Transposable elements regulate thymus development and function

Authors

Jean-David Larouche[1,2], Céline M. Laumont[3,4], Assya Trofimov[1,5,6,7], Krystel Vincent[1], Leslie Hesnard[1], Sylvie Brochu[1], Caroline Côté[1], Juliette Humeau[1], Éric Bonneil[1], Joël Lanoix[1], Chantal Durette[1], Patrick Gendron[1], Jean-Philippe Laverdure[1], Ellen R. Richie[8], Sébastien Lemieux[1,9], Pierre Thibault[1,10], Claude Perreault[1,2]*.

Affiliations

[1] Institute for Research in Immunology and Cancer, Université de Montréal; Montréal, Canada.

[2] Department of Medicine, Université de Montréal; Montréal, Canada.

[3] Deeley Research Centre, BC Cancer; Victoria, Canada.

[4] Department of Medical Genetics, University of British Columbia; Vancouver, Canada.

[5] Department of Computer Science and Operations Research, Université de Montréal; Montréal, Canada.

[6] Department of Physics, University of Washington; Seattle, USA.

[7] Fred Hutchinson Cancer Center; Seattle, USA.

[8] Department of Epigenetics and Molecular Carcinogenesis, University of Texas M.D. Anderson Cancer Center; Houston, USA.

[9] Department of Biochemistry and Molecular Medicine, Université de Montréal; Montréal, Canada.

[10] Department of Chemistry, Université de Montréal; Montréal, Canada.

*Corresponding author. Email: claude.perreault@umontreal.ca

Abstract

Transposable elements (TE) are repetitive sequences representing ~45% of the human and mouse genomes and are highly expressed by medullary thymic epithelial cells (mTEC). In this study, we investigated the role of TEs on T-cell development in the thymus. We performed multi-omic analyses of TEs in human and mouse thymic cells to elucidate their role in T cell development. We report that TE expression in the human thymus is high and shows extensive age- and cell lineage-related variations. TE expression correlates with multiple transcription factors in all cell types of the human thymus. Two cell types express particularly broad TE repertoires: mTECs and plasmacytoid dendritic cells (pDC). In mTECs, transcriptomic data suggest that TEs interact with transcription factors essential for mTEC development and function (e.g., PAX1 and REL), and immunopeptidomic data showed that TEs generate MHC-I-associated peptides implicated in thymocyte education. Notably, AIRE, FEZF2, and CHD4 regulate small yet non-redundant sets of TEs in murine mTECs. Human thymic pDCs homogenously express large numbers of TEs that likely form dsRNA, which can activate innate immune receptors, potentially explaining why thymic pDCs constitutively secrete IFN □/β. This study highlights the diversity of interactions between TEs and the adaptive immune system. TEs are genetic parasites, and the two thymic cell types most affected by TEs (mTEcs and pDCs) are essential to establishing central T-cell tolerance. Therefore, we propose that orchestrating TE expression in thymic cells is critical to prevent autoimmunity in vertebrates.

Introduction

Self/non-self discrimination is a fundamental requirement of life (1). In jawed vertebrates, the thymus is the only site where T lymphocytes can be properly educated to distinguish self from non-self (2, 3). This is vividly illustrated by Oncostatin M-transgenic mice, where T-cell production occurs exclusively in the lymph nodes (4). These mice harbor normal numbers of T-cell receptors (TCR) αβ T cells but present severe autoimmunity and cannot fight infections (5). Intrathymic generation of a functional T-cell repertoire depends on choreographed interactions between the TCRs of thymocytes and peptides presented by major histocompatibility complex (MHC) molecules on various antigen-presenting cells (APC) (6). Positive selection depends on self-antigens presented by cortical thymic epithelial cells (cTEC) and ensures that TCRs recognize antigens in the context of the host's MHC molecules (7, 8). The establishment of central tolerance depends on two main classes of APCs located in the thymic medulla: dendritic cells (DC) and medullary TEC (mTEC) (9-11). Two other APC types have a more limited contribution to central tolerance: thymic fibroblasts and B cells (12, 13). High avidity interactions between thymic APCs and autoreactive thymocytes lead to thymocyte deletion (negative selection) or generation of regulatory T cells (Treg) (14).

The main drivers of central tolerance, mTECs and DCs, display considerable phenotypic and functional heterogeneity. Indeed, recent single-cell RNA-seq (scRNA-seq) studies have identified several subpopulations of mTECs: immature mTEC(I) that stimulate thymocyte migration to the medulla via chemokine secretion (15), mTEC(II) that express high levels of MHC and are essential to tolerance induction, fully differentiated corneocyte-like mTEC(III) that foster a pro-inflammatory microenvironment (16), and finally mimetic mTECs that express peripheral tissue antigens (17). Three different proteins whose loss of function leads to severe autoimmunity, AIRE, FEZF2, and CHD4, have

been shown to drive the expression of non-redundant sets of peripheral tissue antigens in mTECs (18-20). DCs, on the other hand, are separated into three main populations. Conventional DC 1 and 2 (cDC1 and cDC2) have an unmatched ability to present both endogenous antigens and exogenous antigens acquired via cross-presentation or cross-dressing (21). Plasmacytoid DC (pDC) are less effective APCs than cDCs, their primary role being to produce interferon alpha (IFN□) (21). Notably, thymic pDCs originate from intrathymic IRF8^{hi} precursors, and, in contrast to extrathymic pDCs, they constitutively secrete high amounts of IFN□ (22-24). This constitutive IFN□ secretion by thymic pDCs regulates the late stages of thymocyte development by promoting the generation of Tregs and innate CD8 T cells (25-29).

Transposable elements (TE) are repetitive sequences representing ~45% of the human and mouse genomes (30, 31). Most TEs can be grouped into three categories: the long and short interspersed nuclear elements (LINE and SINE, respectively) and the long terminal repeats (LTR). These broad categories are subdivided into over 800 subfamilies based on sequence homology (32). TE expression is typically repressed in host cells to prevent deleterious integrations of TE sequences in protein-coding genes (33). Unexpectedly, TEs were recently found to be expressed at higher levels in human mTECs than in any other MHC-expressing tissues and organs (i.e., excluding the testis) (34, 35), suggesting a role for TEs in thymopoiesis. Since some TEs are translated and generate MHC I-associated peptides (MAP) (34), they might induce TE-specific central tolerance (36). Additionally, TEs provide binding sites to transcription factors (TF) and stimulate cytokine secretion via the formation of double-stranded RNA (dsRNA) (37-41). Hence, TEs could have pleiotropic effects on thymopoiesis. To evaluate the role of TEs in thymopoiesis, we adopted a multipronged strategy beginning with scRNA-seq of human thymi and culminating in MS analyses of the MAP repertoire of mouse mTECs.

Results

LINE, LTR, and SINE expression shows extensive variations during ontogeny of the human thymus

We first profiled TE expression in various thymic cell populations during development. To do so, we quantified the expression of 809 TE subfamilies (classified according to the RepeatMasker annotations) in the scRNA-seq dataset of human thymi created by *Park et al.* (42). Cells were clustered in 19 populations representing the main constituents of the thymic hematolymphoid and stromal compartments (Figure 1a and Figure 1 – figure supplement 1). The expression of TE subfamilies was quantified at all developmental stages available, ranging from 7 post-conception weeks (pcw) to 40 years of age (Supplementary file 1 – Table 1). Unsupervised hierarchical clustering revealed three clusters of TE subfamilies based on their pattern of expression during thymic development (Figure 1b, upper panel): i) maximal expression at early embryonic stages persisting, albeit at lower levels, throughout ontogeny (cluster 1), ii) an expression specific to a given timepoint (cluster 2), or iii) a high expression at early embryonic stages that decreases rapidly at later timepoints (cluster 3). LINE and SINE subfamilies were enriched in cluster 1, whereas LTR subfamilies were significantly enriched in clusters 2 and 3 (Figure 1b, lower panel). Expression of individual LINE and SINE subfamilies was highly shared among different cell types (Figure 1d). In contrast, the LTR subfamilies' expression pattern was shared by fewer cell subsets and adopted a quasi-random distribution (Figure 1d). The pattern of expression assigned to TE subfamilies (Figure 1c, innermost track) was not affected by the proportion of cells of different developmental stages (embryonic or postnatal) (Figure 1c, outermost track, and Figure 1 – figure supplement 2). This suggests that our observations do not result from a bias in the composition of the dataset. To gain further insights into the expression of TE subfamilies, we

studied two biological processes known to regulate TE expression in other contexts: cell proliferation and expression of KRAB zinc-finger proteins (KZFP) (43, 44). Cell cycling scores negatively correlated with TE expression in various thymic cell subsets, particularly for LINE and SINE subfamilies shared among cell types (Figure 1 – figure supplement 3 and Supplementary file 1 – Table 2), whereas analysis of KZFP expression identified ZNF10 as a probable repressor of L1 subfamilies in Th17 and NK cells (Figure 1 – figure supplement 4 and Supplementary file 1 – Table 3). Thus, we conclude that the expression of the three main classes of TEs shows major divergences as a function of age and thymic cell types.

TEs form interactions with transcription factors regulating thymic development and function

TEs provide binding sites to TFs (37, 45, 46), and T-cell development is driven by the coordinated timing of multiple changes in transcriptional regulators (47). We, therefore, investigated interactions between TE subfamilies and TFs during the development of the human thymus. Two criteria defined an interaction: i) a significant and positive correlation between the expression of a TF and a TE subfamily in a given cell population, and ii) the presence of the TF binding motif in the loci of the TE subfamily (Figure 2a). Additionally, we validated the correlations we obtained using a bootstrap procedure to ascertain their reproducibility (see *Material and Methods* for details). This procedure removed weakly correlated TF-TE pairs (Figure 2b). TF-TE interactions were observed in all thymic cell populations (Figure 2c, d, Figure 2 – figure supplement 1, and Supplementary file 1 – Table 4). Numerous TF-TE interactions were conserved between hematolymphoid and stromal cell subsets (Figure 2e). However, the number of interactions and the complexity of the interaction networks were much higher in mTECs than in other cell populations (Figure 2c, d, Figure 2 – figure supplement 1, and Figure 2 – figure supplement 2).

Several TFs instrumental in thymus development and thymopoiesis interacted with TE subfamilies (Figure 2 – figure supplement 2, and Supplementary file 1 – Table 4). These TFs include the *NFKB1* and *REL* subunits of the NF-κB complex and *PAX1* in mTECs (48-50) and *JUND* in thymocytes (51). In DCs, the most notable TF-TE interactions involved interferon regulatory factors (IRF), which regulate the late stages of T-cell maturation, and *TCF4*, which is essential for pDC development (25, 52). This observation is consistent with evidence that TEs have shaped the evolution of IFN signaling networks (37). Finally, we found significant interactions between *CTCF* and TE subfamilies in mTECs and endothelial cells, suggesting that the binding of *CTCF* to TE sequences affects the tridimensional structure of the chromatin in the thymic stroma (53). Interestingly, LINE and SINE subfamilies that occupy more genomic space interacted with higher numbers of transcription factors (Figure 2 – figure supplement 3).

Using data from the ENCODE consortium for hematopoietic cells (54, 55), we looked at the histone marks at the TE loci identified as TF interactors by our analyses (i.e., correlated with TF expression and containing the TF binding motif). The objective was to determine if they could act as promoters or enhancers (Figure 2a and Supplementary file 1 – Table 5). We found several TE promoter and enhancer candidates in all eight hematopoietic cell types analyzed, with a striking overrepresentation of LINE and SINE compared to LTR sequences (Figure 2f and Supplementary file 1 – Table 6). Finally, we analyzed publicly available ChIP-seq data of ETS1, an important TF for NK cell development (56), to confirm its ability to bind TE sequences. Indeed, 19% of ETS1 peaks overlap with TE sequences (Figure 2g). Notably, ETS1 peaks overlapped with TE sequences (Figure 2h, in red) in the promoter regions of PRF1

and KLRD1, two genes critical for NK cells' effector functions (57, 58). Hence, our data suggest that TEs affect thymic development and function by providing binding sites to multiple TFs.

TEs are highly and differentially expressed in human thymic APC subsets

We next sought to determine whether the high expression of TEs reported in mTECs (32, 33) was limited to this cell subset or was found in other thymic cell types. Since several thymic stromal cells reach maturity after birth (59), we selected postnatal samples for the following analyses. We computed two distinct Shannon entropy indices: one for the global diversity of TEs expressed by all cells of a given population, and another for the median value of TE diversity expressed by individual cells of a population (Figure 3a). Then, we computed a linear model to represent the diversity of TEs expressed by a cell population based on the diversity of TEs expressed by individual cells (Figure 3a, blue curve). Two salient findings emerged from this analysis. First, the diversity of TEs expressed in the T-cell lineage decreases during differentiation according to the following hierarchy: DN thymocytes > DP thymocytes > SP thymocytes (Figure 3a, Figure 3 – figure supplement 1). Second, among the populations of thymic APCs implicated in positive and negative selection (Figure 3a, orange dots), cTECs, mTECs, and DCs expressed broader repertoires of TEs than B cells and fibroblasts. While cTECs and DCs expressed highly diverse TE repertoires at both the population and individual cell levels, the breadth of TE expression in mTECs was found only at the population level (Figure 3a). Accordingly, intercellular heterogeneity (i.e., deviation from the linear model) was higher for mTECs than other cell populations (Figure 3b).

We next focused on thymic APCs expressing the broadest TE repertoires: cTECs, mTECs, and DCs (Figure 3a). To this end, we annotated these APC subpopulations based on previously published lists of marker genes (Figure 3c and Figure 3 – figure supplement 2) (42, 60). We performed differential expression analyses to determine whether some TE subfamilies were overexpressed in specific APC subsets. pDCs and mTEC(II) overexpressed a broader TE repertoire than other APCs: 32.01% of subfamilies were overexpressed in pDCs and 10.88% in mTEC(II) (Figure 3d and Supplementary file 1 – Table 7). The nature of the overexpressed TEs differed between pDCs and other thymic APC subsets. Indeed, pDCs overexpressed LTRs, LINEs, and SINEs, including several Alu and L1 subfamilies (Figure 3d and Supplementary file 1 – Table 7). In contrast, other thymic APCs predominantly overexpressed LTRs.

TE expression showed wildly divergent levels of intercellular heterogeneity in APC subsets. Indeed, whereas most TE subfamilies were expressed by <25% of cells of the mTEC(II) population, an important proportion of TEs were expressed by >75% of pDCs (Figure 3e). To evaluate this question further, we compared TE expression between metacells of thymic APCs; metacells are small clusters of cells with highly similar transcription profiles. This analysis revealed that overexpression of TE subfamilies was shared between pDC metacells but not mTEC(II) metacells, reinforcing the idea that TE expression adopts a mosaic pattern in the mTEC(II) population (Figure 3 – figure supplement 3). We conclude that cTECs, mTECs, and DCs express broad TE repertoires. However, two subpopulations of thymic APCs clearly stand out. pDCs express an extremely diversified repertoire of LTRs, SINEs, and LINEs, showing limited intercellular heterogeneity, whereas the mTEC(II) population shows a highly heterogeneous overexpression of LTR subfamilies.

TE expression in human pDCs is associated with dsRNA structures

The high expression of a broad repertoire of TE sequences in thymic pDCs was unexpected (Figure 3d). LINE and SINE subfamilies, in particular, were highly and homogeneously expressed by thymic pDCs (Figure 4a). Constitutive IFN secretion is a feature of thymic pDCs not found in extrathymic pDCs. We, therefore, hypothesized that this constitutive IFN secretion by thymic pDCs might be mechanistically linked to their TE expression profile. We first assessed whether thymic and extrathymic pDCs have similar TE expression profiles by reanalyzing scRNA-seq data from human spleens published by *Madissoon et al.* (61) (Figure 4 – figure supplement 1a, b). This revealed that extrathymic pDCs express TE sequences at similar or lower levels than other splenic cells (Figure 4 – figure supplement 1c, d). We then used pseudobulk RNA-seq methods to perform a differential expression analysis of TE subfamilies between thymic and splenic pDCs. This analysis confirmed that TE expression was globally higher in thymic than in extrathymic pDCs (Figure 4b). Since TE overexpression can lead to the formation of dsRNA (40, 41), we investigated if such structures were found in thymic pDCs. pDCs were magnetically enriched from primary human thymi following labeling with anti-CD303 antibody (a marker of pDCs). Then, pDC-enriched thymic cells were stained with an antibody against CD123 (another marker of pDCs) and the J2 antibody that stains dsRNA. The intensity of the J2 signal was more than 10-fold higher in $CD123^{+}$ relative to $CD123^{-}$ cells (Figure 4c, d). We conclude that thymic pDCs contain large amounts of dsRNAs. To evaluate if these dsRNAs arise from TE sequences, we analyzed in thymic APC subsets the proportion of the transcriptome assigned to two groups of genomic sequences known as important sources of dsRNAs: TEs and mitochondrial genes (62). Strikingly, whereas the percentage of reads from mitochondrial genes was typically lower in pDCs than in other thymic APCs, the proportion of the transcriptome originating from TEs was higher in pDCs (~22%) by several orders of magnitude (Figure 4 – figure supplement 2). Finally, we performed

gene set enrichment analyses to ascertain if the high expression of TEs by thymic pDCs was associated with specific gene signatures. These analyses highlighted signatures of antigen presentation, immune response, and interferon signaling in thymic pDCs (Figure 4e and Supplementary file 1 – Table 8). Notably, thymic pDCs harbored moderate yet significant enrichment of gene signatures of RIG-I and MDA5-mediated IFN □/β signaling compared to all other thymic APCs (Figure 4e and Supplementary file 1 – Table 8). Altogether, these data support a model in which the high and ubiquitous expression of TEs in thymic pDCs would lead to the formation of dsRNAs triggering innate immune sensors, which might explain their constitutive secretion of IFN □/β.

AIRE, CHD4, and FEZF2 regulate distinct sets of TE sequences in murine mTECs

The essential role of mTECs in central tolerance hinges on their ability to ectopically express tissue-restricted genes, whose expression is otherwise limited to specific epithelial lineage (63, 64). This promiscuous gene expression is driven by AIRE, CHD4, and FEZF2 (18-20). We, therefore, investigated the contribution of these three genes to the expression of TE subfamilies in the mTEC(II) population (Figure 3d). First, we validated that mTEC(II) express *AIRE*, *CHD4*, and *FEZF2* in the human scRNA-seq dataset (Figure 5a). Next, we analyzed published murine mTEC RNA-seq data to assess the regulation of TE sequences by AIRE, CHD4, and FEZF2. Differential expression analyses between knock-out (KO) and wild-type (WT) mice showed that these three factors regulate TE sequences, but the magnitude and directionality of this regulation differed (Figure 5b and Supplementary file 1 – Table 9). Indeed, while CHD4 had the biggest impact on TE expression by inducing 433 TE loci and repressing 463, FEZF2's impact was minimal, with 97 TE loci induced and 60 repressed (Figure 5b). Besides, AIRE mainly acted as a repressor of TE sequences, with 326 loci repressed and 171 induced (Figure 5b). Interestingly, there was minimal overlap between the TE sequences regulated by

AIRE, CHD4, and FEZF2, indicating that they have non-redundant roles in TE regulation (Figure 5c). Additionally, AIRE, CHD4, and FEZF2 preferentially targeted LTR and LINE elements, with significant enrichment of specific subfamilies such as MTA_Mm-int and RLTR4_Mm that are induced by Aire and Fezf2, respectively (Figure 5d and Figure 5 – figure supplement 1a). While AIRE and CHD4 preferentially targeted evolutionary young TE sequences, the age of the TE sequence did not seem to affect the regulation by FEZF2 (Figure 5 – figure supplement 1b). We also noticed that the distance between regulated TE loci was smaller than the distributions of randomly selected TEs (Figure 5 – figure supplement 1c). This suggests that AIRE, CHD4, and FEZF2 nonrandomly affect the expression of TE sequences located in specific genomic regions. We observed no significant differences in the genomic localization of TE loci targeted by AIRE, CHD4, and FEZF2 relative to the genomic localization of all TE sequences in the murine genome: most TE loci were located in intronic and intergenic regions (Figure 5 – figure supplement 1d). Enrichment for intronic TEs could not be ascribed to induction of global intron retention: the intron retention ratio was similar for TEs regulated or not by AIRE, CHD4, and FEZF2 (Figure 5 – figure supplement 1e). ChIP-seq-based analysis of permissive histone marks showed that TE loci induced by AIRE, CHD4, and FEZF2 were all marked by H3K4me3 (Figure 5e). As a proof of concept, we validated that 31.42% of AIRE peaks overlap with TE sequences by reanalyzing ChIP-seq data, confirming AIRE's potential to bind TE sequences (Figure 5f). Hence, AIRE, CHD4, and FEZF2 regulate the expression of small yet non-redundant repertoires of TE sequences associated with permissive histone marks.

