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.

19 [10] Department of Chemistry, Université de Montréal; Montréal, Canada.

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

21 **Abstract**

22 Transposable elements (TE) are repetitive sequences representing ~45% of the human and mouse genomes
23 and are highly expressed by medullary thymic epithelial cells (mTEC). In this study, we investigated the
24 role of transposable elements (TE), which are highly expressed by medullary thymic epithelial cells
25 (mTEC), on T-cell development in the thymus. We performed multi-omic analyses of TEs in human and
26 mouse thymic cells to elucidate their role in T cell development. We report that TE expression in the
27 human thymus is high and shows extensive age- and cell lineage-related variations. TEs interact with
28 multiple transcription factors in all cell types of the human thymus. Two cell types express particularly
29 broad TE repertoires: mTECs and plasmacytoid dendritic cells (pDC). In mTECs, TEs interact with
30 transcription factors essential for mTEC development and function (e.g., PAX1 and RELB) and generate
31 MHC-I-associated peptides implicated in thymocyte education. Notably, AIRE, FEZF2, and CHD4
32 regulate non-redundant sets of TEs in murine mTECs. Human thymic pDCs homogenously express large
33 numbers of TEs that lead to the formation of dsRNA, triggering RIG-I and MDA5 signaling and
34 explaining why thymic pDCs constitutively secrete IFN α/β. This study illustrates the diversity of
35 interactions between TEs and the adaptive immune system. TEs are genetic parasites, and the two thymic
36 cell types most affected by TEs (mTEcs and pDCs) are essential to establishing central T-cell tolerance.
37 Therefore, we propose that the orchestration of TE expression in thymic cells is critical to prevent
38 autoimmunity in vertebrates.

39

40

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 (IFNα) (21). Notably, thymic pDCs originate from intrathymic IRF8^{hi} precursors, and, in contrast to extrathymic pDCs, they constitutively secrete high amounts of IFNα (22-24). This constitutive IFNα 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 themselves subdivided into more than 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.

87 **Results**

88 **LINE, LTR, and SINE expression shows extensive variations during thymus ontogeny**

89 We first profiled TEs expression in various thymic cell populations during development. To do so, we
90 quantified the expression of 809 TE subfamilies (classified according to the RepeatMasker annotations)
91 in the scRNA-seq dataset created by *Park et al.* (42). Cells were clustered in 19 populations representing
92 the main constituents of the thymic hematolymphoid and stromal compartments (Figure 1a and Figure 1
93 – figure supplement 1). The expression of TE subfamilies was quantified at all developmental stages
94 available, ranging from 7 post-conception weeks (pcw) to 40 years of age (Supplementary file 1 – Table
95 1). Unsupervised hierarchical clustering revealed three clusters of TE subfamilies based on their pattern
96 of expression during thymic development (Figure 1b, upper panel): i) maximal expression at early
97 embryonic stages persisting, albeit at lower levels, throughout ontogeny (cluster 1), ii) an expression
98 specific to a given timepoint (cluster 2), or iii) a high expression at early embryonic stages that decreases
99 rapidly at later timepoints (cluster 3). LINE and SINE subfamilies were enriched in cluster 1, whereas
100 LTR subfamilies were significantly enriched in clusters 2 and 3 (Figure 1b, lower panel). Expression of
101 individual LINE and SINE subfamilies was highly shared among different cell types (Figure 1d). In
102 contrast, the pattern of expression of LTR subfamilies was shared by fewer cell subsets and adopted a
103 quasi-random distribution (Figure 1d). The pattern of expression assigned to TE subfamilies (Figure 1c,
104 innermost track) was not affected by the proportion of cells of different developmental stages (embryonic
105 or postnatal) (Figure 1c, outermost track, and Figure 1 – figure supplement 2). This suggests that our
106 observations do not result from a bias in the composition of the dataset. To gain further insights into the
107 expression of TE subfamilies, we studied two biological processes known to regulate TE expression in
108 other contexts: cell proliferation and expression of KRAB zinc-finger proteins (KZFP) (43, 44). Cell
109 cycling scores negatively correlated with TE expression in various thymic cell subsets, particularly for

110 LINE and SINE subfamilies shared among cell types (Figure 1 – figure supplement 3 and Supplementary
111 file 1 – Table 2). KZFPs, on the other hand, repressed TE expression strictly in discrete subsets of
112 hematolymphoid cells (Figure 1 – figure supplement 4 and Supplementary file 1 – Table 3). The strongest
113 correlation involved ZNF10, a probable repressor of L1 subfamilies in Th17 and NK cells (Figure 1 –
114 figure supplement 4 and Supplementary file 1 – Table 3). We conclude that the expression of the three
115 main classes of TEs shows major divergences as a function of age and thymic cell types.

116

117 **TEs interact with transcription factors regulating thymic development and function**

118 TEs provide binding sites to TFs (37, 45, 46), and T-cell development is driven by the coordinated timing
119 of multiple changes in transcriptional regulators (47). We, therefore, investigated interactions between
120 TEs and TFs during thymic development. Two criteria defined an interaction: i) a significant and positive
121 correlation between the expression of a TF and a TE subfamily in a given cell population, and ii) the
122 presence of the TF binding motif in the loci of the TE subfamily (Figure 2a). Additionally, we validated
123 the correlations we obtained using a bootstrap procedure to ascertain their reproducibility (see *Material*
124 *and Methods* for details). This procedure removed weakly correlated TF-TE pairs (Figure 2b). TF-TE
125 interactions were observed in all thymic cell populations (Figure 2c, d, Figure 2 – figure supplement 1,
126 and Supplementary file 1 – Table 4). Numerous TF-TE interactions were conserved between
127 hematolymphoid and stromal cell subsets (Figure 2e). However, both the number of interactions and the
128 complexity of the interaction networks were much higher in mTECs than in other cell populations (Figure
129 2c, d and Figure 2 – figure supplement 1).

130

131 Several TFs instrumental in thymus development and thymopoiesis interacted with TE subfamilies (Figure
132 2c, Figure 2 – figure supplement 1, and Supplementary file 1 – Table 4). These TFs include the *NFKB1*
133 and *RELB* subunits of the NF-κB complex as well as *PAX1* in mTECs (48-50), *JUND* in thymocytes (51),
134 and *BCL11A* in B cells (52). In DCs, the most notable TF-TE interactions involved interferon regulatory
135 factors (IRF), which regulate the late stages of T-cell maturation, and *TCF4,* which is essential for pDC
136 development (25, 53). This observation is consistent with evidence that TEs have shaped the evolution of
137 IFN signaling networks (37). Finally, we found significant interactions between *CTCF* and several Alu
138 subfamilies in mTECs, VSMCs, and endothelial cells, suggesting that binding of *CTCF* to TE sequences
139 might affect the tridimensional structure of the chromatin in the thymic stroma (54). Interestingly, LINE
140 and SINE subfamilies that occupy more genomic space interacted with higher numbers of transcription
141 factors (Figure 2 – figure supplement 2).

142

143 Using data from the ENCODE consortium for hematopoietic cells (55, 56), we looked at the histone marks
144 in the TE sequences identified as TF interactors by our analyses. The objective was to determine if they
145 could act as promoters or enhancers (Figure 2a and Supplementary file 1 – Table 5). We found several TE
146 promoter and enhancer candidates in all eight hematopoietic cell types analyzed, with a striking
147 overrepresentation of LINE and SINE compared to LTR sequences (Figure 2f and Supplementary file 1 –
148 Table 6). Finally, we validated the potential of one TF important for NK cell development (57), *ETS1,* to
149 interact with its TE promoter candidates in NK cells by reanalyzing publicly available ChIP-seq data (57).
150 *ETS1* signal was observed in the sequence of two TE promoter candidates (Figure 2g, in red), confirming
151 its potential to bind the TE promoters. Moreover, these TE promoters are located upstream of two genes,
152 *ZNF26* and *MTMR3*, described as ETS1 target genes by the ENCODE Consortium. Hence, our data

153 indicate that TEs can affect the development and function of the thymus by providing binding sites to
154 multiple TFs.

155

156 **TEs are highly and differentially expressed in thymic APC subsets**

157 We next sought to determine whether the high expression of TEs reported in mTECs (32, 33) was limited
158 to this cell subset or was found in other thymic cell types. Since several thymic stromal cells reach maturity
159 after birth (58), we selected postnatal samples for the following analyses. We computed two distinct
160 Shannon entropy indices: one for the global diversity of TEs expressed by all cells of a given population,
161 and another for the median value of TE diversity expressed by individual cells of a population (Figure 3a).
162 Then, we computed a linear model to represent the diversity of TEs expressed by a cell population based
163 on the diversity of TEs expressed by individual cells (Figure 3a, blue curve). Two salient findings emerged
164 from this analysis. First, the diversity of TEs expressed in the T-cell lineage decreases during
165 differentiation according to the following hierarchy: DN thymocytes > DP thymocytes > SP thymocytes
166 (Figure 3a). Second, among the populations of thymic APCs implicated in positive and negative selection
167 (Figure 3a, orange dots), cTECs, mTECs, and DCs expressed broader repertoires of TEs than B cells and
168 fibroblasts. While cTECs and DCs expressed highly diverse TE repertoires at both the population and
169 individual cell levels, the breadth of TE expression in mTECs was found only at the population level
170 (Figure 3a). Accordingly, intercellular heterogeneity (i.e., deviation from the linear model) was higher for
171 mTECs than other cell populations (Figure 3b).

172

173 We next focused on thymic APCs expressing the broadest TE repertoires: cTECs, mTECs, and DCs
174 (Figure 3a). To this end, we annotated these APC subpopulations based on previously published lists of

175 marker genes (Figure 3c and Figure 3 – figure supplement 1) (42, 59). We performed differential
176 expression analyses to determine whether some TE subfamilies were overexpressed in specific APC
177 subsets. pDCs and mTEC(II) overexpressed a broader TE repertoire than other APCs: 32.01% of
178 subfamilies were overexpressed in pDCs and 10.88% in mTEC(II) (Figure 3d and Supplementary file 1 –
179 Table 7). The nature of the overexpressed TEs differed between pDCs and other thymic APC subsets.
180 Indeed, pDCs overexpressed LTRs, LINEs, and SINEs, including several Alu and L1 subfamilies (Figure
181 3d and Supplementary file 1 – Table 7). In contrast, other thymic APCs predominantly overexpressed
182 LTRs.

183

184 TE expression showed wildly divergent levels of intercellular heterogeneity in APC subsets. Indeed,
185 whereas most TE subfamilies were expressed by <25% of cells of the mTEC(II) population, an important
186 proportion of TEs were expressed by >75% of pDCs (Figure 3e). To evaluate this question further, we
187 compared TE expression between metacells of thymic APCs; metacells are small clusters of cells with
188 highly similar transcription profiles. This analysis revealed that overexpression of TE subfamilies was
189 shared between pDC metacells but not mTEC(II) metacells, reinforcing the idea that TE expression adopts
190 a mosaic pattern in the mTEC(II) population (Figure 3 – figure supplement 2). We conclude that cTECs,
191 mTECs, and DCs express broad TE repertoires. However, two subpopulations of thymic APCs clearly
192 stand out. pDCs express an extremely diversified repertoire of LTRs, SINEs, and LINEs, showing limited
193 intercellular heterogeneity, whereas the mTEC(II) population shows a highly heterogeneous
194 overexpression of LTR subfamilies.

195

196 **TE expression in pDCs is associated with dsRNA structures**

197 The high expression of a broad repertoire of TE sequences in thymic pDCs was unexpected (Figure 3d).
198 LINE and SINE subfamilies, particularly, were highly and homogeneously expressed by thymic pDCs
199 (Figure 4a). Constitutive IFNα secretion is a feature of thymic pDCs not found in extrathymic pDCs. We,
200 therefore, hypothesized that this constitutive IFNα secretion by thymic pDCs might be mechanistically
201 linked to their TE expression profile. We first assessed whether thymic and extrathymic pDCs have similar
202 TE expression profiles by reanalyzing scRNA-seq data from human spleens published by *Madissoon et*
203 *al.* (60) (Figure 4 – figure supplement 1a, b). This revealed that extrathymic pDCs express TE sequences
204 at similar or lower levels than other splenic cells (Figure 4 – figure supplement 1c, d). We then used
205 pseudobulk RNA-seq methods to perform a differential expression analysis of TE subfamilies between
206 thymic and splenic pDCs. This analysis confirmed that TE expression was globally higher in thymic than
207 in extrathymic pDCs (Figure 4b). Since TE overexpression can lead to the formation of dsRNA (40, 41),
208 we investigated if such structures were found in thymic pDCs. pDCs were magnetically enriched from
209 primary human thymi following labeling with anti-CD303 antibody. Then, pDC-enriched thymic cells
210 were stained with an antibody against CD123 (a marker of pDCs) and the J2 antibody that stains dsRNA.
211 The intensity of the J2 signal was more than 10-fold higher in $CD123^{+}$ relative to $CD123^{-}$ cells (Figure
212 4c, d). We conclude that thymic pDCs contain large amounts of dsRNAs. Finally, we performed gene set
213 enrichment analyses to ascertain if the high expression of TEs by thymic pDCs was associated with
214 specific gene signatures. These analyses highlighted signatures of antigen presentation, immune response,
215 and interferon signaling in thymic pDCs (Figure 4e and Supplementary file 1 – Table 8). Notably, thymic
216 pDCs harbored gene signatures of RIG-I and MDA5-mediated IFN α/β signaling (Figure 4e and
217 Supplementary file 1 – Table 8). Altogether, these data support a model in which high and ubiquitous
218 expression of LINEs and SINEs in thymic pDCs leads to the formation of dsRNAs recognized by RIG-I
219 and MDA5, causing their constitutive secretion of IFN α/β.

220

221 **Transcription factors involved in promiscuous gene expression regulate distinct sets of TE**
222 **sequences**

223 The essential role of mTECs in central tolerance hinges on their ability to ectopically express tissue-
224 restricted genes, whose expression is otherwise limited to specific epithelial lineage (61, 62). This
225 promiscuous gene expression is driven by AIRE, CHD4, and FEZF2 (18-20). We, therefore, investigated
226 the contribution of these three genes to the expression of TE subfamilies in the mTEC(II) population
227 (Figure 3d). First, we validated that mTEC(II) express *AIRE, CHD4,* and *FEZF2* in the human scRNA-
228 seq dataset (Figure 5a). Next, we analyzed published murine mTEC RNA-seq data to assess the regulation
229 of TE sequences by AIRE, CHD4, and FEZF2. Differential expression analyses between knock-out (KO)
230 and wild-type (WT) mice showed that these three factors regulate TE sequences, but the magnitude and
231 directionality of this regulation differed (Figure 5b and Supplementary file 1 – Table 9). Indeed, while
232 CHD4 had the biggest impact on TE expression by inducing 433 TE loci and repressing 463, FEZF2's
233 impact was minimal, with 97 TE loci induced and 60 repressed (Figure 5b). Besides, AIRE mainly acted
234 as a repressor of TE sequences, with 326 loci repressed and 171 induced (Figure 5b). Interestingly, there
235 was minimal overlap between the sets of TE sequences regulated by AIRE, CHD4, and FEZF2, indicating
236 that they have non-redundant roles in TE regulation (Figure 5c). Additionally, AIRE, CHD4, and FEZF2
237 preferentially targeted LTR and LINE elements, with significant enrichment of specific subfamilies such
238 as MTA_Mm-int and RLTR4_Mm that are induced by Aire and Fezf2, respectively (Figure 5d and Figure
239 5 – figure supplement 1a). We also noticed that the distance between regulated TE loci was smaller than
240 distributions of randomly selected TEs (Figure 5e). This suggests that AIRE, CHD4, and FEZF2
241 nonrandomly affect the expression of TE sequences located in specific genomic regions. While AIRE and
242 CHD4 preferentially targeted evolutionary young TE sequences, the age of the TE sequence did not seem

243 to affect the regulation by FEZF2 (Figure 5 – figure supplement 1b). We observed no significant
244 differences in the genomic localization of TE loci targeted by AIRE, CHD4, and FEZF2 relative to the
245 genomic localization of all TE sequences in the murine genome: most TE loci were located in intronic and
246 intergenic regions (Figure 5 – figure supplement 1c). Enrichment for intronic TEs could not be ascribed
247 to induction of global intron retention: the intron retention ratio was similar for TEs regulated or not by
248 AIRE, CHD4, and FEZF2 (Figure 5 – figure supplement 1d). ChIP-seq-based analysis of permissive
249 histone marks showed that TE loci induced by AIRE, CHD4, and FEZF2 were all marked by H3K4me2
250 and H3K4me3, but the position of those marks differed between the three regulators (Figure 5f and Figure
251 5 – figure supplement 1e). Hence AIRE, CHD4, and FEZF2 regulated the expression of small, yet non-
252 redundant, repertoires of TE sequences associated with permissive histone marks.
253

254 **TEs are translated and presented by MHC class I molecules in thymic APC**

255 Several TEs are translated and generate MAPs (34). Hence, the expression of TEs in cTECs and even
256 more in mTECs raises a fundamental question: do these TEs generate MAPs that would shape the T cell
257 repertoire? Mass spectrometry (MS) is the only method that can faithfully identify MAPs (63-65). Despite
258 its quintessential role in central tolerance, the MAP repertoire of mTECs has never been studied by MS
259 because of the impossibility of obtaining sufficient mTECs for MS analyses: mTECs represent ≤ 1% of
260 thymic cells, and they do not proliferate *in vitro*. To get enough cTECs and mTECs for MS analyses, we
261 used transgenic mice that express cyclin D1 under the control of the keratin 5 promoter (K5D1 mice).
262 These mice develop dramatic thymic hyperplasia, but their thymus is morphologically and functionally
263 normal (66-68). Primary cTECs and mTECs (2 replicates of 70 x 10^6 cells from 121 and 90 mice,
264 respectively) were isolated from the thymi of K5D1 mice as described (69). Following cell lysis and MHC
265 I immunoprecipitation, MAPs were analyzed by liquid chromatography MS/MS (Figure 6a). To identify

266 TE-coded MAPs, we generated a TE proteome by in silico translation of TE transcripts expressed by
267 mTECs or cTECs, and this TE proteome was concatenated with the canonical proteome. MS analyses
268 enabled the identification of a total of 1636 and 1714 MAPs in mTECs and cTECs, respectively. From
269 these, we identified 4 TE-derived MAPs in mTECs and 2 in cTECs, demonstrating that TEs can be
270 translated and presented by MHC I in the thymic cortex and medulla (Figure 6b and Supplementary file 1
271 – Table 10). These MAPs were coded by the three major groups of TE: LINEs (n=1), LTRs (n=1), and
272 SINEs (n=4). Next, we evaluated whether the low number of TE MAPs identified could result from mass
273 spectrometry detection limits (70, 71). We measured the level and frequency of TE expression in two
274 subsets of cTECs (Figure 6c, left) or mTECs (Figure 6c, right) using scRNA-seq data from *Baran-Gale*
275 *et al.* (72). TE subfamilies generating MAPs in cTECs or mTECs are highlighted in red in their respective
276 plots. Strikingly, TECs highly and ubiquitously expressed the MAP-generating TE subfamilies. These
277 results suggest that the contribution of TEs to the MAP repertoire of cTECs and mTECs might be
278 significantly underestimated by the limits of detection of MS. This is particularly true for mTECs because
279 they express high levels of TEs (Figure 3d), but their TE profile displays considerable intercellular
280 heterogeneity (Figure 3e and Figure 3 – figure supplement 2). Nonetheless, our data provide direct
281 evidence that TEs can generate MAPs that are presented by cTECs and mTECs and can therefore
282 contribute to thymocyte education.
283

284 **Discussion**

285 TEs are germline-integrated parasitic DNA elements that comprise about half of mammalian genomes.
286 Over evolutionary timescales, TE sequences have been co-opted for host regulatory functions.
287 Mechanistically, TEs encode proteins and noncoding RNAs that regulate gene expression at multiple
288 levels (32, 73). Regulation of IFN signaling and triggering innate sensors are the best-characterized roles

289 of TEs in the mammalian immune system (36). TEs are immunogenic and can elicit adaptive immune
290 responses implicated in autoimmune diseases (34, 36, 74, 75). Pervasive TE expression in various somatic
291 organs means that co-evolution with their host must depend on establishing immune tolerance, a concept
292 supported by the highly diversified TE repertoire expressed in mTECs (34). This observation provided
293 the impetus to perform multi-omic studies of TE expression in the thymus. At the whole organ level, we
294 found that TE expression showed extensive age- and cell lineage-related variations and was natively
295 regulated by cell proliferation and expression of KZFPs. Additionally, TEs interact with multiple TFs in
296 all thymic cell subsets. This is particularly true for the LINE and SINE subfamilies that occupy larger
297 genomic spaces. Notably, TEs appear to play particularly important roles in two cell types located in the
298 thymic medulla: mTECs and pDCs.

