Rbfox contributes to a conservative program of self-antigen splicing in thymic epithelial cells

Kathrin Jansen[1,2], Noriko Shikama-Dorn[3], Moustafa Attar[1,4], Stefano Maio[2], Maria Lopopolo[4], David Buck[4], Georg A. Holländer[2,3,5†], Stephen N. Sansom[1†*]

1. The Kennedy Institute of Rheumatology, University of Oxford, Oxford, UK;

2. Department of Paediatrics and the Weatherall Institute of Molecular Medicine, University of Oxford, Oxford, UK;

3. The University Children's Hospital of Basel and the Department of Biomedicine, University of Basel, Basel, Switzerland;

4. Wellcome Centre for Human Genetics, University of Oxford, Oxford, UK;

5. Department of Biosystems Science and Engineering, ETH Zurich, Basel, Switzerland

[†] Equal contribution.

***Correspondence and Contact Address:**

Stephen N. Sansom

The Kennedy Institute of Rheumatology,

University of Oxford,

Roosevelt Drive,

Headington, Oxford,

Oxford,

OX3 7FY

Tel: +44 1865 612663

Email: stephen.sansom@kennedy.ox.ac.uk or stephen.sansom@cantab.net

Abstract

Thymic epithelial cells (TEC) guide selection of a T-cell repertoire that is reactive to pathogens but tolerant to self. While this process is known to involve the promiscuous expression of virtually the entire protein-coding gene repertoire by TEC, the extent to which these cells reproduce peripheral isoform structures is unknown. We performed a transcriptomic investigation of transcript structures and splicing factor expression in medullary TEC and 21 peripheral tissues. Our results indicate that TEC re-use a small number of peripheral splicing factors to recreate a limited subset of the peripheral splice isoform repertoire. We found, for example, that TEC lacked both neuronal micro-exons and the splicing factor, *Srrm4*, which promotes their inclusion. We demonstrate a functional role for the Rbfox splicing factors in promoting medullary TEC development and the alternative splicing of promiscuously expressed genes. Our findings have implications for understanding autoimmune diseases and the development of antigen-specific immunotherapies.

Introduction

T cells are essential for the generation and resolution of an adaptive immune response as they are uniquely able to distinguish between benign self and harmful non-self-antigens. The thymus constitutes the primary lymphoid organ for the generation of naïve T cells tolerant to self-antigens but reactive to foreign molecules [1]. The generation of T cells is critically dependent on physical and functional interactions that occur during their development with different thymic stromal cells. The most abundant of these are thymic epithelial cells (TEC) which are classified based on their positional, structural, antigenic and functional characteristics either as cortical (cTEC) or as medullary epithelia (mTEC) [2]. cTECs control the attraction of blood-borne precursor cells to the thymic microenvironment and their commitment to a T cell fate, promote their subsequent growth and initial maturation, and select immature thymocytes that express a T cell antigen receptor (TCR) [1]. Because the receptor specificity is initially generated pseudo-randomly, the functional utility and self-

reactivity of TCRs must be scrutinized during thymocyte development. Thus, a comprehensive and efficient selection mechanism is in place to establish a TCR repertoire that is bespoke to an individual's needs. First, cTEC positively select thymocytes that express a TCR of sufficient affinity for self-MHC [1]. Subsequently, both cTEC and mTEC deplete thymocytes that display a TCR with high affinity for self-antigens, a process designated thymocyte negative selection[3]. In addition, mTECs contribute to a dominant mechanism of immunological self-tolerance by diverting a subset of self-reactive T-cells to a natural T regulatory cell (nTreg) fate [4]. Selection by TECs therefore authorizes the generation of a diversified T cell repertoire with versatile antigen-recognition capacity. In addition to their role in T-cell selection, mTEC also promote the terminal differentiation of thymocytes.

To achieve a broad representation of self-antigens for the sake of thymocyte negative selection, a molecular mirror of an individual's self-antigens is expressed by TEC and presented to maturing thymocytes. This process, designated promiscuous gene expression (PGE), is essential for the avoidance of autoimmunity and involves the transcription of approximately 89% of protein-coding genes by the TEC population [2,5]. The mechanisms by which TEC defy developmental and tissue-specific transcriptional controls to achieve PGE are only incompletely deciphered [6], although the AutoImmune Regulator (Aire) [5] and the transcription factor Fezf2 [7] have been identified to enable PGE. Importantly, comprehensiveness of self-representation by TEC cannot be solely measured by the number of genes which they promiscuously express as the diversity of peripheral self-peptidome is further elaborated by a variety of post-transcriptional mechanisms. These include alternative mRNA splicing [8], the expression of "untranslated" regions [9], and RNA-editing [10]. The resultant proteome is further varied by proteasome mediated splicing [11] and post-translational modifications [12]. In particular, alternative mRNA splicing greatly increases the complexity of the mammalian proteome, with, for example, there being approximately three times more annotated protein-coding transcripts than there are such genes in mice

and humans [13]. Investigation of splice isoform representation in the thymus is important as there is good evidence that absence of tissue-specific splice isoforms in TEC can preclude negative selection of pathogenic T cells able to incite autoimmunity [14].

In the periphery, the generation of alternative splice isoforms is tightly regulated during development and controlled by temporal and context-specific expression of splicing factors. Splicing factors function by recognising intronic sequence motifs in pre-mRNA molecules – which are typically proximal to exon boundaries – and mediating the actions of the spliceosome to dictate exon inclusion or exclusion. The extent to which TEC are able to reproduce peripheral splice isoforms is unclear but it has been claimed that the transcriptome of TEC is unusually complex at the gene level, both in terms of isoform number [15] and splice-junction representation [10]. However, it is thought that representation of developmentally and tissue-restricted splice isoforms in the thymus may be incomplete. Initial surveys that compared TEC with a limited number of peripheral tissues suggest that these cells might produce only one fifth of tissue-restricted isoforms[15] or only 20-60% of tissue-restricted splice junctions [10]. Because of its direct interactions with several splicing-related factors [16], AIRE has also been alleged to effect alternative splicing [15], although molecular evidence in support of such a function is sparse and would only attest to a minor role [10]. It is however appealing to postulate that mTEC may recommission peripheral factors to create tissue-specific splice variants. In one plausible model, mTEC might constitutively express a specific subset of peripheral factors in order to achieve robust coverage of a limited repertoire of peripheral splice variants. In support of this possibility, a limited transcriptomic survey of RNA-binding proteins identified seven RNA-binding factors to be highly and constitutively expressed in murine mTEC [17]. Alternatively, mTEC might adopt a stochastic splicing routine by employing promiscuously expressed splicing factors to achieve a broader coverage of peripheral splice isoforms from all possible developmental and cellular contexts. Such a process would also likely involve the generation of spurious novel isoforms.

Here we set out to explore the splice isoform landscape of TEC and to investigate the mechanisms responsible for shaping it. Our findings represent the first comprehensive assessment of known and novel transcript structures in TEC and show that these cells accurately reproduce a subset of tissue-specific splice isoforms. To do so thymic epithelial cells re-use a small set of peripheral splicing factors that includes members of the Rbfox family.

Results

Transcriptome complexity in TEC is a consequence of promiscuous gene expression

We began our work by investigating how the splicing landscape of TEC compares to that of peripheral (i.e. non thymic) tissues. We performed deep stranded Illumina sequencing of immature and mature mTEC isolated by flow cytometry and compared the resulting data to that of 21 tissues from the mouse ENCODE project [18]. We constructed a new mouse TEC and Tissue (mT&T) transcriptome assembly to avoid biases in transcript annotation between the TEC and tissues (Fig. 1a). In line with previous reports of high transcriptome complexity in these cells[10,15,17], we found that overall, mature mTEC expressed a greater number of transcripts from protein-coding genes than any of the surveyed peripheral tissues (60.1%, Fig. 1b). However, modelling of the relationship between the number of detected genes and transcripts for the TEC and peripheral tissue samples revealed that mature mTEC produced fewer transcripts per gene than was typically found in peripheral tissues (Fig. 1c).

We next sought to understand the extent to which the splicing of promiscuously expressed genes in TEC recreates that found in the normal tissue context. In order to identify genes promiscuously expressed in mTEC, we first defined a set of “tissue-restricted antigen” (TRA) genes (see methods and Supplementary Fig. 1). We stratified the TRA genes according to

their dependence on *Aire* for expression in mTEC (Supplementary Fig. 1). Regardless of their dependence on *Aire,* we detected fewer isoforms from TRA genes in mTEC than was observed in peripheral tissues (Supplementary Fig. 2a-b). In contrast, a similar fraction of isoforms from non-TRA genes was expressed in mTEC and peripheral tissues (Supplementary Fig. 2c). These observations may reflect the lower sequencing coverage of promiscuously expressed genes in the TEC samples, which is a consequence of the low expression of these genes at the mTEC population level [5] (Supplementary Fig. 2d-f). We found however, that for TRA genes, fewer isoforms were detectable in the mTEC samples regardless of gene expression level (Supplementary Fig. 2g-i).

We observed, as previously reported by others [10,15], that the mature mTEC population co-expresses transcript isoforms that normally arise in distinct anatomical locations (Supplementary Fig. 3a-c). In addition, we found evidence that the thyroid and nervous-system specific transcript isoforms of *Calca* were often produced together in the same single cell (Supplementary Fig. 3d). While mature mTEC expressed a higher number of known transcripts (n=57,019) than was observed in peripheral tissues (Fig. 1d), they produced a relatively low number of novel transcripts (n=36,547, Fig. 1d, analysis restricted to protein-coding genes). Finally, we assessed the representation of sets of tissue-restricted isoforms from the different peripheral tissues in mTEC. Transcripts with testis restricted expression were most markedly under-represented in mTEC (52.5% detected), followed by those from the brain (60.7-62.9%), adrenal (62.7%) and ovary (63.6%) (Fig. 1e). We confirmed the reproducibility of these observations in biological replicate sample pools (Supplementary Fig. 4a-d). In addition, the detection of splice junctions was similarly saturated in TEC and peripheral tissues samples (Supplementary Fig. 4e). Together our results indicated that the atypically high number of transcripts present in TEC was not a result of an excess production of known or novel transcripts at the per-gene level. Rather, this phenomenon

appeared to simply reflect the extraordinary high number of genes that TEC promiscuously express.

Genes harbouring novel transcripts in mTEC are associated with T-cell selection

Alternative splicing events are often associated with the evolution and modification of protein function[19]. As the transcriptome of TEC is relatively understudied, we reasoned that novel splicing events specific to TEC might be associated with their specialised functions for T-cell selection. We therefore first validated the existence of the novel transcripts present in the Illumina sequencing based mT&T assembly in immature and mature mTEC using longer-read Oxford Nanopore sequencing (Supplementary Fig. 5). In mature mTEC, we were able to validate the existence 64.3% of novel transcript structures that were expressed at moderate or high levels (> 10 counts, Fig. 2a).

Next, we identified novel transcripts that were uniquely expressed in mature mTEC or in one of the different peripheral anatomical sites surveyed in this study. To mitigate against representation bias, a single tissue exemplar was selected for groups of related tissues (taking, for example, the cerebellum to represent tissues from the brain; Supplementary Fig. 1a). We found that mature mTEC expressed a higher number of unique novel transcripts than did the cerebellum, but substantially fewer than were found in the testis (Fig. 2b and Supplementary Table 3). The majority (85%) of the 1,572 novel transcripts uniquely expressed in mature mTEC were found for loci that do not require *Aire* for their expression (Fig. 2c and Supplementary Table 3).

Overall, TEC-specific novel transcripts were observed for 1,167 protein-coding genes. These genes displayed significant enrichments for Gene Ontology (GO) biological processes such as “regulation of leukocyte differentiation”, “alpha-beta T cell activation”, “regulation of T cell mediated immunity” and “antigen processing and presentation of peptide antigen” (BH adjusted $p < 0.01$, one-sided Fisher tests, Fig. 2d, and Supplementary Table 4) suggesting

that they are likely to encode for non-promiscuously expressed factors that have a functional role in T-cell selection in TEC. Novel transcript structures were detected in genes of well-established importance for thymic epithelial cell function, including *Foxn1* [20] and *Aire* [6] (Supplementary Table 3). Genes that produced transcripts harbouring novel exons or exon skipping events in TEC included *Cdx1*, a transcription factor that has been linked with mTEC maturation [21] (Fig. 2e); *Cd80*, a receptor important for interaction with T-cells via CD28 and CTLA4; the MHC class II gene H2-Aa; *Mill1*, an MHC class I-like molecule that is known to be expressed on a subpopulation of TEC [22] (Fig. 2f) as well as Skint gene family members, including *Skint2* which has been reported to be a novel negative T cell regulator [23].

In summary, our data show that the novel transcript structures detected in thymic epithelial cells arose, in large part, from non-promiscuously expressed genes likely to be relevant for T-cell selection.

Aire contributes to alternative splicing and promotes expression of long transcripts

We next set out to clarify the role of Aire in alternative splicing in mature mTEC. To do so, we generated deep stranded Illumina sequencing data from mature *Aire*-positive and mature *Aire*-knockout mTEC. We discovered 492 significant (5% FDR, rMATS analysis) *Aire*-regulated alternative splicing events in transcripts arising from 459 protein-coding genes (Fig. 3a and Supplementary Table 5). We found, however, a much larger number of significant differences in transcript splicing between the immature and mature wildtype mTEC samples (n=2,236 events in n=1,967 protein-coding genes, 5% FDR, rMATS analysis, Fig. 3a and Supplementary Table 5). Analysis of the differences in splicing events between immature wildtype and mature *Aire*-knockout mTEC confirmed the relatively limited contribution of *Aire* to alternative splicing in these cells (Fig. 3a). Of note, the majority of the *Aire*-controlled and TEC maturity-related alternative splicing events were found in transcripts encoding non-TRA genes (Fig. 3b). In addition, a large number of introns were found to be retained in the transcripts of immature mTEC (n=760, Fig. 3a). Coordinated changes in

intron retention during cellular differentiation are not unusual, and transcripts harbouring retained introns can arise from genes that have specialised cellular functions [24].

Previous studies have reported that AIRE promotes the expression of thousands of distal exons [25] and the release of stalled polymerases [26]. We therefore investigated whether *Aire* might favour the production of long transcripts. Differential transcript usage analysis (Fig. 3c) revealed that *Aire* positively regulated the production of a large number of transcripts (n=4,027, BH adjusted $p < 0.05$, fold change $\gtrapprox 2$; Supplementary Table 6). Transcripts arising from *Aire*-regulated genes had a bi-modal length distribution that comprised of distinct concentrations of shorter (peak at 678bp) and longer transcripts (peak at 2,249bp) (Fig. 4d). *Aire*-regulated transcripts from these genes had a significantly different distribution that consisted of only a unimodal peak of the longer transcripts ($p=2.2 \times 10^{-16}$ Kolmogorov-Smirnov test, Fig. 4d). On average, *Aire*-regulated transcripts were over 1kb longer than their non-*Aire*-regulated counterparts (Wilcoxon test, $p < 0.05$). These findings show that *Aire* plays a major role in promoting the production of long transcripts while only having a limited impact on the generation of alternative splicing events. Hence, splicing factors other than *Aire* must be primarily responsible for the alternative splicing of transcripts in mature mTEC.

Medullary TEC express a distinct set of splicing factors that includes *Rbfox1*

The ability of mTEC to collectively express a large fraction of peripheral tissue-restricted transcripts (Fig. 1e) suggested that these epithelia might recommission mechanisms of peripheral splicing. To investigate this possibility we compiled a set of 674 splicing-related genes using information from the Gene Ontology consortium, the RBPDB database[27] and the literature[17,28-33] (Supplementary Fig. 6a). Using the tissue-specificity metric *tau*, we identified a set of 146 tissue-restricted splicing related factors (TRSF, *tau* > 0.5) (Supplementary Figure 6b and Supplementary Table 7) that included 24 factors with

demonstrated roles in the control of alternative splicing (Fig. 4a, Supplementary Table 8). Of these factors, the frequent (>20% of single cells) and non-promiscuous expression of *Rbfox1*, *Rbm20* and *Msi1* in mature mTEC was interesting because they showed little, if any, expression in skin epithelia (Fig. 4a). The specificity of their expression suggested that they may be of importance for the specialised functions of mTEC. Rbfox1 and Rbm20 have well established roles in alternative splicing in the brain and heart [34,35] while Msi1 is an RNA binding protein that has previously been implicated in regulating splicing in photoreceptors [36]. Meanwhile, it was notable that the majority of the 24 TSRFs, including those restricted in their expression to the brain (such as *Nova1*, *Nova2, Elav2, Elav3, Elav4* and *Rbfox3*), testis (*Brdt*) and heart (*Rbm24*), were expressed at low frequency or showed only promiscuous expression in mature mTEC (blue asterisks, Fig. 4a).

To examine the potential consequences of the absence of many TRSFs from mTEC we also investigated patterns of coding exon inclusion in these cells. In line with the absence of neural and testis-specific splicing factors (Fig. 4a), sets of exons frequently included in transcripts detected in the brain and testis were found to be excluded from mRNA recovered from mTEC (PSI < 0.1, Fig. 4b). In particular, we noted that brain specific micro-exons (≤ 30 bp) were rarely included in mTEC transcripts (Fig. 4c). This observation is likely explained by the low expression of *Srrm4* in mTEC (Fig. 4c) as this factor is known to be responsible for promoting the inclusion of neuronal micro-exons [37,38]. In summary we found that mTEC constitutively express only a small number of peripheral spicing factors. Together with the clear absence of a set of neuronal micro-exons, these results suggest that mTEC may have evolved to represent only specific facets of the peripheral tissue isoform repertoire.

RBFOX is present with AIRE in mTEC nuclei and promotes mTEC development

Amongst the peripheral splicing factors apparently recommissioned by mTEC, *Rbfox1* was of particular interest due to the well-established roles of the Rbfox factors in controlling alternative splicing during the development of tissues such as muscle and brain [34]. We

therefore investigated the expression of the Rbfox factor genes and transcript isoforms in mTEC in more detail (Supplementary Fig. 7). On close inspection, we found that mTEC produced transcripts from *Rbfox1* and *Rbfox2* that predominantly included the neuronal B40 exon and excluded the muscle-specific M43 exon [39] (Supplementary Fig. 7b, c). Confocal image analysis of AIRE-expressing mTEC employing an antibody that recognises the RRM domain common to all Rbfox homologues revealed the molecules' nuclear location and physical proximity to AIRE speckles (Fig. 5a).

Based on the transcriptomic analyses, both *Rbfox1* and *Rbfox2* showed a robust non-promiscuous and tissue-restricted expression in mTEC (at gene or transcript level, Fig. 4a, Supplementary Fig. 7). To investigate their specific roles in transcript splicing in mTEC, we crossed mice in which *Rbfox1* and, or, *Rbfox2* exons were homozygously flanked by loxP sites [40,41] with animals heterozygously expressing Cre recombinase under the transcriptional control of the TEC-specific *Foxn1* locus [42]. The resultant *Rbfox1* thymus knockout (tKO) (*Rbfox1*$^{lox/lox}$:*Foxn1*$^{cre/+}$), *Rbfox2* tKO (*Rbfox2*$^{lox/lox}$:*Foxn1*$^{cre/+}$) and double *Rbfox1*/*2* tKO (*Rbfox1*$^{lox/lox}$: *Rbfox2*$^{lox/lox}$:*Foxn1*$^{cre/+}$) mice were then analysed for changes in thymic cellularity and TEC phenotype. *Rbfox1* tKO mice displayed a 15% increase in thymus cellularity whereas that of age-matched *Rbfox2* tKO mice was unaffected when compared to *Rbfox2*$^{fl/fl}$ cre- littermate controls (Supplementary Fig. 8c, Supplementary Fig. 9c). The phenotype of cTEC (CD45-EpCAM+Ly51+ve, UEA-1-ve), mTEC (CD45-EpCAM+Ly51-ve, UEA-1+ve), immature mTEC (CD45-EpCAM+Ly51-ve, UEA-1+ve CD80-low, MHCII-low), and mature mTEC (CD45-EpCAM+Ly51-ve, UEA-1+ve CD80-high, MHCII-high, either AIRE+ or AIRE-) was unaffected in *Rbfox1* tKO mice (Supplementary Fig. 8). In contrast, *Rbfox2* tKO mice showed an overall reduction in TEC frequency (CD45-EpCAM+, *Rbfox2* tKO 0.12 % vs WT 0.17 % of all thymic cells recovered, $p < 0.05$, two-sided Welch Two Sample t-test, Supplementary Fig. 9d, e). The *Rbfox2* tKO mice also had proportionally more cTEC (10.4 % vs 6.7 % of all TEC, $p = 0.02$, Welch Two Sample t-test) and proportionally fewer mTEC (86.2 % vs 90.8 % of all TEC, $p = 0.03$, Welch Two Sample t-test) present in

their epithelial scaffolds (Fig. 5b-c). *Rbfox1/2* tKO animals showed a TEC population composition that was quantitatively similar to that of *Rbfox2* tKO mice (Supplementary Fig. 10) suggesting that, as in other tissues [41], *Rbfox2* can compensate for the loss of *Rbfox1* in thymic epithelial cells. Thymocyte development and positive selection was quantitatively normal in the Rbfox1/2 tKO animals (Supplementary Fig. 11). In summary these data demonstrated that *Rbfox1* contributes to the regulation of thymic cellularity and that *Rbfox2* acts to promote a medullary TEC cell fate.

Rbfox contributes to the alternative splicing of self-antigen transcripts in mTEC

To investigate the impact of *Rbfox1* and *Rbfox2* on the mTEC transcriptome we performed RNA-sequencing of immature and mature mTEC flow cytometrically isolated from *Rbfox1* tKO, *Rbfox2* tKO and control mice. In mature mTEC loss of *Rbfox1* or *Rbfox2* induced only minor changes in gene expression (Fig. 6a) but caused hundreds of alterations in alternative splicing (*Rbfox1* tKO: n=559 events in 535 genes, FDR < 0.05, |delta PSI|>0.2; *Rbfox2* tKO: n=668 events in 624 genes, FDR <0.05, |delta PSI|>0.02, Fig. 6b, c, and Supplementary Tables 9 and 10). In mature mTEC, *Rbfox1* and *Rbfox2* controlled the alternative splicing of 123 events in 104 TRA genes and 122 events in 110 TRA genes, respectively. In immature mTEC *Rbfox1* and *Rbfox2* exerted similar effects but controlled the alternative splicing of a smaller number of TRA genes (72 events in 51 TRA genes and 43 events in 36 TRA genes, respectively) in keeping with the lower level of promiscuous gene expression in these cells (Supplementary Figure 12 a-c). The splicing of 10 genes was regulated by both *Rbfox1* and *Rbfox2* in mature mTEC (odds ratio=4.64, $p=1.71 \times 10^{-4}$, two-sided Fisher Exact test) while the splicing of 6 genes was regulated by both factors in immature mTEC (odds ratio=18.5, $p=3.23 \times 10^{-6}$, two-sided Fisher Exact test).

We next sought to establish whether the genes spliced by Rbfox in mature mTEC overlapped with those previously predicted to be Rbfox targets in the mouse brain[43]. In

mature mTEC we identified particularly large and significant overlaps between predicted Rbfox target genes and those for which Rbfox2 controlled (i) exon skipping in mTEC (OR=7.8, BH adjusted p=5.4 x 10^{-19}) or (ii) the use of mutually exclusive exons (OR=13.8, BH adjusted p = 1.4 x 10^{-5}) in mTEC (Fig. 6d). In keeping with this observation, many of the TRA genes alternatively spliced by Rbfox2 showed specific expression in the neuronal tissues (Supplementary Fig. 13). GO biological processes over-represented amongst the genes alternatively spliced by Rbfox2 in mature mTEC included “muscle contraction” and “neuron migration” in line with the known roles of Rbfox family members in muscle and neuronal tissues [34] (Fig. 7a). Examples of genes harbouring Rbfox2 mediated alternative splicing events in mature mTEC included *Fn1* and *Insr*, two known Rbfox2 target genes [29,43] as well as *Myom2*, a gene that normally has tissue-restricted expression (Fig. 7b).

