Bioinformatics, 31(7), 2015, 986–990
doi: 10.1093/bioinformatics/btu785
Advance Access Publication Date: 25 November 2014
Discovery Note

OXFORD

Gene expression

Promiscuous mRNA splicing under the control of AIRE in medullary thymic epithelial cells

Peter Keane[1], Rhodri Ceredig[2] and Cathal Seoighe[1,*]

[1]School of Mathematics, Statistics and Applied Mathematics and [2]Regenerative Medicine Institute, National University of Ireland, Galway, Ireland

*To whom correspondence should be addressed.

Associate Editor: Ivo Hofacker

Received on May 12, 2014; revised on August 26, 2014; accepted on November 20, 2014

Abstract

Motivation: The expression of tissue-restricted antigens (TRAs) in the thymus is required to ensure efficient negative selection of potentially auto-reactive T lymphocytes and avoid autoimmune disease. This promiscuous expression is under the control of the autoimmune regulator (AIRE), a transcription factor expressed in medullary thymic epithelial cells (mTECs). Tissue-specific alternative splicing may also produce TRAs but the extent to which splice isoforms that are restricted to specific tissues are expressed in mTECs is yet to be investigated.

Results: We reanalyzed microarray and RNA-Seq datasets from mouse mTECs and other epithelial and non-epithelial cell types and found that the diversity of splice isoforms in mTECs was greater than in any of the other cell types or tissues studied. We identified tissue-specific isoforms from a panel of mouse tissues and found several examples of such isoforms that are expressed in mTECs. The number of isoforms with restricted expression found in mTECs was significantly higher than for comparable cell types. Furthermore, we found evidence that AIRE influences the increased splicing diversity observed in mTECs as the genes for which tissue restricted isoforms are produced in mTECs were significantly more likely than other genes to be differentially spliced between AIRE knock-out and wild-type samples. Our results suggest that developing T lymphocytes are exposed to diverse tissue-restricted splice isoforms in the thymus and that AIRE has a direct or indirect role in this process, representing a novel aspect of its role in the maintenance of immune self-tolerance.

Contact: cathal.seoighe@nuigalway.ie

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

T lymphocytes are cells of the adaptive immune system that are capable of recognizing and binding foreign peptides presented on the surface of other cells by major histocompatibility (MHC) class molecules. Recognition is mediated by the T cell receptor (TCR), a surface bound receptor, generated by a stochastic process of gene-rearrangement during T cell development in the thymus. This somatic rearrangement results in a diverse repertoire of T cells, each expressing a single TCR that is capable of recognizing a limited set of peptide-MHC complexes. However, given the stochastic process by which these TCRs are generated, the possibility arises that a TCR may show an affinity for self-peptide-MHC complexes, leading to an autoimmune response. To avoid this, T cells undergo a process of selection in the thymus, whereby the developing T cell is exposed to a range of self-peptide-MHC complexes. This process, referred to as negative selection, results in the deletion from the T cell repertoire of cells with high affinity for self-peptide-MHC complexes (Klein *et al.*, 2009).

The range of self-peptides to which the developing T cells are exposed includes a number of peptides that are normally only expressed in specific tissues outside the thymus. This ectopic expression of tissue-restricted antigens (TRAs), which helps to prevent

autoimmune responses targeted against TRAs, is mediated by the autoimmune regulator (AIRE), a transcription factor expressed primarily in medullary thymic epithelial cells (mTECs; Anderson *et al.*, 2002). AIRE induces the expression of tissue restricted genes in mTECs, and mutations in *Aire* lead to autoimmune polyendocrinopathy candidiasis ectodermal dystrophy (APECED), an autoimmune disease which affects multiple endocrine organs (Peterson *et al.*, 2004). Although AIRE has been shown to increase the level of spliced versus un-spliced mRNAs (Zumer *et al.*, 2011), little is known about the impact of AIRE on alternative splicing, and the extent to which it can cause expression of alternative splice isoforms, resulting in isoform-specific TRA expression in the thymus. Previous work has shown that AIRE interacts with a number of proteins involved in pre-mRNA processing and spliceosome binding (Abramson *et al.*, 2010), including splicing factors such as SC35, an SR protein involved in transcriptional elongation through the recruitment of the positive transcription elongation factor b (p-TEFb; Lin *et al.*, 2008), and in the commitment of pre-mRNAs to the splicing pathway (Fu, 1993). AIRE can also interact directly with p-TEFb (Oven *et al.*, 2007). A mutation in the transcription activation domain (TAD) of AIRE, found in APECED patients, can prevent this interaction, leading to a decrease in the level of spliced versus un-spliced mRNAs (Abramson *et al.*, 2010; Zumer *et al.*, 2011). In order to ensure efficient negative selection, a diverse range of peptides needs to be presented in the thymus to developing T cells. This includes peptides that are introduced by alternative splicing (Klein *et al.*, 2000), specifically TRAs resulting from tissue-specific alternative splicing.

