

# A transcriptomics-based investigation of the landscape and mechanisms of self-antigen splicing in thymic epithelial cells

Kathrin Jansen

Christ Church

University of Oxford

*A thesis submitted for the degree of  
Doctor of Philosophy*

Hilary 2019

## Abstract

Thymic epithelial cells (TEC) are remarkable for their ability to promiscuously express virtually the entire protein-coding gene repertoire to generate a molecular library of self-antigens for T cell education and selection. This process is essential for the establishment of central T cell tolerance. While the functional relevance of alternative splicing in TEC has been demonstrated for specific genes, the extent and fidelity of peripheral isoform representations of tissue-restricted genes in TEC is only poorly understood. We therefore subjected TEC to deep, stranded, paired-end sequencing and compared their transcriptome to that of peripheral tissues, using matching data from the mouse ENCODE project. In line with previous reports of unusually high splicing entropy in mTEC, I first demonstrated that these cells utilize a broader array of splice junctions than was observed for any of the peripheral tissues examined. Medullary mTEC further transcribed and spliced the highest fraction of known protein-coding gene isoforms. At the same time mTEC do not splice unusually high numbers of novel transcripts identified in a transcriptome assembly of peripheral mouse tissues and mTEC. Using long-read sequencing data from the Oxford Nanopore Technology I confirmed 97 % of highly-expressed novel transcripts expressed in mature mTEC. In this study, I found a limited role for *Aire* in regulating alternative splicing, but observed a role in modulating transcript usage. Investigating tissue-restricted splicing factor expression in mTEC revealed that a subset of peripheral tissue-restricted splicing factors was present in TEC, which might be regulating the unusual splicing complexity in these cells. However, a set of neuronal splicing factors was not detected in mTEC, which was in line with the absence of a set of neuronal microexons in mTEC. To further assess a potential mechanism of tissue-restricted splicing in TEC, I selected the tissue-restricted splicing factor *Rbfox1* for a TEC-targeted knockout mouse model. The conditional single-knockout model showed that *Rbfox1* expression was dispensable for the development and maturation of mTEC. This is to my knowledge the first study validating mTEC transcript structures with long-read sequencing data and investigating a conditional knockout of a tissue-restricted splicing factor in TEC.





# A transcriptomics-based investigation of the landscape and mechanisms of self-antigen splicing in thymic epithelial cells



Kathrin Jansen  
Christ Church  
University of Oxford

Supervisors: Dr. Stephen N. Sansom, Prof. Georg A. Holländer

A thesis submitted for the degree of

*Doctor of Philosophy*

Hilary 2019



# Abstract

Thymic epithelial cells (TEC) are remarkable for their ability to promiscuously express virtually the entire protein-coding gene repertoire to generate a molecular library of self-antigens for T cell education and selection. This process is essential for the establishment of central T cell tolerance. While the functional relevance of alternative splicing in TEC has been demonstrated for specific genes, the extent and fidelity of peripheral isoform representations of tissue-restricted genes in TEC is only poorly understood. We therefore subjected TEC to deep, stranded, paired-end sequencing and compared their transcriptome to that of peripheral tissues, using matching data from the mouse ENCODE project. In line with previous reports of unusually high splicing entropy in mTEC, I first demonstrated that these cells utilize a broader array of splice junctions than was observed for any of the peripheral tissues examined. Medullary mTEC further transcribed and spliced the highest fraction of known protein-coding gene isoforms. At the same time mTEC do not splice unusually high numbers of novel transcripts identified in a transcriptome assembly of peripheral mouse tissues and mTEC. Using long-read sequencing data from the Oxford Nanopore Technology I confirmed 97 % of highly-expressed novel transcripts expressed in mature mTEC. In this study, I found a limited role for *Aire* in regulating alternative splicing, but observed a role in modulating transcript usage. Investigating tissue-restricted splicing factor expression in mTEC revealed that a subset of peripheral tissue-restricted splicing factors was present in TEC, which might be regulating the unusual splicing complexity in these cells. However, a set of neuronal splicing factors was not detected in mTEC, which was in line with the absence of a set of neuronal microexons in mTEC. To further assess a potential mechanism of tissue-restricted splicing in TEC, I selected the tissue-restricted splicing factor *Rbfox1* for a TEC-targeted knockout mouse model. The conditional single-knockout model showed that *Rbfox1* expression was dispensable for the development and maturation of mTEC. This is to my knowledge the first study validating mTEC transcript structures with long-read sequencing data and investigating a conditional knockout of a tissue-restricted splicing factor in TEC.



# Declaration

This thesis is submitted for the degree of Doctor of Philosophy in Genomic Medicine and Statistics at the University of Oxford. No part of this thesis has been accepted or is currently being submitted for any degree, diploma, certificate or other qualification in this University or elsewhere. This thesis is my own work, except where indicated.

Kathrin Jansen  
(March 2019)



# Acknowledgements

First of all I would like to thank Steve and Georg for being excellent supervisors. Steve in particular for all the support in developing my computational skills and for sharing his expertise with me over the past 3.5 years. The suggestions and guidance for both the computational and experimental parts of my thesis were invaluable. I would also like to give a special thanks to Georg for sharing his knowledge of thymus biology, and for directions on experiments and mouse breeding.

Thanks to the all members of the Sansom group for providing such a stimulating environment and for sharing the frustration of debugging pipelines and R code. Furthermore, thanks for the wonderful sweets from all over the world!

Thanks to all present and past members of the Holländer group for sharing their experiences with me. A big thank you goes to Mary and Stefano for all their support. They have always been on hand when issues with mice, the FACS machine and other minor lab panics occurred.

Moreover, special thanks to the BMS staff for their support in maintaining my colony. Thanks to Maria Lolopolo and Dr. David Buck for sharing their expertise with long-read sequencing and for their assistance in generating the data together with Moustafa Attar. A special thanks to Prof. Douglas Black for providing the RRM antibody and the breeding pair of mice. Parts of this project would not have been possible without these generous gifts.

I would also like to thank my friends from Christ Church, the GMS program and abroad who have been a continuous support throughout my PhD. A thanks also goes to my various rowing crews in Christ Church Boat Club but also in Falcon for the enjoyable breaks from work.

A special thanks to my family. Even though you have been in a different country, you have always been a great support and encouraged me regularly with motivational words! Finally, a special thanks to Michael for the continuous encouragement, supportive words, and always having my back. None of this work would have been possible without the time, encouragement and labour of all the people mentioned and I am very grateful for your support.





# Contents

|                                                                                                                            |            |
|----------------------------------------------------------------------------------------------------------------------------|------------|
| <b>List of Figures</b>                                                                                                     | <b>xv</b>  |
| <b>List of Tables</b>                                                                                                      | <b>xix</b> |
| <b>List of Abbreviations</b>                                                                                               | <b>xxi</b> |
| <b>1 Introduction</b>                                                                                                      | <b>1</b>   |
| 1.1 Adaptive immune system . . . . .                                                                                       | 2          |
| 1.2 Central tolerance in the thymus . . . . .                                                                              | 3          |
| 1.2.1 Development of immature thymocytes into single-positive, naïve T cells . . . . .                                     | 5          |
| 1.2.2 Maturation and function of thymic epithelial cells . . . . .                                                         | 6          |
| 1.2.3 The Autoimmune Regulator, Aire, and its role in the thymus . . .                                                     | 8          |
| 1.3 Alternative splicing . . . . .                                                                                         | 12         |
| 1.3.1 Alternative splicing in development and disease . . . . .                                                            | 13         |
| 1.3.2 Cross-species conservation of alternative splicing . . . . .                                                         | 14         |
| 1.3.3 Regulation of alternative splicing . . . . .                                                                         | 16         |
| 1.3.4 Long-read RNA-sequencing methods to assess splicing . . . . .                                                        | 18         |
| 1.3.5 Effects of alternative splicing on proteome complexity . . . . .                                                     | 20         |
| 1.4 Alternative splicing in the thymus and its role in the induction of self-tolerance                                     | 22         |
| 1.4.1 Complexity of the transcriptome and possible underlying mechanisms of splicing in thymic epithelial cells . . . . .  | 22         |
| 1.4.2 The effect of thymic alternative splicing on self-tolerance . . . . .                                                | 24         |
| 1.5 Open questions and hypotheses . . . . .                                                                                | 28         |
| 1.6 Aims . . . . .                                                                                                         | 29         |
| <b>2 Methods</b>                                                                                                           | <b>31</b>  |
| 2.1 Computational methods . . . . .                                                                                        | 32         |
| 2.1.1 Sources of Illumina population RNA-sequencing datasets . . . . .                                                     | 32         |
| 2.1.2 Processing of Illumina RNA-sequencing datasets from ENCODE mouse tissues and mouse thymic epithelial cells . . . . . | 33         |
| 2.1.3 mT&T assembly pipeline . . . . .                                                                                     | 33         |
| 2.1.3.1 Comparison of tissue assemblies . . . . .                                                                          | 35         |
| 2.1.3.2 Multi-gene transcripts in the mT&T assembly . . . . .                                                              | 35         |

|         |                                                                                                              |    |
|---------|--------------------------------------------------------------------------------------------------------------|----|
| 2.1.3.3 | Discovery of novel transcripts from the mT&T assembly                                                        | 36 |
| 2.1.4   | Processing of RNA-sequencing data from thymic epithelial cells with a loss of <i>Rbfox1</i> or <i>Rbfox2</i> | 37 |
| 2.1.5   | Quantification and differential expression analysis at the transcript and gene level                         | 37 |
| 2.1.5.1 | Gene and transcript expression quantification                                                                | 37 |
| 2.1.5.2 | Differential gene expression including definition of <i>Aire</i> -regulated genes                            | 38 |
| 2.1.5.3 | Differential transcript expression including definition of <i>Aire</i> -regulated transcripts                | 38 |
| 2.1.6   | Splicing tools                                                                                               | 40 |
| 2.1.6.1 | Splice junction estimation                                                                                   | 40 |
| 2.1.6.2 | <i>SUPPA</i> for annotation of splice event types and quantification of percent spliced-in values            | 40 |
| 2.1.6.3 | Identification of differential splicing events using <i>rMATS</i>                                            | 41 |
| 2.1.7   | Single-cell RNA-sequencing                                                                                   | 41 |
| 2.1.7.1 | Processing and post-mapping quality assessment of single-cell RNA-sequencing data                            | 42 |
| 2.1.7.2 | Gene expression quantification, normalisation and clustering of single-cell RNA-sequencing data              | 42 |
| 2.1.7.3 | Assessment of differential splicing in single-cell <i>in silico</i> pools using <i>MISO</i>                  | 43 |
| 2.1.8   | Long-read RNA-sequencing using Oxford Nanopore Technology                                                    | 44 |
| 2.1.8.1 | Quality assessment and quantification of long-read sequencing data                                           | 44 |
| 2.1.8.2 | Transcriptome assembly from long-read sequencing data                                                        | 44 |
| 2.1.9   | Statistical analysis and visualisation of processed data                                                     | 46 |
| 2.1.9.1 | Tissue grouping based on transcript expression                                                               | 46 |
| 2.1.9.2 | Geneset over-enrichment testing                                                                              | 47 |
| 2.1.9.3 | Definition of tissue-restricted genes                                                                        | 47 |
| 2.2     | Experimental methods                                                                                         | 49 |
| 2.2.1   | Mice                                                                                                         | 49 |
| 2.2.1.1 | C57BL/6                                                                                                      | 49 |
| 2.2.1.2 | FoxN1-cre                                                                                                    | 49 |
| 2.2.1.3 | <i>Rbfox1</i> <sup>lox/lox</sup> ; <i>Rbfox2</i> <sup>lox/lox</sup>                                          | 49 |
| 2.2.2   | Genotyping and assessment of genomic excision in conditional knockout mice                                   | 50 |
| 2.2.3   | Isolation of thymic epithelial cells                                                                         | 51 |
| 2.2.3.1 | Thymus dissection and digestion                                                                              | 51 |
| 2.2.3.2 | Depletion of CD45 <sup>+</sup> cells                                                                         | 52 |
| 2.2.4   | Flow cytometry                                                                                               | 52 |

|          |                                                                                                                                              |           |
|----------|----------------------------------------------------------------------------------------------------------------------------------------------|-----------|
| 2.2.4.1  | Phenotyping analysis and cell sorting using flow cytometry                                                                                   | 53        |
| 2.2.4.2  | Analysis of flow cytometry data . . . . .                                                                                                    | 53        |
| 2.2.5    | RNA-sequencing . . . . .                                                                                                                     | 55        |
| 2.2.5.1  | Cell sorting and RNA extraction for transcriptomics . .                                                                                      | 55        |
| 2.2.5.2  | Oxford Nanopore Technologies RNA-sequencing . . . . .                                                                                        | 55        |
| 2.2.5.3  | Stranded RNA-sequencing from conditional knockout animals . . . . .                                                                          | 55        |
| 2.2.6    | Immunohistochemistry . . . . .                                                                                                               | 55        |
| 2.2.6.1  | Preparation of tissue sections . . . . .                                                                                                     | 55        |
| 2.2.6.2  | Immunofluorescence staining of thymus sections . . . . .                                                                                     | 56        |
| 2.2.6.3  | Microscopy and data analysis . . . . .                                                                                                       | 56        |
| <b>3</b> | <b>Generation and validation of a reference transcriptome to assess splicing in thymic epithelial cells</b>                                  | <b>57</b> |
| 3.1      | Background . . . . .                                                                                                                         | 58        |
| 3.2      | Strategy to generate the mT&T assembly . . . . .                                                                                             | 60        |
| 3.2.1    | Construction of assemblies from individual mouse tissues . . . . .                                                                           | 62        |
| 3.2.2    | Merging of tissue assemblies and comparison to the reference annotation . . . . .                                                            | 66        |
| 3.2.3    | Assessment of mT&T assembly quality . . . . .                                                                                                | 66        |
| 3.2.3.1  | Comparison of the individual tissue assemblies and of the mT&T assembly with the reference annotation . . . . .                              | 67        |
| 3.2.3.2  | Gene and transcript detection using the mT&T assembly                                                                                        | 69        |
| 3.2.3.3  | Assessment of multi-gene transcripts in the mT&T assembly and the reference annotation using exon networks . .                               | 69        |
| 3.3      | Validation of structures using low-input long-read RNA-sequencing of medullary thymic epithelial cells . . . . .                             | 75        |
| 3.3.1    | Quality assessment of long-read RNA-sequencing data . . . . .                                                                                | 75        |
| 3.3.2    | Building a transcriptome assembly from long-read RNA-sequencing data . . . . .                                                               | 77        |
| 3.3.2.1  | Assessment of reference-guided assembly revealed frequent UTR extensions in novel transcripts . . . . .                                      | 78        |
| 3.3.2.2  | Identification of novel transcripts with potential novel epitopes in an assembly from pre-filtered ONT reads with Illumina support . . . . . | 81        |
| 3.3.3    | Validation of novel transcripts between short and long-read sequencing technologies . . . . .                                                | 85        |
| 3.4      | Discussion . . . . .                                                                                                                         | 87        |

|          |                                                                                                                          |            |
|----------|--------------------------------------------------------------------------------------------------------------------------|------------|
| <b>4</b> | <b>Description of the splicing landscape in thymic epithelial cells in comparison to peripheral tissues</b>              | <b>91</b>  |
| 4.1      | Background . . . . .                                                                                                     | 92         |
| 4.2      | Assessment of transcriptome complexity in TEC . . . . .                                                                  | 94         |
| 4.2.1    | High number of unique splice junctions detected in mature mTEC                                                           | 94         |
| 4.2.2    | Highest fraction of isoforms per protein-coding gene produced by mature mTEC . . . . .                                   | 95         |
| 4.2.3    | More extensive annotated splicing in TEC compared to peripheral tissues . . . . .                                        | 97         |
| 4.2.4    | Description of annotated known splicing events in TEC at the population and single-cell level . . . . .                  | 98         |
| 4.3      | Analysis of peripheral representation of splicing in TEC . . . . .                                                       | 102        |
| 4.3.1    | Highest fraction of tissue-restricted transcripts detected in TEC compared to the testis and cerebellum . . . . .        | 102        |
| 4.3.2    | Transcripts specific to tissues with frequent autoimmunity in APS-1 patients underrepresented in TEC . . . . .           | 103        |
| 4.3.3    | Incomplete representation of neuronal-specific microexons in mTEC                                                        | 105        |
| 4.4      | Characterisation of novel transcripts in TEC . . . . .                                                                   | 107        |
| 4.4.1    | Description of novel transcript examples in TEC . . . . .                                                                | 109        |
| 4.4.2    | Prediction of the impact on protein structure for TEC-specific novel transcripts . . . . .                               | 111        |
| 4.5      | Discussion . . . . .                                                                                                     | 113        |
| <b>5</b> | <b>A computational investigation of possible mechanisms of alternative splicing in thymic epithelial cells</b>           | <b>117</b> |
| 5.1      | Background . . . . .                                                                                                     | 118        |
| 5.2      | Assessment of the influence of AIRE and maturity on splicing in TEC . .                                                  | 121        |
| 5.2.1    | A limited role for AIRE in splicing in mature mTEC . . . . .                                                             | 121        |
| 5.2.2    | Decreased frequency of retained intron events upon TEC maturation                                                        | 122        |
| 5.2.3    | Validation of retained introns and discovery of novel retained introns using long-read sequencing . . . . .              | 124        |
| 5.2.4    | Regulation of transcript expression by AIRE . . . . .                                                                    | 128        |
| 5.3      | Definition of tissue-restricted splicing factors and their expression across mouse tissues and TEC . . . . .             | 131        |
| 5.3.1    | Distinct splicing factor expression in TEC and brain compared to other peripheral tissues . . . . .                      | 131        |
| 5.3.2    | Expression of a subset of peripheral tissue-restricted splicing factors in TEC . . . . .                                 | 134        |
| 5.4      | Prioritisation of a candidate for functional follow-up . . . . .                                                         | 139        |
| 5.4.1    | Differentially expressed splicing factors between TEC populations                                                        | 139        |
| 5.4.2    | Identification of tissue-restricted splicing factors with frequent and robust expression in single mature mTEC . . . . . | 141        |

|          |                                                                                                                                                                   |            |
|----------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------|
| 5.4.3    | Assessment of splicing in single-cell pools stratified by splicing factor expression . . . . .                                                                    | 146        |
| 5.4.3.1  | Comparison of pools stratified by selected tissue-restricted splicing factors with random pools . . . . .                                                         | 146        |
| 5.4.3.2  | Comparison of pools stratified by expression of tissue-restricted splicing factors or tissue-restricted genes . . . .                                             | 148        |
| 5.4.4    | <i>Rbfox1</i> as a candidate splicing factor for <i>in vivo</i> follow-up . . . .                                                                                 | 149        |
| 5.5      | Discussion . . . . .                                                                                                                                              | 151        |
| <b>6</b> | <b><i>In vivo</i> function of <i>Rbfox1</i> and <i>Rbfox2</i> in thymic epithelial cells</b>                                                                      | <b>157</b> |
| 6.1      | Background . . . . .                                                                                                                                              | 158        |
| 6.2      | Validation of <i>Rbfox1</i> and <i>Rbfox2</i> as candidates for TEC-targeted knockout mouse models . . . . .                                                      | 160        |
| 6.2.1    | Validation of proteins from the Rbfox family in TEC . . . . .                                                                                                     | 161        |
| 6.2.2    | Assessment of <i>Rbfox1</i> transcript structures in TEC . . . . .                                                                                                | 162        |
| 6.2.3    | Summary of available mouse models for loxP-cre system . . . . .                                                                                                   | 165        |
| 6.2.4    | Targeted <i>Rbfox1</i> <sup>lox/lox</sup> : <i>Rbfox2</i> <sup>lox/lox</sup> mice showed no difference in thymus architecture compared to wildtype mice . . . . . | 167        |
| 6.3      | TEC-targeted loss of <i>Rbfox1</i> expression . . . . .                                                                                                           | 170        |
| 6.3.1    | Normal thymus architecture in mice lacking <i>Rbfox1</i> expression in TEC . . . . .                                                                              | 170        |
| 6.3.2    | <i>Rbfox1</i> expression in TEC is dispensable for TEC maturation and the expression of AIRE . . . . .                                                            | 170        |
| 6.4      | TEC-targeted loss of <i>Rbfox2</i> expression . . . . .                                                                                                           | 174        |
| 6.4.1    | Unaffected thymus cellularity in TEC with a loss of <i>Rbfox2</i> expression                                                                                      | 174        |
| 6.4.2    | Reduced number of mTEC upon loss of <i>Rbfox2</i> expression . . . .                                                                                              | 176        |
| 6.5      | Transcriptome analysis of mTEC with a loss of <i>Rbfox1</i> or <i>Rbfox2</i> . . . .                                                                              | 180        |
| 6.5.1    | Small number of differentially expressed genes upon loss of <i>Rbfox1</i> and <i>Rbfox2</i> in mature mTEC . . . . .                                              | 184        |
| 6.5.2    | Assessing the effects of <i>Rbfox1</i> or <i>Rbfox2</i> loss on splicing in mature mTEC . . . . .                                                                 | 188        |
| 6.6      | Breeding of the conditional double-knockout mice . . . . .                                                                                                        | 196        |
| 6.7      | Discussion . . . . .                                                                                                                                              | 196        |
| <b>7</b> | <b>General Discussion</b>                                                                                                                                         | <b>203</b> |
|          | <b>Bibliography</b>                                                                                                                                               | <b>215</b> |
|          | <b>Appendices</b>                                                                                                                                                 |            |
| <b>A</b> | <b>Appendix - Computational tools &amp; Experimental materials</b>                                                                                                | <b>253</b> |
| <b>B</b> | <b>Appendix - Chapter 3</b>                                                                                                                                       | <b>261</b> |

|          |                             |            |
|----------|-----------------------------|------------|
| <b>C</b> | <b>Appendix - Chapter 4</b> | <b>273</b> |
| <b>D</b> | <b>Appendix - Chapter 5</b> | <b>279</b> |
| <b>E</b> | <b>Appendix - Chapter 6</b> | <b>295</b> |