Previous studies have identified an enriched Rbfox recognition motif in proximity to Rbfox regulated exons [43]. We investigated the enrichment of Rbfox binding motifs around exons whose inclusion was regulated by Rbfox. In both immature and mature mTEC we observed a significant enrichment of the conserved Rbfox RRM WGCAUGM motif [44] upstream of exons repressed and downstream of exons enhanced by *Rbfox2* in both mature and immature mTEC (Fig. 7c and Supplementary Fig. 12d), following the patterns previously described for this factor [43]. In summary these data provide evidence that Rbfox splicing factors directly regulate the splicing of both promiscuously and non-promiscuously expressed genes in mature mTEC.

Discussion

Our investigations revealed that, at the population level, mature mTEC express an unprecedented number of splice isoforms. We found that this phenomenon involves the generation of nearly sixty percent (58.5%) of all peripheral splice variants, a number higher than that found in other tissues with complex transcriptomes such as the brain or testis. In

contrast to previous reports, we found, however, that isoform production in mTEC is not unusually complex at the gene level. In fact, in mTEC, we detected fewer isoforms per promiscuously expressed gene than were found in the corresponding peripheral tissues in which these genes were normally expressed. While this observation might be explained in part by the lower expression of promiscuously expressed genes at the population level in mTEC, the concept that not all peripheral isoforms are represented in TEC is also supported by a recent targeted qPCR-based study which estimated that a quarter of the genes studied contained epitopes hidden from the thymus[45]. Together with our finding that mTEC favour the production of known rather than novel transcripts this suggests that these cells do not employ "promiscuous" mechanisms to promote and increase splice isoform diversity as has previously been suggested [15]. Contrary to such an idea, the novel transcripts that we observed in mTEC were enriched in genes involved in biological processes known to be important for T-cell selection, which is the principle function of these cells. It therefore seems likely that evolution of the unique abilities of thymic epithelial cells has involved use of alternative splicing to specialise protein function in these cells. Our data may hence provide important clues into the genes and pathways that enable mTEC to produce, traffic and present self-antigens.

Before comparing the representation of transcripts from different tissues in mTEC, we mitigated against evident biases in annotation, sequence duplication and sequencing depth by constructing a common reference transcriptome assembly, de-duplicating reads and performing downsampling. Having done so, we found a remarkably even representation of tissue-restricted transcripts from the surveyed peripheral tissues in mTEC. Notwithstanding, we found a lower representation of transcripts from the testis and brain in mTEC together with a corresponding absence (or only very weak expression) of brain and testis-specific splicing factors in mTEC. Examples of such splicing factors included *Nova1/2* and *Elavl2/3/4* which are highly expressed in the brain and *Brdt* which is strongly expressed in the testis.

One explanation for these observations is that the brain and testis are immune privileged tissues and that it may therefore be less important to educate T cells against epitopes derived from splice-isoforms specific to these tissues in the thymus. Further, the unusually high complexity of splicing in the testes and brain suggests the possibility that immune surveillance may act as a constraint on the evolution of splicing complexity in non-immune privileged peripheral tissues. Our finding that neuronal micro-exons are not frequently spliced into transcripts in mTEC is likely explained by the absence of *Srrm4* expression in these cells, as the SRRM4 splicing factor has been shown to mediate inclusion of neuronal micro-exons [38]. The limited inclusion of neuronal micro-exons in mTEC provides a possible link between the observations that such exons are mis-regulated in the brains of patients with autism[46], and growing evidence that autoimmunity may be involved in autism [47]. Aside from the testis and cerebellum, adrenal and liver tissues were the most incompletely represented in mature TEC. In humans, these tissues are known sites of autoimmunity: autoimmune adrenalitis is the most common cause of Addison's disease, and liver diseases such as autoimmune hepatitis, primary biliary cirrhosis and schelerosing cholangitis are all thought to be the consequence of autoimmunity [48]. Alternative splicing has been implicated in such diseases [49], and our data suggest that lack of thymic representation of isoforms specific to these tissues might contribute to a susceptibility to autoimmune attack.

Our investigations revealed a large number of differences in splicing between immature and mature mTEC. One possible explanation for these observations was the expression of *Aire* in the mature cells. Analysis of *Aire*-knockout mTEC showed however that while *Aire* does promote the use of long transcripts, in line with the concept that it releases stalled RNA polymerases [26], it only has a small role in alternative splicing in mTEC. A comprehensive survey of splicing factor expression revealed that mTEC complement typical epithelial splicing factors such as *Espr1/2* with a small number of peripheral restricted splicing factors. Unlike skin epithelial cells, mature mTEC showed non-promiscuous expression of *Rbm20*,

Msi1 and *Rbfox1*. This observation, along with the apparently incomplete representation of peripheral structures and absence of excessive numbers of novel splice junctions in mTEC, suggests that TEC undertake a conservative program of alternative splicing to ensure the accurate representation of a limited subset of the peripheral splice isoform repertoire. We hence conclude that transcript structures in mTEC are primarily shaped by a small number of splicing and mRNA processing factors[50] rather than being heavily influenced by stochastic use of promiscuously expressed splicing factors. Importantly, if conserved in humans, limited thymic representation of peripheral splice isoforms is consistent with the concept that tissue-specific isoforms are relevant sources of auto-antigens in immune-mediated diseases [51-53]. Furthermore, characterisation of transcript structures in human thymic epithelial cells would be expected to aid the identification of autoantigen transcripts encoding untolerised epitopes[54]. Knowledge of such epitopes is of great value as they can be used for development of antigen-specific therapies for autoimmune disease such as those based on, for example, the use of tolerizing peptides or tolerogenic dendritic cells[55,56].

The discovery of non-promiscuous *Rbfox1* expression in TEC was of particular interest because this factor is otherwise restricted to muscle and neural tissues in which it plays important developmental roles [34]. We detected *Rbfox1* in medullary but not cortical TEC suggesting that it may be important for the development or function of this subpopulation. We therefore performed functional analysis of the role of *Rbfox1* and its homologue *Rbfox2* in mTEC, excluding *Rbfox3* from our investigations as it showed only weak and promiscuous expression in these cells. Phenotypic analysis of TEC in these animals identified a role for *Rbfox2* in promoting the generation of mTEC at the expense of cTEC. Recently, it has been shown that whilst in the embryonic and new-born thymus cortical and medullary TEC arise from a common bipotent progenitors, in the adult mouse mTEC are replenished by lineage-restricted cells [57]. The relative increase that we observed in cTEC numbers at the expense of mTEC in 4-6 week old animals hence suggests non-exclusive roles for *Rbfox2* in favouring the mTEC fate choice of bipotent progenitor cells or in promoting the differentiation

of mTEC restricted progenitor cells. In mature mTEC we found that Rbfox factors shape the splicing of both promiscuously and non-promiscuously expressed genes. The larger role identified for *Rbfox2* in this process was not unexpected, as, in the cerebellum it is known that while *Rbfox2* can largely compensate for loss of *Rbfox1*, *Rbfox1* is less well able to ameliorate an absence of *Rbfox2* [40]. In addition, the changes in alternative splicing identified following loss of *Rbfox2* may involve *Rbfox1* as it was itself differentially spliced in the absence of *Rbfox2*. We found examples of both *Aire*-dependent and *Aire*-independent TRA that were alternatively spliced by *Rbfox2* in mTEC. Given the weak expression of TRA in mTEC and the relatively small amount of biological material sequenced for this analysis, we expect the actual number of TRA regulated by Rbfox factors in mTEC to be substantially higher than that reported here. In conclusion our data suggest that mTEC re-use a small set of peripheral splicing factors that includes Rbfox in order to conservatively reproduce a specific subset of the peripheral splice isoform repertoire.

Methods

Mice

Wildtype C57BL/6 mice were originally obtained from either Harlan Laboratories or Janvier and maintained as a laboratory in-house colony. *Aire*$^{GFP/+}$ mice were previously described [5]. *Rbfox1* and *Rbfox2* mutant mice [40,41] were maintained on a mixed 129S2/Sv x C57BL/6J background. All animals were kept under specific pathogen-free conditions and experiments were carried out in accordance with local and national regulations.

Extraction of thymic epithelial cells and assessment by flow cytometry

Thymic epithelial cells were extracted as previously described[58]. In short, thymic lobes were incubated with Liberase (Roche) and DNase (Roche) in PBS for 30 min at 37 °C. The cells were incubated with magnetic beads for 15 min at room temperature followed by enrichment

of CD45-negative cells using the AutoMACS Pro Seperator (Miltenyl Biotech).
Enriched cells were stained with antibodies against CD45-AF700 (1:1000, 30F11; BioLegend), EpCAM- PerCPCy5.5 (1:1000, G8.8; BioLegend), Ly51- PE (1:200, 6C3; BioLegend), UEA-1-Cy5 (1:500, Vector Laboratories, in-house labelled), MHCII-BV421 (1:1000, M5/114.15.2; BioLegend), CD80-PE-Cy5 (1:1000, 16-10A1, Biolegend), CD86-PE-Cy7 (1:1000, GL-1, Biolegend). Staining was performed at 4ºC in the dark. DAPI or the LIVE/DEAD Fixable Aqua Dead Cell Stain Kit (Thermo Scientific) was used as a live/dead staining (Supplementary Fig. 14). Cells were sorted using FACSAria III (BD Bioscience) and data was analyzed using the FlowJo software (version 10.5.0).

Immunofluorescence

Freshly dissected thymus lobes were frozen in OCT (Tissue Trek). 8 µm tissue sections were cut using a Cryostat (Thermo Scientific CryoStar NX70 with MB DynaSharp Microtome Blade). The tissue sections were fixed for 20 min in 1.4 % PFA (Sigma, in 1xPBS) and for 10 min in Methanol (VWR). Permeabilisation was performed for 10 min in 0.3 % Triton-X (Sigma, in 1xPBS). These steps were followed by one 5 min washing step in 1xPBS, marking of the individual sections by a hydrophobic PAP pen (Sigma) and two further 5 min washing steps of each individual section. The primary antibodies were diluted in 1xPBS containing 10% goat serum, 0.3% Triton-X and incubated for 45 min at 37 °C. Primary antibodies were directed against AIRE (5H12, eBioscience, 1:500) and the RRM domain (1:500). In addition, UEA-1 was used for staining (1:150, Vector Laboratories, in-house labelled). After three washing steps, the secondary antibody was added (diluted 1:500 in 1xPBS, goat α rabbit- AF488 (Invitrogen), goat α rat- AF555 (Invitrogen)) and incubated for 30 min at 37 °C. Subsequently, three washing steps and a 90 min UEA-1 staining at 37°C (1:150 dilution in 1xPBS) were performed. Subsequently, two washing steps were followed by DAPI staining (10 min, 1:10,000 in methanol). After a final washing step slides were mounted using ProLong Gold antifade reagent (Invitrogen).

RNA-sequencing data

For the analysis of immature, mature and *Aire*-knockout mTEC poly(A)+ RNA-sequencing libraries were prepared from cells pooled from multiple mice (1 µg of total RNA; 4 weeks old; n=2 biological replicates) and subjected to 101 bp paired-end stranded Illumina RNA-sequencing. For the analysis of *Rbfox1* tKO and *Rbfox2* tKO animals, immature and mature mTEC (CD45-, EpCAM+, Ly51-, UEA+) were isolated from individual mice and littermate controls (10-15k cells per animal, n=2 biological replicates, 4 weeks old). Stranded RNA-sequencing libraries were prepared (NEB) and subjected to 150bp paired-end Illumina sequencing. Illumina sequencing was performed using the HiSeq 4000 instrument. For further details of the RNA-sequencing datasets see Supplementary Table 1.

For long-read sequencing mature mTEC were obtained from 4-6 week old female wildtype C57BL/6 mice and processed as described above. RNA from EPCAM+, UEA-1+, CD80/86 positive or negative cells was extracted using the RNeasy® Plus Micro Kit (Qiagen). After RNA quality control, 8µl of the extracted RNA was prepared for sequencing using the cDNA-PCR Sequencing Kit (SQK-PCS108, Oxford Nanopore Technologies) following manufacturer's protocol (version: PCS_9035_v108_revF_26Jun2017; update: 31/05/2018). cDNA libraries were amplified for 18 cycles and normalized to 400 fmol before sequencing on Oxford Nanopore MinION flow cells (FLO-MIN 106 R9.4) for 48 hrs.

RNA-sequencing data for the 21 peripheral mouse tissues were obtained from the ENCODE project (GSE36025; 8-week old mice, keeping only the colon samples to represent the large intestine). RNA-sequencing from skin epithelial cells (GSM1094285)[59], cTEC (GSE53111)[5], and single mature mTEC (GSE114713)[21] shown in Fig. 4 was obtained from the Gene Expression Omnibus (GEO).

Computational methods

Generation of the mT&T transcriptome assembly

Sequence reads were trimmed to 76 bp (fastx trimmer v0.0.14) and mapped with Hisat2[60] (v2.1.0; mouse genome plus known splice junctions; Ensembl mm10-v91; settings:"–dta –score -min L,0.0,-0.2 –rna-strandness RF"). To generate a single high-depth sample for each tissue and TEC population replicate samples were combined and downsampled to 200M reads (samtools[61] v1.3.1; Supplementary Table 1). Separate reference-guided assemblies were then constructed for each ENCODE tissue or TEC population using the high-depth samples (StringTie[62] v1.3.3b; Ensembl mm10-v91).

To identify transcripts that were reproducibly detected we first prepared biological replicate sample pools for each tissue and TEC population (n=2; 60M reads/samples, Supplementary Table 1). Expression of the transcripts present in each of the assemblies was then quantified in the relevant tissue or TEC population using the two 60M replicates (Salmon [63], v0.11.3, with parameters: "--incompatPrior=0 --validateMappings --rangeFactorizationBins=4 --seqBias --gcBias -x 0.66"). We then implemented and applied a robust procedure based on computation of the non-parametric Irreproducibility Discovery Rate (npIDR)[18,64] (Supplementary Methods and Supplementary Fig. 15). Transcripts that were reproducibly detected according to this procedure (those expressed above a level determined to correspond to npIDR ≤ 0.1) from each tissue or TEC population were merged into a single unified assembly (StringTie, Ensembl reference annotation guided merge). Transcripts contained in reference introns, possible polymerase run-on fragments, repeats, transcripts overlapping opposite-strand exons or introns and possible pre-mRNA fragments were removed (gffcompare v.0.10.6, class codes 'irpxse'). To be included in the final mT&T assembly, the merged transcript models were additionally required to be reproducibly detected (expressed above a level determined to correspond to npIDR ≤ 0.1; procedure as above) in at least one tissue or TEC population. The final mT&T assembly was used as the annotation for all subsequent steps if not otherwise indicated.

Quantification of gene and transcript expression levels

For comparisons of gene and transcript expression levels between the ENCODE tissues and TEC populations samples were deduplicated (Picard v2.10.9), filtered to exclude unmapped reads, and downsampled to common read depths (Fig. 1, Supplementary Figs 2, 4). Transcripts-per-million (TPM) values were obtained using Salmon with an index created from the new mT&T assembly (k=31 for index, settings: “ISR --gcBias”) and upper-quartile normalized. Reads were counted by FeatureCounts[65] (v1.6.0).

RNA-sequencing reads from the *Rbfox1* and *Rbfox2* tKO mice were not trimmed but were otherwise mapped, deduplicated, filtered to remove unmapped reads, downsampled and quantitated as described above (retaining 14.5 M and 10M paired-end reads/replicate for mature and immature mTEC, respectively).

Identification of tissue-restricted antigen and Aire-regulated genes

Tissue-restricted antigen (TRA) genes were defined as the set of protein-coding genes that showed evidence of tissue restricted expression amongst the ENCODE tissues and wildtype mTEC samples. To avoid representation bias when computing expression specificity, we first identified groups of similar peripheral tissues by hierarchically clustering the tissues according to their transcript expression profiles. One tissue was then selected to represent each of the groups identified (Supplementary Fig. 1a). TRA genes were defined as those with a *tau* [66] value > 0.7 in the representative tissues and wildtype immature and mature mTEC samples (see Supplementary Methods and Supplementary Figure 1b). *Aire*-regulated genes were defined as those significantly downregulated more than 2-fold in the homozygous *Aire*-knockout samples relative to the heterozygous *Aire*-knockout samples (BH adjusted $p < 0.05$, DESeq2 [67] analysis, n=2 biological replicates, Supplementary Figure 1c). For the identification of *Aire*-regulated genes untrimmed, full-depth sequence data for the mTEC populations was mapped (as above) and quantified with FeatureCounts (Ensembl

v91). In total, we identified n=3,889 *Aire*-regulated tissue-restricted antigen genes (*Aire*-TRA), n=5,266 other tissue-restricted antigen genes (non-*Aire* TRA) and n=12,885 non-TRA genes (Supplementary Fig. 1d and Supplementary Table 2).

Identification of novel tissue-restricted transcripts

Novel transcripts were defined as those without a match in the Ensembl annotation (as assessed with gffcompare). Tissue-restricted novel transcripts were identified as those with *tau* > 0.99 [66] in the representative tissues (Supplementary Fig. 1a) and wildtype mature mTEC. In addition, they were required to have an expression level that was > 2 fold higher in a single representative tissue vs all of the other representative tissues (Fig. 2b). *Tau* values were computed using upper-quartile normalised, log2(n+1) transformed TPM values from the representative tissues. Novel transcripts were annotated using the *generateEvents* function in SUPPA [68] (v2.3) (Fig. 2c).

Assessment of differential splicing and splice junctions

Differential splicing events were identified using rMATS [69] (v3.2.5; default settings, filtered for FDR < 0.05; mapping without softclipping; mT&T assembly as annotation). Exon percentage spliced-in (PSI) values were computed using SUPPA (3' and 5'UTR regions excluded from the analysis; event-centric mode). Splice junctions (SJs) were counted using SJcounts (v3.1; settings: "-maxnh 1 -read1 0 -read2 1"). The results were post-processed to split the multi-junction counts into individual junction counts for final quantitation. SJs were assigned to protein coding genes by intersecting their coordinates (±1bp) with those of gene exons (Ensembl v91 annotations, bedtools window (v2.25.0)).

Definition of *Aire*-regulated transcripts

Differential transcript expression between *Aire*-knockout and *Aire*-positive mTEC was assessed using Kallisto[70] (v0.43.1, n=1,000 bootstraps, using Ensembl v91 annotations) and

the Sleuth R package [71] (v0.29). A value of the Sleuth 'b' parameter of log(1.77) was determined to correspond to a 2-fold change (by modelling of actual expression values from Kallisto). Transcript lengths were obtained using the function *transcriptLengths* from the R package Genomic Features (v1.30.3).

Analysis of ONT data

Basecalling of ONT raw tracks was performed using Albacore (v2.1.10) and 'pass' reads (mean quality score of < 7) were trimmed using Porechop (v0.2.3). Merged reads were mapped with minimap2 [72] (v2.9-r720, settings: "-L -ax splice", mm10 genome). Alignments were filtered for mapping quality > 20. The number of splices per read and gap-compressed identity ('de' tag) were extracted from the resulting bam file. Gene expression was quantified using FeatureCounts (settings: "-s 0 -L –fracOverlap 0.8"). Validation of novel mT&T transcripts was performed after first excluding mT&T transcripts with retained introns (defined by SUPPA) and mT&T transcripts arising from genomic loci with overlapping gene models (on opposite strands).

Analysis of single-cell RNA-sequencing data

The data (GSE114713) was mapped (Hisat2, parameters: "--dta --score-min L,0.0,-0.2"), quality controlled and quantitated (Cufflinks v2.21, FeatureCounts, Ensembl v91) using pipeline_scrnaseq.py (https://github.com/sansomlab/scseq). 201 cells with < 50% ERCC spike-in sequences, > 50,000 read pairs, > 2500 genes (Cufflinks), > 5% spliced reads, < 50% duplication rate, <1.3 fold 3' bias and > 70% high-quality reads aligned were retained for further analysis. Fractions of single cells expressing genes were determined based on counts from FeatureCounts.

Geneset over-representation analysis

Geneset over-representation analyses were performed using one-sided Fisher Exact tests (https://github.com/sansomlab/gsfisher). For analysis of genes harbouring novel transcripts

in mature mTEC (Fig. 2d) genes with upper-quartile normalized, log2(n+1)-transformed TPMs > 0.1 were used as the background geneset. For the analysis of differentially spliced events in the *Rbfox2* tKO dataset (Fig. 7a), the foreground geneset was comprised of genes with differentially spliced events (|delta (d)PSI| > 0.2 and FDR < 0.05) and the background geneset comprised of the set of genes that were tested for differential splicing events (rMATS analysis).

Motif enrichment analysis

The tool MATT [73] was used to determine motif enrichment in proximity to significant skipped exon events (*matt rna_maps*, v1.3.0). As input, exons with enhanced inclusion or exclusion (|dPSI| > 0.2, FDR < 0.05) in the *Rbfox2* tKO samples relative to their cre- littermate controls were used. Unregulated exons (|dPSI| < 0.05) were used to compute a control enrichment profile.

Downstream data analysis and data visualisation

Data analysis was performed in Python (jupyter notebooks; pandas v0.17.1) or R (RStudio; R v3.4). Heatmaps were leaf-optimised using the R cba library (v0.2-17). Genomic tracks were visualised using the R package Gviz (v1.22.3) or the *sashimi_plot* function from MISO[74] (v0.5.3).

Data availability

The RNA-sequencing data generated for this study is available at the NCBI Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo/) under the SuperSeries accession number GSE145931.

Acknowledgements

This project was initiated with funding from the Medical Research Council (MRC) CGAT programme [G1000902]. SNS and GAH were supported by funding from the Wellcome Trust

(#066521). SNS and MA are supported by funding from the Kennedy Trust for Rheumatology Research (KTRR). KJ was supported by a Wellcome Trust PhD studentship and is supported by funding from the MRC (MR/S025308/1, MR/S035850/1). We thank Professor Chris Ponting (University of Edinburgh) for support and insightful advice. The anti-RRM antibody and Rbfox mutant animals were kind gifts of Professor Douglas Black and Julia Nikolic (University of California, Los Angeles). We also thank Professor Black for helpful comments and suggestions.

Author contributions

K.J. performed the experiments and computational analyses. N.S.D. generated the RNA-sequencing data from the wildtype and *Aire*-knockout mTEC populations. S.M. assisted with the flow cytometry experiments. M.A. and M.L. performed, and D.B. was responsible for, the Oxford Nanopore library construction and sequencing. S.N.S. conceived the study with input from G.A.H. K.J., G.A.H. and S.N.S. designed the experiments and analyses, interpreted the results and wrote the manuscript. G.A.H. and S.N.S. supervised the study.

Figure legends

Main figures

Figure 1: Comparative analysis of transcript expression in mTEC and peripheral tissues. (a) Generation of a common mouse mTEC and peripheral Tissues (mT&T) transcriptome assembly (for further details see Methods). (b) The curves show the fractions of transcripts from protein-coding genes that were detected in the peripheral tissues and mTEC populations across a range of TPM thresholds (mT&T assembly). (c) The scatter plot shows the relationship between the number of genes and transcripts that were detected in the peripheral tissues and mTEC populations (mT&T assembly). (d) The scatter plot shows the relationship between the number of known (Ensembl version 91) and novel (mT&T assembly) transcripts detected in the peripheral tissues and mTEC populations. (e) The fractions of sets of tissue-restricted transcripts (*tau* ≥ 0.9) from peripheral tissues that were detected in mature mTEC (mT&T assembly). Analyses shown in (b)-(e) were restricted to protein-coding genes and performed using a single high-depth sample per tissue. Similar results were obtained using lower-depth biologically replicate sample pools (n=2, Supplementary Fig. 4). Trend lines shown in (c) and (d) were fitted to all samples except for those from TEC and the testis.

