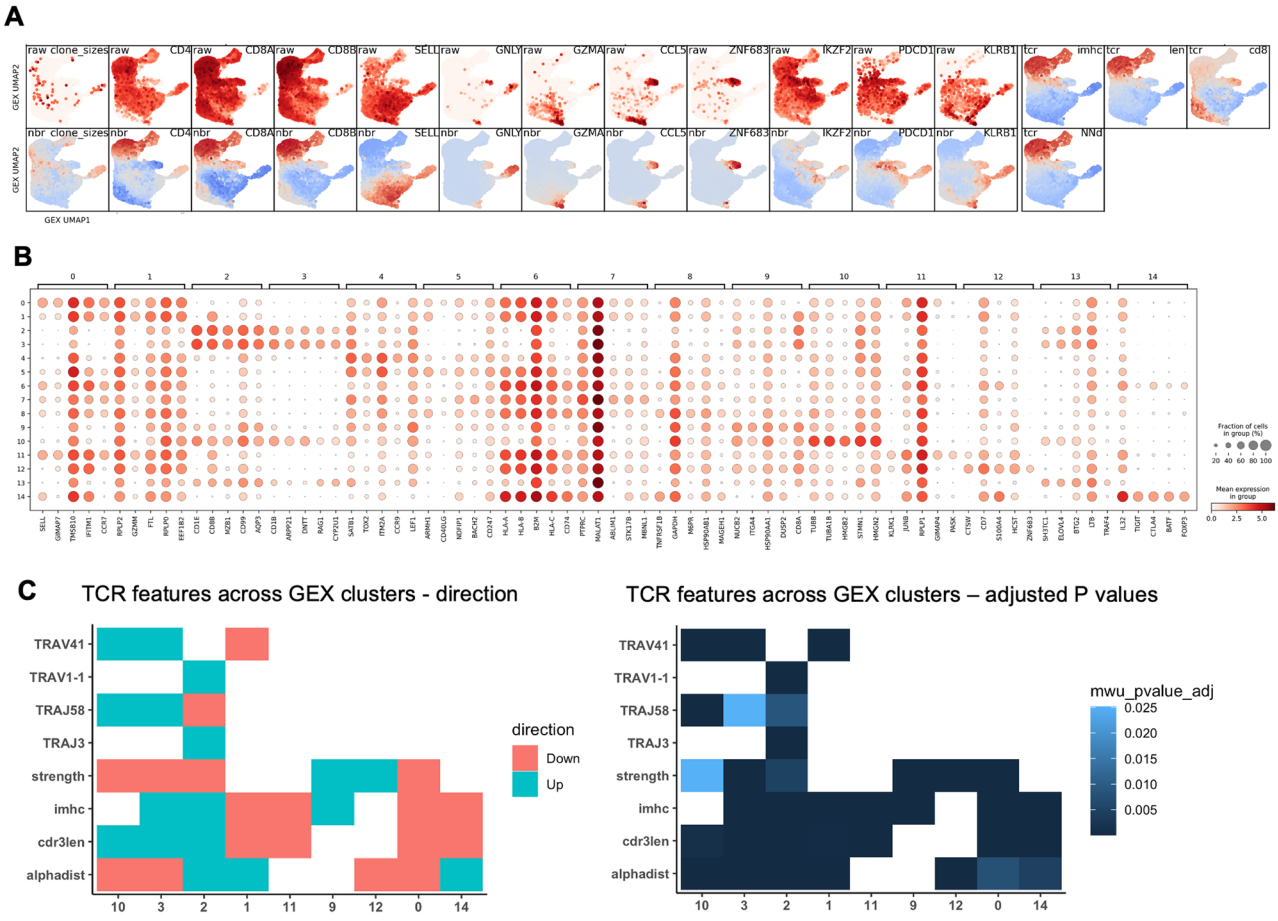

Figure 2. Gene expression and T cell receptor (TCR) features across single cell gene expression (GEX) defined clusters. (**A**) Selected genes and TCR features across the thymic GEX UMAP. Colour indicates expression of selected genes (top row) and Z-score normalised and GEX-neighbourhood averaged expression of the same genes (bottom row). Blue colour indicates negative values, red colour indicates positive values. (**B**) Differentially expressed genes across thymic GEX clusters. Dot colour indicates expression level, and dot size indicates the fraction of clonotypes expressing the gene. (**C**) Selected TCR features exhibiting variation across GEX clusters in a graph-versus-feature analysis. Numerical features indicating properties of the TCR repertoire are mapped onto a GEX cluster graph where clonotypes residing in the same GEX cluster are connected. Variation in the TCR features across the GEX cluster graph is assessed first by an initial t-test, then by a one-sided Mann–Whitney U test for the features with low t-test P values. Colours indicate direction of change (left) or adjusted Mann–Whitney U test P values (right).

clonotypes from the DP(P) GEX cluster 10 and exhibited high usage of the proximal *TRAV41* gene. DP(Q) thymocytes in GEX clusters 2 and 3 participated in several distinct "CoNGA clusters", including one (3:14) biased towards *TRAV41*, and one (2:9) biased towards *TRAV1-2* and *TRAV1-1*.

To further identify variation in TCR features across the thymocyte GEX dataset, we used the CoNGA graph-versus-feature analysis (Fig. 2C, Supplementary Fig. 4, Supplementary table 5). Among the assessed TCR features, CoNGA implements numerical scores representing properties of the TCR, including scores previously developed by others[32,33], and novel scores by the CoNGA developers[26]. For instance, the "alphadist" score represents the ordinal distance between the *TRAV* and *TRAJ* gene segments when the *TCRA* locus is ordered by genomic position[26]. The "alphadist" score was significantly decreased for GEX clusters 10 (DP(P) thymocytes, $P_{adj} = 4.3 \times 10^{-96}$) and 3 (DP(Q) thymocytes, $P_{adj} = 7.0 \times 10^{-80}$), and significantly increased for GEX cluster 2 (DP(Q) thymocytes, $P_{adj} = 9.0 \times 10^{-97}$), fitting with the gene usage patterns noted above.