299

300 As mTECs are the APC population crucial to central tolerance induction, their high and diverse TE
301 expression is poised to profoundly impact the T cell repertoire's formation. The extent and complexity of
302 TF-TE interactions were higher in mTECs than in all other thymic cell subsets. These interactions included
303 *PAX1* and subunits of the NF-κB complex (e.g., *RELB*). PAX1 is essential for the development of TEC
304 progenitors (50), and *RELB* is for the development and differentiation of mTECs (76). RelB-deficient
305 mice have reduced thymic cellularity, markedly fewer mTECs, lack *Aire* expression, and suffer from
306 autoimmunity (49, 77). Under the influence of *Aire, Fezf2,* and *Chd4,* mTECs collectively express almost
307 the entire exome (61, 62). However, the expression of all genes in each mTEC would cause proteotoxic
308 stress (62). Hence, promiscuous expression of tissue-restricted genes in mTECs adopts a mosaic pattern:
309 individual tissue-restricted genes are expressed in a small fraction of mTECs (17, 78). The present work
310 shows that mTECs also express an extensive repertoire of TEs in a mosaic pattern (i.e., with considerable
311 intercellular heterogeneity). *Aire*, *Fezf2*, and *Chd4* regulate non-redundant sets of TEs and preferentially

312 induce TE sequences associated with permissive histone marks. The immunopeptidome of thymic stromal
313 cells is responsible for thymocyte education and represents one of the most fundamental "known
314 unknowns" in immunology. Inferences on the immunopeptidome of thymic stromal cells are based on
315 transcriptomic data. However, i) TCRs interact with MAPs, not transcripts, and ii) the MAP repertoire
316 cannot be inferred from the transcriptome (63, 79, 80). Using K5D1 mice presenting prominent thymic
317 hyperplasia, we conducted MS searches of TE MAPs, identifying 4 TE MAPs in mTECs and 2 in cTECs.
318 These results demonstrate that cTECs and mTECs present TE MAPs and suggest they present different
319 TE MAPs. However, the correlation between transcriptomic and immunopeptidomic data suggests that
320 TECs can present many more TE MAPs. Their profiling will require MS analyses of enormous numbers
321 of TECs or the development of more sensitive MS techniques. As TE MAPs have been detected in normal
322 and neoplastic extrathymic cells (34, 81-83), the presentation of TEs by mTECs is likely essential to
323 central tolerance. In line with vibrant plaidoyers for a collaborative Human Immunopeptidome Project
324 (64, 84), our work suggests that immunopeptidomic studies should not be limited to classical annotated
325 genes (2% of the genome) but also encompass TEs (45% of the genome).

326

327 The second population of cells exhibiting high TE expression, pDCs, are mainly seen as producers of IFN
328 α/β and potentially as APCs (21). Thymic and extrathymic pDCs are ontogenically and functionally
329 different. They develop independently from each other from different precursor cells (23, 24, 85). IFN α/β
330 secretion is inducible in extrathymic pDCs but constitutive in thymic pDCs (21, 22). In line with the
331 location of pDCs in the thymic medulla, their constitutive IFN α/β secretion is instrumental in the terminal
332 differentiation of thymocytes and the generation of Tregs and innate CD8 T cells (25-29). We report here
333 that high TE expression is also a feature of thymic, but not extrathymic, pDCs. Thus, the present study
334 provides a rationale for the constitutive IFN α/β secretion by thymic pDCs: they homogeneously express

335 large numbers of TEs (in particular LINEs and SINEs), leading to the formation of dsRNAs that trigger
336 RIG-I and MDA5 signaling that causes the constitutive secretion of IFN α/β. As such, our data suggest
337 that recognition of TE-derived dsRNAs by innate immune receptors promotes a pro-inflammatory
338 environment favorable to the establishment of central tolerance in the thymic medulla.

339

340 At first sight, the pleiotropic effects of TEs on thymic function may look surprising. It should be reminded
341 that the integration of genetic parasites such as TEs is a source of genetic conflicts with the host. Notably,
342 the emergence of adaptive immunity gave rise to higher-order conflicts between TEs and their vertebrate
343 hosts (36, 86). The crucial challenge for the immune system is developing immune tolerance towards TEs
344 to prevent autoimmune diseases that affect up to 10% of humans (87) without allowing selfish
345 retrotransposition events that hinder genome integrity. The resolution of these conflicts has been proposed
346 to be a determining factor in shaping the function of the immune system (86). Our data suggest that the
347 thymus is the central battlefield for conflict resolution between TEs and T cells in vertebrates. Consistent
348 with the implication of TEs in autoimmunity, more than 90% of putative causal variants associated with
349 autoimmune diseases are in allegedly noncoding regions of the genome (87). In this context, our study
350 illustrates the complexity of interactions between TEs and the vertebrate immune system and should
351 provide impetus to explore them further in health and disease. We see two limitations to our study. First,
352 as with all multi-omic systems immunology studies, our work provides a roadmap for many future
353 mechanistic studies that could not be realized at this stage. Second, our immunopeptidomic analyses of
354 TECs prove that TECs present TE MAPs but certainly underestimate the diversity of TE MAPs presented
355 by cTECs and mTECs.

356 **Methods**

357 **Experimental design**

358 This study aimed to understand better the impacts of TE expression on thymus development and
359 function. Thymic populations are complex and heterogeneous, so we opted for single-cell RNA-seq data
360 to draw a comprehensive profile of TE expression in the thymus. To better understand the impact of
361 AIRE, FEZF2, and CHD4 on TE expression in the mTEC(II) population, RNA-seq data from WT and
362 KO murine mTEC, as well as ChIP-seq for different histone marks in murine mTECs, were reanalyzed
363 to characterize the TE sequences regulated by these three proteins. Unless stated otherwise, studies were
364 done in human cells. For MS analyses, two replicates of 70 million cells from K5D1 mice (66) were
365 injected for both cTECs and mTECs. All experiments were in accordance with the Canadian Council on
366 Animal Care guidelines and approved by the *Comité de Déontologie de l'Expérimentation sur des*
367 *Animaux* of Université de Montréal. Primary human thymi were obtained from 4-month-old to 12-year-
368 old children undergoing cardiovascular surgeries at the CHU Sainte-Justine. This project was approved
369 by the CHU Sainte-Justine Research Ethics Board (protocol and biobank #2126).

370 **Transcriptomic data processing**

371 Preprocessing of the scRNA-seq data was performed with the kallisto and bustools workflow. For human
372 data from *Park et al.* (42), two different indexes were built for the pseudoalignment of reads with kallisto
373 (version 0.46.0) (88): one containing Ensembl 88 (GRCh38.88) transcripts used for the annotation of cell
374 populations, and a second containing Ensembl 88 transcripts and human TE sequences (LINE, LTR,
375 SINE) from RepeatMasker (89) which was used for all subsequent analyses of TE expression. For murine
376 data from *Baran-Gale et al.* (72), cell-type annotations from the original publication were used, and an
377 index containing mm10 transcripts and murine TE sequences from RepeatMasker was used to analyze TE
378 expression. The cell barcodes were corrected, and the feature-barcode matrices were generated with the
379 correct count functions of bustools (version 0.39.3) (90). For murine bulk RNA-seq data, an index

380 composed of mm10 (GRCm38) transcripts and murine TE sequences from RepeatMasker was used for
381 quantification with kallisto.

382 **ChIP-seq data reanalysis**

383 ChIP-seq data for i) ETS1 in human NK cells and ii) several histone marks of mTECs from WT mice were
384 reanalyzed (see "**Data availability**" for the complete list). For ETS1 ChIP-seq, reads were aligned to the
385 reference *Homo sapiens* genome (GRCh38) using bowtie2 (version 2.3.5) (91) with the --very-sensitive
386 and -k 10 parameters. BigWig files were generated using the bamCoverage function of deeptools2, and
387 genomic tracks were visualized in the USCS Genome Browser (92). For the murine histone marks data,
388 reads were aligned to the reference *Mus musculus* genome (mm10) using bowtie2 with the --very-sensitive
389 and -k 10 parameters. Read coverage at the sequence body and flanking regions (+/- 3000 base pairs) of
390 TE loci induced by AIRE, FEZF2, and CHD4 was visualized using ngs.plot.r (version 2.63) (93). Input
391 samples were used as a negative control.

392 **Cell population annotation**

393 Feature-barcode matrices were imported in R with SingleCellExperiment (version 1.12.0) (94). As a
394 quality control, cells with less than 2000 UMI detected, less than 500 genes detected, or more than 5%
395 reads assigned to mitochondrial genes were considered low quality and removed from the dataset with
396 scuttle (version 1.0.4) (95). Cells with more than 7000 genes detected were considered doublets and
397 removed. Normalization of cell size factors was performed with scran (version 1.18.7) (96), and log-
398 normalization of read counts was done with scuttle with default parameters. Variable regions of TCR and
399 IG genes, as well as ribosomal and cell cycle genes (based on *Park et al*. (42)), were removed, and highly
400 variable features were selected based on a mean-variance trend based on a Poisson distribution of noise
401 with scran. Adjustment of sequencing depths between batches and mutual nearest neighbors (MNN)
402 correction were computed with batchelor (version 1.6.3) (97). Cell clustering was performed with scran

403 using the Jaccard index for edge weighting and the Louvain method for community detection. Lists of
404 marker genes for human thymic cell populations and TEC subsets were taken from *Park et al.* (42) and
405 *Bautista et al.* (59), whereas marker genes of splenic populations were based on *Madissoon et al.* (60).

406 **TE expression throughout thymic development**

407 The expression of TE subfamilies was obtained by summing the read counts of loci based on the
408 RepeatMasker annotations. For each TE subfamily in each cell population, expression levels amongst
409 developmental stages were normalized by dividing them with the maximal expression value. Next, the
410 Euclidean distance between each TE subfamily in each cell population (based on their normalized
411 expression across developmental stages) was computed, followed by unsupervised hierarchical clustering.
412 The tree was then manually cut into three clusters, and enrichment of LINE, LTR, and SINE elements in
413 these three clusters was determined using Fisher's exact tests. The cluster assigned to each TE subfamily
414 in each cell population was visualized in a circos plot using the circlize package (version 0.4.14) (98) in
415 R, and the percentage of each cell population found in embryonic or postnatal samples. Finally, we
416 computed the frequency that each TE family was assigned to the three clusters, and the maximal value
417 was kept. The distributions of LINE, LTR, and SINE elements were compared to a random distribution
418 (n = 809) with Kolmogorov-Smirnov tests.

419 **Regulation of TE expression by cell proliferation and KZFPs**

420 Proliferation scores were generated for each dataset cell using the CellCycleScoring function of Seurat
421 (version 4.1.0). As per *Cowan et al.* (99), we combined previously published lists of G2M and S phase
422 marker genes (100) to compute the proliferation scores. For each thymic cell population, we calculated
423 the Spearman correlation between proliferation scores and the expression of TE subfamilies. The
424 Benjamini-Hochberg method was applied to correct for multiple comparisons. Correlations were
425 considered positive if the correlation coefficient was ≥ 0.2 and the adjusted p-value≤ 0.05, and negative if

426 the coefficient was ≤-0.2 and the adjusted p-value≤0.05. We also computed the median of all correlation
427 coefficients for each cell population. We then assigned the class of each TE subfamily correlated with cell
428 proliferation and compared this distribution to the distribution of classes of all TE subfamilies in the
429 human genome. The percentage of overlap of the sets of TE subfamilies significantly correlated with cell
430 proliferation was determined. A list of 401 human KZFPs was downloaded from *Imbeault et al.* (44).
431 Spearman correlations between KZFP and TE expression were independently computed in each cell
432 population with the same methodology as the cell proliferation analysis, and Benjamini-Hochberg
433 correction for multiple comparisons was applied. The information on the enrichment of KZFPs within TE
434 subfamilies was downloaded from *Imbeault et al.* (44). Sharing of KZFP-TE pairs between cell
435 populations was represented using the circlize package.

436 **Estimation of TE sequences' age**

437 The sequence divergence (defined as the number of mismatches per thousand) was given by the milliDiv
438 value in RepeatMasker. The milliDiv values of each TE locus were divided by the substitution rate of its
439 host's genome ($2{,}2 \times 10^{-9}$ mutation/year for *Homo sapiens* and 4.5×10^{-9} mutation/year for *Mus musculus*
440 (101, 102)). Finally, the age of each TE subfamily was determined by averaging the age of all loci of the
441 subfamily.

442 **Interactions between TE subfamilies and transcription factors**

443 We downloaded a list of 1638 transcription factors (TF) manually curated by *Lambert et al.* (103). For
444 each cell population of the thymus, Spearman correlations were computed for each possible pair of TF
445 and TE subfamily, and the Benjamini-Hochberg method was applied to correct the p-values for multiple
446 comparisons. Correlations were considered significant if i) the correlation coefficient was ≥0.2, ii) the
447 adjusted p-value was ≤0.05, and iii) the TF was expressed by ≥10% of the cells of the population. The
448 correlations were validated using a bootstrap procedure (1000 iterations) to ensure their reproducibility.

449 Briefly, we randomly selected n cells out of the n cells of a given population (while allowing cells to be
450 selected multiple times). The empirical p-value was determined by dividing the number of iterations with
451 a correlation coefficient <0.2 by the total number of iterations (1000). In parallel, the curated binding
452 motifs of 945 TFs were downloaded from the JASPAR database. We then used the *Find Individual Motif*
453 *Occurences* (FIMO) software (104) to identify the 100 000 genomic positions with the most significant
454 matches for the TF binding motif. These lists of binding motif positions were then intersected with the
455 positions of TE loci with the intersect function of BEDTools (version 2.29.2) (105), and the percentage of
456 TE loci of each subfamily harboring TF binding motifs was determined. Thus, in a specific cell population
457 of the thymus, a TF was considered as interacting with a TE subfamily if it satisfied two criteria: i) its
458 expression was correlated with the one of the TE family (spearman coefficient ≥0.2, adjusted p-value
459 ≤0.05 and expression of TF in ≥10% of cells), and ii) at least one locus of the TE subfamily contained a
460 binding motif of the TF. For each cell population, networks of interactions between TF and TE subfamilies
461 were generated with the network package (version 1.17.1) (106) in R and represented with the ggnetwork
462 package. For the sake of clarity, only the most significant interactions were illustrated for each cell type
463 (i.e., correlation coefficient ≥0.3, TF binding sites in ≥1% of the loci of the TE subfamily, and TF
464 expression in ≥10% of cells of the population). Sharing of TF-TE interactions between cell populations
465 was represented with a chord diagram using the circlize package. For each TE subfamily, the number of
466 interactions with TFs and the number of loci of the TE subfamily in the human genome were determined.
467 Wilcoxon-Mann-Whitney tests were used to compare the number of interactions with TF of LTR, LINE,
468 and SINE elements, whereas Kendall tau correlation was calculated between the number of interactions
469 with TF and the number of loci of TE subfamilies.

470 **Identification of TE promoter and enhancer candidates**

471 From the previously identified list of TF-TE interactions, we isolated the specific loci containing TF
472 binding sites from the subfamilies whose expression was positively correlated with the TF. To determine
473 if these TE loci could act as promoters or enhancers, we used histone ChIP-seq data from the ENCODE
474 consortium for H3K27ac, H3K4me1, and H3K4me3. BED files from the ENCODE consortium were
475 downloaded for eight immune cell populations: B cells, CD4 Single Positive T cells (CD4 SP), CD8 Single
476 Positive T cells (CD8 SP), dendritic cells (DC), monocytes and macrophages (Mono/Macro), NK cells,
477 Th17, and Treg. TE loci colocalizing with peaks in histone ChIP-seq data were identified using the
478 intersect function of BEDTools (version 2.29.2). To be considered as enhancer candidates, TE loci had to
479 colocalize with H3K27ac and H3K4me1, but not H3K4me3. To be considered as promoter candidates,
480 TE loci had to colocalize with H3K27ac and H3K4me3, but not H3K4me1, and be located at ≤ 1000
481 nucleotides from a transcription start site (TSS) annotated in the refTSS database (107).

482 **Diversity of TE expression**

483 The human thymic scRNA-seq dataset was subsampled to retain only postnatal cells, as it was shown by
484 *Bornstein et al.* (56) that thymic APCs are mainly found in postnatal samples. The diversity of TE
485 sequences expressed by thymic populations was assessed using Shannon entropy. Using the vegan
486 package (version 2.5-7) (108) in R, two distinct Shannon entropy metrics were computed for each cell
487 population. First, the Shannon entropy was computed based on the expression level (i.e., log(read count))
488 of TE subfamilies for each cell individually. The median entropy was calculated for each cell population.
489 In parallel, the diversity of TE sequences expressed by an entire population was also assessed. For this
490 purpose, a binary code was generated to represent the expression status of TE subfamilies in each cell
491 (where 1 is expressed and 0 is not expressed). For each population separately, the binary codes of
492 individual cells were summed to obtain the frequency of expression of each TE subfamily in the
493 population, which was used to compute the Shannon entropy of TE sequences expressed by the population.

494 A linear model was generated with the lm function of the stats package in R to summarize the data
495 distribution. The deviation (Δy) from the observed population's TE diversity and the one expected by the
496 linear model was computed for each cell population.

497 **TE expression in thymic APC**

498 A differential expression analysis of TE subfamilies between the subsets of thymic APC was performed
499 with the FindAllMarkers function with default parameters of Seurat (109) with the MAST model. Finally,
500 the heterogeneity of TE expression inside thymic APC subsets was evaluated with the MetaCell package
501 (version 0.3.5) (110). The composition of the metacells was validated based on manual annotation (see
502 the "Single-cell RNA-seq preprocessing" section), and only metacells with >50% of cells belonging to
503 the same subset of thymic APCs were kept. Differential expression of TE subfamilies between metacells
504 was performed as described above, and the percentage of overlap between the sets of TEs overexpressed
505 by the different metacells was computed.

506 **Isolation of human thymic pDCs and immunostaining of dsRNAs**

507 Primary human thymi were obtained from 4-month-old to 12-year-old children undergoing cardiovascular
508 surgeries at the CHU Sainte-Justine. This project was approved by the CHU Sainte-Justine Research
509 Ethics Board (protocol and biobank #2126). Thymi from 4-month-old to 12-year-old individuals were
510 cryopreserved in liquid nitrogen in the following solution: 95% (PBS-5% Dextran 40 (Sigma-Aldrich)) –
511 5% DMSO (Fisher Scientific). Protocol for thymic pDCs isolation was based on *Stoeckle et al.* (111).
512 Briefly, thymic samples were cut in ~2mm pieces, followed by three rounds of digestion (40 min, 180
513 RPM at 37°C) in RPMI 1640 (Gibco) supplemented with 2mg/mL of Collagenase A (Roche) and
514 0.1mg/mL of DNase I (Sigma-Aldrich). APCs were then enriched using Percoll (Sigma-Aldrich) density
515 centrifugation (3500g, 35min at 4°C), followed by an FBS cushion density gradient (5mL of RPMI 1640
516 containing enriched APCs layered on 5mL of heat-inactivated FBS (Invitrogen, 12483020), 1000RPM for

517 10min at 4°C) to remove cell debris. Finally, thymic pDCs were magnetically enriched using the
518 QuadroMACS Separator (Miltenyi). Cells were stained with a CD303 (BDCA-2) MicroBead Kit
519 (Miltenyi), and labeled cells were loaded on LS columns (Miltenyi) for magnetic-activated cell sorting.

520 Purified thymic pDCs were pipetted on poly-L-lysine (Sigma-Aldrich, 1:10 in dH_2O) coated 15µ-Slide 8
521 well (ibid) and incubated for 2h at 37°C in RPMI 1640 supplemented with 10% BSA (Sigma-Aldrich).
522 Cells were fixed using 1% [w/v] paraformaldehyde (PFA, Sigma-Aldrich) in PBS 1X (Sigma-Aldrich) for
523 30min at room temperature. Cells were permeabilized for 30min at room temperature with 0.1% [v/v]
524 Triton X-100 (Sigma-Aldrich) in PBS 1X, followed by blocking using 5% [w/v] BSA (Sigma-Aldrich) in
525 PBS 1X for 30min at room temperature. Immunostaining was performed in four steps to avoid unspecific
526 binding of the secondary antibodies: i) incubation overnight at 4°C with the mouse monoclonal IgG2a J2
527 antibody anti-dsRNA (Jena Bioscience, cat. RNT-SCI-10010500, dilution 1:200), ii) incubation with the
528 donkey anti-mouse IgG (H+L) antibody coupled to Alexa Fluor 555 (Invitrogen, cat. A-31570, dilution
529 1:500) for 30min at room temperature, iii) incubation with the mouse monoclonal IgG1 clone 6H6 anti-
530 CD123 (eBioscience, cat. 14-1239-82, 1:100) for one hour at room temperature, and iv) incubation with
531 the goat anti-mouse IgG1 polyclonal Alexa Fluor 488 antibody (Invitrogen, cat. A-21121, 1:1000) for
532 30min at room temperature. Finally, cells were stained with DAPI (Invitrogen, cat. D3571, 1:1000) for
533 5min at room temperature. All antibodies and DAPI were diluted in a blocking solution. Image acquisition
534 was made with an LSM 700 laser scanning confocal microscope (Zeiss) using a 40x oil objective (Zeiss,
535 Plan-Neofluar N.A. 1.4) and the ZEN software. Using the whiteTopHat function of the EBImage package
536 and the sigmoNormalize function of the MorphoR package in R, the background of the DAPI signal was
537 removed. The nuclei were segmented on the resulting images as circular shapes based on the DAPI signal.
538 The mean intensity of CD123 and J2 staining was determined for each cytoplasm, defined as 19nm rings
539 around nuclei. Based on the distribution of the CD123 signal across cells, a threshold between $CD123^-$

540 and CD123^{+} cells was set up for each replicate independently. J2 signal intensity was compared between
541 CD123^{-} and CD123^{+} cells using the Wilcoxon Rank Sum test in R.