Figure 2: Identification and characterisation of novel TEC-specific transcripts. (a) Validation of novel mT&T transcripts using Oxford Nanopore Technology (ONT) RNA-sequencing. The fraction of novel transcripts that are supported by ONT reads are shown for mature mTEC (red) and immature mTEC (yellow) (see Methods). (b) The number of novel transcripts that were “uniquely” detected in each of the mature mTEC and representative peripheral tissue samples (see Methods and Supplementary Fig. 1a). (c) Breakdown of the TEC-specific novel splicing events by event type and promiscuous expression status

(Supplementary Fig. 1). Each event was analysed separately, hence one novel transcript can contribute to event counts in multiple categories. SE = skipped exon, RI = retained intron, MX = mutually exclusive exon, A3/A5 = alternative 3'/5' splice site, AF/AL = alternative first/last exon. (d) Selected GO biological processes over-represented amongst the protein-coding genes (n=1,167) that contained the mTEC-specific novel splice variants (one-sided Fisher Exact tests, BH adjusted $p < 0.01$).
(e) and (f) Examples of novel TEC-specific transcripts (red) in the genes *Cdx1* and *Mill1*. The existence of the novel transcripts in mature mTEC was supported by both the Illumina (sashimi plots) and long-read ONT (selected reads) sequencing data. Novel transcript identifiers are indicated by the "_N" suffix. Novel transcript names given in red represent TEC-specific novel transcripts.

Figure 3: *Aire* promotes the generation of long transcripts. (a) The barplot shows the numbers of protein-coding genes (x axis) in which different categories of differential splicing events (y axis) were detected between immature and mature mTEC (blue), immature mTEC and *Aire*-knockout mTEC (red) and *Aire*-knockout and *Aire*-positive mature mTEC (yellow) (rMATS, per-analysis FDR < 0.05, |delta PSI| > 0.2, n=2 replicates per sample). (b) Breakdown of the identified splicing events by event type and promiscuous expression status (Supplementary Fig. 1). SE = skipped exon, RI = retained intron, MXE = mutually exclusive exon, A3SS/A5SS = alternative 3'/5' splice site. (c) The MA plot shows differential transcript expression in *Aire*-knockout compared to *Aire*-positive mTEC. Transcripts regulated by *Aire* are shown in red (Sleuth, Wald test, qval < 0.05, fc $\gtrapprox$ 2, n=2 replicates per sample). Transcripts from house-keeping genes are shown in blue. (d) The histograms show the length distributions of *Aire*-regulated and non-*Aire*-regulated transcripts in *Aire*-regulated genes (analysis limited to genes that contained at least one significantly *Aire*-regulated transcript (c)).

Figure 4: Selective expression of peripheral splicing factors and exons in mTEC.

(a) The heatmap shows the expression of a set of tissue-restricted (*tau* > 0.5) genes that encode for functionally studied splicing factors (Supplementary Table 8) in mTEC populations, skin epithelia, cTEC and the ENCODE peripheral tissues. The bar plot (left) shows the fraction of a set of single mature mTEC that express each factor. Significant differences in expression between immature vs mature (*Aire*-KO) mTEC or *Aire*-KO vs *Aire*-positive mTEC are indicated by red and blue asterisks, respectively (BH adjusted p-value < 0.05, |fc| > 2, DESeq2 analysis of population RNA-sequencing data, n=2 biological replicates/condition). (b) The heatmap shows protein-coding exons (>50bp in length) that are included in transcripts from peripheral tissues but not mTEC. Exons were selected if they were included in at least one peripheral tissue (PSI > 0.5) but had a low inclusion rate in the mTEC samples (mean PSI < 0.1; max PSI < 0.2) (c) The heatmap shows micro-exon (≤ 30 bp) inclusion in transcripts from protein-coding genes in the mTEC population and ENCODE peripheral tissue samples.

Figure 5: RBFOX is present with AIRE in mTEC nuclei and promotes mTEC development. (a) Confocal immunofluorescence imaging analysis of the localisation of AIRE (green), RBFOX (anti-RRM domain antibody; red) in the medulla of the thymus. The cells were also stained for the mTEC cell surface marker UEA-1 (yellow) and nuclei are labelled with DAPI (blue). Two representative sections are shown in the upper and lower panels. (b) Representative flow cytometric analysis of cTEC and mTEC frequencies amongst thymic epithelial cells (live, EPCAM+, CD45-) extracted from *Rbfox2* tKO animals. (c) Quantification of cTEC and mTEC frequency in *Rbfox2* tKO animals (n=2 independent experiments with n=6 4-6 weeks old mice per experiment). Both experiments were combined for statistical analysis as indicated. The mean ± SE is shown in bar graphs; * p-value < 0.05 for two-sided Welch Two Sample t-test.

Figure 6: Rbfox factors regulate alternative splicing in thymic epithelial cells. (a) The scatterplot shows changes in gene expression between the *Rbfox1* tKO vs *Rbfox1*$^{fl/fl}$ mature

mTEC (x axis) and *Rbfox2* tKO vs *Rbfox2*$^{fl/fl}$ mature mTEC (y axis). Genes significantly differentially expressed (n=2 biological replicates, DESeq2, BH adjusted $p < 0.05$, $|fc| > 1.5$) in both knockouts are colored in orange, in the *Rbfox1* tKO alone in green and in the *Rbfox2* tKO alone in blue. (b) The numbers of genes containing differential splicing events identified in mature mTEC extracted from *Rbfox1* tKO and *Rbfox2* tKO animals vs cre- littermate controls (n=2 biological replications, rMATS, FDR < 0.05, |delta PSI| > 0.2, protein-coding genes only). (c) Breakdown of the identified splicing events (b) by event type and promiscuous expression status (Supplementary Fig. 1). (d) The barplots show the enrichments (odds ratios) of previously predicted Rbfox target genes [43] in the sets of genes that were found to be differentially spliced in the *Rbfox1* tKO or *Rbfox2* tKO mature mTEC (Two-sided Fisher Exact tests, BH adjusted p-values). * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, SE = skipped exon, RI = retained intron, MXE = mutually exclusive exon, A3SS/A5SS = alternative 3'/5' splice site.

Figure 7: Characterisation of Rbfox2 regulated skipped exon events in mature mTEC

(a) Selected gene ontology (GO) biological processes that were found to be significantly over-represented (one-sided Fisher Exact tests, BH adjusted p-values < 0.05) in protein-coding genes that contained significantly Rbfox2 regulated skipped exon events in mature mTEC. (b) Three examples of significantly Rbfox2 regulated splicing events in mature mTEC. *Fn1* and *Insr* are known Rbfox target genes [29,43]. *Myom2* is an example of a non-*Aire* TRA gene. In all cases, significantly Rbfox2 regulated skipped exon events were identified as those with |delta PSI| > 0.2 and FDR < 0.05 in mature mTEC from *Rbfox2* tKO vs cre- littermate controls (Fig. 6b). (c) Enrichment of the Rbfox recognition motif (M159/M017) [44] in the sequences surrounding exons that were found to be significantly regulated by Rbfox2 in mature mTEC (Fig. 6b). The lines show enrichments for the sets of exons that were found to be enhanced (blue) or repressed (red) or not significantly regulated (green) by Rbfox2. Thicker lines indicate regions of statistically significant enrichment (FDR ≤ 0.05, n=1,000 permutations).

References

1 Klein, L., Kyewski, B., Allen, P. M. & Hogquist, K. A. Positive and negative selection of the T cell repertoire: what thymocytes see (and don't see). *Nat Rev Immunol* **14**, 377-391, doi:10.1038/nri3667 (2014).

2 Abramson, J. & Anderson, G. Thymic Epithelial Cells. *Annu Rev Immunol* **35**, 85-118, doi:10.1146/annurev-immunol-051116-052320 (2017).

3 Stritesky, G. L. *et al.* Murine thymic selection quantified using a unique method to capture deleted T cells. *Proc Natl Acad Sci U S A* **110**, 4679-4684, doi:10.1073/pnas.1217532110 (2013).

4 Stritesky, G. L., Jameson, S. C. & Hogquist, K. A. Selection of self-reactive T cells in the thymus. *Annu Rev Immunol* **30**, 95-114, doi:10.1146/annurev-immunol-020711-075035 (2012).

5 Sansom, S. N. *et al.* Population and single-cell genomics reveal the Aire dependency, relief from Polycomb silencing, and distribution of self-antigen expression in thymic epithelia. *Genome Res* **24**, 1918-1931, doi:10.1101/gr.171645.113 (2014).

6 Abramson, J. & Goldfarb, Y. AIRE: From promiscuous molecular partnerships to promiscuous gene expression. *Eur J Immunol* **46**, 22-33, doi:10.1002/eji.201545792 (2016).

7 Takaba, H. *et al.* Fezf2 Orchestrates a Thymic Program of Self-Antigen Expression for Immune Tolerance. *Cell* **163**, 975-987, doi:10.1016/j.cell.2015.10.013 (2015).

8 Matera, A. G. & Wang, Z. A day in the life of the spliceosome. *Nat Rev Mol Cell Biol* **15**, 108-121, doi:10.1038/nrm3742 (2014).

9 Starck, S. R. & Shastri, N. Nowhere to hide: unconventional translation yields cryptic peptides for immune surveillance. *Immunol Rev* **272**, 8-16, doi:10.1111/imr.12434 (2016).

10 Danan-Gotthold, M., Guyon, C., Giraud, M., Levanon, E. Y. & Abramson, J. Extensive RNA editing and splicing increase immune self-representation diversity in medullary thymic epithelial cells. *Genome Biol* **17**, 219, doi:10.1186/s13059-016-1079-9 (2016).

11 Granados, D. P., Laumont, C. M., Thibault, P. & Perreault, C. The nature of self for T cells-a systems-level perspective. *Curr Opin Immunol* **34**, 1-8, doi:10.1016/j.coi.2014.10.012 (2015).

12 Raposo, B. *et al.* T cells specific for post-translational modifications escape intrathymic tolerance induction. *Nat Commun* **9**, 353, doi:10.1038/s41467-017-02763-y (2018).

13 Zerbino, D. R. *et al.* Ensembl 2018. *Nucleic Acids Res* **46**, D754-D761, doi:10.1093/nar/gkx1098 (2018).

14 Klein, L., Klugmann, M., Nave, K. A., Tuohy, V. K. & Kyewski, B. Shaping of the autoreactive T-cell repertoire by a splice variant of self protein expressed in thymic epithelial cells. *Nat Med* **6**, 56-61, doi:10.1038/71540 (2000).

15 Keane, P., Ceredig, R. & Seoighe, C. Promiscuous mRNA splicing under the control of AIRE in medullary thymic epithelial cells. *Bioinformatics* **31**, 986-990, doi:10.1093/bioinformatics/btu785 (2015).

16 Abramson, J., Giraud, M., Benoist, C. & Mathis, D. Aire's partners in the molecular control of immunological tolerance. *Cell* **140**, 123-135, doi:10.1016/j.cell.2009.12.030 (2010).

17 St-Pierre, C., Trofimov, A., Brochu, S., Lemieux, S. & Perreault, C. Differential Features of AIRE-Induced and AIRE-Independent Promiscuous Gene Expression in Thymic Epithelial Cells. *J Immunol* **195**, 498-506, doi:10.4049/jimmunol.1500558 (2015).
18 Pervouchine, D. D. *et al.* Enhanced transcriptome maps from multiple mouse tissues reveal evolutionary constraint in gene expression. *Nat Commun* **6**, 5903, doi:10.1038/ncomms6903 (2015).
19 Kelemen, O. *et al.* Function of alternative splicing. *Gene* **514**, 1-30, doi:10.1016/j.gene.2012.07.083 (2013).
20 Zuklys, S. *et al.* Foxn1 regulates key target genes essential for T cell development in postnatal thymic epithelial cells. *Nat Immunol* **17**, 1206-1215, doi:10.1038/ni.3537 (2016).
21 Handel, A. E. *et al.* Comprehensively Profiling the Chromatin Architecture of Tissue Restricted Antigen Expression in Thymic Epithelial Cells Over Development. *Front Immunol* **9**, 2120, doi:10.3389/fimmu.2018.02120 (2018).
22 Kajikawa, M. *et al.* MHC class I-like MILL molecules are beta2-microglobulin-associated, GPI-anchored glycoproteins that do not require TAP for cell surface expression. *J Immunol* **177**, 3108-3115, doi:10.4049/jimmunol.177.5.3108 (2006).
23 Yang, Y. *et al.* Characterization of B7S3 as a novel negative regulator of T cells. *J Immunol* **178**, 3661-3667 (2007).
24 Jacob, A. G. & Smith, C. W. J. Intron retention as a component of regulated gene expression programs. *Hum Genet* **136**, 1043-1057, doi:10.1007/s00439-017-1791-x (2017).
25 Meredith, M., Zemmour, D., Mathis, D. & Benoist, C. Aire controls gene expression in the thymic epithelium with ordered stochasticity. *Nat Immunol* **16**, 942-949, doi:10.1038/ni.3247 (2015).
26 Giraud, M. *et al.* Aire unleashes stalled RNA polymerase to induce ectopic gene expression in thymic epithelial cells. *Proc Natl Acad Sci U S A* **109**, 535-540, doi:10.1073/pnas.1119351109 (2012).
27 Cook, K. B., Kazan, H., Zuberi, K., Morris, Q. & Hughes, T. R. RBPDB: a database of RNA-binding specificities. *Nucleic Acids Res* **39**, D301-308, doi:10.1093/nar/gkq1069 (2011).
28 Barbosa-Morais, N. L., Carmo-Fonseca, M. & Aparicio, S. Systematic genome-wide annotation of spliceosomal proteins reveals differential gene family expansion. *Genome Res* **16**, 66-77, doi:10.1101/gr.3936206 (2006).
29 Chen, M. & Manley, J. L. Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches. *Nat Rev Mol Cell Biol* **10**, 741-754, doi:10.1038/nrm2777 (2009).
30 Grosso, A. R. *et al.* Tissue-specific splicing factor gene expression signatures. *Nucleic Acids Res* **36**, 4823-4832, doi:10.1093/nar/gkn463 (2008).
31 Han, H. *et al.* MBNL proteins repress ES-cell-specific alternative splicing and reprogramming. *Nature* **498**, 241-245, doi:10.1038/nature12270 (2013).
32 Jangi, M. & Sharp, P. A. Building robust transcriptomes with master splicing factors. *Cell* **159**, 487-498, doi:10.1016/j.cell.2014.09.054 (2014).
33 Merkin, J., Russell, C., Chen, P. & Burge, C. B. Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. *Science* **338**, 1593-1599, doi:10.1126/science.1228186 (2012).

34 Conboy, J. G. Developmental regulation of RNA processing by Rbfox proteins. *Wiley Interdiscip Rev RNA* **8**, doi:10.1002/wrna.1398 (2017).
35 Guo, W. *et al.* RBM20, a gene for hereditary cardiomyopathy, regulates titin splicing. *Nat Med* **18**, 766-773, doi:10.1038/nm.2693 (2012).
36 Murphy, D., Cieply, B., Carstens, R., Ramamurthy, V. & Stoilov, P. The Musashi 1 Controls the Splicing of Photoreceptor-Specific Exons in the Vertebrate Retina. *PLoS Genet* **12**, e1006256, doi:10.1371/journal.pgen.1006256 (2016).
37 Quesnel-Vallieres, M. *et al.* Misregulation of an Activity-Dependent Splicing Network as a Common Mechanism Underlying Autism Spectrum Disorders. *Mol Cell* **64**, 1023-1034, doi:10.1016/j.molcel.2016.11.033 (2016).
38 Quesnel-Vallieres, M., Irimia, M., Cordes, S. P. & Blencowe, B. J. Essential roles for the splicing regulator nSR100/SRRM4 during nervous system development. *Genes Dev* **29**, 746-759, doi:10.1101/gad.256115.114 (2015).
39 Damianov, A. & Black, D. L. Autoregulation of Fox protein expression to produce dominant negative splicing factors. *RNA* **16**, 405-416, doi:10.1261/rna.1838210 (2010).
40 Gehman, L. T. *et al.* The splicing regulator Rbfox2 is required for both cerebellar development and mature motor function. *Genes Dev* **26**, 445-460, doi:10.1101/gad.182477.111 (2012).
41 Gehman, L. T. *et al.* The splicing regulator Rbfox1 (A2BP1) controls neuronal excitation in the mammalian brain. *Nat Genet* **43**, 706-711, doi:10.1038/ng.841 (2011).
42 Gordon, J. *et al.* Specific expression of lacZ and cre recombinase in fetal thymic epithelial cells by multiplex gene targeting at the Foxn1 locus. *BMC Dev Biol* **7**, 69, doi:10.1186/1471-213X-7-69 (2007).
43 Weyn-Vanhentenryck, S. M. *et al.* HITS-CLIP and integrative modeling define the Rbfox splicing-regulatory network linked to brain development and autism. *Cell Rep* **6**, 1139-1152, doi:10.1016/j.celrep.2014.02.005 (2014).
44 Ray, D. *et al.* A compendium of RNA-binding motifs for decoding gene regulation. *Nature* **499**, 172-177, doi:10.1038/nature12311 (2013).
45 Shilov, E. S., Gorshkova, E. A., Minnegalieva, A. R. & Potashnikova, D. M. [Splicing Pattern of mRNA in Thymus Epithelial Cells Limitsthe Transcriptome Available for Negative Selection of Autoreactive T Cells]. *Mol Biol (Mosk)* **53**, 109-119, doi:10.1134/S0026898419010154 (2019).
46 Irimia, M. *et al.* A highly conserved program of neuronal microexons is misregulated in autistic brains. *Cell* **159**, 1511-1523, doi:10.1016/j.cell.2014.11.035 (2014).
47 Hughes, H. K., Mills Ko, E., Rose, D. & Ashwood, P. Immune Dysfunction and Autoimmunity as Pathological Mechanisms in Autism Spectrum Disorders. *Front Cell Neurosci* **12**, 405, doi:10.3389/fncel.2018.00405 (2018).
48 Decock, S., McGee, P. & Hirschfield, G. M. Autoimmune liver disease for the non-specialist. *BMJ* **339**, b3305, doi:10.1136/bmj.b3305 (2009).
49 Webster, N. J. G. Alternative RNA Splicing in the Pathogenesis of Liver Disease. *Front Endocrinol (Lausanne)* **8**, 133, doi:10.3389/fendo.2017.00133 (2017).
50 Guyon, C. *et al.* Aire-dependent genes undergo Clp1-mediated 3'UTR shortening associated with higher transcript stability in the thymus. *bioRxiv* (2019).

51 Evsyukova, I., Somarelli, J. A., Gregory, S. G. & Garcia-Blanco, M. A. Alternative splicing in multiple sclerosis and other autoimmune diseases. *RNA Biol* **7**, 462-473, doi:10.4161/rna.7.4.12301 (2010).
52 Juan-Mateu, J., Villate, O. & Eizirik, D. L. MECHANISMS IN ENDOCRINOLOGY: Alternative splicing: the new frontier in diabetes research. *Eur J Endocrinol* **174**, R225-238, doi:10.1530/EJE-15-0916 (2016).
53 Newman, J. R. B. *et al.* Disease-specific biases in alternative splicing and tissue-specific dysregulation revealed by multitissue profiling of lymphocyte gene expression in type 1 diabetes. *Genome Res* **27**, 1807-1815, doi:10.1101/gr.217984.116 (2017).
54 Ng, B. *et al.* Increased noncanonical splicing of autoantigen transcripts provides the structural basis for expression of untolerized epitopes. *J Allergy Clin Immunol* **114**, 1463-1470, doi:10.1016/j.jaci.2004.09.006 (2004).
55 Mosanya, C. H. & Isaacs, J. D. Tolerising cellular therapies: what is their promise for autoimmune disease? *Ann Rheum Dis* **78**, 297-310, doi:10.1136/annrheumdis-2018-214024 (2019).
56 Pozsgay, J., Szekanecz, Z. & Sarmay, G. Antigen-specific immunotherapies in rheumatic diseases. *Nat Rev Rheumatol* **13**, 525-537, doi:10.1038/nrrheum.2017.107 (2017).
57 Ohigashi, I. *et al.* Adult Thymic Medullary Epithelium Is Maintained and Regenerated by Lineage-Restricted Cells Rather Than Bipotent Progenitors. *Cell Rep* **13**, 1432-1443, doi:10.1016/j.celrep.2015.10.012 (2015).
58 Dhalla, F. *et al.* Biologically indeterminate yet ordered promiscuous gene expression in single medullary thymic epithelial cells. *EMBO J* **39**, e101828, doi:10.15252/embj.2019101828 (2020).
59 St-Pierre, C. *et al.* Transcriptome sequencing of neonatal thymic epithelial cells. *Sci Rep* **3**, 1860, doi:10.1038/srep01860 (2013).
60 Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* **37**, 907-915, doi:10.1038/s41587-019-0201-4 (2019).
61 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).
62 Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* **33**, 290-295, doi:10.1038/nbt.3122 (2015).
63 Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods* **14**, 417-419, doi:10.1038/nmeth.4197 (2017).
64 Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21, doi:10.1093/bioinformatics/bts635 (2013).
65 Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923-930, doi:10.1093/bioinformatics/btt656 (2014).
66 Kryuchkova-Mostacci, N. & Robinson-Rechavi, M. A benchmark of gene expression tissue-specificity metrics. *Brief Bioinform* **18**, 205-214, doi:10.1093/bib/bbw008 (2017).

67 Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550, doi:10.1186/s13059-014-0550-8 (2014).
68 Alamancos, G. P., Pages, A., Trincado, J. L., Bellora, N. & Eyras, E. Leveraging transcript quantification for fast computation of alternative splicing profiles. *RNA* **21**, 1521-1531, doi:10.1261/rna.051557.115 (2015).
69 Shen, S. *et al.* rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proc Natl Acad Sci U S A* **111**, E5593-5601, doi:10.1073/pnas.1419161111 (2014).
70 Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* **34**, 525-527, doi:10.1038/nbt.3519 (2016).
71 Pimentel, H., Bray, N. L., Puente, S., Melsted, P. & Pachter, L. Differential analysis of RNA-seq incorporating quantification uncertainty. *Nat Methods* **14**, 687-690, doi:10.1038/nmeth.4324 (2017).
72 Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094-3100, doi:10.1093/bioinformatics/bty191 (2018).
73 Gohr, A. & Irimia, M. Matt: Unix tools for alternative splicing analysis. *Bioinformatics* **35**, 130-132, doi:10.1093/bioinformatics/bty606 (2019).
74 Katz, Y., Wang, E. T., Airoldi, E. M. & Burge, C. B. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat Methods* **7**, 1009-1015, doi:10.1038/nmeth.1528 (2010).