In agreement, analysis by Immunarch (Fig. 3A, Supplementary Fig. 5) revealed a high frequency of *TRAV41* for DP(P), in addition to a high frequency of *TRAJ57* and *TRAJ58*. We finally noted diversity in gene usage among the innate-like population labelled as Natural Killer T (NKT) cells, in agreement with reports of a $CD8^+$ innate-like polyclonal thymic cell population from mice[34], and NK-like $CD8^+$ T cells in humans[35]. In sum, DP(P) thymocytes exhibited a bias toward proximal *TRA* genes, while shift from proximal to distal *TRA* gene usage occurred at the DP(Q) thymocyte stage.

Early thymocytes have longer CDR3aa sequences compared to late or agonist thymocytes

We next investigated CDR3aa length across thymocytes. In the CoNGA graph-versus-feature analysis (Fig. 2C, Supplementary Fig. 4, Supplementary table 5), DP thymocytes (GEX clusters 10, 2, and 3) exhibited significantly increased CDR3aa lengths ($P_{adj} = 0.0011$ for GEX cluster 10, 7.7×10^{-61} for GEX cluster 2, 2.2×10^{-48} for GEX cluster 3) compared to remaining clusters. By contrast, CDR3aa length was significantly decreased for the more mature $CD8^+$ (GEX cluster 1, $P_{adj} = 3.0 \times 10^{-4}$ and GEX cluster 11, $P_{adj} = 8.1 \times 10^{-6}$) and $CD4^+$ SP (GEX cluster 0, $P_{adj} = 3.7 \times 10^{-7}$) thymocyte stages.

To determine whether the same conclusion was supported by Immunarch, we grouped unique *TRA* or *TRB* CDR3aa sequences of early stage, late stage, and agonist selected thymocytes (Supplementary table 2). We tested for differences in CDR3aa length between early stage thymocytes and either late stage or agonist thymocytes by Welsh's t-test, under an assumption of a normal distribution (Fig. 3B, Supplementary table 3). Early stage thymocytes exhibited significantly longer CDR3aa lengths compared to late stage thymocytes ($P_{adj} = 1.7 \times 10^{-17}$ for *TRA*, 7.1×10^{-220} for *TRB*) or agonist selected thymocytes ($P_{adj} = 5.2 \times 10^{-16}$ for *TRA*, 2.9×10^{-138} for *TRB*). In sum, both pipelines highlight a decrease in CDR3 length after the DP(Q) stage.

PDCD1$^+$ thymocytes exhibit previously reported *ZNF683*$^+$-associated TCR biases

We further examined the potential biases in the TCR repertoire among thymocytes developing towards CD8$\alpha\alpha$ T cells, as CoNGA identified specific biases for both thymic and PBMC *ZNF683*$^+$ CD8$\alpha\alpha$ T cells in a previous report[26]. While a "CoNGA cluster" (12:2) was formed by clonotypes from the *ZNF683*$^+$ GEX cluster 12 and TCR cluster 2, the presence of additional 1:2 and 11:2 "CoNGA clusters" could indicate similar TCR features among other $CD8^+$ T cells (Supplementary Fig. 3). However, the graph-versus-feature analysis indicated a significantly increased score for "strength" in GEX cluster 12 ($P_{adj} = 2.5 \times 10^{-5}$), a measure reflecting a CDR3aa composition mediating strong TCR interactions based on an estimated interaction potential[26,32,33,36] (Fig. 2C, Supplementary Fig. 4, Supplementary table 5).

GEX cluster 12 (*ZNF683*$^+$ CD8$\alpha\alpha$) also exibited several features previously observed among thymic and peripheral *ZNF683*$^+$ T cells[26], including enrichement of large and positively charged amino acid residues ("volume" score, reflecting amino acid size, and "charge" score, reflecting amino acid charge[26,32,33], Supplementary Fig. 4, Supplementary table 5). Still, no increase was observed in the score termed "independent of peptide:MHC" ("imhc"), which summarises the reported *ZNF683*$^+$ T cell-associated TCR features as a weighted linear combination of several TCR sequence features, including "volume" and "charge" scores among others[26].

By contrast, "imhc" was significantly upregulated for the *PDCD1*-expressing GEX cluster 9 (CD8$\alpha\alpha$ precursor, $P_{adj} = 3.3 \times 10^{-15}$), and for the DP(Q) thymocytes in GEX clusters 3 ($P_{adj} = 3.2 \times 10^{-35}$) and 2 ($P_{adj} = 1.18 \times 10^{-51}$). Like GEX cluster 12, GEX cluster 9 was also significantly enriched for CDR3aa sequences mediating high affinity TCR interactions ("strength" score, $P_{adj} = 1.3 \times 10^{-19}$) (Fig. 2C, Supplementary Fig. 4, Supplementary table 5).

While T_{reg} progenitors mirroring the previously reported $T_{(agonist)}$ and T_{reg}(diff) populations were not identified in the CoNGA analysis, potentially as an effect of the employed cluster resolution, mature T_{regs} could be observed in GEX cluster 14 (Figs. 1B, 2B). This cluster exhibited reduced "imhc" and "charge" scores (Fig. 2C, Supplmentary table 5), thus constrasting with the observations from the CD8$\alpha\alpha$ T cell lineage.

Overall, thymic *PDCD1*$^+$ and *ZNF683*$^+$ cells exhibited TCR features reported to be associated with *ZNF683*$^+$ T cells. For the *PDCD1*$^+$ cluster, this resulted in a significant upregulation of the summarising "imhc" score, suggesting a TCR repertoire bearing similarities to previously characterized *ZNF683*$^+$ T cells.