TEs are translated and presented by MHC class I molecules in murine TECs

Several TEs are translated and generate MAPs (34). Hence, the expression of TEs in cTECs and even more in mTECs raises a fundamental question: do these TEs generate MAPs that would shape the T cell

repertoire? Mass spectrometry (MS) is the only method that can faithfully identify MAPs (65-67). Despite its quintessential role in central tolerance, the MAP repertoire of mTECs has never been studied by MS because of the impossibility of obtaining sufficient mTECs for MS analyses: mTECs represent ≤ 1% of thymic cells, and they do not proliferate *in vitro*. To get enough cTECs and mTECs for MS analyses, we used transgenic mice that express cyclin D1 under the control of the keratin 5 promoter (K5D1 mice). These mice develop dramatic thymic hyperplasia, but their thymus is morphologically and functionally normal (68-70). Primary cTECs and mTECs (2 replicates of 70 x 10^6 cells from 121 and 90 mice, respectively) were isolated from the thymi of K5D1 mice as described (71). Following cell lysis and MHC I immunoprecipitation, MAPs were analyzed by liquid chromatography MS/MS (Figure 6a). To identify TE-coded MAPs, we generated a TE proteome by in silico translation of TE transcripts expressed by mTECs or cTECs, and this TE proteome was concatenated with the canonical proteome. MS analyses enabled the identification of a total of 1636 and 1714 MAPs in mTECs and cTECs, respectively. From these, we identified 4 TE-derived MAPs in mTECs and 2 in cTECs, demonstrating that TEs can be translated and presented by MHC I in the thymic cortex and medulla (Figure 6b and Supplementary file 1 – Table 10). These MAPs were coded by the three major groups of TE: LINEs (n=1), LTRs (n=1), and SINEs (n=4). Next, we evaluated whether the low number of TE MAPs identified could result from mass spectrometry detection limits (72, 73). We measured the level and frequency of TE expression in two subsets of cTECs (Figure 6c, left) or mTECs (Figure 6c, right) using scRNA-seq data from *Baran-Gale et al.* (74). TE subfamilies generating MAPs in cTECs or mTECs are highlighted in red in their respective plots. Strikingly, TECs highly and ubiquitously expressed the MAP-generating TE subfamilies. These results suggest that the contribution of TEs to the MAP repertoire of cTECs and mTECs might be significantly underestimated by the limits of detection of MS. This is particularly true for mTECs because they express high levels of TEs (Figure 3d), but their TE

profile displays considerable intercellular heterogeneity (Figure 3e and Figure 3 – figure supplement 2). Nonetheless, our data provide direct evidence that TEs can generate MAPs presented by cTECs and mTECs, which can contribute to thymocyte education.

Discussion

TEs are germline-integrated parasitic DNA elements that comprise about half of mammalian genomes. Over evolutionary timescales, TE sequences have been co-opted for host regulatory functions. Mechanistically, TEs encode proteins and noncoding RNAs that regulate gene expression at multiple levels (32, 75). Regulation of IFN signaling and triggering innate sensors are the best-characterized roles of TEs in the mammalian immune system (36). TEs are immunogenic and can elicit adaptive immune responses implicated in autoimmune diseases (34, 36, 76, 77). Pervasive TE expression in various somatic organs means that co-evolution with their host must depend on establishing immune tolerance, a concept supported by the highly diversified TE repertoire expressed in mTECs (34). This observation provided the impetus to perform multi-omic studies of TE expression in the thymus. At the whole organ level, we found that TE expression showed extensive age- and cell lineage-related variations and was negatively correlated with cell proliferation and expression of KZFPs. The negative correlation between TE expression and cell cycle scores in the thymus is coherent with recent data showing that transcriptional activity of L1s is increased in senescent cells (78). A potential rationale for this could be to prevent deleterious transposition events during DNA replication and cell division. On the other hand, the contribution of KZFPs to TE regulation in the thymus is likely underestimated due to their typically low expression (79) and scRNA-seq detection limit. Additionally, TEs interact with multiple TFs in all thymic cell subsets. This is particularly true for the LINE and SINE subfamilies that occupy larger

genomic spaces. Notably, TEs appear to play particularly important roles in two cell types located in the thymic medulla: mTECs and pDCs.

As mTECs are the APC population crucial to central tolerance induction, their high and diverse TE expression is poised to impact the T cell repertoire's formation profoundly. The extent and complexity of TF-TE interactions were higher in mTECs than in all other thymic cell subsets. These interactions included *PAX1* and subunits of the NF-κB complex (e.g., *RELB*). PAX1 is essential for the development of TEC progenitors (50), and *RELB* is for the development and differentiation of mTECs (80). RelB-deficient mice have reduced thymic cellularity, markedly fewer mTECs, lack *Aire* expression, and suffer from autoimmunity (49, 81). Under the influence of *Aire*, *Fezf2*, and *Chd4*, mTECs collectively express almost the entire exome (63, 64). However, the expression of all genes in each mTEC would cause proteotoxic stress (64). Hence, promiscuous expression of tissue-restricted genes in mTECs adopts a mosaic pattern: individual tissue-restricted genes are expressed in a small fraction of mTECs (17, 82). The present work shows that mTECs also express an extensive repertoire of TEs in a mosaic pattern (i.e., with considerable intercellular heterogeneity). *Aire*, *Fezf2*, and *Chd4* regulate non-redundant sets of TEs and preferentially induce TE sequences associated with permissive histone marks. The immunopeptidome of thymic stromal cells is responsible for thymocyte education and represents one of the most fundamental “known unknowns” in immunology. Inferences on the immunopeptidome of thymic stromal cells are based on transcriptomic data. However, i) TCRs interact with MAPs, not transcripts, and ii) the MAP repertoire cannot be inferred from the transcriptome (65, 83, 84). Using K5D1 mice presenting prominent thymic hyperplasia, we conducted MS searches of TE MAPs, identifying 4 TE MAPs in mTECs and 2 in cTECs. These results demonstrate that cTECs and mTECs present TE MAPs and suggest they present different TE MAPs. However, the correlation between

transcriptomic and immunopeptidomic data suggests that TECs can present many more TE MAPs. Their profiling will require MS analyses of enormous numbers of TECs or the development of more sensitive MS techniques. As TE MAPs have been detected in normal and neoplastic extrathymic cells (34, 85-87), the presentation of TEs by mTECs is likely essential to central tolerance. In line with vibrant plaidoyers for a collaborative Human Immunopeptidome Project (66, 88), our work suggests that immunopeptidomic studies should not be limited to protein-coding genes (2% of the genome) but also encompass non-coding sequences such as TEs.

The second population of cells exhibiting high TE expression, pDCs, are mainly seen as producers of IFN □/β and potentially as APCs (21). Thymic and extrathymic pDCs are ontogenically and functionally different. They develop independently from each other from different precursor cells (23, 24, 89). IFN □/β secretion is inducible in extrathymic pDCs but constitutive in thymic pDCs (21, 22). In line with the location of pDCs in the thymic medulla, their constitutive IFN □/β secretion is instrumental in the terminal differentiation of thymocytes and the generation of Tregs and innate CD8 T cells (25-29). We report here that high TE expression is also a feature of thymic, but not extrathymic, pDCs. Thus, the present study provides a rationale for the constitutive IFN □/β secretion by thymic pDCs: they homogeneously express large numbers of TEs (in particular LINEs and SINEs), leading to the formation of dsRNAs that trigger RIG-I and MDA5 signaling that causes the constitutive secretion of IFN □/β. As such, our data suggest that recognition of TE-derived dsRNAs by innate immune receptors promotes a pro-inflammatory environment favorable to the establishment of central tolerance in the thymic medulla.

At first sight, the pleiotropic effects of TEs on thymic function may look surprising. It should be reminded that the integration of genetic parasites such as TEs is a source of genetic conflicts with the host. Notably, the emergence of adaptive immunity gave rise to higher-order conflicts between TEs and their vertebrate hosts (36, 90). The crucial challenge for the immune system is developing immune tolerance towards TEs to prevent autoimmune diseases that affect up to 10% of humans (91) without allowing selfish retrotransposition events that hinder genome integrity. The resolution of these conflicts has been proposed to be a determining factor in shaping the function of the immune system (90). Our data suggest that the thymus is the central battlefield for conflict resolution between TEs and T cells in vertebrates. Consistent with the implication of TEs in autoimmunity, more than 90% of putative causal variants associated with autoimmune diseases are in allegedly noncoding regions of the genome (91). In this context, our study illustrates the complexity of interactions between TEs and the vertebrate immune system and should provide impetus to explore them further in health and disease. We see two limitations to our study. First, as with all multi-omic systems immunology studies, our work provides a roadmap for many future mechanistic studies that could not be realized at this stage. Second, our immunopeptidomic analyses of TECs prove that TECs present TE MAPs but certainly underestimate the diversity of TE MAPs presented by cTECs and mTECs.

Methods

Experimental design

This study aimed to understand better the impacts of TE expression on thymus development and function. Thymic populations are complex and heterogeneous, so we opted for single-cell RNA-seq data to draw a comprehensive profile of TE expression in the thymus. To better understand the impact of AIRE, FEZF2, and CHD4 on TE expression in the mTEC(II) population, RNA-seq data from WT and KO murine mTEC, as well as ChIP-seq for different histone marks in murine mTECs, were reanalyzed

to characterize the TE sequences regulated by these three proteins. Unless stated otherwise, studies were done in human cells. For MS analyses, two replicates of 70 million cells from K5D1 mice (68) were injected for both cTECs and mTECs. All experiments were in accordance with the Canadian Council on Animal Care guidelines and approved by the *Comité de Déontologie de l'Expérimentation sur des Animaux* of Université de Montréal. Primary human thymi were obtained from 4-month-old to 12-year-old children undergoing cardiovascular surgeries at the CHU Sainte-Justine. This project was approved by the CHU Sainte-Justine Research Ethics Board (protocol and biobank #2126).

Transcriptomic data processing

Preprocessing of the scRNA-seq data was performed with kallisto (92), which uses an expectation-maximization algorithm to reassign multimapping reads based on the frequency of unique mappers at each sequence and bustools workflow. For human thymic data from *Park et al.* (42) and splenic data from *Madissoon et al.* (61), two different indexes were built for the pseudoalignment of reads with kallisto (version 0.46.0): one containing Ensembl 88 (GRCh38.88) transcripts used for the annotation of cell populations, and a second containing Ensembl 88 transcripts and human TE sequences (LINE, LTR, SINE) from RepeatMasker (93) which was used for all subsequent analyses of TE expression. For murine data from *Baran-Gale et al.* (74), cell-type annotations from the original publication were used, and an index containing mm10 transcripts and murine TE sequences from RepeatMasker was used to analyze TE expression. The cell barcodes were corrected, and the feature-barcode matrices were generated with the correct count functions of bustools (version 0.39.3) (94). For murine bulk RNA-seq data, an index composed of mm10 (GRCm38) transcripts and murine TE sequences from RepeatMasker was used for quantification with kallisto.

ChIP-seq data reanalysis

ChIP-seq data for i) ETS1 in human NK cells, ii) AIRE in murine mTECs, and iii) several histone marks of mTECs from WT mice were reanalyzed (see "**Availability of data and materials**" for the complete list). ETS1 ChIP-seq reads were aligned to the reference *Homo sapiens* genome (GRCh38) using bowtie2 (version 2.3.5) (95) with the --very-sensitive parameter. Multimapping reads were removed using the samtools view function with the -q 10 parameter, and duplicate reads were removed using the samtools markdup function with the -r parameter (96). Peak calling was performed with macs2 with the -m 5 50 parameter (97). Peaks overlapping with the ENCODE blacklist regions (98) were removed with bedtools intersect (99) with default parameters. Overlap of ETS1 peaks with TE sequences was determined using bedtools intersect with default parameters. BigWig files were generated using the bamCoverage function of deeptools2 (100), and genomic tracks were visualized in the USCS Genome Browser (101). For the murine histone marks and AIRE data, reads were aligned to the reference *Mus musculus* genome (mm10) using bowtie2 with the --very-sensitive parameter. Multimapping reads were removed using the samtools view function with the -q 10 parameter, and duplicate reads were removed using the samtools markdup function with the -r parameter. For histone marks, read coverage at the sequence body and flanking regions (+/- 3000 base pairs) of TE loci induced by AIRE, FEZF2, and CHD4 was visualized using ngs.plot.r (version 2.63) (102). For AIRE, peaks overlapping with the ENCODE blacklist regions were removed with bedtools intersect, and overlap of peaks with TE sequences was determined using bedtools intersect with default parameters.

Cell population annotation

Feature-barcode matrices were imported in R with SingleCellExperiment (version 1.12.0) (103). As a quality control, cells with less than 2000 UMI detected, less than 500 genes detected, or more than 5% reads assigned to mitochondrial genes were considered low quality and removed from the dataset with scuttle (version 1.0.4) (104). Cells with more than 7000 genes detected were considered doublets and

removed. Normalization of cell size factors was performed with scran (version 1.18.7) (105), and log-normalization of read counts was done with scuttle with default parameters. Variable regions of TCR and IG genes, as well as ribosomal and cell cycle genes (based on *Park et al*. (42)), were removed, and highly variable features were selected based on a mean-variance trend based on a Poisson distribution of noise with scran. Adjustment of sequencing depths between batches and mutual nearest neighbors (MNN) correction were computed with batchelor (version 1.6.3) (106). Cell clustering was performed with scran using the Jaccard index for edge weighting and the Louvain method for community detection. Lists of marker genes for human thymic cell populations and TEC subsets were taken from *Park et al.* (42) and *Bautista et al.* (60), whereas marker genes of splenic populations were based on *Madissoon et al*. (61).

TE expression throughout thymic development

The expression of TE subfamilies was obtained by summing the read counts of loci based on the RepeatMasker annotations. For each TE subfamily in each cell population, expression levels amongst developmental stages were normalized by dividing them with the maximal expression value. Next, the Euclidean distance between each TE subfamily in each cell population (based on their normalized expression across developmental stages) was computed, followed by unsupervised hierarchical clustering. The tree was then manually cut into three clusters, and enrichment of LINE, LTR, and SINE elements in these three clusters was determined using Fisher's exact tests. The cluster assigned to each TE subfamily in each cell population was visualized in a circos plot using the circlize package (version 0.4.14) (107) in R, and the percentage of each cell population found in embryonic or postnatal samples. Finally, we computed the frequency that each TE family was assigned to the three clusters, and the maximal value was kept. As a control, a random distribution of the expression of 809 TE subfamilies in 18 cell populations was generated. A cluster (cluster 1, 2, or 3) was randomly attributed for each

combination of TE subfamily and cell type, and the maximal occurrence of a given cluster across cell types was then computed for each TE subfamily. Finally, the LINE, LTR, and SINE elements distributions were compared to the random distribution with Kolmogorov-Smirnov tests.

Regulation of TE expression by cell proliferation and KZFPs

Proliferation scores were generated for each dataset cell using the CellCycleScoring function of Seurat (version 4.1.0). As per *Cowan et al.* (108), we combined previously published lists of G2M and S phase marker genes (109) to compute the proliferation scores. For each thymic cell population, we calculated the Spearman correlation between proliferation scores and the expression of TE subfamilies. The Benjamini-Hochberg method was applied to correct for multiple comparisons. Correlations were considered positive if the correlation coefficient was ≥ 0.2 and the adjusted p-value≤ 0.05, and negative if the coefficient was ≤ -0.2 and the adjusted p-value≤ 0.05. We also computed the median of all correlation coefficients for each cell population. We then assigned the class of each TE subfamily correlated with cell proliferation and compared this distribution to the distribution of classes of all TE subfamilies in the human genome. The percentage of overlap of the sets of TE subfamilies significantly correlated with cell proliferation was determined. A list of 401 human KZFPs was downloaded from *Imbeault et al.* (44). Spearman correlations between KZFP and TE expression were independently computed in each cell population with the same methodology as the cell proliferation analysis, and Benjamini-Hochberg correction for multiple comparisons was applied. The information on the enrichment of KZFPs within TE subfamilies was downloaded from *Imbeault et al.* (44). Sharing of KZFP-TE pairs between cell populations was represented using the circlize package.

Estimation of TE sequences' age

The sequence divergence (defined as the number of mismatches per thousand) was given by the milliDiv value in RepeatMasker. The milliDiv values of each TE locus were divided by the substitution rate of its

host's genome ($2{,}2 \times 10^{-9}$ mutation/year for *Homo sapiens* and 4.5×10^{-9} mutation/year for *Mus musculus* (110, 111)). Finally, the age of each TE subfamily was determined by averaging the age of all loci of the subfamily.

Interactions between TE subfamilies and transcription factors

We downloaded a list of 1638 transcription factors (TF) manually curated by *Lambert et al.* (112). For each cell population of the thymus, Spearman correlations were computed for each possible pair of TF and TE subfamily, and the Benjamini-Hochberg method was applied to correct the p-values for multiple comparisons. Correlations were considered significant if i) the correlation coefficient was ≥0.2, ii) the adjusted p-value was ≤0.05, and iii) the TF was expressed by ≥10% of the cells of the population. The correlations were validated using a bootstrap procedure (1000 iterations) to ensure their reproducibility. Briefly, we randomly selected n cells out of the n cells of a given population (while allowing cells to be selected multiple times). The empirical p-value was determined by dividing the number of iterations with a correlation coefficient <0.2 by the total number of iterations (1000). In parallel, the curated binding motifs of 945 TFs were downloaded from the JASPAR database. We then used the *Find Individual Motif Occurences* (FIMO) software (113) to identify the 100 000 genomic positions with the most significant matches for the TF binding motif. These lists of binding motif positions were then intersected with the positions of TE loci with the intersect function of BEDTools (version 2.29.2) (99), and the percentage of TE loci of each subfamily harboring TF binding motifs was determined. Thus, in a specific cell population of the thymus, a TF was considered as interacting with a TE subfamily if it satisfied two criteria: i) its expression was correlated with the one of the TE family (spearman coefficient ≥0.2, adjusted p-value ≤0.05 and expression of TF in ≥10% of cells), and ii) at least one locus of the TE subfamily contained a binding motif of the TF. For each cell population, networks of interactions between TF and TE subfamilies were generated with the network package (version 1.17.1)

(114) in R and represented with the ggnetwork package. For the sake of clarity, only the most significant interactions were illustrated for each cell type (i.e., correlation coefficient ≥0.3, TF binding sites in ≥1% of the loci of the TE subfamily, and TF expression in ≥10% of cells of the population). Sharing of TF-TE interactions between cell populations was represented with a chord diagram using the circlize package. For each TE subfamily, the number of interactions with TFs and the number of loci of the TE subfamily in the human genome were determined. Wilcoxon-Mann-Whitney tests were used to compare the number of interactions with TF of LTR, LINE, and SINE elements, whereas Kendall tau correlation was calculated between the number of interactions with TF and the number of loci of TE subfamilies.