# List of Figures

|      |                                                                                                                  |     |
|------|------------------------------------------------------------------------------------------------------------------|-----|
| 1.1  | Schematic overview of the thymus. . . . .                                                                        | 4   |
| 1.2  | Stages of mTEC maturation. . . . .                                                                               | 7   |
| 1.3  | Alternative splicing event types. . . . .                                                                        | 13  |
| 1.4  | Activity of splicing factors depends on binding position relative to the intron. . . . .                         | 17  |
| 1.5  | Scheme of an example where short-read sequencing was insufficient to resolve transcript structures. . . . .      | 19  |
| 1.6  | Schematic summary of responses to induction of EAE in different mouse strains. . . . .                           | 25  |
| 3.1  | Schematic overview of the pipeline to generate the mT&T assembly. . . . .                                        | 61  |
| 3.2  | Results of reproducibility analysis using non-parametric IDR. . . . .                                            | 64  |
| 3.3  | Statistics on npIDR calculation. . . . .                                                                         | 65  |
| 3.4  | Assembly statistics for the mT&T assembly. . . . .                                                               | 68  |
| 3.5  | Detection of genes and transcripts using the mT&T assembly. . . . .                                              | 70  |
| 3.6  | Graph-based analysis to assess multi-gene transcripts. . . . .                                                   | 72  |
| 3.7  | Long-read sequencing of mTEC using Oxford Nanopore Technology. . . . .                                           | 76  |
| 3.8  | Stand-alone assembly of ONT sequencing reads. . . . .                                                            | 80  |
| 3.9  | Outline to generate ONT assembly from filtered reads. . . . .                                                    | 82  |
| 3.10 | Results of the ONT assembly from filtered reads. . . . .                                                         | 83  |
| 3.11 | Examples of novel transcripts from ONT assembly. . . . .                                                         | 84  |
| 3.12 | Validation of novel transcripts between technologies. . . . .                                                    | 86  |
| 4.1  | Splicing complexity in mTEC and peripheral tissues based on splice junctions. . . . .                            | 95  |
| 4.2  | Splicing complexity in mTEC and peripheral tissues based on the fraction of isoforms per expressed gene. . . . . | 96  |
| 4.3  | Annotated splice variants in mTEC and peripheral tissues. . . . .                                                | 98  |
| 4.4  | Examples of known tissue-specific alternative splicing events. . . . .                                           | 101 |
| 4.5  | Representation of tissue-restricted isoforms in mTEC. . . . .                                                    | 104 |
| 4.6  | Heatmap of the percentage spliced-in (PSI) values for exons in protein-coding genes. . . . .                     | 106 |
| 4.7  | Assessment of novel transcripts in the mT&T assembly. . . . .                                                    | 108 |
| 4.8  | Examples of novel TEC-specific transcripts. . . . .                                                              | 110 |
| 4.9  | Peptide prediction for TEC-specific novel transcripts. . . . .                                                   | 112 |
| 5.1  | Differential splicing analysis using <i>rMATS</i> . . . . .                                                      | 123 |

|      |                                                                                                                                                                                                                |     |
|------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| 5.2  | Validation of a retained intron in a known and a mT&T novel transcript using long-read sequencing data. . . . .                                                                                                | 126 |
| 5.3  | ONT novel transcripts with retained introns in mature mTEC. . . . .                                                                                                                                            | 127 |
| 5.4  | Definition of <i>Aire</i> -regulated transcripts. . . . .                                                                                                                                                      | 129 |
| 5.5  | Identification of a tissue-restricted splicing factor set. . . . .                                                                                                                                             | 131 |
| 5.6  | Principal component analysis using all splicing factors across mTEC and peripheral tissues . . . . .                                                                                                           | 133 |
| 5.7  | t-SNE analysis of splicing factor expression profile across mTEC and peripheral tissues. . . . .                                                                                                               | 134 |
| 5.8  | Heatmap of 50 most highly variable tissue-restricted splicing factors. . . .                                                                                                                                   | 136 |
| 5.9  | Heatmap of selected tissue-restricted splicing factors with a described role in the literature. . . . .                                                                                                        | 138 |
| 5.10 | Differential expression analysis of splicing factors between different TEC types. . . . .                                                                                                                      | 140 |
| 5.11 | Principal component analysis, t-SNE and clustering of single mature mTEC. .                                                                                                                                    | 143 |
| 5.12 | Heatmap of identified cluster markers using <i>Seurat</i> . . . . .                                                                                                                                            | 144 |
| 5.13 | Splicing factor expression in single mature mTEC. . . . .                                                                                                                                                      | 145 |
| 5.14 | Differential splicing events in <i>in silico</i> single-cell pools. . . . .                                                                                                                                    | 147 |
| 5.15 | Heatmap to prioritise tissue-restricted splicing factors for functional follow-up study. . . . .                                                                                                               | 150 |
| 6.1  | Expression of <i>Rbfox</i> family members at the transcript and protein level. .                                                                                                                               | 162 |
| 6.2  | Flow cytometric analysis of RBFOX family expression in wildtype mTEC. .                                                                                                                                        | 163 |
| 6.3  | Sashimi plot for <i>Rbfox1</i> in heart, cerebellum and mature mTEC. . . . .                                                                                                                                   | 164 |
| 6.4  | Information on <i>Rbfox1</i> - and <i>Rbfox2</i> -targeted mouse models. . . . .                                                                                                                               | 166 |
| 6.5  | Comparative analysis of the thymus cellularity and the TEC compartment of wildtype and <i>Rbfox1</i> <sup>lox/lox</sup> ; <i>Rbfox2</i> <sup>lox/lox</sup> mice. . . . .                                       | 168 |
| 6.6  | Analysis of mTEC subpopulations and AIRE expression between wildtype animals and <i>Rbfox1</i> <sup>lox/lox</sup> ; <i>Rbfox2</i> <sup>lox/lox</sup> mice. . . . .                                             | 169 |
| 6.7  | Assessment of thymus anatomy and cellularity in <i>Rbfox1</i> <sup>lox/lox</sup> ; <i>cre</i> <sup>-/-</sup> and <i>Rbfox1</i> <sup>lox/lox</sup> ; <i>cre</i> <sup>+/-</sup> . . . . .                        | 171 |
| 6.8  | Phenotyping of thymic epithelial cells from thymi of <i>Rbfox1</i> <sup>lox/lox</sup> ; <i>cre</i> <sup>-/-</sup> and <i>Rbfox1</i> <sup>lox/lox</sup> ; <i>cre</i> <sup>+/-</sup> mice. . . . .               | 172 |
| 6.9  | Analysis of TEC subpopulations and AIRE expression across these in <i>Rbfox1</i> <sup>lox/lox</sup> ; <i>cre</i> <sup>-/-</sup> and <i>Rbfox1</i> <sup>lox/lox</sup> ; <i>cre</i> <sup>+/-</sup> mice. . . . . | 173 |
| 6.10 | Assessment of thymus anatomy and cellularity in <i>Rbfox2</i> <sup>lox/lox</sup> ; <i>cre</i> <sup>-/-</sup> and <i>Rbfox2</i> <sup>lox/lox</sup> ; <i>cre</i> <sup>+/-</sup> . . . . .                        | 175 |
| 6.11 | Phenotyping of thymic epithelial cells from thymi of <i>Rbfox2</i> <sup>lox/lox</sup> ; <i>cre</i> <sup>-/-</sup> compared to <i>Rbfox2</i> <sup>lox/lox</sup> ; <i>cre</i> <sup>+/-</sup> mice. . . . .       | 177 |
| 6.12 | Analysis of thymic epithelial cell subtypes in thymi from <i>Rbfox2</i> <sup>lox/lox</sup> ; <i>cre</i> <sup>-/-</sup> compared to <i>Rbfox2</i> <sup>lox/lox</sup> ; <i>cre</i> <sup>+/-</sup> . . . . .      | 178 |



|      |                                                                                                                                                                                                                                                                          |     |
|------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| 6.13 | Analysis of TEC subpopulations and AIRE expression across these in Rbfox2 <sup>lox/lox</sup> :cre <sup>-/-</sup> and Rbfox2 <sup>lox/lox</sup> :cre <sup>+/-</sup> mice. . . . .                                                                                         | 179 |
| 6.14 | Quality assessment of RNA-sequencing data for immature and mature mTEC from Rbfox1 <sup>lox/lox</sup> :cre <sup>-/-</sup> and Rbfox1 <sup>lox/lox</sup> :cre <sup>+/-</sup> mice. . . . .                                                                                | 181 |
| 6.15 | Quality assessment of RNA-sequencing data for immature and mature mTEC from Rbfox2 <sup>lox/lox</sup> :cre <sup>-/-</sup> and Rbfox2 <sup>lox/lox</sup> :cre <sup>+/-</sup> mice. . . . .                                                                                | 183 |
| 6.16 | Principal Component Analysis using gene expression data for immature and mature mTEC from both single-knockout models. . . . .                                                                                                                                           | 184 |
| 6.17 | Differentially expressed genes between Rbfox1 <sup>lox/lox</sup> :cre <sup>+/-</sup> or Rbfox2 <sup>lox/lox</sup> :cre <sup>+/-</sup> samples and the corresponding cre <sup>-/-</sup> mice. . . . .                                                                     | 186 |
| 6.18 | Differential splicing events between mature mTEC from Rbfox1 <sup>lox/lox</sup> :cre <sup>+/-</sup> or Rbfox2 <sup>lox/lox</sup> :cre <sup>+/-</sup> and the corresponding cre <sup>-/-</sup> mice or from a comparison between both cre <sup>-/-</sup> samples. . . . . | 189 |
| 6.19 | Assessment of genes affected by differentially spliced events in mature mTEC with a loss of <i>Rbfox1</i> or <i>Rbfox2</i> or mature mTEC from both cre <sup>-/-</sup> samples. . . . .                                                                                  | 190 |
| 6.20 | Examples of Rbfox target genes with differentially spliced events in mature mTEC with a loss of <i>Rbfox1</i> or <i>Rbfox2</i> . . . . .                                                                                                                                 | 193 |
| 6.21 | Differential splicing event in <i>Fgfr2</i> in mature mTEC with a loss of <i>Rbfox2</i> . . . . .                                                                                                                                                                        | 194 |
| B.1  | Number of uniquely mapping reads per tissue before combining and subsampling replicates from each tissue. . . . .                                                                                                                                                        | 261 |
| B.2  | IDR results from other approaches. . . . .                                                                                                                                                                                                                               | 265 |
| B.3  | Example of a multi-gene transcript occurring in both Ensembl 91 annotation and mT&T assembly. . . . .                                                                                                                                                                    | 266 |
| B.4  | Correlation of gene expression quantified using Ensembl 91 annotation or mT&T assembly. . . . .                                                                                                                                                                          | 267 |
| B.5  | Identification of <i>Aire</i> -regulated genes. . . . .                                                                                                                                                                                                                  | 268 |
| B.6  | Comparison of <i>Aire</i> -regulated genes with a previously defined set. . . . .                                                                                                                                                                                        | 268 |
| B.7  | Quality assessment of ONT data. . . . .                                                                                                                                                                                                                                  | 270 |
| B.8  | Definition of tissue-restricted genes using <i>tau</i> . . . . .                                                                                                                                                                                                         | 271 |
| C.1  | Definition of tissue groups. . . . .                                                                                                                                                                                                                                     | 273 |
| C.2  | Transcriptome complexity in TEC. . . . .                                                                                                                                                                                                                                 | 274 |
| C.3  | Peripheral representation of tissue-restricted transcripts. . . . .                                                                                                                                                                                                      | 275 |
| C.4  | Examples of TEC-specific novel transcripts. . . . .                                                                                                                                                                                                                      | 277 |
| C.5  | Protein prediction for novel transcript in <i>Thmd</i> . . . . .                                                                                                                                                                                                         | 278 |
| C.6  | Protein prediction for novel transcript in <i>Ccdc50</i> . . . . .                                                                                                                                                                                                       | 278 |
| D.1  | Characterisation of genes containing significantly differently spliced retained intron events in immature mTEC compared to mature mTEC. . . . .                                                                                                                          | 280 |

|      |                                                                                                                                                                                                 |     |
|------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| D.2  | <i>Tbata</i> as an additional example of retained introns with significantly higher inclusion in immature mTEC. . . . .                                                                         | 282 |
| D.3  | MA plot of <i>Sleuth</i> -based identification of <i>Aire</i> -regulated genes. . . . .                                                                                                         | 284 |
| D.4  | Determination of the log2 fold change corresponding to b from <i>Sleuth</i> . . .                                                                                                               | 284 |
| D.5  | Comparison of <i>Aire</i> -regulated transcripts and genes with splicing changes affected by <i>Aire</i> between different methods. . . . .                                                     | 285 |
| D.6  | Assessment of genes containing <i>Aire</i> -regulated transcripts but not <i>Aire</i> -regulated genes and <i>Aire</i> -regulated transcripts not part of <i>Aire</i> -regulated genes. . . . . | 286 |
| D.7  | Histograms of selected quality metrics for 285 single mature mTEC . . .                                                                                                                         | 287 |
| D.8  | PCA of quality metrics in single mature mTEC . . . . .                                                                                                                                          | 288 |
| D.9  | Pipeline scheme and additional data for splicing analysis between <i>in silico</i> single-cell pools . . . . .                                                                                  | 290 |
| D.10 | Quality metrics of <i>in silico</i> single-cell pools. . . . .                                                                                                                                  | 291 |
| E.1  | Example gating schemes for live cells. . . . .                                                                                                                                                  | 295 |
| E.2  | Immunofluorescence staining without primary antibody in mouse thymus sections. . . . .                                                                                                          | 296 |
| E.3  | Heatmap of PSI values for skipped exon events in the <i>cre</i> <sup>-/-</sup> replicates from the SKO1 and SKO2 datasets. . . . .                                                              | 302 |

# List of Tables

|      |                                                                                                                                                     |     |
|------|-----------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| 2.1  | PCR programs. . . . .                                                                                                                               | 51  |
| 2.2  | Antibody panels for different types of experiments. . . . .                                                                                         | 54  |
| 3.1  | Class codes assigned by <i>gffcompare</i> after the merging step. . . . .                                                                           | 67  |
| 3.2  | Summary of transcripts and their class codes assigned by <i>gffcompare</i> in the final assembly. . . . .                                           | 67  |
| 3.3  | Number of individual graphs per annotation. . . . .                                                                                                 | 71  |
| 3.4  | Graphs with more than five genes in the Ensembl annotation (v91) using the exon graph approach. . . . .                                             | 73  |
| 3.5  | Graphs with more than five genes in the mT&T assembly using the exon graph approach. . . . .                                                        | 73  |
| 3.6  | Summary of transcripts and their class codes in reference-guided ONT assembly. . . . .                                                              | 79  |
| 5.1  | List of manually curated TRSF. . . . .                                                                                                              | 137 |
| 6.1  | Summary of differentially expressed genes in mature and immature mTEC with a loss of <i>Rbfox1</i> or <i>Rbfox2</i> . . . . .                       | 187 |
| 6.2  | Summary of results of thymus cellularity and TEC phenotyping for mice with TEC-targeted loss of <i>Rbfox1</i> and <i>Rbfox2</i> expression. . . . . | 197 |
| A.1  | List of software packages with corresponding version numbers. . . . .                                                                               | 254 |
| A.2  | Overview of quality metrics included in the <i>scseq</i> pipeline. . . . .                                                                          | 255 |
| A.3  | List of R packages with corresponding version numbers. . . . .                                                                                      | 256 |
| A.4  | Filters applied to <i>gsfisher</i> output before manual filtering. . . . .                                                                          | 256 |
| A.5  | List of primary antibodies used for immunofluorescence stainings. . . . .                                                                           | 257 |
| A.6  | List of secondary antibodies used for immunofluorescence stainings. . . . .                                                                         | 257 |
| A.7  | List of FACS antibodies. . . . .                                                                                                                    | 258 |
| A.8  | List of primers. . . . .                                                                                                                            | 259 |
| A.9  | List of in-house buffers. . . . .                                                                                                                   | 260 |
| A.10 | List of commercially available reagents and kits. . . . .                                                                                           | 260 |
| B.1  | Overview of unpublished datasets used for analyses. . . . .                                                                                         | 262 |
| B.2  | Overview of population RNA-sequencing samples before subsampling or merging. . . . .                                                                | 263 |

|     |                                                                                                                                                                                                     |     |
|-----|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| B.3 | Overview of population RNA-sequencing samples including which replicates were merged. . . . .                                                                                                       | 264 |
| B.4 | Overview table of all ONT samples. . . . .                                                                                                                                                          | 269 |
| D.1 | Number of events per splice event type in <i>rMATS</i> analysis of mTEC. . .                                                                                                                        | 281 |
| D.2 | Validation of novel transcripts containing retained introns between Illumina and ONT data. . . . .                                                                                                  | 283 |
| D.3 | Number of <i>Aire</i> -regulated transcripts for <i>DESeq2</i> and <i>Sleuth</i> . . . . .                                                                                                          | 283 |
| D.4 | Loadings for the first 5 PCs for the PCA of all quality metrics in single mTEC. . . . .                                                                                                             | 289 |
| D.5 | Overview of events for all tissue-restricted splicing factor comparisons per event type using single-cell pooling approach. . . . .                                                                 | 292 |
| D.6 | Overview of events for other tissue-restricted gene comparisons per event type using single-cell pooling approach. . . . .                                                                          | 292 |
| D.7 | Overview table summarising the characteristics of selected tissue-restricted splicing factors ( $\tau > 0.5$ ) used for prioritisation, part 1. . . . .                                             | 293 |
| D.8 | Overview table summarising the characteristics of selected tissue-restricted splicing factors ( $\tau > 0.5$ ) used for prioritisation, part 2. . . . .                                             | 294 |
| E.1 | Overview table of all RNA-sequencing samples from conditional knockout mouse model. . . . .                                                                                                         | 296 |
| E.2 | Loadings for the first 4 PCs for the PCA of all quality metrics in RNA-sequencing data from immature and mature mTEC of $Rbfox1^{lox/lox};cre^{+/-}$ and $Rbfox1^{lox/lox};cre^{-/-}$ mice. . . . . | 297 |
| E.3 | Loadings for the first 4 PCs for the PCA of all quality metrics in RNA-sequencing data from immature and mature mTEC of $Rbfox2^{lox/lox};cre^{+/-}$ and $Rbfox2^{lox/lox};cre^{-/-}$ mice. . . . . | 297 |
| E.4 | Summary of differentially expressed genes between $Rbfox1^{lox/lox};cre^{+/-}$ mice and $Rbfox1^{lox/lox};cre^{-/-}$ mice. . . . .                                                                  | 298 |
| E.5 | Summary of differentially expressed genes between $Rbfox2^{lox/lox};cre^{+/-}$ mice and $Rbfox2^{lox/lox};cre^{-/-}$ mice. . . . .                                                                  | 299 |

# List of Abbreviations

|                                    |                                                                                 |
|------------------------------------|---------------------------------------------------------------------------------|
| <b>Ataq-seq</b>                    | Assay for Transposase-Accessible Chromatin using sequencing.                    |
| <b>ATP</b>                         | Adenosine tri-phosphate.                                                        |
| <b>ChIP-seq</b>                    | Chromatin immunoprecipitation followed by sequencing.                           |
| <b>cTEC</b>                        | Cortical thymic epithelial cells.                                               |
| <b>DP</b>                          | Double-positive (thymocyte).                                                    |
| <b>FACS</b>                        | Fluorescence-activated cell sorting.                                            |
| <b>FPKM</b>                        | Fragments Per Kilobase Per Million.                                             |
| <b>HITS-CLIP</b>                   | High-throughput sequencing of RNA isolated by crosslinking immunoprecipitation. |
| <b>k</b>                           | thousand.                                                                       |
| <b>kbp</b>                         | kilo base pairs.                                                                |
| <b>lincRNA</b>                     | long intergenic non-coding RNA.                                                 |
| <b>M</b>                           | million / molar in Chapter 2.                                                   |
| <b>MGland</b>                      | Mammary Gland.                                                                  |
| <b>MHC</b>                         | Major histocompatibility complex.                                               |
| <b>mTEC</b>                        | Medullary thymic epithelial cells.                                              |
| <b>mT&amp;T assembly</b>           | mouse tissue & TEC assembly.                                                    |
| <b>mtech<sub>i</sub>_gfpposhet</b> | mature AIRE-positive mTEC.                                                      |
| <b>mtech<sub>i</sub>_gfpposko</b>  | mature <i>Aire</i> -KO mTEC.                                                    |
| <b>mtech<sub>i</sub></b>           | mature mTEC.                                                                    |
| <b>mtecto</b>                      | immature mTEC.                                                                  |
| <b>MXE/MX</b>                      | Mutually Exclusive Exon.                                                        |
| <b>ONT</b>                         | Oxford Nanopore Technology.                                                     |
| <b>PCA</b>                         | Principal Component Analysis.                                                   |
| <b>PCR</b>                         | Polymerase Chain Reaction.                                                      |
| <b>PGE</b>                         | Promiscuous Gene Expression.                                                    |
| <b>PSI</b>                         | Percentage spliced-in.                                                          |

|                    |                                               |
|--------------------|-----------------------------------------------|
| <b>RI</b>          | Retained Intron.                              |
| <b>Ribo-seq</b>    | Ribosomal footprinting plus sequencing.       |
| <b>RT-PCR</b>      | Reverse Transcription PCR.                    |
| <b>SE</b>          | Skipped Exon.                                 |
| <b>SJ</b>          | Splice junction.                              |
| <b>SKO1</b>        | Rbfox1 <sup>lox/lox</sup> :cre <sup>+/-</sup> |
| <b>SKO2</b>        | Rbfox2 <sup>lox/lox</sup> :cre <sup>+/-</sup> |
| <b>SP</b>          | Single-positive (thymocytes).                 |
| <b>SmIntestine</b> | Small Intestine.                              |
| <b>SubcFatPad</b>  | Subcutaneous fat pad.                         |
| <b>TEC</b>         | Thymic epithelial cells.                      |
| <b>TCR</b>         | T cell receptor.                              |
| <b>TPM</b>         | Transcripts Per Million.                      |
| <b>TRA</b>         | Tissue-restricted antigen.                    |
| <b>TRSF</b>        | Tissue-restricted splicing factor.            |
| <b>t-SNE</b>       | t-distributed stochastic neighbour embedding. |
| <b>UTR</b>         | Untranslated region.                          |
| <b>WT</b>          | cre <sup>-/-</sup> samples.                   |

# 1

## Introduction

### Contents

---

|            |                                                                                                                     |           |
|------------|---------------------------------------------------------------------------------------------------------------------|-----------|
| <b>1.1</b> | <b>Adaptive immune system . . . . .</b>                                                                             | <b>2</b>  |
| <b>1.2</b> | <b>Central tolerance in the thymus . . . . .</b>                                                                    | <b>3</b>  |
| 1.2.1      | Development of immature thymocytes into single-positive, naïve T cells . . . . .                                    | 5         |
| 1.2.2      | Maturation and function of thymic epithelial cells . . . . .                                                        | 6         |
| 1.2.3      | The Autoimmune Regulator, Aire, and its role in the thymus . . . . .                                                | 8         |
| <b>1.3</b> | <b>Alternative splicing . . . . .</b>                                                                               | <b>12</b> |
| 1.3.1      | Alternative splicing in development and disease . . . . .                                                           | 13        |
| 1.3.2      | Cross-species conservation of alternative splicing . . . . .                                                        | 14        |
| 1.3.3      | Regulation of alternative splicing . . . . .                                                                        | 16        |
| 1.3.4      | Long-read RNA-sequencing methods to assess splicing . . . . .                                                       | 18        |
| 1.3.5      | Effects of alternative splicing on proteome complexity . . . . .                                                    | 20        |
| <b>1.4</b> | <b>Alternative splicing in the thymus and its role in the induction of self-tolerance . . . . .</b>                 | <b>22</b> |
| 1.4.1      | Complexity of the transcriptome and possible underlying mechanisms of splicing in thymic epithelial cells . . . . . | 22        |
| 1.4.2      | The effect of thymic alternative splicing on self-tolerance . . . . .                                               | 24        |
| <b>1.5</b> | <b>Open questions and hypotheses . . . . .</b>                                                                      | <b>28</b> |
| <b>1.6</b> | <b>Aims . . . . .</b>                                                                                               | <b>29</b> |

---

## 1.1 Adaptive immune system

Protection against pathogens such as bacteria or viruses is an essential process in mammals involving a large variety of cell types. The processes of immune defense can be split into innate and adaptive immunity. The main cell types involved in adaptive immune responses are T cells and B cells. Due to the presence of antigen receptors on their surfaces, they can exhibit specific responses after encountering an antigen. While B cells respond by producing and secreting antibodies, T cells differentiate into different types of effector T cells.

For T cell-mediated immune responses two important surface receptors are required, namely the T cell receptor (TCR) with a unique specificity and one of the two co-receptors termed CD4 or CD8. The expressed co-receptor results in a restriction for a given T cell to recognise major histocompatibility complex I (MHCI) in the case of CD8 or MHCII in the case of CD4 (reviewed by Gaud *et al.* [1]). Upon activation, CD8 T cells differentiate into CD8 cytotoxic T cells and destroy the target cells. An example of target cells are pathogen-infected cells showing expression of MHCI bound with pathogenic peptides on the cell surface. These MHCI:peptide complexes are recognised by CD8 T cells resulting in the killing of the infected cell. To ensure specificity of such a reaction, a variety of co-stimulatory signals from for instance dendritic cells and CD4 T cells are required. One form of co-stimulatory signal is the secretion of the cytokine interleukin-2 (IL-2) by CD4 T cells upon stimulation by an antigen-presenting cell (APC) such as dendritic cells (reviewed by Boyman *et al.* [2]).

In contrast to the cytotoxic activity of CD8 T cells, CD4 T cells are (i) activated by MHCII:peptide complexes and (ii) function through recruitment and activation of other cell types (reviewed by Wan *et al.* [3]). The differentiation of the naive T cells into at least one of five established effector cell types, i.e. T helper cell type 1 ( $T_H1$ ),  $T_H2$ ,  $T_H17$ ,  $T_H22$ , T follicular helper cell ( $T_{FH}$ ), T regulatory cell ( $T_{reg}$ ), depends on the locally present cytokines. Each effector cell type subsequently evokes the secretion of specific



effector molecules to activate downstream effector cells of the innate and/or adaptive immune system.

Given the destructive power especially of the adaptive immune system, an essential prerequisite of this arm of the defence system is the ability to effectively distinguish between benign 'Self' and potentially harmful 'Non-Self'. This capacity is instructed during T cell development, when a repertoire of T cells is selected, that collectively expresses an antigen receptor repertoire reactive to foreign antigens, but remains non-responsive to tissue-restricted antigens (TRA a.k.a. self-antigens) (reviewed in Klein *et al.* [4]).