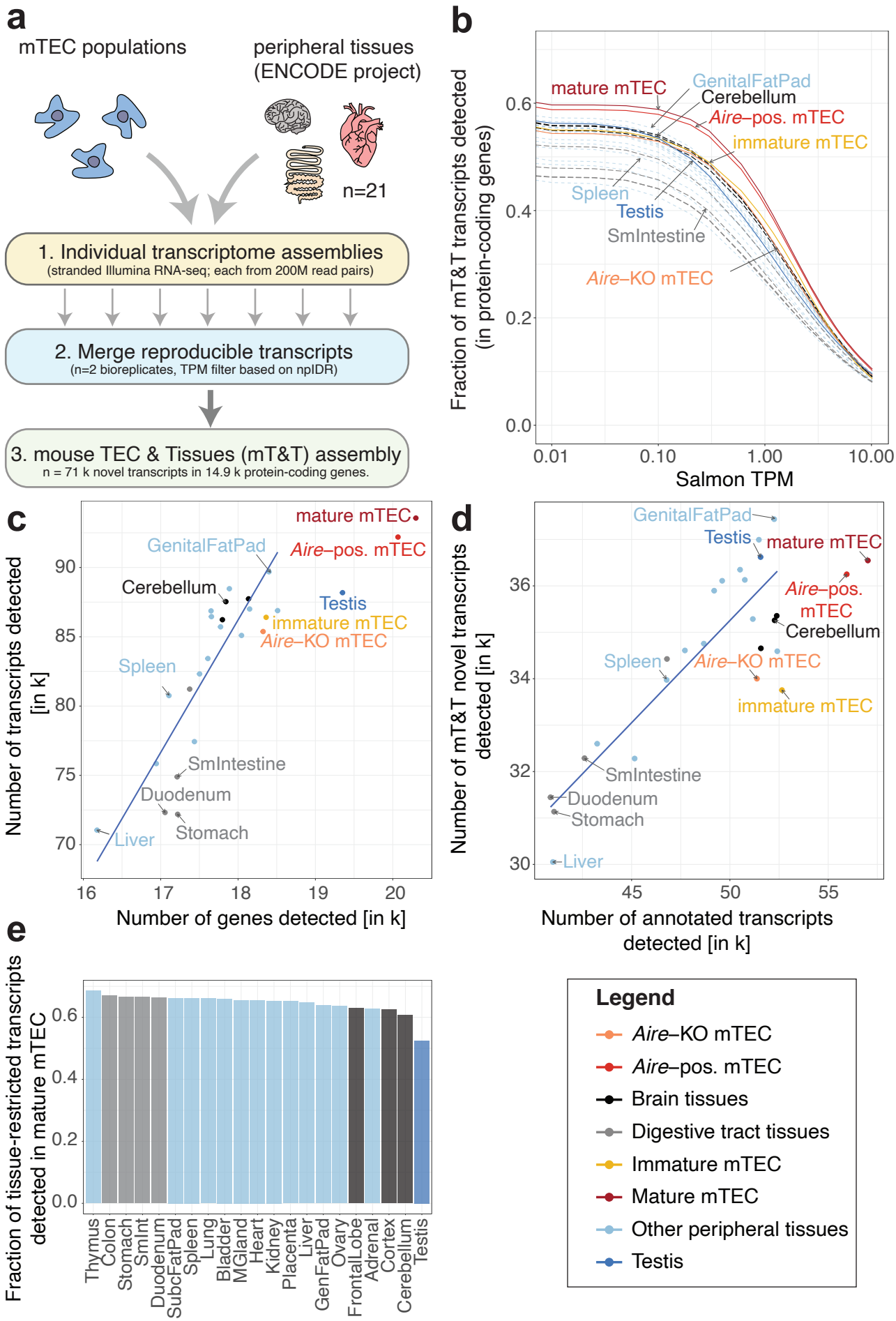

a
mTEC populations
peripheral tissues
(ENCODE project)
n=21
1. Individual transcriptome assemblies
(stranded Illumina RNA-seq; each from 200M read pairs)
2. Merge reproducible transcripts
(n=2 bioreplicates, TPM filter based on npIDR)
3. mouse TEC & Tissues (mT&T) assembly
n = 71 k novel transcripts in 14.9 k protein-coding genes.
b
Fraction of mT&T transcripts detected
(in protein-coding genes)
Salmon TPM
mature mTEC
GenitalFatPad
Cerebellum
Aire–pos. mTEC
immature mTEC
Spleen
Testis
SmIntestine
Aire–KO mTEC
c
Number of transcripts detected
[in k]
Number of genes detected [in k]
mature mTEC
GenitalFatPad
Aire–pos. mTEC
Cerebellum
Testis
immature mTEC
Aire–KO mTEC
Spleen
SmIntestine
Duodenum
Stomach
Liver
d
Number of mT&T novel transcripts
detected [in k]
Number of annotated transcripts
detected [in k]
GenitalFatPad
Testis
mature mTEC
Aire–pos.
mTEC
Cerebellum
Spleen
Aire–KO mTEC
immature mTEC
SmIntestine
Duodenum
Stomach
Liver
e
Fraction of tissue-restricted transcripts
detected in mature mTEC
Thymus
Colon
Stomach
SmInt
Duodenum
SubcFatPad
Spleen
Lung
Bladder
MGland
Heart
Kidney
Placenta
Liver
GenFatPad
Ovary
FrontalLobe
Adrenal
Cortex
Cerebellum
Testis
Legend
Aire–KO mTEC
Aire–pos. mTEC
Brain tissues
Digestive tract tissues
Immature mTEC
Mature mTEC
Other peripheral tissues
Testis

Figure 1

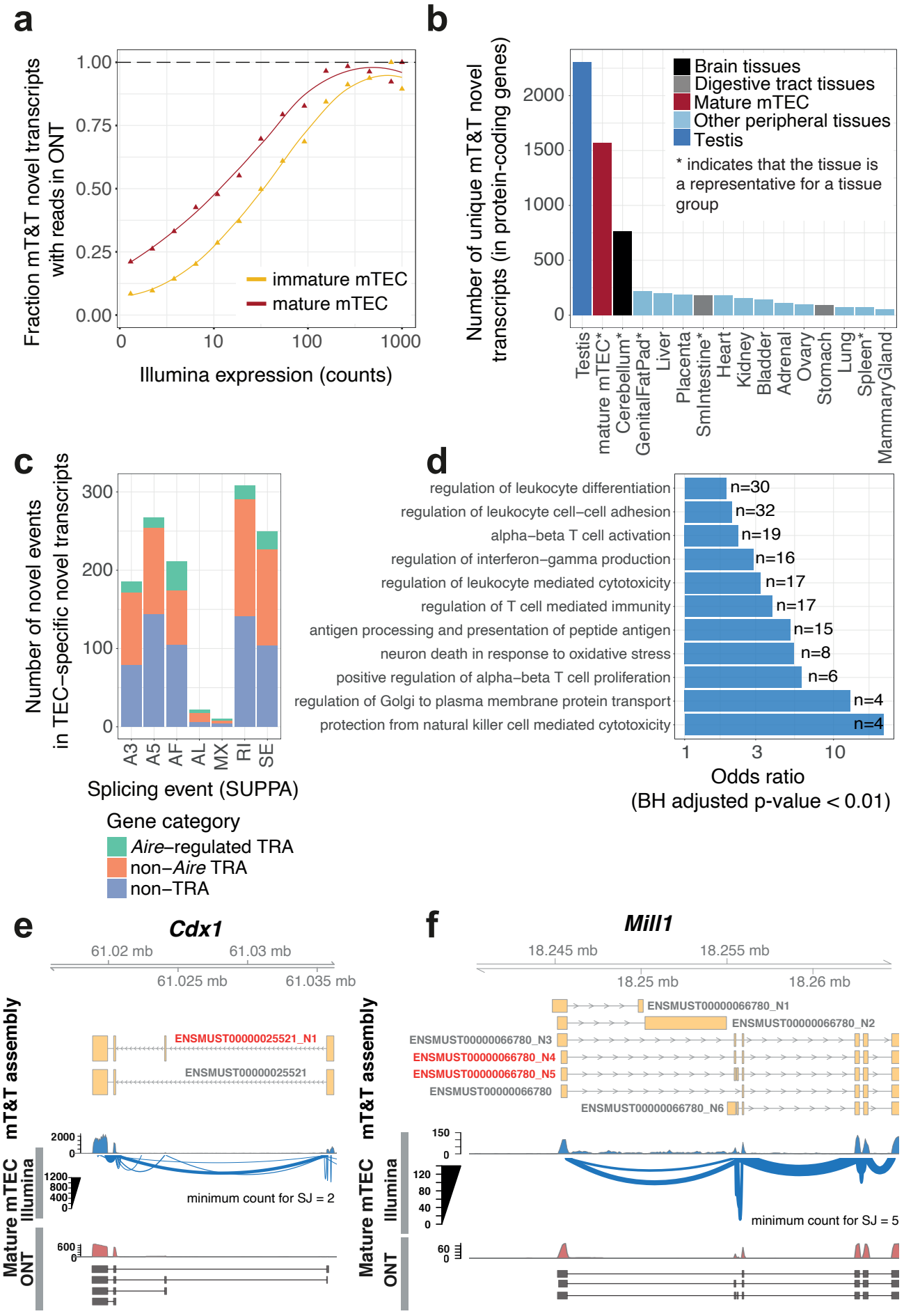

a
Fraction mT&T novel transcripts with reads in ONT
1.00
0.75
0.50
0.25
0.00
0
10
100
1000
Illumina expression (counts)
immature mTEC
mature mTEC
b
Number of unique mT&T novel transcripts (in protein-coding genes)
2000
1500
1000
500
0
Brain tissues
Digestive tract tissues
Mature mTEC
Other peripheral tissues
Testis
* indicates that the tissue is a representative for a tissue group
Testis
mature mTEC*
Cerebellum*
GenitalFatPad*
Liver
Placenta
SmIntestine*
Heart
Kidney
Bladder
Adrenal
Ovary
Stomach
Lung
Spleen*
MammaryGland
c
Number of novel events in TEC−specific novel transcripts
300
200
100
0
A3
A5
AF
AL
MX
RI
SE
Splicing event (SUPPA)
Gene category
Aire−regulated TRA
non−Aire TRA
non−TRA
d
regulation of leukocyte differentiation
regulation of leukocyte cell−cell adhesion
alpha−beta T cell activation
regulation of interferon−gamma production
regulation of leukocyte mediated cytotoxicity
regulation of T cell mediated immunity
antigen processing and presentation of peptide antigen
neuron death in response to oxidative stress
positive regulation of alpha−beta T cell proliferation
regulation of Golgi to plasma membrane protein transport
protection from natural killer cell mediated cytotoxicity
n=30
n=32
n=19
n=16
n=17
n=17
n=15
n=8
n=6
n=4
n=4
1
3
10
Odds ratio
(BH adjusted p-value < 0.01)
e
Cdx1
61.02 mb
61.03 mb
61.025 mb
61.035 mb
mT&T assembly
ENSMUST00000025521_N1
ENSMUST00000025521
Mature mTEC Illumina
2000
0
1200
800
400
0
minimum count for SJ = 2
Mature mTEC ONT
600
0
f
Mill1
18.245 mb
18.255 mb
18.25 mb
18.26 mb
mT&T assembly
ENSMUST00000066780_N1
ENSMUST00000066780_N2
ENSMUST00000066780_N3
ENSMUST00000066780_N4
ENSMUST00000066780_N5
ENSMUST00000066780
ENSMUST00000066780_N6
Mature mTEC Illumina
150
0
120
80
40
0
minimum count for SJ = 5
Mature mTEC ONT
60
0

Figure 2

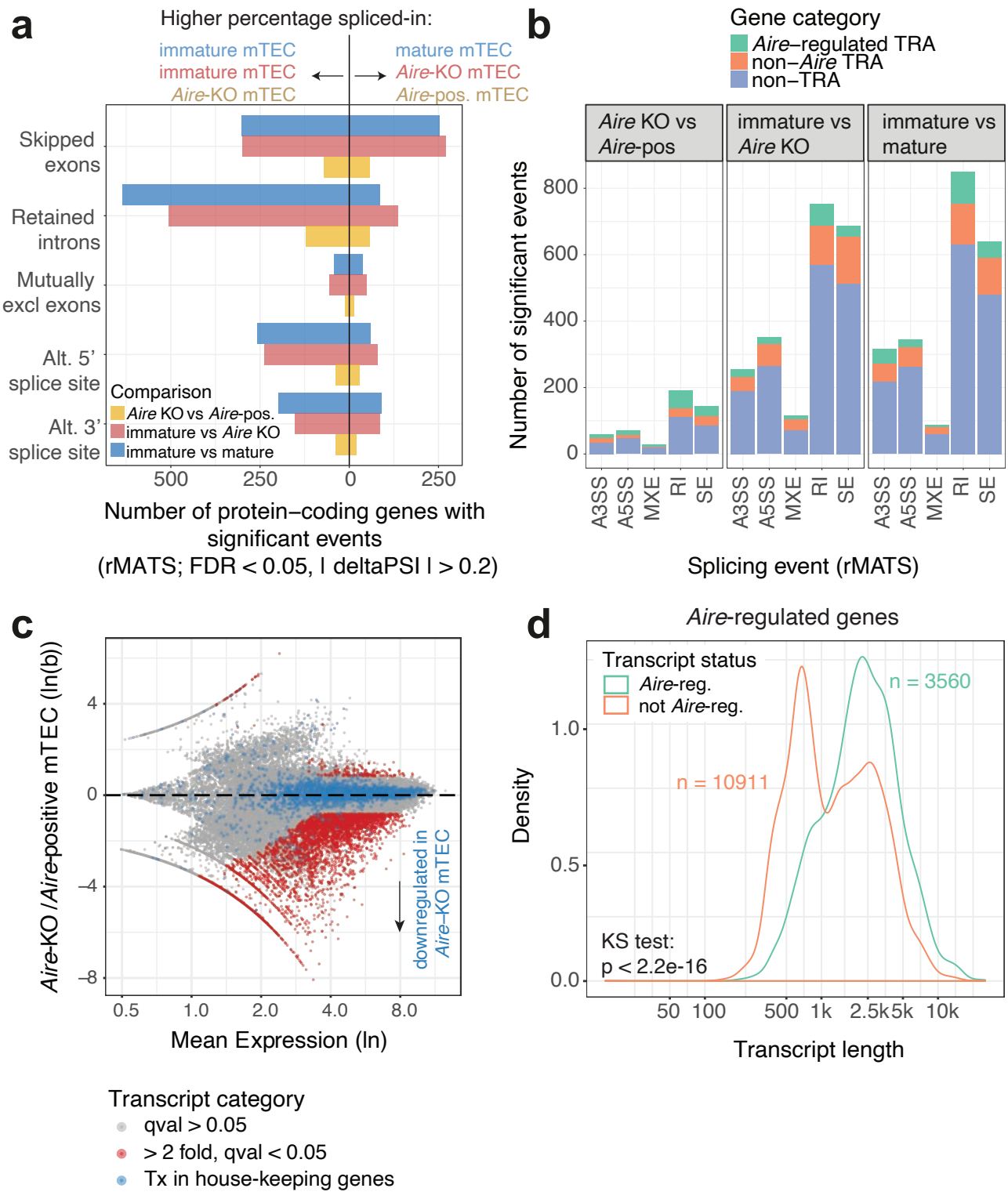

a
Higher percentage spliced-in:
immature mTEC
immature mTEC
Aire-KO mTEC
mature mTEC
Aire-KO mTEC
Aire-pos. mTEC
Skipped exons
Retained introns
Mutually excl exons
Alt. 5' splice site
Alt. 3' splice site
Comparison
Aire KO vs Aire-pos.
immature vs Aire KO
immature vs mature
500
250
0
250
Number of protein–coding genes with significant events
(rMATS; FDR < 0.05, | deltaPSI | > 0.2)
b
Gene category
Aire–regulated TRA
non–Aire TRA
non–TRA
Aire KO vs Aire-pos
immature vs Aire KO
immature vs mature
Number of significant events
800
600
400
200
0
A3SS
A5SS
MXE
RI
SE
Splicing event (rMATS)
c
Aire-KO / Aire-positive mTEC (ln(b))
4
0
–4
–8
0.5
1.0
2.0
4.0
8.0
Mean Expression (ln)
downregulated in Aire–KO mTEC
Transcript category
qval > 0.05
> 2 fold, qval < 0.05
Tx in house-keeping genes
d
Aire-regulated genes
Transcript status
Aire-reg.
not Aire-reg.
n = 3560
n = 10911
Density
1.0
0.5
0.0
KS test:
p < 2.2e-16
50
100
500
1k
2.5k
5k
10k
Transcript length

Figure 3

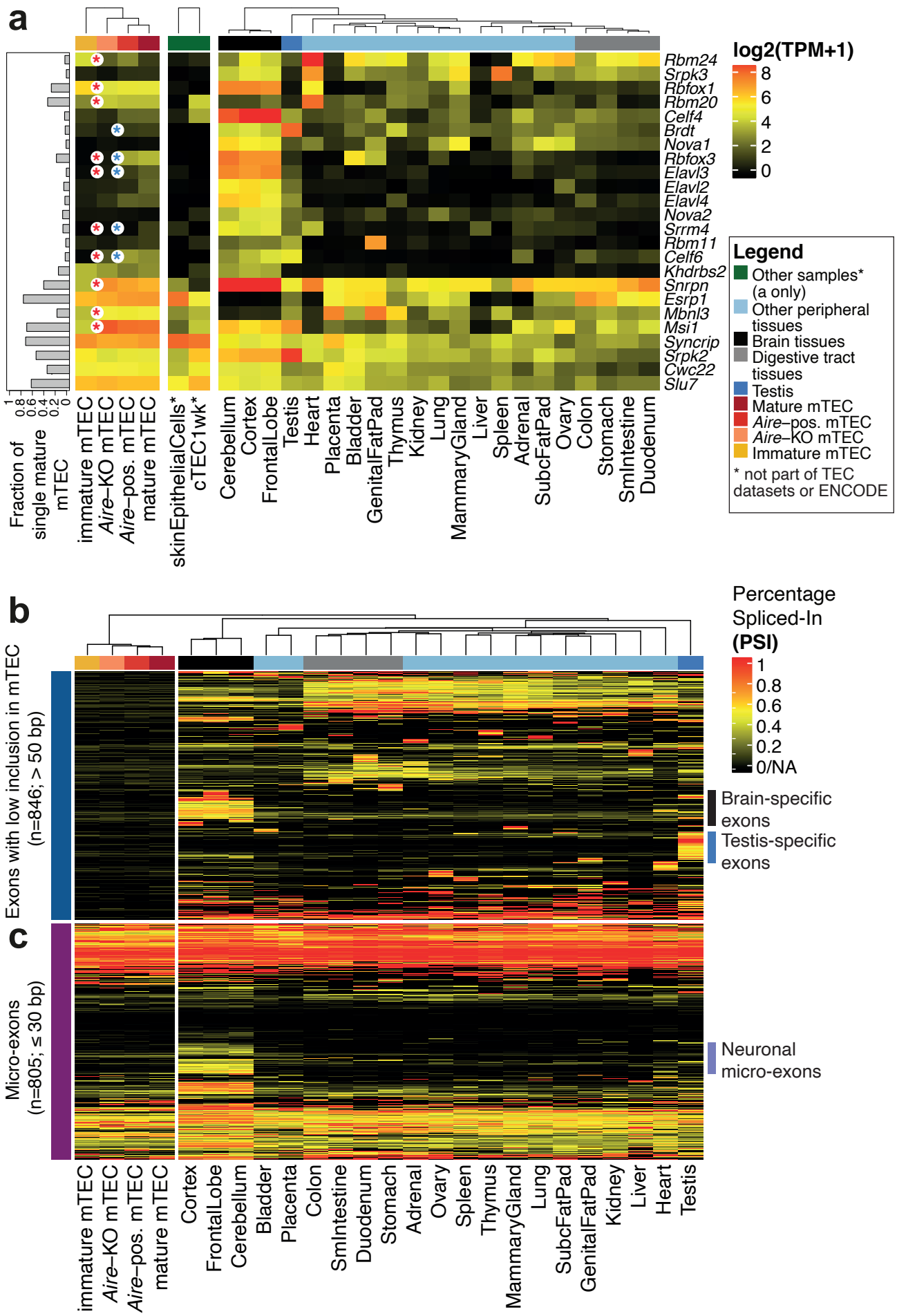

a
Fraction of single mature mTEC
immature mTEC
Aire–KO mTEC
Aire–pos. mTEC
mature mTEC
skinEpithelialCells*
cTEC1wk*
Cerebellum
Cortex
FrontalLobe
Testis
Heart
Placenta
Bladder
GenitalFatPad
Thymus
Kidney
Lung
MammaryGland
Liver
Spleen
Adrenal
SubcFatPad
Ovary
Colon
Stomach
SmIntestine
Duodenum
Rbm24
Srpk3
Rbfox1
Rbm20
Celf4
Brdt
Nova1
Rbfox3
Elavl3
Elavl2
Elavl4
Nova2
Srrm4
Rbm11
Celf6
Khdrbs2
Snrpn
Esrp1
Mbnl3
Msi1
Syncrip
Srpk2
Cwc22
Slu7
log2(TPM+1)
8
6
4
2
0
Legend
Other samples* (a only)
Other peripheral tissues
Brain tissues
Digestive tract tissues
Testis
Mature mTEC
Aire–pos. mTEC
Aire–KO mTEC
Immature mTEC
* not part of TEC datasets or ENCODE
b
Exons with low inclusion in mTEC (n=846; > 50 bp)
Percentage Spliced-In (PSI)
1
0.8
0.6
0.4
0.2
0/NA
Brain-specific exons
Testis-specific exons
c
Micro-exons (n=805; ≤ 30 bp)
Neuronal micro-exons
immature mTEC
Aire–KO mTEC
Aire–pos. mTEC
mature mTEC
Cortex
FrontalLobe
Cerebellum
Bladder
Placenta
Colon
SmIntestine
Duodenum
Stomach
Adrenal
Ovary
Spleen
Thymus
MammaryGland
Lung
SubcFatPad
GenitalFatPad
Kidney
Liver
Heart
Testis

Figure 4

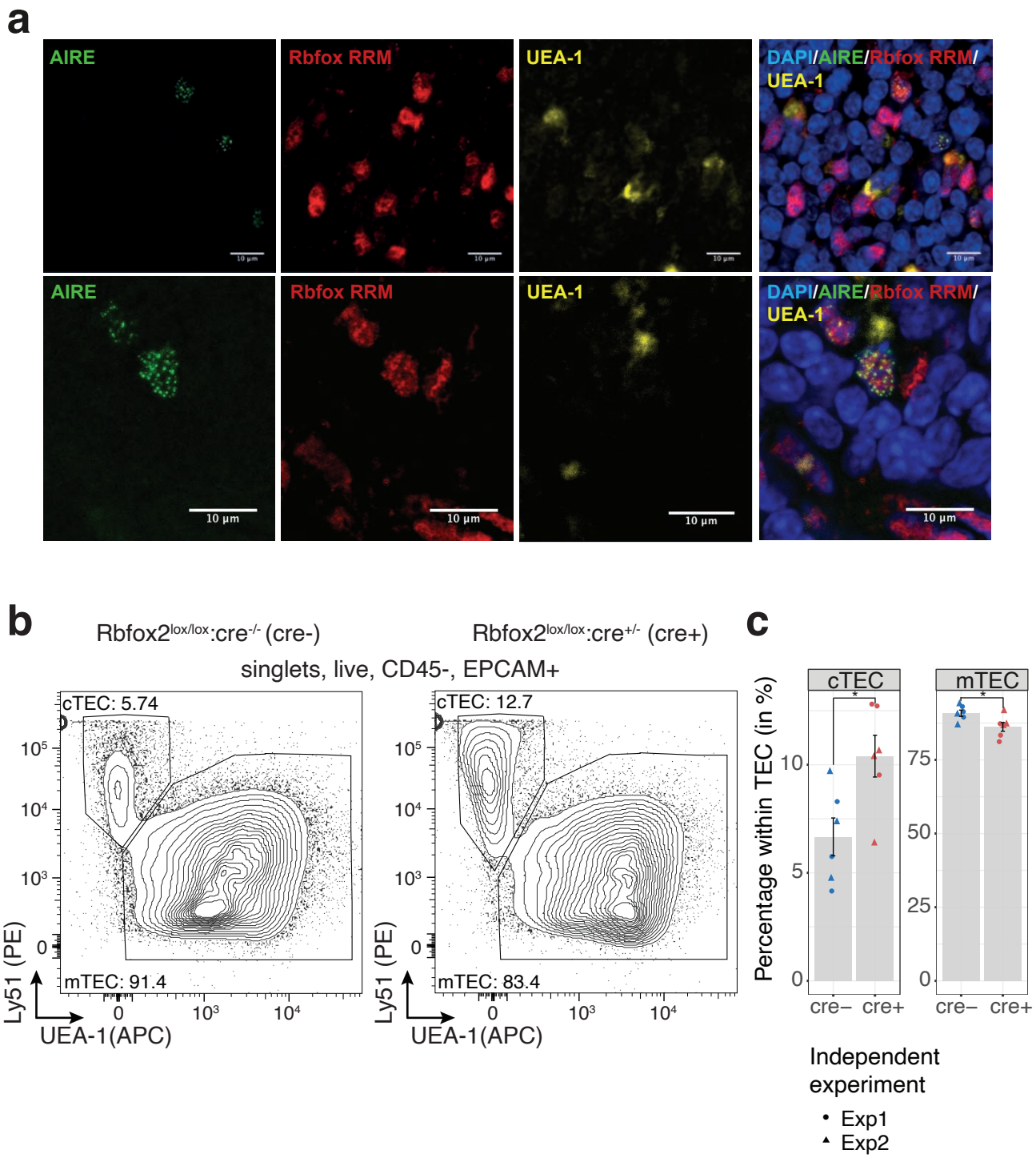

a
AIRE
Rbfox RRM
UEA-1
DAPI/AIRE/Rbfox RRM/
UEA-1
10 µm
b
Rbfox2lox/lox:cre-/- (cre-)
Rbfox2lox/lox:cre+/- (cre+)
singlets, live, CD45-, EPCAM+
cTEC: 5.74
mTEC: 91.4
cTEC: 12.7
mTEC: 83.4
Ly51 (PE)
UEA-1(APC)
c
cTEC
mTEC
Percentage within TEC (in %)
cre−
cre+
Independent experiment
Exp1
Exp2