Peripheral *ZNF683*$^+$ cells resemble their thymic counterparts with respect to TCR biases

We next applied CoNGA to the peripheral T cell TCR and GEX dataset (Fig. 4A–C, Supplementary Fig. 2B, Supplementary Fig. 6, Supplementary table 6). We observed clonotypes residing within overlapping neighbourhoods in the GEX and TCR datasets, indicated by low "CoNGA scores", among $CD8^+$ T cells (GEX clusters 4 and 7) (Fig. 4A). GEX cluster 7 expressed *CCL5* and consisted of *GNLY*$^+$ and *ZNF683*$^+$ subsets, with low "CoNGA scores" largely mapping to the *GNLY*$^+$ subset (Fig. 4A, B). The clonotypes of GEX cluster 7 participated in two "CoNGA clusters" (7:5 and 7:14), which were marked by *TRAV1-2* and *TRAJ33* usage[37,38] and a corresponding high score defining TCR sequences matching that of a consensus sequence for the receptor of mucosal-associated

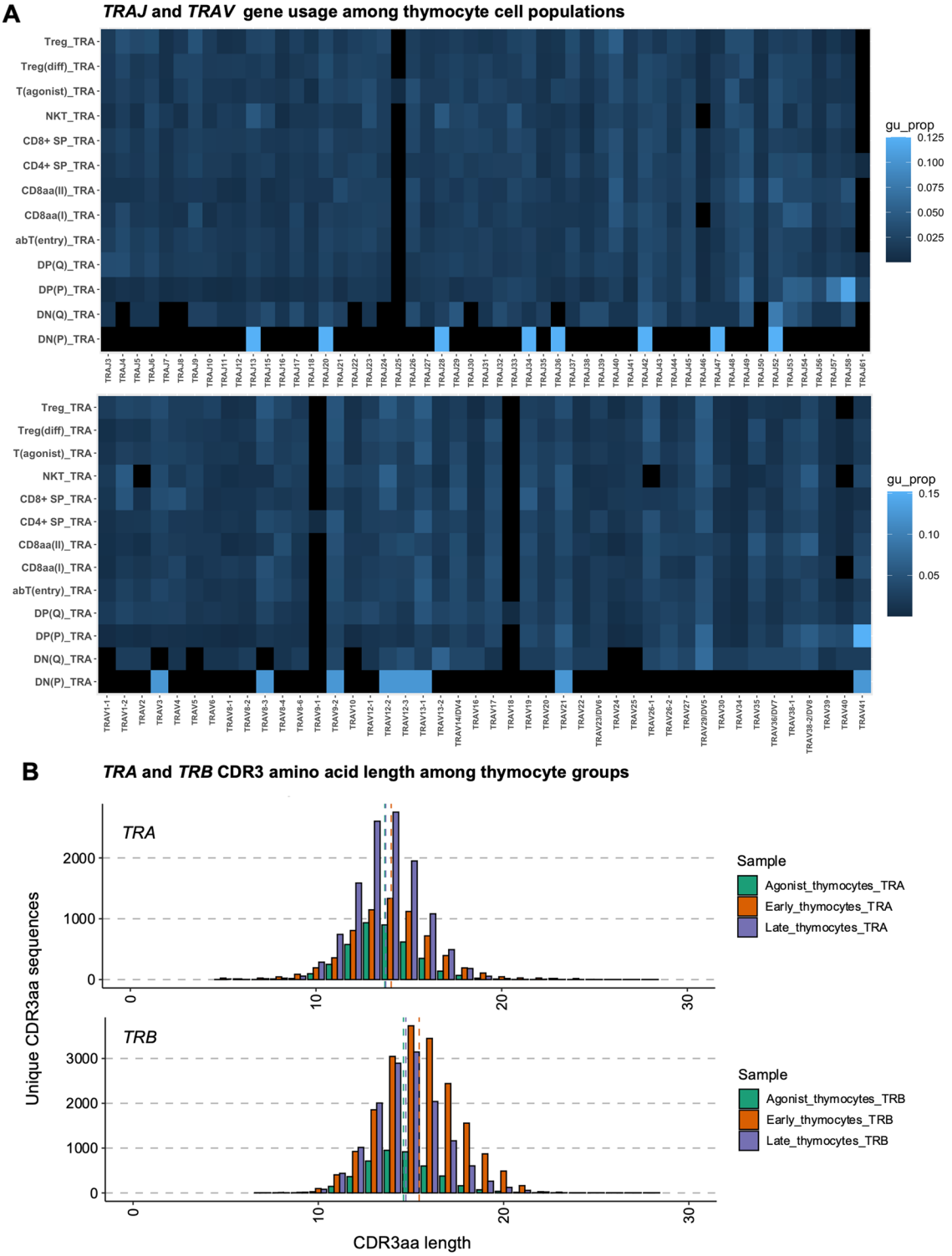

Figure 3. Gene usage proportions and distribution of CDR3aa lengths among thymocytes. (**A**) *TRAJ* and *TRAV* gene usage across thymocytes. Colour indicates proportion among CellRanger-derived clonotypes weighted by clonal counts. (**B**) Length of unique *TRA* and *TRB* CDR3aa sequences among agonist (CD8 $\alpha\alpha$(I), CD8$\alpha\alpha$ (II), T(agonist), T_{reg}(diff), T_{reg}), early (DN(P), DN(Q), DP(P), DP(Q)), and late ($\alpha\beta$T(entry), $CD4^+$ SP, $CD8^+$ SP) thymocytes, dotted vertical lines indicate group means.