## 1.2 Central tolerance in the thymus

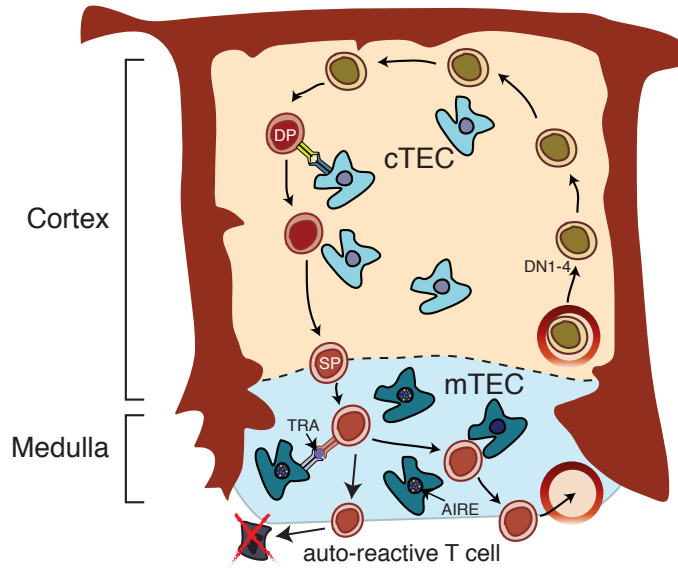
The thymus is an important primary lymphoid organ indispensable for the generation of a functional T cell repertoire, which includes the maturation of immature T cells (thymocytes) to naïve CD4 or CD8 T cells and their appropriate selection (Figure 1.1). The stringent process of T cell receptor selection deletes auto-reactive T cells and prevents their eventual exit from the thymus into the periphery. TCR specificity is randomly chosen and therefore requires a stringent selection process during thymocyte maturation to avoid that potentially harmful T cells populate peripheral lymphoid tissues.

The TCR comprises of two major components, namely  $\alpha$  and  $\beta$  chains, which both have a constant and a variable region (reviewed by Gaud *et al.* [1]). In addition to these TCR specificity subunits, a signal-transduction complex, namely the CD3 complex, is associated with the  $\alpha$  and  $\beta$  chains. The CD3 complex consists of two heterodimers containing CD3 $\gamma, \epsilon$  and CD3 $\delta, \epsilon$ , and a central homodimer of CD3 $\zeta, \zeta$ .

The variable region in both  $\alpha$  and  $\beta$  chains is produced via somatic recombination of a single gene from the family of variable (V), diversity (D; only for  $\beta$  chain) and joining (J) gene segments. A large repertoire of TCR specificities is achieved via this process of DNA rearrangements referred to as V(D)J-rearrangement. This repertoire is actively selected during thymic T cell maturation via two sequential processes that gauge the

TCR affinity for MHC:antigen complexes.

The stromal microenvironment in the thymus is essential for the development of functional T cells. The thymic epithelium comprises of cortical thymic epithelial cells (cTEC) located in the outer region designated cortex and medullary thymic epithelial cells (mTEC) present in the inner compartment known as the medulla (Figure 1.1). The two compartments can be easily distinguished in hematoxylin & eosin (H&E) staining of tissue sections due to the high density of cell nuclei in the cortex compared to a less densely populated medulla [5].



**Figure 1.1:** Schematic overview of the thymic compartments and the stages of intrathymic T cell development. Immature double-negative (DN) thymocytes enter the thymus, where they first become  $CD4^+CD8^+$  double-positive (DP) thymocytes. After restriction to the CD4 or CD8 T cell lineage, the single-positive (SP) T cells undergo positive and negative selection. Positive selection is based on the presence of a functioning T cell receptor and performed by cortical thymic epithelial cells (cTEC). The subsequent negative selection process is mainly regulated by medullary thymic epithelial cells (mTEC), which present complexes of MHCII and tissue-restricted antigens (TRA) on their cell surfaces. Figure modified from a review by Takahama [6].

### 1.2.1 Development of immature thymocytes into single-positive, naïve T cells

T cell precursors, that are displaying a  $CD4^-CD8^-CD44^{+}ckit^{+}$  phenotype (designated as double-negative 1, DN1), derived from the bone marrow, access the thymus at the cortico-medullary junction [6] (Figure 1.1). The sequential developmental stages (DN1 to DN4) can be characterised by the cell surface expression of CD25 and CD44. The DN2 stage is defined by the additional expression of CD25, whereas the DN3 stage is marked by the downregulation of CD44 and the persistence of CD25 on the cell surface. The presence of the pre-TCR, which consists of a non rearranging pre-TCR  $\alpha$  chain, a rearranged TCR  $\beta$  chain and the associated CD3 complex, on DN3 cells represents an important checkpoint step. The presence of the pre-TCR with a sufficiently rearranged TCR  $\beta$  chain on DN3 cells leads to enhanced proliferation followed by downregulation of CD25 in DN4. In the subsequent development stage, CD4 and CD8 are concomitantly expressed at the cell surface identifying these cells as DP thymocytes. During the DP stage, the VJ-recombination of the  $\alpha$  locus results in the generation of a complete  $\alpha\beta$  TCR, which is subsequently selected for its antigen specificity.

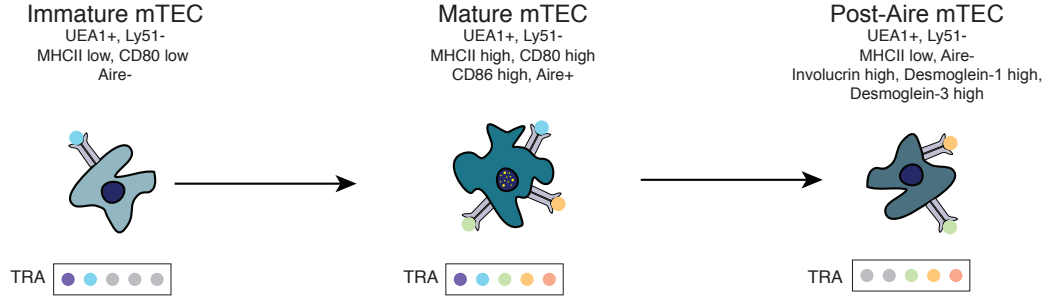
In a first, positive selection, immature thymocytes are chosen that express a functional TCR as assessed by interaction with peptide:MHC complexes expressed on cTEC. TEC in the cortex have the ability to present peptides complexed to MHCI and MHCII molecules on their cell surface, that have been processed from endogenous protein via proteasomal degradation. A specific thymoproteasome containing the  $\beta 5t$  subunit is involved in the generation of peptides loaded onto MHCI. The lack of such a thymoproteasome results in defective positive selection of  $CD8^{+}$  T cells [7]. Ninety percent of thymocytes do not survive positive selection, as, consequent to the lack of a survival signal, they undergo apoptosis in a phenomenon known as 'death by neglect' [4]. Depending on the affinity for the complex of peptide presented by MHCI or MHCII molecules, DP thymocytes develop into CD8 or CD4 single-positive (SP) T cells, respectively. The first wave of negative selection is mediated by cTEC in the cortex, where single-positive T cells expressing TCRs with high affinity to self-antigens are deleted [8, 9]. This process of negative thymocyte

selection results in the programmed cell death of these thymocytes. Single-positive T cells eventually express the chemokine receptor CCR7 and are enabled to migrate into the medulla. In a second selection step, mTEC present self-antigens on MHCII complexes to the SP T cells resulting in negative selection if high affinity between TCR and self-antigens occurs. Thymocytes expressing a TCR with an affinity that is not high enough to execute programmed cell death, but substantially higher than that required for positive selection will adopt a fate different to apoptosis, namely that of regulatory T cells.

In the thymic medulla, the first parts of T cell maturation is defined as three stages, namely semi-mature (SM), mature 1 (M1) and mature 2 (M2) stages [10]. While SM cells are susceptible to apoptosis induction, this susceptibility is lost in M1 cells. Additionally, M1 cells become proliferation competent and upregulate MHCI and the transcription factor KLF2. Before exiting from the thymus into the periphery, M2 cells show expression of a range of molecules including the egress marker S1P1 and the lymph node homing receptor CD62L, the expression of which is promoted by KLF2 [11, 10].

### 1.2.2 Maturation and function of thymic epithelial cells

The two phenotypically different TEC lineages - cTEC and mTEC - develop from a common endoderm-derived precursor arising from the third pharyngeal pouch in mice and humans [12]. While the initial commitment to a TEC fate does not require the expression of the transcription factor Forkhead box N1 (FoxN1), later stages of the cell's development require this master regulator and involve the formation of a functional 3D epithelial scaffold [13, 14]. Both TEC types are characterised by their expression of the epithelial marker EPCAM. Cortical TEC are further marked by LY51 expression on the cell surface. The expression of CD80 and the capacity to bind the lectin UEA-1 (*Ulex europaeus agglutinin 1*) are characteristic features of mTEC, whereas CD86 and AIRE are only expressed by a subpopulation of mature mTEC (Figure 1.2).



**Figure 1.2:** Stages of mTEC maturation. Distinct stages of mTEC maturation can be defined by the levels of CD80, CD86, MHCII and AIRE expression. The latter controls the expression of thousands of the tissue-restricted antigens (TRA) detected in mTEC. In terminally differentiated mTEC, the expression of AIRE is lost, while markers such as Desmoglein-1 and Desmoglein-3 are expressed. The TRA profile of these terminally differentiated mTEC is not as comprehensive as in mature mTEC [15]. Scheme is modified from Abramson & Anderson [16].

The maturation from immature ( $CD80^{\text{low}}$ ,  $CD86^-$ ,  $MHCII^{\text{low}}$ ; termed mTEC<sub>lo</sub>) to mature mTEC ( $CD80^{\text{high}}$ ,  $CD86^+$ ,  $MHCII^{\text{high}}$ ; termed mTEC<sub>hi</sub>) is regulated by RANK ligand, lymphotoxin  $\beta$  and CD40 ligand signalling. It is further demarcated by the upregulation of MHCII and CD80 as well as the expression of CD86 and the Autoimmune Regulator (Aire). To achieve the expression of almost the entire protein-coding gene repertoire, mTEC employ a program termed 'promiscuous gene expression (PGE)', that is in part, but not exclusively, regulated by AIRE (reviewed by Anderson & Su [17]; Section 1.2.3). Another protein named FEZ family zinc finger protein 2 (FEZF2) has also been recently described as a regulator of PGE [18].

Recently two groups have identified a novel subset of TEC with high similarity to peripheral tuft cells [19, 20]. Bornstein *et al.* identified the cell type as a distinct mTEC cluster (mTEC IV) using single-cell RNA-sequencing that showed shared gene expression with intestinal tuft cells (for instance *Pou2f3*) [19]. In addition to the shared set of genes between the mTEC IV cluster and intestinal tuft cells, a set of genes including *Krt17* and *Gp2* was only found on the mTEC IV cluster. In a mouse model with a loss of *Pou2f3*, the authors further observed that the mTEC IV cluster of tuft cells was absent.

In addition to the mTEC subtypes described above, heterogeneity was observed among mTEC that express low levels of MHCII on their cell surface, as cells with this phenotype are either immature or terminally differentiated (i.e. post-AIRE mTEC, Figure 1.2). The development from mature to post-AIRE mTEC is regulated by lymphotoxin  $\beta$  receptor mediated signals and marked by the expression of Involucrin and CCL21 [21, 22]. Moreover, these cells showed reduced expression of *Aire*-dependent genes as well as expression of keratinocyte-related antigens, Desmoglein-1 and Desmoglein-3 [23, 15].

### 1.2.3 The Autoimmune Regulator, Aire, and its role in the thymus

AIRE is required in mature mTEC to ensure comprehensive expression of TRA for T cell selection. In humans, loss-of-function mutations of *AIRE* cause an immunodeficiency termed autoimmune-polyglandular-syndrome-type-1 (APS-1, also known as autoimmune-polyendocrinopathy-candidiasis-ectodermal dystrophy, APECED; OMIM #240300). The documented *AIRE* mutations are numerous and spread across the entire open-reading frame of the *AIRE* gene (reviewed by Bruserud *et al.* [24]). APS-1 patients show a wide range of severe organ-specific autoimmune pathologies, which underscore the significance of AIRE mediating PGE and its effect on the induction of T cell tolerance to 'Self'. Tissues frequently affected in APS-1 patients are the parathyroid, the adrenal glands and reproductive organs (reviewed by Bruserud *et al.* [24]).

In 2002, two independently generated *Aire*-knockout mouse models were reported to develop severe autoimmune phenotypes [25, 26], demonstrating that Aire's role was conserved between mouse and human. While Ramsey *et al.* interpreted the findings as a 'disturbance of homeostasis of the immune system' [25], the report by Anderson *et al.* showed in detail, that the absence of AIRE in mTEC results in a loss of normal TRA expression and consequently establishes a population of T cells that contain TCR specificities directed against self-antigens.

AIRE expression in mature mTEC increases the fraction of expressed protein-coding genes from 76% to 87%, the highest number reported for any cell type so far [27]. AIRE is an unconventional transcriptional activator which does not bind directly to DNA, but instead activates tissue-restricted genes characterised by inactive or repressive chromatin marks. AIRE has been shown to bind unmethylated H3K4 (H3K4me0) via its plant homeodomains 1 (PHD1) finger domain [28, 29]. However, in an *in vitro* system, global demethylation of H3K27 was not sufficient to modify TRA expression [29], thus further regulatory mechanisms apart from binding to H3K4me0 are in place to regulate PGE. To further assess the mechanism and identify additional regulators of PGE, an *in vitro* pull-down assay was used to identify AIRE binding partners, which can be categorised into four major groups, namely 'transcription', 'chromatin binding', 'pre-mRNA processing' and 'nuclear transport' [30]. Interaction partners identified by Abramson *et al.* and relevant to this study were for instance three members of the serine-arginine-rich splicing factor family (SRSF1, SRSF2 and SRSF3).

In addition to the chromatin mark H3K4me0, another post-translational histone modification has been correlated with AIRE-mediated PGE: *Aire*-regulated genes are marked by the presence of H3K27me3 and absence of the active H3K4me3 mark at their transcriptional start sites [27]. AIRE itself was shown to interact with two genes known to associate with the Polycomb Repressive Complex 2 (PRC2), namely Chromodomain-Helicase-DNA members 4 (CHD4) and 6 (CHD6) [31, 32].

In addition to interactions with other proteins via the two PHD domains, AIRE's SAND domain interacts with a complex of ATF7IP and MBD1 [33]. Hence, AIRE indirectly recognises methylated DNA as MBD1 binds methylated CpG dinucleotides and maintains heterochromatin. Consequently, the targeted loss of MBD1 expression results in spontaneous autoimmunity in mice. AIRE has also been shown to release stalled RNA polymerase II via the binding to the positive transcription elongation factor b (pTEBb) and heterogeneous nuclear ribonucleoprotein L (HNRNPL) [34, 35, 36]. The effect of pTEBb was identified using *in vitro* assays and further validated in a knockout mouse

model by deleting the pTEBb subunit, CycT1. These mice showed lymphocytic infiltrates that affected salivary glands [34]. In 2011, a human *AIRE* mutation was mapped to the domain binding pTEBb and the binding of AIRE to pTEBb validated using *in vitro* assays [35]. AIRE binding to pTEBb results in a higher rate of spliced compared to unspliced RNA, suggesting co-transcriptional splicing and the recruitment of splicing factors by AIRE at least *in vitro*. The binding of AIRE to HNRNPL was identified along-side other splicing factors such as SFRS1/SRSF1 or DDX6 using a luciferase-based RNAi screening in the human TEC cell line 4D6 and validated in lentigenic knockdown mice [36].

Due to the number of mTEC in a single mouse (estimated to be 300 k in young C57BL/6 mice [37]), assays to assess the molecular basis of AIRE function in mTEC are challenging as a high input cell number is required for standard molecular biology assays. For instance assays such as ATAC-sequencing to assess open chromatin or HITS-CLIP to examine RNA binding protein sites depend on a high number of input cells. Attempts have been made to understand the temporal and spatial characteristics of PGE. Initially, PGE was suggested to be a stochastic mechanism as co-expression patterns of a set of peripherally co-expressed genes did not occur in mTEC [38, 39].

Recent advances in single-cell RNA-sequencing have opened new avenues to examining PGE. Consistent with the observations from qPCR, Sansom *et al.* did not find clear sub-groups of *Aire*-positive mature mTEC following clustering of a small number (n=141) of such cells based on their expression of *Aire*-regulated TRA [27]. In contrast to population RNA-sequencing where low levels of *Aire*-regulated genes were observed, infrequent (median 1.1 % of single mature mTEC) but 16-fold (median) higher expression of *Aire*-regulated genes was described at the single-cell level compared to population level. In further single-cell RNA-sequencing studies, Brennecke *et al.* described heterogeneity between single mTEC with co-expression patterns between TRA [40]. Such co-expressed genes were located in close proximity of each other in the genome. Another study described the regulation process of PGE as 'ordered stochasticity' [41]. The authors also found gene expression microclusters of specific TRA within single cells from the medulla. Meredith *et al.* further confirmed the observation that *Aire*-induced transcripts were



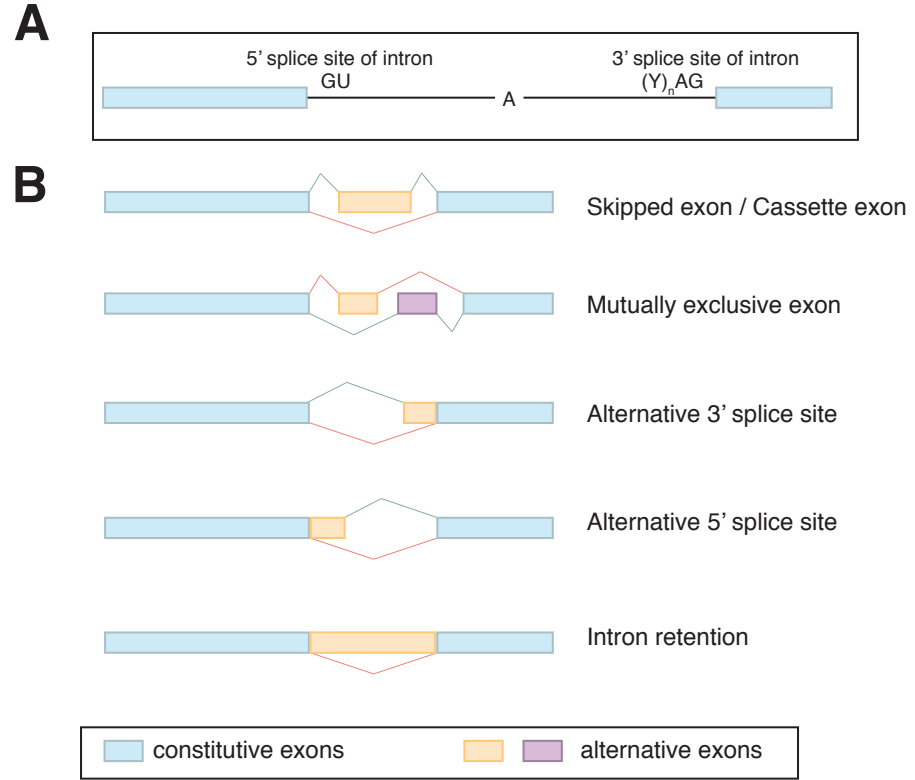
expressed in a low frequency of single mTEC cells.

The expression of *Aire* in mTEC is regulated by engagement of essential, upstream, non-coding sequences, which are binding elements for nuclear factor- $\kappa$ B (NF- $\kappa$ B) complexes [42, 43]. Deletion of the binding element results in the absence of AIRE and as a consequence spontaneous autoimmunity in mice [42]. Furthermore, these sequences are highly conserved across species [43]. This observation links RANK signalling via NF- $\kappa$ B to regulation of *Aire* expression. Apart from TEC, expression of *Aire* has only been reported in embryonic stem cells, B cells and the testis [44, 45, 46]. Additionally, a rare *Aire*-expressing cell type, referred to as 'extrathymic *Aire* expressing cells' (eTACs), is present in secondary lymphoid tissues in both mice and humans [47, 48]. The expression of only a subset of TRA is regulated by AIRE [49, 27, 50], suggesting that additional mechanisms regulating PGE are likely to exist.

## 1.3 Alternative splicing

In 1977, the process of splicing was first described by the groups of Phillip Sharp and Richard Roberts [51, 52]. Phillip Sharp's group used electron microscopy to assess the 5' terminus of adenovirus 2 RNA, where they observed a split of genes in the adenovirus RNA. This observation led to the suggestion of intramolecular joining of precursor molecule parts to a mature RNA product. Such a process explained unexpected phenomena such as the presence of long RNA molecules in the nucleus compared to short RNA molecules in the cytoplasm. Since this ground-breaking discovery, a vast amount of research in the field of splicing has been conducted. The early molecular biology work on individual genes was followed by genome-wide approaches. The advances in next-generation sequencing technologies (and even more so third generation sequencing) have made it possible to easily discover and quantify isoform usage across the whole transcriptome. Additionally, the formation of big consortia such as ENCODE has increased the number of publicly-available high-depth RNA-sequencing datasets, which can be used to study splicing.

The process of splicing is generally divided in two classes, namely constitutive and alternative splicing. Alternatively spliced exons have been defined as exons that are 'included in mature mRNA in certain cellular contexts but excluded in others' (definition from Chen & Manley [53]). This definition reflects the tissue-specific nature of alternative splicing or the temporal differences such as found, for instance, during cellular differentiation. While the definition is focussed on exons, other types of alternative splicing such as retained introns have also been described (Figure 1.3). Constitutive splicing is regulated by core splicing machinery that most typically recognises the conserved splicing sequences GT-AG, where GT and AG are located on the 5' and 3' splice sites of the intron, respectively [54]. Another core splicing signal is the branch point sequence (indicated by 'A' in Figure 1.3A), which interacts with spliceosome components during the splicing process and facilitates the formation of the so-called RNA lariat structure [55].



**Figure 1.3:** Alternative splicing event types. **A:** Scheme of intron definition via highly conserved splicing motifs near intron ends. The base 'A' in the middle of the intron represents the branch point. **B:** Five common alternative splicing event types. (Y)<sub>n</sub> = polypyrimidine tract.

The process of alternative splicing involves competition between multiple pairs of possible 3' and 5' splice sites in a process known as 'exon definition' (reviewed by Fu & Ares [56]). This process is regulated by a complex interplay of *cis*-regulatory sequence elements, splice site strength, and by *trans*-acting sequence-specific RNA-binding proteins known as splicing factors. Splicing factors that regulate alternative splicing often showed tissue- or celltype-specific expression patterns (reviewed by Chen & Manley [53] or Baralle & Giudice [57]).

### 1.3.1 Alternative splicing in development and disease

Important processes regulated by alternative splicing include for instance developmental processes, such as neuronal differentiation, or the general tissue functionality, such as that found in the heart, where alternative splicing is important for modulated muscle elasticity (reviewed by Baralle & Giudice [57]).

Multiple diseases have also been found to arise due to splicing factor mutations that result in aberrant splicing. For instance, mutations in RBFOX1 or dysregulation of SRRM4 have been connected to Autism Spectrum Disorders (ASD) [58, 59, 60, 61]. Another well-described example is the misregulation of *TITIN* splicing by dis-functioning RBM20, which causes dilated cardiomyopathy [62, 63].

### 1.3.2 Cross-species conservation of alternative splicing

Two comprehensive studies assessing alternative splicing across evolution using RNA-sequencing were published in 2012 [64, 65]. Barbosa-Morais *et al.* reported a larger divergence in alternative splicing than in overall gene expression between human and frog. Remarkably, the comparison of percentage spliced-in (PSI) for conserved exons showed a higher pairwise correlation between the organs in one species than between the same organ of different tetrapod species. Barbosa-Morais *et al.* further found evidence that *cis*-regulatory elements were the main regulator of species-specific alternative splicing. Nucleic acid binding proteins were found to be significantly enriched for species-specific alternative splicing events. However within those genes, the events tended to fall outside the nucleic acid binding domains. The authors highlighted a species-classifying exon in the *trans*-acting splicing factor PTBP1, which contained an alternatively spliced and 'activity-modulating' exon. This exon had a high disorder rate and was frequently skipped in mammalian species. Overall, the authors found species-classifying alternative splicing events to be enriched in disordered domains of *trans*-acting regulators, which could result in altered protein-protein interaction networks.