Identification of TE promoter and enhancer candidates

From the previously identified list of TF-TE interactions, we isolated the specific loci containing TF binding sites from the subfamilies whose expression was positively correlated with the TF. To determine if these TE loci could act as promoters or enhancers, we used histone ChIP-seq data from the ENCODE consortium for H3K27ac, H3K4me1, and H3K4me3. BED files from the ENCODE consortium were downloaded for eight immune cell populations: B cells, CD4 Single Positive T cells (CD4 SP), CD8 Single Positive T cells (CD8 SP), dendritic cells (DC), monocytes and macrophages (Mono/Macro), NK cells, Th17, and Treg. TE loci colocalizing with peaks in histone ChIP-seq data were identified using the intersect function of BEDTools (version 2.29.2). To be considered enhancer candidates, TE loci had to colocalize with H3K27ac and H3K4me1 but not H3K4me3. To be considered as promoter candidates, TE loci had to colocalize with H3K27ac and H3K4me3, but not H3K4me1, and be located at ≤ 1000 nucleotides from a transcription start site (TSS) annotated in the refTSS database (115).

Diversity of TE expression

The human thymic scRNA-seq dataset was subsampled to retain only postnatal cells, as it was shown by *Bornstein et al.* (56) that thymic APCs are mainly found in postnatal samples. The diversity of TE

sequences expressed by thymic populations was assessed using Shannon entropy. Using the vegan package (version 2.5-7) (116) in R, two distinct Shannon entropy metrics were computed for each cell population. First, the Shannon entropy was computed based on the expression level (i.e., log(read count)) of TE subfamilies for each cell individually. The median entropy was calculated for each cell population. In parallel, the diversity of TE sequences expressed by an entire population was also assessed. For this purpose, a binary code was generated to represent the expression status of TE subfamilies in each cell (where 1 is expressed and 0 is not expressed). For each population separately, the binary codes of individual cells were summed to obtain the frequency of expression of each TE subfamily in the population, which was used to compute the Shannon entropy of TE sequences expressed by the population. A linear model was generated with the lm function of the stats package in R to summarize the data distribution. The deviation (Δy) from the observed population's TE diversity and the one expected by the linear model was computed for each cell population.

TE expression in thymic APC

A differential expression analysis of TE subfamilies between the subsets of thymic APC was performed with the FindAllMarkers function with default parameters of Seurat (117) with the MAST model. Finally, the heterogeneity of TE expression inside thymic APC subsets was evaluated with the MetaCell package (version 0.3.5) (118). The composition of the metacells was validated based on manual annotation (see the "Single-cell RNA-seq preprocessing" section), and only metacells with >50% of cells belonging to the same subset of thymic APCs were kept. Differential expression of TE subfamilies between metacells was performed as described above, and the percentage of overlap between the sets of TEs overexpressed by the different metacells was computed.

Isolation of human thymic pDCs and immunostaining of dsRNAs

Primary human thymi were obtained from 4-month-old to 12-year-old children undergoing cardiovascular surgeries at the CHU Sainte-Justine. This project was approved by the CHU Sainte-Justine Research Ethics Board (protocol and biobank #2126). Thymi from 4-month-old to 12-year-old individuals were cryopreserved in liquid nitrogen in the following solution: 95% (PBS-5% Dextran 40 (Sigma-Aldrich)) – 5% DMSO (Fisher Scientific). Protocol for thymic pDCs isolation was based on *Stoeckle et al*. (119). Briefly, thymic samples were cut in ~2mm pieces, followed by three rounds of digestion (40 min, 180 RPM at 37□) in RPMI 1640 (Gibco) supplemented with 2mg/mL of Collagenase A (Roche) and 0.1mg/mL of DNase I (Sigma-Aldrich). APCs were then enriched using Percoll (Sigma-Aldrich) density centrifugation (3500g, 35min at 4□), followed by an FBS cushion density gradient (5mL of RPMI 1640 containing enriched APCs layered on 5mL of heat-inactivated FBS (Invitrogen, 12483020), 1000RPM for 10min at 4□) to remove cell debris. Finally, thymic pDCs were magnetically enriched using the QuadroMACS Separator (Miltenyi). Cells were stained with a CD303 (BDCA-2) MicroBead Kit (Miltenyi), and labeled cells were loaded on LS columns (Miltenyi) for magnetic-activated cell sorting.

Purified thymic pDCs were pipetted on poly-L-lysine (Sigma-Aldrich, 1:10 in dH_2O) coated 15µ-Slide 8 well (ibid) and incubated for 2h at 37□ in RPMI 1640 supplemented with 10% BSA (Sigma-Aldrich). Cells were fixed using 1% [w/v] paraformaldehyde (PFA, Sigma-Aldrich) in PBS 1X (Sigma-Aldrich) for 30min at room temperature. Cells were permeabilized for 30min at room temperature with 0.1% [v/v] Triton X-100 (Sigma-Aldrich) in PBS 1X, followed by blocking using 5% [w/v] BSA (Sigma-Aldrich) in PBS 1X for 30min at room temperature. Immunostaining was performed in four steps to avoid unspecific binding of the secondary antibodies: i) incubation overnight at 4□ with the mouse monoclonal IgG2a J2 antibody anti-dsRNA (Jena Bioscience, cat. RNT-SCI-10010500, dilution 1:200), ii) incubation with the donkey anti-mouse IgG (H+L) antibody coupled to Alexa Fluor 555 (Invitrogen,

cat. A-31570, dilution 1:500) for 30min at room temperature, iii) incubation with the mouse monoclonal IgG1 clone 6H6 anti-CD123 (eBioscience, cat. 14-1239-82, 1:100) for one hour at room temperature, and iv) incubation with the goat anti-mouse IgG1 polyclonal Alexa Fluor 488 antibody (Invitrogen, cat. A-21121, 1:1000) for 30min at room temperature. Finally, cells were stained with DAPI (Invitrogen, cat. D3571, 1:1000) for 5 minutes at room temperature. All antibodies and DAPI were diluted in a blocking solution. Image acquisition was made with an LSM 700 laser scanning confocal microscope (Zeiss) using a 40x oil objective (Zeiss, Plan-Neofluar N.A. 1.4) and the ZEN software. Using the whiteTopHat function of the EBImage package and the sigmoNormalize function of the MorphoR package in R, the background of the DAPI signal was removed. The nuclei were segmented on the resulting images as circular shapes based on the DAPI signal. The mean intensity of CD123 and J2 staining was determined for each cytoplasm, defined as 19nm rings around nuclei. Based on the distribution of the CD123 signal across cells, a threshold between $CD123^-$ and $CD123^+$ cells was set up for each replicate independently. J2 signal intensity was compared between $CD123^-$ and $CD123^+$ cells using the Wilcoxon Rank Sum test in R.

Gene set enrichment analysis

Gene set enrichment analyses were performed to determine which biological processes are enriched in mTEC(II) and pDCs. Differential gene expression analyses were performed between each possible pair of thymic APCs subsets using MAST with the FindMarkers function of Seurat. The gene set enrichment analysis was performed using the iDEA package (version 1.0.1) (120) in R. As per *Ma et al.* (120), the fold change and standard error of gene expression were used as input for iDEA, in addition to predefined lists of gene sets compiled in the iDEA package. Gene sets associated with antigen presentation, interferon signaling, and immune response were manually annotated. iDEA was launched with default parameters, except for the 500 iterations of the Markov chain Monte Carlo algorithm, and p-values were

corrected with the Louis method. We also visualized the expression of *AIRE*, *FEZF2*, and *CHD4* in the TEC lineage to validate their expression in mTEC(II).

TE loci regulated by AIRE, FEZF2, and CHD4

A differential expression analysis of TE subfamilies between WT and *Aire*-, *Fezf2*-, or *Chd4*-KO mice was performed with the voom method of the limma package (version 3.46.0) (121, 122). Stringent criteria (i.e., an expression below 2 transcripts per million (TPM) in all samples) were applied to remove lowly expressed TEs. TE subfamilies with i) a fold change ≥2 and an adjusted p-value ≤0.05 or ii) a fold change ≤-2 and an adjusted p-value ≤0.05 were considered as induced and repressed, respectively. The percentage of overlap between the sets of TE loci induced or repressed by AIRE, FEZF2, and CHD4 was computed. The class and subfamily were assigned to each regulated TE locus, and the distributions of classes and subfamilies across all TE sequences of the murine genome were used as controls. Significant enrichment of classes or subfamilies was determined with Chi-squared tests, and a Bonferroni correction for multiple comparisons was performed to enrich subfamilies in induced or repressed TEs. The distance between TE loci induced or repressed by AIRE, FEZF2, or CHD4 was defined as the minimal distance between the middle position of TE loci on the same chromosome. As a control, distributions of randomly selected TE loci whose expression is independent of AIRE, FEZF2, and CHD4 and equal size to the sets of regulated TEs were generated (for example, if 433 TE loci are induced by CHD4, 433 independent TE loci were randomly selected). Wilcoxon rank-sum tests were used to compare random and regulated distributions. Genomic positions of exons, introns 3’ and 5’ untranslated transcribed region (UTR) were downloaded from the UCSC Table Browser. The genomic localization of regulated TEs was determined using the intersect mode of the BEDTools suite version 2.29.2. TE loci not located in exons, introns, 3’UTR, or 5’UTR were considered intergenic. The percentage of regulated TE loci in each type of genomic region was determined and compared to the

genomic localization of all TE loci in the murine genome with chi-squared tests. Finally, we estimated the frequency of intron retention events for introns containing TE loci regulated by AIRE, FEZF2, or CHD4 with S-IRFindeR (123). Sequencing reads were aligned to the reference *Mus musculus* genome (mm10) using STAR version 2.7.1a (124) with default parameters. Each intron's Stable Intron Retention ratio (SIRratio) was computed with the computeSIRratio function of S-IRFindeR. Introns containing TE loci induced by AIRE, FEZF2, or CHD4 were filtered using BEDTools intersect. Random distributions of equivalent sizes of introns containing TE sequences independent of AIRE, FEZF2, and CHD4 were generated as control. A SIRratio of 0.1 was used as a threshold of significant intron retention events.

Enzymatic digestion and isolation of murine TECs

Thymic stromal cell enrichment was performed as previously described (71, 125). Briefly, thymi from 16- to 22-week-old K5D1 mice were mechanically disrupted and enzymatically digested with papain (Worthington Biochemical Corporation), DNase I (Sigma-Aldrich), and collagenase IV (Sigma-Aldrich) at 37°C. Next, the single-cell suspension obtained after enzymatic digestion was maintained at 4°C in FACS buffer (PBS, 0.5% [w/v] BSA, 2mM EDTA) and enriched in thymic epithelial cells using anti-EpCAM (CD326) or anti-CD45 microbeads (mouse, Miltenyi) and LS columns (Miltenyi). Then, the enriched epithelial cell suspension was stained for flow cytometry cell sorting with the following antibodies and dyes: anti-EpCAM-APC-Cy7 clone G8.8 (BioLegend, cat. 118218), anti-CD45-APC clone 30-F11 (BD Biosciences, cat. 559864), anti-UEA1-biotinylated (Vector Laboratories, cat. B-1065), anti-I-A/I-E-Alexa Fluor 700 clone M5/114.15.2 (BioLegend, cat. 107622), anti-Ly51-FITC clone 6C3 (BioLegend, cat. 553160), anti-streptavidin-PE-Cy7 (BD Biosciences, cat. 557598), and 7-AAD (BD Biosciences, cat. 559925). Cell sorting was performed using a BD FACSAria (BD Biosciences), and data were analyzed using the FACSDiva. TECs were defined as $EpCAM^{+}CD45^{-}$, while the cTEC and mTEC subsets were defined as $UEA1^{-}Ly51^{+}$ and $UEA1^{+}Ly51^{-}$ TEC, respectively.

RNA-Sequencing

Total RNA from 80 000 mTECs or cTECs was isolated using TRIzol and purified with an RNeasy micro kit (Qiagen). Total RNA was quantified using Qubit (Thermo Scientific), and RNA quality was assessed with the Agilent 2100 Bioanalyzer (Agilent Technologies). Transcriptome libraries were generated using a KAPA RNA HyperPrep kit (Roche) using a poly(A) selection (Thermo Scientific). Sequencing was performed on the Illumina NextSeq 500, obtaining ~200 million paired-end reads per sample.

Preparation of CNBR-activated Sepharose beads for MHC I immunoprecipitation

CNBR-activated Sepharose 4B beads (Sigma-Aldrich, cat. 17-0430-01) were incubated with 1 mM HCl at a ratio of 40 mg of beads per 13.5 ml of 1 mM HCl for 30 minutes with tumbling at room temperature. Beads were spun at 215g for 1 minute at 4°C, and supernatants were discarded. 40 mg of beads were resuspended with 4 ml of coupling buffer (0.1M $NaHC0_3$/0.5M NaCl pH 8.3), spun at 215g for 1 minute at 4°C, and the supernatants were discarded. Mouse antibodies Pan-H2 (clone M1/42), H2-K^b (clone Y-3), and H2-D^b (clone 28-14-8S) were coupled to beads at a ratio of 1 mg of antibody to 40 mg of beads in coupling buffer for 120 minutes with tumbling at room temperature. Beads were spun at 215g for 1 minute at 4°C, and supernatants were discarded. 40 mg of beads were resuspended with 1 ml of blocking buffer (0.2M glycine), incubated for 30 minutes with tumbling at room temperature, and the supernatants were discarded. Beads were washed by centrifugation twice with PBS pH 7.2, resuspended at a concentration of 1 mg of antibody per ml of PBS pH 7.2, and stored at 4°C.

Immuno-isolation of MAPs

Frozen pellets of mTECs (90 mice, 191 million cells total) and cTECs (121 mice, 164 million cells total) were thawed, pooled, and resuspended with PBS pH 7.2 up to 4 ml and then solubilized by adding 4 mL of detergent buffer containing PBS pH 7.2, 1% (w/v) CHAPS (Sigma, cat. C9426-5G) supplemented

with Protease inhibitor cocktail (Sigma, cat. P8340-5mL). Solubilized cells were incubated for 60 minutes with tumbling at 4°C and then spun at 16,600g for 20 minutes at 4°C. Supernatants were transferred into new tubes containing 1.5 mg of Pan-H2, 0.5 mg of H2-K^b, and 0.5 mg of H2-D^b antibodies covalently-cross-linked CNBR-Sepharose beads per sample and incubated with tumbling for 180 minutes at 4°C. Samples were transferred into BioRad Poly prep chromatography columns and eluted by gravity. Beads were first washed with 11.5 mL PBS, then with 11.5 mL of 0.1X PBS, and finally with 11.5 mL of water. MHC I complexes were eluted from the beads by acidic treatment using 1% trifluoroacetic acid (TFA). Acidic filtrates containing peptides were separated from MHC I subunits (HLA molecules and β-2 microglobulin) using home-made stage tips packed with two 1 mm diameter octadecyl (C-18) solid phase extraction disks (EMPORE). Stage tips were pre-washed with methanol, then with 80% acetonitrile (ACN) in 0.1% TFA, and finally with 1% TFA. Samples were loaded onto the stage tips and washed with 1% TFA and 0.1% TFA. Peptides were eluted with 30% ACN in 0.1% TFA, dried using vacuum centrifugation, and then stored at -20 °C until MS analysis.

MS analyses

Peptides were loaded and separated on a home-made reversed-phase column (150-µm i.d. by 200 mm) with a 106-min gradient from 10 to 38% B (A: formic acid 0.1%, B: 80% CAN 0.1% formic acid) and a 600-nl/min flow rate on an Easy nLC-1200 connected to an Orbitrap Exploris 480 (Thermo Fisher Scientific). Each full MS spectrum acquired at a resolution of 240,000 was followed by tandem-MS (MS-MS) spectra acquisition on the most abundant multiply charged precursor ions for a maximum of 3s. Tandem-MS experiments were performed using higher energy collision-induced dissociation (HCD) at a collision energy of 34%. The generation of the personalized proteome containing TE sequences, as well as the identification of TE-derived MAPs, was performed as per *Larouche et al.* (34) with the following modifications: the mm10 murine reference genome was downloaded from the UCSC Genome

Browser, the annotations for murine genes and TE sequences were downloaded from the UCSC Table Browser, and the Uniprot mouse database (16 977 entries) was used for the canonical proteome. MAPs were identified using PEAKS X Pro (Bioinformatics Solutions, Waterloo, ON). The level and frequency of expression of TE subfamilies generating MAPs or not were determined in thymic epithelial cells were determined by averaging the expression values across cells of a TEC subset and dividing the number of cells with a positive (i.e.,> 0) expression of the TEs by the total number of cells of the TEC subset, respectively.

Availability of data and materials:

scRNA-seq data of human thymi and spleen were downloaded from ArrayExpress (accession number E-MTAB-8581) and the NCBI BIOPROJECT (accession code PRJEB31843), respectively. scRNA-seq data of murine thymi were downloaded from ArrayExpress (accession number E-MTAB-8560). RNA-seq data from WT, Aire-KO, Fezf2-KO, and Chd4-KO murine mTECs were downloaded from the Gene Expression Omnibus (GEO) under the accession code GSE144880. ChIP-seq data of ETS1 in human NK cells and AIRE in murine mTECs were downloaded from GEO (accession codes GSE124104 and GSE92654, respectively). ChIP-seq data for different histone marks in murine mTECs were also downloaded from GEO: H3K4me3 for mTECs (GSE53111); H3K4me1 and H3K27ac from MHCII^{hi} mTECs (GSE92597); H3K4me2 in mTEC-II (GSE103969); and H3K4ac and H3K9ac in mTECs (GSE114713). Transcriptomic and immunopeptidomic data of K5D1 mice mTECs and cTECs generated in this study are available on the Gene Expression Omnibus (GEO) under the accession GSE232011 and on the Proteomics Identification Database (PRIDE) under the accession PXD042241, respectively.

Author contributions:

Conceptualization: JDL, CML, AT, KV, CP

Methodology: JDL, CML, AT, KV

Investigation: JDL, LH, CC, SB, JH, EB, JL, CD, PG, JPL

Visualization: JDL, CML, AT

Funding acquisition: JDL, CP

Project administration: JDL, KV, CP

Supervision: SL, PT, CP

Writing – original draft: JDL, KV, CP

Writing – review & editing: JDL, CML, AT, KV, LH, JH, SB, CC, EB, JL, CD, PG, ERR, BHN, SL, PT, CP

Acknowledgments:

The authors thank Christian Charbonneau and Raphaëlle Lambert from IRIC's bio-imaging and genomics platforms, respectively. We also thank Allan Sauvat for the help with the microscopy quantification. We thank Mathilde Soulez, Bernhard Lehnertz, Biljana Culjkovic, Brian Wilhelm, and Michaël Imbeault for insightful discussions. We are indebted to Kathie Béland and Elie Haddad from the CHU Sainte-Justine Research Center for providing the primary thymic samples.

References

1. Boehm T. Evolution of vertebrate immunity. Curr Biol. 2012;22(17):R722-32.
2. Boehm T, Swann JB. Origin and evolution of adaptive immunity. Annu Rev Anim Biosci. 2014;2:259-83.
3. Suo C, Dann E, Goh I, Jardine L, Kleshchevnikov V, Park JE, et al. Mapping the developing human immune system across organs. Science. 2022;376(6597):eabo0510.
4. Terra R, Louis I, Le Blanc R, Ouellet S, Zuniga-Pflucker JC, Perreault C. T-cell generation by lymph node resident progenitor cells. Blood. 2005;106(1):193-200.
5. Blais ME, Brochu S, Giroux M, Belanger MP, Dulude G, Sekaly RP, et al. Why T cells of thymic versus extrathymic origin are functionally different. J Immunol. 2008;180(4):2299-312.
6. Zuniga-Pflucker JC, Longo DL, Kruisbeek AM. Positive selection of CD4-CD8+ T cells in the thymus of normal mice. Nature. 1989;338(6210):76-8.