Figure 5

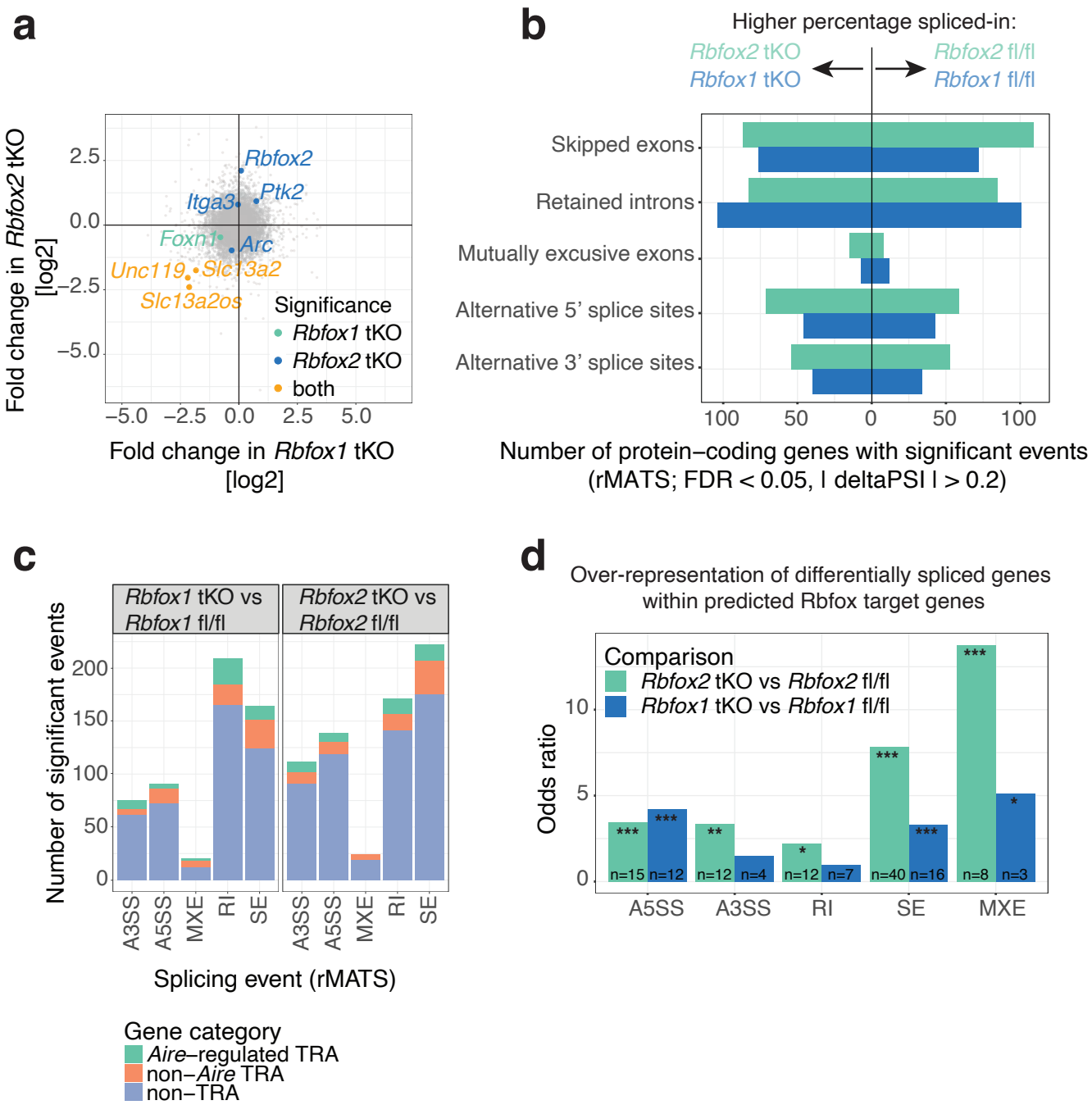
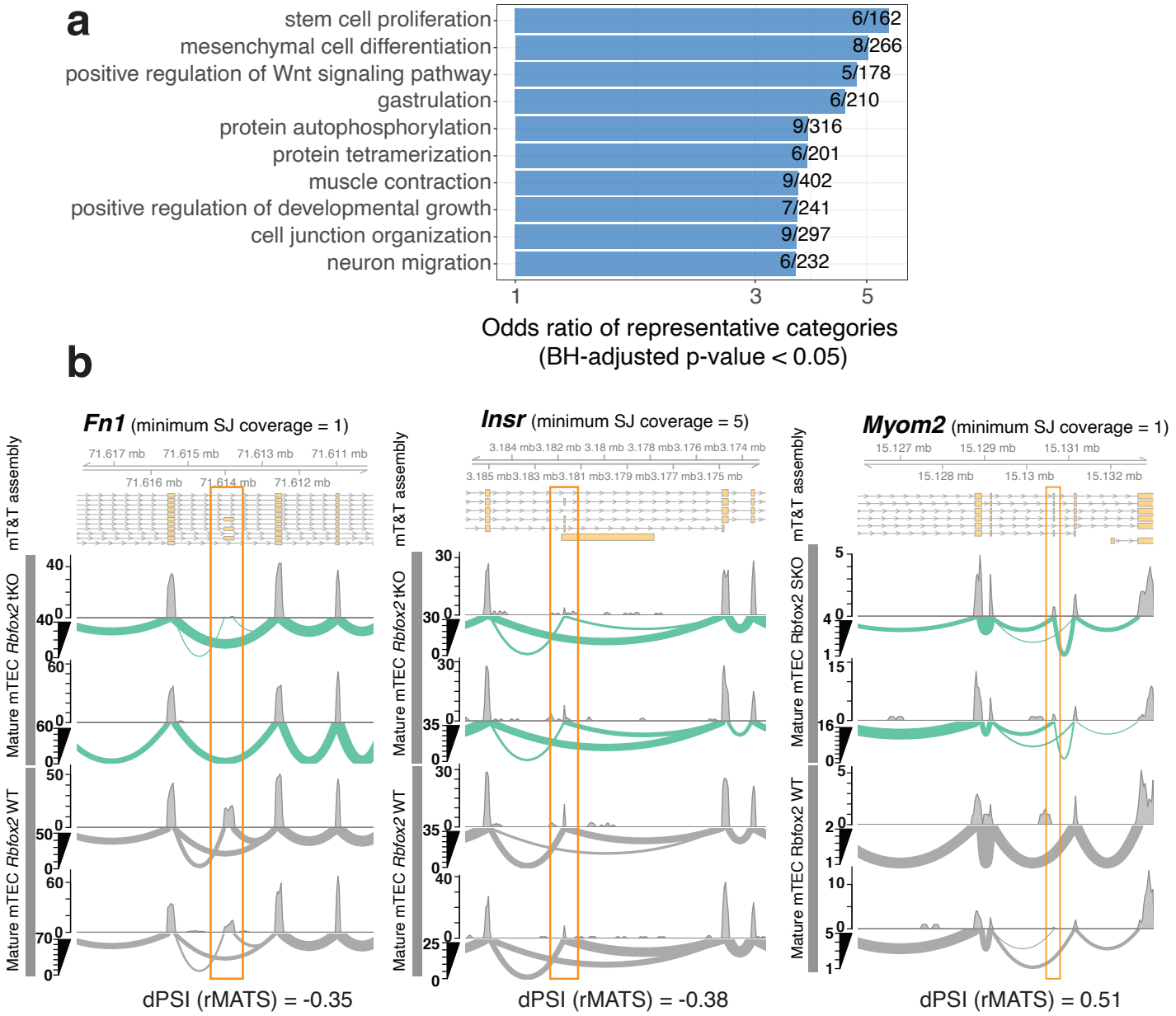

a
Fold change in Rbfox2 tKO [log2]
Fold change in Rbfox1 tKO [log2]
Rbfox2
Ptk2
Itga3
Foxn1
Arc
Unc119
Slc13a2
Slc13a2os
Significance
Rbfox1 tKO
Rbfox2 tKO
both
b
Higher percentage spliced-in:
Rbfox2 tKO
Rbfox1 tKO
Rbfox2 fl/fl
Rbfox1 fl/fl
Skipped exons
Retained introns
Mutually excusive exons
Alternative 5' splice sites
Alternative 3' splice sites
Number of protein–coding genes with significant events
(rMATS; FDR < 0.05, | deltaPSI | > 0.2)
c
Rbfox1 tKO vs Rbfox1 fl/fl
Rbfox2 tKO vs Rbfox2 fl/fl
Number of significant events
A3SS
A5SS
MXE
RI
SE
Splicing event (rMATS)
Gene category
Aire–regulated TRA
non–Aire TRA
non–TRA
d
Over-representation of differentially spliced genes
within predicted Rbfox target genes
Comparison
Rbfox2 tKO vs Rbfox2 fl/fl
Rbfox1 tKO vs Rbfox1 fl/fl
Odds ratio
n=15 n=12 n=12 n=4 n=12 n=7 n=40 n=16 n=8 n=3
A5SS
A3SS
RI
SE
MXE

Figure 6


a
stem cell proliferation
mesenchymal cell differentiation
positive regulation of Wnt signaling pathway
gastrulation
protein autophosphorylation
protein tetramerization
muscle contraction
positive regulation of developmental growth
cell junction organization
neuron migration
6/162
8/266
5/178
6/210
9/316
6/201
9/402
7/241
9/297
6/232
1
3
5
Odds ratio of representative categories
(BH-adjusted p-value < 0.05)
b
Fn1 (minimum SJ coverage = 1)
Insr (minimum SJ coverage = 5)
Myom2 (minimum SJ coverage = 1)
mT&T assembly
Mature mTEC Rbfox2 tKO
Mature mTEC Rbfox2 WT
Mature mTEC Rbfox2 SKO
dPSI (rMATS) = -0.35
dPSI (rMATS) = -0.38
dPSI (rMATS) = 0.51

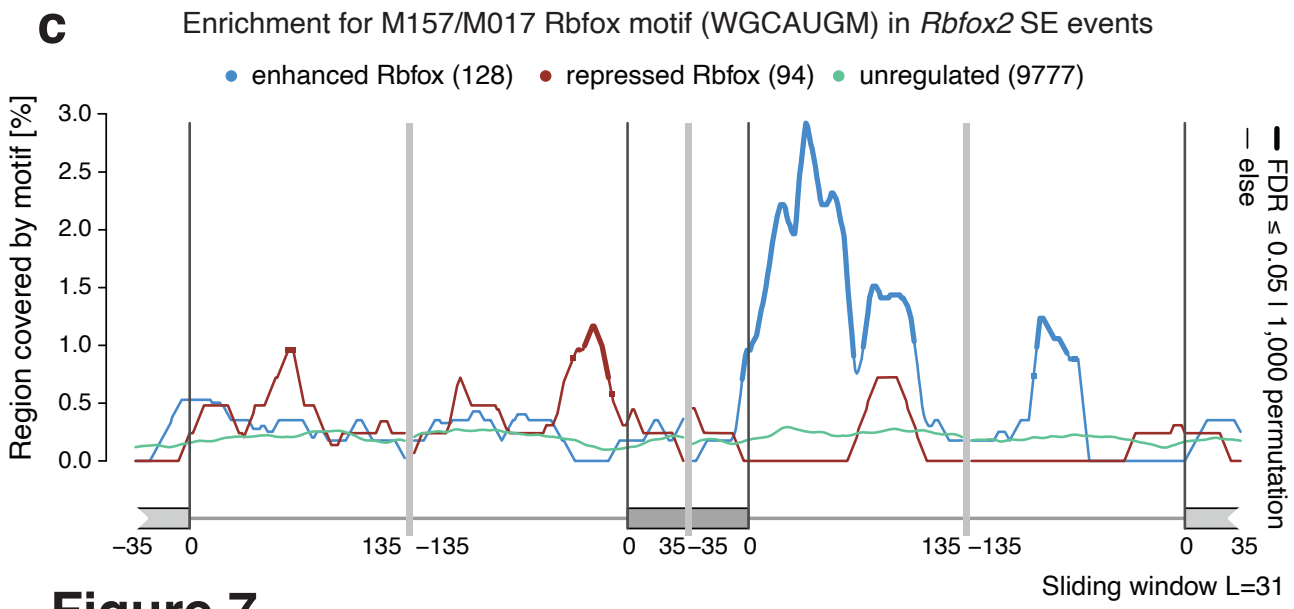

c
Enrichment for M157/M017 Rbfox motif (WGCAUGM) in Rbfox2 SE events
enhanced Rbfox (128)
repressed Rbfox (94)
unregulated (9777)
Region covered by motif [%]
FDR ≤ 0.05 | 1,000 permutation
else
Sliding window L=31

Figure 7

Supplementary Material

Contents

I. Supplementary Methods

II. Supplementary Figure Legends

III. Supplementary Tables

IV. Supplementary References

Supplementary Methods

Isolation of thymocytes and analysis by flow cytometry

Thymi and spleens were dissected from knockout and wildtype mice and isolated by gently dissociating the tissues between two frosted glass slides. Cells were filtered and resuspended in PBS containing 2% FCS (Merck) and stained with a combination of the markers given in Supplementary Table 11. The staining for surface markers was performed for 20 min at 4 ºC in the dark. The cell viability was assessed using the LIVE/DEAD Fixable Aqua Dead Cell Stain Kit (ThermoFisher Scientific) according to manufacturer's instructions. Cells were acquired using FACSAria III (BD Bioscience) and data was analyzed using FlowJo software (version 10.5.0).

Definition of tissue-restricted splicing-related factors

To assess the expression of splicing-related genes in ENCODE tissues and TEC populations, we compiled a list of known splicing related genes from (i) literature sources[1-7], (ii) the RNA binding protein database (RBPDB) [8] and (iii) relevant Gene Ontology (GO) categories. From the RBPDB database (http://rbpdb.ccbr.utoronto.ca/, downloaded 1st June

2018) we included the mouse RNA binding protein genes. Genes were included from GO categories that matched the string 'splic' (but not 'tRNA' or 'protein splicing'). GO data was retrieved from AmiGO or the GO Online SQL Environment (GOOSE) database [9] (downloaded 29th May 2018). Tissue-restricted splicing-related genes were then identified as those with *tau* > 0.5 in the TEC and peripheral tissue samples.

Procedure for identification of reproducibly detected transcripts

To identify reproducibly detected transcripts, we implemented a robust procedure based on the npIDR statistic [10,11]. First TPM values were log10 transformed (after addition of a small pseudocount = 0.001) and extreme values ($x < 0.5$ or $x > 0.95$ expression quantile) removed to avoid issues arising from data sparsity. Data were then binned (n=50 bins) and npIDR values for each bin computed as previously described [11]. To determine the expression level above which transcripts could be reliably detected we modelled the TPM vs npIDR relationship by LOESS regression. The fitted curve was used to determine the TPM value that corresponded to npIDR ≤ 0.1. The analysis was performed separately for each TEC and peripheral tissue (with n = 2 biological replicate sample pools). Determination of the TPM threshold above which transcripts could be reproducibly detected in the adrenal samples is shown in Supplementary Fig 15a-b. The TPM thresholds determined for each of the TEC and tissue samples are shown in Supplementary Fig 15c. For our datasets this procedure was more consistent and conservative than the use of per-transcript npIDR values (data not shown).

Supplementary Figure Legends

Supplementary Figure 1: Definition of tissue groups and identification of promiscuously expressed (tissue-restricted) genes (a) Definition of groups of similar peripheral tissues. The mouse ENCODE tissue samples were hierarchically clustered by

expression of known transcripts from protein-coding genes that showed variable expression (top n=24,664 most highly variable known transcripts, Manhattan distance, complete linkage, n=2 replicate pools per tissue). Tissue groups (red boxes) were defined by cutting the dendrogram at a fixed height. The tissues taken as the representatives of the groups with multiple tissues are shown in bold font and underlined. (b) The histogram shows the distribution of *tau*[12] values for protein-coding genes as computed using the selected tissues (a) and the wildtype immature and mature mTEC samples. Genes with a *tau* value of > 0.7 (as indicated by the dashed vertical line) were identified as tissue-restricted antigens (TRA). (c) Identification of *Aire*-regulated genes by differential expression analysis of protein-coding genes between *Aire*-knockout and *Aire*-positive mature mTEC (n=2 biological replicates). Genes with a significant, >2-fold downregulation in the *Aire*-knockout were defined as *Aire*-regulated genes (BH-adjusted $p < 0.05$). Reads were quantitated with FeatureCounts [13] (Ensembl v91 annotations) and differential expression analysis performed using DESeq2 [14]. (d) Venn diagram showing the overlap between tissue-restricted genes (*tau* > 0.7) and *Aire*-regulated genes. The three categories of *Aire*-regulated TRA (*Aire*-TRA), non-*Aire*-regulated TRA (non-*Aire* TRA) and non-TRA defined here are used throughout the manuscript.

Supplementary Figure 2: Numbers of transcript isoforms detected in peripheral tissues and TEC. (a-c) The fraction of isoforms detected (TPM>0) per multi-isoform protein-coding gene in mTEC and peripheral tissues. The boxplots show the fractions for *Aire*-regulated TRA genes (a), non-*Aire* TRA genes (b) and non-TRA genes (c) (see Supplementary Fig. 1d). The boxes correspond to the 25th to 75th percentile, while the upper and lower whiskers extend to 1.5*inter-quartile range (distance between the first and third quartile). Outliers are data points outside the whisker range and are plotted as individual points. (d-f) Gene expression level distributions for the testis, sets of peripheral tissues (for legibility mean TPM values are shown for all brain tissues, all digestive tract tissues and all other peripheral tissues) and the immature and mature mTEC cell populations. The histograms show the distributions for *Aire*-regulated TRA genes (d), non-*Aire* TRA genes (e)

and non-TRA genes (f). The three dashed lines indicate 1, 10 and 100 TPM. (g-i) The relationship between gene expression level (x axis) and isoform detection (y axis) for the testis, sets of peripheral tissues (sets of tissues defined as for d-e; mean isoform fractions) and immature and mature mTEC. The plots show the LOESS regression curves for *Aire*-regulated TRA genes (g), non-*Aire* regulated TRA genes (h) and non-TRA genes (i). The analyses for all panels was performed using the high-depth samples (n=1). For the analyses of TRA in peripheral tissues (a, b, d, e, g, h) gene statistics were only counted for the peripheral tissue in which the gene was most highly expressed. For the analyses of TRA in TEC (a, b, d, e, g, h) the full sets of TRA genes were quantitated in each of the TEC populations.

Supplementary Figure 3: Examples of known tissue-specific alternative splicing events recapitulated in TEC. (a) Transcript models, read coverage and sashimi plots for the gene *Actinin1*. Mature mTEC express both the muscular and non-muscular isoforms of *Actinin1* which are differentiated by expression of the indicated mutually exclusive exon pair (red box) [15] (b) Transcript models, read coverage and sashimi plots for the gene *Tjp1*. Mature mTEC express isoforms with and without exon 20 (red box). In the periphery inclusion of exon 20 is known to be highly tissue specific [6]. (c) Transcript models, read coverage and sashimi plots for the gene *Calca*. The shorter isoform, in which inclusion of exon 4 (red box) introduces a premature stop codon, is specific to the thyroid and produces the Calcitonin (CT) peptide [2,16]. In the nervous system, exon 4 (red box) is skipped to form the longer α-CGRP isoform. Mature mTEC express both of these isoforms. For the Illumina sequencing, the sashimi plots in a-c were generated from the single high-depth sample of mature mTEC. Only splice junctions with coverage of ≥ 5 reads are shown. (d) Detection of exon 4 inclusion and exclusion in *Calca* transcripts in single mature mTEC. 35 % of the cells where *Calca* is detected produced *Calca* transcripts both with and without this exon.

Supplementary Figure 4: Comparison of transcript and gene expression in TEC and mouse ENCODE tissues. The values shown in (a-d) represent the mean of n=2 biologically replicate sample pools (60M reads/sample as described in the Methods and Supplementary Table 1) (a) The curves show the fractions of transcripts from protein-coding genes that were detected in the peripheral tissues and mTEC populations across a range of TPM thresholds (mT&T assembly). (b) The scatter plot shows the relationship between the number of genes and transcripts that were detected in the peripheral tissues and mTEC populations (mT&T assembly). (c) The scatter plot shows the relationship between the number of known (Ensembl version 91) and novel (mT&T assembly) transcripts detected in the peripheral tissues and mTEC populations. Trend lines shown in (b) and (c) were fitted to all samples except for those from TEC and the testis. (d) The fractions of sets of tissue-restricted transcripts (*tau* ≥ 0.9) from peripheral tissues that were detected in mature mTEC (mT&T assembly). (e) Saturation analysis of splice-junction detection in protein-coding genes in TEC and peripheral tissues. Numbers of unique splice junctions (>0 reads, y axis) detected in sub-samples of sequencing reads (x axis) (replicates were combined prior to sub-sampling).

Supplementary Figure 5: Long-read sequencing of mTEC using Oxford Nanopore Technology. (a) Histograms of read length for Oxford Nanopore Technology (ONT) experiments. Immature and mature mTEC samples shown here represent the merged reads from n=3 biological replicates. (b) Histograms of number of splices detected per read for the samples of merged ONT reads from immature and mature mTEC. (c) Gap-compressed divergence as a measure of error (0 representing no errors in alignment) from Minimap2 for the merged ONT reads from the immature and mature mTEC samples. (d) The numbers of protein-coding genes detected in merged ONT read samples from mature and immature mTEC (after normalising read numbers between the populations by downsampling). (e) Number of *Aire*-regulated genes or tissue-specific genes (*tau* > 0.7) detected in the merged sample of ONT reads from mature mTEC.

Supplementary Figure 6: Identification of a set of tissue-restricted splicing factors. (a) Compilation of a list of tissue-restricted splicing-related factor genes (see Supplementary Methods). (b) Heatmap of expression level of the 50 tissue-restricted splicing-related factors with the highest *tau* values across the included peripheral tissue and TEC samples (excluding predicted genes (those beginning with 'Gm') and RIKEN clones).

Supplementary Figure 7: Gene and transcript expression of the Rbfox family across peripheral tissues and TEC. (a) Heatmap of the expression of the Rbfox family across different mouse ENCODE tissue samples and TEC populations (columns are hierarchically clustered). (b) and (c) transcript models, read coverage and sashimi plots for *Rbfox1* and *Rbfox2* in the mature mTEC population and mouse ENCODE cortex and heart samples (generated with MISO [17] from a single high-depth, deduplicated sample per tissue).

Supplementary Figure 8: Phenotyping of thymi and thymic epithelial cells from Rbfox1$^{lox/lox}$:cre-/- (cre-) and Rbfox1$^{lox/lox}$:cre+/- (cre+). (a) PCR analysis of genomic exons 11 and 12 of *Rbfox1*. (b) Representative image of gross anatomy of the thymus in cre- and *Rbfox1* tKO mice. (c) Total thymus cellularity is shown with symbols indicating the sex of each mouse. (d) Representative graph demonstrating gating for TEC (EPCAM+, CD45-) among non-enriched cells isolated from the indicated mouse strains and (e) their frequencies. (f) Representative graph demonstrating gating for TEC subtypes from enriched cells (EPCAM+, CD45-) and (g) their frequencies. (h) Subsequent gating for mTEC subtypes defined by MHCII and CD80 status and (i) their frequencies. In (g) and (i), the two independent experiments are indicated by different symbols in the bar graphs. Data presented in bar graphs represents the combined results from n=2 independent experiments with n=6 mice in each experiment (4-6 weeks old). Significant differences of p-value < 0.05

between Rbfox1$^{lox/lox}$:cre-/- and Rbfox1$^{lox/lox}$:cre+/- using a two-sided Welch Two Sample t-test are indicated with * (mean ± SE shown in bar graphs).