542 **Gene set enrichment analysis**

543 Gene set enrichment analyses were performed to determine which biological processes are enriched in
544 mTEC(II) and pDCs. Differential gene expression analyses were performed between each possible pair of
545 thymic APCs subsets using MAST with the FindMarkers function of Seurat. The gene set enrichment
546 analysis was performed using the iDEA package (version 1.0.1) (112) in R. As per *Ma et al.* (112), the
547 fold change and standard error of gene expression were used as input for iDEA, in addition to predefined
548 lists of gene sets compiled in the iDEA package. Gene sets associated with antigen presentation, interferon
549 signaling, and immune response were manually annotated. iDEA was launched with default parameters,
550 except for the 500 iterations of the Markov chain Monte Carlo algorithm, and p-values were corrected
551 with the Louis method. We also visualized the expression of *AIRE, FEZF2,* and *CHD4* in the TEC lineage
552 to validate their expression in mTEC(II).

553 **TE loci regulated by AIRE, FEZF2, and CHD4**

554 A differential expression analysis of TE subfamilies between WT and *Aire*-, *Fezf2*-, or *Chd4*-KO mice
555 was performed with the voom method of the limma package (version 3.46.0) (113, 114). Stringent criteria
556 (i.e., an expression below 2 transcripts per million (TPM) in all samples) were applied to remove lowly
557 expressed TEs. TE subfamilies with i) a fold change ≥ 2 and an adjusted p-value ≤ 0.05 or ii) a fold change
558 ≤ -2 and an adjusted p-value ≤ 0.05 were considered as induced and repressed, respectively. The percentage
559 of overlap between the sets of TE loci induced or repressed by AIRE, FEZF2, and CHD4 was computed.
560 The class and subfamily were assigned to each regulated TE locus, and the distributions of classes and
561 subfamilies across all TE sequences of the murine genome were used as controls. Significant enrichment
562 of classes or subfamilies was determined with Chi-squared tests, and a Bonferroni correction for multiple

563 comparisons was performed to enrich subfamilies in induced or repressed TEs. The distance between TE
564 loci induced or repressed by AIRE, FEZF2, or CHD4 was defined as the minimal distance between the
565 middle position of TE loci on the same chromosome. As a control, distributions of randomly selected TE
566 loci whose expression is independent of AIRE, FEZF2, and CHD4 and equal size to the sets of regulated
567 TEs were generated (for example, if 433 TE loci are induced by CHD4, 433 independent TE loci were
568 randomly selected). Wilcoxon rank-sum tests were used to compare random and regulated distributions.
569 Genomic positions of exons, introns 3' and 5' untranslated transcribed region (UTR) were downloaded
570 from the UCSC Table Browser. The genomic localization of regulated TEs was determined using the
571 intersect mode of the BEDTools suite version 2.29.2. TE loci not located in exons, introns, 3'UTR, or
572 5'UTR were considered intergenic. The percentage of regulated TE loci in each type of genomic region
573 was determined and compared to the genomic localization of all TE loci in the murine genome with chi-
574 squared tests. Finally, we estimated the frequency of intron retention events for introns containing TE loci
575 regulated by AIRE, FEZF2, or CHD4 with S-IRFindeR (115). Sequencing reads were aligned to the
576 reference *Mus musculus* genome (mm10) using STAR version 2.7.1a (116) with default parameters. Each
577 intron's Stable Intron Retention ratio (SIRratio) was computed with the computeSIRratio function of S-
578 IRFindeR. Introns containing TE loci induced by AIRE, FEZF2, or CHD4 were filtered using BEDTools
579 intersect. Random distributions of equivalent sizes of introns containing TE sequences independent of
580 AIRE, FEZF2, and CHD4 were generated as control. A SIRratio of 0.1 was used as a threshold of
581 significant intron retention events.

582 **Enzymatic digestion and isolation of murine TECs**

583 Thymic stromal cell enrichment was performed as previously described (69, 117). Briefly, thymi from 16-
584 to 22-week-old K5D1 mice were mechanically disrupted and enzymatically digested with papain
585 (Worthington Biochemical Corporation), DNase I (Sigma-Aldrich), and collagenase IV (Sigma-Aldrich)

586 at 37°C. Next, the single-cell suspension obtained after enzymatic digestion was maintained at 4°C in
587 FACS buffer (PBS, 0.5% [w/v] BSA, 2mM EDTA) and enriched in thymic epithelial cells using anti-
588 EpCAM (CD326) or anti-CD45 microbeads (mouse, Miltenyi) and LS columns (Miltenyi). Then, the
589 enriched epithelial cell suspension was stained for flow cytometry cell sorting with the following
590 antibodies and dyes: anti-EpCAM-APC-Cy7 clone G8.8 (BioLegend, cat. 118218), anti-CD45-APC clone
591 30-F11 (BD Biosciences, cat. 559864), anti-UEA1-biotinylated (Vector Laboratories, cat. B-1065), anti-
592 I-A/I-E-Alexa Fluor 700 clone M5/114.15.2 (BioLegend, cat. 107622), anti-Ly51-FITC clone 6C3
593 (BioLegend, cat. 553160), anti-streptavidin-PE-Cy7 (BD Biosciences, cat. 557598), and 7-AAD (BD
594 Biosciences, cat. 559925). Cell sorting was performed using a BD FACSAria (BD Biosciences), and data
595 were analyzed using the FACSDiva. TECs were defined as $EpCAM^{+}CD45^{-}$, while the cTEC and mTEC
596 subsets were defined as $UEA1^{-}Ly51^{+}$ and $UEA1^{+}Ly51^{-}$ TEC, respectively.

597 **RNA-Sequencing**

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

603 **Preparation of CNBR-activated Sepharose beads for MHC I immunoprecipitation**

604 CNBR-activated Sepharose 4B beads (Sigma-Aldrich, cat. 17-0430-01) were incubated with 1 mM HCl
605 at a ratio of 40 mg of beads per 13.5 ml of 1 mM HCl for 30 minutes with tumbling at room temperature.
606 Beads were spun at 215g for 1 minute at 4°C, and supernatants were discarded. 40 mg of beads were
607 resuspended with 4 ml of coupling buffer (0.1M $NaHC0_3$/0.5M NaCl pH 8.3), spun at 215g for 1 minute
608 at 4°C, and the supernatants were discarded. Mouse antibodies Pan-H2 (clone M1/42), H2-K^b (clone Y-

609 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
610 coupling buffer for 120 minutes with tumbling at room temperature. Beads were spun at 215g for 1 minute
611 at 4°C, and supernatants were discarded. 40 mg of beads were resuspended with 1 ml of blocking buffer
612 (0.2M glycine), incubated for 30 minutes with tumbling at room temperature, and the supernatants were
613 discarded. Beads were washed by centrifugation twice with PBS pH 7.2, resuspended at a concentration
614 of 1 mg of antibody per ml of PBS pH 7.2, and stored at 4°C.

615 **Immuno-isolation of MAPs**

616 Frozen pellets of mTECs (90 mice, 191 million cells total) and cTECs (121 mice, 164 million cells total)
617 were thawed, pooled, and resuspended with PBS pH 7.2 up to 4 ml and then solubilized by adding 4 mL
618 of detergent buffer containing PBS pH 7.2, 1% (w/v) CHAPS (Sigma, cat. C9426-5G) supplemented with
619 Protease inhibitor cocktail (Sigma, cat. P8340-5mL). Solubilized cells were incubated for 60 minutes with
620 tumbling at 4°C and then spun at 16,600g for 20 minutes at 4°C. Supernatants were transferred into new
621 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-
622 linked CNBR-Sepharose beads per sample and incubated with tumbling for 180 minutes at 4°C. Samples
623 were transferred into BioRad Poly prep chromatography columns and eluted by gravity. Beads were first
624 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
625 complexes were eluted from the beads by acidic treatment using 1% trifluoroacetic acid (TFA). Acidic
626 filtrates containing peptides were separated from MHC I subunits (HLA molecules and β-2 microglobulin)
627 using home-made stage tips packed with two 1 mm diameter octadecyl (C-18) solid phase extraction disks
628 (EMPORE). Stage tips were pre-washed with methanol, then with 80% acetonitrile (ACN) in 0.1% TFA,
629 and finally with 1% TFA. Samples were loaded onto the stage tips and washed with 1%TFA and 0.1%
630 TFA. Peptides were eluted with 30% ACN in 0.1% TFA, dried using vacuum centrifugation, and then
631 stored at -20°C until MS analysis.

632 **MS analyses**

633 Peptides were loaded and separated on a home-made reversed-phase column (150-µm i.d. by 200 mm)
634 with a 106-min gradient from 10 to 38% B (A: formic acid 0.1%, B: 80% CAN 0.1% formic acid) and a
635 600-nl/min flow rate on an Easy nLC-1200 connected to an Orbitrap Exploris 480 (Thermo Fisher
636 Scientific). Each full MS spectrum acquired at a resolution of 240,000 was followed by tandem-MS (MS-
637 MS) spectra acquisition on the most abundant multiply charged precursor ions for a maximum of 3s.
638 Tandem-MS experiments were performed using higher energy collision-induced dissociation (HCD) at a
639 collision energy of 34%. The generation of the personalized proteome containing TE sequences, as well
640 as the identification of TE-derived MAPs, was performed as per *Larouche et al.* (34) with the following
641 modifications: the mm10 murine reference genome was downloaded from the UCSC Genome Browser,
642 the annotations for murine genes and TE sequences were downloaded from the UCSC Table Browser, and
643 the Uniprot mouse database (16 977 entries) was used for the canonical proteome. MAPs were identified
644 using PEAKS X Pro (Bioinformatics Solutions, Waterloo, ON). The level and frequency of expression of
645 TE subfamilies generating MAPs or not were determined in thymic epithelial cells were determined by
646 averaging the expression values across cells of a TEC subset and dividing the number of cells with a
647 positive (i.e.,> 0) expression of the TEs by the total number of cells of the TEC subset, respectively.

648 **Availability of data and materials:**

649 scRNA-seq data of human thymi and spleen were downloaded from ArrayExpress (accession number E-
650 MTAB-8581) and the NCBI BIOPROJECT (accession code PRJEB31843), respectively. scRNA-seq data
651 of murine thymi were downloaded from ArrayExpress (accession number E-MTAB-8560). RNA-seq data
652 from WT, Aire-KO, Fezf2-KO, and Chd4-KO murine mTECs were downloaded from the Gene
653 Expression Omnibus (GEO) under the accession code GSE144880. ChIP-seq data for different histone
654 marks in murine mTECs were also downloaded from GEO: H3K4me3 for mTECs (GSE53111);

655 H3K4me1 and H3K27ac from MHCII^{hi} mTECs (GSE92597); H3K4me2 in mTEC-II (GSE103969); and
656 H3K4ac and H3K9ac in mTECs (GSE114713). Transcriptomic and immunopeptidomic data of K5D1
657 mice mTECs and cTECs generated in this study are available on the Gene Expression Omnibus (GEO)
658 under the accession GSE232011 and on the Proteomics Identification Database (PRIDE) under the
659 accession PXD042241, respectively.

660 **Author contributions:**

661 Conceptualization: JDL, CML, AT, KV, CP

662 Methodology: JDL, CML, AT, KV

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

664 Visualization: JDL, CML, AT

665 Funding acquisition: JDL, CP

666 Project administration: JDL, KV, CP

667 Supervision: SL, PT, CP

668 Writing – original draft: JDL, KV, CP

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

671 **Acknowledgments:**

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

677

678 **References**

679 1. Boehm T. Evolution of vertebrate immunity. Curr Biol. 2012;22(17):R722-32.
680 2. Boehm T, Swann JB. Origin and evolution of adaptive immunity. Annu Rev Anim Biosci.
681 2014;2:259-83.
682 3. Suo C, Dann E, Goh I, Jardine L, Kleshchevnikov V, Park JE, et al. Mapping the developing
683 human immune system across organs. Science. 2022;376(6597):eabo0510.
684 4. Terra R, Louis I, Le Blanc R, Ouellet S, Zuniga-Pflucker JC, Perreault C. T-cell generation by
685 lymph node resident progenitor cells. Blood. 2005;106(1):193-200.
686 5. Blais ME, Brochu S, Giroux M, Belanger MP, Dulude G, Sekaly RP, et al. Why T cells of
687 thymic versus extrathymic origin are functionally different. J Immunol. 2008;180(4):2299-312.
688 6. Zuniga-Pflucker JC, Longo DL, Kruisbeek AM. Positive selection of CD4-CD8+ T cells in the
689 thymus of normal mice. Nature. 1989;338(6210):76-8.
690 7. Breed ER, Lee ST, Hogquist KA. Directing T cell fate: How thymic antigen presenting cells
691 coordinate thymocyte selection. Semin Cell Dev Biol. 2018;84:2-10.
692 8. Dervovic D, Zuniga-Pflucker JC. Positive selection of T cells, an in vitro view. Semin Immunol.
693 2010;22(5):276-86.
694 9. Lebel ME, Coutelier M, Galipeau M, Kleinman CL, Moon JJ, Melichar HJ. Differential
695 expression of tissue-restricted antigens among mTEC is associated with distinct autoreactive T cell fates.
696 Nat Commun. 2020;11(1):3734.
697 10. Srinivasan J, Lancaster JN, Singarapu N, Hale LP, Ehrlich LIR, Richie ER. Age-Related
698 Changes in Thymic Central Tolerance. Front Immunol. 2021;12:676236.
699 11. Cheng M, Anderson MS. Thymic tolerance as a key brake on autoimmunity. Nat Immunol.
700 2018;19(7):659-64.
701 12. Perera J, Zheng Z, Li S, Gudjonson H, Kalinina O, Benichou JIC, et al. Self-Antigen-Driven
702 Thymic B Cell Class Switching Promotes T Cell Central Tolerance. Cell Rep. 2016;17(2):387-98.
703 13. Nitta T, Ohigashi I, Nakagawa Y, Takahama Y. Cytokine crosstalk for thymic medulla
704 formation. Curr Opin Immunol. 2011;23(2):190-7.
705 14. Malhotra D, Linehan JL, Dileepan T, Lee YJ, Purtha WE, Lu JV, et al. Tolerance is established
706 in polyclonal CD4(+) T cells by distinct mechanisms, according to self-peptide expression patterns. Nat
707 Immunol. 2016;17(2):187-95.
708 15. Lkhagvasuren E, Sakata M, Ohigashi I, Takahama Y. Lymphotoxin beta receptor regulates the
709 development of CCL21-expressing subset of postnatal medullary thymic epithelial cells. J Immunol.
710 2013;190(10):5110-7.
711 16. Laan M, Salumets A, Klein A, Reintamm K, Bichele R, Peterson H, et al. Post-Aire Medullary
712 Thymic Epithelial Cells and Hassall's Corpuscles as Inducers of Tonic Pro-Inflammatory
713 Microenvironment. Front Immunol. 2021;12:635569.
714 17. Michelson DA, Hase K, Kaisho T, Benoist C, Mathis D. Thymic epithelial cells co-opt lineage-
715 defining transcription factors to eliminate autoreactive T cells. Cell. 2022.
716 18. Ramsey C, Winqvist O, Puhakka L, Halonen M, Moro A, Kampe O, et al. Aire deficient mice
717 develop multiple features of APECED phenotype and show altered immune response. Hum Mol Genet.
718 2002;11(4):397-409.
719 19. Takaba H, Morishita Y, Tomofuji Y, Danks L, Nitta T, Komatsu N, et al. Fezf2 Orchestrates a
720 Thymic Program of Self-Antigen Expression for Immune Tolerance. Cell. 2015;163(4):975-87.

721 20. Tomofuji Y, Takaba H, Suzuki HI, Benlaribi R, Martinez CDP, Abe Y, et al. Chd4 choreographs
722 self-antigen expression for central immune tolerance. Nat Immunol. 2020;21(8):892-901.
723 21. Ginhoux F, Guilliams M, Merad M. Expanding dendritic cell nomenclature in the single-cell era.
724 Nat Rev Immunol. 2022;22(2):67-8.
725 22. Colantonio AD, Epeldegui M, Jesiak M, Jachimowski L, Blom B, Uittenbogaart CH. IFN-alpha
726 is constitutively expressed in the human thymus, but not in peripheral lymphoid organs. PLoS One.
727 2011;6(8):e24252.
728 23. Lavaert M, Liang KL, Vandamme N, Park JE, Roels J, Kowalczyk MS, et al. Integrated scRNA-
729 Seq Identifies Human Postnatal Thymus Seeding Progenitors and Regulatory Dynamics of
730 Differentiating Immature Thymocytes. Immunity. 2020;52(6):1088-104 e6.
731 24. Le J, Park JE, Ha VL, Luong A, Branciamore S, Rodin AS, et al. Single-Cell RNA-Seq Mapping
732 of Human Thymopoiesis Reveals Lineage Specification Trajectories and a Commitment Spectrum in T
733 Cell Development. Immunity. 2020;52(6):1105-18 e9.
734 25. Xing Y, Wang X, Jameson SC, Hogquist KA. Late stages of T cell maturation in the thymus
735 involve NF-kappaB and tonic type I interferon signaling. Nat Immunol. 2016;17(5):565-73.
736 26. Hanabuchi S, Ito T, Park WR, Watanabe N, Shaw JL, Roman E, et al. Thymic stromal
737 lymphopoietin-activated plasmacytoid dendritic cells induce the generation of FOXP3+ regulatory T
738 cells in human thymus. J Immunol. 2010;184(6):2999-3007.
739 27. Martin-Gayo E, Sierra-Filardi E, Corbi AL, Toribio ML. Plasmacytoid dendritic cells resident in
740 human thymus drive natural Treg cell development. Blood. 2010;115(26):5366-75.
741 28. Martinet V, Tonon S, Torres D, Azouz A, Nguyen M, Kohler A, et al. Type I interferons regulate
742 eomesodermin expression and the development of unconventional memory CD8(+) T cells. Nat
743 Commun. 2015;6:7089.
744 29. Epeldegui M, Blom B, Uittenbogaart CH. BST2/Tetherin is constitutively expressed on human
745 thymocytes with the phenotype and function of Treg cells. Eur J Immunol. 2015;45(3):728-37.
746 30. Treangen TJ, Salzberg SL. Repetitive DNA and next-generation sequencing: computational
747 challenges and solutions. Nat Rev Genet. 2011;13(1):36-46.
748 31. Deniz O, Frost JM, Branco MR. Regulation of transposable elements by DNA modifications.
749 Nat Rev Genet. 2019;20(7):417-31.
750 32. Bourque G, Burns KH, Gehring M, Gorbunova V, Seluanov A, Hammell M, et al. Ten things
751 you should know about transposable elements. Genome Biol. 2018;19(1):199.
752 33. Argueso JL, Westmoreland J, Mieczkowski PA, Gawel M, Petes TD, Resnick MA. Double-
753 strand breaks associated with repetitive DNA can reshape the genome. Proc Natl Acad Sci U S A.
754 2008;105(33):11845-50.
755 34. Larouche JD, Trofimov A, Hesnard L, Ehx G, Zhao Q, Vincent K, et al. Widespread and tissue-
756 specific expression of endogenous retroelements in human somatic tissues. Genome Med.
757 2020;12(1):40.
758 35. Carter JA, Stromich L, Peacey M, Chapin SR, Velten L, Steinmetz LM, et al. Transcriptomic
759 diversity in human medullary thymic epithelial cells. Nat Commun. 2022;13(1):4296.
760 36. Kassiotis G. The Immunological Conundrum of Endogenous Retroelements. Annu Rev
761 Immunol. 2023;41:99-125.
762 37. Chuong EB, Elde NC, Feschotte C. Regulatory evolution of innate immunity through co-option
763 of endogenous retroviruses. Science. 2016;351(6277):1083-7.
764 38. Bogdan L, Barreiro L, Bourque G. Transposable elements have contributed human regulatory
765 regions that are activated upon bacterial infection. Philos Trans R Soc Lond B Biol Sci.
766 2020;375(1795):20190332.