7. Breed ER, Lee ST, Hogquist KA. Directing T cell fate: How thymic antigen presenting cells coordinate thymocyte selection. Semin Cell Dev Biol. 2018;84:2-10.
8. Dervovic D, Zuniga-Pflucker JC. Positive selection of T cells, an in vitro view. Semin Immunol. 2010;22(5):276-86.
9. Lebel ME, Coutelier M, Galipeau M, Kleinman CL, Moon JJ, Melichar HJ. Differential expression of tissue-restricted antigens among mTEC is associated with distinct autoreactive T cell fates. Nat Commun. 2020;11(1):3734.
10. Srinivasan J, Lancaster JN, Singarapu N, Hale LP, Ehrlich LIR, Richie ER. Age-Related Changes in Thymic Central Tolerance. Front Immunol. 2021;12:676236.
11. Cheng M, Anderson MS. Thymic tolerance as a key brake on autoimmunity. Nat Immunol. 2018;19(7):659-64.
12. Perera J, Zheng Z, Li S, Gudjonson H, Kalinina O, Benichou JIC, et al. Self-Antigen-Driven Thymic B Cell Class Switching Promotes T Cell Central Tolerance. Cell Rep. 2016;17(2):387-98.
13. Nitta T, Ohigashi I, Nakagawa Y, Takahama Y. Cytokine crosstalk for thymic medulla formation. Curr Opin Immunol. 2011;23(2):190-7.
14. Malhotra D, Linehan JL, Dileepan T, Lee YJ, Purtha WE, Lu JV, et al. Tolerance is established in polyclonal CD4(+) T cells by distinct mechanisms, according to self-peptide expression patterns. Nat Immunol. 2016;17(2):187-95.
15. Lkhagvasuren E, Sakata M, Ohigashi I, Takahama Y. Lymphotoxin beta receptor regulates the development of CCL21-expressing subset of postnatal medullary thymic epithelial cells. J Immunol. 2013;190(10):5110-7.
16. Laan M, Salumets A, Klein A, Reintamm K, Bichele R, Peterson H, et al. Post-Aire Medullary Thymic Epithelial Cells and Hassall's Corpuscles as Inducers of Tonic Pro-Inflammatory Microenvironment. Front Immunol. 2021;12:635569.
17. Michelson DA, Hase K, Kaisho T, Benoist C, Mathis D. Thymic epithelial cells co-opt lineage-defining transcription factors to eliminate autoreactive T cells. Cell. 2022.
18. Ramsey C, Winqvist O, Puhakka L, Halonen M, Moro A, Kampe O, et al. Aire deficient mice develop multiple features of APECED phenotype and show altered immune response. Hum Mol Genet. 2002;11(4):397-409.
19. Takaba H, Morishita Y, Tomofuji Y, Danks L, Nitta T, Komatsu N, et al. Fezf2 Orchestrates a Thymic Program of Self-Antigen Expression for Immune Tolerance. Cell. 2015;163(4):975-87.
20. Tomofuji Y, Takaba H, Suzuki HI, Benlaribi R, Martinez CDP, Abe Y, et al. Chd4 choreographs self-antigen expression for central immune tolerance. Nat Immunol. 2020;21(8):892-901.
21. Ginhoux F, Guilliams M, Merad M. Expanding dendritic cell nomenclature in the single-cell era. Nat Rev Immunol. 2022;22(2):67-8.
22. Colantonio AD, Epeldegui M, Jesiak M, Jachimowski L, Blom B, Uittenbogaart CH. IFN-alpha is constitutively expressed in the human thymus, but not in peripheral lymphoid organs. PLoS One. 2011;6(8):e24252.
23. Lavaert M, Liang KL, Vandamme N, Park JE, Roels J, Kowalczyk MS, et al. Integrated scRNA-Seq Identifies Human Postnatal Thymus Seeding Progenitors and Regulatory Dynamics of Differentiating Immature Thymocytes. Immunity. 2020;52(6):1088-104 e6.
24. Le J, Park JE, Ha VL, Luong A, Branciamore S, Rodin AS, et al. Single-Cell RNA-Seq Mapping of Human Thymopoiesis Reveals Lineage Specification Trajectories and a Commitment Spectrum in T Cell Development. Immunity. 2020;52(6):1105-18 e9.
25. Xing Y, Wang X, Jameson SC, Hogquist KA. Late stages of T cell maturation in the thymus involve NF-kappaB and tonic type I interferon signaling. Nat Immunol. 2016;17(5):565-73.

26. Hanabuchi S, Ito T, Park WR, Watanabe N, Shaw JL, Roman E, et al. Thymic stromal lymphopoietin-activated plasmacytoid dendritic cells induce the generation of FOXP3+ regulatory T cells in human thymus. J Immunol. 2010;184(6):2999-3007.
27. Martin-Gayo E, Sierra-Filardi E, Corbi AL, Toribio ML. Plasmacytoid dendritic cells resident in human thymus drive natural Treg cell development. Blood. 2010;115(26):5366-75.
28. Martinet V, Tonon S, Torres D, Azouz A, Nguyen M, Kohler A, et al. Type I interferons regulate eomesodermin expression and the development of unconventional memory CD8(+) T cells. Nat Commun. 2015;6:7089.
29. Epeldegui M, Blom B, Uittenbogaart CH. BST2/Tetherin is constitutively expressed on human thymocytes with the phenotype and function of Treg cells. Eur J Immunol. 2015;45(3):728-37.
30. Treangen TJ, Salzberg SL. Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat Rev Genet. 2011;13(1):36-46.
31. Deniz O, Frost JM, Branco MR. Regulation of transposable elements by DNA modifications. Nat Rev Genet. 2019;20(7):417-31.
32. Bourque G, Burns KH, Gehring M, Gorbunova V, Seluanov A, Hammell M, et al. Ten things you should know about transposable elements. Genome Biol. 2018;19(1):199.
33. Argueso JL, Westmoreland J, Mieczkowski PA, Gawel M, Petes TD, Resnick MA. Double-strand breaks associated with repetitive DNA can reshape the genome. Proc Natl Acad Sci U S A. 2008;105(33):11845-50.
34. Larouche JD, Trofimov A, Hesnard L, Ehx G, Zhao Q, Vincent K, et al. Widespread and tissue-specific expression of endogenous retroelements in human somatic tissues. Genome Med. 2020;12(1):40.
35. Carter JA, Stromich L, Peacey M, Chapin SR, Velten L, Steinmetz LM, et al. Transcriptomic diversity in human medullary thymic epithelial cells. Nat Commun. 2022;13(1):4296.
36. Kassiotis G. The Immunological Conundrum of Endogenous Retroelements. Annu Rev Immunol. 2023;41:99-125.
37. Chuong EB, Elde NC, Feschotte C. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. Science. 2016;351(6277):1083-7.
38. Bogdan L, Barreiro L, Bourque G. Transposable elements have contributed human regulatory regions that are activated upon bacterial infection. Philos Trans R Soc Lond B Biol Sci. 2020;375(1795):20190332.
39. Adoue V, Binet B, Malbec A, Fourquet J, Romagnoli P, van Meerwijk JPM, et al. The Histone Methyltransferase SETDB1 Controls T Helper Cell Lineage Integrity by Repressing Endogenous Retroviruses. Immunity. 2019;50(3):629-44 e8.
40. Lefkopoulos S, Polyzou A, Derecka M, Bergo V, Clapes T, Cauchy P, et al. Repetitive Elements Trigger RIG-I-like Receptor Signaling that Regulates the Emergence of Hematopoietic Stem and Progenitor Cells. Immunity. 2020;53(5):934-51 e9.
41. Lima-Junior DS, Krishnamurthy SR, Bouladoux N, Collins N, Han SJ, Chen EY, et al. Endogenous retroviruses promote homeostatic and inflammatory responses to the microbiota. Cell. 2021.
42. Park JE, Botting RA, Dominguez Conde C, Popescu DM, Lavaert M, Kunz DJ, et al. A cell atlas of human thymic development defines T cell repertoire formation. Science. 2020;367(6480).
43. Brocks D, Chomsky E, Mukamel Z, Lifshitz A, Tanay A. Single cell analysis reveals dynamics of transposable element transcription following epigenetic de-repression. bioRxiv. 2018:462853.
44. Imbeault M, Helleboid PY, Trono D. KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. Nature. 2017;543(7646):550-4.

45. Kunarso G, Chia NY, Jeyakani J, Hwang C, Lu X, Chan YS, et al. Transposable elements have rewired the core regulatory network of human embryonic stem cells. Nat Genet. 2010;42(7):631-4.
46. Sundaram V, Cheng Y, Ma Z, Li D, Xing X, Edge P, et al. Widespread contribution of transposable elements to the innovation of gene regulatory networks. Genome Res. 2014;24(12):1963-76.
47. Hosokawa H, Rothenberg EV. How transcription factors drive choice of the T cell fate. Nat Rev Immunol. 2021;21(3):162-76.
48. Baik S, Sekai M, Hamazaki Y, Jenkinson WE, Anderson G. Relb acts downstream of medullary thymic epithelial stem cells and is essential for the emergence of RANK(+) medullary epithelial progenitors. Eur J Immunol. 2016;46(4):857-62.
49. Akiyama T, Shimo Y, Yanai H, Qin J, Ohshima D, Maruyama Y, et al. The tumor necrosis factor family receptors RANK and CD40 cooperatively establish the thymic medullary microenvironment and self-tolerance. Immunity. 2008;29(3):423-37.
50. Yamazaki Y, Urrutia R, Franco LM, Giliani S, Zhang K, Alazami AM, et al. PAX1 is essential for development and function of the human thymus. Sci Immunol. 2020;5(44).
51. Meixner A, Karreth F, Kenner L, Wagner EF. JunD regulates lymphocyte proliferation and T helper cell cytokine expression. EMBO J. 2004;23(6):1325-35.
52. Cisse B, Caton ML, Lehner M, Maeda T, Scheu S, Locksley R, et al. Transcription factor E2-2 is an essential and specific regulator of plasmacytoid dendritic cell development. Cell. 2008;135(1):37-48.
53. Choudhary MN, Friedman RZ, Wang JT, Jang HS, Zhuo X, Wang T. Co-opted transposons help perpetuate conserved higher-order chromosomal structures. Genome Biol. 2020;21(1):16.
54. Consortium EP. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489(7414):57-74.
55. Luo Y, Hitz BC, Gabdank I, Hilton JA, Kagda MS, Lam B, et al. New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. Nucleic Acids Res. 2020;48(D1):D882-D9.
56. Taveirne S, Wahlen S, Van Loocke W, Kiekens L, Persyn E, Van Ammel E, et al. The transcription factor ETS1 is an important regulator of human NK cell development and terminal differentiation. Blood. 2020;136(3):288-98.
57. Kim N, Kim M, Yun S, Doh J, Greenberg PD, Kim TD, et al. MicroRNA-150 regulates the cytotoxicity of natural killers by targeting perforin-1. J Allergy Clin Immunol. 2014;134(1):195-203.
58. Gunturi A, Berg RE, Forman J. The role of CD94/NKG2 in innate and adaptive immunity. Immunol Res. 2004;30(1):29-34.
59. Bornstein C, Nevo S, Giladi A, Kadouri N, Pouzolles M, Gerbe F, et al. Single-cell mapping of the thymic stroma identifies IL-25-producing tuft epithelial cells. Nature. 2018;559(7715):622-6.
60. Bautista JL, Cramer NT, Miller CN, Chavez J, Berrios DI, Byrnes LE, et al. Single-cell transcriptional profiling of human thymic stroma uncovers novel cellular heterogeneity in the thymic medulla. Nat Commun. 2021;12(1):1096.
61. Madissoon E, Wilbrey-Clark A, Miragaia RJ, Saeb-Parsy K, Mahbubani KT, Georgakopoulos N, et al. scRNA-seq assessment of the human lung, spleen, and esophagus tissue stability after cold preservation. Genome Biol. 2019;21(1):1.
62. Sadeq S, Al-Hashimi S, Cusack CM, Werner A. Endogenous Double-Stranded RNA. Noncoding RNA. 2021;7(1).
63. Sansom SN, Shikama-Dorn N, Zhanybekova S, Nusspaumer G, Macaulay IC, Deadman ME, et al. Population and single-cell genomics reveal the Aire dependency, relief from Polycomb silencing, and distribution of self-antigen expression in thymic epithelia. Genome Res. 2014;24(12):1918-31.

64. St-Pierre C, Morgand E, Benhammadi M, Rouette A, Hardy MP, Gaboury L, et al. Immunoproteasomes Control the Homeostasis of Medullary Thymic Epithelial Cells by Alleviating Proteotoxic Stress. Cell Rep. 2017;21(9):2558-70.
65. Shapiro IE, Bassani-Sternberg M. The impact of immunopeptidomics: From basic research to clinical implementation. Semin Immunol. 2023;66:101727.
66. Vizcaino JA, Kubiniok P, Kovalchik KA, Ma Q, Duquette JD, Mongrain I, et al. The Human Immunopeptidome Project: A Roadmap to Predict and Treat Immune Diseases. Mol Cell Proteomics. 2020;19(1):31-49.
67. Kubiniok P, Marcu A, Bichmann L, Kuchenbecker L, Schuster H, Hamelin DJ, et al. Understanding the constitutive presentation of MHC class I immunopeptidomes in primary tissues. iScience. 2022;25(2):103768.
68. Robles AI, Larcher F, Whalin RB, Murillas R, Richie E, Gimenez-Conti IB, et al. Expression of cyclin D1 in epithelial tissues of transgenic mice results in epidermal hyperproliferation and severe thymic hyperplasia. Proc Natl Acad Sci U S A. 1996;93(15):7634-8.
69. Klug DB, Crouch E, Carter C, Coghlan L, Conti CJ, Richie ER. Transgenic expression of cyclin D1 in thymic epithelial precursors promotes epithelial and T cell development. J Immunol. 2000;164(4):1881-8.
70. Ohigashi I, Tanaka Y, Kondo K, Fujimori S, Kondo H, Palin AC, et al. Trans-omics Impact of Thymoproteasome in Cortical Thymic Epithelial Cells. Cell Rep. 2019;29(9):2901-16 e6.
71. Dumont-Lagace M, Daouda T, Depoers L, Zumer J, Benslimane Y, Brochu S, et al. Qualitative Changes in Cortical Thymic Epithelial Cells Drive Postpartum Thymic Regeneration. Front Immunol. 2019;10:3118.
72. Ghosh M, Gauger M, Marcu A, Nelde A, Denk M, Schuster H, et al. Guidance Document: Validation of a High-Performance Liquid Chromatography-Tandem Mass Spectrometry Immunopeptidomics Assay for the Identification of HLA Class I Ligands Suitable for Pharmaceutical Therapies. Mol Cell Proteomics. 2020;19(3):432-43.
73. Nanaware PP, Jurewicz MM, Clement CC, Lu L, Santambrogio L, Stern LJ. Distinguishing Signal From Noise in Immunopeptidome Studies of Limiting-Abundance Biological Samples: Peptides Presented by I-A(b) in C57BL/6 Mouse Thymus. Front Immunol. 2021;12:658601.
74. Baran-Gale J, Morgan MD, Maio S, Dhalla F, Calvo-Asensio I, Deadman ME, et al. Ageing compromises mouse thymus function and remodels epithelial cell differentiation. Elife. 2020;9.
75. Frank JA, Feschotte C. Co-option of endogenous viral sequences for host cell function. Curr Opin Virol. 2017;25:81-9.
76. Groger V, Cynis H. Human Endogenous Retroviruses and Their Putative Role in the Development of Autoimmune Disorders Such as Multiple Sclerosis. Front Microbiol. 2018;9:265.
77. Volkman HE, Stetson DB. The enemy within: endogenous retroelements and autoimmune disease. Nat Immunol. 2014;15(5):415-22.
78. De Cecco M, Ito T, Petrashen AP, Elias AE, Skvir NJ, Criscione SW, et al. L1 drives IFN in senescent cells and promotes age-associated inflammation. Nature. 2019;566(7742):73-8.
79. Huntley S, Baggott DM, Hamilton AT, Tran-Gyamfi M, Yang S, Kim J, et al. A comprehensive catalog of human KRAB-associated zinc finger genes: insights into the evolutionary history of a large family of transcriptional repressors. Genome Res. 2006;16(5):669-77.
80. Mouri Y, Nishijima H, Kawano H, Hirota F, Sakaguchi N, Morimoto J, et al. NF-kappaB-inducing kinase in thymic stroma establishes central tolerance by orchestrating cross-talk with not only thymocytes but also dendritic cells. J Immunol. 2014;193(9):4356-67.

81. O'Sullivan BJ, Yekollu S, Ruscher R, Mehdi AM, Maradana MR, Chidgey AP, et al. Autoimmune-Mediated Thymic Atrophy Is Accelerated but Reversible in RelB-Deficient Mice. Front Immunol. 2018;9:1092.
82. Klein L, Kyewski B, Allen PM, Hogquist KA. Positive and negative selection of the T cell repertoire: what thymocytes see (and don't see). Nat Rev Immunol. 2014;14(6):377-91.
83. Caron E, Vincent K, Fortier MH, Laverdure JP, Bramoulle A, Hardy MP, et al. The MHC I immunopeptidome conveys to the cell surface an integrative view of cellular regulation. Mol Syst Biol. 2011;7:533.
84. Admon A. The biogenesis of the immunopeptidome. Semin Immunol. 2023;67:101766.
85. Laumont CM, Vincent K, Hesnard L, Audemard E, Bonneil E, Laverdure JP, et al. Noncoding regions are the main source of targetable tumor-specific antigens. Sci Transl Med. 2018;10(470).
86. Burbage M, Rocanin-Arjo A, Baudon B, Arribas YA, Merlotti A, Rookhuizen DC, et al. Epigenetically controlled tumor antigens derived from splice junctions between exons and transposable elements. Sci Immunol. 2023;8(80):eabm6360.
87. Shah NM, Jang HJ, Liang Y, Maeng JH, Tzeng SC, Wu A, et al. Pan-cancer analysis identifies tumor-specific antigens derived from transposable elements. Nat Genet. 2023;55(4):631-9.
88. Shao W, Pedrioli PGA, Wolski W, Scurtescu C, Schmid E, Vizcaino JA, et al. The SysteMHC Atlas project. Nucleic Acids Res. 2018;46(D1):D1237-D47.
89. Weijer K, Uittenbogaart CH, Voordouw A, Couwenberg F, Seppen J, Blom B, et al. Intrathymic and extrathymic development of human plasmacytoid dendritic cell precursors in vivo. Blood. 2002;99(8):2752-9.
90. Boehm T, Morimoto R, Trancoso I, Aleksandrova N. Genetic conflicts and the origin of self/nonself-discrimination in the vertebrate immune system. Trends Immunol. 2023;44(5):372-83.
91. Harroud A, Hafler DA. Common genetic factors among autoimmune diseases. Science. 2023;380(6644):485-90.
92. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. 2016;34(5):525-7.
93. Smit A, Hubley, R & Green, P. RepeatMasker Open-4.0. 2013-2015.
94. Melsted P, Booeshaghi AS, Gao F, Beltrame E, Lu L, Hjorleifsson KE, et al. Modular and efficient preprocessing of single-cell RNA-seq. bioRxiv. 2019:673285.
95. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9(4):357-9.
96. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. Gigascience. 2021;10(2).
97. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 2008;9(9):R137.
98. Amemiya HM, Kundaje A, Boyle AP. The ENCODE Blacklist: Identification of Problematic Regions of the Genome. Sci Rep. 2019;9(1):9354.
99. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6):841-2.
100. Ramirez F, Ryan DP, Gruning B, Bhardwaj V, Kilpert F, Richter AS, et al. deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Res. 2016;44(W1):W160-5.
101. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The human genome browser at UCSC. Genome Res. 2002;12(6):996-1006.
102. Shen L, Shao N, Liu X, Nestler E. ngs.plot: Quick mining and visualization of next-generation sequencing data by integrating genomic databases. BMC Genomics. 2014;15:284.