Consistent with Barbosa-Morais *et al.*, Merkin *et al.* observed organ-specific clustering for gene expression, but mainly species-specific clustering for alternatively spliced exons based a hierarchical clustering of PSI values [65]. However, three tissues were found to be exceptions in the PSI clustering: The brain and combinations of heart and muscle samples from the different species clustered in an organ-specific rather than species-specific manner, suggesting that the alternative splicing program in these organs might be highly

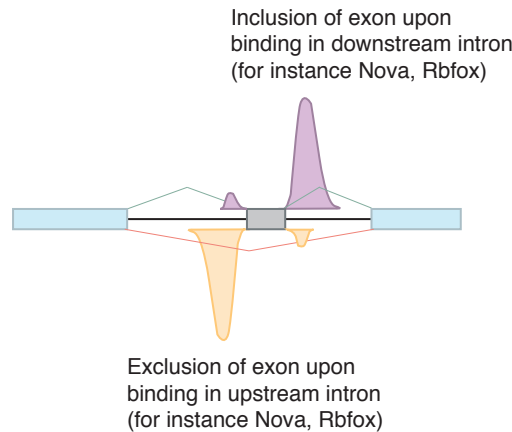
tissue-specific. In addition, Merkin *et al.* found highly conserved 5-nucleotide sequences in downstream introns of alternatively spliced exons, matching with binding sites for known splicing regulator families such as Mbnl, Rbfox or Ptb. The authors suggested that signalling networks were affected by alternatively spliced exons as they found that conserved tissue-regulated exons frequently contained phosphorylation sites. The two reports above suggest an important role of *cis*-regulatory sequences in alternative splicing regulation and potentially a less important role for *trans*-acting splicing factors. The regulatory sequences were also mapped to a small set of splicing factor families. Given these observations, and the overlapping specificities of splicing factors within families [66, 67, 68], it could be speculated that it is not essential for all splicing factors to be expressed in TEC to regulate alternative splicing in this cell type.

An important pre-requisite for the assessment of splicing is a comprehensive transcriptome annotation. However, until recently the mouse transcriptome was poorly annotated compared to the human transcriptome. Given that knockout mouse models are essential tools for biomedical research, the issue of poor transcriptome annotation in the mouse was addressed by the mouse ENCODE project, which generated a large RNA-sequencing dataset including adult and embryonic mouse tissues [69]. Using this dataset, an assembly containing 200 k novel transcripts with CAGE support from the FANTOM dataset was generated [70]. Pervouchine *et al.* further defined genes with small variation in expression levels between cell types and species as 'constrained' genes. While previous studies had defined general house-keeping genes in different species, the definition of constrained genes by Pervouchine *et al.* included an evolutionary component. The evolutionary constrained genes were marked by active epigenetic marks from mouse and human ChIP-seq data. In addition, the transcript abundances of such constrained genes were found to be influenced by splicing to a greater extent than was observed for non-constrained genes. Finally, a high number of mouse splice junctions in protein-coding genes was found to be conserved in human protein-coding genes (200 k out of 298 k mouse splice junctions).

### 1.3.3 Regulation of alternative splicing

Splicing regulation is a complex interplay between *cis*-regulatory RNA sequences and *trans*-acting splicing factors or RNA-binding proteins. A method to assess the protein-RNA interactions genome-wide, namely crosslinking and immunoprecipitation followed by high-throughput sequencing (CLIP-seq or HITS-CLIP), revealed that factors such as Nova or Rbfox regulated inclusion or exclusion of a target exon depending on the location of the binding motif (Figure 1.4) [71, 72]. These insights into regulatory binding sites motivated the modelling of splicing networks with machine-learning approaches in order to help interpret the 'splicing code' and predict the transcripts regulated by a splicing factor of interest. Such networks including target exons were generated for Nova and Rbfox [71, 73]. The Bayesian model for the splicing network of the Rbfox family based on *in vivo* HITS-CLIP data, junction microarray data from *Rbfox1* knockout mouse brains, shRNA knockdown of *Rbfox2* in HeLa cells, evolutionary signatures and bioinformatic predictions of motif sites revealed that around 1,000 exons were regulated by Rbfox. Furthermore, Weyn-Vanhentenryck *et al.* identified Rbfox-regulated splicing events in a set of 48 genes that had previously been associated with susceptibility to autism.

In addition to *cis*-regulatory elements, combinations of splicing factors can regulate events in a synergistic or antagonistic fashion (reviewed by Raj & Blencowe [74]). For instance Rbfox and Nova have been shown to synergistically regulate alternative splicing of the GABA receptor gamma 2 (*Gabrg2*). The binding sites in this gene have been identified using HITS-CLIP data from mouse brains for each gene individually [75, 73] and an increase in exon inclusion in the presence of both splicing factors was confirmed using a minigene assay with reverse transcription polymerase chain reaction (RT-PCR) in the human embryonic kidney 293T cell line [75]. Using a similar approach, antagonistic effects for Nova and Rbfox have been observed for example in the gene *Arhgef12*. Another example of co-regulation by two important splicing factors, namely Rbfox and Ptbp, was found for microexons. These short (3-50 bp) exons were identified as neuronal-specific and frequently misregulated in autism-spectrum disorder using RNA-sequencing data from ASD patient brains compared to controls [58]. Moreover, Irimia *et al.* identified the



**Figure 1.4:** Activity of splicing factors depends on binding position relative to the intron. Schematic locus showing alternative splicing factor binding locations represented by HITS-CLIP peaks. Binding of splicing factors such as Nova or Rbfox in the upstream intron of an exon tends to result in skipping of the alternative exon, while binding in the downstream intron generally promotes the inclusion of the alternative exon [71, 73].

neuronal splicing factor nSR100 (SRRM4) as a regulator of microexon inclusion. The inclusion of microexons often impacted domains involved in protein-protein interactions, for example leading to stronger neuronal-specific protein interactions between AP1S2 and AP1B1. In parallel, Li *et al.* had found high conservation of intronic sequences flanking alternatively spliced brain-specific microexons [76], which showed enrichment for Rbfox binding motifs. The binding of Rbfox in the 3' flanking regions of alternatively but not constitutively-spliced microexons was confirmed using HITS-CLIP data from mouse brains. PTBP1 was found to regulate the exclusion of microexons in an *in vitro* knockdown experiment. Consistent with Irimia *et al.*, Li *et al.* showed that alternatively spliced microexons tended to localise to disordered protein domains, which likely affected protein-protein interactions.

Although the complex regulation of splicing via the so-called 'splicing code' has not been solved fully, a catalogue of well-described alternative splicing examples and factors regulating them is available (reviews such as Chen & Manley [53] or Jangi & Sharp [77]). An initial study to systematically define splicing factor families in humans [78] was followed by a microarray-based study to identify mouse orthologues and focus on tissue-restricted splicing factors [79]. The latter identified 104 tissue-specific splicing-related

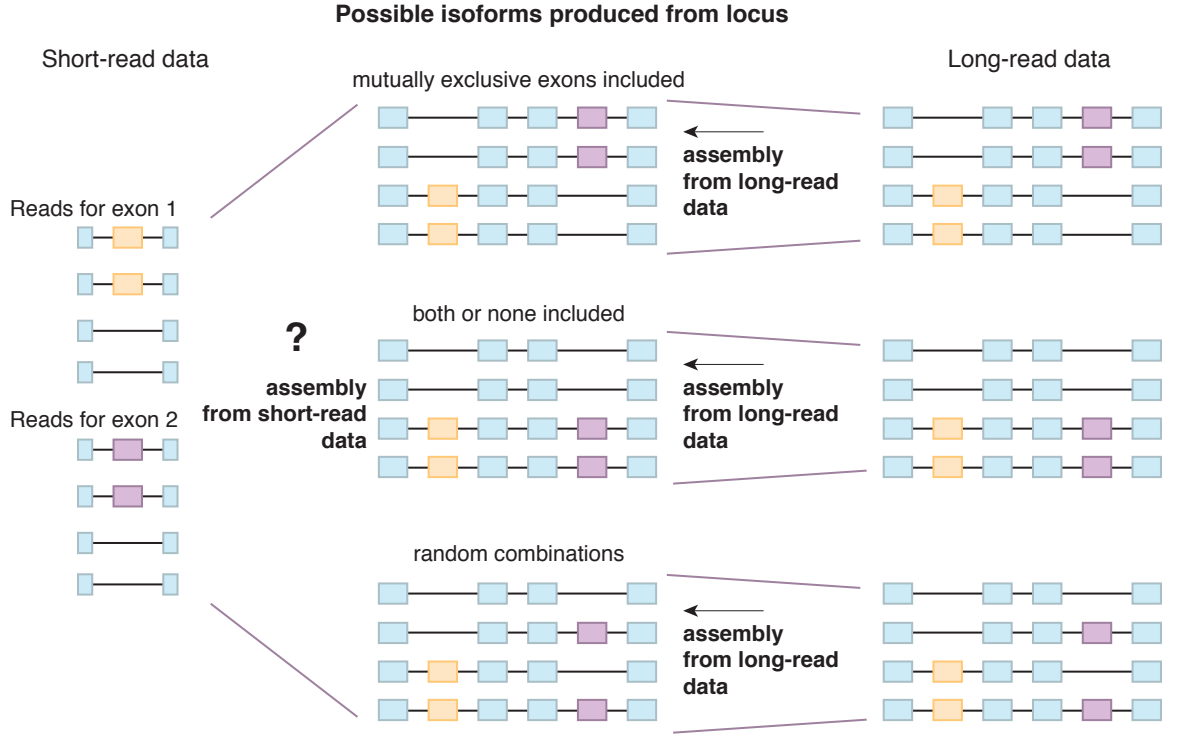
genes using mouse, human and chimpanzee microarray datasets ( $> 1.5$ -fold difference in at least one species). The organs with the highest number of tissue-restricted splicing factors were the brain and testis. Known brain-specific splicing factors such as *Nova1*, *Rbfox1* or members of the Celf (Brunol) and Elavl families were found in the brain-regulated set.

### 1.3.4 Long-read RNA-sequencing methods to assess splicing

The current next-generation sequencing technologies are producing vast amounts of data from short-reads, but to unambiguously resolve the genomic sequence of repetitive regions for genome assemblies or isoform structures for transcriptome assemblies, long-read sequencing technologies are required. For such applications, Oxford Nanopore Technology (ONT) sequencing and Pacific Bioscience (PacBio) sequencing were developed. The former method uses nanopores to measure the changes in electric current occurring while a strand of nucleotides travels through the pore (reviewed by Jain *et al.* [80]). Fluctuations in the recorded electric current can be converted into bases using a basecalling software. In contrast, PacBio sequencing uses single-molecule real time (SMRT) sequencing. In this method, a complementary nucleotide strand is produced and the fluorescent emissions of fluorophore-labelled nucleotides are measured as they get incorporated (reviewed by Roberts *et al.* [81]). An example of a situation, where short-read sequencing gives insufficient information about exon connectivity and resulting isoform structures, is given in Figure 1.5. The short reads on the left can be assembled to any of the three isoform combinations in the middle, while use of long-reads would allow the actual isoform combinations to be easily determined. The further development and accessibility of long-read sequencing technologies will improve transcriptome assemblies and help to resolve isoform structures unambiguously.

In 2017, two studies performed RNA-sequencing using modified SmartSeq2 protocols to convert RNA to cDNA and run ONT sequencing from bulk [83] or single cell samples [84]. The average read length using such cDNA-ONT sequencing was around 1,500 bp in both studies. On the single-cell level, Byrne *et al.* observed high diversity in B cell





**Figure 1.5:** Scheme of an example where short-read sequencing was insufficient to resolve transcript structures. Short Illumina reads (left) cannot resolve the connectivity between the exons correctly and three possible assembly scenarios are possible (middle). In contrast, using long-read sequencing the correct combination of transcripts can be resolved unambiguously (right). Scheme modified from Tilgner *et al.* [82].

surface receptor isoforms (*CD37*) using consensus isoforms generated from ONT reads. Furthermore, novel transcriptional start and end sites with CAGE or polyA signals were identified suggesting that the long-read data supplied additional information about the transcriptome. Targeted approaches using pre-amplification of certain genes revealed the existence of almost 8,000 isoforms from only four *Drosophila* genes including the highly complex *Dscam1* gene [85].

In 2018, Garalde *et al.* published the first study on direct RNA-sequencing of the yeast transcriptome using the ONT platform [86]. A high Spearman correlation of transcript expression between cDNA and direct RNA (dRNA) methods as well as Illumina short-read data and dRNA was observed. Using Spike-in RNA Variant Control Mixes (SIRVs) to assess the assembly and splice variant detection, a Spearman correlation of 0.8 was found at the transcript level (SIRV E2 panel). In a final analysis, they showed the ability to

detect RNA modifications in artificially modified sequences with a Hidden Markov Model, which was trained to recognise the difference in electric current produced by a modified base. To further compare cDNA and dRNA methods on the ONT platform, a world-wide RNA consortium generated 10 M direct RNA and 15 M cDNA reads from the human cell line GM12878 with the MinION platform [87]. The average error rate across all samples was 14 % suggesting an improvement from 20-30 % error rates as seen in earlier studies [88, 83]. Workman *et al.* analysed all reads using a pipeline termed FLAIR (Full-Length Alternative Isoform Analysis of RNA, [89]) and identified a set of 30 k high-confidence transcripts, that had not been annotated in GENCODE (v24). In the total set of 50 k isoforms, they found a significantly higher number of isoforms per-gene in non-coding genes than in protein-coding gene. The work on long-read sequencing by Workman *et al.* found a high number of previously unknown transcripts, which suggested that a transcriptome assessment using long-read sequencing would reveal a more comprehensive view of the transcriptome.

In addition to improvements in the ONT chemistry and therefore in read quality, novel software tools such as aligners or genome assemblers have been developed to handle long but erroneous reads with specific error characteristics [90, 91]. However, hybrid tools using both short and long-read sequencing data are currently the most reliable methods and the optional usage of both data types has recently been implemented in the FLAIR and SQANTI pipelines [89, 92].

### 1.3.5 Effects of alternative splicing on proteome complexity

In humans, 95 % of multi-exonic genes are alternatively spliced increasing the transcriptional complexity [93, 94]. A recent study, however, suggested that this increased transcriptome complexity did not result in increased proteome complexity [95]. That report was criticised due to the coverage and sensitivity of mass-spectrometry methods used [96], which constitutes a major challenge for the proteomic assessment of alternative splicing. Furthermore, the probability to detect a part of the alternatively spliced sequence was lower for instance if only a novel splice junction existed. Through use of a more

sensitive method (integrated RNA-sequencing and Ribo-seq), some 75 % of medium to high abundance human transcripts with skipped exons events were found in ribosomes [97], indicating that alternatively spliced transcripts are translated.

## 1.4 Alternative splicing in the thymus and its role in the induction of self-tolerance

Alternative splicing is relevant for self-tolerance, as peptides containing exons and splice junctions of different TRA isoforms may need to be presented to developing T cells in order to achieve complete tolerance. Previous studies of the mTEC transcriptome, the level of alternative splicing in TEC and mechanistic studies using mouse models are summarised below.

### 1.4.1 Complexity of the transcriptome and possible underlying mechanisms of splicing in thymic epithelial cells

A few reports have described splicing in thymic epithelial cells including an initial report by Keane *et al.*, which used previously published RNA-sequencing from mTEC and peripheral tissues as well as microarray data. The authors made use of neonatal mTEC RNA-sequencing data from a report by St-Pierre *et al.* [98], together with RNA-sequencing data from nine mouse tissues that was generated by Merkin *et al.* [65]. Additionally, Keane *et al.* further used microarray data from Giraud *et al.* to estimate the number of exons detected per gene [99]. A second report by St-Pierre *et al.* used deeply sequenced RNA-sequencing data from wildtype and *Aire*-knockout TEC [50]. The most recent study by Danan-Gotthold *et al.* describing alternative splicing and RNA editing in TEC, used a range of publicly available datasets including the peripheral mouse tissues generated by Merkin *et al.* [65], skin epithelial cells and cTEC from St-Pierre *et al.* [98] as well as immature and mature mTEC from Chuprin *et al.* [100]. In addition, Danan-Gotthold *et al.* generated a new RNA-sequencing dataset from *Aire*-knockout mice. In this study, a read depth of 50 M reads was used (80 bp read length) due to the read depth across the samples.

Initially, splicing complexity in TEC was assessed using splicing entropy, a metric which reflects the diversity of transcripts in the sample [101]. The splicing entropy was computed as the Shannon entropy of transcript abundances as previously used by Ritchie *et al.* [102]. In comparison to peripheral tissues, a significantly higher splicing entropy was found in mTEC [101]. Keane *et al.* further observed a larger number of tissue-restricted

isoforms from a variety of peripheral tissues in mTEC compared to cTEC or skin epithelial cells. Furthermore, St-Pierre *et al.* found (i) a higher number of *Aire*-independent TRA existed than *Aire*-induced TRA and (ii) a higher complexity with respect to their splicing in *Aire*-independent TRA [50].

The most recent study describing alternative splicing and RNA editing in TEC revealed a lower representation of TRA genes from the brain and testis in mature mTEC [103]. This observation could possibly be explained by a high complexity of splicing in these tissues [64, 65, 104] or by their status as immunoprivileged tissues [105, 106]. A 'leave-one-out' analysis showed that TEC expressed the highest fraction of TRA across a range of expression values [103]. With respect to splicing, a higher number of junctions per gene was detected for mature mTEC. Furthermore, Danan-Gotthold *et al.* reported the presence of both isoforms, normally restricted to the brain or the haematopoietic system, of the gene *Fyn* in mature mTEC. In this report, TRA junctions were defined as splice junctions that were only found in one tissue, while the expression of the respective gene was detected in at least one other tissue. The percentages of TRA junctions from the peripheral tissues detected in mature mTEC were lower than TRA gene expression (20-60 % detected).

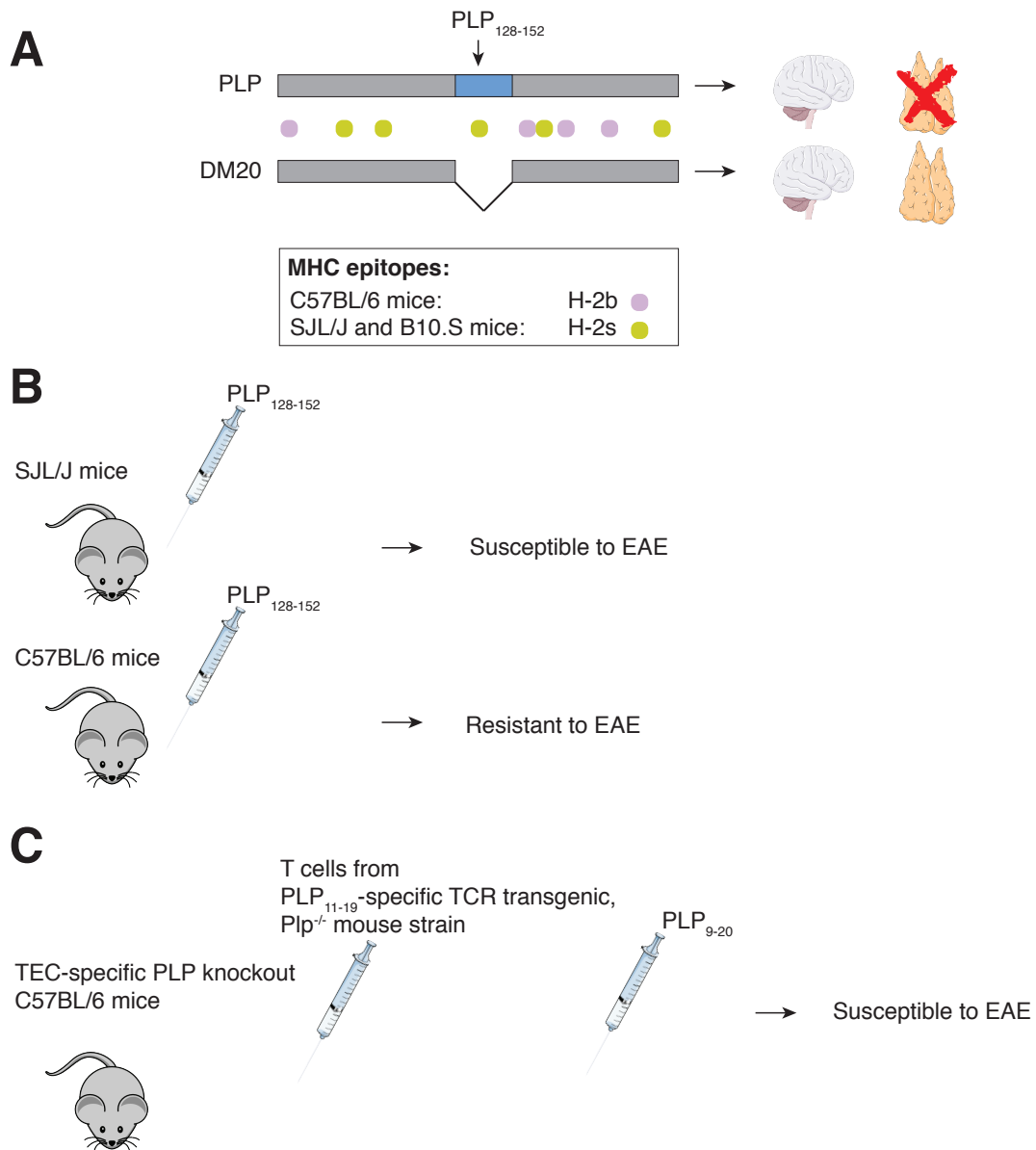
The studies discussed above have also addressed a direct role for AIRE in splicing regulation in TEC. Keane *et al.* found a significantly higher number of exons per gene in wildtype compared to *Aire*-knockout mTEC [101]. Together with the high splicing complexity in mTEC, the authors concluded that AIRE is involved in splicing regulation in mTEC and that splicing in mTEC is a 'promiscuous' process similar to PGE. Independently, Meredith *et al.* identified exons that were upregulated by AIRE without changes in gene expression level using RNA-sequencing data from wildtype and *Aire*-knockout mTEC [41]. In addition, St. Pierre *et al.* observed a higher splicing complexity in wildtype, mature mTEC than in *Aire*-knockout mTEC or other TEC types. In contrast to these findings, Danan-Gotthold *et al.* suggested a limited role for AIRE in the regulation of TRA splicing as the authors observed no significant alterations in the percentage of TRA junctions from peripheral tissues in *Aire*-knockout TEC compared to

mature wildtype mTEC (except for kidney and testis).

Alternative mechanisms underlying the high splicing complexity in TEC could involve the expression of tissue-restricted and non-constitutive splicing factors or the recruitment of splicing factors by PGE mediators such as AIRE or FEZF2. St-Pierre *et al.* reported seven *Aire*-independent mRNA pre-processing factors from a database of RNA-binding proteins [107], which were at least 2-fold higher in mature compared to immature mTEC. The two genes with highest expression in mature mTEC were *Synj2* and *Msi1*, while *Snrpn* showed the most marked difference between cTEC and mature mTEC. The authors suggested that irrespective of AIRE, these other factors were likely involved in splicing the numerous TRA present in mature mTEC. As discussed above, a protein-protein interaction study by Abramson *et al.* identified pre-mRNA processing factors as a major category of proteins that interact with AIRE including eight spliceosome binding proteins [30]. Furthermore, the RNAi screen by Giraud *et al.* identified a large set of genes categorised as related to 'RNA processing / splicing' [36]. For instance SFRS1 (SRSF1) and SFRS3 (SRSF3) were identified as AIRE-binding partners in both *in vitro* studies.

### 1.4.2 The effect of thymic alternative splicing on self-tolerance

Evidence that the absence of specific splice variant(s) in the thymus can result in potentially auto-reactive T cells in the periphery was published in 2000 by Klein *et al.* [108]. The initial observation was a different susceptibility of mouse strains to experimental autoimmune encephalomyelitis (EAE), which is widely used as a model for multiple sclerosis (MS) (Figure 1.6). EAE can be induced by a variety of peptides including purified peptides from for instance myelin proteolipid protein (PLP) or myelin basic protein (MBP) [109]. Mouse strains such as SJL/J or Balb/c mice with the MHC epitope H-2<sup>s</sup> are very susceptible to the induction of EAE, while C57BL/6 mice that carry the H-2<sup>b</sup> MHC epitope are resistant [110].



**Figure 1.6:** Schematic summary of responses to induction of EAE in different mouse strains. **A:** The isoforms structures of PLP in the brain compared to the thymus including the MHC epitopes are shown. **B:** The susceptibility to induction of EAE using a specific PLP peptide in the SJL/J and C57BL/6 mice is shown [108]. **C:** C57BL/6 mice with a TEC-targeted loss of *Plp* showed susceptibility to EAE induction after T cell transfer and injection of PLP peptide [111].