To investigate the diversity of alternative splicing in the thymus, we calculated splicing entropy (Ritchie *et al.*, 2008) using existing RNA-Seq data from mTECs and compared it with the splicing entropy of other tissues. Entropy can be considered as a measure of uncertainty, and splicing entropy reflects the diversity of observed splice isoforms in a given sample. We also compared mRNA splicing between wild-type and knock-out mTECs. Finally, we investigated the ectopic expression of tissue-specific splice isoforms in mTECs. We suggest that AIRE plays a role in promiscuous splicing in mTECs, resulting in the expression of TRAs in the thymus for the purpose of negative selection of T cells. This is likely to contribute to the role of AIRE in inducing tolerance to self-antigens.

2 Methods

2.1 Splicing entropy

Splicing entropy is a measure of the diversity of observed transcript isoforms in a given sample. Ritchie *et al.* (2008) used splicing entropy to measure splicing disorder using isoform expression values obtained from expressed sequence tag (EST) data. Here, we used isoform expression values calculated using RNA-Seq data. Splicing entropy was calculated for each gene G, as

$$-\sum_{i=1}^{g} P_i \log_2(P_i) \quad (1)$$

where g is the number of isoforms that make up G, and P_i is the proportion that each isoform i contributes to the overall expression of G. We calculated this for each sample independently and summarized each sample using the median splicing entropy value obtained.

2.2 Microarray data analysis

Affymetrix MoGene 1.0 ST microarray data from AIRE wild-type and knock-out mTECs, generated by Giraud *et al.* (2012), was downloaded from the Gene Expression Omnibus (GEO) under accession GSE33878. This microarray platform contains probesets targeting individual exons, allowing for exon level splicing analysis. Detected above background (DABG) probabilities and exon level expression estimates were calculated with Affymetrix Power Tools using the probe logarithmic intensity error (PLIER) algorithm with GC background correction. Unexpressed genes were removed from the dataset. A gene was considered expressed if over 50% of its probesets could be detected with a DABG P-value < 0.05 in over 50% of all samples. Genes with single probesets were discarded. We applied the alternative splicing robust prediction based on entropy method (ARH) (Rasche and Herwig, 2010), to detect differences in splicing between experimental conditions.

2.3 RNA-Seq data analysis

We downloaded RNA-Seq data generated by St-Pierre *et al.* (2013; GEO accession GSE44945). This data consisted of paired-end sequences from mTECs as well as from cortical thymic epithelial cells (cTECs) and skin epithelial cells (sECs). Raw reads were processed using Trimmomatic v0.30 (Bolger *et al.*, 2014) to remove low-quality sequences. These were then mapped to the mouse genome (mm9) with TopHat v1.4.1 (Trapnell *et al.*, 2009) using default parameters and RefSeq transcript expression levels were calculated with Cufflinks v1.3.0 (Trapnell *et al.*, 2010) using the following parameters (–multi-read-correct –min-isoform-fraction 0). Genes with only one annotated isoform were discarded from further analysis. Splicing entropy was then calculated for each gene. Bootstrap analysis of the splicing entropy distribution was conducted in R v3.0.1.

To compare splicing entropy across different tissues, we used the mouse 9-tissue dataset generated by Merkin *et al.* (2012; GEO accession GSE41637). These RNA-Seq datasets contain reads of varying lengths. For comparability across experiments, all reads were trimmed to a maximum length of 50 bp before mapping to the mouse genome with TopHat using the following parameters (–mate-inner-dist 200 –mate-std-dev 30). RefSeq transcript abundances were calculated with Cufflinks and splicing entropy was then calculated for each gene.