103. Amezquita RA, Lun ATL, Becht E, Carey VJ, Carpp LN, Geistlinger L, et al. Orchestrating single-cell analysis with Bioconductor. Nat Methods. 2020;17(2):137-45.
104. McCarthy DJ, Campbell KR, Lun AT, Wills QF. Scater: preprocessing, quality control, normalization and visualization of single-cell RNA-seq data in R. Bioinformatics. 2017;33(8):1179-86.
105. Lun AT, McCarthy DJ, Marioni JC. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. F1000Res. 2016;5:2122.
106. Haghverdi L, Lun ATL, Morgan MD, Marioni JC. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. Nat Biotechnol. 2018;36(5):421-7.
107. Gu Z, Gu L, Eils R, Schlesner M, Brors B. circlize Implements and enhances circular visualization in R. Bioinformatics. 2014;30(19):2811-2.
108. Cowan JE, Malin J, Zhao Y, Seedhom MO, Harly C, Ohigashi I, et al. Myc controls a distinct transcriptional program in fetal thymic epithelial cells that determines thymus growth. Nat Commun. 2019;10(1):5498.
109. Kowalczyk MS, Tirosh I, Heckl D, Rao TN, Dixit A, Haas BJ, et al. Single-cell RNA-seq reveals changes in cell cycle and differentiation programs upon aging of hematopoietic stem cells. Genome Res. 2015;25(12):1860-72.
110. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409(6822):860-921.
111. Mouse Genome Sequencing C, Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, et al. Initial sequencing and comparative analysis of the mouse genome. Nature. 2002;420(6915):520-62.
112. Lambert SA, Jolma A, Campitelli LF, Das PK, Yin Y, Albu M, et al. The Human Transcription Factors. Cell. 2018;172(4):650-65.
113. Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. Bioinformatics. 2011;27(7):1017-8.
114. Butts CT. network: A Package for Managing Relational Data in R. Journal of Statistical Software. 2008;24(2):1 - 36.
115. Abugessaisa I, Noguchi S, Hasegawa A, Kondo A, Kawaji H, Carninci P, et al. refTSS: A Reference Data Set for Human and Mouse Transcription Start Sites. J Mol Biol. 2019;431(13):2407-22.
116. Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D, et al. vegan: Community Ecology Package. 2020;https://CRAN.R-project.org/package=vegan.
117. Satija R, Farrell JA, Gennert D, Schier AF, Regev A. Spatial reconstruction of single-cell gene expression data. Nat Biotechnol. 2015;33(5):495-502.
118. Baran Y, Bercovich A, Sebe-Pedros A, Lubling Y, Giladi A, Chomsky E, et al. MetaCell: analysis of single-cell RNA-seq data using K-nn graph partitions. Genome Biol. 2019;20(1):206.
119. Stoeckle C, Rota IA, Tolosa E, Haller C, Melms A, Adamopoulou E. Isolation of myeloid dendritic cells and epithelial cells from human thymus. J Vis Exp. 2013(79):e50951.
120. Ma Y, Sun S, Shang X, Keller ET, Chen M, Zhou X. Integrative differential expression and gene set enrichment analysis using summary statistics for scRNA-seq studies. Nat Commun. 2020;11(1):1585.
121. Law CW, Chen Y, Shi W, Smyth GK. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol. 2014;15(2):R29.
122. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43(7):e47.
123. Broseus L, Ritchie W. S-IRFindeR: stable and accurate measurement of intron retention. bioRxiv. 2020:2020.06.25.164699.

124. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15-21.
125. Kim MJ, Miller CM, Shadrach JL, Wagers AJ, Serwold T. Young, proliferative thymic epithelial cells engraft and function in aging thymuses. J Immunol. 2015;194(10):4784-95.

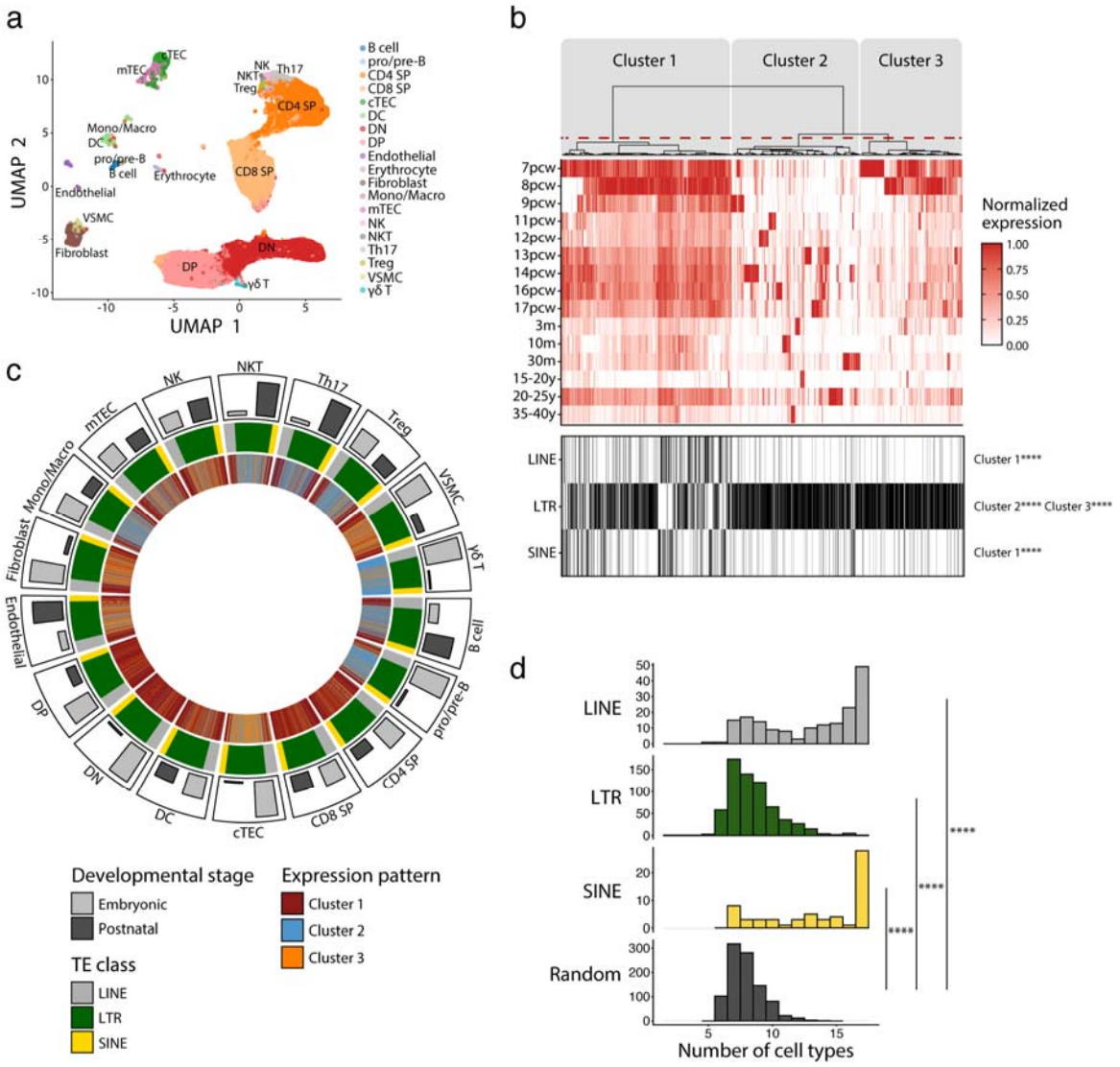


Figure 1. LINEs, SINEs, and LTRs exhibit distinct expression profiles in human thymic cell populations. (a) UMAP depicting the cell populations present in human thymi (CD4 SP, CD4 single positive thymocytes; CD8 SP, CD8 single positive thymocytes; cTEC, cortical thymic epithelial cells; DC, dendritic cells; DN, double negative thymocytes; DP, double positive thymocytes; Mono/Macro,

monocytes and macrophages; mTEC, medullary thymic epithelial cells; NK, natural killer cells; NKT, natural killer T cells; pro/pre-B, pro-B and pre-B cells; Th17, T helper 17 cells; Treg, regulatory T cells; VSMC, vascular smooth muscle cell). Cells were clustered in 19 populations based on the expression of marker genes from *Park et al.* (40). **(b)** *Upper panel:* Heatmap of TE expression during thymic development, with each column representing the expression of one TE subfamily in one cell type. Unsupervised hierarchical clustering was performed, and the dendrogram was manually cut into 3 clusters (red dashed line). *Lower panel:* The class of TE subfamilies and significant enrichments in the 3 clusters (Fisher's exact tests; ****$p \leq 0.0001$). (pcw, post-conception week; m, month; y, year). **(c)** Circos plot showing the expression pattern of TE subfamilies across thymic cells. From outermost to innermost tracks: i) proportion of cells in embryonic and postnatal samples, ii) class of TE subfamilies, iii) expression pattern of TE subfamilies identified in (b). TE subfamilies are in the same order for all cell types. **(d)** Histograms showing the number of cell types sharing the same expression pattern for a given TE subfamily. LINE (n=171), LTR (n=577), and SINE (n=60) were compared to a randomly generated distribution (n=809) (Kolmogorov-Smirnov tests, ****$p \leq 0.0001$).

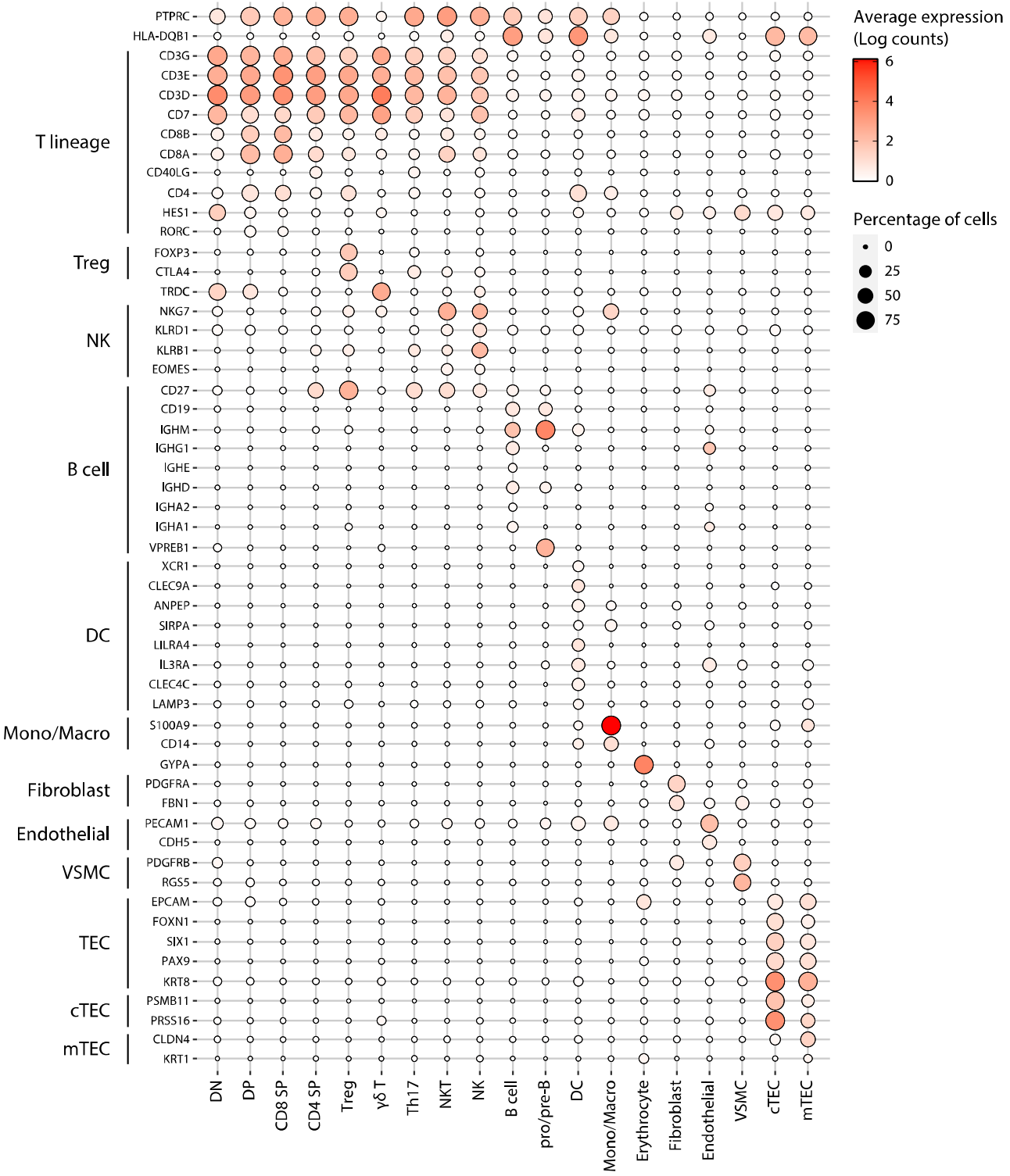


Figure 1 – figure supplement 1. Annotation of human thymic cell populations.

Dot plot depicting the expression of marker genes in the annotated cell types of the thymus. The average expression and percentage of cells expressing the gene are represented by the color and size of the dot, respectively (DN, double negative thymocytes; DP, double positive thymocytes; CD8 SP, CD8 single positive thymocytes; CD4 SP, CD4 single positive thymocytes, Treg, regulatory T cells; NKT, natural killer T cells; NK, natural killer cells; DC, dendritic cells; Mono/Macro, monocytes and macrophages; VSMC, vascular smooth muscle cells; cTEC, cortical thymic epithelial cells; mTEC, medullary thymic epithelial cells).

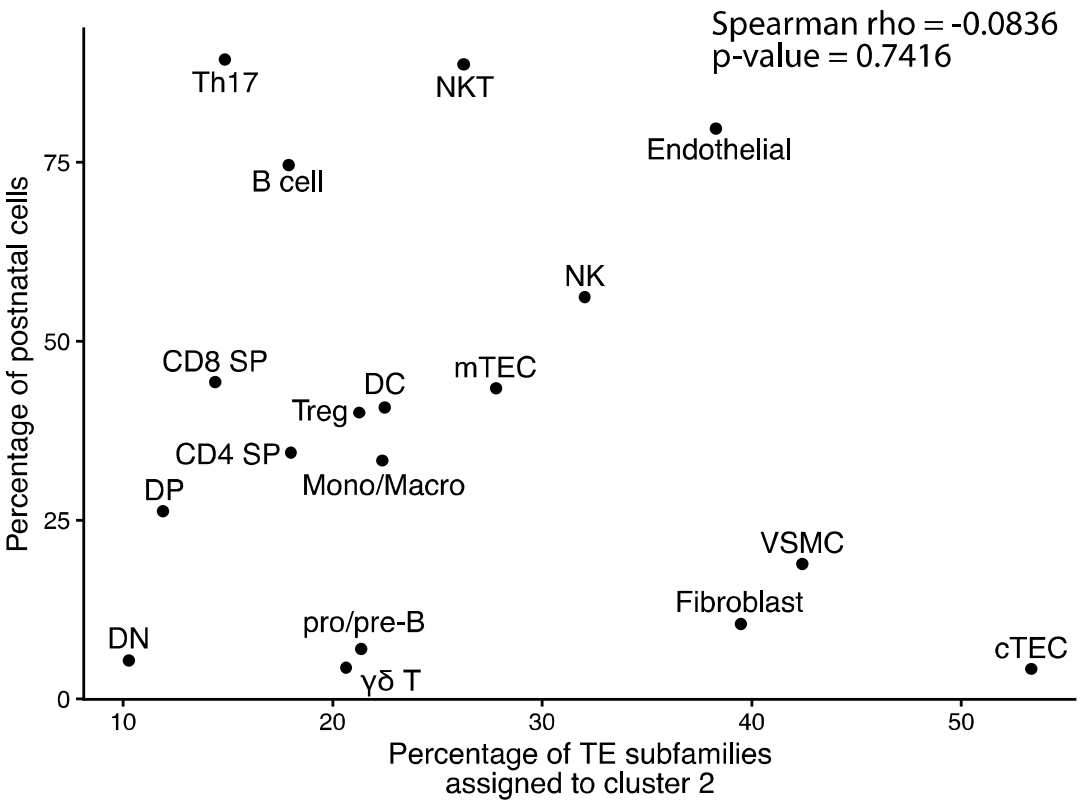


Figure 1 – figure supplement 2. Assignment to cluster 2 is independent of the developmental stage of cells.

The graph depicts the correlation between the proportion of cells of a population originating from a postnatal sample and the proportion of TE subfamilies assigned to cluster 2 by the hierarchical clustering in Figure 1B.

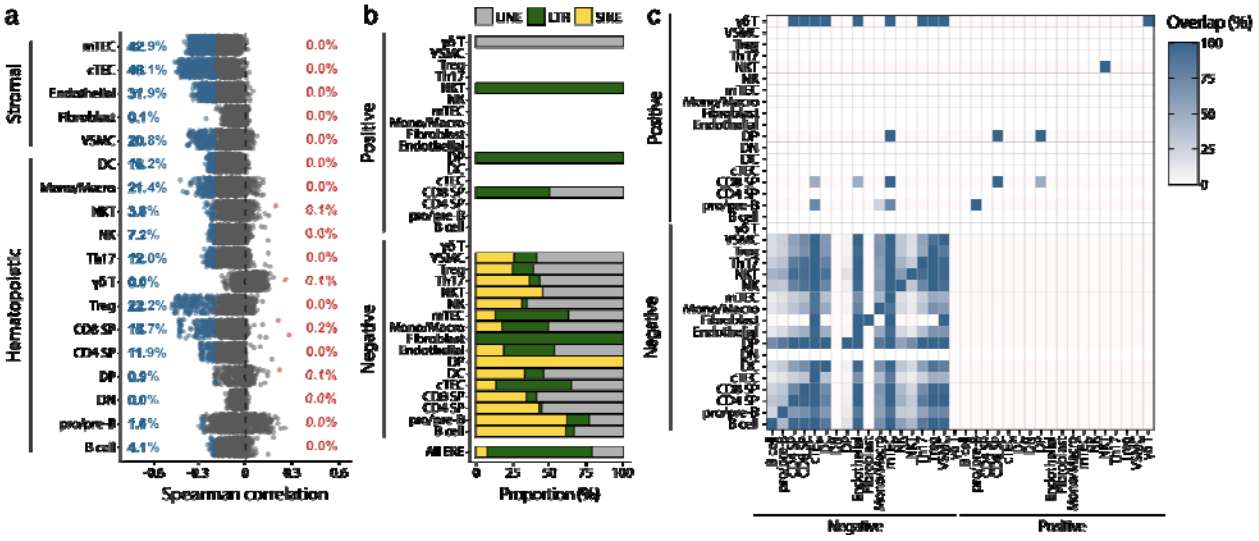


Figure 1 – figure supplement 3. TE expression is negatively correlated with cell proliferation. **(a)** Spearman correlation between the expression of TE subfamilies and cell cycle scores. Positively ($r \geq 0.2$ and adj. $p \leq 0.01$) and negatively ($r \leq -0.2$ and adj. $p \leq 0.01$) correlated subfamilies are red and blue, respectively. P-values were corrected for multiple comparisons with the Benjamini-Hochberg method). **(b)** Proportion of subfamilies positively or negatively correlated with cell proliferation belonging to each TE class. **(c)** Percentage of overlap of TE subfamilies positively or negatively correlated with cell proliferation between cell types.