To assess whether central tolerance induction plays a role in susceptibility to EAE, Klein *et al.* first showed by RT-PCR that expression of PLP in the thymus of both mouse strains is comparable, but in both strains restricted to a shorter splice variant named DM20. The observation was also confirmed in a human thymus sample. Testing different PLP peptides in lymph nodes cells from PLP immunized mice *in vitro*, a strong

response to a PLP peptide (PLP<sub>128-152</sub>) was observed for the cells isolated from SJL/J mice only. Comparing the location of the auto-reactive peptide to the location of H-2<sup>s</sup> MHC-presented epitopes revealed a H-2<sup>s</sup>-specific epitope in the exon missing from the shorter DM20 (thymic) splice variant. Together, these observations suggested that the absence of the central nervous system specific exon in TEC resulted in the presence of potentially auto-reactive T cells in the periphery, which can readily react to immunisation with the PLP peptide spanning amino acids 128-152. However, in mouse strains with the H-2<sup>b</sup> MHC haplotype such as C57BL/6 no MHC epitope is present in the spliced region, thus these mouse strains are resistant to the induction of EAE using the peptide PLP<sub>128-152</sub>. In the same year, Anderson *et al.* published a study focussing on the dominant immune response of SJL/J mice to the PLP<sub>139-151</sub> peptide compared to PLP<sub>178-191</sub> [49]. Consistent with Klein *et al.*, they found that the peptide PLP<sub>139-151</sub> was located within the spliced-out exon and that the shorter DM20 splice variant was present in the thymus. Moreover, Anderson *et al.* introduced Ig-PLP<sub>139-151</sub> into thymi of embryos by *in utero* injection resulting in a diminished proliferation of neonate thymocytes upon stimulation with PLP<sub>139-151</sub> *in vitro*. Both reports support the concept that lack of negative selection due to the absence of certain splice isoforms in the thymus can result in the presence of potentially auto-reactive T cells, which can be activated upon stimulation.

Prior to the discovery of the thymic absence of a specific exon in the *Plp* gene, a difference in susceptibility to EAE of the SJL/J and B10.S mouse strains, which share the H-2<sup>s</sup> MHC haplotype, had been observed [112, 113]. One potential explanation was the deficiency of interleukin-12 (IL-12) signaling to induce T<sub>H</sub>1 cells in B10.S mice. T cells extracted from B10.S mice challenged with MBP to induce EAE, were treated with IL-12 *in vitro* resulting in an increased interferon- $\gamma$  production. Segal *et al.* further showed that injection of these cultured cells into B10.S recipient caused EAE in 100 % of cases (although not as severe as in SJL/J mice) [113]. Another study focused on the resistance of B10.S mice to the peptide specific to the missing exon in the thymus (PLP<sub>139-151</sub>). Use of tetramers comprised of MHCII and PLP<sub>139-151</sub>, revealed an increase in PLP-specific CD4<sup>+</sup>CD25<sup>+</sup> T cell numbers in the B10.S mice compared to SJL/J mice [114]. Upon *in*



*in vivo* blockade of CD25 prior to immunization, an increase in *in vitro* proliferative capacity of T cells upon restimulation with PLP<sub>139-151</sub> as well as in EAE symptoms *in vivo* could be observed. However, the symptoms were only observed in one third of anti-CD25 treated mice, thus suggesting that other mechanisms - such as deficient IL12 signaling - are involved in the resistance of B10.S mice to PLP<sub>139-151</sub> induced EAE in addition to the higher numbers of PLP<sub>139-151</sub>-specific regulatory T cells that they possess. The authors further discussed that the higher level of regulatory T cells in B10.S mice could be due to the higher expression level of *Plp* as measured by RT-PCR in the thymi of B10.S mice compared to SJL/J mice.

In addition to research into MS antigens, splicing variants of antigens potentially causing beta cell destruction in type 1 diabetes were observed. In 2006, Dogra *et al.* reported the absence of full-length transcripts in *G6PC2*, the gene encoding for islet-specific glucose-6-phosphatase catalytic subunit-related protein (IGRP), in human thymus samples by RT-PCR (n=9) [115]. The matching pancreas samples from all individuals showed the presence of the full-length transcript. Most thymus samples also lacked several of the five major pancreatic splice variants assessed which were consistently detected in matching pancreas samples. However, this study was based on frozen whole thymus tissue, and therefore it remained unclear whether gene expression in mTEC or thymocytes (which greatly outnumber mTEC in the thymus) was assessed.

To further investigate whether splice variants in *G6PC2* could contribute to the development of type 1 diabetes, de Jong *et al.* assessed the splice variants of this gene present in pancreatic islets and the thymus. All of the known splice variants were detected in the thymus samples and, additionally, several novel and potentially immunogenic islet-specific isoforms were identified [116]. However, assessing auto-antibody reactivity against IGRP variants and presence of IGRP-specific CD8<sup>+</sup> T cells in diabetes patients compared to healthy individuals did not reveal a significant difference. Thus, the authors concluded that self-reactive T cells against IGRP were not necessarily involved in disease onset.

An additional diabetes-related gene, namely IA-2, was found to have differential alternative splicing between the thymus and pancreas as shown by RT-PCR using tissue from healthy individuals [117, 118]. The absence of full-length IA-2 transcripts and transcripts lacking exon 14 were shown in the thymus. While the RT-PCR was based on whole thymus tissue and could therefore be representative of thymocytes, immunohistochemistry was also shown for thymus section confirming that the expression originated from mTEC. Several antibody and T cell epitopes found in diabetes patients could be mapped to exon 13 and exon 14, suggesting that the absence of these exons might be involved in the development of type 1 diabetes.

## 1.5 Open questions and hypotheses

A comprehensive description of the splicing landscape in TEC was missing at the start of this project. While the comprehensive representation of gene expression was shown in TEC, it remained unclear how completely the peripheral transcriptome is recreated in TEC at the isoform level. It was further unknown whether the structures in TEC mirror the ones observed in the periphery or whether TEC produce a high number of novel transcripts. Another open question, that could have been addressed with recently developed single-cell RNA-sequencing datasets, was whether different transcript structures of the same gene were produced by the same single mTEC. It was proposed by Keane *et al.* that splicing is a promiscuous process regulated by AIRE [101], but another recent report found the role of AIRE in splicing regulation to be limited [103]. Although splicing factors were observed as AIRE binding partners, so far, only *in vitro* knockdown studies of splicing factors had been performed. Therefore, the role of alternative splicing factors in TEC *in vivo* remains undefined. Finally, mTEC produce a high number of epitopes for the presentation on MHCII surface molecules to developing T cells. A limited number of examples of thymic alternative splicing with an effect on T cell selection were reported, but the extent to which such MHCII-restricted epitopes are produced from alternative splice sites remains unknown.

Given previous descriptions of splicing in TEC, I hypothesized that splicing could either be a promiscuous process or that it could have an important role for TEC function and be tightly regulated to achieve a high complexity transcriptome of a specific composition. An indication for the latter would be for instance the expression of a set of *Aire*-independent, tissue-restricted splicing factors, which would suggest that the 'directed' expression of these splicing factors is required for shaping the splice isoform landscape in TEC. Along with this hypothesis, the direct influence of AIRE on splicing would be expected to be small. Based on the transcriptome complexity in TEC, it could be hypothesized that the absence of a TRSF in TEC would result in defects in TEC biology and/or in the absence of spliced transcripts and the epitopes produced from such splice variants.

## 1.6 Aims

The presented study had five main aims. First, matched, deep RNA-sequencing from mTEC and peripheral mouse tissues was used to generate a common reference transcriptome in order to facilitate comparison of splicing between mTEC and peripheral tissues (Chapter 3). The identified TEC-specific novel transcripts were validated using long-read RNA-sequencing from mTEC. Second, based on the common reference transcriptome, I aimed to elucidate the splicing landscape in mTEC including representation of tissue-specific splice isoforms and the existence of novel transcript structures in mTEC (Chapter 4). Third, the hypothesis regarding promiscuous splicing and AIRE's role in splicing were addressed using RNA-sequencing from an *Aire*-knockout mouse model (Chapter 5). Fourth, I aimed to identify splicing factors that might be involved in the regulation of alternative splicing in mTEC and suggest potential mechanisms for splicing regulation in TEC (Chapter 5). Finally, from the assessment of splicing factor expression in TEC a candidate factor was prioritised and probed in a TEC-specific conditional knockout mouse model (Chapter 6). The *in vivo* model was used to analyse whether the splicing factor is indispensable for TEC development, for TEC maturation, for PGE and for the splicing landscape in TEC (Chapter 6).



# 2

## Methods

### Contents

---

|            |                                                                                                              |           |
|------------|--------------------------------------------------------------------------------------------------------------|-----------|
| <b>2.1</b> | <b>Computational methods</b>                                                                                 | <b>32</b> |
| 2.1.1      | Sources of Illumina population RNA-sequencing datasets                                                       | 32        |
| 2.1.2      | Processing of Illumina RNA-sequencing datasets from ENCODE mouse tissues and mouse thymic epithelial cells   | 33        |
| 2.1.3      | mT&T assembly pipeline                                                                                       | 33        |
| 2.1.4      | Processing of RNA-sequencing data from thymic epithelial cells with a loss of <i>Rbfox1</i> or <i>Rbfox2</i> | 37        |
| 2.1.5      | Quantification and differential expression analysis at the transcript and gene level                         | 37        |
| 2.1.6      | Splicing tools                                                                                               | 40        |
| 2.1.7      | Single-cell RNA-sequencing                                                                                   | 41        |
| 2.1.8      | Long-read RNA-sequencing using Oxford Nanopore Technology                                                    | 44        |
| 2.1.9      | Statistical analysis and visualisation of processed data                                                     | 46        |
| <b>2.2</b> | <b>Experimental methods</b>                                                                                  | <b>49</b> |
| 2.2.1      | Mice                                                                                                         | 49        |
| 2.2.2      | Genotyping and assessment of genomic excision in conditional knockout mice                                   | 50        |
| 2.2.3      | Isolation of thymic epithelial cells                                                                         | 51        |
| 2.2.4      | Flow cytometry                                                                                               | 52        |
| 2.2.5      | RNA-sequencing                                                                                               | 55        |
| 2.2.6      | Immunohistochemistry                                                                                         | 55        |

---

## 2.1 Computational methods

All computational analyses were performed on the High-Performance Computing (HPC) cluster at the Kennedy Institute of Rheumatology. Pipelines were written using the CGAT workflow management system [119] ([github.com/cgat-developers/cgat-core](https://github.com/cgat-developers/cgat-core)) and ruffus [120]. Selected python functions and classes from the CGAT code collection were also used for custom scripts. All software was run with default settings unless otherwise indicated. The software and R packages including version numbers are given in Appendix Tables A.1 and A.3.

### 2.1.1 Sources of Illumina population RNA-sequencing datasets

Illumina RNA-sequencing data from four mTEC populations (as analysed in Chapters 3, 4 and 5) was generated by Dr. Noriko Shikama-Dorn, Dr. Saulius Zukyls and Katrin Hafen (University of Basel). For this dataset, the details on the populations sorted are given in Appendix Table B.1. The Fluorescence-activated cell sorting (FACS) gating schemes were used as previously described in Sansom *et al.* [27].

In brief, two biological replicate ribosome-depleted, polyA+ RNA-sequencing libraries were generated from different mTEC populations (1 µg of total RNA; 4-6 weeks old mice). The library preparation was performed using a custom dUTP-based protocol as implemented by the Wellcome Trust Centre for Human Genetics and 101 bp paired-end stranded RNA-sequencing was performed on Illumina HiSeq 2000. In addition to the mTEC samples, RNA-sequencing data for 22 mouse tissues were obtained from the mouse ENCODE project (GSE36025; 8 weeks old mice). For splicing factor expression analysis two additional datasets were included, which were RNA-sequencing from skin epithelial cells and cTEC. The skin epithelial cells of newborn mice were sequenced at 100 bp and paired-end [98] (GSM1094285). The cTEC dataset from 1 week old mice was sequenced at 36 bp and single-end [27] (GSE53111). The datasets for skin epithelial cells and cTEC were not processed with the other peripheral tissues and were not included in the transcriptome assembly (Section 2.1.3).

### 2.1.2 Processing of Illumina RNA-sequencing datasets from ENCODE mouse tissues and mouse thymic epithelial cells

The quality of raw reads was initially examined using the *readqc* pipeline from the CGAT apps code collection ([github.com/cgat-developers/cgat-apps](https://github.com/cgat-developers/cgat-apps)). The reads of the population data were trimmed to 76 bp to match the ENCODE dataset (*fastx\_trimmer*; [hannonlab.cshl.edu/fastx\\_toolkit/](http://hannonlab.cshl.edu/fastx_toolkit/)). The trimmed population data was further processed using the *scseq* pipeline ([github.com/snsansom/scseq](https://github.com/snsansom/scseq)). The in-house pipeline consisted of strand-aware mapping by *HISAT2* [121] (mouse genome plus known splice junctions; Ensembl mm10-v91; settings: '*--dta --score-min L,0.0,-0.2 --rna-strandness RF*'), calculation of post-mapping quality metrics by *Picard* and gene expression quantification using *featureCounts* and *Salmon*. Post-mapping quality metrics inspected included for instance number of mapped reads, number of read pairs, percentage of aligned high-quality reads, number of genes expressed, 3' bias, duplication rate and fraction of spliced reads (summarised in Appendix Table A.2). Transcripts-per-million (TPM) were generated with *Salmon* in paired-end mode with an index created from Ensembl annotations (mm10-v91; other settings: '*--gcBias*'). To test for differential expression between cell types, counts per gene were computed using *featureCounts* (default settings). For the initial quality assessment, *Cufflinks* quantification was included (default settings). After inspecting the individual Illumina RNA-sequencing datasets, for each tissue (except cTEC and skin epithelial cells) a high-depth sample at 200 million (M) reads was generated by merging all reads from the respective tissue and subsampling to 200 M reads (*samtools* view *-s* option; [122]). Furthermore, the read depth at which two replicates per tissue could be generated was determined to be 60 M reads (Appendix Tables B.2 and B.3). Thus, two replicates indicated as replicate A and B in Appendix Table B.3 were generated by first combining the indicated replicates per tissue into bio-replicate pools (where necessary) and subsequently randomly downsampled to 60 M reads.

### 2.1.3 mT&T assembly pipeline

To compare splicing between the periphery and TEC, a high-confidence assembly of the transcriptome from **mouse TEC** and **Tissues (mT&T)** was generated (pipeline

overview in Figure 3.1). First, uniquely-mapped and non-duplicated reads (*Picard*) were selected from each sample. Second, reference-guided assemblies of all individual tissues were generated using the high depth 200 M read samples. A strand-specific *stringTie* assembly (`'--rf'`) was used in reference-guided mode with Ensembl (mm10-v91) as reference. Subsequently, the per-tissue assembly was used as an index for quantification with *Salmon* in the same tissue (settings: `'--incompatPrior=0 --validateMappings --rangeFactorizationBins=4 --seqBias --gcBias -x 0.66'`) of the two 60 M read bio-replicate pools [123]. Following the quantification, a non-parametric Irreproducibility Discovery Rate (npIDR) was calculated based on a modified script from Pervouchine *et al.* [70]. The lower and upper 5% quantiles of the log-normal distribution of TPM values were merged into single bins due to data sparsity, the TPM values were binned and the npIDR was calculated as described previously [124, 70]. For each tissue or TEC sample an IDR curve was used to determine a TPM threshold (at  $IDR \leq 0.1$ ), that was used to filter the transcripts (Figure 3.2). The filtered transcripts from all tissue and TEC samples were merged into a new reference transcriptome using the Ensembl reference annotation without further filtering (*stringTie merge*; settings: `'-F 0 -T 0 -f 0.0'`). The *stringTie merge* filters for TPM (`-T`), FPKM (`-F`) and minimum isoform fraction (`-f`) were disabled to avoid further filtering by *stringTie merge* as the transcripts were already filtered in the npIDR process. To assess how the transcripts in the obtained mT&T assembly related to the reference annotation, the transcripts were compared to the Ensembl annotation (mm10-v91) using *gffcompare*. The tool *gffcompare* assigned class codes to each transcript. The definitions for each class code are given in Table 3.1 corresponding to [ccb.jhu.edu/software/stringtie/gffcompare.shtml](http://ccb.jhu.edu/software/stringtie/gffcompare.shtml). The class codes were used to exclude likely non-relevant, for instance likely non-coding, transcripts from the assembly for downstream analysis. Such excluded transcripts were located on the opposite strand (exonic [`'x'`] or intronic [`'s'`] overlap) or were located fully within a reference intron [`'i'`]. Furthermore, transcripts in repetitive regions [`'r'`], transcripts that represented polymerase run-on fragments without any overlap [`'p'`] or pre-mRNA fragments [`'e'`] were removed from the mT&T assembly. Transcripts with these class codes were excluded due to their strandedness [`'x'`, `'s'`], which could result from reads being mis-mapped to the opposite



strand or non-coding anti-sense transcription. Additionally, the genomic location ['i', 'r'] could be a reason for exclusion, as for instance correctly assembling transcripts in repetitive regions is challenging. Moreover, assembled transcripts resulting from possible artefacts such as pre-mRNA, genomic DNA contamination or polymerase run-on events ['p', 'e'] should be excluded.

The class codes were used to define novel transcripts, which could either be overlapping with the reference on the same strand ['o'] or extending the reference ['k']. Transcripts with multiple exons and at least one junction overlapping the reference were also categorised as novel transcripts ['j']. Finally, novel transcripts generated by retained introns were included ['m', 'n']. The novel transcripts (class code: ['jkmno']) were renamed to their corresponding Ensembl transcript ID plus a suffix *\_NX*, where *X* is the number of novel transcripts for the respective Ensembl transcript ID. Unknown or intergenic transcripts ['u'] and transcripts containing reference transcripts within introns ['y'] were also included and the *stringTie merge* names used.

The final version of the mT&T assembly was used as the transcriptome annotation for all subsequent steps if not otherwise indicated.

### 2.1.3.1 Comparison of tissue assemblies

To compare the individual tissue assemblies, the harmonic mean between sensitivity and precision termed F1 score was used (Equation 2.1, Figure 3.4A). The sensitivity and precision were calculated by *gffcompare* at the transcript, exon and base level. The sensitivity describes the fraction of transcripts that can also be found in the reference annotation, while precision is the fraction of 'true' transcripts out of all transcripts in the assembled dataset.

$$F1score = 2 \times \frac{precision \times sensitivity}{precision + sensitivity} \quad (2.1)$$

### 2.1.3.2 Multi-gene transcripts in the mT&T assembly

The inspection of transcripts from the mT&T assembly revealed some multi-gene transcripts and overlapping loci on opposite strands, which were further investigated. Two is-

sues related to this observation have been open on the *stringTie* github page ([github.com/gpertea/stringtie/issues/178](https://github.com/gpertea/stringtie/issues/178), [github.com/gpertea/stringtie/issues/19](https://github.com/gpertea/stringtie/issues/19)). A fusion event was also observed with *stringTie* by Venturini *et al.* [125].

To identify multi-gene transcripts and to understand the connectivity of exons in the genome, a network graph containing all connected exons was generated (Figure 3.6). As a first step, custom Python scripts were used to identify all connected exons either within the Ensembl annotation or within the mT&T assembly. From each transcript all connected exons were summarised as connected exon pairs in the respective GTF file, using the CGAT *GTF* class ([github.com/cgat-developers/cgat-apps](https://github.com/cgat-developers/cgat-apps)). Finally, a network graph was built from the connected exons (Python package *networkX*) [126]. Here, every pair of connected exons in a transcript was treated as a pair in the graph. Thus, every gene or gene locus with no connections to other genes would be represented by one individual graph (Figure 3.6A, C).

To identify novel multi-gene transcripts, I classified all novel transcripts in graphs that contained one or more Ensembl genes in the mT&T graphs as novel multi-gene transcripts. It is known that some annotated transcripts connect more than one gene as discussed in Section 3.2.3.3.

### 2.1.3.3 Discovery of novel transcripts from the mT&T assembly

To analyse novel transcripts across the tissue and mTEC samples and to avoid over-representation of similar tissues, the tissues were first grouped based on the abundances of most highly variable transcripts in protein-coding genes (excluding novel transcripts; Section 2.1.9.1). For groups of similar tissues one representative tissue was chosen as indicated in Appendix Figure C.1. The upper-quartile normalized,  $\log_2(n+1)$ -transformed TPM values of novel transcripts (*Salmon*) were used to calculate *tau* (Section 2.1.9.3) in order to classify the novel transcripts according to their tissue (group) of origin. Additionally, a filter for  $> 2$ -fold difference to the tissue with the next highest expression of the novel transcript was included in order to conservatively define tissue-restricted novel transcripts as shown in Figure 4.7. A geneset over-representation analysis was performed on the known protein-coding genes for which I discovered novel transcripts

specific to mature mTEC (protein-coding genes with TPM > 0.1 in these cells were used as the 'universe'; Section 2.1.9.2).

#### 2.1.4 Processing of RNA-sequencing data from thymic epithelial cells with a loss of *Rbfox1* or *Rbfox2*

The paired-end, 150 bp RNA-sequencing data generated from immature and mature mTEC from *Rbfox1*<sup>lox/lox</sup>:FoxN1-cre<sup>+/-</sup> and *Rbfox2*<sup>lox/lox</sup>:FoxN1-cre<sup>+/-</sup> and the respective wildtype counterparts were processed using the *readqc* and *scseq* pipelines as described in Section 2.1.2 (additional *HISAT2* setting: '*--no-softclip*'). After quality assessment (Figures 6.14 and 6.15), all 12 samples were downsampled to a read depth of 40 M reads. Differential gene expression analysis was performed between the single-knockout and respective wildtype samples using *featureCounts* gene-level quantification (settings: '*-p -s 2*') and *DESeq2* (settings: size-factor normalisation using '*shorth*'; dispersion fit using '*local*') [127, 128]. The differentially expressed genes are shown in Figure 6.17 (Appendix Tables E.4 and E.5) and characterised in Table 6.1.

#### 2.1.5 Quantification and differential expression analysis at the transcript and gene level

##### 2.1.5.1 Gene and transcript expression quantification

*Salmon* was used in quasi-mapping mode with either an index (k=31) created from Ensembl annotation (mm10-v91) or from the new mT&T assembly to quantify TPM. Gene expression was calculated by aggregating the transcript TPM values for each gene. These TPM values were subsequently upper-quartile normalized and log2(TPM+1)-transformed [129]. To assess the fraction of isoforms per gene, the average gene-level TPM of two replicates at 60 M reads was calculated for multi-isoform genes. Per protein-coding gene, the number of expressed isoforms was divided by the total number of isoforms for the respective gene to obtain the fraction of isoforms per gene. Within each gene, the isoforms were additionally stratified into known (included in Ensembl mm10-v91) or novel (in the mT&T assembly) isoforms. The mean of the obtained fraction for genes expressed above a TPM filter (average of all tissue TPM filters determined by npIDR) is shown in Figure 4.2.

### 2.1.5.2 Differential gene expression including definition of *Aire*-regulated genes

Compared to previous reports, the RNA-sequencing data for mTEC samples used here comprised of longer reads (100 bp), stranded, paired-end and deeper sequenced (170 M reads). Therefore, I redefined *Aire*-regulated genes using the Illumina RNA-sequencing dataset and Ensembl annotation (v91). The untrimmed, full read depth data for the four mTEC populations were mapped as above. Reads were counted by *featureCounts* (settings: '-p -s 2') for differential expression analysis with *DESeq2* (settings: size-factor normalisation using '*shorth*'; dispersion fit using '*local*'). *Aire*-regulated genes were defined as more than 2-fold and significantly (Benjamini-Hochberg adjusted p-value  $< 0.05$ ) downregulated in the two homozygous *Aire*-knockout samples compared to heterozygous *Aire*-knockout (Appendix Figure B.5).

Sansom *et al.* had previously identified 3,980 genes as induced or positively regulated by *Aire* (Ensembl v67), out of which 3,862 were protein-coding according to Ensembl v91. In contrast, I identified 5,059 protein-coding genes as significantly and more than 2-fold downregulated in *Aire*-knockout mTEC compared to *Aire*-positive mTEC (Appendix Figure B.6). In comparison to the previous definition, I classified an additional 1,462 genes as *Aire*-regulated but did not identify 265 genes as *Aire*-regulated.

An alternative method to identify differentially expressed genes was to use *Kallisto* and *Sleuth* [130, 131]. These tools were developed for transcript quantification, but *Sleuth* has been adapted to define differentially expressed genes [132]. The *Sleuth* gene-level approach was used in order to compare the numbers of *Aire*-regulated transcripts (Section 2.1.5.3) and *Aire*-regulated genes using the same method (Figure 5.4B and D, Appendix Figure D.3).