2.4 Tissue-specific alternative splicing

We used the nine-tissue dataset from Merkin *et al.* (2012) to identify tissue-specific splice isoforms. An isoform I, of a gene G, was considered to be specific to a tissue, if the proportion of transcripts from G corresponding to I was significantly >0.1 in all replicates of that tissue and in no replicate of any other tissue. Expression of an isoform, as a proportion of gene expression, was considered to be significantly greater if the lower bound of the 95% confidence interval for the isoform proportion exceeded 0.1. In order to avoid including tissue-specific genes, we considered only genes with fragments per kilobase of transcript per million mapped reads (FPKM) >0.5 in more than one of the tissues examined.

3 Results and discussion

3.1 Splicing entropy in mTECs compared with other cell types

The thymic epithelium creates a specialized environment that supports the development and maturation of T lymphocytes. This process involves interactions of developing lymphocytes with epithelial cells in the medulla and cortex of the thymus, called mTECs and cTECs, respectively. These cell types arise from a common bi-potent progenitor cell (Rossi *et al.*, 2006), but assume distinct roles in T cell

development (Anderson and Takahama, 2012). At a mature stage of development, mTECs express AIRE, a transcription factor that induces TRA expression in the thymic medulla (Anderson *et al.*, 2002). We hypothesized that there may exist a mechanism to express tissue-restricted splice isoforms in mTECs, to ensure negative selection of T cells that react to TRAs resulting from tissue-specific splicing. A prediction of this hypothesis is that thymic epithelial cells should express a diverse set of splice isoforms, including isoforms that are normally restricted to specific tissues. To investigate this prediction we calculated splicing entropy based on previously published RNA-Seq data (St-Pierre *et al.*, 2013) from mouse mTECs, cTECs and sECs, the latter being an example of a non-thymic epithelial cell type. Only genes that were expressed in all three cell-types and for which more than one annotated isoform was available were considered for the calculation of splicing entropy. We found that the median splicing entropy across genes was higher in the mTEC than in the cTEC and sEC samples ($P = 1 \times 10^{-36}$, $P = 6 \times 10^{-56}$, respectively, Wilcoxon signed-rank test; Fig. 1A and B). It should be noted that just one sample was available per cell type and the statistical significance reported here reflects consistency in the differences in splicing diversity across genes, rather than replication in multiple samples. Interestingly, the difference in splicing entropy was more apparent for AIRE-induced genes [as identified by Giraud *et al.* (2012)] than for AIRE-independent genes (Supplementary Table S1), which is suggestive of a role for AIRE in the diverse splicing output of mTECs.

We next compared thymic epithelial cells to a range of non-epithelial cell types using data from nine mouse tissues from Merkin *et al.* (2012). This dataset contains RNA-Seq samples from brain, colon, heart, kidney, liver, lung, skeletal muscle (SkM), spleen and testes. For each tissue, three samples (denoted mouse A, B and C) were available, with the exception of heart for which only two samples were available. Each sample originates from a different mouse strain (DBA/2J, C57BL/6 and CD1 for samples A, B and C, respectively). Differences in alternative splicing have been observed across mouse strains (Su *et al.*, 2008), which may contribute to some of the differences in splicing entropy observed across samples from the same tissue. The median (across genes) of the splicing entropy in mTECs was higher than for any of the tissues examined and for all samples, except for one of the mouse brain samples (Mouse_B_Brain) the difference was highly statistically significant ($P < 0.0001$ for all pair-wise comparisons, other than Mouse_B_Brain for which $P = 0.19$; Wilcoxon signed-rank test; Fig. 1C). Although based on a single available sample of mTECs these results suggest that mTECs express a wider range of isoforms compared with other cell types. This is consistent with the proposal that tissue-restricted splice isoforms are produced in the thymus and presented to T cells for the purpose of negative selection. The absence of a