767 39. Adoue V, Binet B, Malbec A, Fourquet J, Romagnoli P, van Meerwijk JPM, et al. The Histone
768 Methyltransferase SETDB1 Controls T Helper Cell Lineage Integrity by Repressing Endogenous
769 Retroviruses. Immunity. 2019;50(3):629-44 e8.
770 40. Lefkopoulos S, Polyzou A, Derecka M, Bergo V, Clapes T, Cauchy P, et al. Repetitive Elements
771 Trigger RIG-I-like Receptor Signaling that Regulates the Emergence of Hematopoietic Stem and
772 Progenitor Cells. Immunity. 2020;53(5):934-51 e9.
773 41. Lima-Junior DS, Krishnamurthy SR, Bouladoux N, Collins N, Han SJ, Chen EY, et al.
774 Endogenous retroviruses promote homeostatic and inflammatory responses to the microbiota. Cell.
775 2021.
776 42. Park JE, Botting RA, Dominguez Conde C, Popescu DM, Lavaert M, Kunz DJ, et al. A cell atlas
777 of human thymic development defines T cell repertoire formation. Science. 2020;367(6480).
778 43. Brocks D, Chomsky E, Mukamel Z, Lifshitz A, Tanay A. Single cell analysis reveals dynamics
779 of transposable element transcription following epigenetic de-repression. bioRxiv. 2018:462853.
780 44. Imbeault M, Helleboid PY, Trono D. KRAB zinc-finger proteins contribute to the evolution of
781 gene regulatory networks. Nature. 2017;543(7646):550-4.
782 45. Kunarso G, Chia NY, Jeyakani J, Hwang C, Lu X, Chan YS, et al. Transposable elements have
783 rewired the core regulatory network of human embryonic stem cells. Nat Genet. 2010;42(7):631-4.
784 46. Sundaram V, Cheng Y, Ma Z, Li D, Xing X, Edge P, et al. Widespread contribution of
785 transposable elements to the innovation of gene regulatory networks. Genome Res. 2014;24(12):1963-
786 76.
787 47. Hosokawa H, Rothenberg EV. How transcription factors drive choice of the T cell fate. Nat Rev
788 Immunol. 2021;21(3):162-76.
789 48. Baik S, Sekai M, Hamazaki Y, Jenkinson WE, Anderson G. Relb acts downstream of medullary
790 thymic epithelial stem cells and is essential for the emergence of RANK(+) medullary epithelial
791 progenitors. Eur J Immunol. 2016;46(4):857-62.
792 49. Akiyama T, Shimo Y, Yanai H, Qin J, Ohshima D, Maruyama Y, et al. The tumor necrosis factor
793 family receptors RANK and CD40 cooperatively establish the thymic medullary microenvironment and
794 self-tolerance. Immunity. 2008;29(3):423-37.
795 50. Yamazaki Y, Urrutia R, Franco LM, Giliani S, Zhang K, Alazami AM, et al. PAX1 is essential
796 for development and function of the human thymus. Sci Immunol. 2020;5(44).
797 51. Meixner A, Karreth F, Kenner L, Wagner EF. JunD regulates lymphocyte proliferation and T
798 helper cell cytokine expression. EMBO J. 2004;23(6):1325-35.
799 52. Yu Y, Wang J, Khaled W, Burke S, Li P, Chen X, et al. Bcl11a is essential for lymphoid
800 development and negatively regulates p53. J Exp Med. 2012;209(13):2467-83.
801 53. Cisse B, Caton ML, Lehner M, Maeda T, Scheu S, Locksley R, et al. Transcription factor E2-2 is
802 an essential and specific regulator of plasmacytoid dendritic cell development. Cell. 2008;135(1):37-48.
803 54. Choudhary MN, Friedman RZ, Wang JT, Jang HS, Zhuo X, Wang T. Co-opted transposons help
804 perpetuate conserved higher-order chromosomal structures. Genome Biol. 2020;21(1):16.
805 55. Consortium EP. An integrated encyclopedia of DNA elements in the human genome. Nature.
806 2012;489(7414):57-74.
807 56. Luo Y, Hitz BC, Gabdank I, Hilton JA, Kagda MS, Lam B, et al. New developments on the
808 Encyclopedia of DNA Elements (ENCODE) data portal. Nucleic Acids Res. 2020;48(D1):D882-D9.
809 57. Taveirne S, Wahlen S, Van Loocke W, Kiekens L, Persyn E, Van Ammel E, et al. The
810 transcription factor ETS1 is an important regulator of human NK cell development and terminal
811 differentiation. Blood. 2020;136(3):288-98.

812 58. Bornstein C, Nevo S, Giladi A, Kadouri N, Pouzolles M, Gerbe F, et al. Single-cell mapping of
813 the thymic stroma identifies IL-25-producing tuft epithelial cells. Nature. 2018;559(7715):622-6.
814 59. Bautista JL, Cramer NT, Miller CN, Chavez J, Berrios DI, Byrnes LE, et al. Single-cell
815 transcriptional profiling of human thymic stroma uncovers novel cellular heterogeneity in the thymic
816 medulla. Nat Commun. 2021;12(1):1096.
817 60. Madissoon E, Wilbrey-Clark A, Miragaia RJ, Saeb-Parsy K, Mahbubani KT, Georgakopoulos N,
818 et al. scRNA-seq assessment of the human lung, spleen, and esophagus tissue stability after cold
819 preservation. Genome Biol. 2019;21(1):1.
820 61. Sansom SN, Shikama-Dorn N, Zhanybekova S, Nusspaumer G, Macaulay IC, Deadman ME, et
821 al. Population and single-cell genomics reveal the Aire dependency, relief from Polycomb silencing, and
822 distribution of self-antigen expression in thymic epithelia. Genome Res. 2014;24(12):1918-31.
823 62. St-Pierre C, Morgand E, Benhammadi M, Rouette A, Hardy MP, Gaboury L, et al.
824 Immunoproteasomes Control the Homeostasis of Medullary Thymic Epithelial Cells by Alleviating
825 Proteotoxic Stress. Cell Rep. 2017;21(9):2558-70.
826 63. Shapiro IE, Bassani-Sternberg M. The impact of immunopeptidomics: From basic research to
827 clinical implementation. Semin Immunol. 2023;66:101727.
828 64. Vizcaino JA, Kubiniok P, Kovalchik KA, Ma Q, Duquette JD, Mongrain I, et al. The Human
829 Immunopeptidome Project: A Roadmap to Predict and Treat Immune Diseases. Mol Cell Proteomics.
830 2020;19(1):31-49.
831 65. Kubiniok P, Marcu A, Bichmann L, Kuchenbecker L, Schuster H, Hamelin DJ, et al.
832 Understanding the constitutive presentation of MHC class I immunopeptidomes in primary tissues.
833 iScience. 2022;25(2):103768.
834 66. Robles AI, Larcher F, Whalin RB, Murillas R, Richie E, Gimenez-Conti IB, et al. Expression of
835 cyclin D1 in epithelial tissues of transgenic mice results in epidermal hyperproliferation and severe
836 thymic hyperplasia. Proc Natl Acad Sci U S A. 1996;93(15):7634-8.
837 67. Klug DB, Crouch E, Carter C, Coghlan L, Conti CJ, Richie ER. Transgenic expression of cyclin
838 D1 in thymic epithelial precursors promotes epithelial and T cell development. J Immunol.
839 2000;164(4):1881-8.
840 68. Ohigashi I, Tanaka Y, Kondo K, Fujimori S, Kondo H, Palin AC, et al. Trans-omics Impact of
841 Thymoproteasome in Cortical Thymic Epithelial Cells. Cell Rep. 2019;29(9):2901-16 e6.
842 69. Dumont-Lagace M, Daouda T, Depoers L, Zumer J, Benslimane Y, Brochu S, et al. Qualitative
843 Changes in Cortical Thymic Epithelial Cells Drive Postpartum Thymic Regeneration. Front Immunol.
844 2019;10:3118.
845 70. Ghosh M, Gauger M, Marcu A, Nelde A, Denk M, Schuster H, et al. Guidance Document:
846 Validation of a High-Performance Liquid Chromatography-Tandem Mass Spectrometry
847 Immunopeptidomics Assay for the Identification of HLA Class I Ligands Suitable for Pharmaceutical
848 Therapies. Mol Cell Proteomics. 2020;19(3):432-43.
849 71. Nanaware PP, Jurewicz MM, Clement CC, Lu L, Santambrogio L, Stern LJ. Distinguishing
850 Signal From Noise in Immunopeptidome Studies of Limiting-Abundance Biological Samples: Peptides
851 Presented by I-A(b) in C57BL/6 Mouse Thymus. Front Immunol. 2021;12:658601.
852 72. Baran-Gale J, Morgan MD, Maio S, Dhalla F, Calvo-Asensio I, Deadman ME, et al. Ageing
853 compromises mouse thymus function and remodels epithelial cell differentiation. Elife. 2020;9.
854 73. Frank JA, Feschotte C. Co-option of endogenous viral sequences for host cell function. Curr
855 Opin Virol. 2017;25:81-9.
856 74. Groger V, Cynis H. Human Endogenous Retroviruses and Their Putative Role in the
857 Development of Autoimmune Disorders Such as Multiple Sclerosis. Front Microbiol. 2018;9:265.

858 75. Volkman HE, Stetson DB. The enemy within: endogenous retroelements and autoimmune
859 disease. Nat Immunol. 2014;15(5):415-22.
860 76. Mouri Y, Nishijima H, Kawano H, Hirota F, Sakaguchi N, Morimoto J, et al. NF-kappaB-
861 inducing kinase in thymic stroma establishes central tolerance by orchestrating cross-talk with not only
862 thymocytes but also dendritic cells. J Immunol. 2014;193(9):4356-67.
863 77. O'Sullivan BJ, Yekollu S, Ruscher R, Mehdi AM, Maradana MR, Chidgey AP, et al.
864 Autoimmune-Mediated Thymic Atrophy Is Accelerated but Reversible in RelB-Deficient Mice. Front
865 Immunol. 2018;9:1092.
866 78. Klein L, Kyewski B, Allen PM, Hogquist KA. Positive and negative selection of the T cell
867 repertoire: what thymocytes see (and don't see). Nat Rev Immunol. 2014;14(6):377-91.
868 79. Caron E, Vincent K, Fortier MH, Laverdure JP, Bramoulle A, Hardy MP, et al. The MHC I
869 immunopeptidome conveys to the cell surface an integrative view of cellular regulation. Mol Syst Biol.
870 2011;7:533.
871 80. Admon A. The biogenesis of the immunopeptidome. Semin Immunol. 2023;67:101766.
872 81. Laumont CM, Vincent K, Hesnard L, Audemard E, Bonneil E, Laverdure JP, et al. Noncoding
873 regions are the main source of targetable tumor-specific antigens. Sci Transl Med. 2018;10(470).
874 82. Burbage M, Rocanin-Arjo A, Baudon B, Arribas YA, Merlotti A, Rookhuizen DC, et al.
875 Epigenetically controlled tumor antigens derived from splice junctions between exons and transposable
876 elements. Sci Immunol. 2023;8(80):eabm6360.
877 83. Shah NM, Jang HJ, Liang Y, Maeng JH, Tzeng SC, Wu A, et al. Pan-cancer analysis identifies
878 tumor-specific antigens derived from transposable elements. Nat Genet. 2023;55(4):631-9.
879 84. Shao W, Pedrioli PGA, Wolski W, Scurtescu C, Schmid E, Vizcaino JA, et al. The SysteMHC
880 Atlas project. Nucleic Acids Res. 2018;46(D1):D1237-D47.
881 85. Weijer K, Uittenbogaart CH, Voordouw A, Couwenberg F, Seppen J, Blom B, et al. Intrathymic
882 and extrathymic development of human plasmacytoid dendritic cell precursors in vivo. Blood.
883 2002;99(8):2752-9.
884 86. Boehm T, Morimoto R, Trancoso I, Aleksandrova N. Genetic conflicts and the origin of
885 self/nonself-discrimination in the vertebrate immune system. Trends Immunol. 2023;44(5):372-83.
886 87. Harroud A, Hafler DA. Common genetic factors among autoimmune diseases. Science.
887 2023;380(6644):485-90.
888 88. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification.
889 Nat Biotechnol. 2016;34(5):525-7.
890 89. Smit A, Hubley, R & Green, P. RepeatMasker Open-4.0. 2013-2015.
891 90. Melsted P, Booeshaghi AS, Gao F, Beltrame E, Lu L, Hjorleifsson KE, et al. Modular and
892 efficient pre-processing of single-cell RNA-seq. bioRxiv. 2019:673285.
893 91. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods.
894 2012;9(4):357-9.
895 92. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The human genome
896 browser at UCSC. Genome Res. 2002;12(6):996-1006.
897 93. Shen L, Shao N, Liu X, Nestler E. ngs.plot: Quick mining and visualization of next-generation
898 sequencing data by integrating genomic databases. BMC Genomics. 2014;15:284.
899 94. Amezquita RA, Lun ATL, Becht E, Carey VJ, Carpp LN, Geistlinger L, et al. Orchestrating
900 single-cell analysis with Bioconductor. Nat Methods. 2020;17(2):137-45.
901 95. McCarthy DJ, Campbell KR, Lun AT, Wills QF. Scater: pre-processing, quality control,
902 normalization and visualization of single-cell RNA-seq data in R. Bioinformatics. 2017;33(8):1179-86.

903 96. Lun AT, McCarthy DJ, Marioni JC. A step-by-step workflow for low-level analysis of single-
904 cell RNA-seq data with Bioconductor. F1000Res. 2016;5:2122.
905 97. Haghverdi L, Lun ATL, Morgan MD, Marioni JC. Batch effects in single-cell RNA-sequencing
906 data are corrected by matching mutual nearest neighbors. Nat Biotechnol. 2018;36(5):421-7.
907 98. Gu Z, Gu L, Eils R, Schlesner M, Brors B. circlize Implements and enhances circular
908 visualization in R. Bioinformatics. 2014;30(19):2811-2.
909 99. Cowan JE, Malin J, Zhao Y, Seedhom MO, Harly C, Ohigashi I, et al. Myc controls a distinct
910 transcriptional program in fetal thymic epithelial cells that determines thymus growth. Nat Commun.
911 2019;10(1):5498.
912 100. Kowalczyk MS, Tirosh I, Heckl D, Rao TN, Dixit A, Haas BJ, et al. Single-cell RNA-seq reveals
913 changes in cell cycle and differentiation programs upon aging of hematopoietic stem cells. Genome Res.
914 2015;25(12):1860-72.
915 101. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and
916 analysis of the human genome. Nature. 2001;409(6822):860-921.
917 102. Mouse Genome Sequencing C, Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, et
918 al. Initial sequencing and comparative analysis of the mouse genome. Nature. 2002;420(6915):520-62.
919 103. Lambert SA, Jolma A, Campitelli LF, Das PK, Yin Y, Albu M, et al. The Human Transcription
920 Factors. Cell. 2018;172(4):650-65.
921 104. Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif.
922 Bioinformatics. 2011;27(7):1017-8.
923 105. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features.
924 Bioinformatics. 2010;26(6):841-2.
925 106. Butts CT. network: A Package for Managing Relational Data in R. Journal of Statistical
926 Software. 2008;24(2):1 - 36.
927 107. Abugessaisa I, Noguchi S, Hasegawa A, Kondo A, Kawaji H, Carninci P, et al. refTSS: A
928 Reference Data Set for Human and Mouse Transcription Start Sites. J Mol Biol. 2019;431(13):2407-22.
929 108. Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D, et al. vegan: Community
930 Ecology Package. 2020;https://CRAN.R-project.org/package=vegan.
931 109. Satija R, Farrell JA, Gennert D, Schier AF, Regev A. Spatial reconstruction of single-cell gene
932 expression data. Nat Biotechnol. 2015;33(5):495-502.
933 110. Baran Y, Bercovich A, Sebe-Pedros A, Lubling Y, Giladi A, Chomsky E, et al. MetaCell:
934 analysis of single-cell RNA-seq data using K-nn graph partitions. Genome Biol. 2019;20(1):206.
935 111. Stoeckle C, Rota IA, Tolosa E, Haller C, Melms A, Adamopoulou E. Isolation of myeloid
936 dendritic cells and epithelial cells from human thymus. J Vis Exp. 2013(79):e50951.
937 112. Ma Y, Sun S, Shang X, Keller ET, Chen M, Zhou X. Integrative differential expression and gene
938 set enrichment analysis using summary statistics for scRNA-seq studies. Nat Commun.
939 2020;11(1):1585.
940 113. Law CW, Chen Y, Shi W, Smyth GK. voom: Precision weights unlock linear model analysis
941 tools for RNA-seq read counts. Genome Biol. 2014;15(2):R29.
942 114. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential
943 expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43(7):e47.
944 115. Broseus L, Ritchie W. S-IRFindeR: stable and accurate measurement of intron retention.
945 bioRxiv. 2020:2020.06.25.164699.
946 116. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal
947 RNA-seq aligner. Bioinformatics. 2013;29(1):15-21.

117. 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.

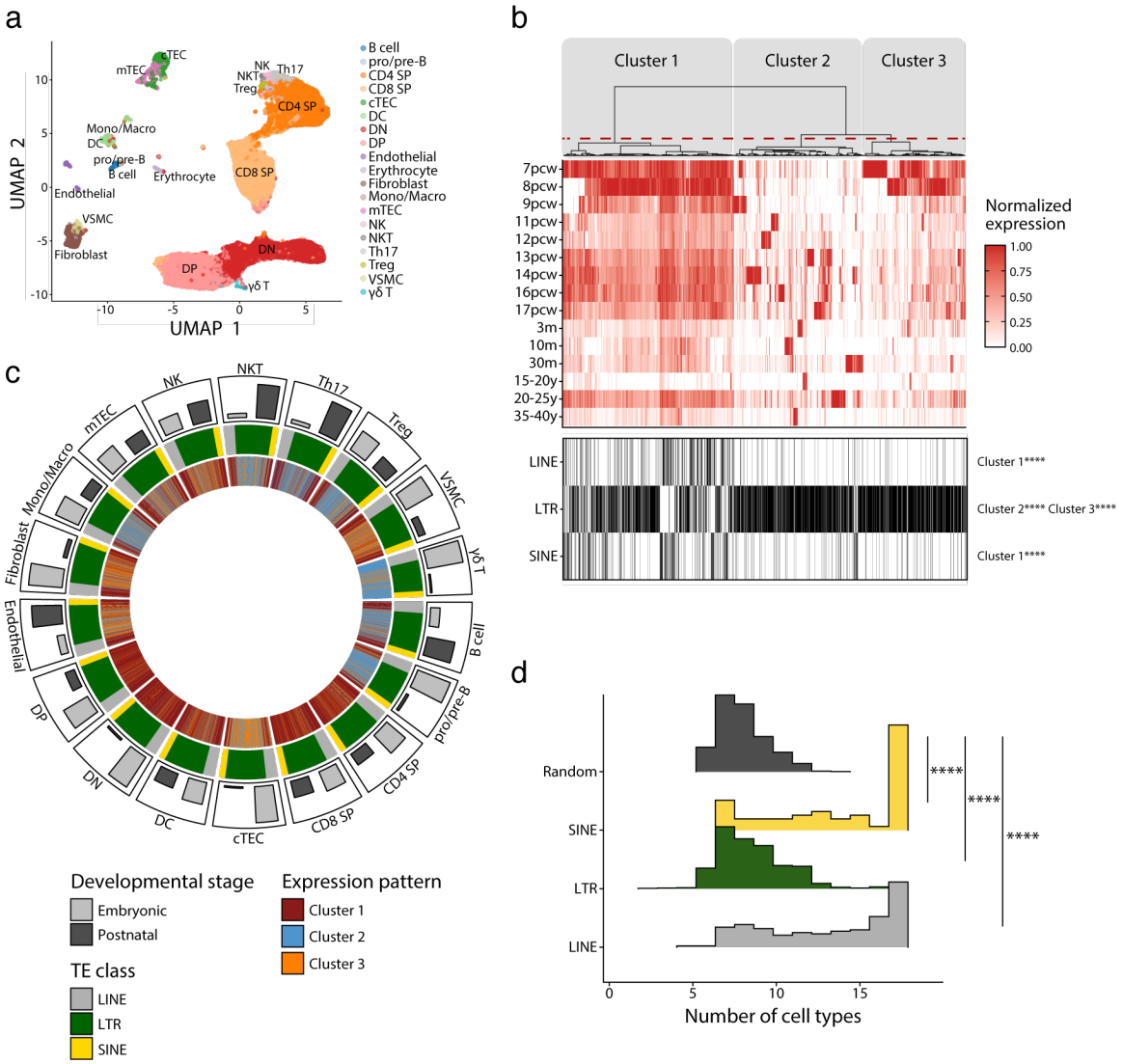


Figure 1. LINEs, SINEs, and LTRs exhibit distinct expression profiles in thymic cell populations. **(a)** UMAP depicting the cell populations present in human thymi (CD4 SP, CD4 single positive thymocytes; CD8 SP, CD8 single positive thymocytes; cTEC, cortical thymic epithelial cells; DC, dendritic cells;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