Supplementary Figure 9: Phenotyping of thymi and thymic epithelial cells from Rbfox2$^{lox/lox}$:cre-/- (cre-) and Rbfox2$^{lox/lox}$:cre+/- (cre+). (a) PCR analysis of genomic exons 6 and 7 of *Rbfox2*. (b) Representative image of gross anatomy of the thymus in cre- and *Rbfox2* tKO mice. (c) Total thymus cellularity is shown with symbols indicating the sex of each mouse. (d) Representative graph demonstrating gating for TEC (EPCAM+, CD45-) among non-enriched cells isolated from the indicated mouse strains and (e) their frequencies. The symbols indicate the sex of each mouse. (f) Representative graphs demonstrating the gatings for the mTEC subtypes as defined by MHCII and CD80 status (EPCAM+, CD45-, UEA-1+, Ly51-) and (g) their frequencies. A representative gating for mTEC versus cTEC based on UEA-1 and Ly51 is shown in Fig. 5b. The two independent experiments are indicated by different symbols in the bar graph. Data presented in bar graphs represents the combined results from n=2 independent experiments with n=6 mice in each experiment (4-6 weeks old). Significant differences of p-value < 0.05 between Rbfox2$^{lox/lox}$:cre-/- and Rbfox2$^{lox/lox}$:cre+/- using a two-sided Welch Two Sample t-test are indicated with * (mean ± SE shown in bar graphs).

Supplementary Figure 10: Phenotyping of thymi and thymic epithelial cells from Rbfox1$^{lox/lox}$:Rbfox2$^{lox/lox}$:cre-/- and Rbfox1$^{lox/lox}$: Rbfox2$^{lox/lox}$:cre+/- (cre+). (a) PCR analysis of genomic exons 11 and 12 of *Rbfox1* (left) and exons 6 and 7 of *Rbfox2* (right). (b) Representative image of gross anatomy of the thymus in cre- and *Rbfox1*:*Rbfox2* tKO mice. (c) Total thymus cellularity is shown with symbols indicating the sex of each mouse. (d) Representative graph demonstrating gating for TEC (EPCAM+, CD45-) among non-enriched cells isolated from the indicated mouse strains and (e) their frequencies. The symbols indicate the sex of each mouse. (f) Representative graphs demonstrating the gatings for the TEC subtypes from enriched cells (EPCAM+, CD45-) and (g) their

frequencies. (h) Subsequent gating for mTEC subtypes defined by MHCII and CD80 status and (i) their frequencies. In (g) and (i), the n=3 independent experiments are indicated by different symbols. Data presented in bar graphs represents the combined results from n=3 independent experiments (4-6 weeks old; n=8 animals per genotype; equal numbers of males and females per genotype). Significant differences of p-value < 0.05 or < 0.01 between Rbfox1$^{lox/lox}$:Rbfox2$^{lox/lox}$:cre-/- and Rbfox1$^{lox/lox}$:Rbfox2$^{lox/lox}$:cre+/- using a two-sided Welch Two Sample t-test are indicated with * or **, respectively (mean ± SE shown in bar graphs).

Supplementary Figure 11: Phenotyping of developing thymocytes from Rbfox1$^{lox/lox}$:Rbfox2$^{lox/lox}$:cre-/- (cre-) and Rbfox1$^{lox/lox}$: Rbfox2$^{lox/lox}$:cre+/- (cre+). (a) Gating strategy to define developmental stages and selection of thymocytes within the thymus. DN = double-negatives, SP = single-positive, DP = double-positive. (b) Representative graphs demonstrating the gatings for CD4 and CD8 in total, live thymocytes from the two genotypes. The gates define CD4 single-positive (SPCd4), CD8 single-positive (SPCd8), double-negative (DNs) and double-positive (DP) thymocytes. (c) Frequencies of cell types as defined by gating in (b) from two independent experiments. (d) Representative graphs to demonstrate the gating for CD69 and TCR-β in total, live thymocytes from the two genotypes. The stages represent positive selection of thymocytes, which is characterised by upregulation of CD69 indicating an interaction between the TCR and MHC molecules. (e) Frequencies of cell types as defined by gating in (d) from two independent experiments. (f) Representative graph demonstrating gating for CD44 and CD25 in CD24+ double-negative thymocytes. The four quadrants represent the double-negative stages, where Q1 corresponds to DN1 (CD44+CD25-), Q2 to DN2 (CD44+CD25+), Q3 to DN3 (CD44-CD25+) and Q4 to DN4 (CD44-CD25-). (g) Frequencies of cells in DN stages as defined by gating in (f) from two independent experiments.

Supplementary Figure 12: Differential gene expression and splicing in immature mTEC from *Rbfox1* tKO and *Rbfox2* tKO. (a) The scatterplot shows changes in gene expression between the immature mTEC from *Rbfox1* tKO vs cre- littermates (x axis) and immature mTEC from Rbfox2 tKO vs cre- littermates (y axis). Genes significantly differentially expressed (n=2 biological replicates, DESeq2, BH adjusted $p < 0.05$, $|fc| > 1.5$) in both knockouts are colored in orange, in the *Rbfox1* tKO alone in green and in the *Rbfox2* tKO alone in blue. (b) The numbers of protein-coding genes containing significantly differentially spliced events identified in immature mTEC extracted from *Rbfox1* tKO and *Rbfox2* tKO animals vs respective littermate controls (n=2 biological replicates, rMATS, FDR < 0.05, |delta (d)PSI| > 0.2, protein-coding genes only). (c) Breakdown of the identified splicing events (b) by event type and promiscuous expression status (Supplementary Fig. 1). (d) Enrichment of the M159/M017 Rbfox RRM recognition motif in the sequence around exons found to be significantly regulated by Rbfox2 in immature mTEC (|dPSI| > 0.2, FDR < 0.05). The lines show enrichments for the sets of exons that were found to be enhanced (blue) or repressed (red) or not significantly regulated (green) by Rbfox2. Thicker lines indicate regions of statistically significant enrichment (FDR ≤ 0.05, n=1,000 permutations).

Supplementary Figure 13: The peripheral tissue expression of genes differential spliced by *Rbfox2* in mature mTEC. The heatmap shows the expression of all TRA genes with differential splicing events in *Rbfox2* tKO vs littermate controls across the peripheral tissues. The source of highest expression in the periphery and *Aire* regulation status are shown as colored side bars.

Supplementary Figure 14: Example gating strategy for live cells. (a) From left to right: Representative gating to define cells based on FSC (area) and SSC (area), to define single cells based on FSC (height) and SSC (height) and to define live cells based on the absence of DAPI staining. (b) From left to right: Representative gating to define cells based on FSC

(area) and SSC (area), to define single cells based on FSC (height) and SSC (height) and to define live cells based on the absence of AQUA staining.

Supplementary Figure 15: Procedure for identification of reproducibly detected transcripts. (a) The scatter plot shows the relationship between expression level and npIDR for the adrenal samples. The relationship was modelled by LOESS regression (green curve). The fitted curve was then used to determine the TPM threshold corresponding to npIDR of 0.1 (dashed lines). (b) The scatter plot shows transcript expression levels in the two adrenal samples. Transcripts (dots) are colored according to whether they pass (red) or fail (blue) the determined TPM expression level filter. (c) The TPM thresholds determined to correspond to npIDR $\leq$ 0.1 for each of the TEC and peripheral tissue samples.

Supplementary Tables

Tissue	Data source	Genotype	No. bio-replicates	High-depth sample	Replicate pool 1	Replicate pool 2
Adrenal	mouse ENCODE project	wildtype	6	all replicates	R1; R5	R2; R4
Colon		wildtype	6	all replicates	R1; R4	R2; R5
Duodenum		wildtype	7	all replicates	R1; R5	R2; R3
Genital Fat Pad		wildtype	4	all replicates	R3	R4
Heart		wildtype	3	all replicates	R1	R3
Kidney		wildtype	6	all replicates	R1	R4
Liver		wildtype	6	all replicates	R1; R6	R3; R4
Lung		wildtype	4	all replicates	R1	R2; R3
Mammary Gland		wildtype	6	all replicates	R3	R2; R6
Ovary		wildtype	9	all replicates	R1; R10; R2	R4; R5; R6; R7
Small Intestine		wildtype	9	all replicates	R1; R2	R5; R6; R7
Spleen		wildtype	6	all replicates	R1; R2	R3; R4
Stomach		wildtype	6	all replicates	R4; R5	R1; R2
Subc Fat Pad		wildtype	4	all replicates	R1	R4
Testis		wildtype	4	all replicates	R1	R3
Thymus		wildtype	6	all replicates	R4	R2; R3
Frontal Lobe		wildtype	2	all replicates	R1	R2
Bladder		wildtype	2	all replicates	R1	R2
Cortex		wildtype	2	all replicates	R1	R2
Placenta		wildtype	2	all replicates	R1	R2
Cerebellum		wildtype	2	all replicates	R1	R2
Immature mTEC	this study	wildtype	2	all replicates	R1	R2
Mature mTEC		wildtype	2	all replicates	R1	R2
Aire-positive mTEC		$Aire^{wildtype/GFP}$ (GFP+ve)	2	all replicates	R1	R2
Aire-Knockout mTEC		$Aire^{GFP/GFP}$ (GFP+ve)	2	all replicates	R1	R2

Supplementary Table 1: Summary of samples used for the mT&T assembly. The "High-depth sample", "Replicate pool 1" and "Replicate pool 2" columns give the identifiers of the bio-replicates that were merged to generate the 200M read high-depth samples and the two 60M read replicate pools. To maintain biological replication the two 60M replicate pools were constructed from non-overlapping sets.

Supplementary Table 2: Classification of protein-coding genes according to their *Aire*-regulation status and tissue- specificity.

(xlsx file)

Supplementary Table 3: Novel transcripts specific to mature mTEC (Fig. 2b-c).

(xlsx file)

Supplementary Table 4: Gene ontology categories enriched in protein-coding genes with TEC-specific novel transcripts (Fig. 2d).

(xlsx file)

Supplementary Table 5: Splicing events associated with maturation and *Aire* in TEC (Fig. 3a-b).

(xlsx file)

Supplementary Table 6: *Aire* regulated transcripts in mTEC (Fig. 3c-d).

(xlsx file)

Supplementary Table 7: Compilation of splicing-related genes (Fig. 4a).

(xlsx file)

Splicing factor	Tissue described in	Reference (PMID)
Brdt	Testis	15261828
Celf4, Celf6	Neurons	22180311
Cwc22	involved in spliceosome	23236153
Elavl2, Elavl3, Elavl4	Neurons	9096138
Esrp1	Epithelium	26371508
Khdrbs2	Neurons	24469635
Mbnl3	Muscle	25183524
Msi1	NSC, cancer EMT, photoreceptors	25380226, 27541351
Nova1, Nova2	Neurons	17065982
Rbfox1, Rbfox3	Neurons	27748060
Rbm11	Neurons, Testis	21984414
Rbm20, Rbm24	Muscle, Heart	22466703, 25313962
Slu7	Liver	24865429
Snrpn	embryo, neuronal (imprinted region)	25238490
Srpk2	Testis	18653532
Srpk3	Muscle, Heart, Spleen	16140986
Srrm4	Neurons	25525873, 25838543
Syncrip	Neurons	30649277

Supplementary Table 8: Tissue restricted and functionally investigated splicing factors (Fig. 4a).

Supplementary Table 9: Genes regulated by *Rbfox1* and *Rbfox2* in mTEC (Fig. 6a).

(xlsx file)

Supplementary Table 10: Splicing events regulated by *Rbfox1* and *Rbfox2* in immature and mature mTEC (Figs 6, 7, Supplementary Fig. 12).

(xlsx file)

Antigen	Fluorophore	Concentration	Clone	Company
CD8a	AF700	1/200	53-6.7	BioLegend
CD4	APC-Cy7	1/1000	RM4-5	BioLegend
TCRb	PE	1/1000	H57-597	eBioscience
CD24	FITC	1/1000	M1/69	BioLegend
CD25	eFluor450	1/500	PC61.5	eBioscience
CD44	PE TR (PE eFluor 605)	1/1000	IM7	eBioscience
CD69	PE-Cy5	1/1000	H1.2F3	BioLegend
Streptavidin	APC	1/1000		BioLegend
CD11b	Biotin	1/1000	M1/70	BioLegend
CD11c	Biotin	1/1000	N418	BioLegend
Gr1	Biotin	1/1000	RB6-8C5	BioLegend
CD19	Biotin	1/1000	1D3	eBioscience
CD49b	Biotin	1/1000	DX5	BioLegend
F4/80	Biotin	1/1000	BM8	BioLegend
NK1.1	Biotin	1/500	PK136	BioLegend
TCRgd	Biotin	1/1000	eBioGL3 (GL-3, GL3)	eBioscience
Ter119	Biotin	1/1000	TER-119	BioLegend

Supplementary Table 11: List of antibodies used for Supplementary Fig. 11. The columns indicate the target antigen, the conjugated fluorophore, the concentration used in experiments, the antibody clone and the company.

Supplementary References

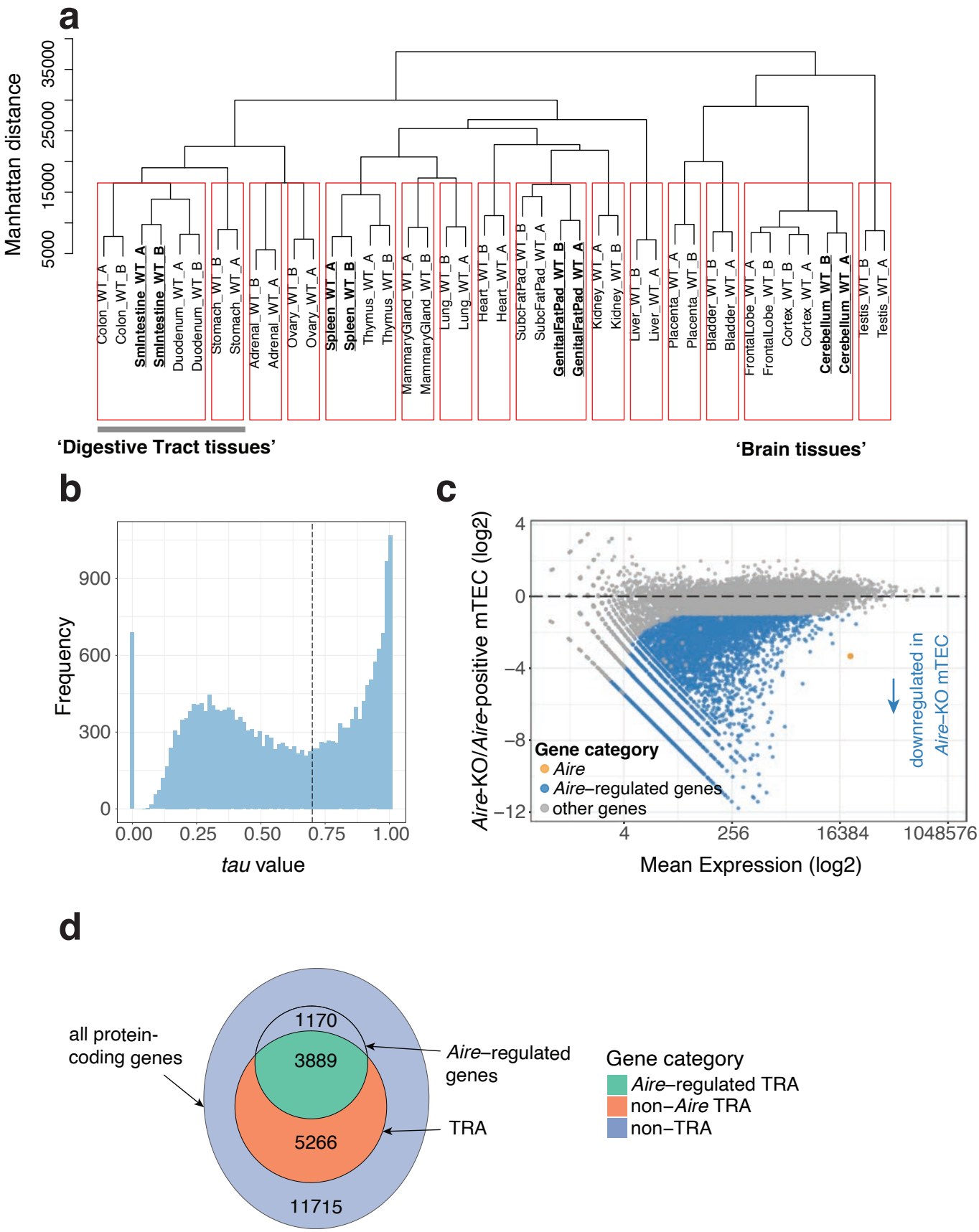
1 Barbosa-Morais, N. L., Carmo-Fonseca, M. & Aparicio, S. Systematic genome-wide annotation of spliceosomal proteins reveals differential gene family expansion. *Genome Res* **16**, 66-77, doi:10.1101/gr.3936206 (2006).

2 Chen, M. & Manley, J. L. Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches. *Nat Rev Mol Cell Biol* **10**, 741-754, doi:10.1038/nrm2777 (2009).

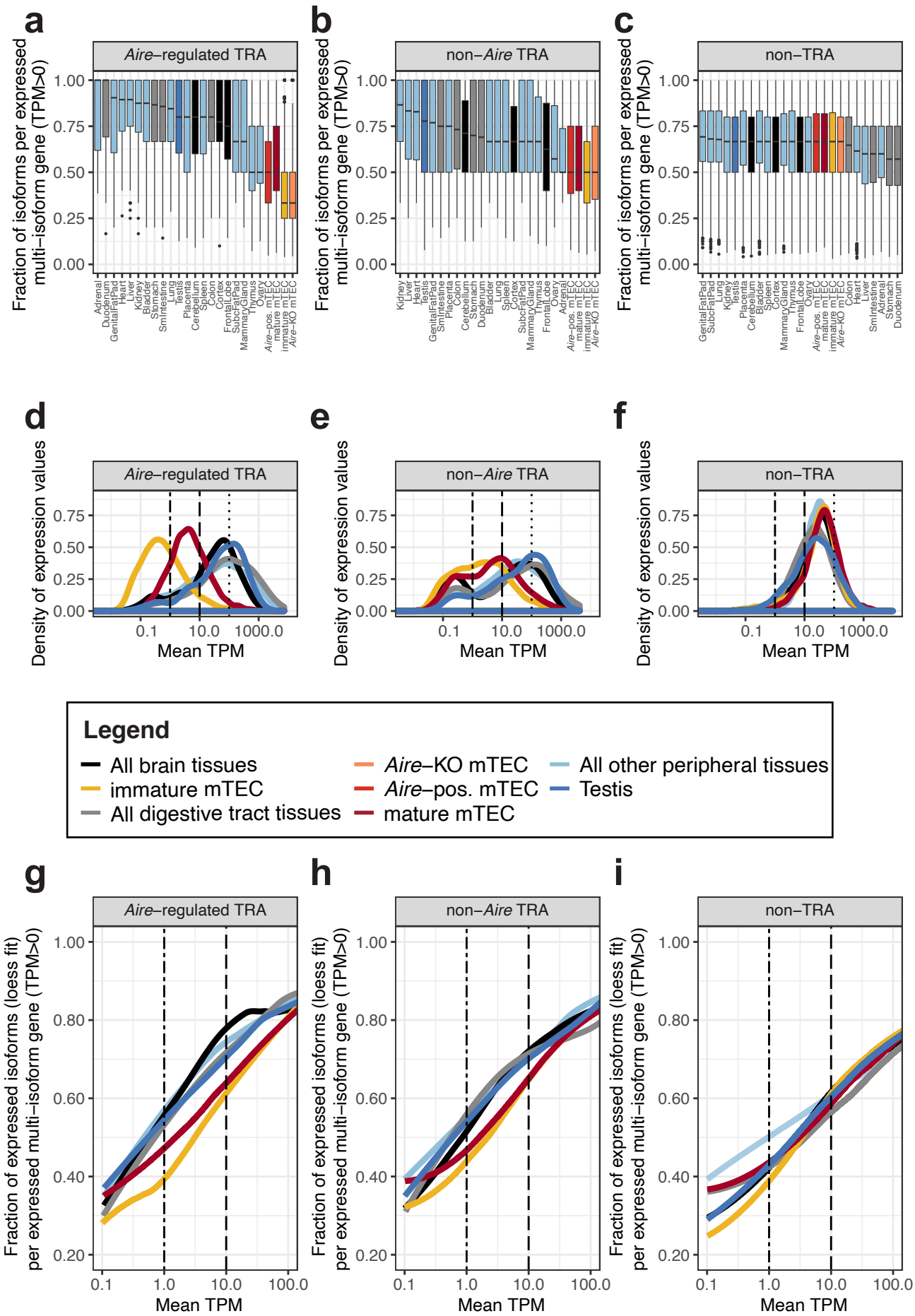
3 Grosso, A. R. *et al.* Tissue-specific splicing factor gene expression signatures. *Nucleic Acids Res* **36**, 4823-4832, doi:10.1093/nar/gkn463 (2008).

4 Han, H. *et al.* MBNL proteins repress ES-cell-specific alternative splicing and reprogramming. *Nature* **498**, 241-245, doi:10.1038/nature12270 (2013).

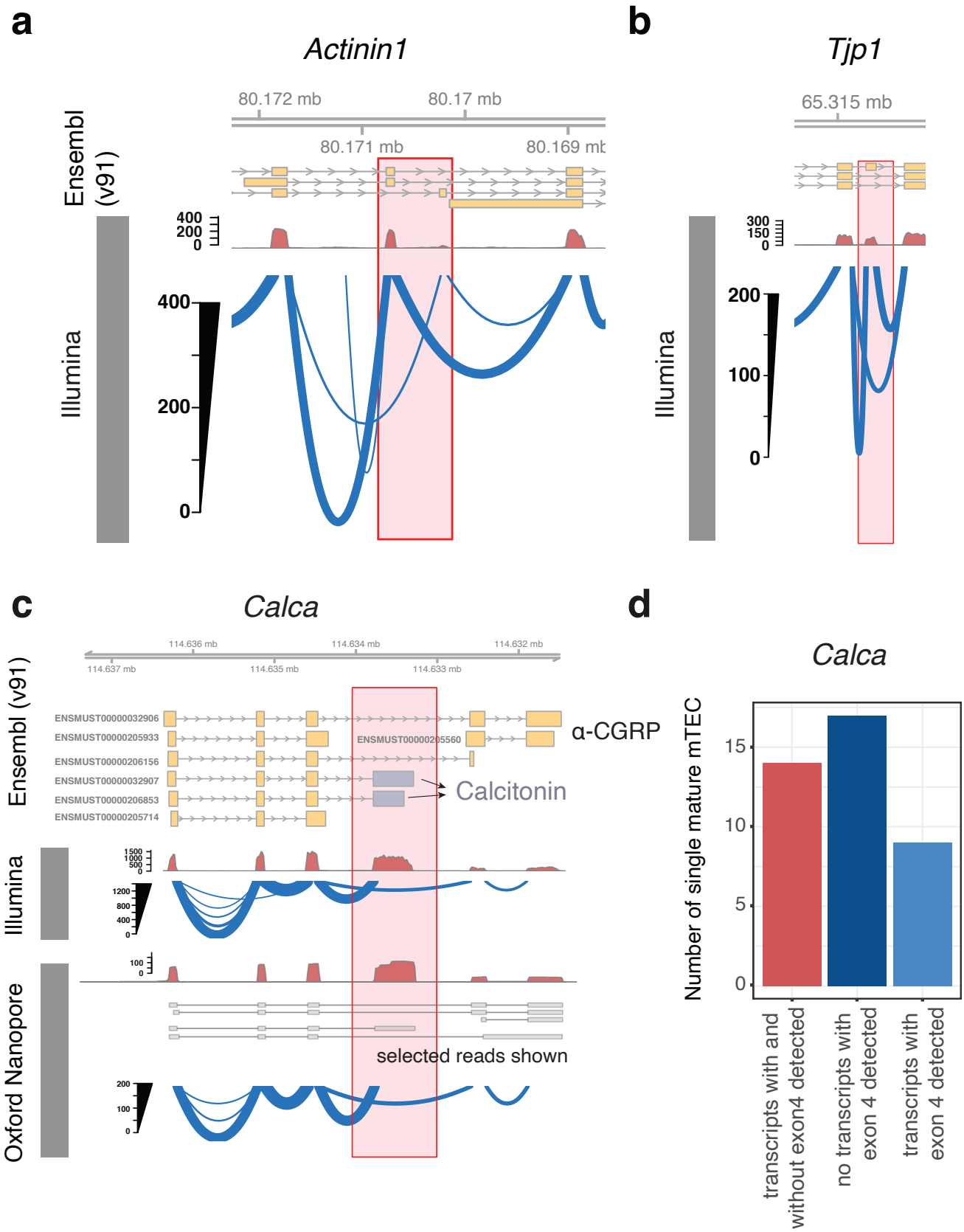
5 Jangi, M. & Sharp, P. A. Building robust transcriptomes with master splicing factors. *Cell* **159**, 487-498, doi:10.1016/j.cell.2014.09.054 (2014).
6 Merkin, J., Russell, C., Chen, P. & Burge, C. B. Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. *Science* **338**, 1593-1599, doi:10.1126/science.1228186 (2012).
7 St-Pierre, C., Trofimov, A., Brochu, S., Lemieux, S. & Perreault, C. Differential Features of AIRE-Induced and AIRE-Independent Promiscuous Gene Expression in Thymic Epithelial Cells. *J Immunol* **195**, 498-506, doi:10.4049/jimmunol.1500558 (2015).
8 Cook, K. B., Kazan, H., Zuberi, K., Morris, Q. & Hughes, T. R. RBPDB: a database of RNA-binding specificities. *Nucleic Acids Res* **39**, D301-308, doi:10.1093/nar/gkq1069 (2011).
9 Carbon, S. *et al.* AmiGO: online access to ontology and annotation data. *Bioinformatics* **25**, 288-289, doi:10.1093/bioinformatics/btn615 (2009).
10 Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21, doi:10.1093/bioinformatics/bts635 (2013).
11 Pervouchine, D. D. *et al.* Enhanced transcriptome maps from multiple mouse tissues reveal evolutionary constraint in gene expression. *Nat Commun* **6**, 5903, doi:10.1038/ncomms6903 (2015).
12 Kryuchkova-Mostacci, N. & Robinson-Rechavi, M. A benchmark of gene expression tissue-specificity metrics. *Brief Bioinform* **18**, 205-214, doi:10.1093/bib/bbw008 (2017).
13 Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923-930, doi:10.1093/bioinformatics/btt656 (2014).
14 Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550, doi:10.1186/s13059-014-0550-8 (2014).
15 Gromak, N., Matlin, A. J., Cooper, T. A. & Smith, C. W. Antagonistic regulation of alpha-actinin alternative splicing by CELF proteins and polypyrimidine tract binding protein. *RNA* **9**, 443-456, doi:10.1261/rna.2191903 (2003).
16 Chew, S. L. Alternative splicing of mRNA as a mode of endocrine regulation. *Trends Endocrinol Metab* **8**, 405-413, doi:10.1016/s1043-2760(97)00167-7 (1997).
17 Katz, Y., Wang, E. T., Airoldi, E. M. & Burge, C. B. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat Methods* **7**, 1009-1015, doi:10.1038/nmeth.1528 (2010).