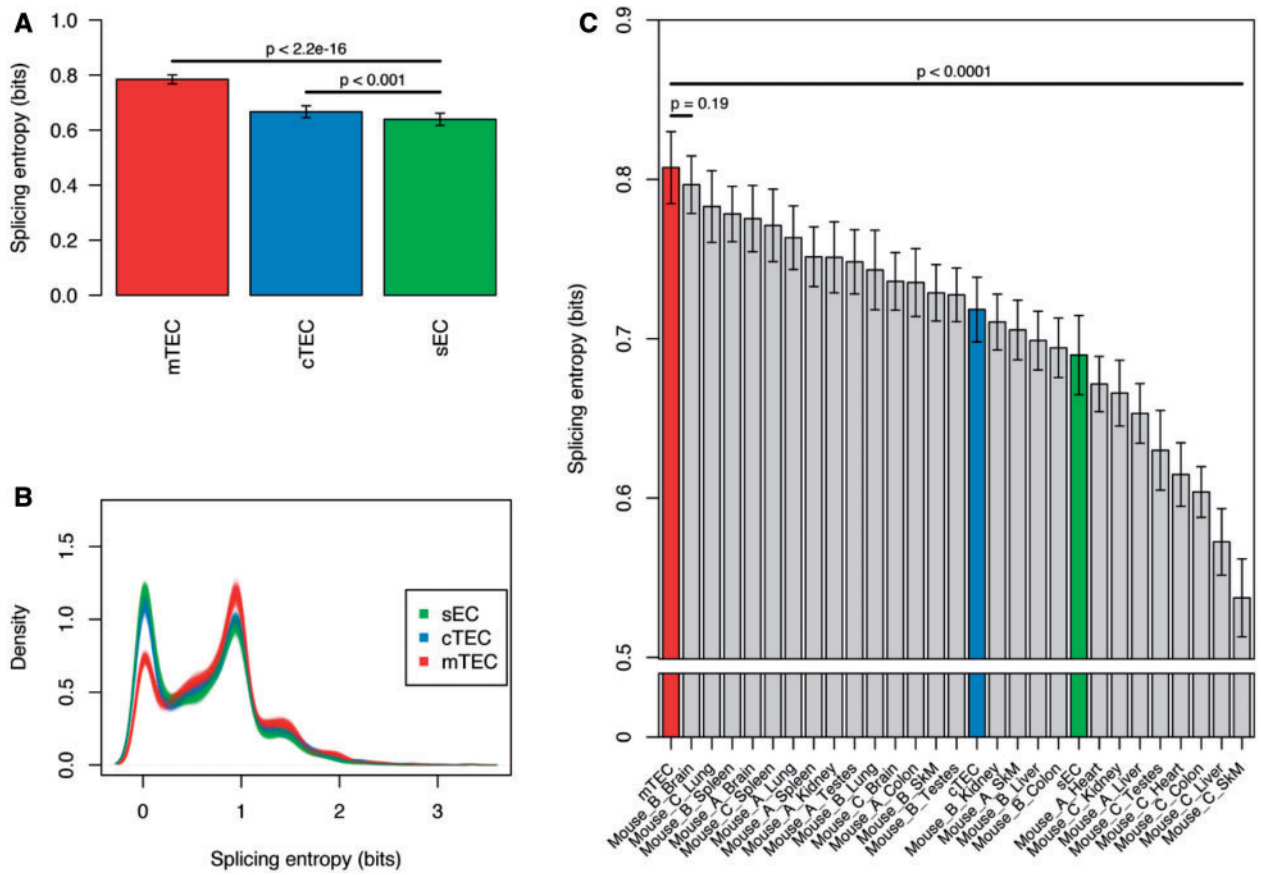

Fig. 1. (**A**) Median splicing entropy in mTECs (black), cTECs (blue) and sECs (green). Error bars correspond to the standard error. (**B**) Density plot of splicing entropy in mTECs, cTECs and sECs. We carried out 1000 bootstrap resamples of genes and plotted separate lines for each pseudoreplicate of the data. The results of the bootstrap procedure can be seen as a blurred thickening of the lines on the density plot. (**C**) Median splicing entropy across a range of cell types, including mTECs, cTECs and sEC as well as tissues from the nine mouse tissue dataset. The latter included three replicates per tissue, denoted Mouse A (DBA/2J mice), B (C57BL/6 mice) and C (CD1 mice), with the exception of heart samples that had only two replicates. Error bars denote standard error of the median. Only genes expressed across all epithelial samples are included in A and only genes expressed across all samples are included in B

Table 1. Summary of DABG analysis of microarray data

Sample	Mean number of exons detected per gene
AIRE KO rep 1	7.29
AIRE KO rep 2	7.30
AIRE KO rep 3	7.18
AIRE WT rep 1	7.95
AIRE WT rep 2	7.99
AIRE WT rep 3	8.59

particular isoform gives rise to the potential for T cells capable of directing auotimmity towards that isoform to escape the thymus, causing autoimmunity.