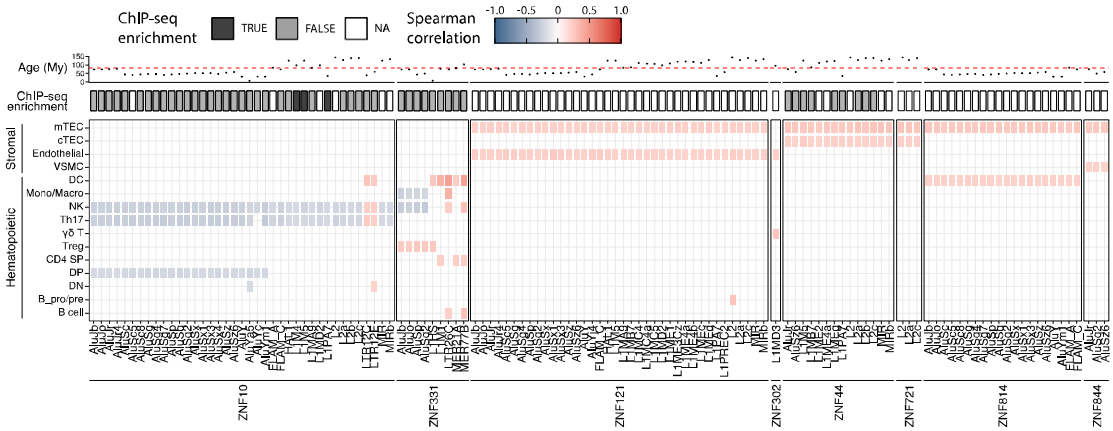
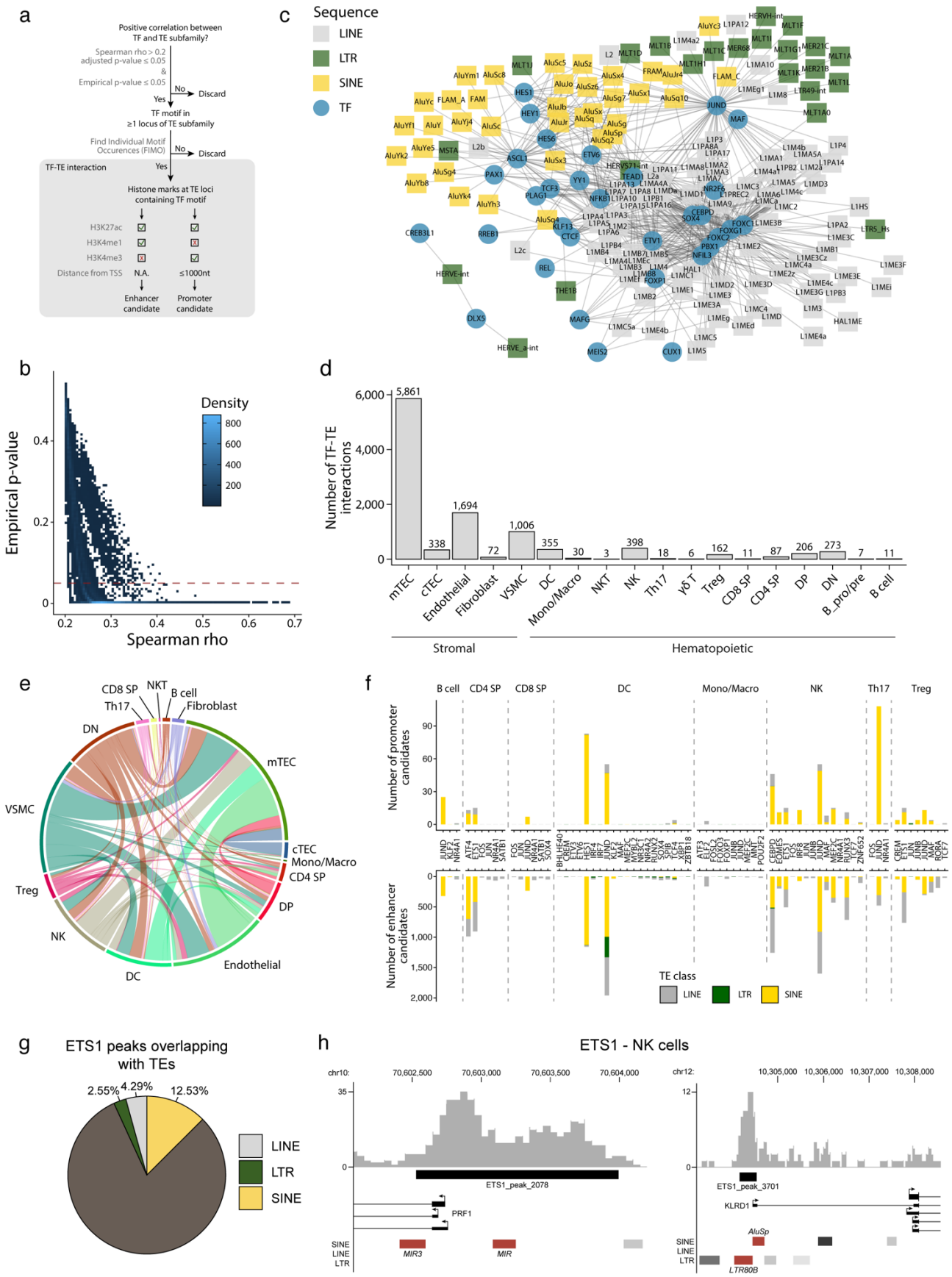


Figure 1 – figure supplement 4. KZFPs repress TE expression in the hematopoietic lineage of the human thymus.

Lower panel: pairs of TE subfamilies and KZFPs significantly correlated in at least two cell types (significant correlation: r>0.2 and adj. p≤0.05, or r<-0.2 and adj. p≤0.05, p-values corrected for multiple comparisons with the Benjamini-Hochberg method). *Middle panel*: Enrichment of the KZFP in the sequence of the correlated TE subfamily in ChIP-seq data from *Imbeault et al.* (44). *Upper panel*: Age of TE subfamilies in millions of years (My). The dashed line indicates the estimated time of divergence between primates and rodents (82 million years ago).


a
Positive correlation between TF and TE subfamily?
Spearman rho > 0.2 adjusted p-value ≤ 0.05 & Empirical p-value ≤ 0.05
Yes
No
Discard
TF motif in ≥1 locus of TE subfamily
Find Individual Motif Occurences (FIMO)
TF-TE interaction
Histone marks at TE loci containing TF motif
H3K27ac
H3K4me1
H3K4me3
Distance from TSS
N.A.
≤1000nt
Enhancer candidate
Promoter candidate
b
Density
Empirical p-value
Spearman rho
c
Sequence
LINE
LTR
SINE
TF
d
Number of TF-TE interactions
5,861
338
1,694
72
1,006
355
30
3
398
18
6
162
11
87
206
273
7
11
mTEC
cTEC
Endothelial
Fibroblast
VSMC
DC
Mono/Macro
NKT
NK
Th17
γδ T
Treg
CD8 SP
CD4 SP
DP
DN
B_pro/pre
B cell
Stromal
Hematopoietic
e
CD8 SP
NKT
B cell
Th17
Fibroblast
DN
mTEC
VSMC
cTEC
Mono/Macro
CD4 SP
Treg
DP
NK
Endothelial
DC
f
B cell
CD4 SP
CD8 SP
DC
Mono/Macro
NK
Th17
Treg
Number of promoter candidates
Number of enhancer candidates
TE class
LINE
LTR
SINE
g
ETS1 peaks overlapping with TEs
2.55%
4.29%
12.53%
LINE
LTR
SINE
h
ETS1 - NK cells
chr10:
70,602,500
70,603,000
70,603,500
70,604,000
ETS1_peak_2078
PRF1
MIR3
MIR
chr12:
10,305,000
10,306,000
10,307,000
10,308,000
ETS1_peak_3701
KLRD1
AluSp
LTR80B

1060

Figure 2. TEs shape complex gene regulatory networks in human thymic cells. (a) The flowchart depicts the decision tree for each TE promoter or enhancer candidate. **(b)** Density heatmap representing the correlation coefficient and the empirical p-value determined by bootstrap for TF and TE pairs in each cell type of the dataset. The color code shows density (i.e., the occurrence of TF-TE pairs at a specific point). **(c)** Connectivity map of interactions between TEs and TFs in mTECs. For visualization purposes, only TF-TE pairs with high positive correlations (Spearman correlation coefficient ≥ 0.3 and p-value adjusted for multiple comparisons with the Benjamini-Hochberg procedure ≤ 0.05) and TF binding sites in $\geq 1\%$ of TE loci are shown. **(d)** Number of TF-TE interactions for each thymic cell population. **(e)** Sharing of TF-TE pairs between thymic cell types. **(f)** Number of promoter (*top*) or enhancer (*bottom*) TE candidates per transcription factor in hematopoietic cells of the thymus. **(g)** The proportion of statistically significant peaks overlapping with TE sequences in ETS1 ChIP-seq data from NK cells. **(h)** Genomic tracks depicting the colocalization of ETS1 occupancy (i.e., read coverage) and TE sequences (*in red)* in the upstream region of two genes in ETS1 ChIP-seq data from NK cells. Statistically significant ETS1 peaks are indicated by the black rectangles.

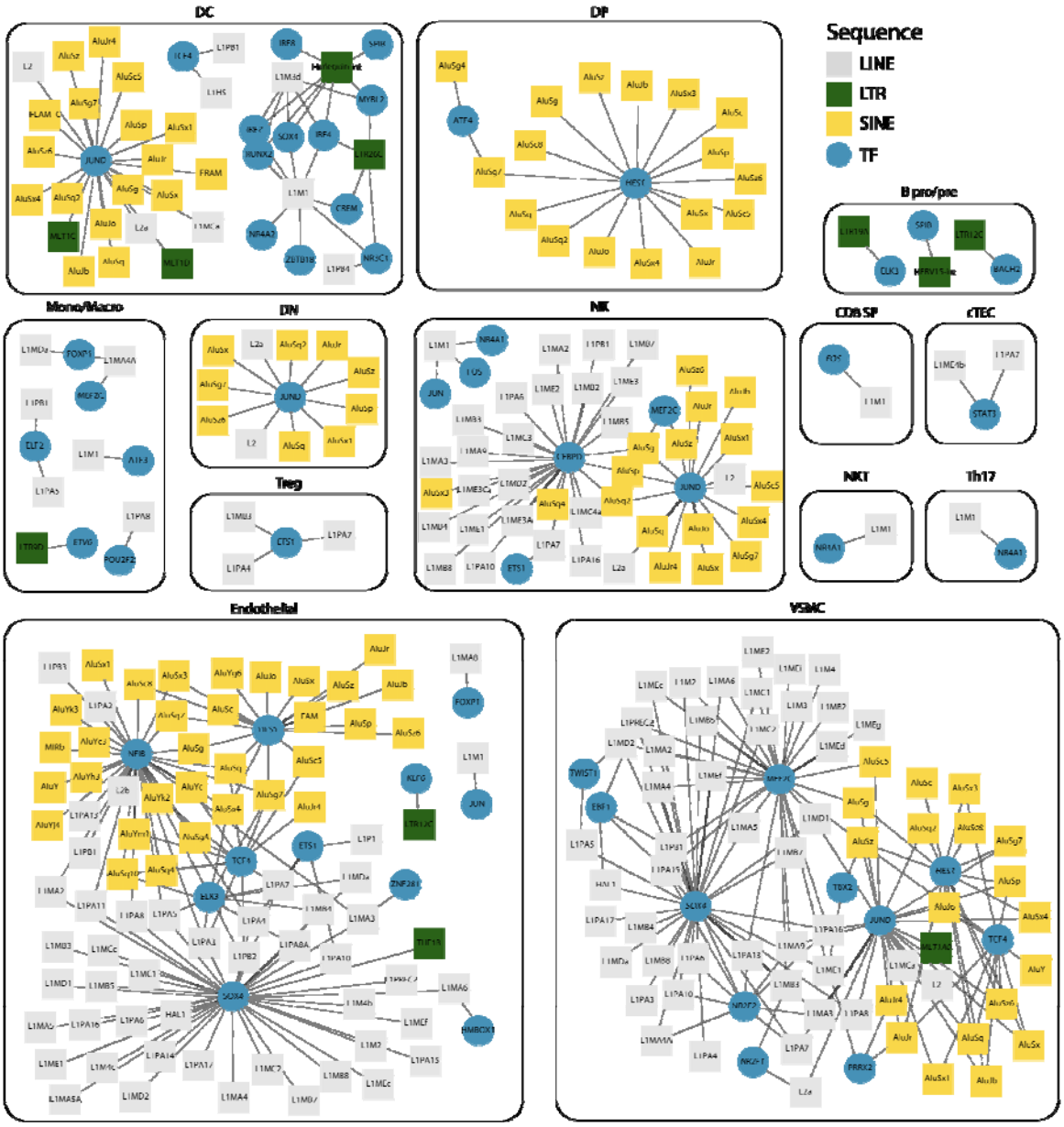
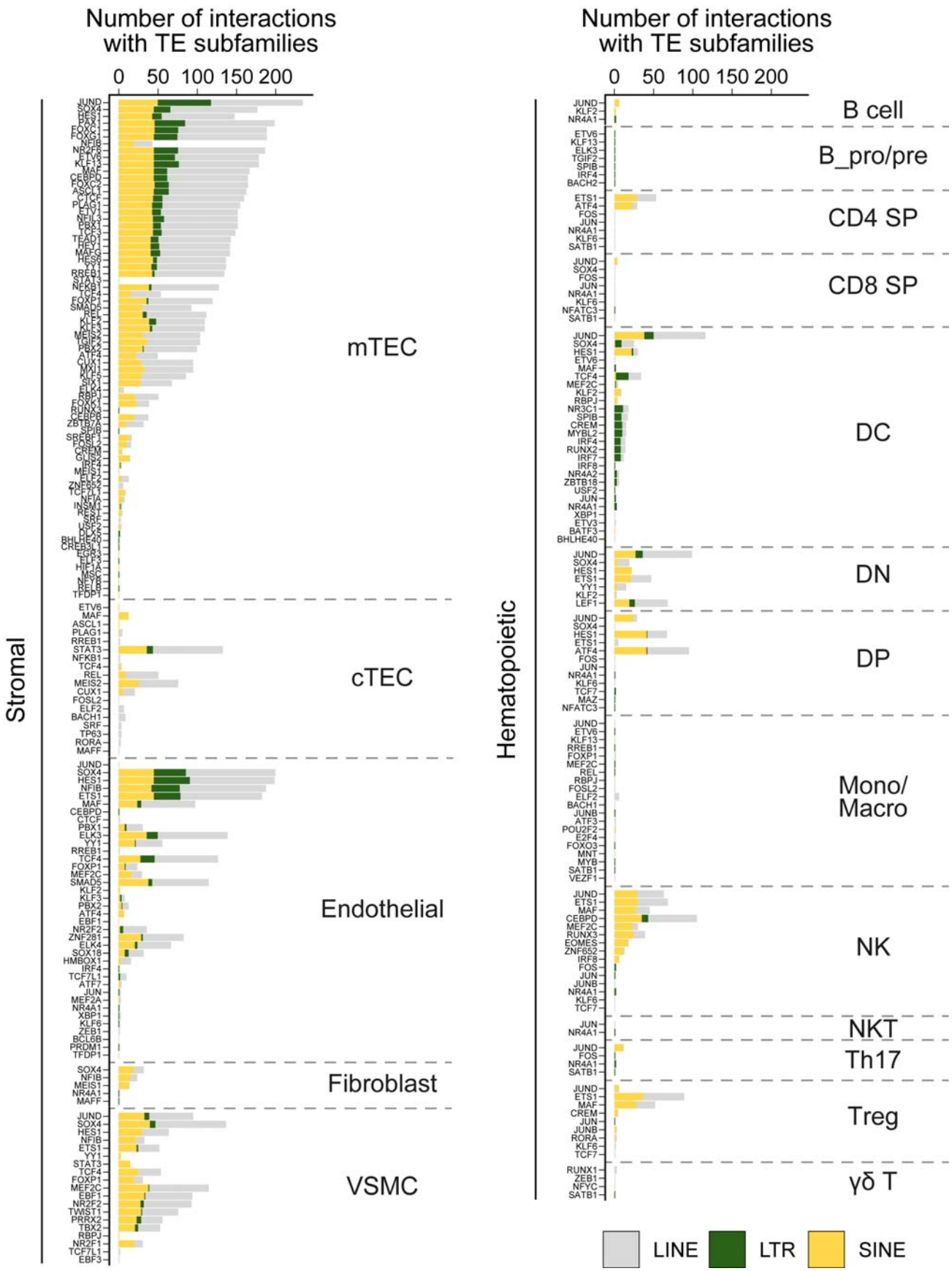


Figure 2 – figure supplement 1. Interaction networks between transcription factors and TE subfamilies.

For each cell type, networks illustrate the interactions between TF and TE subfamilies. Pairs of TF and TE are connected by edges when i) their expressions are significantly correlated (Spearman correlation coefficient ≥ 0.2) and ii) the TF binding motifs are found in the loci of the TE subfamily. TE subfamilies are colored based on the class of TE subfamily (LINE, LTR, and SINE).


Number of interactions with TE subfamilies
0 50 100 150 200
Stromal
mTEC
cTEC
Endothelial
Fibroblast
VSMC
Number of interactions with TE subfamilies
0 50 100 150 200
Hematopoietic
B cell
B_pro/pre
CD4 SP
CD8 SP
DC
DN
DP
Mono/ Macro
NK
NKT
Th17
Treg
γδ T
LINE
LTR
SINE

1082

Figure 2 – figure supplement 2. Frequency of interactions between transcription factors and TE subfamilies in thymic cells. For each cell type of the stromal (*left*) or hematopoietic (*right*) compartments of the thymus, the graph shows the number of interactions between transcription factors and TE subfamilies of the LINE, LTR, or SINE groups.

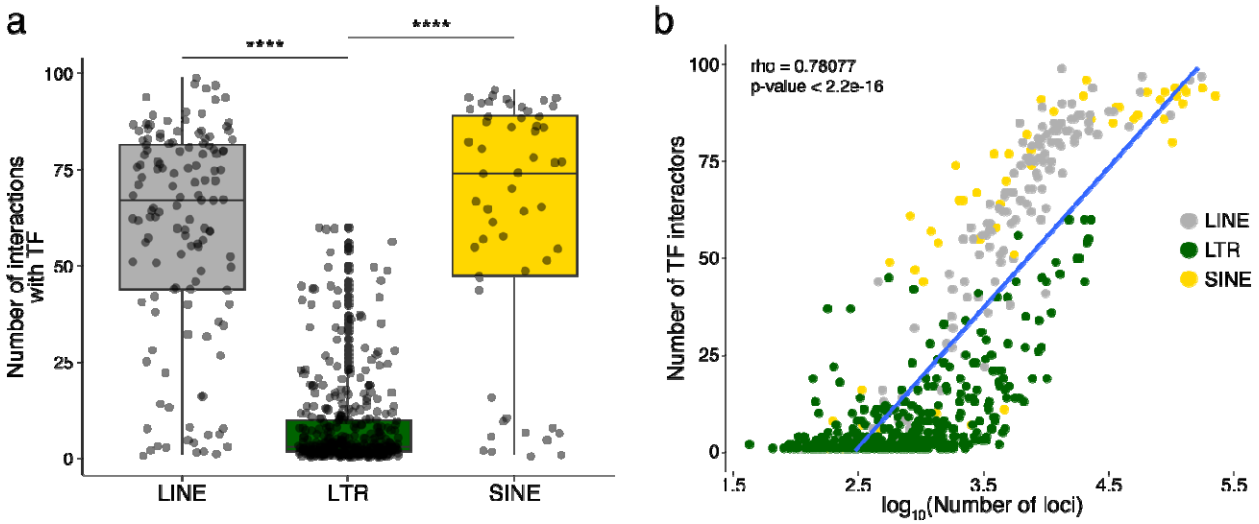


Figure 2 – figure supplement 3. TE subfamilies occupying larger genomic spaces interact more frequently with TF.