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

970

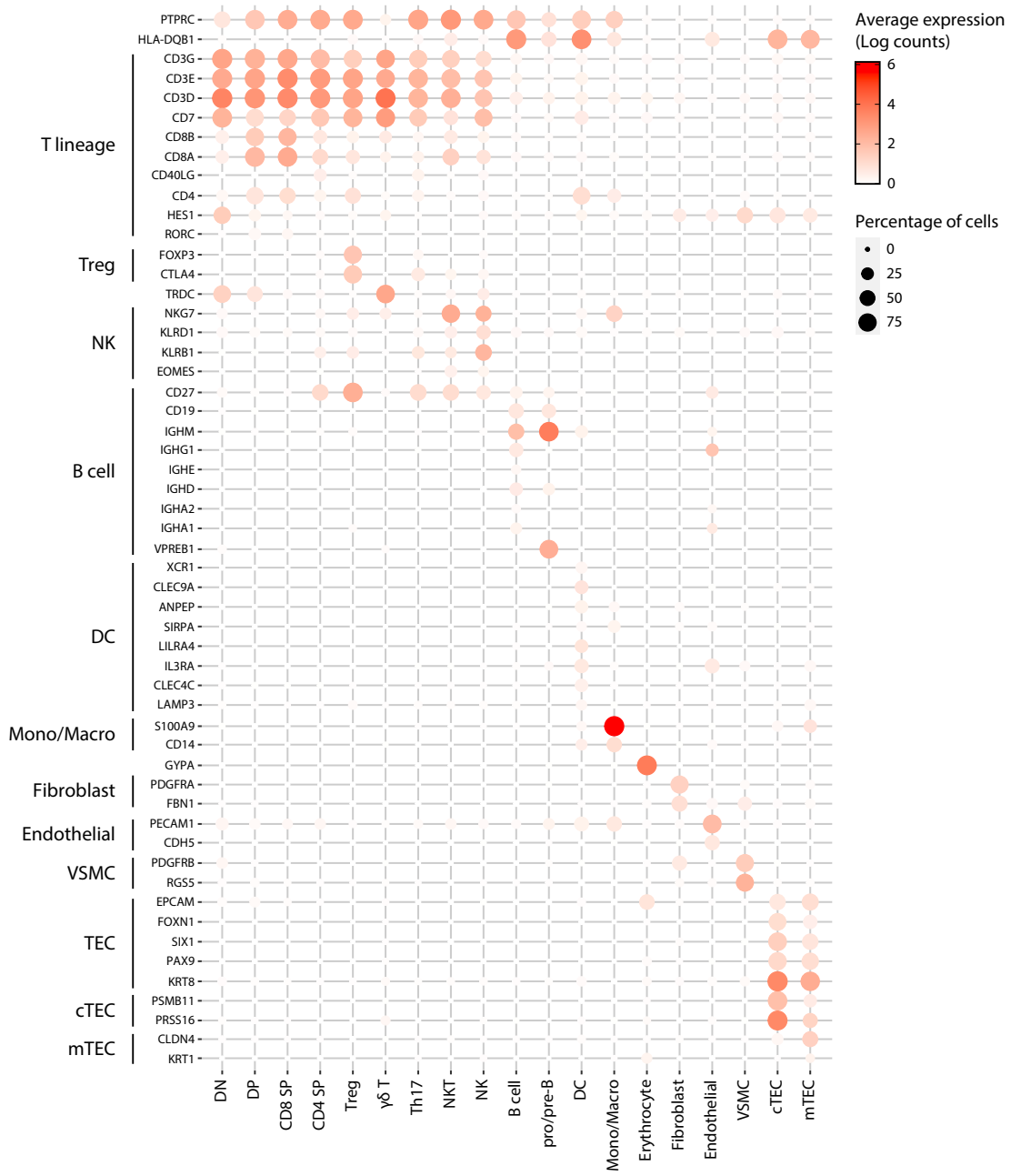
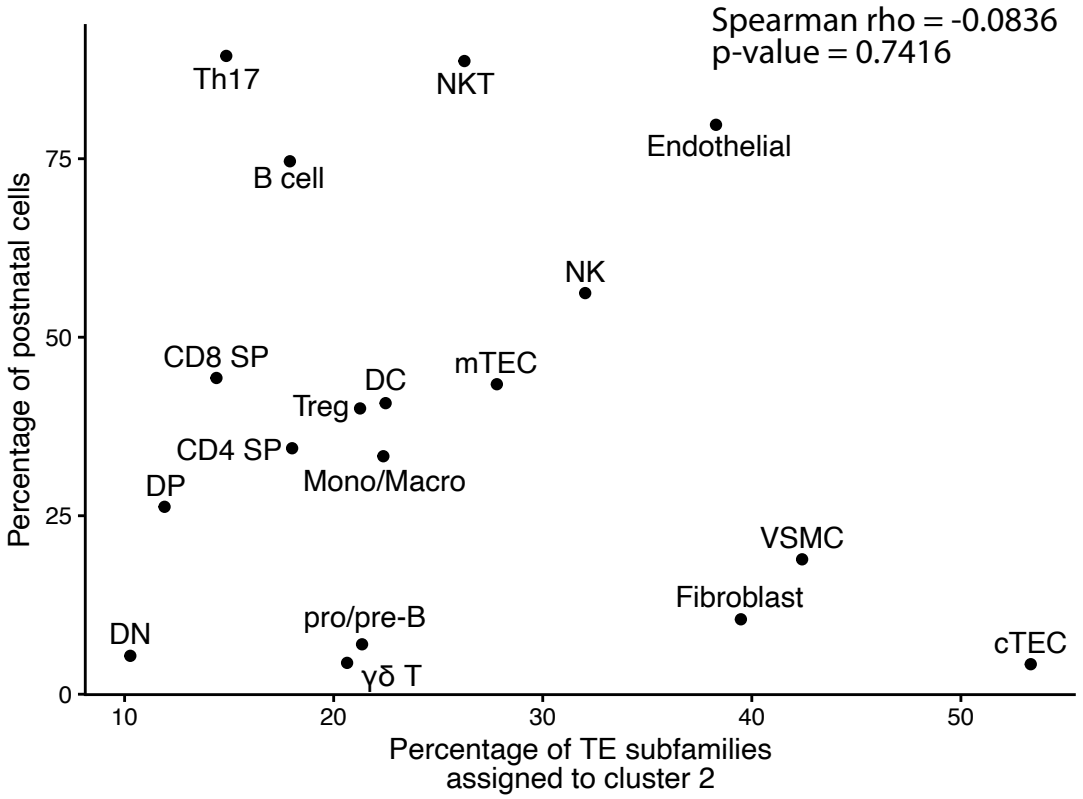


Figure 1 – figure supplement 1. Annotation of human thymic cell populations.

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

977 killer T cells; NK, natural killer cells; DC, dendritic cells; Mono/Macro, monocytes and macrophages;
978 VSMC, vascular smooth muscle cells; cTEC, cortical thymic epithelial cells; mTEC, medullary thymic
979 epithelial cells).

980



981

982 **Figure 1 – figure supplement 2. Assignation to cluster 2 is independent of the developmental stage**
983 **of cells.**

984 Correlation between the proportion of cells of a population originating from a postnatal sample and the
985 proportion of TE subfamilies assigned to the cluster 2 by the hierarchical clustering in Figure 1B.

986

987

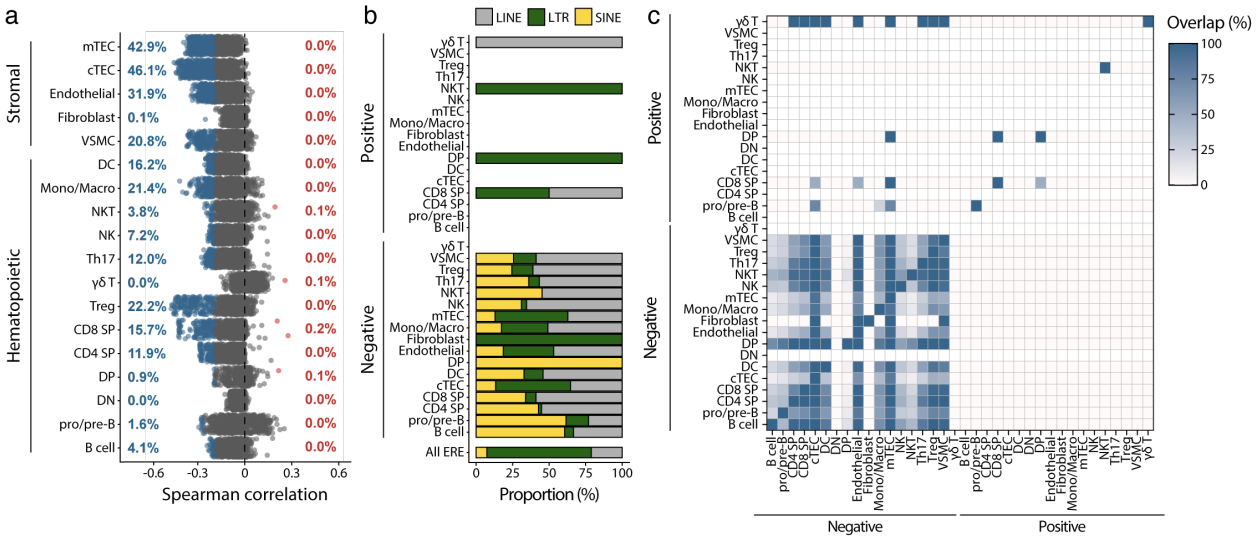


Figure 1 – figure supplement 3. TE expression is negatively correlated with cell proliferation. **(a)** Spearman correlation between the expression of TE subfamilies and cell cycle scores. Positively ($r \geq 0.2$ and adj. $p \leq 0.01$) and negatively ($r \leq -0.2$ and adj. $p \leq 0.01$) correlated 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.

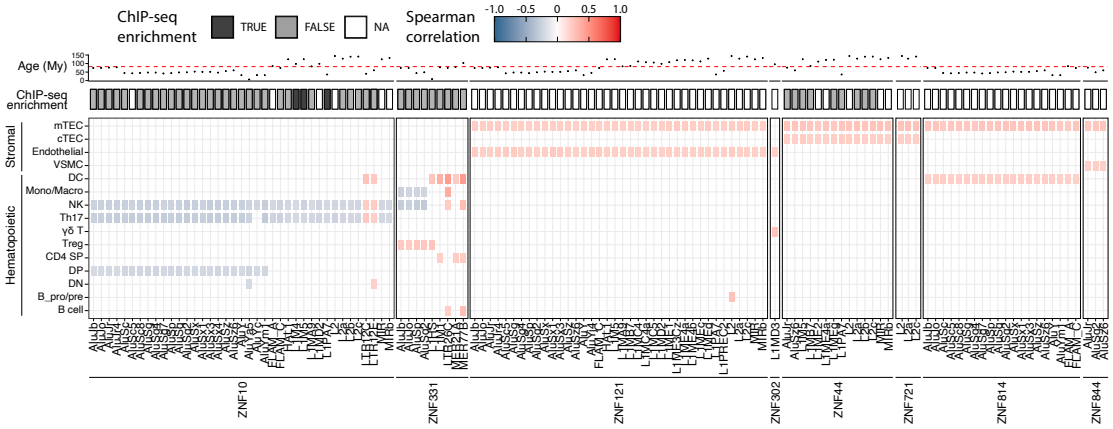
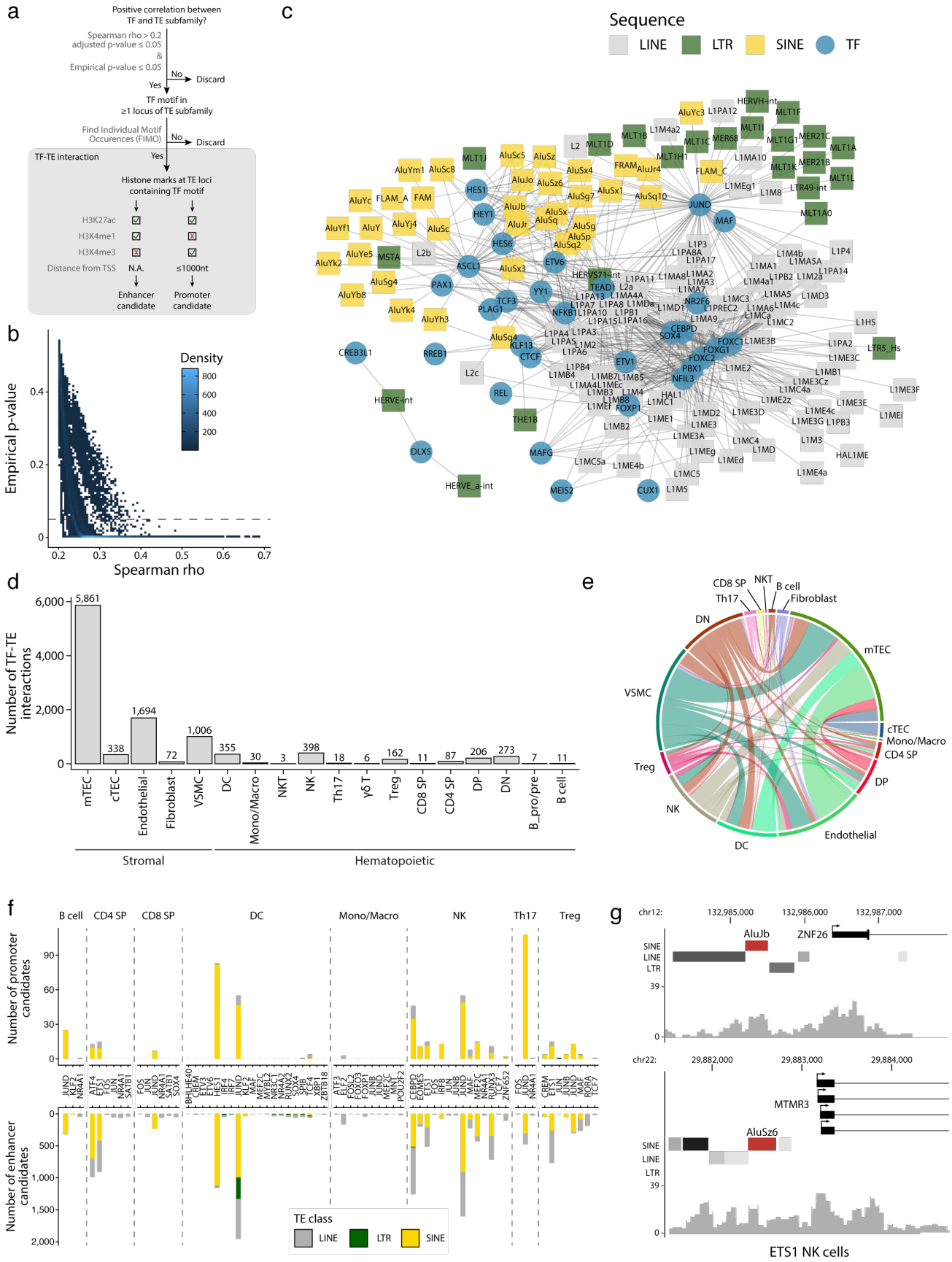


Figure 1 – figure supplement 4. KZFPs repress TE expression in the hematopoietic lineage of the thymus.

Lower panel: pairs of TE subfamilies and KZFPs significantly correlated in at least 2 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 estimated time of divergence between primates and rodents (82 million years ago) is indicated by the dashed line.


a
Positive correlation between TF and TE subfamily?
Spearman rho > 0.2
adjusted p-value ≤ 0.05
&
Empirical p-value ≤ 0.05
No
Discard
Yes
TF motif in ≥1 locus of TE subfamily
Find Individual Motif Occurences (FIMO)
No
Discard
Yes
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
800
600
400
200
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
chr12:
132,985,000
132,986,000
132,987,000
AluJb
ZNF26
SINE
LINE
LTR
chr22:
29,882,000
29,883,000
29,884,000
MTMR3
AluSz6
ETS1 NK cells

1009

1010 **Figure 2. TEs shape complex gene regulatory networks in thymic cells. (a)** Flowchart depicting the
1011 decision tree for each TE promoter or enhancer candidate. **(b)** Density heatmap representing the
1012 correlation coefficient and the empirical p-value determined by bootstrap for TF and TE pairs in each cell
1013 type of the dataset. The color code shows density (i.e., the occurrence of TF-TE pairs at a specific point).
1014 **(c)** Connectivity map of interactions between TEs and TFs in mTECs. For visualization purposes, only
1015 TF-TE pairs with high positive correlations (Spearman correlation coefficient ≥ 0.3 and p-value adjusted
1016 for multiple comparisons with the Benjamini-Hochberg procedure ≤ 0.05) and TF binding sites in $\geq 1\%$
1017 of TE loci are shown. **(d)** Number of TF-TE interactions for each thymic cell population. **(e)** Sharing of
1018 TF-TE pairs between thymic cell types. **(f)** Number of promoter (*top*) or enhancer (*bottom*) TE candidates
1019 per transcription factor in hematopoietic cells of the thymus. **(g)** Genomic tracks depicting ETS1
1020 occupancy (i.e., read coverage) of two identified TE promoter candidates (*in red)* in ETS1 ChIP-seq data
1021 from NK cells.

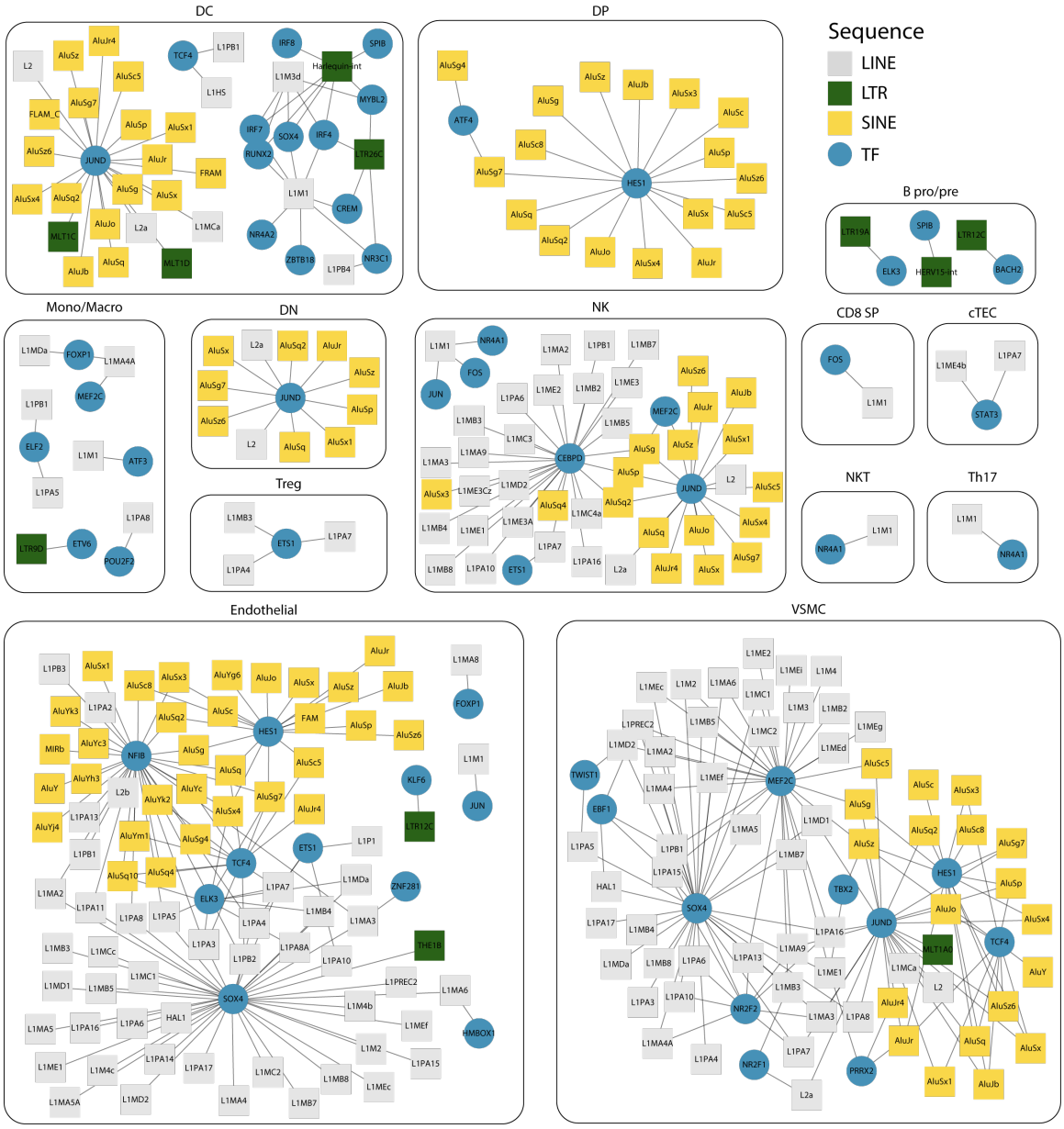


Figure 2 – figure supplement 1. Interaction networks between transcription factors and TE subfamilies.

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).

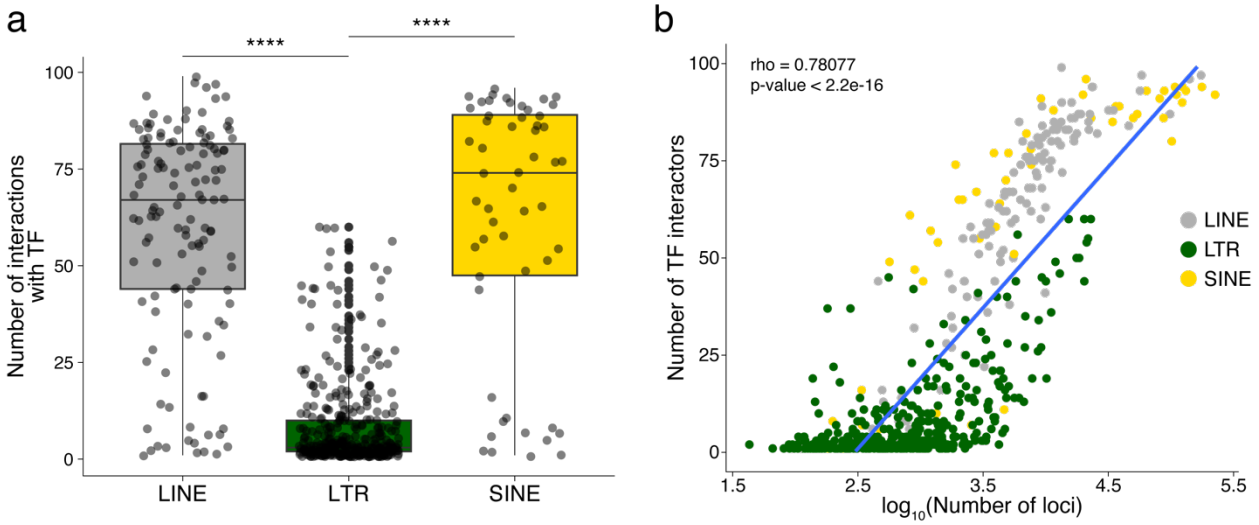


Figure 2 – figure supplement 2. TE subfamilies occupying larger genomic spaces interact more frequently with TF.

(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.