3.2 AIRE plays a role in alternative splicing in mTECs

To measure directly the impact of AIRE on alternative splicing, we applied the ARH method (Rasche and Herwig, 2010) to AIRE knock-out and wild-type mTECS, using microarray data obtained from Giraud *et al.* (2012). These data were generated using the Affymetrix MoGene 1.0 ST array platform, which contains probesets targeting individual exons. Genes that were either not expressed or to which only one probeset was mapped were removed prior to calculation of ARH values. We found that 4572 genes show differences in alternative splicing between AIRE knock-out and wild-type mTECs with an ARH P value < 0.05 (Supplementary Table S2). Interestingly, this list of genes was significantly enriched for genes with tissue-restricted isoforms that are expressed in mTECs ($P = 0.01$, from logistic regression, with gene expression level and the number of expressed isoforms as covariates). We also counted the number of exons detected in each sample. An exon was considered detected if it was assigned a DABG P-value < 0.05. The mean number of exons detected per expressed gene was significantly higher for the wild-type samples than for the AIRE knock-out samples ($P = 0.04$, two sided t-test; Table 1). Together, these results show that AIRE influences alternative splicing in mTECs, increasing the number of exons detected in AIRE wild-type compared with knock-out, consistent with a role for AIRE in the diverse splicing output of mTECs. Although AIRE has previously been shown to increase the level of spliced versus un-spliced mRNAs (Zumer *et al.*, 2011), a role for AIRE in alternative splicing has not been established. Here we show that AIRE plays a role in alternative splicing and increases the inclusion of exons, and therefore isoforms containing these exons and this is likely to be an aspect of the role of AIRE in negative selection.

3.3 Ectopic expression of tissue-specific isoforms in mTECs

Given that alternative splicing can give rise to tissue-specific isoforms (Xu *et al.*, 2002), we asked if ectopic expression of tissue-specific isoforms could be detected in mTECs. We identified a set of 339 tissue-specific isoforms using the nine mouse tissue dataset (Merkin *et al.*, 2012; see Supplementary Table S3 for the list of all detected TS isoforms). Of these, 72 were also expressed in mTECs, compared to 26 in cTECs and 21 in sECs ($P = 2 \times 10^{-7}$ and $P = 3 \times 10^{-9}$, respectively, for the comparison of the proportion of tissue-specific isoforms expressed in mTECs to the proportions in cTECs and sECs, respectively; two-proportion z-test; Fig. 2A; see Supplementary Table S4–S6 for TS isoforms detected in mTECs, cTECs and sECs, respectively). We found that 50 tissue-specific

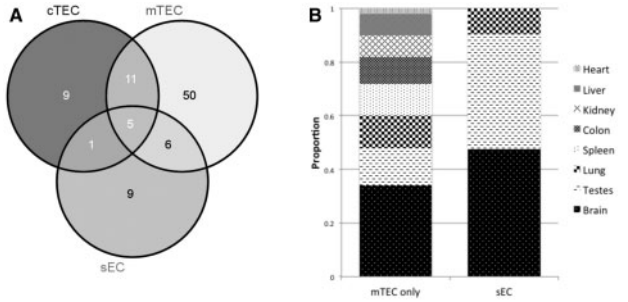

Fig. 2. (**A**) Number of tissue-specific isoforms detected in mTECs, cTECs and sECs. (**B**) Distribution of tissues represented by tissue-specific splice isoforms found in mTECs only (i.e. found in mTECs but not in cTECs or sECs), the second column includes all 21 tissue-specific isoforms found in sECs, whether or not they are expressed in mTECs or cTECs