(a) Number of interactions formed with TFs for each TE subfamily of the LINE, LTR, and SINE classes (Wilcoxon-Mann-Whitney tests, ****$p \leq 0.0001$). **(b)** Scatterplot depicting the Kendall tau correlation between the number of interactions with TFs of a TE subfamily and the number of loci of that subfamily in the human genome. The color code indicates the class of TE subfamilies.

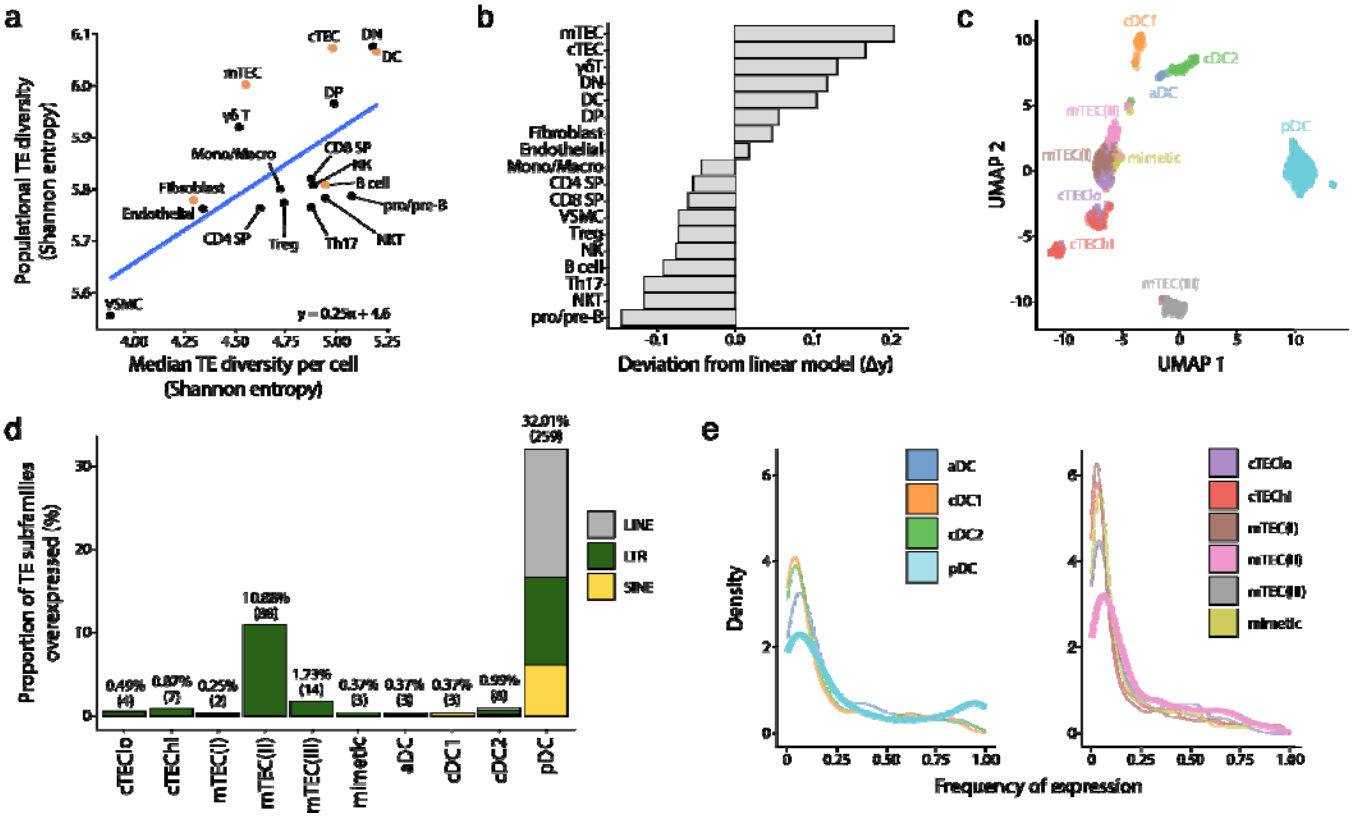


Figure 3. Human pDCs and mTEC(II) express diverse and distinct repertoires of TE sequences. **(a)** Diversity of TEs expressed by thymic populations measured by Shannon entropy. The x and y axes represent the median diversity of TEs expressed by individual cells in a population and the global diversity of TEs expressed by an entire population, respectively. The equation and blue curve represent a linear model summarizing the data. Thymic APC subsets are indicated in orange. **(b)** Difference between the observed diversity of TEs expressed by cell populations and the one expected by the linear model in (A). **(c)** UMAP showing the subsets of thymic APCs (aDC, activated DC; cDC1, conventional DC1; cDC2, conventional DC2; pDC, plasmacytoid DC). **(d)** Bar plot showing the number and class of differentially expressed TE subfamilies between APC subsets. **(e)** Frequency of expression of TE subfamilies by the different APC subsets. The distributions for pDCs and mTEC(II) are highlighted in bold.

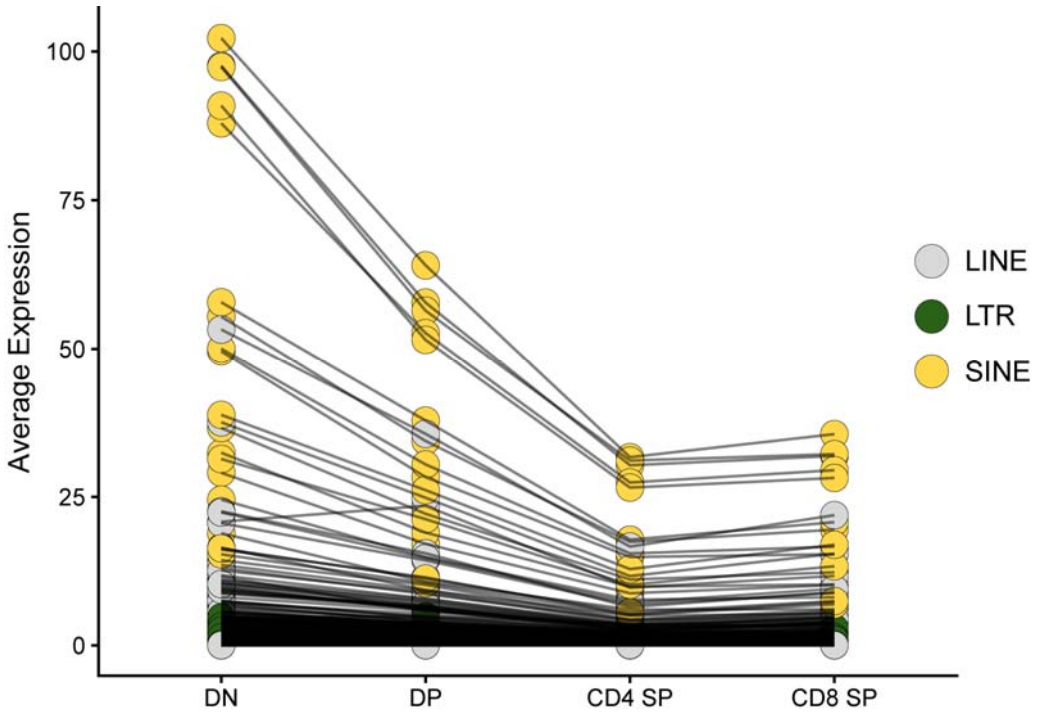


Figure 3 – figure supplement 1. TE expression decreases during thymocyte differentiation. The average expression level of TE subfamilies across cells of the four main populations of thymocytes is shown: DN, DP, CD4 SP, and CD8 SP. Black lines between thymocyte subsets connect expression values for the same TE subfamily.

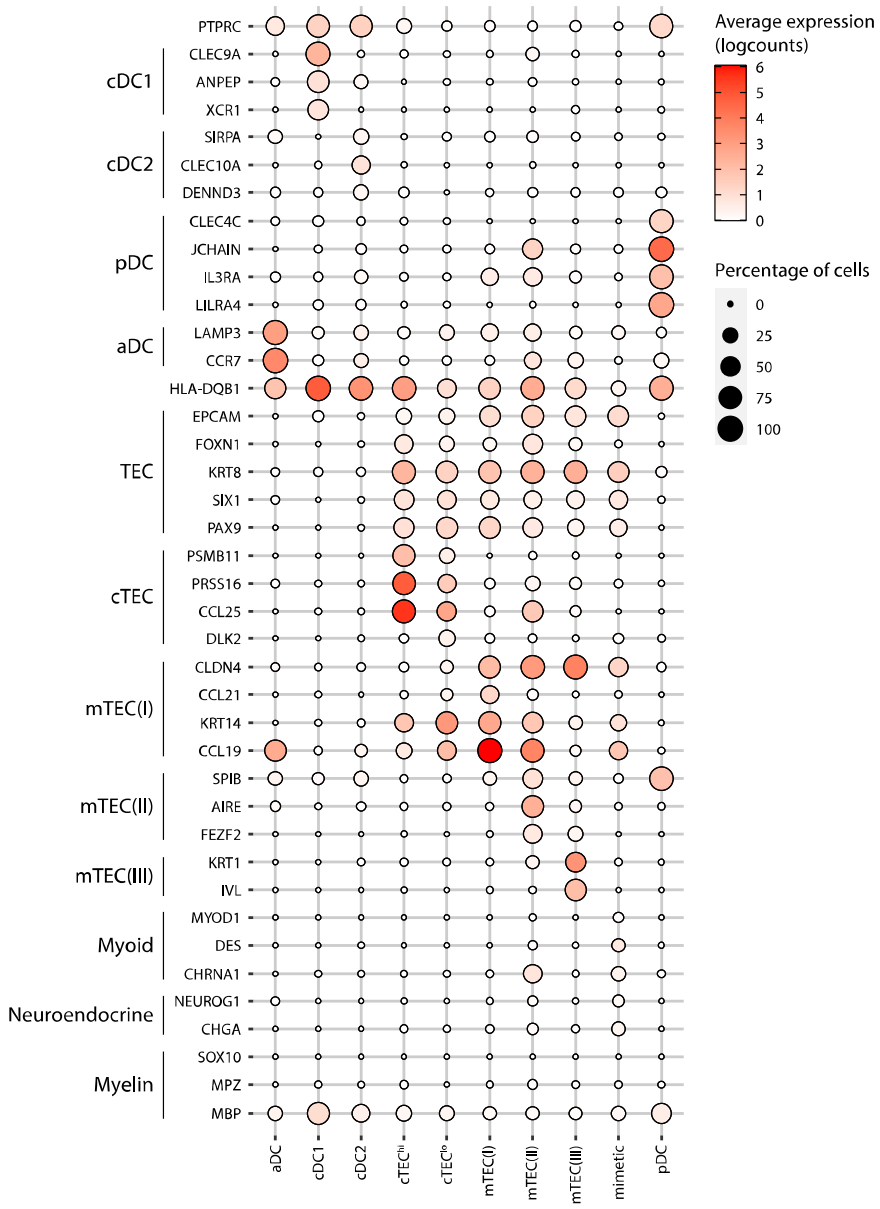


Figure 3 – figure supplement 2. Annotation of human thymic antigen-presenting cell subsets. Dot plot depicting the expression of marker genes in the annotated cell types of the thymus. The average expression and percentage of cells expressing the gene are represented by the color and size of the dot, respectively. Myoid-, myeloid- and neuroendocrine-related genes are used as markers of mimetic mTEC. (aDC, activated dendritic cell; cDC1, conventional dendritic cell 1; cDC2, conventional

dendritic cell 2; cTEC, cortical thymic epithelial cell; mTEC, medullary thymic epithelial cell; pDC, plasmacytoid dendritic cell).

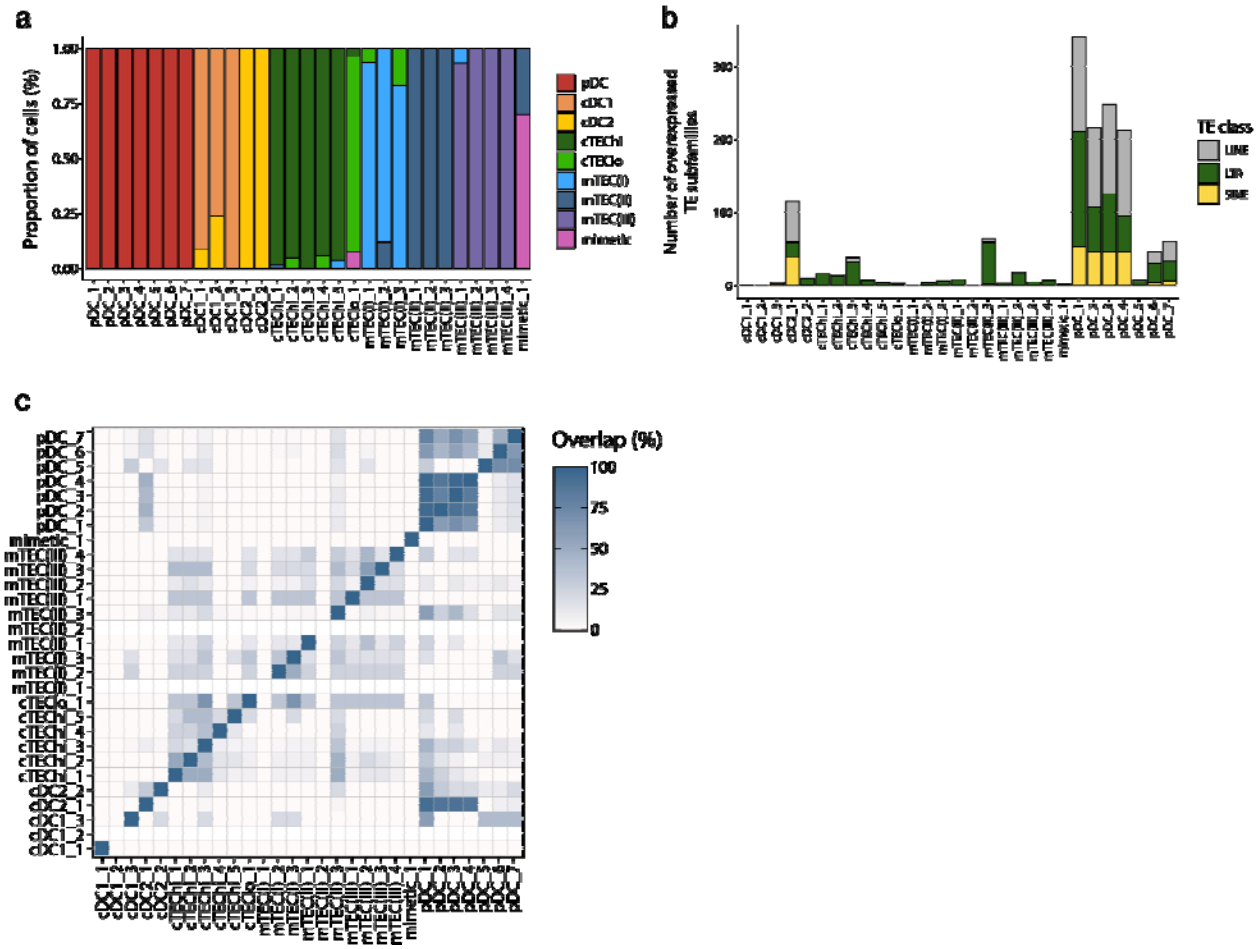


Figure 3 – figure supplement 3. Differential TE expression in metacells of human thymic antigen-presenting cells.

(a) Cellular composition of the metacells (x-axis) based on the manual annotation of the thymic cell populations (see Fig. S1). **(b)** Number of TE subfamilies overexpressed expressed between the metacells. TE subfamilies are colored based on class (LINE, LTR, and SINE). **(c)** Percentage of overlap of the TE subfamilies overexpressed by each metacell.

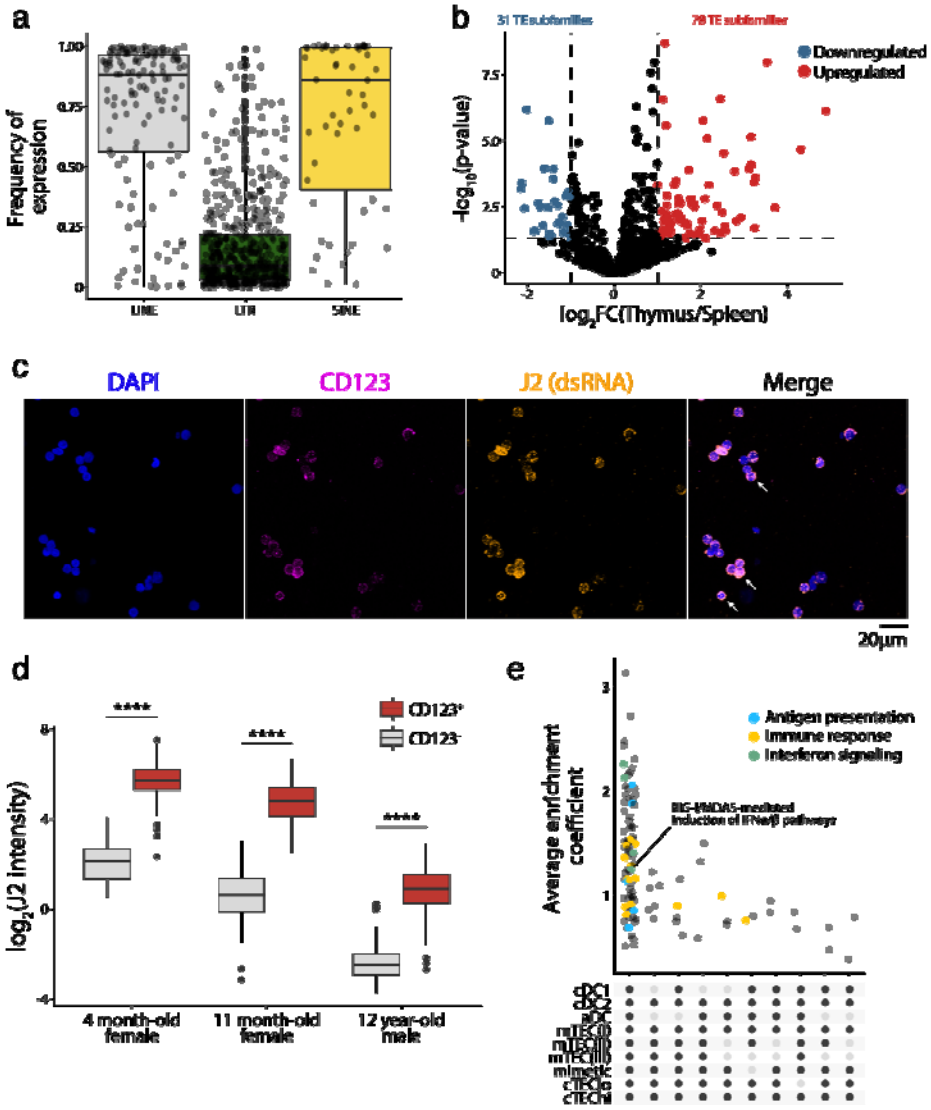


Figure 4. TE expression in human pDCs is associated with dsRNA formation and type I IFN signaling. (a) Frequency of LINE, LTR, and SINE subfamilies expression in thymic pDCs. **(b)** Differential expression of TE subfamilies between splenic and thymic pDCs. TE subfamilies significantly upregulated or downregulated by thymic pDCs are indicated in red and blue, respectively (Upregulated, $\log_2$(Thymus/Spleen)≥1 and adj. p≤0.05; Downregulated, $\log_2$(Thymus/Spleen)≤-1 and adj. p≤0.05). **(c,d)** Immunostaining of dsRNAs in human thymic pDCs ($CD123^+$) using the J2 antibody (n=3). **(c)** One representative experiment. Three examples of CD123 and J2 colocalization are shown with white arrows. **(d)** J2 staining intensity in $CD123^+$ and $CD123^-$ cells from three human thymi

(Wilcoxon Rank Sum test, ****p-value≤0.0001). **(e)** UpSet plot showing gene sets enriched in pDCs compared to the other populations of thymic APCs. On the lower panel, black dots represent cell populations for which gene signatures are significantly depleted compared to pDCs. All comparisons where gene signatures were significantly enriched in pDCs are shown.