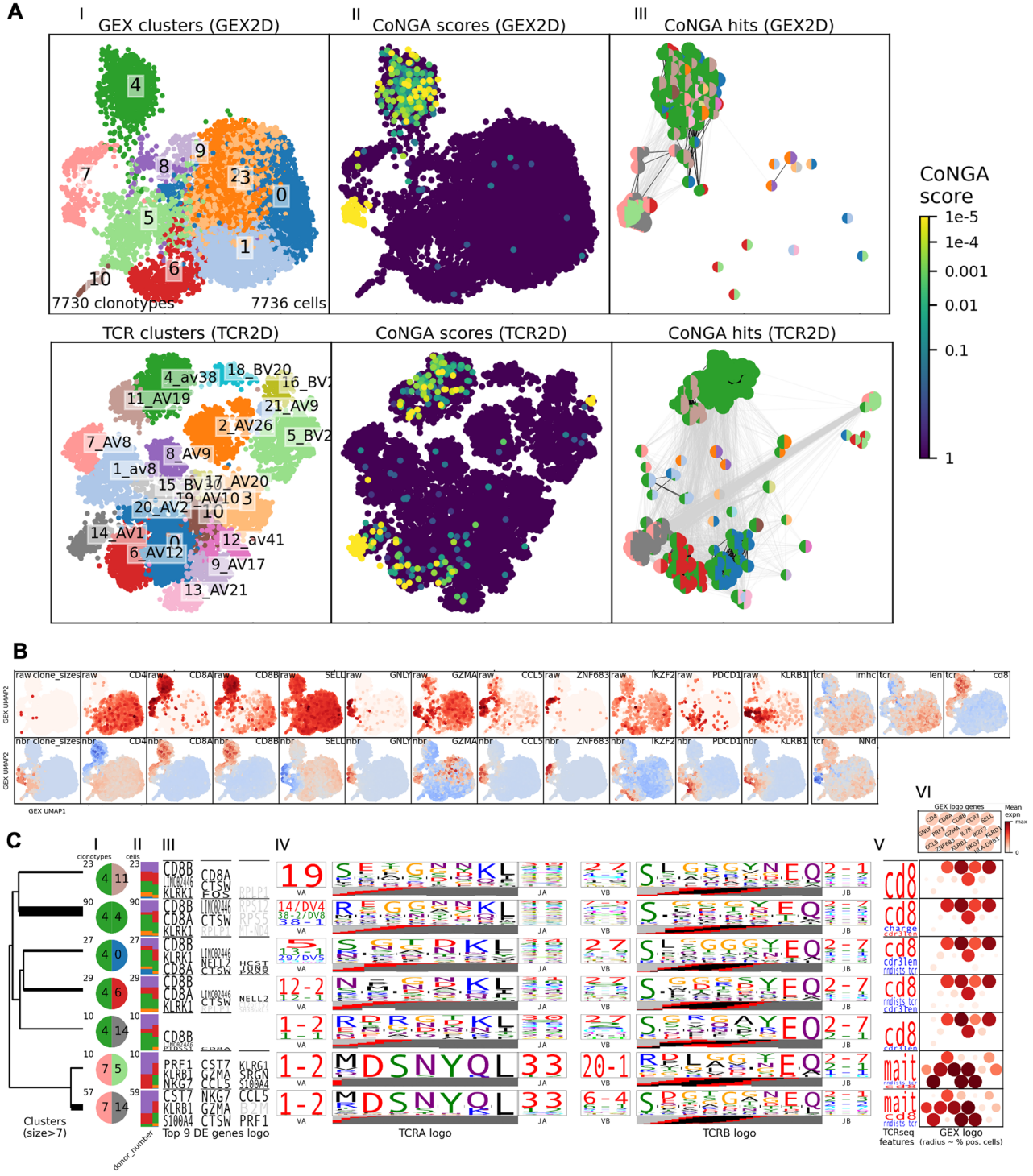

Figure 4. Clonotype Neighbour Graph Analysis (CoNGA) for identification of correlations between peripheral T cell gene expression (GEX) and T cell receptor (TCR) datasets (N = 5). (**A**) (I) UMAP projections coloured by GEX (top row) and TCR (bottom row) clusters. (II) UMAP projections coloured by "CoNGA scores", indicating clonotypes residing in overlapping neighbourhoods in both the GEX and TCR datasets. (III) Bi-coloured discs indicating the GEX (left half) and TCR (right half) cluster assignment of clonotypes with low "CoNGA scores" ("CoNGA hits"). (**B**) Raw and Z-score normalised and GEX-neighbourhood averaged (nbr) expression of selected genes and TCR features across the thymic GEX UMAP. Blue colour indicates negative values, red colour indicates positive values. (**C**) Gene expression and TCR features among clonotypes grouped into "CoNGA clusters". Shown are (I) bi-coloured discs indicating GEX (left half) and TCR (right half) cluster assignment, (II) donor origin of clonotypes, (III) differentially expressed genes, (IV) frequently used TCR gene segments and amino acid sequences, (V) properties of the TCR repertoire (represented as numerical TCR feature scores, red colour indicates increased scores, blue colour indicates decrease scores), and (VI) expression of selected genes.

invariant T cells ("mait" score)[26] (Fig. 4C, Supplementary table 6). Plotting on the GEX UMAP indicated that the mait cell-associated TCR sequences mainly related to the *GNLY*$^{+}$-expressing subset (Supplementary Fig. 6, Supplementary table 6). Clonotypes of GEX cluster 7 exhibited enrichment for CDR3aa mediating strong interactions, indicated as a significant increase in the "strength" score ($P_{adj} = 0.024$), but not in the "imhc" score indicating *ZNF683*$^{+}$-associated features. Further, we note that the increase in "strength" was also observed for GEX cluster 4 ($P_{adj} = 4.7 \times 10^{-7}$).

In sum, biases in the peripheral TCR repertoire were largely attributed to peripheral mait cells, with the *ZNF683*$^{+}$ subset exhibiting no increase in the previously reported "imhc" score, in resemblance to the thymic *ZNF683*$^{+}$ cells.

TRA chains exhibit a higher degree of clonal overlap compared to *TRB* chains

We next determined clonal overlap of *TRA* and *TRB* chains across donors and cell populations in thymus and blood, using the Morisita-Horn index as implemented in Immunarch (Fig. 5A–D). While the extent of overlap was limited, we observed a trend towards higher overlap for *TRA* compared to *TRB*, in agreement with previous reports[39].