isoforms were expressed in the mTEC sample and not in either of the cTEC or sEC samples. These represented tissue-specific splice isoforms from eight of the nine tissues examined. This is compared with three tissues represented in all tissue-specific isoforms detected in sECs (Fig. 2B). We found that the range of tissue-specific isoforms expressed in mTECs represents almost all of the tissues studied. This is consistent with previous findings that mTECs express a diverse range of tissue-restricted genes which represent most tissues of the body (Derbinski *et al.*, 2005). The tissue-specific isoforms that were found in mTECs and not in cTECs or sECs included the long form of the gene *Oxr1* (Fig. 3A), which is known to undergo brain-specific alternative splicing (Wakamatsu *et al.*, 2009). The *Oxr1* gene was expressed across all samples (Fig. 3B) but the long form of *Oxr1*, NM_001130166, was expressed only in brain and in mTECs (Fig. 3B and Supplementary Fig. S1). We also found that mTECs express a tissue-specific splice isoform of sperm associated antigen 16 (*Spag16*). An alternative splice isoform of *Spag16* has been identified as a B cell autoimmune target in multiple sclerosis (Bock *et al.*, 2014). It has been found that isoform-specific regions of autoantigens that undergo alternative splicing often encode MHC class I and II restricted epitopes (Ng *et al.*, 2004), and therefore *Spag16* may also represent a T cell target in autoimmune disease. The ectopic expression of tissue-specific isoforms is likely to contribute to negative selection, by providing an additional source of TRAs in the thymus. The importance of TRAs for this purpose is well established; however, previous studies have focused on TRAs derived from tissue-restricted genes (Derbinski *et al.*, 2001). Promiscuous splicing may be an important factor in establishing T cell tolerance to tissue-specific isoforms. Experimental validation would be required to formally show that T cells which are specific to isoform-specific peptides do escape the thymus and cause autoimmunity in the absence of AIRE and confirm our conclusion that AIRE introduces isoform-specific TRAs for the purpose of negative selection.

4 Conclusion

It is well established that AIRE induces the expression of tissue restricted genes in the thymus. Here, we show that AIRE also influences alternative splicing in mTECs, increasing the number of exons detectable per gene and allowing antigens on these exons to be produced. mTECs express a range of tissue-specific splice isoforms and knock-out of AIRE affects splicing of the genes for which tissue-restricted splice isoforms are found in mTECs. We suggest that promiscuous splicing in mTECs is under the control of AIRE, and that a

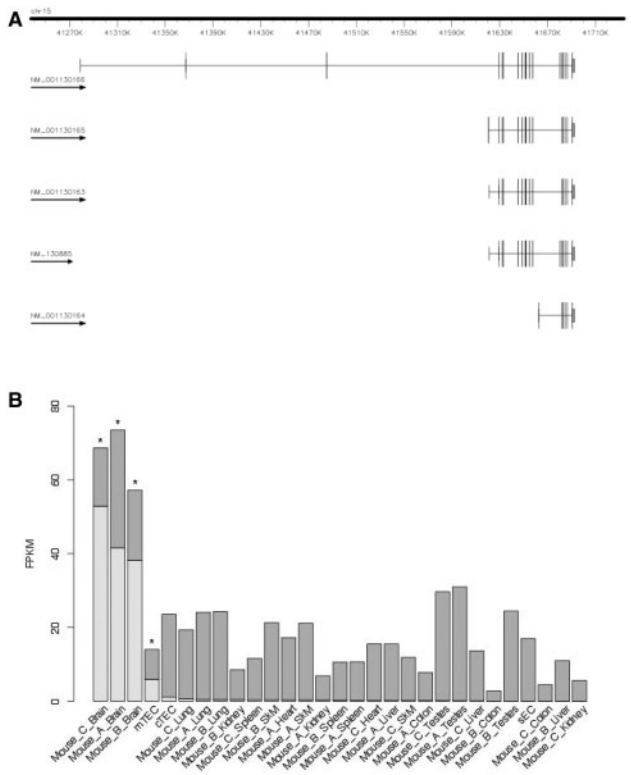

Fig. 3. (**A**) RefSeq annotated isoforms of the *Oxr1* gene, including the long form of the gene, NM_00130166. (**B**) FPKM values of the *Oxr1* gene in all tissues, and also in mTECs, cTECs and sECs. FPKM values of the long transcript, NM_001130166, are shown with a lighter coloured bar. The *Oxr1* gene is expressed across all samples. The isoform NM_01130166 is highly expressed in brain, and also in mTECs. Samples in which this isoform is significantly expressed are denoted by an asterix (see Methods)

breakdown in this mechanism would result in the absence of isoform-specific TRAs in the thymus, allowing T cells capable of directing autoimmunity towards tissue-specific isoforms to escape the thymus and contribute to autoimmine disease.

Funding

P.K. is funded under the Simulation Science (SimSci) PhD programme. SimSci is funded under the Programme for Research in Third Level Institutions and co-funded under the European Regional Development Fund. Rh.C. is funded by Science Foundation Ireland [grant number SFI09/SRC/B1794 and SFI07/SK/B1233b]. C.S. is funded by Science Foundation Ireland [grant number 07/SK/M1211b].