### 2.1.5.3 Differential transcript expression including definition of *Aire*-regulated transcripts

Differential transcript expression between *Aire*-knockout and *Aire*-positive mTEC was assessed by two different methods, namely *DESeq2* and *Sleuth*, and compared to the

*Aire*-regulated genes defined by the respective method (Figure 5.4). For the *DESeq2* approach, transcript counts were obtained from *featureCounts* (run including a genome fasta; settings: `'-p -s 2'`). Different size-factor normalisations and dispersion fits were tested for *DESeq2*, and the settings `'shorth'` and `'local'` were chosen for size-factor normalisation and dispersion fit, respectively. To define *Aire*-regulated transcripts, the same filters were applied as for *Aire*-regulated genes above (2-fold downregulated in *Aire*-knockout mTEC, Benjamini-Hochberg adjusted p-value  $< 0.05$ ).

The alternative approach to identify *Aire*-regulated genes included running the transcript quantifier *Kallisto* with 1 k bootstraps to obtain transcript abundances. These were further analysed using the R package *Sleuth*. A Wald Test was used to test for differentially abundant transcripts. However, the output of *Sleuth* is not a fold change but a metric called 'b', which was described as corresponding to a fold change on the natural log scale. In order to determine the value of 'b' that corresponds to a log2 fold change, I generated a linear model based on all significant 'b' values and fold changes calculated from the *Kallisto* transcript quantification, which means only transcripts that were significantly differentially expressed are included in the model (Appendix Figure D.4). The model was additionally only fitted between a 'b' of -20 and +20 due to the spread of the fold changes outside these boundaries. The model was then used to determine 'b' at a log2-fold change of 2 (threshold used in *DESeq2* analysis).

For both methods, the differentially expressed transcripts were overlaid with transcripts in house-keeping genes, which were not expected to be differentially expressed. From a set of human house-keeping genes defined by de Jonge *et al.* [133], 468 1-to-1 orthologues were found in the mouse genome. Zero or four of these genes contained differentially expressed transcripts in the *DESeq2* and *Sleuth* approaches, respectively (Appendix Table D.3). Furthermore, for both approaches a geneset over-enrichment analysis (Section 2.1.9.2) was performed for the protein-coding genes containing *Aire*-regulated transcripts, but that were not defined as *Aire*-regulated genes, and the resulting over-enriched categories are presented in Appendix Figure D.6.

## 2.1.6 Splicing tools

### 2.1.6.1 Splice junction estimation

In order to estimate the saturation of splice junction (SJ) detection (Section 4.2.1), SJ were quantified from BAM files using *SJcounts* (settings: `'-maxnh 1 -read1 0 -read2 1'`) [70]. Reads crossing multiple SJ were split using a custom Python script and the counts added to the individual SJ. To map the SJ to genes, an intersection with protein-coding genes (Ensembl mm10-v91) was performed by generating 1 bp intervals from each site of the SJ, which were then intersected with protein-coding genes (*bedtools window*, [134]). The above process was repeated for a range of read depths and the total number of SJ in protein-coding genes with a minimum of 5 reads coverage is reported in Figure 4.1.

For Figure 4.3 the number of unique, annotated SJ was determined by using the annotation status definition by the *ipsa* pipeline, which is based on *SJcounts* ([github.com/pervouchine/ipsa-full](https://github.com/pervouchine/ipsa-full); including *maptools*). This approach was chosen as it determined an annotation status for each pair of splice sites, which takes into account the annotation status of both splice sites. Three steps of the *ipsa* pipeline including read aggregation, read annotation compared to a given reference (Ensembl mm10-v91) and strand definition were run (*aggregate reads*, *annotate reads*, *strand-specific reads*). Here, SJ were included, where both splice sites were annotated, but the intron was either not annotated (annotation code = 2) or was annotated (annotation code = 3).

### 2.1.6.2 SUPPA for annotation of splice event types and quantification of percent spliced-in values

A transcript can generate a variety of splicing events such as skipped exons or retained intron. To assess the type of splicing event generated by a transcript, the annotation function (*generateEvents*) from *SUPPA* was used to analyse the transcript structures for local alternative splicing events (settings: `'-f ioe -e SE RI MX FL SS'`) [135]. This tool was applied to examine the local alternative splicing event produced by novel transcripts in the mT&T assembly (Figure 4.7B) or the assembly from filtered Oxford Nanopore Technology (ONT) reads (Figure 3.10A).

Additionally, the percentage spliced-in (PSI) for skipped exons (Figure 4.6) was computed using the function *psiPerEvent* in *SUPPA* (using events mode with *Salmon* quantification).

### 2.1.6.3 Identification of differential splicing events using *rMATS*

To identify differential splicing events between two samples, the data was re-mapped without softclipping as *rMATS* required reads with identical read length (*HISAT2* with additional setting: `'--no-softclip'`). Subsequently, *rMATS* was run between two samples with two 60 M replicate pools each (default settings with strand setting `'fr-firststrand'`). The output was filtered for  $\text{FDR} < 0.05$  and for  $\text{delta PSI} > 0.2$ . The comparison between the different mTEC populations are shown in Figure 5.1A and Appendix Table D.1.

Differential splicing analysis was also performed between three 40 M replicates of mature mTEC for *Rbfox1* and *Rbfox2* single-knockouts compared to the respective wildtype sample using *rMATS* (Figure 6.18). In addition, a comparison between both wildtype samples was performed. The *rMATS* output was filtered for  $\text{FDR} < 0.05$  and absolute  $\text{delta PSI} > 0.2$ . The genes containing differential skipped exon events were compared to a published list of predicted Rbfox targets (Figure 6.19A, [73]). For the over-enrichment analysis, the original list of genes used for the prediction model of Rbfox targets was not available as a background list. Therefore, I limited the analysis to genes reported to have cassette exons ([zhanglab.c2b2.columbia.edu/index.php/Cortex\\_AS](http://zhanglab.c2b2.columbia.edu/index.php/Cortex_AS)) that were expressed in mTEC. Equivalent to geneset over-enrichment analysis as detailed in Section 2.1.9.2, the over-enrichment for predicted Rbfox targets in genes with skipped exon events from each single-knockout was tested using Fisher's Exact tests. Furthermore, for genes containing differentially skipped exon events geneset over-enrichment was assessed (Section 2.1.9.2, Figure 6.19).

### 2.1.7 Single-cell RNA-sequencing

Single mature mTEC ( $\text{CD45}^-$ ;  $\text{EPCAM}^+$ ;  $\text{Ly51}^-$ ;  $\text{UEA-1}^+$ ;  $\text{MHCII}^{\text{high}}$ ) were sorted from 4-6 weeks old mice by Stefano Maio into three 96-well plates (details in Section 2.2.3). SmartSeq2 libraries [136] were generated by Dr. Annina Grädel and Dr. Iain Macaulay and were sequenced on Illumina HiSeq 2500 (125 bp paired-end; GSE114713 [137]).

### 2.1.7.1 Processing and post-mapping quality assessment of single-cell RNA-sequencing data

The single-cell data were pre-processed from CRAM files using the *cram2fastq* pipeline with quality trimming (minimum quality: 20; minimum length: 50 bp; [github.com/snsansom/scseq](https://github.com/snsansom/scseq)). The data was processed using the *scseq* pipeline as described for the population data above (settings for *HISAT2*: `'--dta --score-min L,0.0,-0.2'`; mm10). Ensembl v83 was used as annotation for the initial quality assessment (Appendix Figures D.7, D.8 and single-cell pooling outlined in Section 2.1.7.3), while for all other analysis Ensembl v91 was used. Cells were filtered for several quality metrics including fraction of reads mapping to ERCC spike-ins  $< 0.5$ , number of pairs  $> 50,000$  and number of genes expressed (*Cufflinks*)  $> 2500$  genes. In addition, fraction of spliced reads  $> 0.05$ , duplication rate  $< 0.5$ , 3' bias  $< 1.3$  and percentage of high-quality reads aligned  $> 0.7$  were applied. Applying these filters to all 285 cells, resulted in 201 single mature mTEC passing quality control (Appendix Figure D.7).

To assess the overall quality of the single cells, a principal component analysis (PCA) of quality metrics was performed. The first 3 principal components (PC) were examined in more detail (Appendix Figure D.8). PC1 and PC2 split the cells failing quality assessment from the ones that were passing the filters. The loadings for PC1 were mainly quality metrics with respect to number of genes expressed (Appendix Table D.4), while PC2 was loaded by total number of reads or number of pairs. 3' bias in addition to read number distinguished cells across PC3.

### 2.1.7.2 Gene expression quantification, normalisation and clustering of single-cell RNA-sequencing data

The quantification of gene expression in single cells was performed using count data from *featureCounts* (Ensembl v91). The count data were normalised using the R package *Seurat* [138], where the total number of counts and cell cycle phase were regressed out using the function *ScaleData*. The cell cycle scores and phase were determined using the *Seurat* function *CellCycleScoring*, which used gene lists for S-phase and G2M phase. The PCA was performed on the most highly variable genes ( $n=2,101$  genes). After inspection of the



scree plot (Figure 5.11D), PC1 to PC20 were selected for the subsequent t-Distributed Stochastic Neighbour Embedding (t-SNE, perplexity = 30, Figure 5.11E). The clusters were defined using the *Seurat* approach of building a shared nearest neighbour (SNN) graph from 20 PCs and grouping cells via Louvain clustering (implemented in *FindClusters* function). The resolution parameter was set to 1. Cluster markers were identified using *FindAllMarkers* testing the differences between clusters with a Wilcoxon Rank Sum test. All markers were filtered for an adjusted p-value < 0.05 (Benjamini-Hochberg correction). The top 15 cluster markers for each cluster are shown in Figure 5.12.

### 2.1.7.3 Assessment of differential splicing in single-cell *in silico* pools using *MISO*

To assess the effects of tissue-restricted splicing factor (TRSF) expression in single mature mTEC, an analysis workflow based on *in silico* construction of pools of single mTEC was developed (Appendix Figure D.9A). Firstly, all single cells were subsampled to a common read depth to minimise potential coverage differences (3 M reads per cell; *samtools view* with '-s' option). Secondly, two types of *in silico* pools were generated using *samtools merge*. The single cells were either stratified by expression of a TRSF, *Aire* or other tissue-restricted antigens (TRA, n=50 cells) or 50 randomly selected cells were merged into each of a total of 2 k 'random pools'. 3 k runs of *MISO* [139] with unique combinations of two random pools were performed in order to simulate a control distribution of differentially spliced events (Figure 5.14A). Alternatively, *MISO* was run between the pool with high expression of the selected gene (gene<sup>+ve</sup>) compared to the pool with low/no expression of the respective gene (gene<sup>-ve</sup>). To obtain a more robust estimate of the difference between the gene<sup>+ve</sup> and gene<sup>-ve</sup> pools, in total 10 gene<sup>-ve</sup> pools were generated and the median of the 10 iterations was used (Appendix Figure D.9C). Due to the high frequency of *Aire* expression in the single mature mTEC, the *Aire*<sup>+ve</sup> pools contained cells with the highest expression of *Aire*, whereas *Aire*<sup>-ve</sup> pools were generated from cells with the lowest expression of *Aire*. Additionally, only 44 *Rbfox1* and *Rbfox2* double-positive mTEC were present, thus six single-positive cells were used for the *Rbfox1/Rbfox2* double-positive pool. Similarly, three cells with low expression of either *Rbfox1* or *Rbfox2* were added to the negative/low pool.

Finally, the *MISO* output was filtered for a difference in PSI of  $> 0.1$  and, following practise established by Katz *et al.* a Bayes factor  $> 5$  [139]. The latter was described as a confidence estimate for the occurrence of differential regulation between the two samples. For an event to be included into the model, at least one read for inclusion or exclusion plus a minimum of five reads for the overall event were required. The results of the comparison between TRSF and TRA pools for individual alternative splicing events are shown in Figure 5.14C.

## 2.1.8 Long-read RNA-sequencing using Oxford Nanopore Technology

### 2.1.8.1 Quality assessment and quantification of long-read sequencing data

Long-read sequencing of immature and mature mTEC was performed using the Oxford Nanopore Technology platform *MinION* (Section 2.2.5.2, Appendix Table B.4). The basecalling was performed using *Albacore* and the passing reads (quality score  $> 7$ ) were trimmed using *Porechop* (default settings, [github.com/rrwick/Porechop](https://github.com/rrwick/Porechop)). Both steps were performed by Hannah Roberts and Duncan Parkes (Wellcome Trust Centre for Human Genetics). The subsequent steps were performed in a custom ONT-specific pipeline: for each sample, the trimmed reads were merged into one FASTQ file and mapped against the genome using *Minimap2* (settings: `'-ax splice'`, reference genome: mm10) [90]. The BAM file was filtered for primary alignments with a minimum mapping quality of 20. As part of the pipeline, the number of splices per read was calculated from the BAM files using the CIGAR string (Appendix Figure B.7B). The quality of the FASTQ and BAM files was assessed with *NanoPack* [140](Appendix Figure B.7C). As a measure of error rate, the gap-compressed sequence divergence of each alignment was extracted from *Minimap2* (tag 'de', Appendix Figure B.7D). Read counting at the gene and transcript level was performed using *featureCounts* (settings: `'-L --fracOverlap 0.8'`).

### 2.1.8.2 Transcriptome assembly from long-read sequencing data

To generate an assembly from long-read RNA-sequencing data, two approaches were tested. Initially, *stringTie* was updated to v1.3.5 for the ONT analyses. In this release one default setting changed and therefore the setting `'-M 0'` was used to match the default of

v1.3.3b. In the first assembly approach, the transcriptome was assembled from all ONT reads using a reference-guided *stringTie* assembly (Figure 3.8). Inspecting the assembled transcripts revealed that a large fraction of novel transcripts had extended untranslated regions (UTR) and was therefore not interesting with respect to novel transcripts that could potentially produce novel epitopes.

To limit the assembly to transcripts from novel isoforms containing novel SJ, an initial read filtering step was included to identify informative reads (Figure 3.9). The filtering was based on assessing the potential of a read to generate a novel transcript. The tool *paftools* (part of *Minimap2*) was used to annotate splice junctions in a mapped read with a known or novel ('false positive') status. All reads with at least one known and one novel junction (10 bp apart from another annotated splice junction; custom Python scripts) were further filtered to only retain reads containing one splice junction that was also detected in the Illumina data from mature mTEC (using custom Python scripts and *SJcounts* output from Illumina data). A 5 bp window around the splice junction was allowed to account for errors in the ONT data. Finally, the filtered reads were subjected to a *stringTie* assembly (settings: `'-M 0 -c 0.001 -f 0'`). The coverage filter (`'-c'`) was decreased from 2.5 (default) to a low value (0.001) to avoid additional filtering by *stringTie*. To inactivate filtering for isoforms with high isoform fraction, the parameter `'-f'` was set to 0. To obtain mappings of novel transcripts to genes, the assembly was compared to the Ensembl (mm10-v91) annotation using *gffcompare*. This strategy identified novel transcripts in 5.6 k protein-coding genes in mature mTEC (Figure 3.9).

To validate novel transcripts identified in the mT&T assembly, the ONT dataset was quantified using *featureCounts* with the mT&T assembly as annotation (Figure 3.12A). Non-validated novel transcripts with high expression in Illumina data, were frequently novel transcripts containing retained introns (defined by SUPPA, Section 2.1.6.2) and novel transcripts, where overlapping loci on the opposite strand were observed. These two categories were removed from the set of novel mT&T transcripts. For the quantification of unstranded data, overlapping loci on opposite strands might be challenging, hence a

set of such opposite strand overlapping loci was identified as outlined below. All exons from the mT&T assembly and Ensembl annotation (v91) were selected and the resulting GTF files were compared using an unstranded *bedtools intersect* (settings: *-wa -wb*). Comparisons were run between mT&T exons and Ensembl exons as well as between Ensembl exons with itself. The resulting file with both GTF entries was scanned using a custom Python script, which assessed each entry for the *transcript\_id* in entry A and the *gene\_id* in entry B. Such an entry indicates overlap between the exon from query file A with an exon in query file B. Overlapping loci were defined as the ones overlapping with two *gene\_ids*. As a final step, the same validation of novel ONT transcripts was performed by using the ONT assembly to quantify the Illumina data (Figure 3.12B).

### 2.1.9 Statistical analysis and visualisation of processed data

Data visualisation and statistical analyses were carried out in *Jupyter* notebooks (including the Python package *pandas*, *rpy2*) or in *RMarkdown*. The versions for the R packages used are given in Appendix Table A.3. Plots were generated using the R packages *ggplot2*, *Gviz* or *ComplexHeatmap*. For hierarchical clustering analyses Manhattan distance, complete linkage and leaf-optimisation (package *cba*) were used if not otherwise indicated. Labels were placed in plots using *ggrepel*. The R versions of the Two Sample Welch Test, Fisher's Exact tests or Wilcoxon Rank Sum Test were used for significance testing. The obtained p-values were multiple testing corrected using Benjamini-Hochberg procedure where appropriate.

PCAs were performed based on R function *prcomp*. The PCA based on gene expression was performed on variance-stabilised counts (*varianceStabilizingTransformation* from *DESeq2* using a 'local' fit) including only genes which were expressed above the 25 % quantile in at least three samples (Figure 6.16). Sashimi plots and coverage tracks were generated with the R packages *AnnotationDbi*, *GenomicRanges* and *Gviz*.

#### 2.1.9.1 Tissue grouping based on transcript expression

To avoid over-representation when comparing transcriptomes between tissues, I performed hierarchical clustering on the most highly variable transcripts in protein-coding genes

(Appendix Figure C.1). Manhattan distance and average linkage were used for the hierarchical clustering and one representative tissue per group was chosen. The reproducibility of clusters was assessed using the bootstrapping method implemented in the R package *pvclust* [141], which provides an approximate unbiased (AU) p-value. The majority of tissue clusters had an AU p-value of 95 % or more suggesting that these were stable clusters. In addition to selecting one representative tissue for each group (underlined in Appendix Figure C.1), I removed one of the two large intestine samples (named LgIntestine and Colon), retaining the one that was most similar to the other intestinal samples (Colon).

#### 2.1.9.2 Geneset over-enrichment testing

Fishers' Exact tests were used to identify Gene Ontology (GO) categories over-represented amongst sets of genes containing differentially spliced events, differentially used transcripts or containing novel transcripts (R package *gsfisher*, [github.com/sansomlab/gsfisher](https://github.com/sansomlab/gsfisher)). Protein-coding genes with TPM > 0.1 from the same tissue were used as the 'universe'. The categories were subsequently filtered by adjusted p-value (Benjamini-Hochberg correction), odds ratio and number of genes per category (Appendix Table A.4). Finally the over-represented categories were manually curated in order to avoid over-reporting of similar categories (similarity determined by inspection of genes involved in the enrichment).

#### 2.1.9.3 Definition of tissue-restricted genes

While several metrics have been applied to define tissue-restricted genes, the metric called *tau* was shown to be the most robust method showing reproducible results in different sets of tissues [142]. Using the tissue groups from ENCODE mouse tissues plus the two wildtype mTEC samples I calculated *tau* from upper-quartile normalised and log2(n+1)-transformed TPM values for all protein-coding genes [143]. The *tau* equation (Equation 2.2) includes the number of tissues,  $n$ , and the gene expression level,  $x_i$ , across samples.

$$tau = \frac{\sum_{i=1}^n (1 - \hat{x}_i)}{n - 1}; \hat{x}_i = \frac{x_i}{\max_{1 \leq i \leq n} (x_i)} \quad (2.2)$$

To confirm the robustness of  $\tau$  and to ensure over-representation of tissues did not result in misleading  $\tau$  values, I calculated it based on the set of representative tissues identified in Section 2.1.9.1. This analysis resulted in 9,155 genes defined as tissue-specific ( $\tau > 0.7$ ; Appendix Figure B.8A) compared to 9,158 tissue-restricted genes with all tissues included. The comparison of the two sets of tissues used for the calculation revealed that the majority of  $\tau$  values was not affected with only some values being slightly lower if representative tissues were used (Appendix Figure B.8C). Interestingly, for 40 genes a very high tissue-specificity ( $\tau > 0.9$ ) was found using all tissues, while the  $\tau$  value was 0 for the grouped tissue analysis. This observation suggested that these 40 genes were specific to tissues that were not selected to represent the groups. Taken together, I could confirm that  $\tau$  is a robust measure and the  $\tau$  values for genes based on the grouped tissues were used for downstream analyses.

The number of tissue-restricted genes identified from the grouped tissues here was higher than the 7,300 tissue-restricted genes found using microarray data and a dynamic step model described in Sansom *et al.* [27] (Figure B.8B). Nevertheless, the majority of genes identified as tissue-restricted were also defined as *Aire*-regulated, confirming that AIRE induced the expression of TRA in mTEC (Figure B.8B).

## 2.2 Experimental methods

The buffers or reagents prepared using in-house protocols are described in Appendix Table A.9. The manufacturing companies for all commercially available reagents or kits are listed in Appendix Table A.10.

### 2.2.1 Mice

Animals were housed and maintained in University of Oxford Biomedical Services Facilities, unless stated otherwise. All breedings and experiments were performed under the relevant project license as well as under UK Home Office Regulations. Mice were sacrificed following 'Schedule 1' methods. All experiments were performed at 4-6 weeks of age. Where possible, sex-matched mice from one litter were used. If necessary, age-matched and sex-matched littermates from another litter were added.

#### 2.2.1.1 C57BL/6

Wildtype C57BL/6 mice were originally obtained from Harlan Laboratories and maintained as a laboratory in-house colony. Additional wildtype C57BL/6 mice were purchased from the University of Oxford in-house colony.

#### 2.2.1.2 FoxN1-cre

Previously, the iCre sequence was introduced into a P1 artificial chromosome (PAC) containing the *FoxN1* gene. After injection of the modified FoxN1-PAC into pronuclei of fertilized B6D2F1 oocytes, a stable FoxN1-cre<sup>+/-</sup> line was maintained on the C57BL/6 background [144]. Due to the infertility of FoxN1-cre homozygous animals, heterozygous animals were used for breeding with *Rbfox1*<sup>lox/lox</sup>:*Rbfox2*<sup>lox/lox</sup>.

#### 2.2.1.3 *Rbfox1*<sup>lox/lox</sup>:*Rbfox2*<sup>lox/lox</sup>

The double-targeted animals were a kind gift of Prof. Douglas Black and not generated as part of this project. A breeding pair was received from his laboratory at the University of California, Los Angeles. Conditional single- and double-knockouts using Nestin-cre had been previously generated by Prof. Black's group [145, 146].

### 2.2.2 Genotyping and assessment of genomic excision in conditional knockout mice

DNA extractions for genotyping were performed using the *Phire Animal Tissue Direct PCR kit* according to the manufacturer's protocol. Alternatively, DNA was purified using the following protocol: Earclips were digested in lysis buffer with proteinase K at 55 °C overnight (Appendix Table A.9). After 15 min centrifugation at maximum speed, the supernatants were transferred into new Eppendorf tubes, 600 µl isopropanol added and the tube mixed to precipitate the DNA. Another 15 min centrifugation step was followed by washing the pellet with 600 µl 70 % ethanol. The DNA pellet was air-dried and resuspended in 30 µl TE buffer (Appendix Table A.9).

To assess the genotype of mice, 1 µl of purified DNA (or 3 µl of kit-extracted DNA) were used for polymerase chain reaction (PCR) in a thermocycler (Applied Biosystems). PCR reactions for iCre amplification were performed using the *Taq* kit, while *Rbfox1* and *Rbfox2* were amplified using the *Taq* polymerase or *Sequal* PCR kits. The conditions for iCre (*Taq*), *Rbfox1* and *Rbfox2* (*Taq/Sequal*) amplifications are given in Table 2.1. For the *Taq* polymerase PCR reaction, the template DNA (or control) and *Taq* buffer was mixed with dNTPs (final concentration: 200 µM each), both primers (final concentration: 0.2 µM each) and *Taq* polymerase (final concentration: 0.02 u/µl).

The alternative PCR reaction, namely *Sequal* PCR, included a *Sequal* buffer including dNTPs, 0.4 µl DMSO, *Sequal* Enhancer A, primers (final concentration: 10 µM each) and *Sequal* polymerase (final concentration: 1.8 U). The primer sequences are given in Appendix Table A.8. The PCR products were examined on a 1.5 % Agarose gel (in 1xTBE plus 1:20,000 *GelRed*). The amplified PCR products were diluted with 5 µl of 6x Loading Dye and 10 µl loaded on the gel. As a reference the 1 kb Plus ladder was used. Finally, the PCR products were visualised using the Gel Doc XR+ system (Bio-Rad).