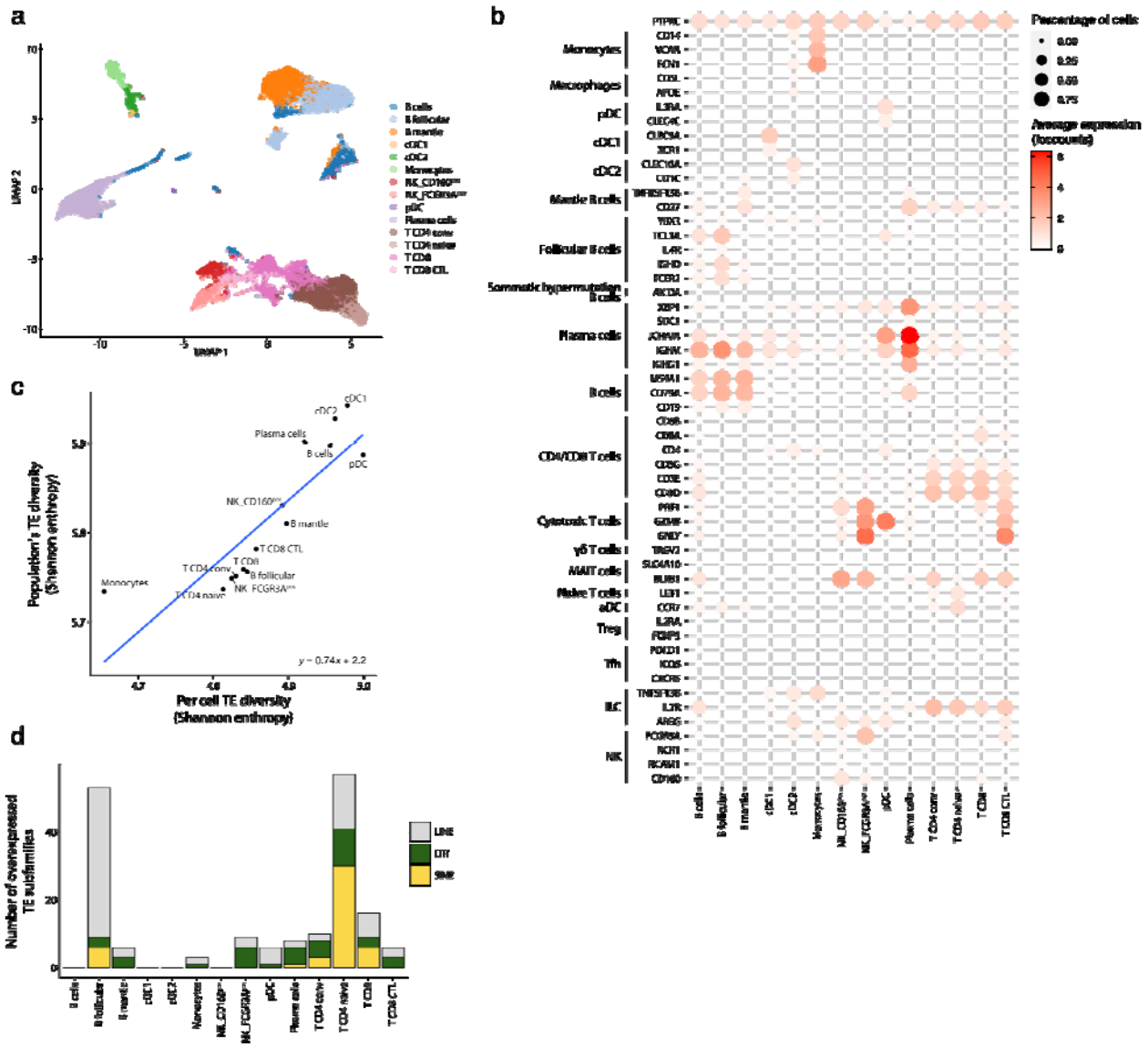


Figure 4 – figure supplement 1. TE expression in human splenic pDCs.

(a) UMAP depicting the cell populations present in the human spleen. **(b)** Dot plot showing the expression of marker genes in the annotated cell types of the spleen. The average expression and percentage of cells expressing the gene are represented by the color and size of the dot, respectively. **(c)** Diversity of TE expressed by splenic populations measured by Shannon entropy. The x and y axes represent the median diversity of TE expressed by individual cells of a population and the global diversity of TE expressed by discrete populations, respectively. The equation and blue curve represent a linear model summarizing the data. **(d)** Bar plot showing the number (y-axis) and class (color) of differentially expressed TE subfamilies between splenic cell populations.

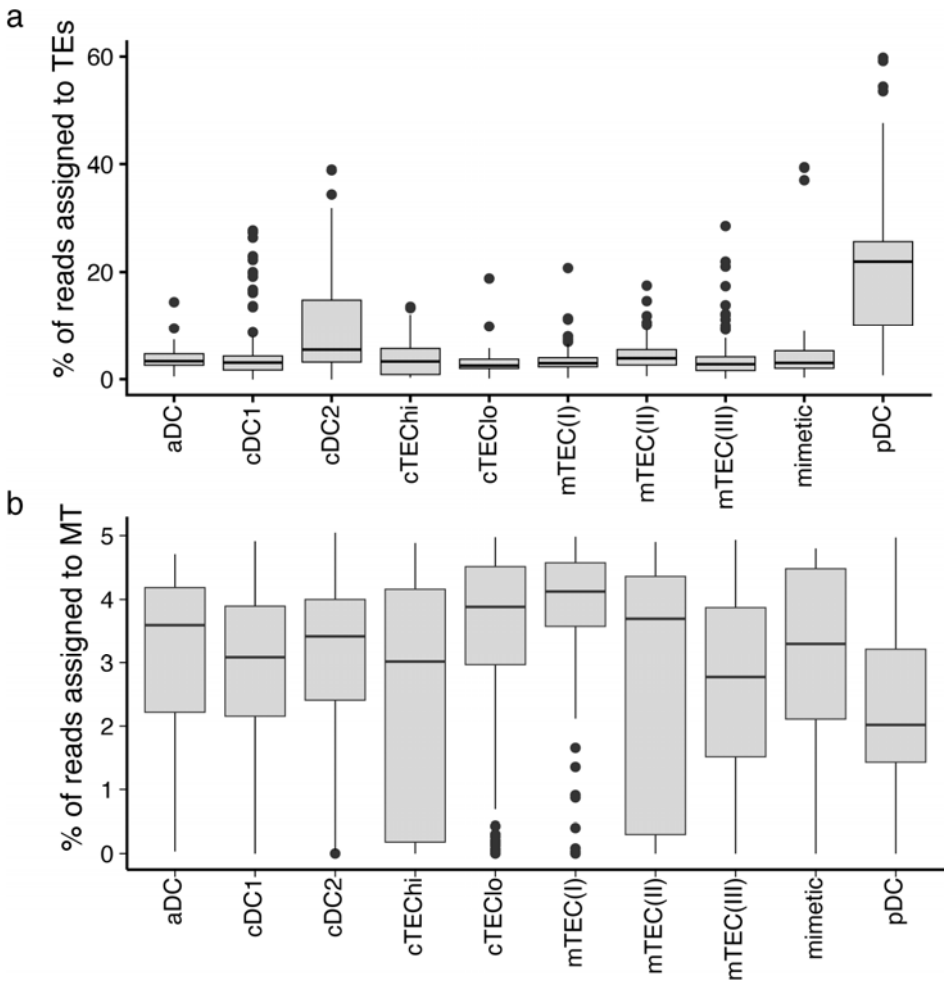


Figure 4 – figure supplement 2. A higher proportion of reads originates from TEs in pDCs than in other thymic APCs. Boxplots depicting the percentage of reads assigned to **(a)** TE sequences or **(b)** mitochondrial reads in the different subpopulations of thymic APCs.

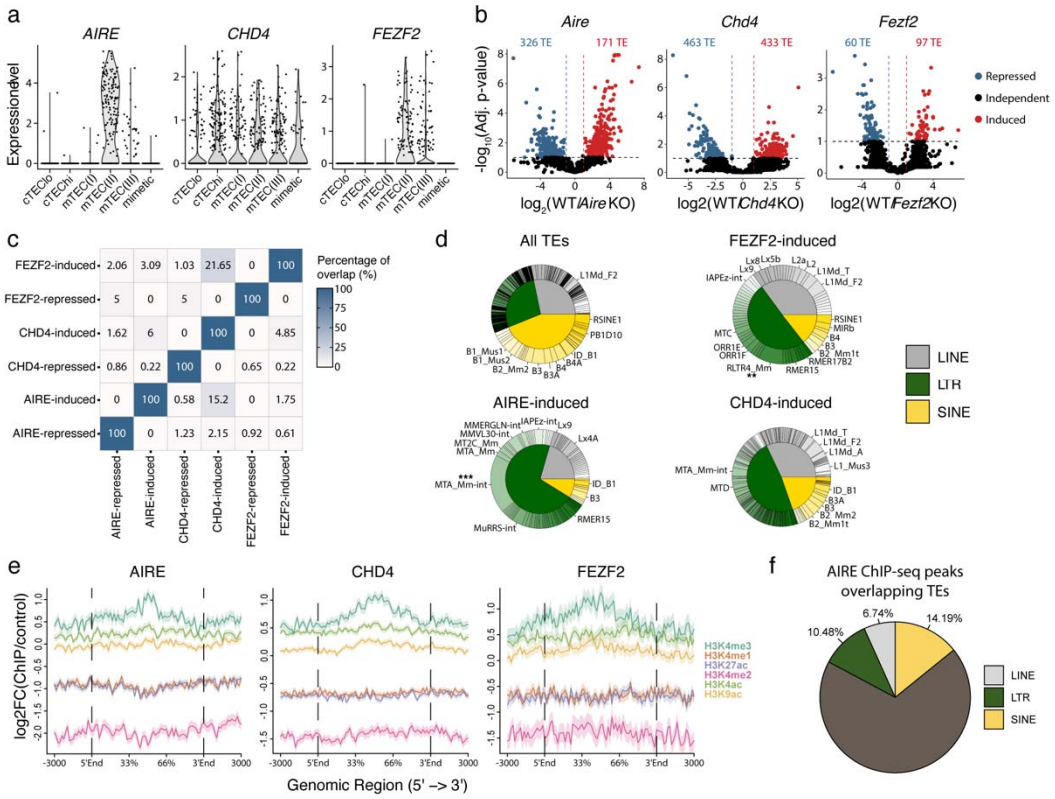


Figure 5. *AIRE*, *FEZF2*, and *CHD4* regulate non-redundant sets of TEs in murine mTECs. (a) Expression of *AIRE*, *CHD4*, and *FEZF2* in human TEC subsets. **(b)** Differential expression of TE loci between wild-type (WT) and *Aire*-, *Chd4*- or *Fezf2*-knockout (KO) mice (Induced, $\log_2$(WT/KO)≥2 and adj. p≤0.05; Repressed, $\log_2$(WT/KO)≤-2 and adj. p≤0.05). P-values were corrected for multiple comparisons with the Benjamini-Hochberg procedure. The numbers of induced (red) and repressed (blue) TE loci are indicated on the volcano plots. **(c)** Overlap of TE loci repressed or induced by AIRE, FEZF2, and CHD4. **(d)** Proportion of TE classes and subfamilies in the TE loci regulated by AIRE,

FEZF2, or CHD4, as well as all TE loci in the murine genome for comparison (Chi-squared tests with Bonferroni correction, **adj. p≤0.01, ***adj. p≤0.001). **(e)** Plots for the tag density of H3K4me3 and H3K4me2 on the sequence and flanking regions (3000 base pairs) of TE loci induced by AIRE, FEZF2, and CHD4. **(f)** Proportion of statistically significant peaks overlapping TE sequences in AIRE ChIP-seq data from murine mTECs.

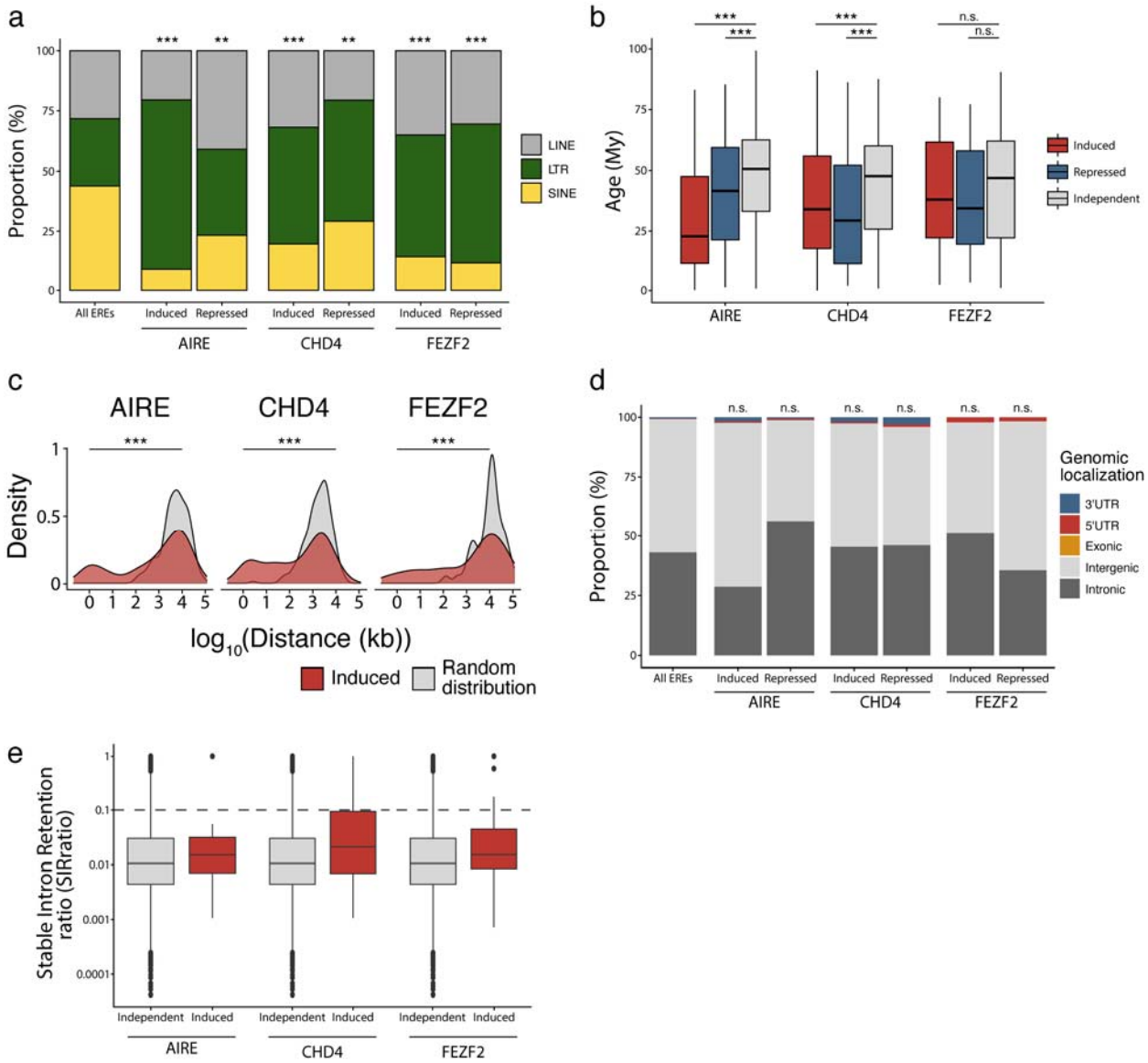


Figure 5 – figure supplement 1. Characterization of TE subfamilies regulated by AIRE, CHD4, and FEZF2 in murine mTECs.

(a) Class of TEs induced or repressed by AIRE, CHD4, and FEZF2. Distributions were compared to the proportion of LINEs, LTRs, and SINEs amongst all TE sequences of the murine genome with Chi-squared tests (**p≤0.01, ***p≤0.001). **(b)** Age of TEs induced, repressed, or independent of AIRE, CHD4, and FEZF2 (Wilcoxon-Mann-Whitney test, *p≤0.05, ***p≤0.001) (My, millions of years). **(c)** Distance between TE loci induced by AIRE, FEZF2, and CHD4, and random selections of TE loci (Wilcoxon rank-sum tests, ***p≤0.001). **(d)** Genomic localization of the TE loci induced or repressed by AIRE, CHD4, and FEZF2. **(e)** Intron retention ratio of intronic TE induced or independent of AIRE, CHD4, and FEZF2. The dashed line represents intron retention events occurring in at least 10% of transcripts.

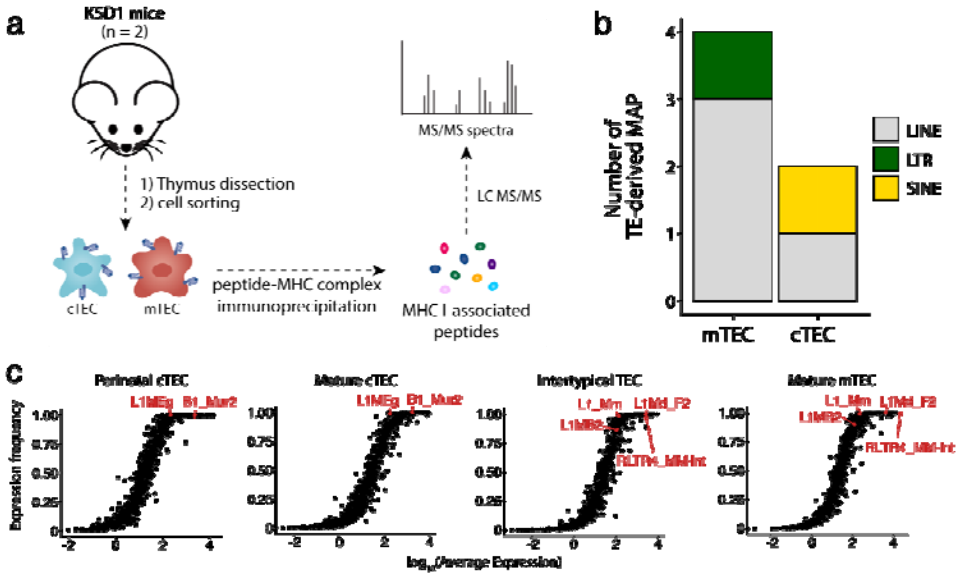


Figure 6. Murine cTECs and mTECs present TE MAPs. (a) mTECs and cTECs were isolated from the thymi of K5D1 mice (n=2). The peptide-MHC I complexes were immunoprecipitated independently for both populations, and MAPs were sequenced by MS analyses. **(b)** Number of LINE-, LTR-, and

SINE-derived MAPs in mTECs and cTECs from K5D1 mice. **(c)** Distributions of TE subfamilies in murine TECs subsets based on expression level (*x-axis*) and frequency of expression (*y-axis*).