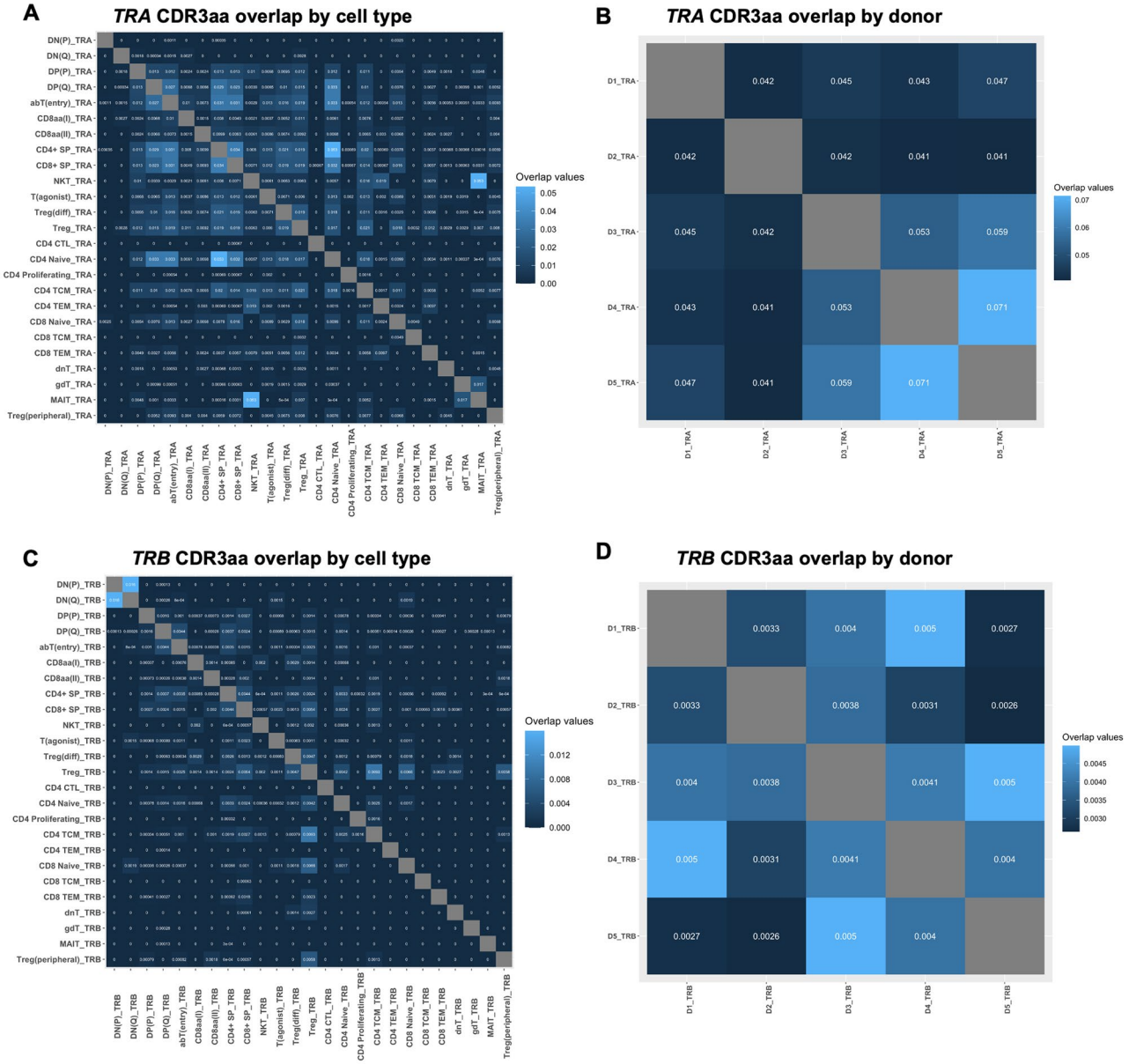

Figure 5. Clonal overlap in TCR data. Overlap (Morisita-Horn index) of unique *TRA* (top row) and *TRB* (bottom row) CDR3aa sequences among thymocytes and peripheral T cells across either cell types (left panels) or donors (right panels).

Late stage thymocytes were enriched for pathology-specific CDR3aa *TRB* sequences

To identify pathology-associated CDR3aa sequences, we investigated overlap between our TCR data and CDR3aa sequences from the "autoimmune" or "pathogens" categories of the McPAS-TCR database[24]. *TRA* and *TRB* CDR3aa sequences were assessed separately for early stage thymocytes, agonist selected thymocytes, late stage thymocytes, peripheral CD4 T cells (CD4 CTL, CD4 naive, CD4 Proliferating, CD4 TCM, CD4 TEM), and peripheral CD8 T cells (CD8 naive, CD8 TCM, CD8 TEM) (Supplementary table 4)[25]. Enrichment of pathology-associated sequences was particularly pronounced for *TRB* sequences of late stage thymocytes, with significant (Bonferroni corrected Fisher's exact t-test) enrichment of Influenza—($P_{adj} = 0.013$), Celiac disease—($P_{adj} = 1.7 \times 10^{-4}$), and Inflammatory bowel disease (IBD)—($P_{adj} = 5.0 \times 10^{-5}$) associated sequences, implying the presence of both pathogen—and autoimmune-associated CDR3aa sequences among post-selection thymocytes.

Discussion

In this work, we have performed scAIRR-seq on human thymus and blood from young paediatric donors, at an age prior to extensive thymic involution. We have previously reported GEX profiling of the same single cells, with cell type annotations[17]. Here we explored potential differences in the immune cell receptor repertoires across the identified cell populations. The use of single cell technology permitted coupling of immune receptor sequences to the transcriptional profile of each cell, facilitating assessment of TCR profiles across distinct thymic cell populations. Two distinct analytical approaches were used, where one was based on the detailed cell type annotations of the GEX dataset and the Immunarch workflow. The second approach was based on the CoNGA pipeline, which introduces a new integrative clustering using both gene expression and TCR data.

Both approaches identified dynamic changes in the TCR repertoire during thymocyte development. First, both approaches indicated expression of *TRA* as early as the DP(P) thymocyte stage, with a favouring of proximal *TRA* genes, before a shift towards distal *TRA* genes at the DP(Q) stage. This contrast with the established model of *TRA* recombination at the DP(Q) stage, and has previously been observed by others in human thymic single cell TCR sequencing data[14]. Previous work has revealed that premature *TRA* expression may occur in transgenic cell lines, resulting in expansion of CD8$\alpha\alpha$ IELs[40]. Early *TRA* expression has also been reported in non-transgenic murine thymocytes[41]. However, these studies reported a preference for the distal *TRAV1* gene, potentially induced from the Eδ rather than the Eα enhancer, while we observed a preference for the proximal *TRA* gene. Again, our observations aligned with previous single cell human TCR data[14]. A preference for proximal *TRA* genes for more immature thymocyte populations, followed by a shift towards distal genes as development progresses, would also be in agreement with the suggested processive mechanism of repeated *TRA* recombination events[42].