Conflict of Interest: none declared.

References

Abramson,J. *et al.* (2010) AIRE's partners in the molecular control of immunological tolerance. *Cell*, **140**, 123–135.

Anderson,G. and Takahama,Y. (2012) Thymic epithelial cells: working class heroes for T cell development and repertoire selection. *Trends Immunol.*, **33**, 256–263.

Anderson,M.S. *et al.* (2002) Projection of an immunological self shadow within the thymus by the AIRE protein. *Science*, **298**, 1395–1401.

Bock,L.D. *et al.* (2014) Sperm-associated antigen 16 is a novel target of the humoral autoimmune response in multiple sclerosis. *J. Immunol.*, **193**, 2147–2156.

Bolger,A.M. *et al.* (2014) Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*, **30**, 2114–2120.

Derbinski,J. *et al.* (2001) Promiscuous gene expression in medullary thymic epithelial cells mirrors the peripheral self. *Nat. Immunol.*, **2**, 1032–1039.

Derbinski,J. *et al.* (2005) Promiscuous gene expression in thymic epithelial cells is regulated at multiple levels. *J. Exp. Med.*, **202**, 33–45.

Fu,X.D. (1993) Specific commitment of different pre-mRNAs to splicing by single SR proteins. *Nature*, **365**, 82–85.

Giraud,M. *et al.* (2012) Aire unleashes stalled RNA polymerase to induce ectopic gene expression in thymic epithelial cells. *Proc. Natl Acad. Sci. USA*, **109**, 535–540.

Klein,L. *et al.* (2000) Shaping of the autoreactive t-cell repertoire by a splice variant of self protein expressed in thymic epithelial cells. *Nat. Med.*, **6**, 56–61.

Klein,L. *et al.* (2009) Antigen presentation in the thymus for positive selection and central tolerance induction. *Nat. Rev. Immunol.*, **9**, 833–844.

Lin,S. *et al.* (2008) The splicing factor SC35 has an active role in transcriptional elongation. *Nat. Struct. Mol. Biol.*, **15**, 819–826.

Merkin,J. *et al.* (2012) Evolutionary dynamics of gene and isoform regulation in mammalian tissues. *Science*, **338**, 1593–1599.

Ng,B. *et al.* (2004) Increased noncanonical splicing of autoantigen transcripts provides the structural basis for expression of untolerized epitopes. *J. Allergy Clin. Immunol.*, **114**, 1463–1470.

Oven,I. *et al.* (2007) AIRE recruits p-TEFb for transcriptional elongation of target genes in medullary thymic epithelial cells. *Mol. Cell. Biol.*, **27**, 8815–8823.

Peterson,P. *et al.* (2004) Autoimmune polyendocrinopathy candidiasis ectodermal dystrophy (APECED): a model disease to study molecular aspects of endocrine autoimmunity. *Clin. Exp. Immunol.*, **135**, 348–357.

Rasche,A. and Herwig,R. (2010) ARH: predicting splice variants from genome-wide data with modified entropy. *Bioinformatics*, **26**, 84–90.

Ritchie,W. *et al.* (2008) Entropy measures quantify global splicing disorders in cancer. *PLoS Comput. Biol.*, **4**, e1000011.

Rossi,S.W. *et al.* (2006) Clonal analysis reveals a common progenitor for thymic cortical and medullary epithelium. *Nature*, **441**, 988–991.

St-Pierre,C. *et al.* (2013) Transcriptome sequencing of neonatal thymic epithelial cells. *Sci. Rep.*, **3**, 1860.

Su,W-L. *et al.* (2008) Exon and junction microarrays detect widespread mouse strain- and sex-bias expression differences. *BMC Genomics*, **9**, 273.

Trapnell,C. *et al.* (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, **25**, 1105–1111.

Trapnell,C. *et al.* (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.*, **28**, 511–515.

Wakamatsu,A. *et al.* (2009) Identification and functional analyses of 11 769 full-length human cDNAs focused on alternative splicing. *DNA Res.*, **16**, 371–383.

Xu,Q. *et al.* (2002) Genome-wide detection of tissue-specific alternative splicing in the human transcriptome. *Nucleic Acids Res.*, **30**, 3754–3766.

Zumer,K. *et al.* (2011) Patient mutation in AIRE disrupts p-TEFb binding and target gene transcription. *Nucleic Acids Res.*, **39**, 7908–7919.