**Table 2.1:** PCR programs for iCre and Rbfox1/2 genotyping as well as Rbfox1/2 genomic excision PCR. min = minutes, s = seconds.

| Number of cycles                   | Temperature                     | Time  |
|------------------------------------|---------------------------------|-------|
| <b>iCre program</b>                |                                 |       |
|                                    | 94 °C                           | 4 min |
| 35 x                               | 94 °C                           | 30 s  |
|                                    | 59 °C                           | 1 min |
|                                    | 72 °C                           | 45 s  |
|                                    | 72 °C                           | 2 min |
|                                    | 4 °C                            | hold  |
| <b>Rbfox <i>Taq</i> program</b>    |                                 |       |
|                                    | 98 °C                           | 2 min |
| 33 x                               | 98 °C                           | 30 s  |
|                                    | 53 °C (Rbfox1) / 50 °C (Rbfox2) | 30 s  |
|                                    | 72 °C                           | 40 s  |
|                                    | 72 °C                           | 6 min |
|                                    | 4 °C                            | hold  |
| <b>Rbfox <i>Sequal</i> program</b> |                                 |       |
|                                    | 94 °C                           | 2 min |
| 33 x                               | 94 °C                           | 10 s  |
|                                    | 53 °C (Rbfox1) / 50 °C (Rbfox2) | 30 s  |
|                                    | 68 °C                           | 3 min |
|                                    | 72 °C                           | 5 min |
|                                    | 4 °C                            | hold  |

## 2.2.3 Isolation of thymic epithelial cells

### 2.2.3.1 Thymus dissection and digestion

Thymi were dissected followed by removal of fat and connective tissue under the microscope. The tissue was digested using three rounds of mixed enzymatic and mechanic digestions as follows: In the first round, two thymus lobes were placed in 1.5 ml tissue digestion buffer (Appendix Table A.9) and gently pipetted with a large-cut tip. An incubation at 37 °C for 5 min was followed by another round of pipetting to mechanically support tissue digestion. The first round of digestion was finished by transferring the supernatant into a tube with 35 ml FACS buffer (Appendix Table A.9), which was stored on ice. The subsequent two rounds of digestion were following the same approach, but 1 or 0.8 ml tissue digestion buffer as well as small-cut or normal tips were used, respectively. In the final round of digestion, additional DNase enzyme could be added optionally. After the whole thymus was digested, the cell suspension was centrifuged (1,500 rpm for

5 min) and the supernatant poured off. The cell pellet was resuspended in 1 ml FACS buffer and filtered through a 70  $\mu\text{m}$  filter, which was washed with an additional 3 ml FACS buffer. The number of cell was counted using a haemocytometer (1:100 dilution).

### 2.2.3.2 Depletion of CD45<sup>+</sup> cells

To enrich the cells for CD45<sup>-</sup> cells, the cell suspension was incubated with magnetic beads coated in  $\alpha$  CD45 (5  $\mu\text{l}$  per 10 million cells). After the 15 min incubation, the cells were washed with 35 ml FACS buffer including a centrifugation step (1,500 rpm for 5 min). Subsequently, the cells were resuspended in FACS buffer and concentrations adjusted to a maximum of 100 million cells per ml (2 ml total volume). Finally, the cells were filtered using a 30  $\mu\text{m}$  filter, 80  $\mu\text{l}$  DNase added and tubes were loaded onto the AutoMACS machine (Miltenyl Biotec) for magnetic-activated cell sorting (program: *depletes* with *clean* steps between tubes).

### 2.2.4 Flow cytometry

In order to stain surface markers or intracellular markers on different cell types, the appropriate master mixes of flow cytometry antibodies were prepared (Table 2.2, Appendix Table A.7). For extracellular panels, cells were incubated with 1  $\mu\text{l}$  of antibody master mix for 20 min at 4 °C in the dark followed by a washing step. All washing steps consisted of addition of 2 ml FACS buffer, centrifugation (1,500 rpm for 5 min), removal of the supernatant and resuspension of cells. For intracellular stainings, a live-dead staining using the AQUA kit was performed: Cells were incubated in 1 ml 1xPBS plus 1  $\mu\text{l}$  AQUA reagent for 30 min (4 °C in the dark). Subsequently, the cells were fixed and permeabilised using the eBioscience Fix-permeabilisation kit, which includes a 30 min fixation step with 1 ml fixation buffer plus a washing step with permeabilisation buffer. All subsequent washing steps were performed using permeabilisation buffer. Finally, intracellular protein stainings were performed. The antibodies for intracellular proteins such as AIRE were incubated for 60 min at 4 °C in the dark. After a washing step, if necessary, the secondary antibody was added (for instance for RRM) and incubated for 60 min at 4 °C in the dark. Finally, the cells were washed before proceeding to the flow

cytometer. Depending on the antibody panel (Table 2.2), 1  $\mu$ l of DAPI (13.7  $\mu$ g/ml) was added for live-dead staining of the cells.

Prior to analysis of samples or sorting on the machine, a compensation using splenocytes was performed. In short, a spleen was squashed and red blood cells were lysed using red blood cell lysis buffer (Appendix Table A.9, 5 min at room temperature). Subsequently, cells were washed with 50 ml FACS buffer and centrifuged (1,500 rpm for 5 min). The resuspended cells were filtered using a 100  $\mu$ m filter, which was washed up to a final volume of 10 ml. About 1 million cells were stained using a compensation panel with the same fluorophors as used for experimental samples (antibodies given in Appendix Table A.7).

#### **2.2.4.1 Phenotyping analysis and cell sorting using flow cytometry**

Cell phenotyping as well as cell sorting were performed on a BD FACS Aria III (BD Biosciences). Before loading tubes on the flow cytometer, cells were filtered through a 40  $\mu$ m filter (using a 1 ml syringe) and vortexed. Cells were either sorted into FCS-coated Eppendorf tubes or into 75  $\mu$ l RLT Plus buffer (from RNeasy Micro Plus kit) in Eppendorf tubes (4-Way Purity). A cooling system to keep the samples at 4 °C (Thermo Scientific) as well as filter system (Bufallo filter) were used for sorting cells. Depending on downstream experiments, cells were either pelleted and snap frozen in liquid nitrogen or directly subjected to RNA extraction using the RNeasy Micro Plus kit.

#### **2.2.4.2 Analysis of flow cytometry data**

Flow cytometry data was analysed using *FlowJo* (version 10.5.0). Analyses included adjustment of compensation, gating schemes as well as extraction of geometric mean fluorescence intensity (gMFI) as a measure of protein expression. All gating schemes are shown as 5 % contour plots with outliers, unless otherwise stated.

**Table 2.2:** Antibody panels for different types of experiments. The details about antibodies and concentrations are given in Appendix Table A.7. For experiments with extracellular staining only, DAPI was used for live-dead staining, while AQUA was used for experiments with intracellular stainings.

| Master mix                                                 | Antibody        | Fluorophore  |
|------------------------------------------------------------|-----------------|--------------|
| <b>TEC sort &amp; TEC phenotyping (undepleted samples)</b> |                 |              |
| Extracellular staining                                     | CD45            | AF700        |
|                                                            | EpCAM           | PerCp-Cy5.5  |
|                                                            | Ly51            | PE           |
|                                                            | UEA-1           | Cy5          |
|                                                            | MHCII           | FITC         |
|                                                            | CD80            | PE-Cy5       |
|                                                            | CD86            | PE-Cy7       |
| <b>TEC phenotyping (depleted samples)</b>                  |                 |              |
| Extracellular staining                                     | CD45            | AF700        |
|                                                            | EpCAM           | PerCp-Cy5.5  |
|                                                            | Ly51            | PE           |
|                                                            | UEA-1           | Cy5          |
|                                                            | MHCII           | BV421        |
|                                                            | CD80            | PE-Cy5       |
|                                                            | CD86            | PE-Cy7       |
| Intracellular staining                                     | AIRE            | AF488        |
| <b>RRM experiments</b>                                     |                 |              |
| Extracellular staining                                     | CD45            | AF700        |
|                                                            | EpCAM           | PerCp-Cy5.5  |
|                                                            | Ly51            | PE           |
|                                                            | MHCII           | BV421        |
|                                                            | CD80            | PE-Cy5       |
|                                                            | CD86            | PE-Cy7       |
| Intracellular staining                                     | AIRE            | AF647        |
|                                                            | RRM (primary)   | unconjugated |
|                                                            | RRM (secondary) | PE-Texas Red |

## **2.2.5 RNA-sequencing**

### **2.2.5.1 Cell sorting and RNA extraction for transcriptomics**

For Oxford Nanopore Technology transcriptomics (Chapter 3) as well as for stranded Illumina RNA-sequencing (Chapter 6), mTEC were sorted into 75 µl RLT Plus buffer and RNA was extracted immediately (RNeasy Micro Plus kit). Total RNA was eluted in 14 µl RNase-free water and RNA quality was assessed using the Tapestation (Agilent 4200 Tapestation) or Bioanalyzer (Agilent 2100 Bioanalyzer) according to manufacturer's instructions. A minimum RIN of 7.7 (6.8 for pilot) was used for the ONT samples (Chapter 3), while a minimum RIN of 7.1 was used for the stranded Illumina RNA-sequencing (Chapter 6). Finally, samples were transported to the Wellcome Trust Centre for Human Genetics for further processing.

### **2.2.5.2 Oxford Nanopore Technologies RNA-sequencing**

For the RNA-sequencing using the MinION platform (ONT), the following steps were performed by Moustafa Attar and Maria Lopopolo (Wellcome Trust Centre for Human Genetics). 8 µl of the extracted RNA were prepared for sequencing using the cDNA-PCR sequencing kit following manufacturer's instructions. cDNA libraries were amplified for 18 cycles and normalized to 400 fmol before sequencing on MinION flow cells (FLO-MIN 106 R9.4) for 48 hours.

### **2.2.5.3 Stranded RNA-sequencing from conditional knockout animals**

For stranded RNA-sequencing the samples were processed by the Oxford Genomics Centre sequencing service in the Wellcome Trust Centre for Human Genetics. Libraries were generated using a stranded, paired-end protocol including polyA-enrichment (NEB), subsequently pooled and sequenced on a HiSeq 4000 (150 bp read length).

## **2.2.6 Immunohistochemistry**

### **2.2.6.1 Preparation of tissue sections**

After tissues were dissected from mice, fat as well as connective tissue were removed under a microscope. The tissue was washed with 1xPBS, followed by a wash with OCT and added to a suitable mould. The mould was filled with OCT and left in a butanol

and dry-ice mixture to freeze. Samples were stored at -80 °C.

Eight  $\mu\text{m}$  tissue sections were cut using a Cryostat (Thermo Scientific CryoStar NX70 with MB DynaSharp Microtome Blade). The sample and blade temperatures were set to -22 and -18 °C, respectively. Two tissue sections were placed on one SuperFrost Plus slide. The slides were stored at -20 °C.

### 2.2.6.2 Immunofluorescence staining of thymus sections

The tissue sections were fixed for 20 min in 1.4 % PFA (in 1xPBS) and for 10 min in Methanol. Permeabilisation was performed for 10 min in 0.3 % Triton-X (in 1xPBS). These steps were followed by one 5 min washing step in 1xPBS, marking of the individual sections by a hydrophobic PAP pen and two further 5 min washing steps (1xPBS) 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. Slides were placed in a humid box to avoid drying of the slides. The individual dilutions and information on primary antibodies are given in Appendix Table A.5. After three washing steps (5 min, 1xPBS), the secondary antibody was added (diluted 1:500 in 1xPBS; Appendix Table A.6) and incubated for 30 min at 37 °C.

If UEA-1 staining was added, the secondary antibody was followed by three washing steps (5 min, 1xPBS) and a 90 min UEA-1 staining at 37 °C (1:150 dilution in 1xPBS). If no UEA-1 staining was added or after the UEA-1 staining, the following steps were performed: Two washing steps (5 min, 1xPBS) were followed by DAPI staining (10 min, 1:10,000 in methanol). Finally, one final washing step (5 min, 1xPBS) was followed by mounting using ProLong Gold antifade reagent.

### 2.2.6.3 Microscopy and data analysis

The slides were imaged on a confocal microscope using a 63x lens (with oil; Zeiss LSM780 inverted) with the ZEN software (ZEN Black 2011 SP7 LSM). Images were analysed using *Fiji* (version 1.0).

# 3

## Generation and validation of a reference transcriptome to assess splicing in thymic epithelial cells

### Contents

---

|            |                                                                                                                              |           |
|------------|------------------------------------------------------------------------------------------------------------------------------|-----------|
| <b>3.1</b> | <b>Background . . . . .</b>                                                                                                  | <b>58</b> |
| <b>3.2</b> | <b>Strategy to generate the mT&amp;T assembly . . . . .</b>                                                                  | <b>60</b> |
| 3.2.1      | Construction of assemblies from individual mouse tissues . . .                                                               | 62        |
| 3.2.2      | Merging of tissue assemblies and comparison to the reference<br>annotation . . . . .                                         | 66        |
| 3.2.3      | Assessment of mT&T assembly quality . . . . .                                                                                | 66        |
| <b>3.3</b> | <b>Validation of structures using low-input long-read RNA-<br/>sequencing of medullary thymic epithelial cells . . . . .</b> | <b>75</b> |
| 3.3.1      | Quality assessment of long-read RNA-sequencing data . . . . .                                                                | 75        |
| 3.3.2      | Building a transcriptome assembly from long-read RNA-sequencing<br>data . . . . .                                            | 77        |
| 3.3.3      | Validation of novel transcripts between short and long-read<br>sequencing technologies . . . . .                             | 85        |
| <b>3.4</b> | <b>Discussion . . . . .</b>                                                                                                  | <b>87</b> |

---

## 3.1 Background

In the thymus, medullary thymic epithelial cells (mTEC) are responsible for a large part of the negative selection of developing thymocytes in order to avoid the escape of auto-reactive T cells into the periphery. These specialised antigen-presenting cells have been shown to express a remarkably large number of protein-coding genes, which include a variety of tissue-restricted antigens (TRA) [37]. The mechanism of TRA expression in TEC is termed 'promiscuous gene expression' and can occur via different pathways. Of these, the most highly studied process is the regulation of TRA expression by the Autoimmune Regulator, AIRE (Section 1.2.3). In addition to the unusually high number of protein-coding genes in mTEC, a high complexity with respect to splice isoform diversity has been reported in this cell type [101]. Keane *et al.* further hypothesized that the complex transcriptome of TEC could be generated by promiscuous splicing in these cells. Recent reports have, however, focussed on exon expression (using microarrays [101]) or splice junctions (using RNA-sequencing data [103]). One reason for the limited information on individual transcript expression in TEC could be the absence of long-read, stranded, deep RNA-sequencing datasets, which would be needed to have enough power to detect differences in transcript usage. Recent studies suggested that 100-400 M paired-end Illumina reads are needed, and now typically used, for the characterisation of splicing or splicing differences in mammalian transcriptomes [70, 147, 148]. Previously, one group generated deep, paired-end RNA-sequencing datasets from different mTEC subtypes (average read depth 134 M reads; unstranded) [50]. The generated RNA-sequencing dataset used in this project had on average higher sequencing depth plus strand information (average 170 M reads per sample) .

In recent years there have been efforts to improve the annotation of the mouse transcriptome, which had been poorly annotated compared to the human transcriptome. The sequencing of an extensive set of mouse adult and embryonic tissues has aimed to extend the mouse transcriptome annotation [69, 70, 149]. The RNA-sequencing data from the mouse tissues were used to generate an assembly of the mouse transcriptome



[70] which revealed 200 k novel transcripts that were also supported by CAGE data from the FANTOM project [150]. The assembly pipeline used by Pervouchine *et al.* included *Cufflinks* to generate transcript models followed by quantification and filtering using the non-parametric irreproducibility discovery rate (npIDR) [151, 152]. Additionally, Pervouchine *et al.* required novel transcripts to have CAGE support.

In this chapter, the RNA-sequencing data from peripheral mouse tissues from the mouse ENCODE project were used to generate a common reference transcriptome including peripheral mouse tissues and mTEC. The assembly generated by Pervouchine *et al.* was not suitable for the analyses here as it did not include a transcriptome assembly from mTEC. However, the approach by Pervouchine *et al.* was partially adopted for a combined assembly from mouse peripheral tissues and mTEC. In addition to the ENCODE mouse tissues, other datasets including more tissues such as retina or salivary gland were available (GNF Mouse GeneAtlas V3, [153]). However, this dataset contains microarray data which is only partially suitable for analysis of splicing, depending on whether probes are spanning splice junctions. Another dataset, that was available when this project was started, was RNA-sequencing from eight mouse tissues (read length: 36/50/80 bp) [65]. Due to the limited set of tissues and shorter read length, I decided to perform the analysis on the ENCODE mouse tissues.

To resolve transcript structures in detail, short-read RNA-sequencing data would ideally be complemented with full-length transcript information. Using short-read RNA-sequencing data, transcript structures can often not be disentangled fully due to overlaps between the structures, where reads cannot be uniquely placed on one structure. With newer long-read sequencing technologies such as Pacific Bioscience sequencing or Oxford Nanopore Technology (ONT) sequencing, very long reads up to full transcript length can be achieved (Section 1.3.4 and Figure 1.5) allowing unambiguous determination of the splice isoforms that are present in a cDNA library.

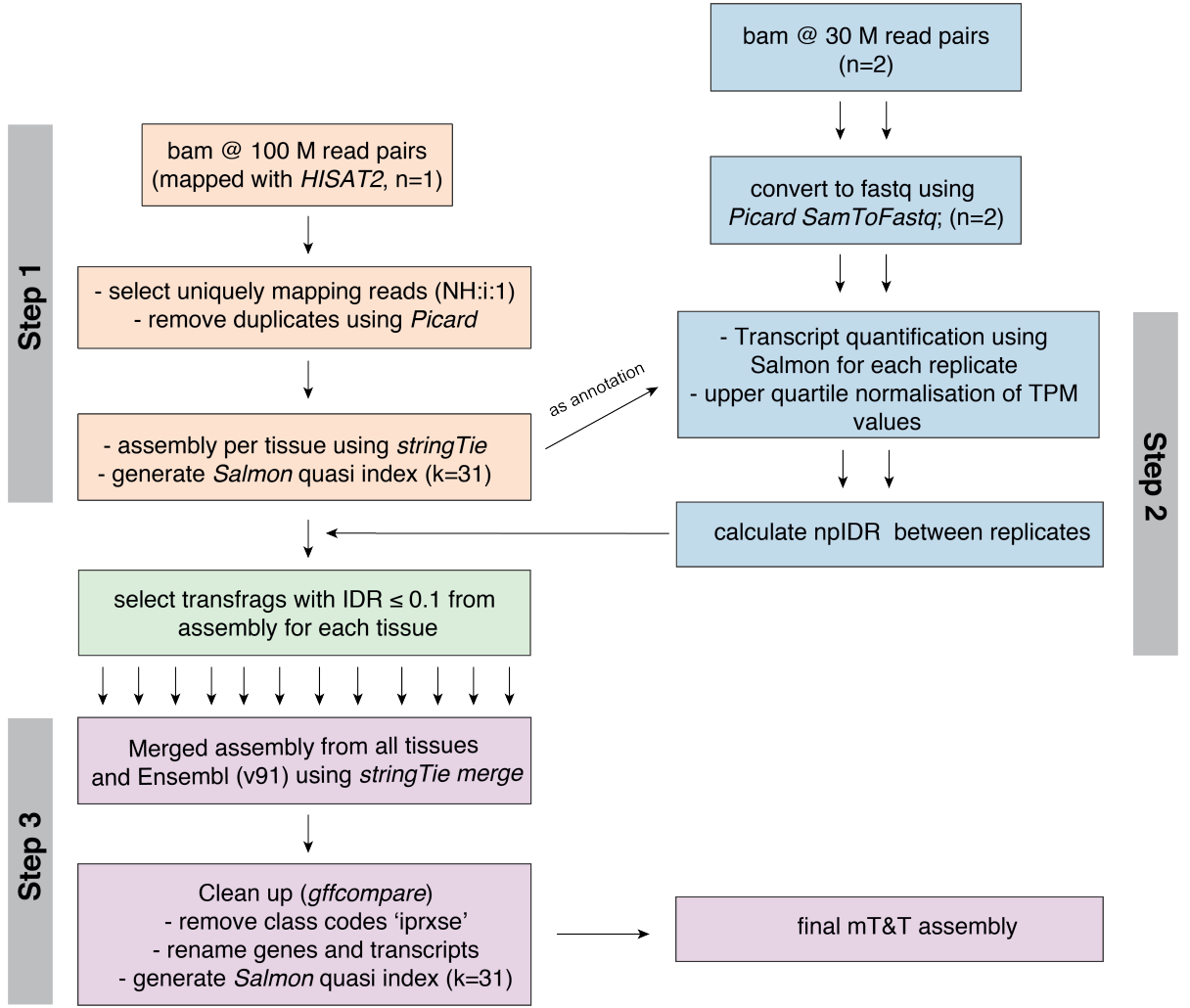
## Aims

Publicly-available annotations likely under-represent the transcriptomes of specific cell types such as TEC. The transcriptome of TEC has not been defined at the splice isoform level and therefore needs to be assembled. In order to perform a fair comparison, assemblies of similar quality are required from all peripheral tissues. Finally, to facilitate comparison between the individual tissue and mTEC assemblies a unified assembly is necessary as a common reference point. In this chapter, I first aimed to use high-depth RNA-sequencing data from a variety of mouse peripheral tissues and mTEC to build an assembly, which was termed **m**ouse **T**issue & **T**EC assembly (mT&T assembly). In addition to serving as a common reference, the assembly was also used to assess the presence and structure of novel transcripts in TEC (Section 4.4). Second, I aimed to validate assembled structures using long-read sequencing from small bulks of mTEC.

## 3.2 Strategy to generate the mT&T assembly

In order to compare splicing in mTEC to peripheral tissues, I built a new transcriptome assembly using RNA-sequencing data from 22 ENCODE mouse tissues as well as four TEC population ([70], Appendix Tables B.1, B.3). For the TEC populations, stranded, paired-end and deep RNA-sequencing datasets were generated from immature (MHCII low) and mature (MHCII high) wildtype mTEC as well as from heterozygous (mature *Aire*-positive) and homozygous *Aire*-knockout (mature *Aire*-KO) mTEC samples. In the mouse model used here enhanced GFP is under the transcriptional control of the *Aire* promoter due to the insertion of GFP into exon 1 of *Aire* ([27]). For both models, the GFP+ mature mTEC were sorted. These cells should either have one functional (heterozygous) or no functional (homozygous) copy of AIRE.

The RNA-sequencing data from the peripheral tissues and mTEC samples were trimmed to a common read length of 76 bp in order to match the read length of the mouse ENCODE samples. After mapping with *HISAT2*, the read depths of all samples were assessed (Appendix Figure B.1, Appendix Table B.2). To obtain two biological replicate



**Figure 3.1:** Schematic overview of the pipeline to generate the mT&T assembly.

sample pools per tissue, 60 M reads (30 M read pairs) was the highest read depth available across all tissues (Appendix Table B.3). In addition, a high-depth sample for assembly (200 M reads; drawn from all available replicates) was generated from each tissue. The aligner *HISAT2* was chosen as it has been shown to provide fast and accurate alignments [121]. The 'new Tuxedo' protocol for transcript quantification of RNA-sequencing data includes *HISAT2* as well as *stringTie* and *Ballgown* [154].

The pipeline to generate the assembly was modified from the pipeline used by Pervouchine *et al.* [70], and consisted of three steps (Figure 3.1): an initial assembly per tissue, assessment of reproducibility within each tissue and merging of the tissue assemblies including filtering of the merged assembly.