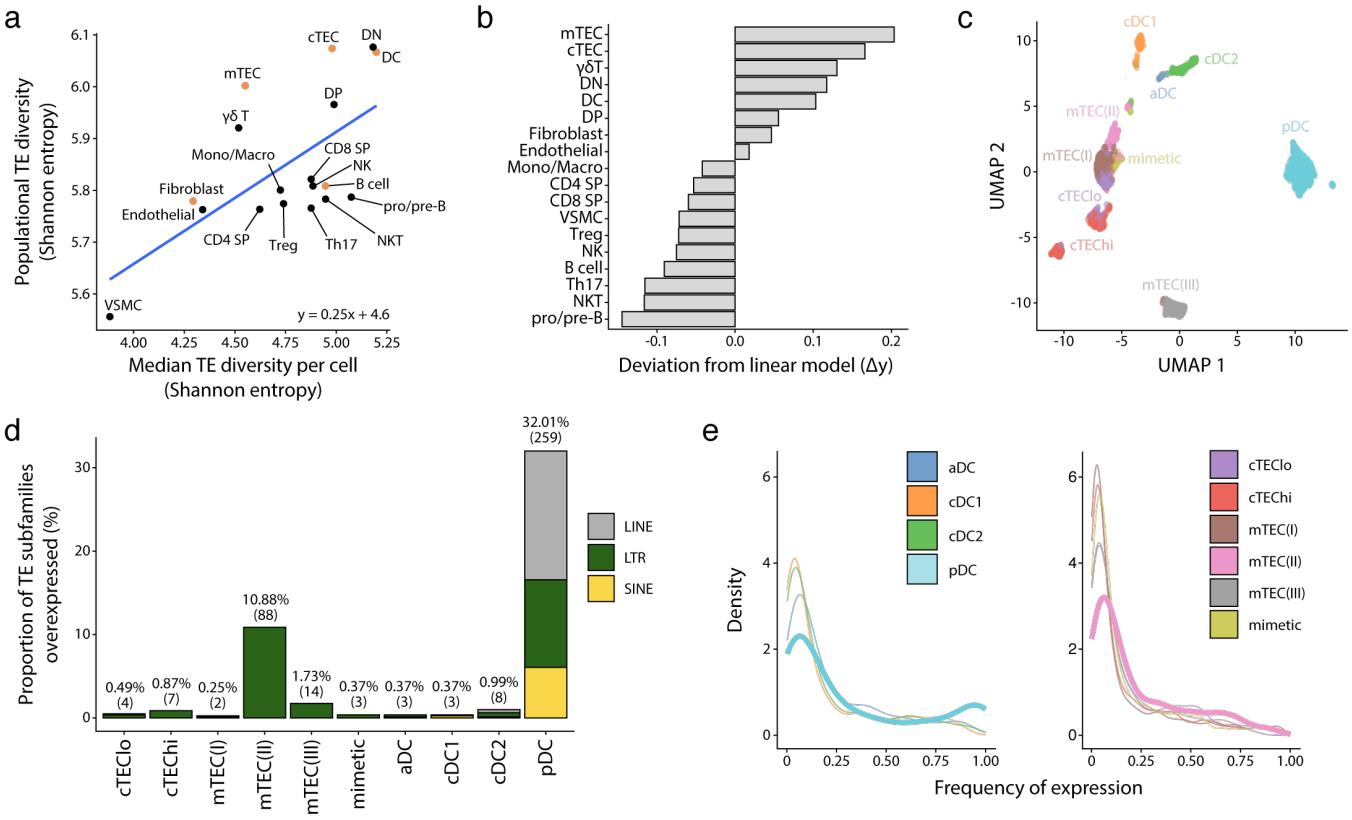


Figure 3. pDCs and mTEC(II) express diverse and distinct repertoires of TE sequences. (a) Diversity of TEs expressed by thymic populations measured by Shannon entropy. The x and y axes represent the median diversity of TEs expressed by individual cells in a population and the global diversity of TEs 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.

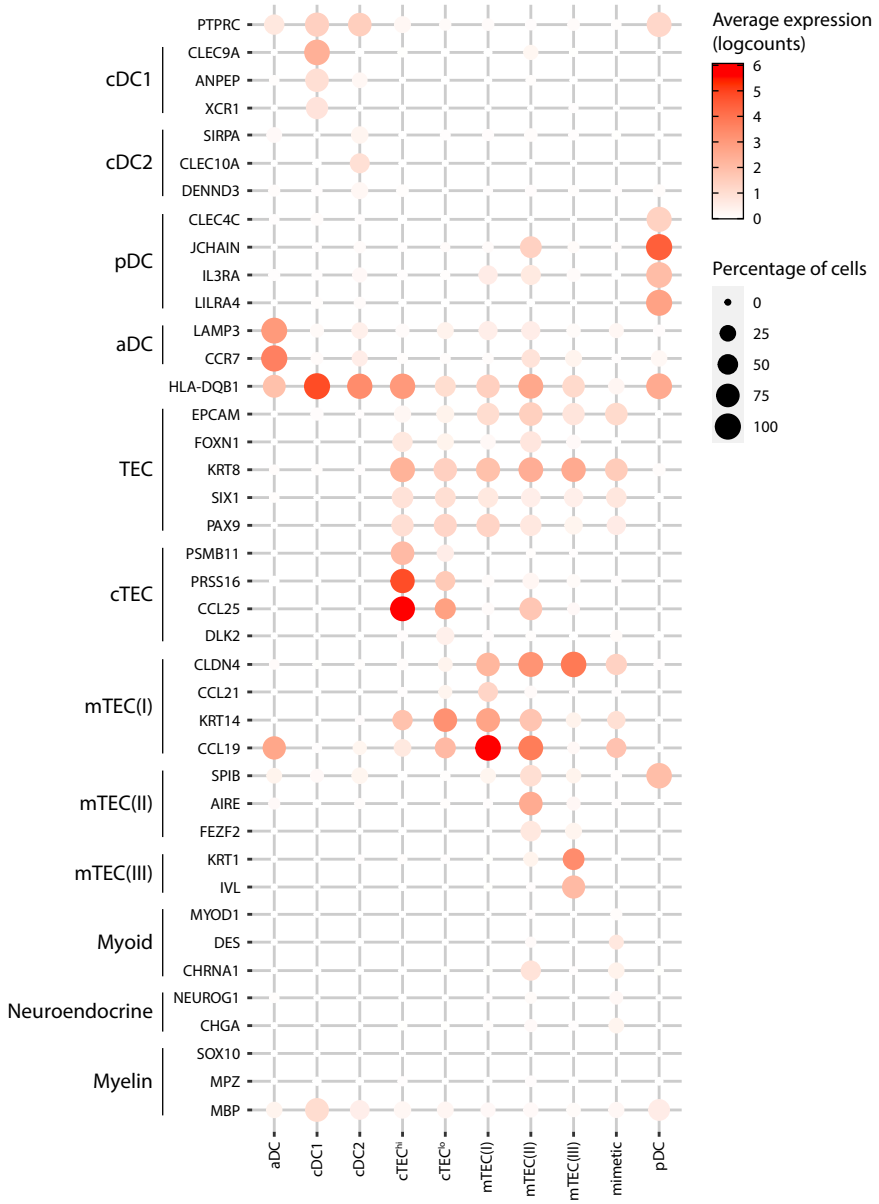


Figure 3 – figure supplement 1. Annotation of human thymic antigen presenting cell subsets. Dot plot depicting the expression of marker genes in the annotated cell types of the thymus. The average expression and percentage of cells expressing the gene are represented by the color and size of the dot, 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).

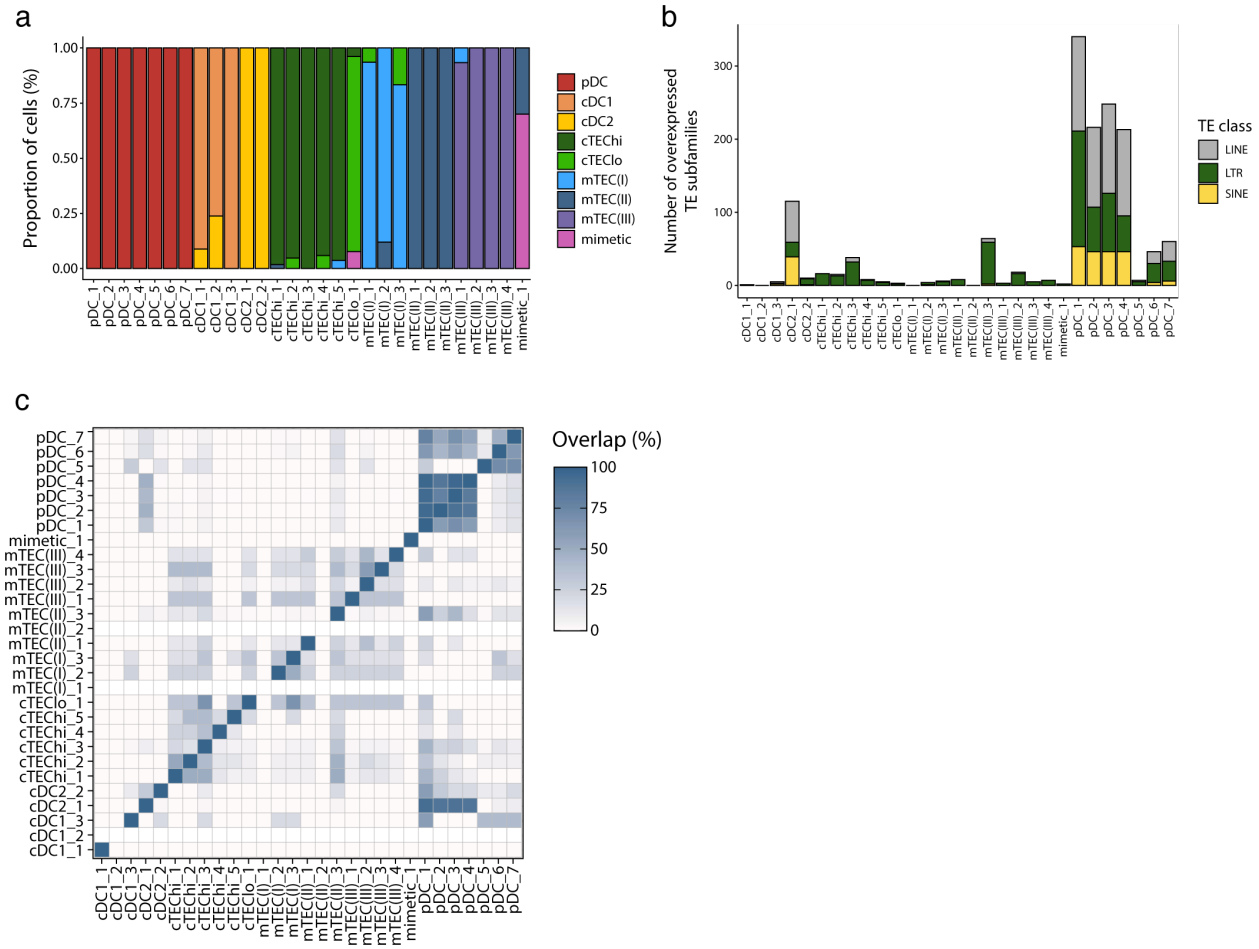


Figure 3 – figure supplement 2. Differential TE expression in metacells of thymic antigen presenting cells.

(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 their class (LINE, LTR, and SINE). **(c)** Percentage of overlap of the TE subfamilies overexpressed by each metacell.

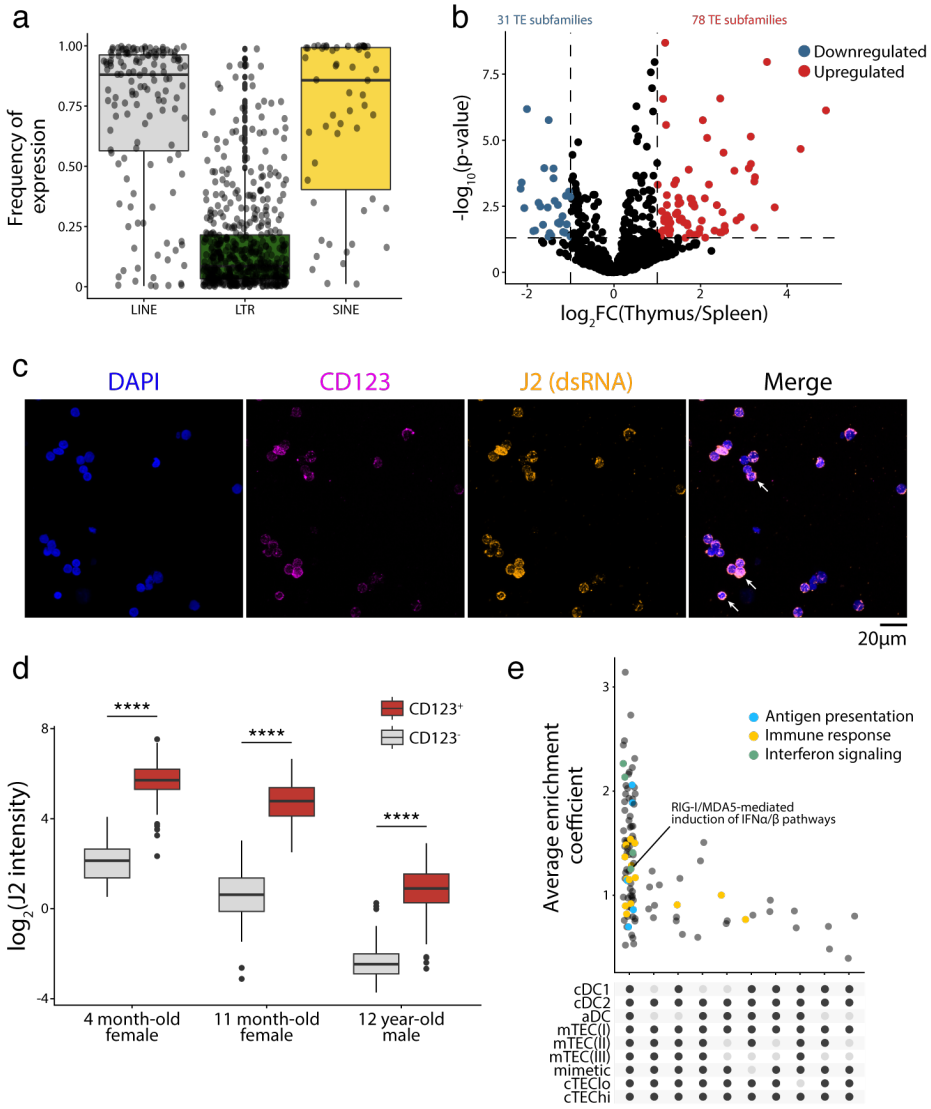
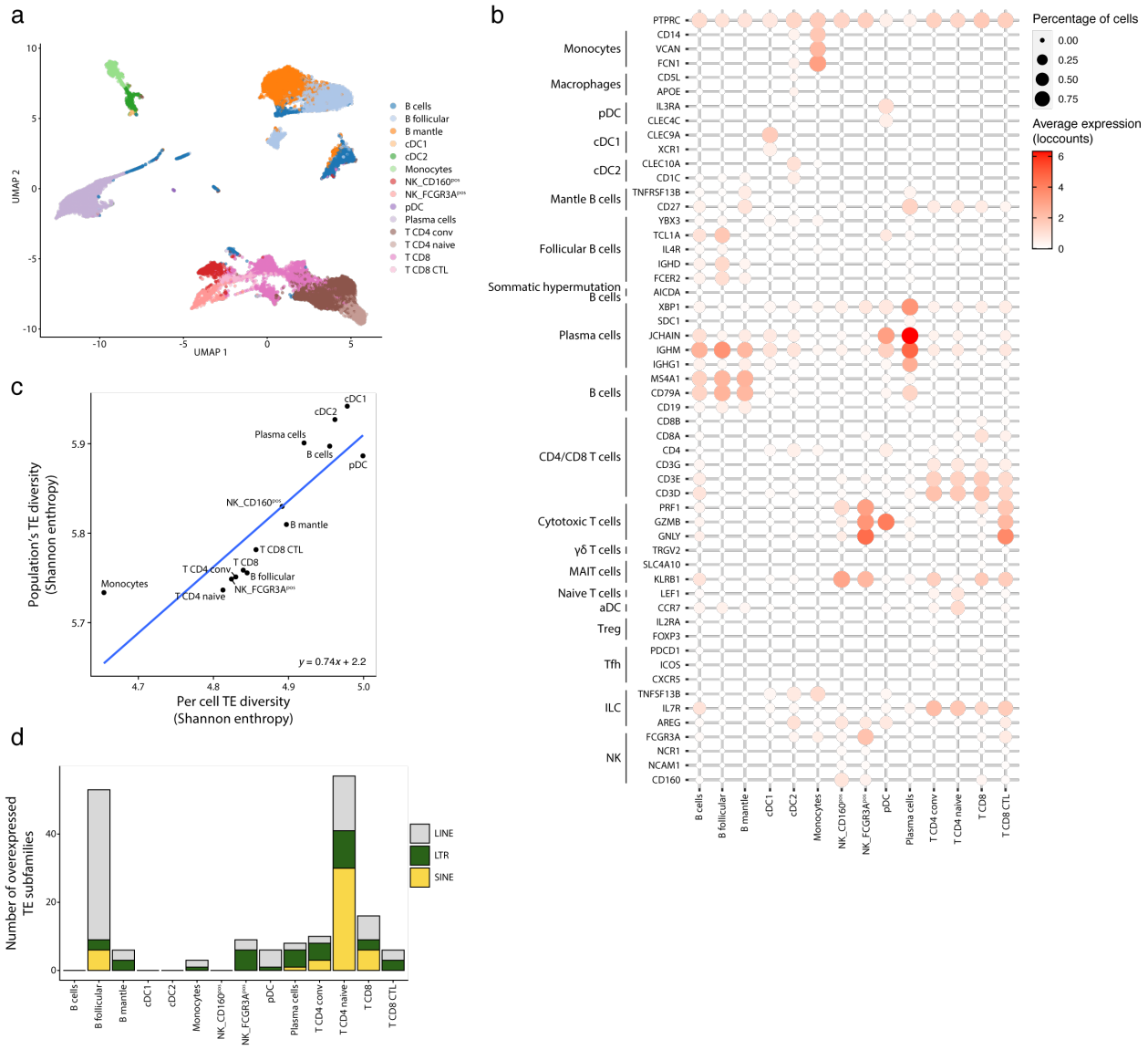


Figure 4. TE expression in pDCs leads to dsRNA formation and type I IFN signaling. (a) Frequency of expression of LINE, LTR, and SINE subfamilies in thymic pDCs. **(b)** Differential expression of TE subfamilies between splenic and thymic pDCs. TE subfamilies significantly upregulated or downregulated 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)**

1073 UpSet plots showing gene sets enriched in pDCs compared to the other populations of thymic APCs. On
1074 the lower panel, black dots represent cell populations for which gene signatures are significantly depleted
1075 compared to pDCs. All comparisons where gene signatures were significantly enriched in pDCs are
1076 shown.

1077



1078

1079 **Figure 4 – figure supplement 1. TE expression in 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 a population, respectively. A linear model summarizing the data is represented by the equation and blue curve. **(d)** Bar plot showing the number (y axis) and class (color) of differentially expressed TE subfamilies between splenic cell populations.

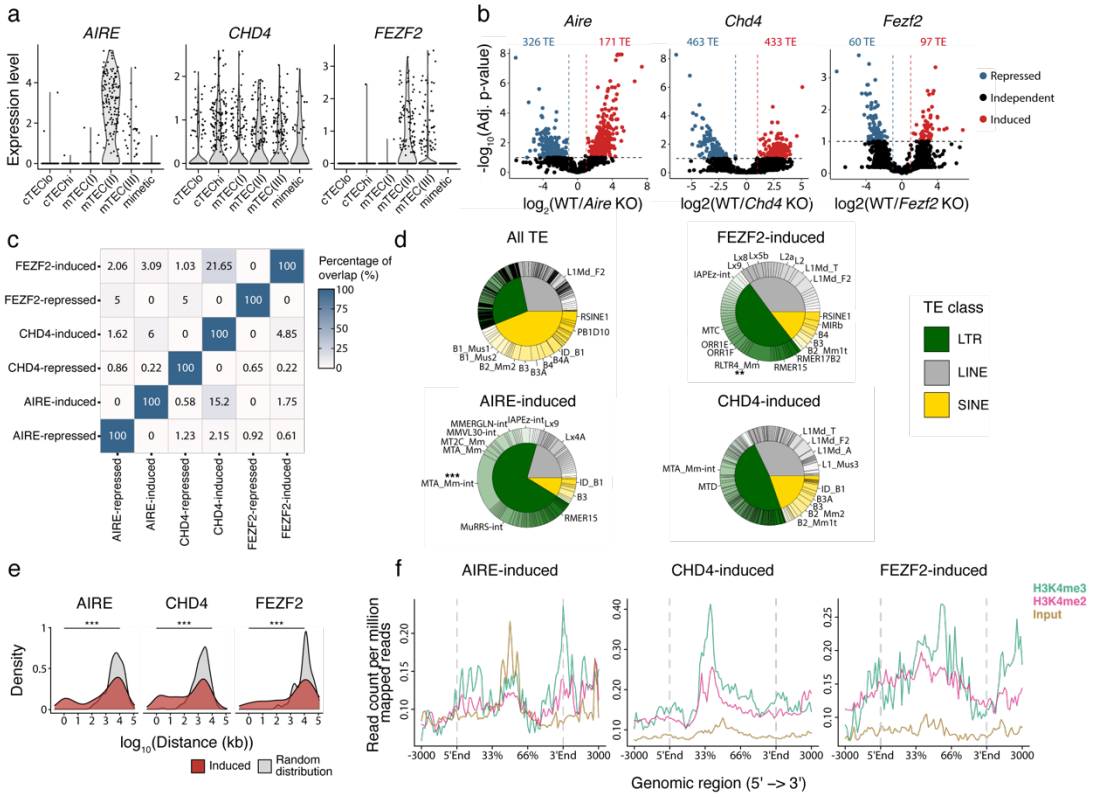


Figure 5. *AIRE*, *FEZF2*, and *CHD4* regulate non-redundant sets of TEs in mTECs. (a) Expression of *AIRE*, *CHD4*, and *FEZF2* in human TEC subsets. **(b)** Differential expression of TE loci between wild-

1092 type (WT) and *Aire*-, *Chd4*- or *Fezf2*-knockout (KO) mice (Induced, $\log_2$(WT/KO)≥2 and adj. p≤0.05;
1093 Repressed, $\log_2$(WT/KO)≤-2 and adj. p≤0.05). P-values were corrected for multiple comparisons with the
1094 Benjamini-Hochberg procedure. The numbers of induced (red) and repressed (blue) TE loci are indicated
1095 on the volcano plots. **(c)** Overlap of TE loci repressed or induced by AIRE, FEZF2, and CHD4. **(d)**
1096 Proportion of TE classes and subfamilies in the TE loci regulated by AIRE, FEZF2, or CHD4, as well as
1097 all TE loci in the murine genome for comparison (Chi-squared tests with Bonferroni correction, **adj.
1098 p≤0.01, ***adj. p≤0.001). **(e)** Distance between TE loci induced by AIRE, FEZF2, and CHD4, and
1099 random selections of TE loci (Wilcoxon rank-sum tests, ***p≤0.001). **(f)** Plots for the tag density of
1100 H3K4me3 and H3K4me2 on the sequence and flanking regions (3000 base pairs (bp)) of TE loci induced
1101 by AIRE, FEZF2, and CHD4.