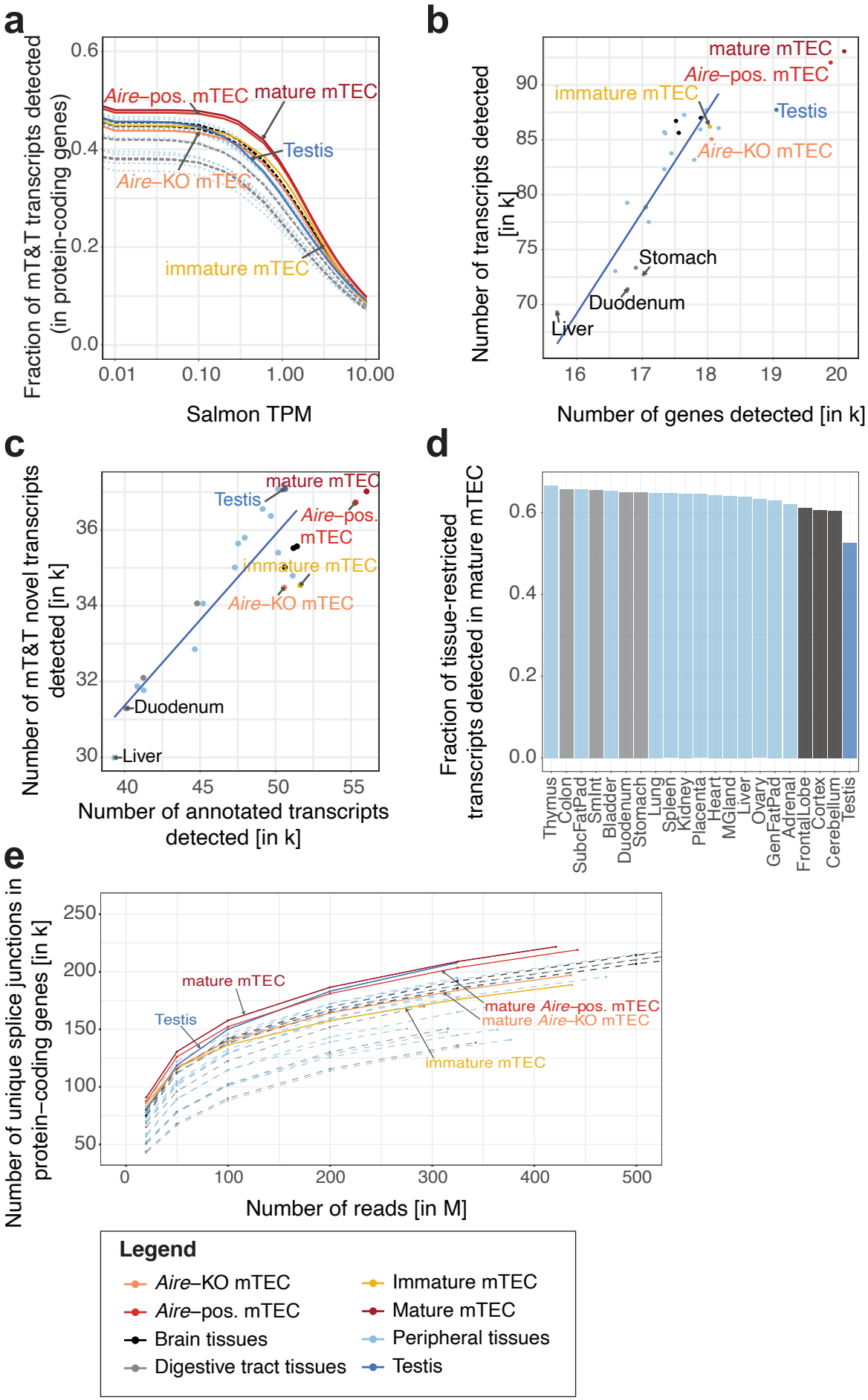
Supplementary Figure 1



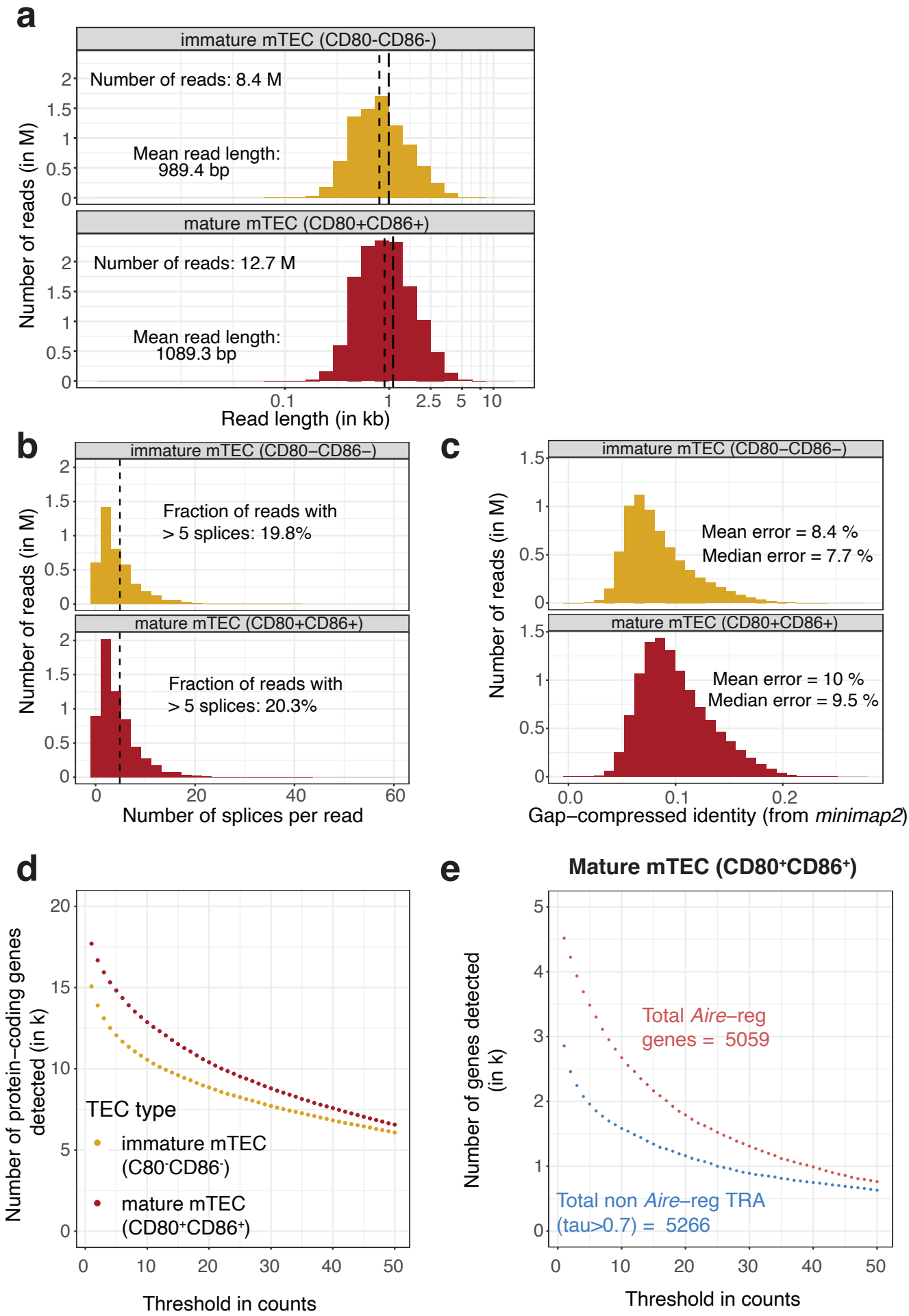
Supplementary Figure 2



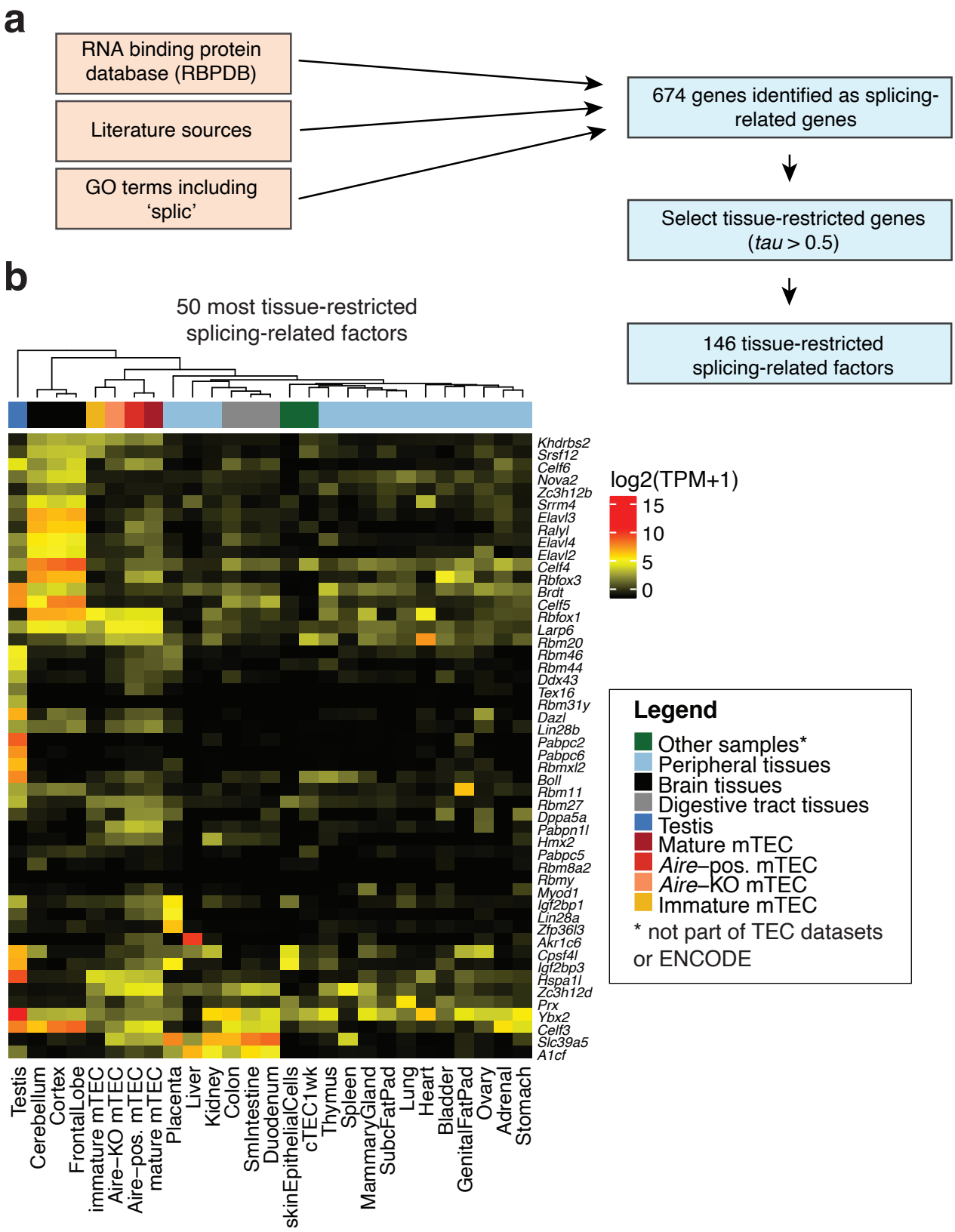
Supplementary Figure 3



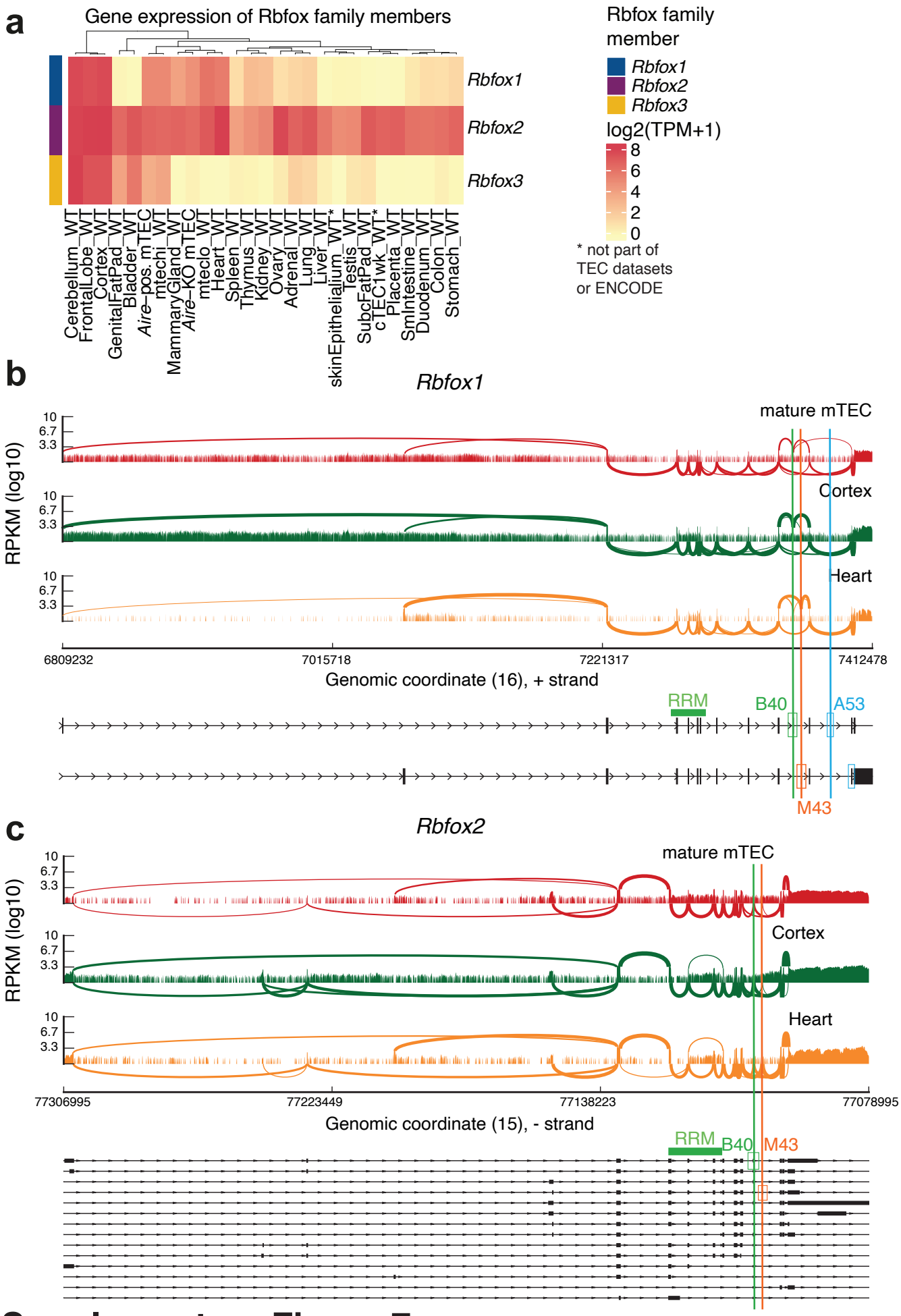
Supplementary Figure 4



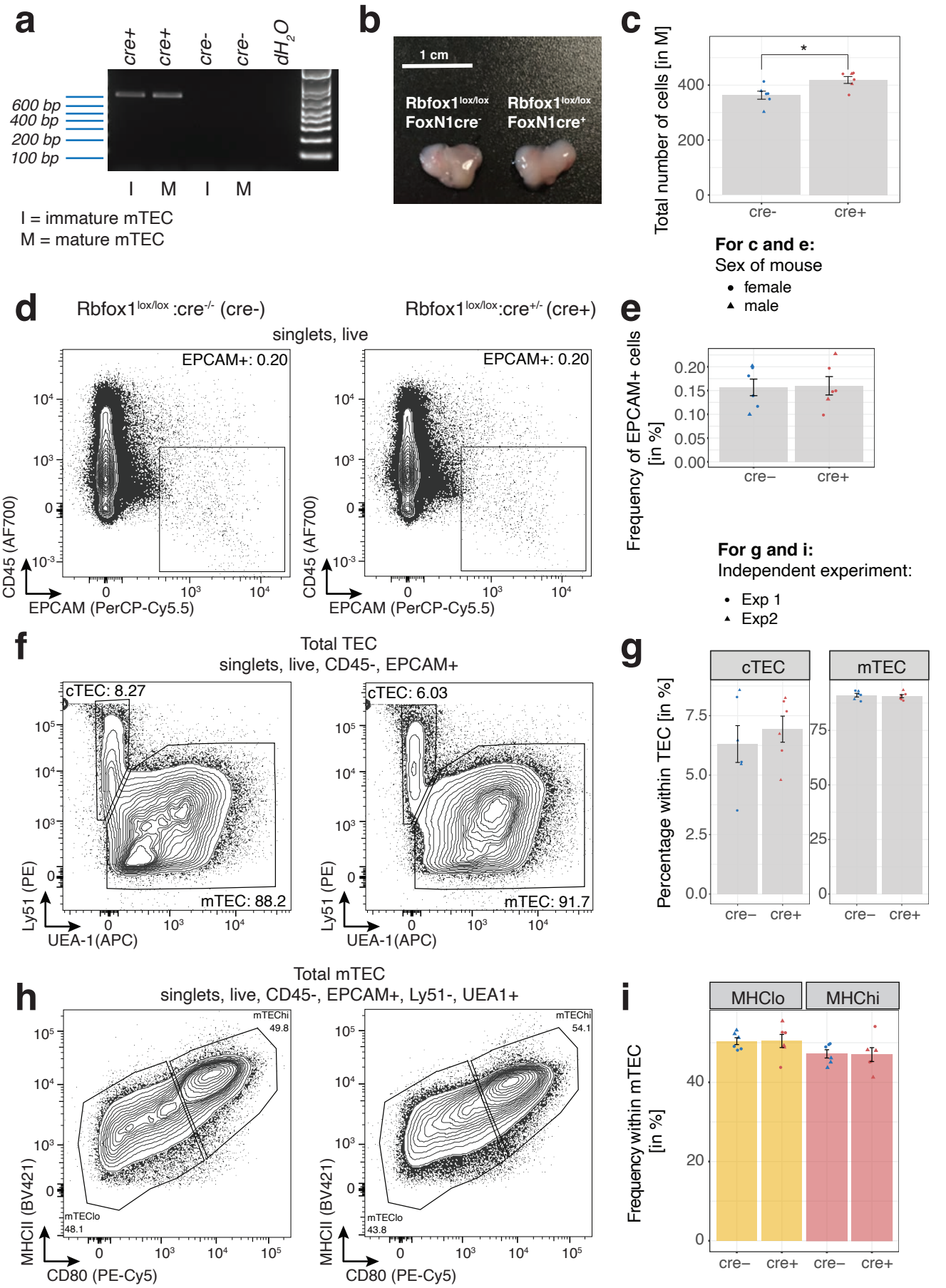
Supplementary Figure 5



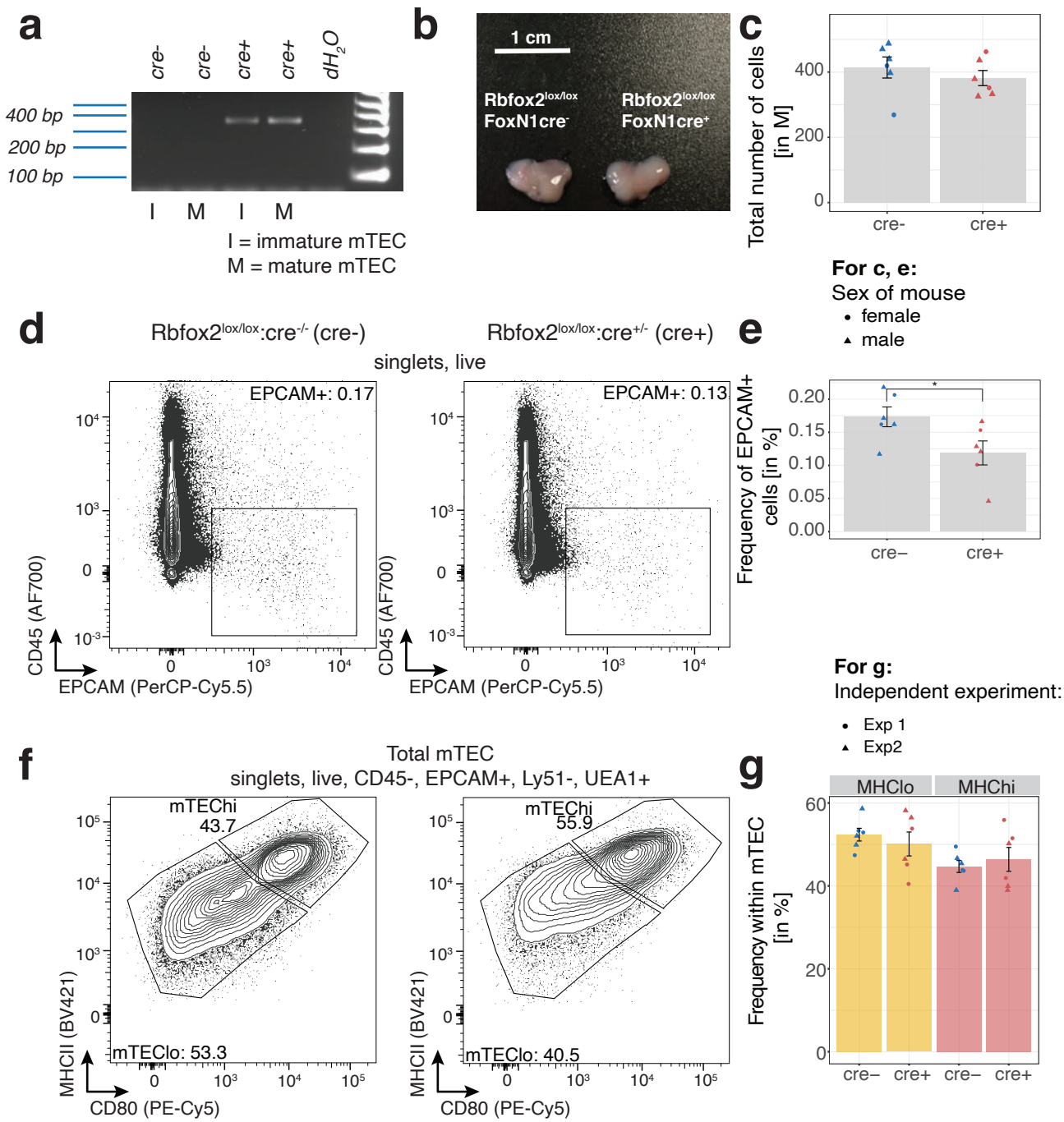
Supplementary Figure 6



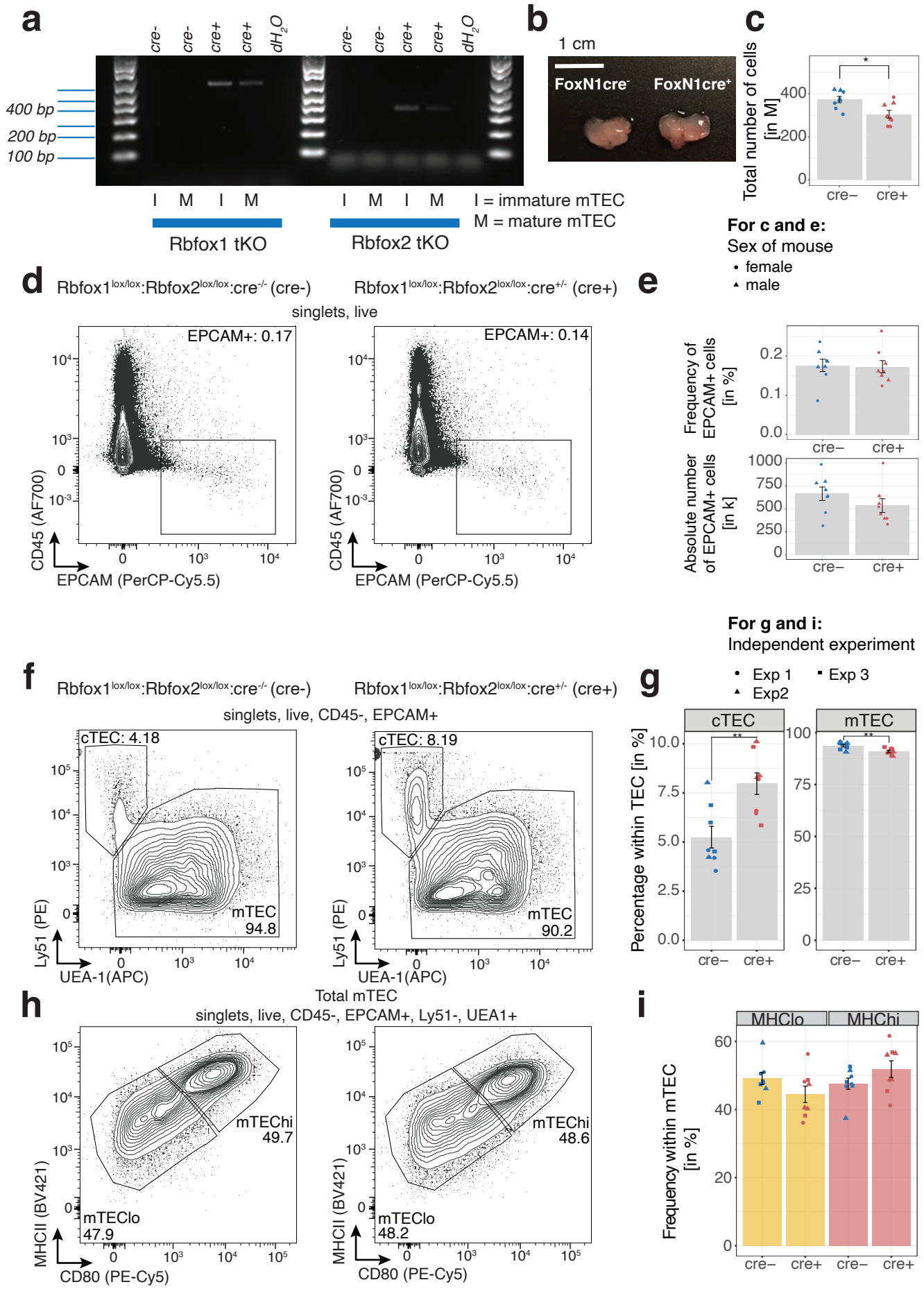