Both our analytical approaches further implied a shortening of the CDR3 region as thymocytes progress from DP to more mature stages, which potentially could be explained by biases inferred by selection events at the $\alpha\beta$T(entry) stage[43]. As such, both pipelines resulted in similar observations consistent with previous literature, highlighting robustness across approaches.

For the more mature thymocytes and T cells, clonotypes that resided in overlapping neighbourhoods in the GEX and TCR datasets were largely attributed to $CD8^+$ rather than $CD4^+$ populations. This is in agreement with the previous CoNGA analyses of both PBMC and thymic samples[26], potentially explained by a more restricted repertoire for $CD8^+$ T cells compared to $CD4^+$ T cells[44]. Alternatively, this could indicate an increased suitability for the CoNGA approach in identification of $CD8^+$ compared to $CD4^+$-derived TCR biases. Nevertheless, CoNGA appeared well suited for elucidation of the choice between conventional $CD8^+$ SP selection or CD8$\alpha\alpha$ lineage differentiation.

Previously, CoNGA has identified specific TCR features associated with thymic and peripheral *ZNF683*$^+$ CD8$\alpha\alpha$ T cells[14,26], including CDR3aa sequences enriched for large, positively charged amino acids. However, we did not observe a significant increase in the "imhc" score, which summarises the *ZNF683*$^+$-associated features, for the *ZNF683*$^+$-expressing clusters in our dataset. We did, however, notice variation in several of the included TCR features among *ZNF683*$^+$ cells, and we cannot exclude the possibility that the lack of a significant increase in "imhc" resulted from the limited data available.

Intriguingly, biases in TCR features were striking for a thymic *PDCD1*$^+$ expressing cluster, including enrichment for large, positively charged amino acids associated with strong peptide-MHC:TCR interactions, and in this case resulting in a significantly upregulated "imhc" score. In our previous report, we observed a branch point between conventional and agonist selected thymocytes at a developmental time point prior to the SP stage, with the agonist selected lineages upregulating markers of a strong TCR response and signalling to APC subsets. The observation of *ZNF683*$^+$-associated TCR biases in the *PDCD1*$^+$ cluster, together with reports of a *PDCD1*$^+$ precursor population for CD8$\alpha\alpha$ T cells[31], further supports this branch point. However, this must be corroborated by functional studies, and additionally, the role played by the developmental timing of *TRA* recombination for the choice between agonist or conventional positive selection needs further elucidation. We further note that the "imhc" score was upregulated also in DP thymocyte clusters, potentially implying that the TCR features reported to be associated with *ZNF683*$^+$ T cells are also broadly present among less mature thymocyte stages or thymocytes that have not yet undergone conventional selection.

Plasticity between conventional and agonist selected lineages is exemplified by the ability of conventional, mature $CD4^+$ T cells to differentiate to induced T_{regs} in the periphery. Induced T_{regs} appear to have a reduced lineage stability compared to thymically derived, natural T_{regs}, an aspect that must be considered in the development of T_{reg}-derived therapies[45–47]. Mirroring this plasticity, peripheral $CD4^+$ T cells may also differentiate to become $CD4^+$CD8$\alpha\alpha$ IELs[48]. However, a model of TCR-specific biases for the *PDCD1*$^+$ thymic population, suggested as a thymic CD8$\alpha\alpha$ IEL precursor diverging prior to $CD4^+$ lineage commitment, could imply cell-intrinsic differences between thymic-derived and peripherally induced CD8$\alpha\alpha$ IELs, with particular implications in immune-mediated diseases in the gut such as IBD[9]. As such, further insights into differences between

IELs of different developmental origin would increase the understanding of the role of the IEL compartment in immune regulation and disease.

Altogether, our study supported that *PDCD1*$^{+}$ thymocytes exhibit a TCR repertoire bearing similarity to previously reported *ZNF683*$^{+}$ populations, fitting with a branch point between conventional and agonist-selected thymocyte populations and potentially implying a developmentally timed role for the TCR in human CD8$\alpha\alpha$ thymic T cell differentiation.

Data availability

Datasets are available from the Gene Expression Omnibus (GEO), accession number GSE227408.

Code availability

Code central to the conclusions of the paper is available from Zenodo, https://doi.org/10.5281/zenodo.8260575. Additional code is available from the corresponding author upon request.

Received: 30 March 2023; Accepted: 11 October 2023
Published online: 18 October 2023