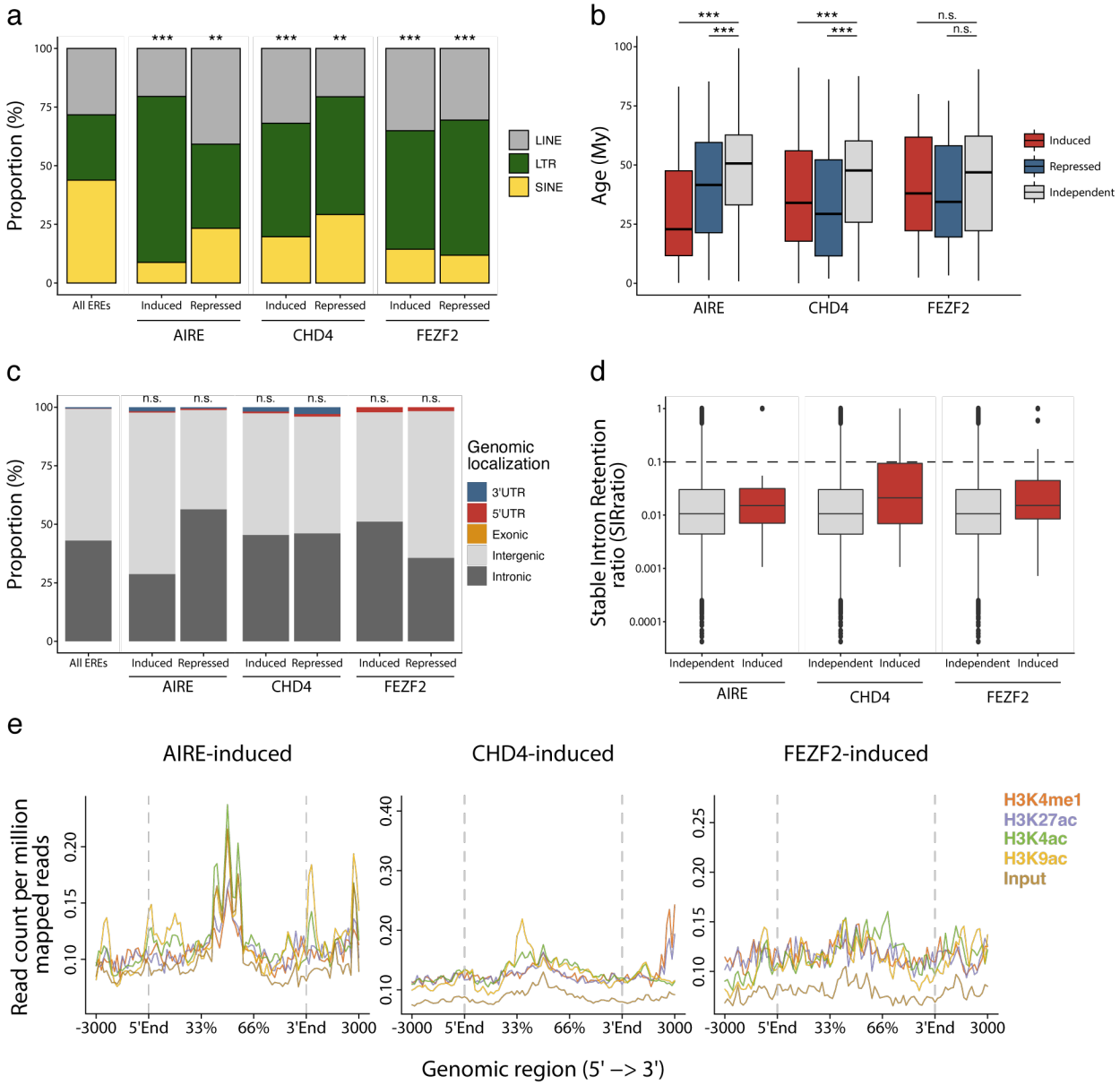


Figure 5 – figure supplement 1. Characterization of TE subfamilies regulated by AIRE, CHD4 and FEZF2.

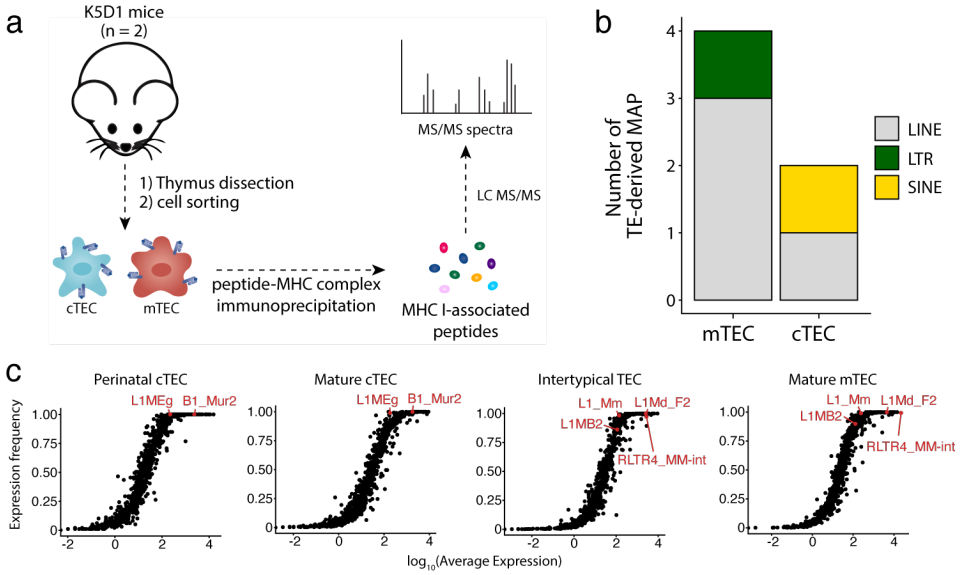
(a) Class of TE induced or repressed by AIRE, CHD4 and FEZF2. Distributions were compared to the proportion of LINE, LTR, and SINE amongst all TE sequences of the murine genome with Chi-squared tests (**p≤0.01, ***p≤0.001). **(b)** Age of TE induced, repressed or independent of AIRE, CHD4 and FEZF2 (Wilcoxon-Mann-Whitney test, *p≤0.05, ***p≤0.001) (My, millions of years). **(c)** Genomic localization of the TE loci induced or repressed by AIRE, CHD4 and FEZF2. **(d)** Intron retention ratio of

1110 intronic TE induced or independent of AIRE, CHD4 and FEZF2. Dashed line represents intron retention

1111 events occurring in at least 10% of transcripts. **(e)** Plots for the tag density of histone marks on the sequence

1112 and flanking regions (3000bp) of TE loci induced by AIRE, CHD4, and FEZF2.

1113



1114

1115 **Figure 6. TE MAPs are presented by cTECs and mTECs. (a)** mTECs and cTECs were isolated from

1116 the thymi of K5D1 mice (n=2). The peptide-MHC I complexes were immunoprecipitated for both

1117 populations independently, and MAPs were sequenced by MS analyses. **(b)** Number of LINE-, LTR-, and

1118 SINE-derived MAPs in mTECs and cTECs from K5D1 mice. **(c)** Distributions of TE subfamilies in

1119 murine TECs subsets based on expression level (*x-axis*) and frequency of expression (*y-axis*).

Fig. S1. 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).

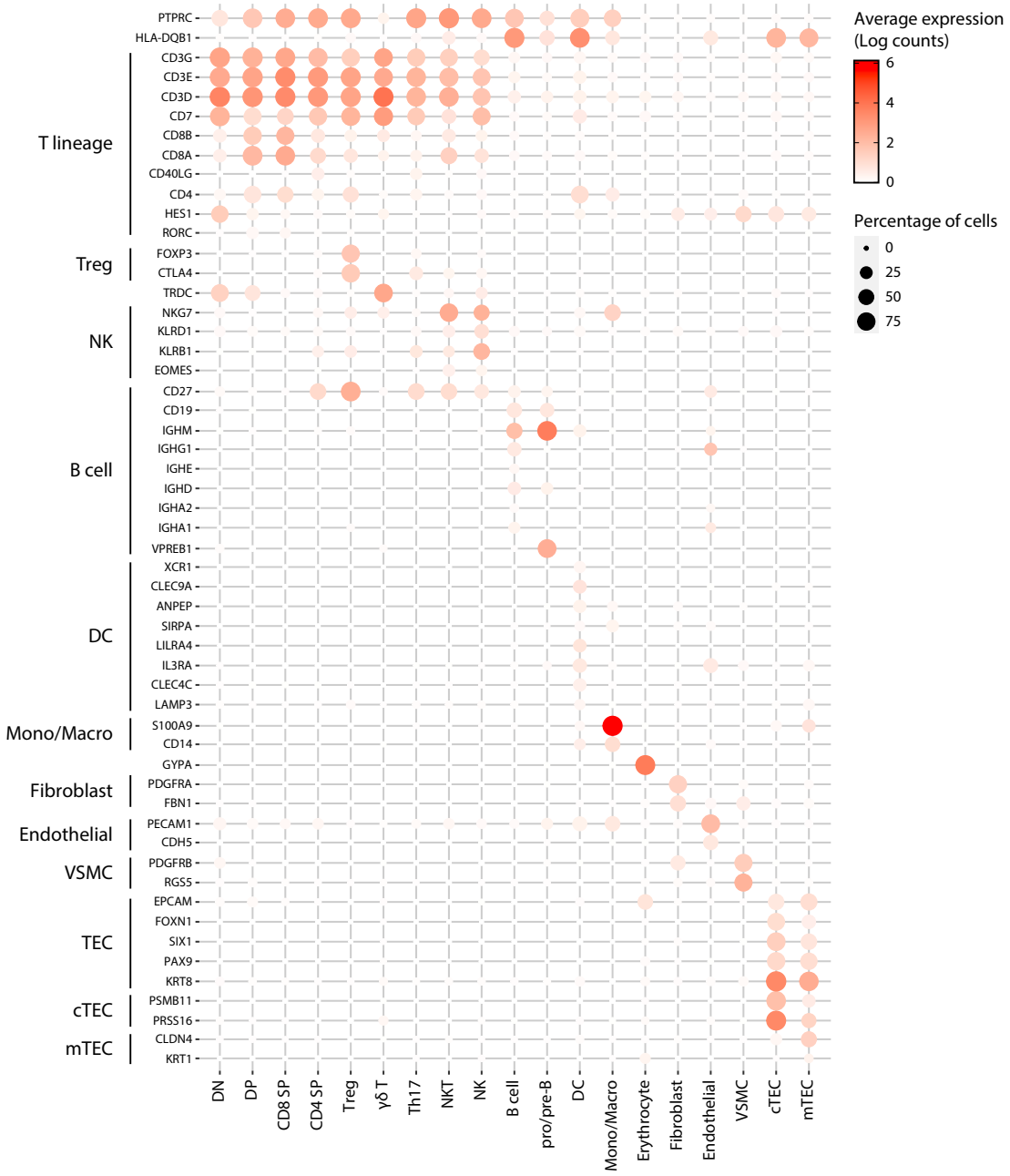


Fig. S2. Assignation to cluster 2 is independent of the developmental stage of cells.
Correlation between the proportion of cells of a population originating from a postnatal sample and the proportion of TE subfamilies assigned to the cluster 2 by the hierarchical clustering in Fig. 1B.

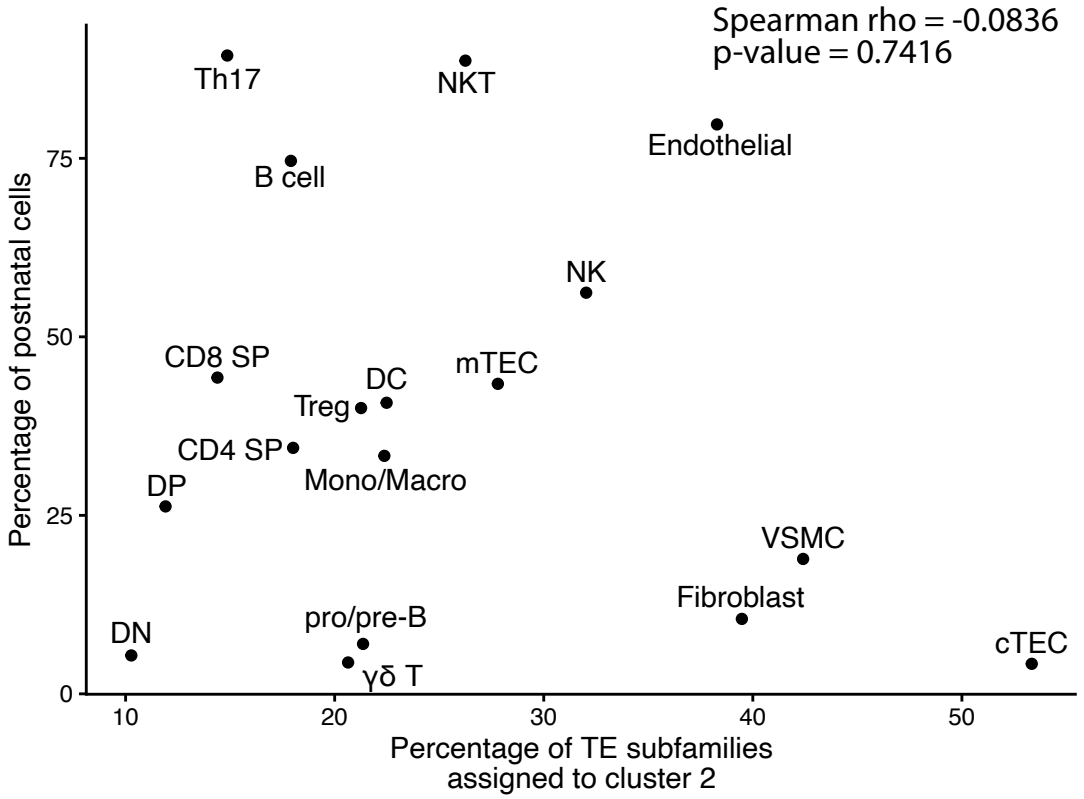


Fig. S3. TE expression is negatively correlated with cell proliferation.
a Spearman correlation between the expression of TE subfamilies and cell cycle scores. Positively (r≥0.2 and adj. p≤0.01) and negatively (r≤-0.2 and adj. p≤0.01) correlated subfamilies are red and blue, respectively. p-values were 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.

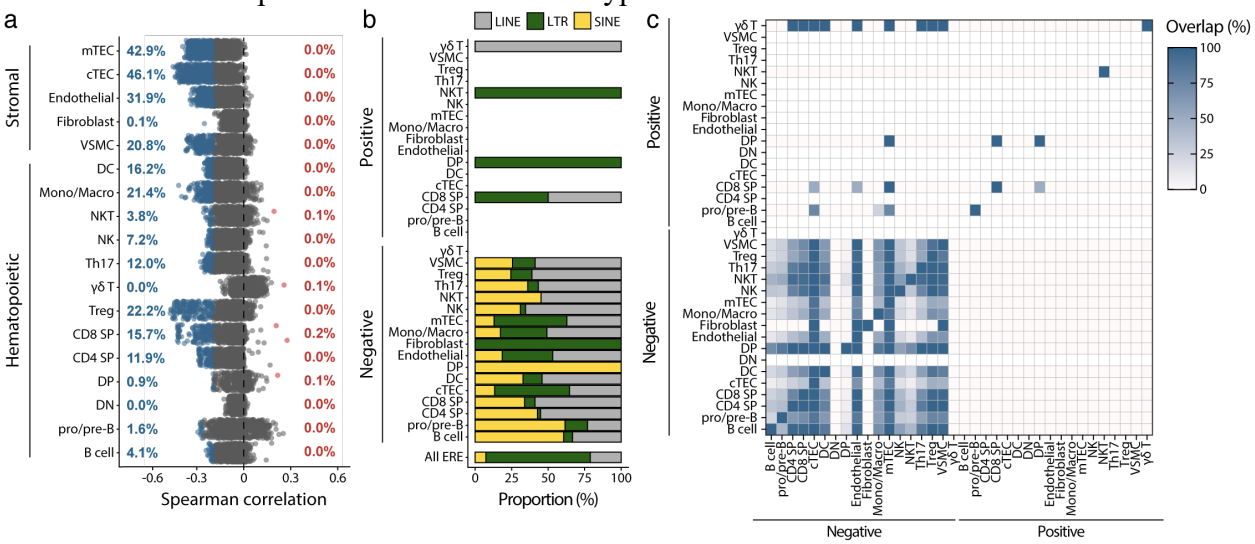


Fig. S4. KZFPs repress TE expression in the hematopoietic lineage of the thymus.
Lower panel: pairs of TE subfamilies and KZFPs significantly correlated in at least 2 cell types (significant correlation: r>0.2 and adj. p≤0.05, or r<-0.2 and adj. p≤0.05, p-values corrected for multiple comparisons with 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* (1). *Upper panel*: Age of TE subfamilies in millions of years (My). The estimated time of divergence between primates and rodents (82 million years ago) is indicated by the dashed line.

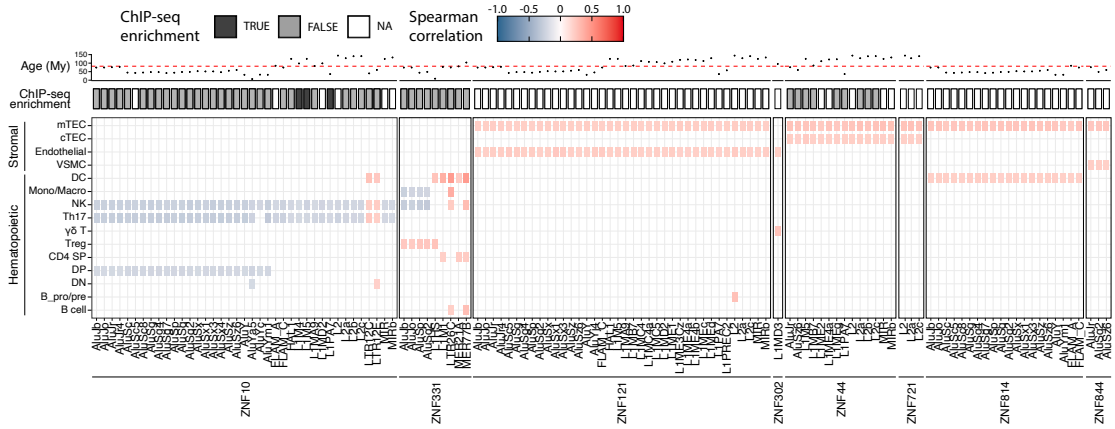


Fig. S5. Interaction networks between transcription factors and TE subfamilies.
For each cell type, networks illustrate the interactions between TF and TE subfamilies. Pairs of TF and TE are connected by edges when i) their expressions are significantly correlated (Spearman correlation coefficient ≥ 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).