Supplementary Figure 7



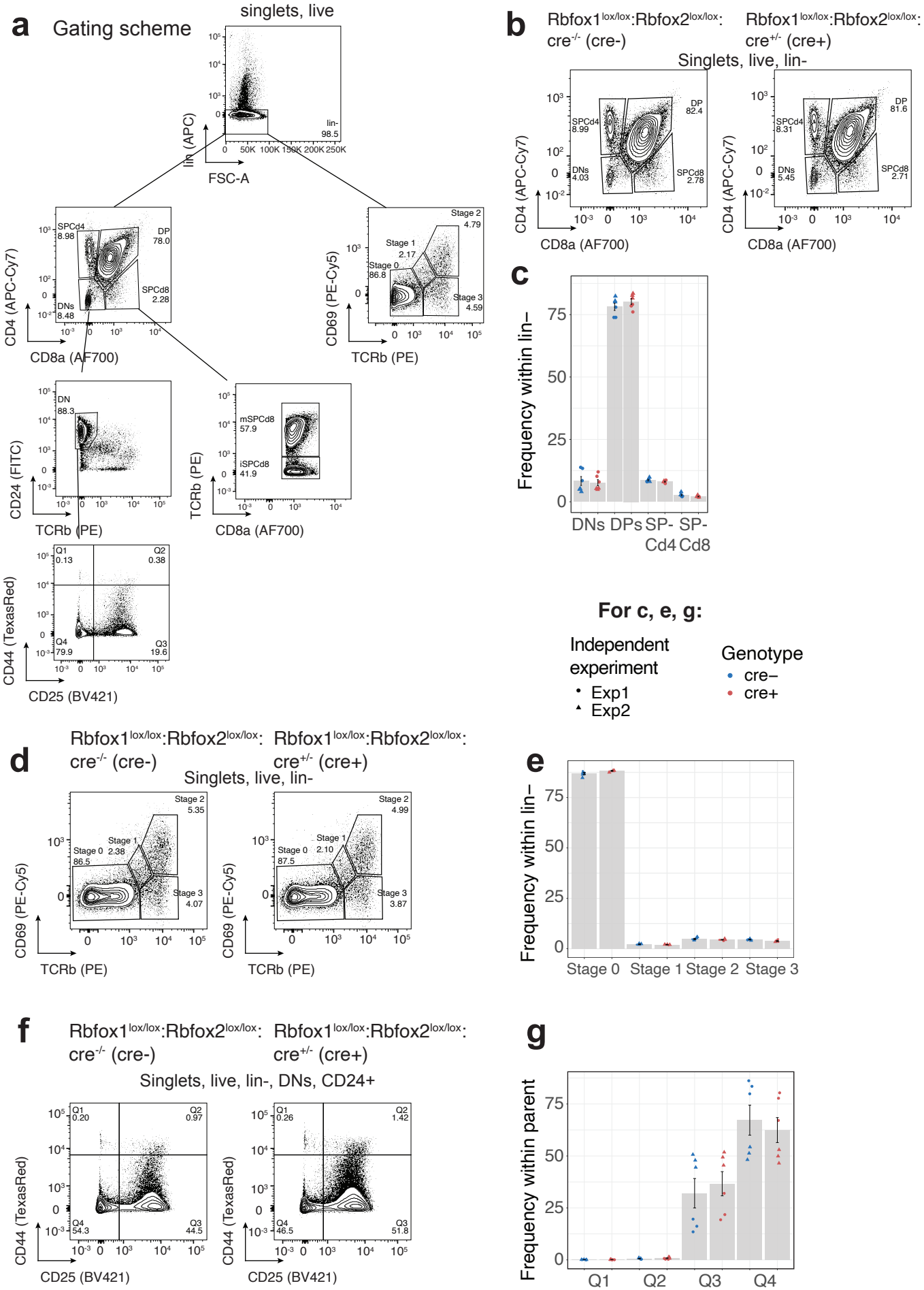
Supplementary Figure 8



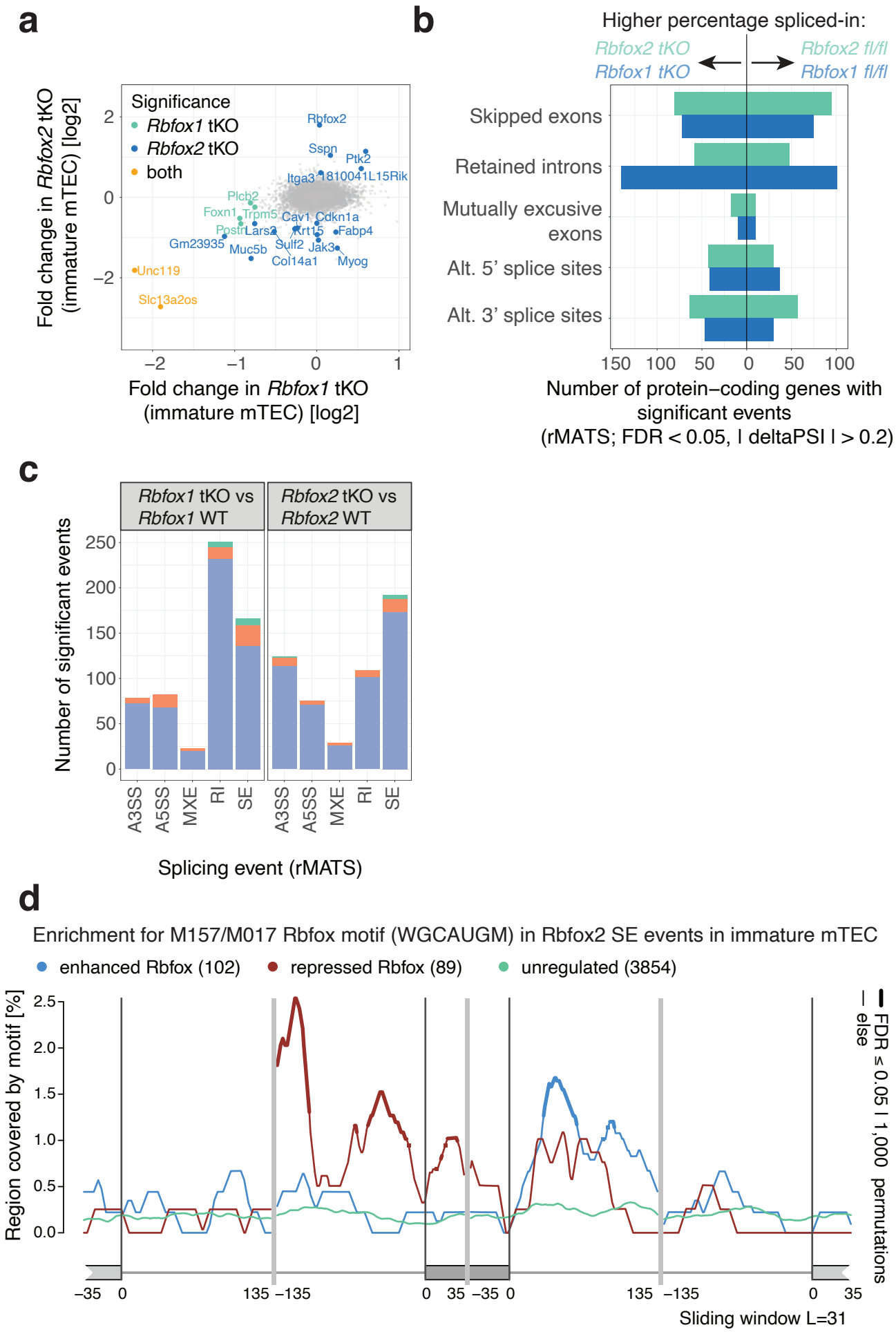
Supplementary Figure 9



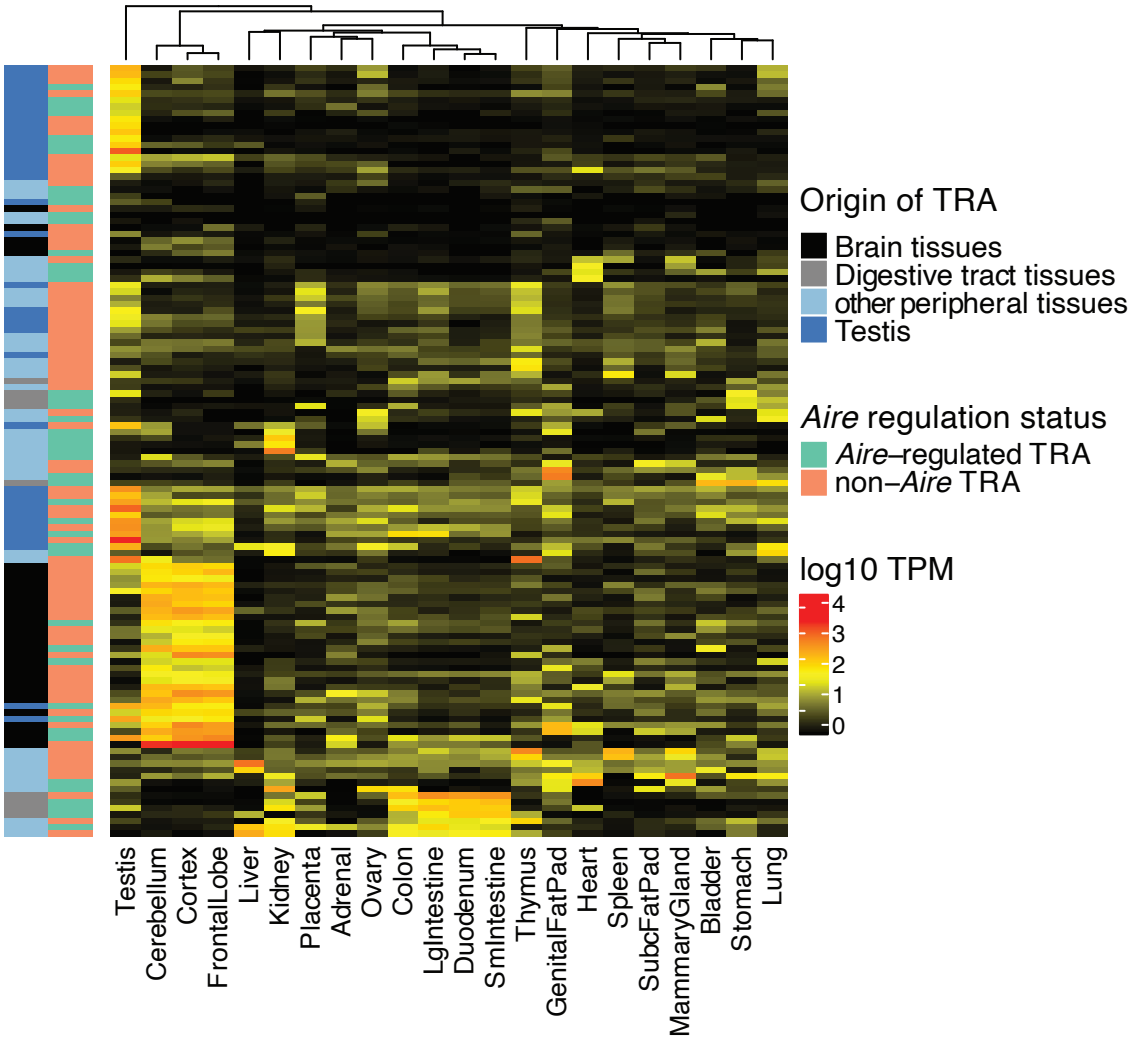
Supplementary Figure 10



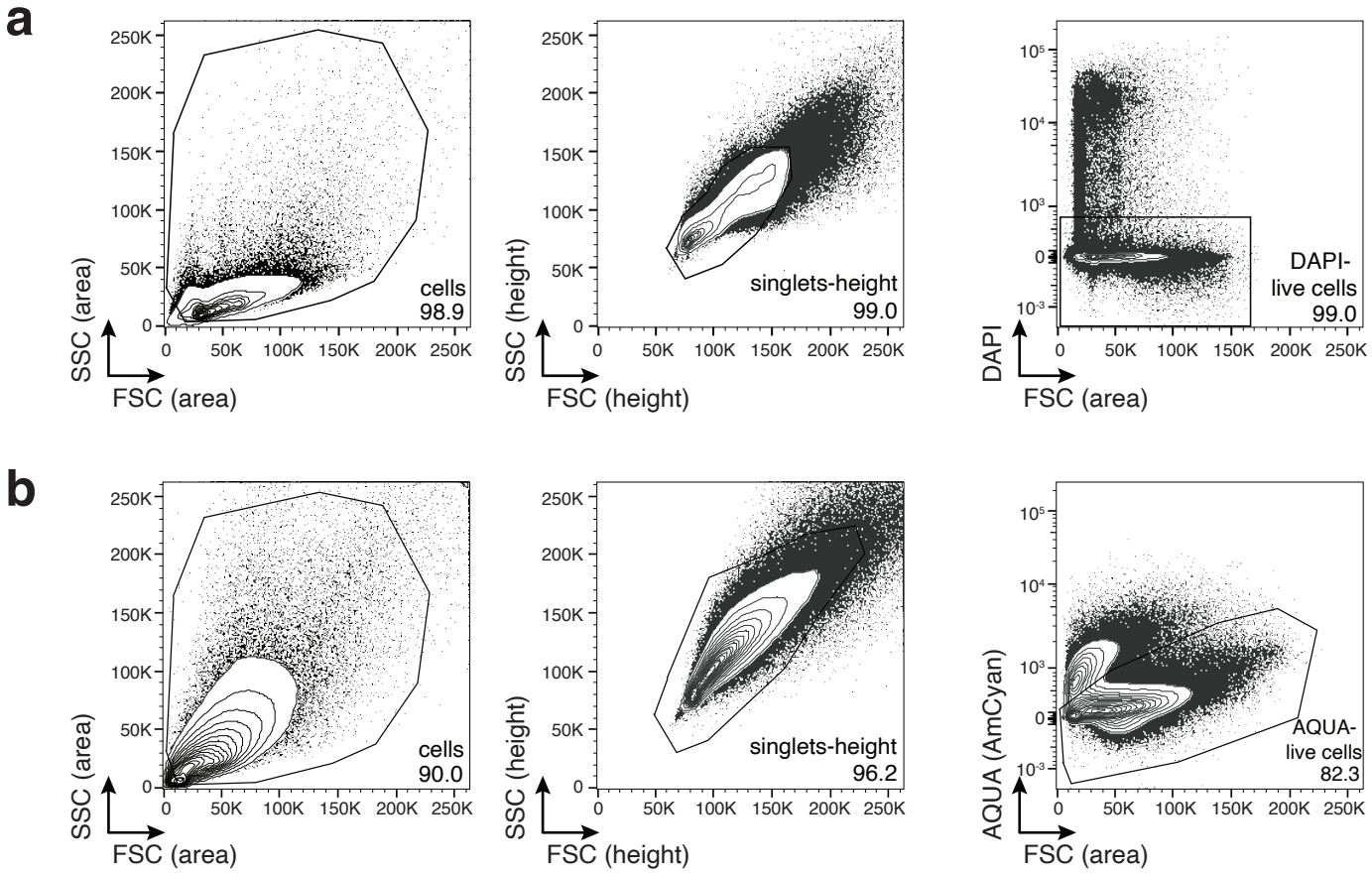
Supplementary Figure 11



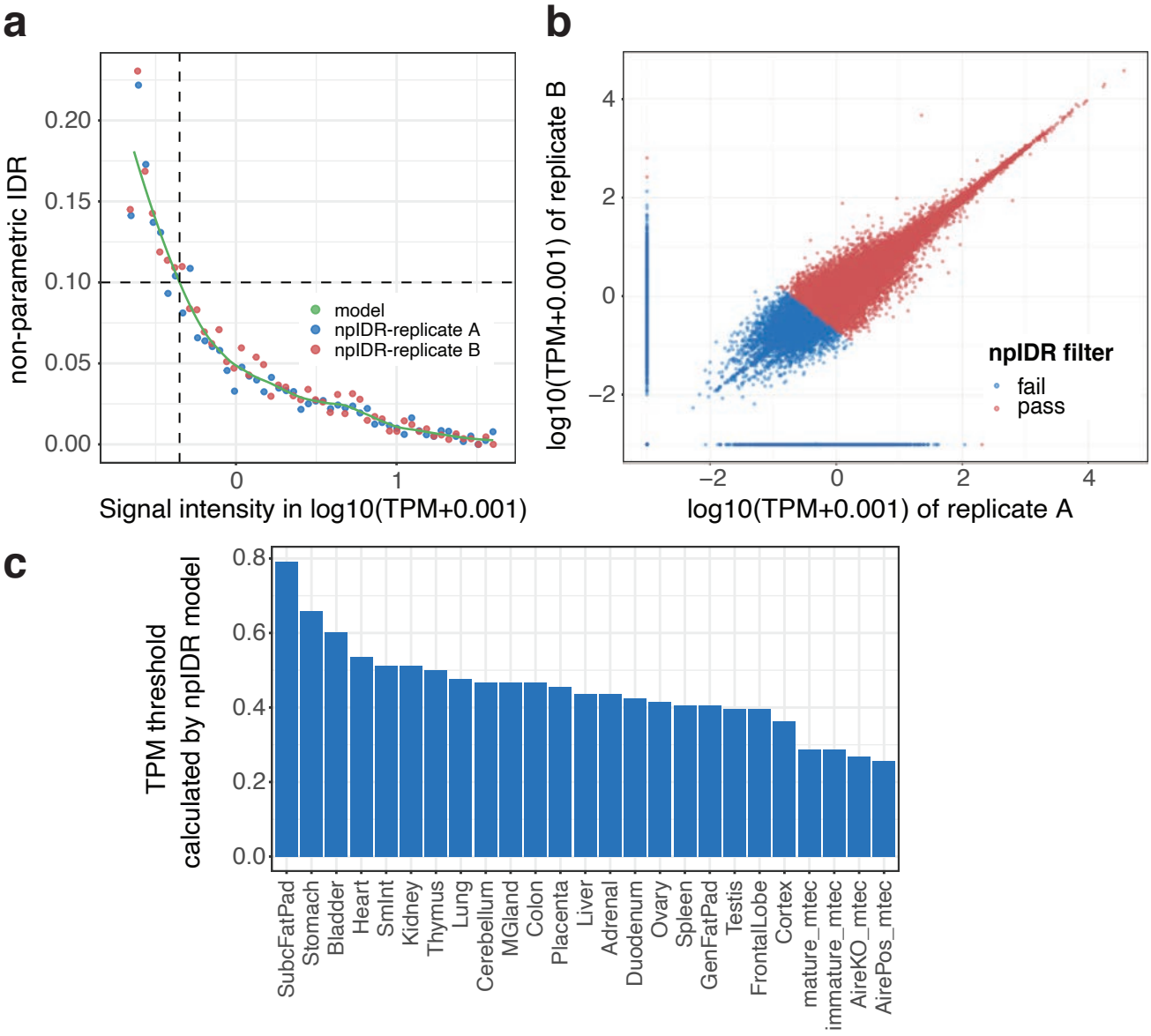
Supplementary Figure 12



Supplementary Figure 13



Supplementary Figure 14



Supplementary Figure 15