References

1. Krangel, M. S. Mechanics of T cell receptor gene rearrangement. *Curr. Opin. Immunol.* **21**, 133–139. https://doi.org/10.1016/j.coi.2009.03.009 (2009).
2. Miho, E. *et al.* Computational strategies for dissecting the high-dimensional complexity of adaptive immune repertoires. *Front. Immunol.* **9**, 224. https://doi.org/10.3389/fimmu.2018.00224 (2018).
3. Wilson, A., Held, W. & MacDonald, H. R. Two waves of recombinase gene expression in developing thymocytes. *J. Exp. Med.* **179**, 1355–1360. https://doi.org/10.1084/jem.179.4.1355 (1994).
4. Irla, M. Instructive cues of thymic T cell selection. *Annu. Rev. Immunol.* **40**, 95–119. https://doi.org/10.1146/annurev-immunol-101320-022432 (2022).
5. Bouneaud, C., Kourilsky, P. & Bousso, P. Impact of negative selection on the T cell repertoire reactive to a self-peptide: A large fraction of T cell clones escapes clonal deletion. *Immunity* **13**, 829–840. https://doi.org/10.1016/s1074-7613(00)00080-7 (2000).
6. Yin, Y., Li, Y., Kerzic, M. C., Martin, R. & Mariuzza, R. A. Structure of a TCR with high affinity for self-antigen reveals basis for escape from negative selection. *EMBO J.* **30**, 1137–1148. https://doi.org/10.1038/emboj.2011.21 (2011).
7. Jordan, M. S. *et al.* Thymic selection of CD4+CD25+ regulatory T cells induced by an agonist self-peptide. *Nat. Immunol.* **2**, 301–306. https://doi.org/10.1038/86302 (2001).
8. Bluestone, J. A. *et al.* Type 1 diabetes immunotherapy using polyclonal regulatory T cells. *Sci Transl Med* **7**, 315–389. https://doi.org/10.1126/scitranslmed.aad4134 (2015).
9. Das, G. *et al.* An important regulatory role for CD4+CD8 alpha alpha T cells in the intestinal epithelial layer in the prevention of inflammatory bowel disease. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 5324–5329. https://doi.org/10.1073/pnas.0831037100 (2003).
10. Ma, H., Qiu, Y. & Yang, H. Intestinal intraepithelial lymphocytes: Maintainers of intestinal immune tolerance and regulators of intestinal immunity. *J. Leukoc. Biol.* **109**, 339–347. https://doi.org/10.1002/JLB.3RU0220-111 (2021).
11. Dijke, I. E. *et al.* Discarded human thymus is a novel source of stable and long-lived therapeutic regulatory T cells. *Am. J. Transplant* **16**, 58–71. https://doi.org/10.1111/ajt.13456 (2016).
12. Verstichel, G. *et al.* The checkpoint for agonist selection precedes conventional selection in human thymus. *Sci. Immunol.* **2**, 4232. https://doi.org/10.1126/sciimmunol.aah4232 (2017).
13. Morgana, F. *et al.* Single-cell transcriptomics reveals discrete steps in regulatory T cell development in the human thymus. *J. Immunol.* **208**, 384–395. https://doi.org/10.4049/jimmunol.2100506 (2022).
14. Park, J. E. *et al.* A cell atlas of human thymic development defines T cell repertoire formation. *Science* **367**, eaay3224 (2020). https://doi.org/10.1126/science.aay3224
15. Kurd, N. S. *et al.* Factors that influence the thymic selection of CD8alphaalpha intraepithelial lymphocytes. *Mucosal. Immunol.* **14**, 68–79. https://doi.org/10.1038/s41385-020-0295-5 (2021).
16. Leishman, A. J. *et al.* Precursors of functional MHC class I- or class II-restricted CD8alphaalpha(+) T cells are positively selected in the thymus by agonist self-peptides. *Immunity* **16**, 355–364. https://doi.org/10.1016/s1074-7613(02)00284-4 (2002).
17. Heimli, M. *et al.* Multimodal human thymic profiling reveals trajectories and cellular milieu for T agonist selection. *Front. Immunol.* **13**, 1092028. https://doi.org/10.3389/fimmu.2022.1092028 (2022).
18. 10x Genomics. *Build notes for Reference Packages*. https://support.10xgenomics.com/single-cell-gene-expression/software/release-notes/build#GRCh38_2020A (2020).
19. 10x Genomics. *Gene Expression Algorithms Overview*. https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/algorithms/overview (2020).
20. R Core Team. R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria*. https://www.R-project.org/, (2022).
21. Nazarov, V. I. *et al. Immunarch: Bioinformatics Analysis of T-Cell and B-Cell Immune Repertoires*. https://immunarch.com/, https://github.com/immunomind/immunarch, (2022).
22. Kidman, J. *et al.* Characteristics of TCR repertoire associated with successful immune checkpoint therapy responses. *Front. Immunol.* **11**, 587014. https://doi.org/10.3389/fimmu.2020.587014 (2020).
23. Rempala, G. A. & Seweryn, M. Methods for diversity and overlap analysis in T-cell receptor populations. *J. Math. Biol.* **67**, 1339–1368. https://doi.org/10.1007/s00285-012-0589-7 (2013).
24. 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. https://doi.org/10.1093/bioinformatics/btx286 (2017).
25. Amoriello, R. *et al.* TCR repertoire diversity in Multiple Sclerosis: High-dimensional bioinformatics analysis of sequences from brain, cerebrospinal fluid and peripheral blood. *EBioMedicine* **68**, 103429. https://doi.org/10.1016/j.ebiom.2021.103429 (2021).
26. Schattgen, S. A. *et al.* Integrating T cell receptor sequences and transcriptional profiles by clonotype neighbor graph analysis (CoNGA). *Nat. Biotechnol.* **40**, 54–63. https://doi.org/10.1038/s41587-021-00989-2 (2022).
27. Polanski, K. *et al.* BBKNN: Fast batch alignment of single cell transcriptomes. *Bioinformatics* **36**, 964–965. https://doi.org/10.1093/bioinformatics/btz625 (2020).
28. McInnes, L., Healy, J. & Melville, J. *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. 1802.03426v3 (2018).
29. Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: Guaranteeing well-connected communities. *Sci. Rep.* **9**, 5233. https://doi.org/10.1038/s41598-019-41695-z (2019).
30. Wickham, H. & Sievert, C. *Ggplot2: Elegant Graphics for Data Analysis* (Springer, 2016). https://doi.org/10.1007/978-3-319-24277-4.