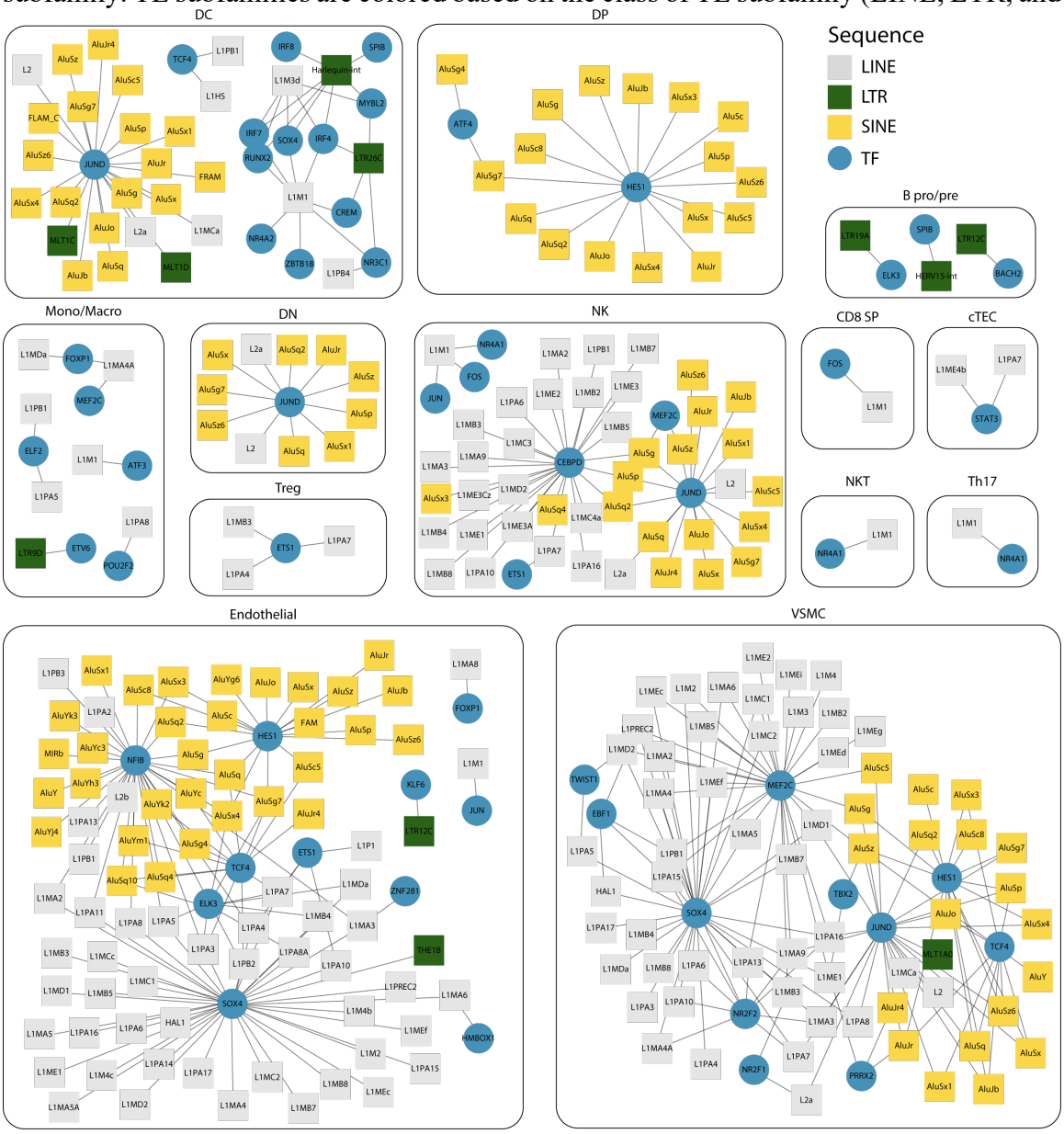


Fig. S6. TE subfamilies occupying larger genomic spaces interact more frequently with TF. **a** Number of interactions formed with TFs for each TE subfamily of the LINE, LTR, and SINE classes (Wilcoxon-Mann-Whitney tests, ****p≤0.0001). **b** Scatterplot depicting the Kendall tau correlation between 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.

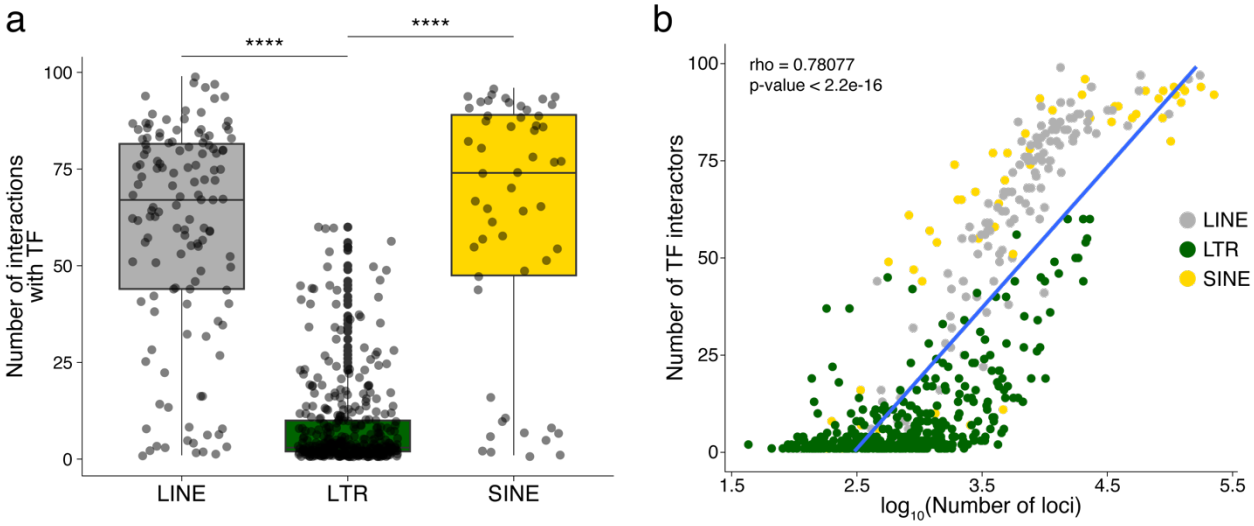


Fig. S7. Annotation of human thymic antigen presenting cell subsets.
Dot plot depicting the expression of marker genes in the annotated cell types of the thymus. The average expression and percentage of cells expressing the gene are represented by the color and size of the dot, respectively. Myoid-, 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).

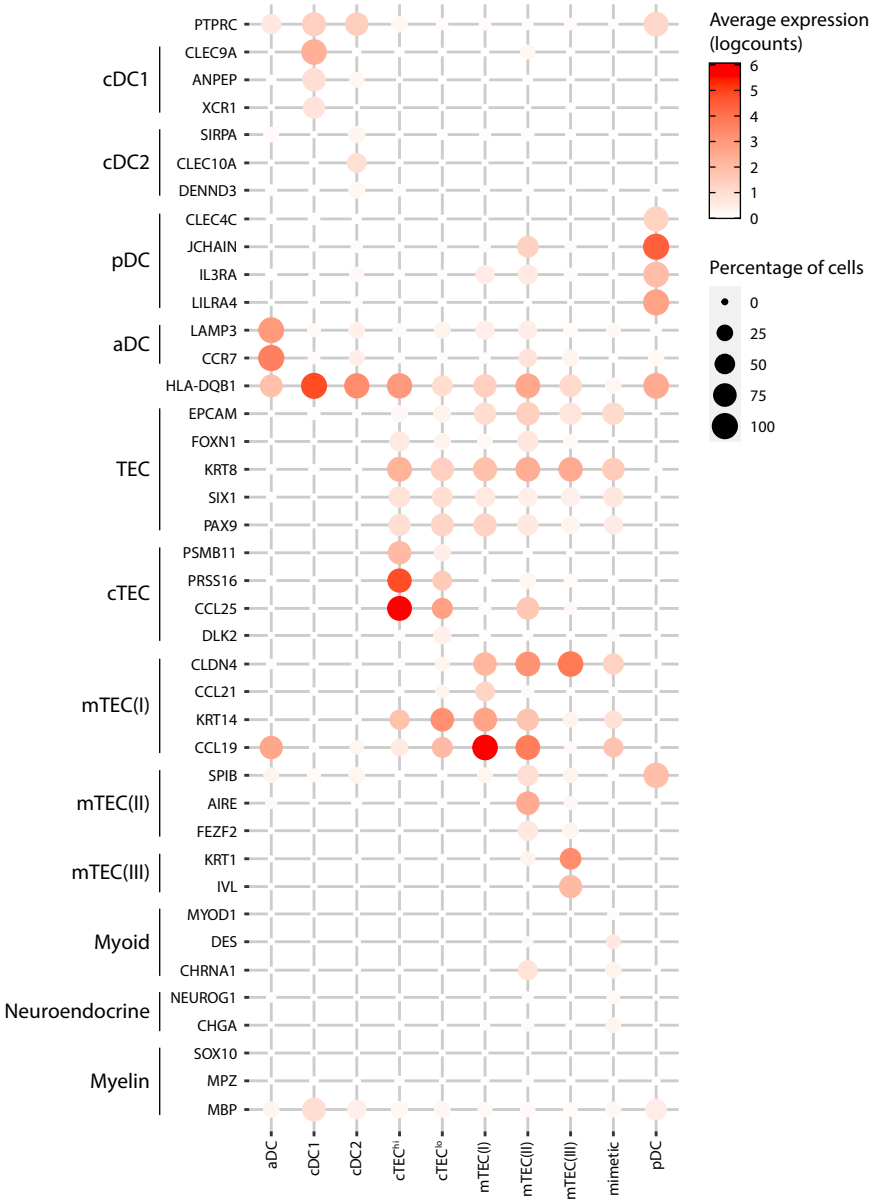


Fig. S8. Differential TE expression in metacells of thymic antigen presenting cells.
a Cellular composition of the metacells (x axis) based on the manual annotation of the thymic cell populations (see Fig. S1). **b** Number of TE subfamilies overexpressed expressed between the metacells. TE subfamilies are colored based on their class (LINE, LTR, and SINE). **c** Percentage of overlap of the TE subfamilies overexpressed by each metacell.

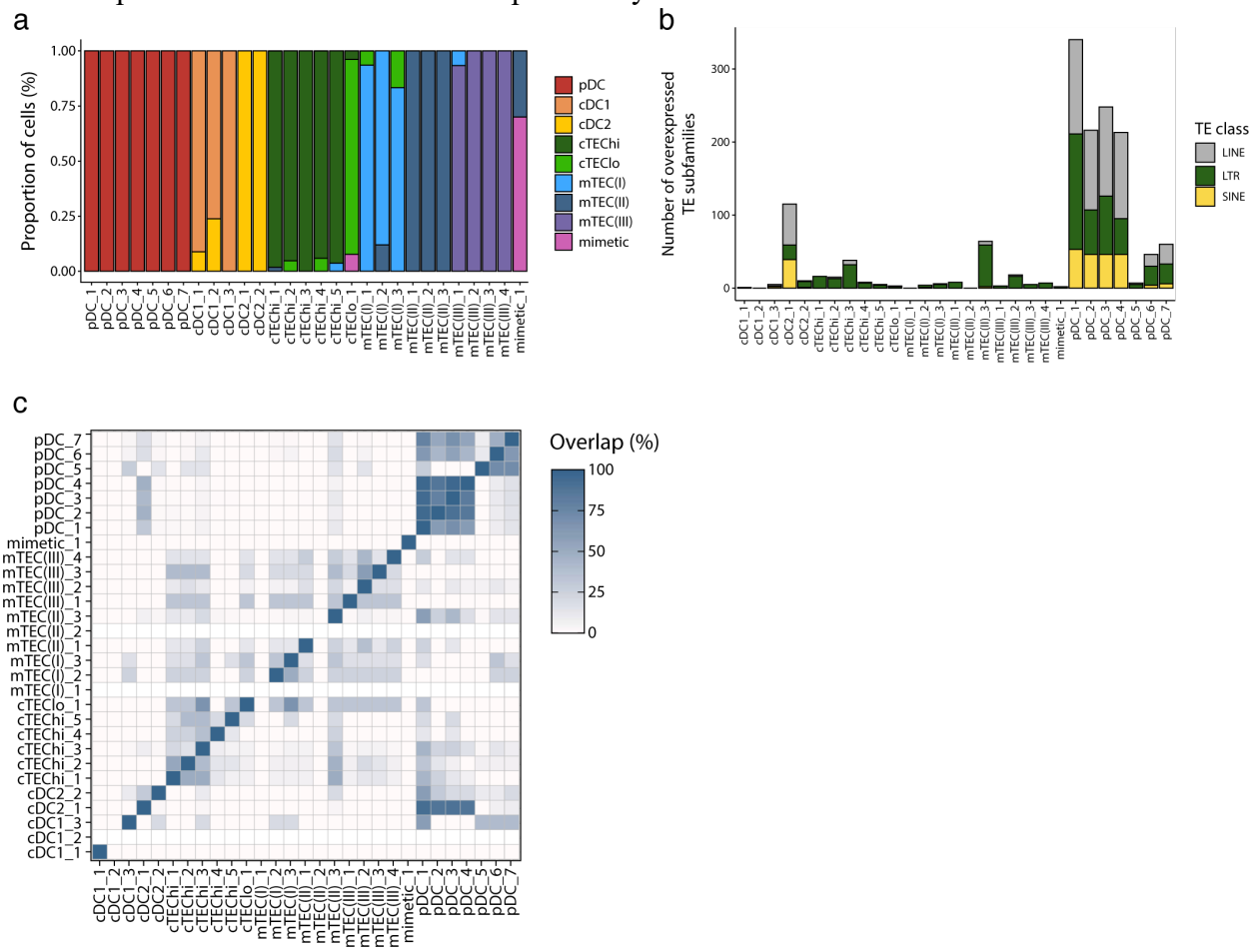


Fig. S9. TE expression in splenic pDCs.

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 a population, respectively. A linear model summarizing the data is represented by the equation and blue curve. **d** Bar plot showing the number (y axis) and class (color) of differentially expressed TE subfamilies between splenic cell populations.

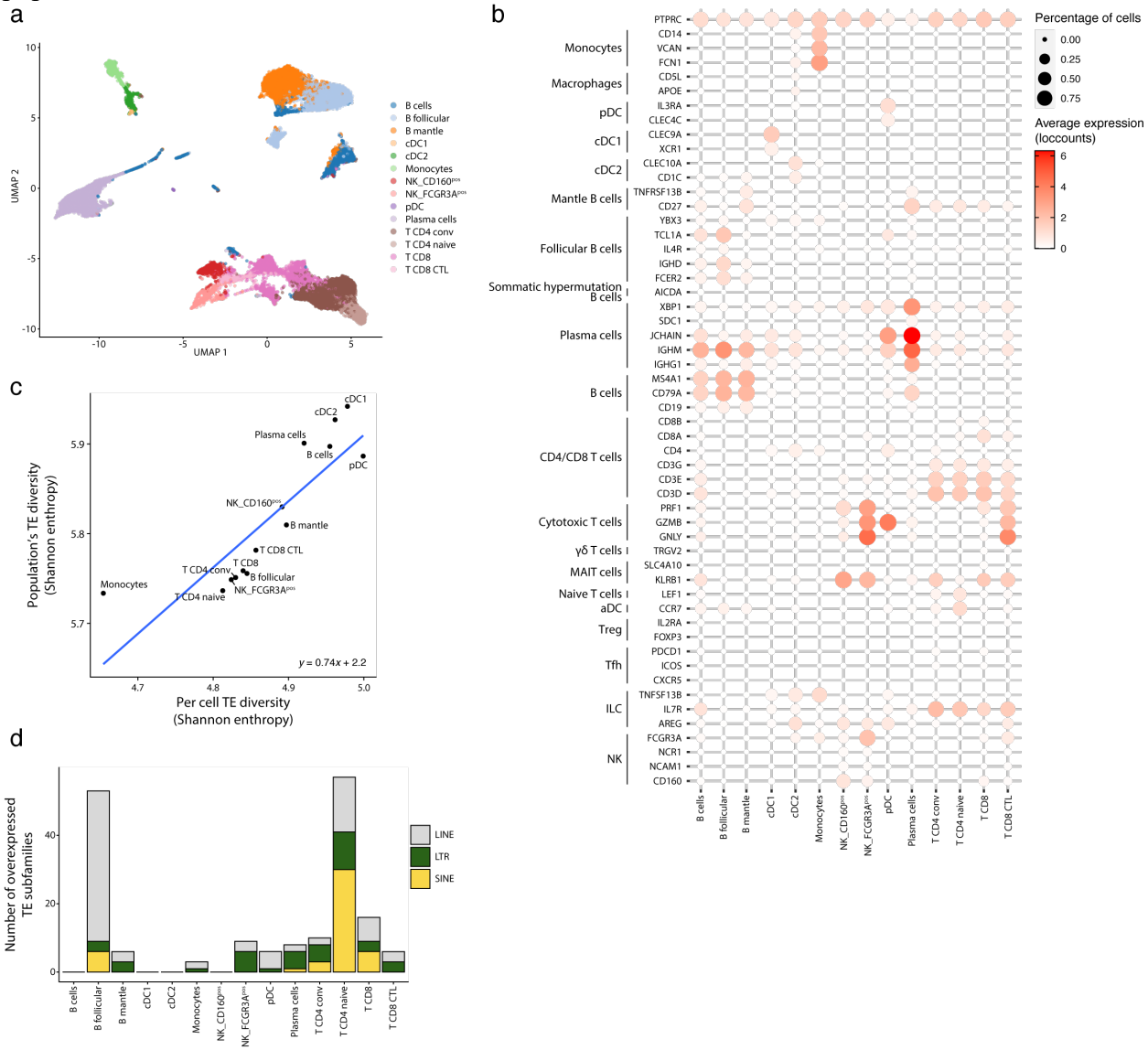
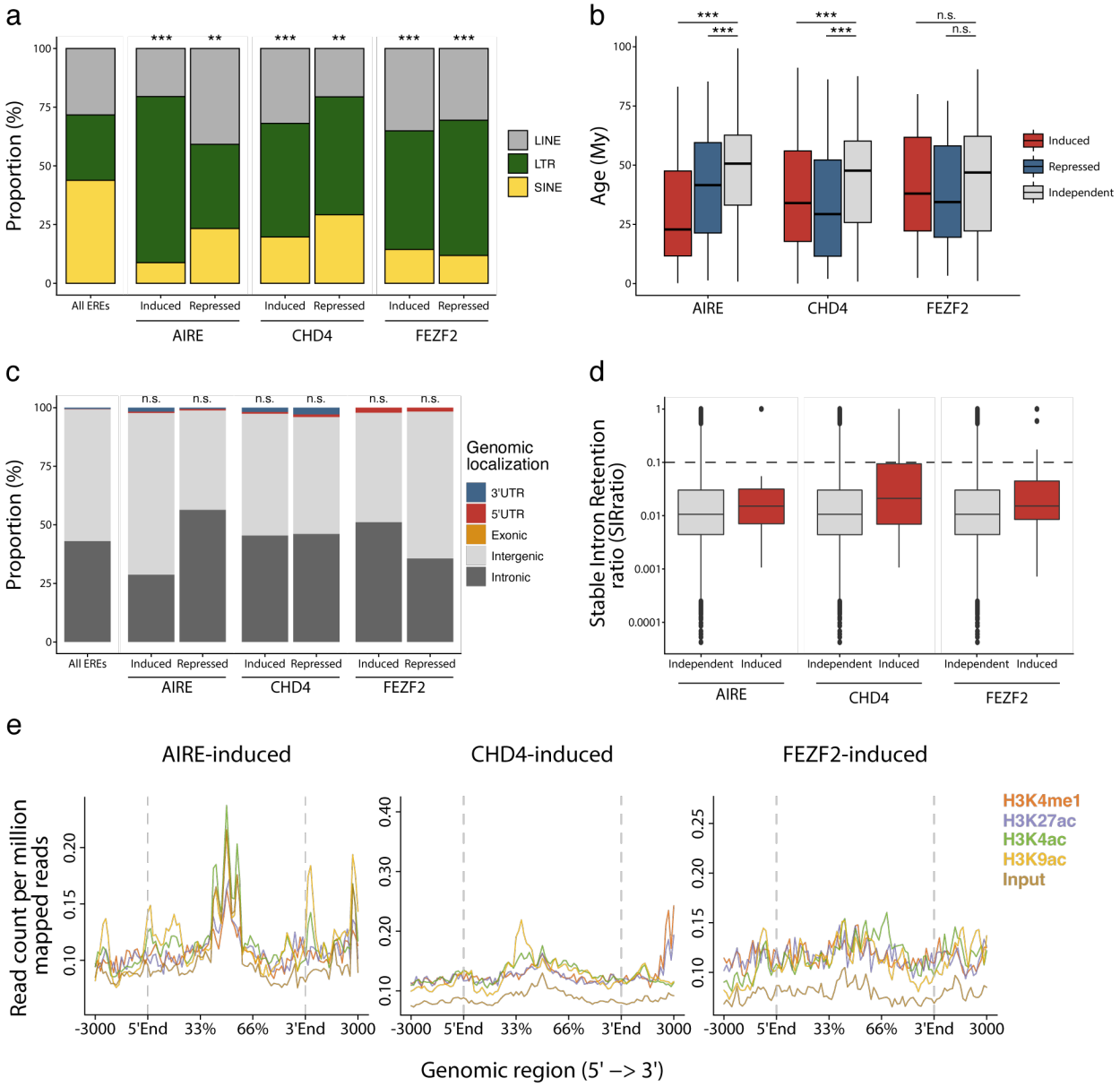



Fig. S10. Characterization of TE subfamilies regulated by AIRE, CHD4 and FEZF2. **a** Class of TE induced or repressed by AIRE, CHD4 and FEZF2. Distributions were compared to the proportion of LINE, LTR, and SINE amongst all TE sequences of the murine genome with Chi-squared tests (**p≤0.01, ***p≤0.001). **b** Age of TE induced, repressed or independent of AIRE, CHD4 and FEZF2 (Wilcoxon-Mann-Whitney test, *p≤0.05, ***p≤0.001) (My, millions of years). **c** Genomic localization of the TE loci induced or repressed by AIRE, CHD4 and FEZF2. **d** Intron retention ratio of intronic TE induced or independent of AIRE, CHD4 and FEZF2. Dashed line represents intron retention events occurring in at least 10% of transcripts. **e** Plots for the tag density of histone marks on the sequence and flanking regions (3000bp) of TE loci induced by AIRE, CHD4, and FEZF2.