31. Ruscher, R., Kummer, R. L., Lee, Y. J., Jameson, S. C. & Hogquist, K. A. CD8alphaalpha intraepithelial lymphocytes arise from two main thymic precursors. *Nat. Immunol.* **18**, 771–779. https://doi.org/10.1038/ni.3751 (2017).
32. Shugay, M. *et al.* VDJtools: Unifying post-analysis of T cell receptor repertoires. *PLoS Comput. Bio.l* **11**, e1004503. https://doi.org/10.1371/journal.pcbi.1004503 (2015).
33. Shugay, M. *vdjtools Documentation: Release snapshot*. https://readthedocs.org/projects/vdjtools-doc/downloads/pdf/master/ (2018).
34. Rafei, M. *et al.* Development and function of innate polyclonal TCRalphabeta+ CD8+ thymocytes. *J. Immunol.* **187**, 3133–3144. https://doi.org/10.4049/jimmunol.1101097 (2011).
35. Pita-Lopez, M. L., Pera, A. & Solana, R. Adaptive memory of human NK-like CD8(+) T-cells to aging, and viral and tumor antigens. *Front. Immunol.* **7**, 616. https://doi.org/10.3389/fimmu.2016.00616 (2016).
36. Miyazawa, S. & Jernigan, R. L. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J. Mol. Biol.* **256**, 623–644. https://doi.org/10.1006/jmbi.1996.0114 (1996).
37. Tilloy, F. *et al.* An invariant T cell receptor alpha chain defines a novel TAP-independent major histocompatibility complex class Ib-restricted alpha/beta T cell subpopulation in mammals. *J. Exp. Med.* **189**, 1907–1921. https://doi.org/10.1084/jem.189.12.1907 (1999).
38. Treiner, E. *et al.* Selection of evolutionarily conserved mucosal-associated invariant T cells by MR1. *Nature* **422**, 164–169. https://doi.org/10.1038/nature01433 (2003).
39. Kitaura, K., Shini, T., Matsutani, T. & Suzuki, R. A new high-throughput sequencing method for determining diversity and similarity of T cell receptor (TCR) alpha and beta repertoires and identifying potential new invariant TCR alpha chains. *BMC Immunol.* **17**, 38. https://doi.org/10.1186/s12865-016-0177-5 (2016).
40. Baldwin, T. A., Sandau, M. M., Jameson, S. C. & Hogquist, K. A. The timing of TCR alpha expression critically influences T cell development and selection. *J. Exp. Med.* **202**, 111–121. https://doi.org/10.1084/jem.20050359 (2005).
41. Aifantis, I. *et al.* The E delta enhancer controls the generation of CD4- CD8- alphabetaTCR-expressing T cells that can give rise to different lineages of alphabeta T cells. *J. Exp. Med.* **203**, 1543–1550. https://doi.org/10.1084/jem.20051711 (2006).
42. Carico, Z. M., Roy-Choudhury, K., Zhang, B., Zhuang, Y. & Krangel, M. S. Tcrd rearrangement redirects a processive Tcra recombination program to expand the Tcra repertoire. *Cell Rep.* **19**, 2157–2173. https://doi.org/10.1016/j.celrep.2017.05.045 (2017).
43. Hou, X. *et al.* Shorter TCR beta-chains are highly enriched during thymic selection and antigen-driven selection. *Front. Immunol.* **10**, 299. https://doi.org/10.3389/fimmu.2019.00299 (2019).
44. Li, H. M. *et al.* TCRbeta repertoire of CD4+ and CD8+ T cells is distinct in richness, distribution, and CDR3 amino acid composition. *J. Leukoc. Biol.* **99**, 505–513. https://doi.org/10.1189/jlb.6A0215-071RR (2016).
45. Yadav, M. *et al.* Neuropilin-1 distinguishes natural and inducible regulatory T cells among regulatory T cell subsets in vivo. *J. Exp. Med.* **209**, 1713–1722. https://doi.org/10.1084/jem.20120822 (2012).
46. Koenecke, C. *et al.* Alloantigen-specific de novo-induced Foxp3+ Treg revert in vivo and do not protect from experimental GVHD. *Eur. J. Immunol.* **39**, 3091–3096. https://doi.org/10.1002/eji.200939432 (2009).
47. Kanamori, M., Nakatsukasa, H., Okada, M., Lu, Q. & Yoshimura, A. Induced regulatory T cells: Their development, stability, and applications. *Trends Immunol.* **37**, 803–811. https://doi.org/10.1016/j.it.2016.08.012 (2016).
48. Bilate, A. M. *et al.* T cell receptor is required for differentiation, but not maintenance, of intestinal CD4(+) intraepithelial lymphocytes. *Immunity* **53**, 1001-1014.e20. https://doi.org/10.1016/j.immuni.2020.09.003 (2020).

Acknowledgements

We are highly grateful to donors and parents for their participation. Karl-Andreas Dumont and staff at the Department of Cardiothoracic Surgery at the Oslo University Hospital facilitated sample collection. The sequencing service was provided by the Norwegian Sequencing Centre (www.sequencing.uio.no), a national technology platform hosted by the University of Oslo and supported by the "Functional Genomics" and "Infrastructure" programs of the Research Council of Norway and the Southeastern Regional Health Authorities. Computational resources were provided by the Norwegian Research Infrastructure Services (NRIS) and Tjenester for Sensitive Data (TSD) facilities, owned by the University of Oslo.

Author contributions

Conceptualisation, M.H., S.T.F., H.S.H., B.A.L.; Methodology, M.H., S.T.F., H.S.H., M.C., Q.K.L., V.G., B.A.L.; Formal analysis, M.H, P.M.B.; Investigation, M.H., S.T.F., H.S.H.; Data curation, M.H., P.M.B., B.A.L.; Writing – Original Draft, M.H., V.G., B.A.L.; Writing – Review & Editing, M.H., S.T.F., H.S.H., P.M.B, M.C., Q.K.L., X.T, V.G., B.A.L.; Visualisation, M.H., P.M.B., B.A.L.; Supervision, X.T., V.G., B.A.L.; Project Administration, B.A.L.; Funding acquisition, B.A.L.

Funding

The Norwegian Research Council (#214280, #301536, and #274718 to B.A.L; #300740, #331890 and #31341 to V.G), the Norwegian Diabetes Association (to B.A.L), the South-Eastern Norway Regional Health Authorities (#2021017 to B.A.L.), UiO: LifeScience Convergence Environment Immunolingo (to V.G.), EU Horizon 2020 iReceptorplus (#825821) (to V.G.).

Competing interests

V.G. declares advisory board positions in aiNET GmbH, Enpicom B.V, Absci, Omniscope, and Diagonal Therapeutics. V.G. is a consultant for Adaptyv Biosystems, Specifica Inc, Roche/Genentech, immunai, and LabGenius. Remaining authors declare that they have no competing interest.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-44693-4.

Correspondence and requests for materials should be addressed to B.A.L.