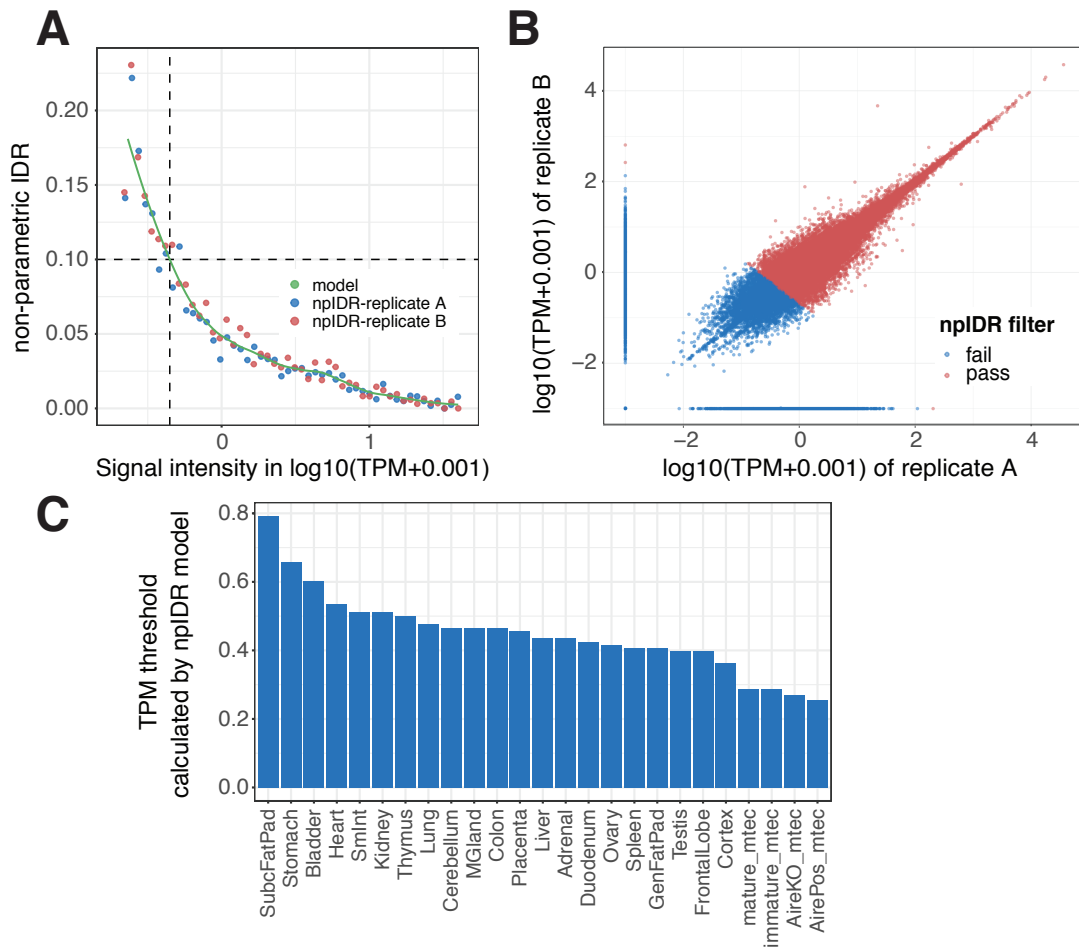
### 3.2.1 Construction of assemblies from individual mouse tissues

Firstly, for each mTEC or tissue sample a reference-guided *stringTie* assembly was built from the high-depth sample (200 M reads). Given the overall aim of extending the known annotation to include the TEC-specific transcripts, a reference-guided assembly was chosen with Ensembl (v91) annotation as the reference. The assembler *stringTie* was a new fast and accurate assembler at the time of analysis and it was therefore used in combination with *HISAT2*. The *stringTie* assembly process includes the generation of a splice graph followed by finding the heaviest path through the graph [155]. Applying this new method, the authors showed that it produced a more complete transcriptome assembly than the previously widely-used assembler *Cufflinks* [155]. Another group has tested a range of assembly tools on different species, where *stringTie* ranked better than *Cufflinks* [125]. In this study, the ranking was based on a combination of sensitivity and precision (F1 score) for assembly feature such as exons or intron-chains across the species examined.

Secondly, I aimed to assess whether - for each tissue - the assembled transcripts were reproducible between the two replicates. The metric Irreproducibility Discovery Rate (IDR) has been used extensively for ChIP-sequencing data analysis [152]. This approach uses a copula mixture model to classify signals into reproducible and irreproducible. More recently, a metric termed non-parametric Irreproducibility Discovery Rate (npIDR) was suggested for RNA-sequencing analysis by Dobin *et al.* [124]. According to the publication, the signals in each replicate should be binned and then the fraction of for instance transcripts that are present in replicate 1, but are not detected in replicate 2, should be calculated. The same should be repeated vice versa and the npIDR values for each bin averaged. As this metric was also used in the pipeline from the Pervouchine *et al.* assembly [70], I followed the described workflow. The per-tissue assemblies here were converted to per-tissue *Salmon* indices and used to quantify transcript expression in the two replicates at 60 M reads from the same tissue. *Salmon* was used for quantification due to its speed and high accuracy at the transcript level [123]. In addition to the npIDR script by Pervouchine *et al.*, the R package *IDR* applying the copula mixture model was

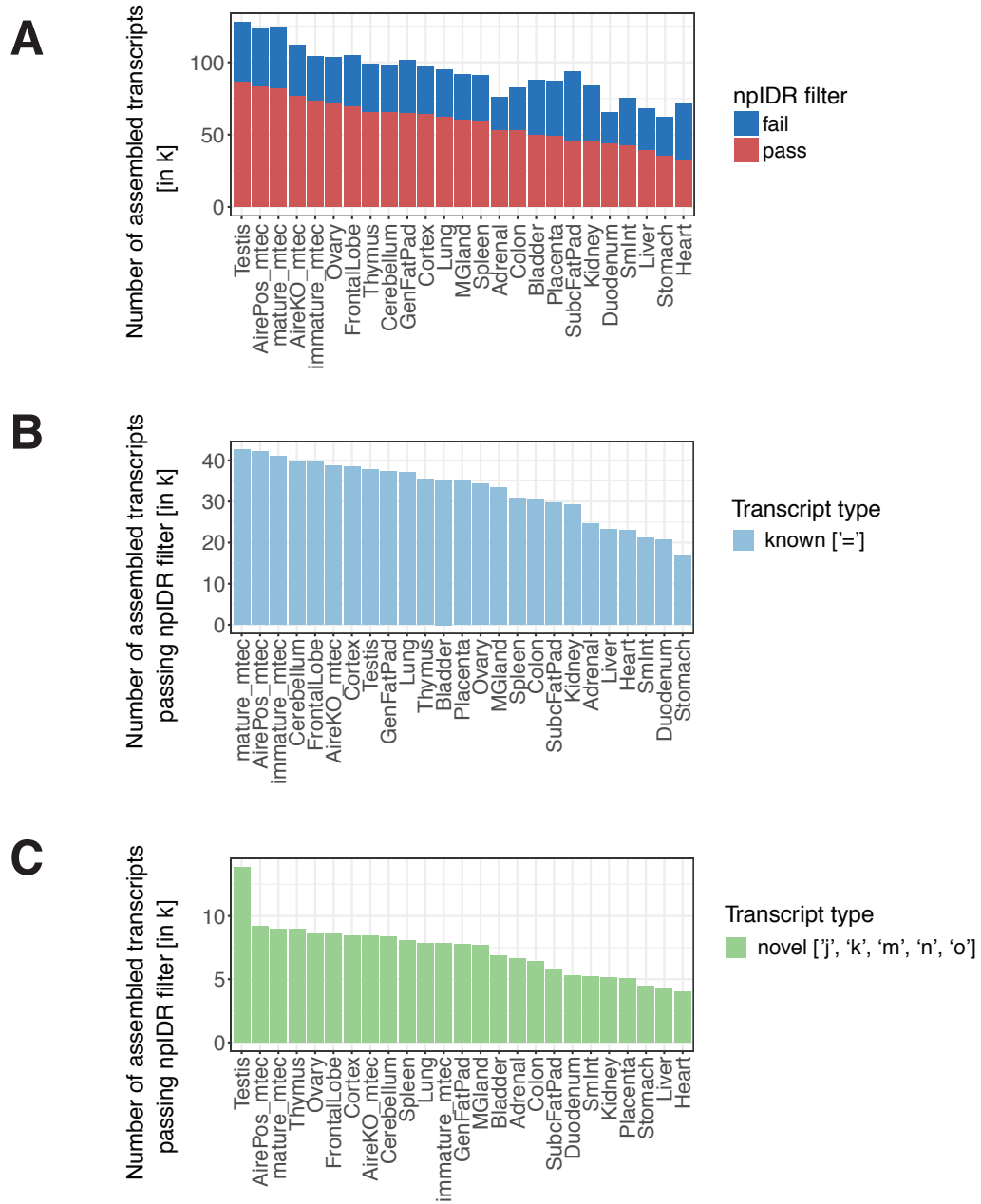
tested to define reproducible transcripts. The analysis using the R package showed very few reproducible transcripts between the two replicates (Appendix Figure B.2A), while the script from Pervouchine *et al.* excluded very few assembled transcripts (Appendix Figure B.2B). When applied to our dataset, an additional issue with the latter approach was that lowly expressed transcripts were not excluded from the downstream analysis. This could likely be caused by binning issues, where by chance bins are empty due to sparsity of the data at the extremities of the distribution. Due to this observation, I developed a different binning strategy and filtering process. After upper-quartile normalisation and log-transformation of the TPM values, the TPM distributions were cut at the 0.05 and 0.95 quantiles. All data above or below each 0.05 quantile were pooled into one bin on each side to avoid data sparsity issues, and the described npIDR was calculated per replicate. Instead of an average of the individual npIDRs from the two replicates, they were plotted separately against the signal intensity (TPM here, Figure 3.2A) and a loess curve was fitted through all data points. The resulting curve was used to determine the signal threshold for  $\text{npIDR} \leq 0.1$ . Finally, only transcripts passing this TPM threshold (mean of both replicates) were kept for the merging step (Figure 3.2B). In summary, my updated npIDR approach for TPM data successfully excluded transcripts that show high variability between the replicates (potentially due to low expression level). The TPM thresholds used for the filtering ranged from 0.26 to 0.79 TPM (Figure 3.2C).

Applying this strategy to all peripheral tissue and TEC samples resulted in 33 k to 87 k transcripts passing per tissue (Figure 3.3A). The highest number of transcripts was assembled and passed the TPM threshold in the testis followed by the four mTEC types (Figure 3.3A). Transcripts were classified into known (full intron-chain match, class code [=]) or novel transcripts (class codes [j'+k'+m'+n'+o']) using the transcript categories defined in *gffcompare* (definitions in Table 3.1). This annotation of the transcripts passing the TPM threshold, revealed 16.6 k (stomach) to 42.7 k (mature mTEC) known transcripts across the TEC and tissue samples (Figure 3.3B). Interestingly, the number of known transcripts was higher in mature and *Aire*-positive mTEC than in the Testis indicating that more assembled transcripts in the testis were novel transcripts



**Figure 3.2:** Results of reproducibility analysis using non-parametric IDR. The representative peripheral tissue shown here is adrenal. **A:** Model fit using npIDR and binned gene expression value in TPM. **B:** Transcripts passing npIDR using updated npIDR strategy. **C:** TPM threshold for each tissue determined by updated npIDR approach.

or falling into transcript categories, that were not included here. Consistent with Figure 3.3A, the testis had the highest number of novel transcripts passing (13.9 k), while the lowest number was observed for the heart (4 k) (Figure 3.3C).



**Figure 3.3:** Statistics on npIDR calculation. **A:** Number of transcripts passing or failing npIDR in different tissue assemblies. Tissues are ordered by the number of passing transcripts. **B** and **C:** Number of known (B) and novel (C) transcripts for all passing transcripts per tissue assembly. Known and novel transcripts were defined by *gffcompare* using the indicated class codes.

### 3.2.2 Merging of tissue assemblies and comparison to the reference annotation

As a third step, the transcripts that passed npIDR in each tissue, and the overall Ensembl annotation (v91) were merged. *StringTie merge* was used for merging and the default filters were deactivated due to the filtering already performed with npIDR. The resulting merged assembly was compared to the Ensembl annotation using *gffcompare*. Transcripts were filtered for the assigned class codes (Table 3.1): If the transcripts fell entirely within a reference intron ['i'] or overlapped with the reference on the opposite strand (exonic ['x'] or intronic ['s'] overlap), they were excluded from the assembly. Additionally, transcripts that represented potential polymerase run-on fragments ['p'] or single exon, pre-mRNA fragments ['e'] were not included. In total, around 53 k transcripts were excluded due to their class codes.

Several categories were summarised as novel transcripts including multi-exonic transcripts with at least one junction overlap ['j'], extension of the reference ['k'] or same strand overlap with reference transcripts ['o']. Furthermore, two class codes ['m', 'n'] with various overlaps including retained introns were included into the novel transcripts here. These transcript categories represented variations in the coding sequence and were therefore likely a source of TEC-specific peptides, which could be presented to developing T cells. Finally, all these novel transcripts were renamed according to the transcript they are part of plus '\_NX', where X is the number of novel transcripts assigned to the reference transcript by *gffcompare*. The categories ['u', 'y'] were kept in the assembly, but not renamed or assigned to a known gene or transcript as these might be novel genes (e.g. lincRNA). The final mT&T assembly included 243 k transcripts of which 71 k were high-confidence novel transcripts (Table 3.2).

### 3.2.3 Assessment of mT&T assembly quality

After creating a reference assembly containing all transcripts from the Ensembl annotation (v91) and additional transcripts that were assembled in the peripheral tissues or mTEC



**Table 3.1:** Class codes assigned by *gffcompare* after the merging step. The transcripts with class codes in the bottom section of the table were excluded from the assembly. Transcripts with class code [‘r’] were marked for exclusion, but no transcripts occurred in this category. N = Number, SJ = splice junction.

| Class code | N transcripts<br>(after npIDR) | Description of class code<br>(from <i>gffcompare</i> ) |
|------------|--------------------------------|--------------------------------------------------------|
| =          | 133,847                        | full intron-chain match                                |
| j          | 45,710                         | multi-exonic, 1 SJ match                               |
| k          | 4,021                          | containment of reference                               |
| m          | 5,548                          | RI, full intron-chain match or overlap                 |
| n          | 12,768                         | RI, partial match                                      |
| o          | 2,960                          | other overlap with reference                           |
| u          | 38,339                         | unknown                                                |
| y          | 403                            | contains reference within introns                      |
| i          | 19,037                         | fully within reference intron                          |
| e          | 3                              | single-exonic partially in intron                      |
| x          | 13,759                         | exonic overlap on opposite strand                      |
| s          | 15,712                         | intron-chain match on opposite strand                  |
| p          | 4,944                          | no overlap, possible polymerase run-on                 |

**Table 3.2:** Summary of transcripts and their class codes assigned by *gffcompare* in the final assembly.

| Category | Class code        | Number of transcripts |
|----------|-------------------|-----------------------|
| Known    | =                 | 133,847               |
| Novel    | j + k + m + n + o | 71,007                |
| Unknown  | u + y             | 38,742                |
| Total    |                   | 243,596               |

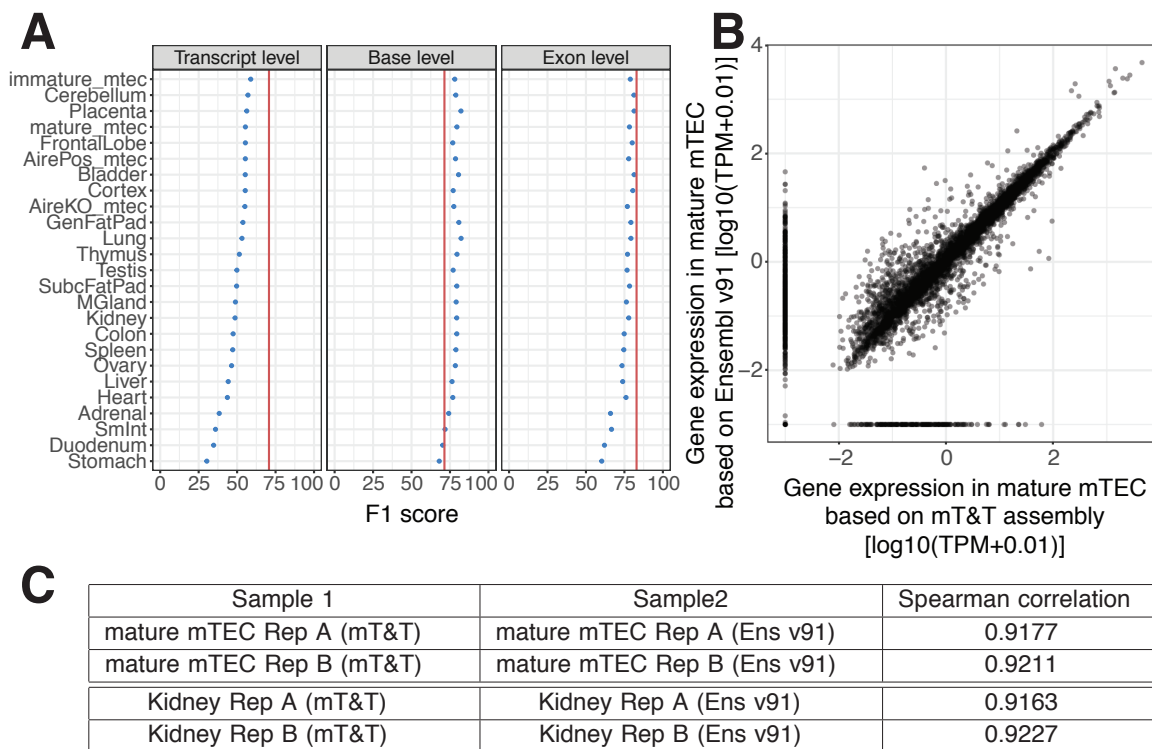
samples, the quality of the mT&T assembly was assessed to ensure it is suitable for the comparison between splice variants in mTEC and peripheral tissues.

### 3.2.3.1 Comparison of the individual tissue assemblies and of the mT&T assembly with the reference annotation

To compare the assemblies between the peripheral tissues and mTEC, the sensitivity and precision calculated by *gffcompare* were combined into the F-score for each tissue assembly (Equation 2.1). The F1 score for immature and mature mTEC were amongst the Top 5 scoring tissues at the base, exon and transcript level (Figure 3.4A). This could

be an indication of the transcriptome complexity in mTEC and supported the observation of high-complexity splicing in mTEC (Chapter 4). In contrast, the F1 score for stomach, duodenum and adrenal were amongst the lowest for all three categories.

To examine whether gene expression levels between a quantification using the ENCODE annotation (v19) or quantification using the mT&T assembly were comparable, the correlation was assessed for example tissues (Figure 3.4B). For mature mTEC and a peripheral tissue (kidney chosen as an example) correlations of around 0.92 were obtained (Table 3.4C) suggesting that using the mT&T assembly had limited effects on the gene-level quantification.



**Figure 3.4:** Assembly statistics for the mT&T assembly. **A:** F1 scores for tissue assemblies and final mT&T assembly. Scores are calculated from sensitivity and precision values from *gffcompare*. **B:** Correlation of gene expression quantified using Ensembl 91 annotation or mT&T assembly for one example tissue. 10,000 randomly selected genes are shown. **C:** Table with correlations between the quantification using the mT&T assembly and quantification using Ensembl v91 (Ens v91).

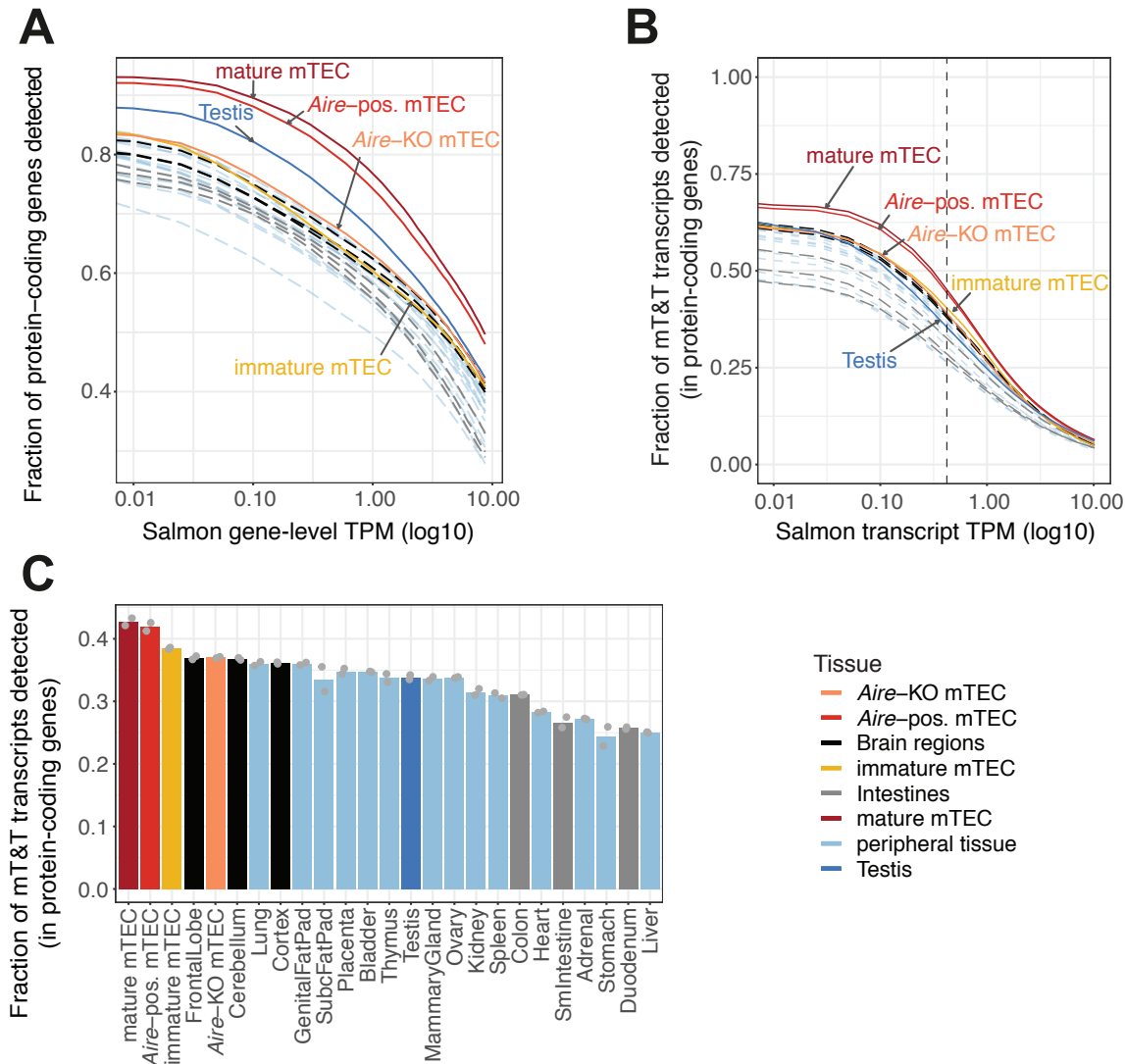
### 3.2.3.2 Gene and transcript detection using the mT&T assembly

Using the mT&T assembly, I found the highest fraction of protein-coding genes to be expressed in mature mTEC (Figure 3.5A), which was consistent with a previous publication from our groups and another report, which detected 89 % and 85 % of protein-coding genes in mature mTEC, respectively [27, 103]. An even higher fraction of 93 % of protein-coding genes was found with the mT&T assembly and the deeply sequenced RNA-sequencing dataset.

The observation of a high fraction of protein-coding genes detected also holds true on the transcript-level, where I found the highest fraction of transcripts in protein-coding genes in mature mTEC (67 %) followed by *Aire*-positive mTEC (Figure 3.5B). The other two mTEC types, namely *Aire*-KO mTEC and immature mTEC showed a lower fraction of mT&T transcripts. Inspecting the fraction of transcripts at the TPM threshold of 0.45 (average of TPM thresholds determined using the npIDR approach, Figure 3.2), revealed that all four mTEC types and all three brain regions are amongst the tissues with the highest fraction of transcripts detected (Figure 3.5C), which agreed with previous observations of large numbers of transcripts in the brain [65, 64].

### 3.2.3.3 Assessment of multi-gene transcripts in the mT&T assembly and the reference annotation using exon networks

During an assembly from short-read RNA-sequencing data, artefactual transcripts can occur. One of such artefacts could be caused by repetitive regions or regions with similar sequences, where short reads cannot be unambiguously mapped. Given the stringent filtering of reads before assembly (uniquely mapping, deduplicated reads), together with the use of stranded, long (76 bp), paired-end RNA-sequencing data, artefacts arising from mismapped reads were expected to be minimal [156]. Another type of artefactual transcripts in an assembly could be transcripts connecting two genes, termed 'multi-gene transcripts' here. However, it is unclear whether these are biologically meaningful or artefacts of the sequencing data or the assembly.



**Figure 3.5:** Detection of genes and transcripts using the mT&T assembly. **A:** Fraction of protein-coding genes detected across different TPM thresholds. **B:** Fraction of transcripts in protein-coding genes detected across different TPM thresholds. The dashed line indicates the threshold used for the bargraph in C and throughout this thesis (TPM = 0.45). The threshold represented the mean of the TPM thresholds from all tissues determined by npIDR (Figure 3.2). **B:** Bargraph showing the fraction of protein-coding transcripts detected at an expression threshold of 0.45 TPM. The TPM for genes and transcripts were generated using *Salmon* from two replicates per tissue (60 M reads). In A and B the mean for both replicates is shown, while the replicates are indicated as grey dots in C. The color legend for all panels is given in the bottom right.

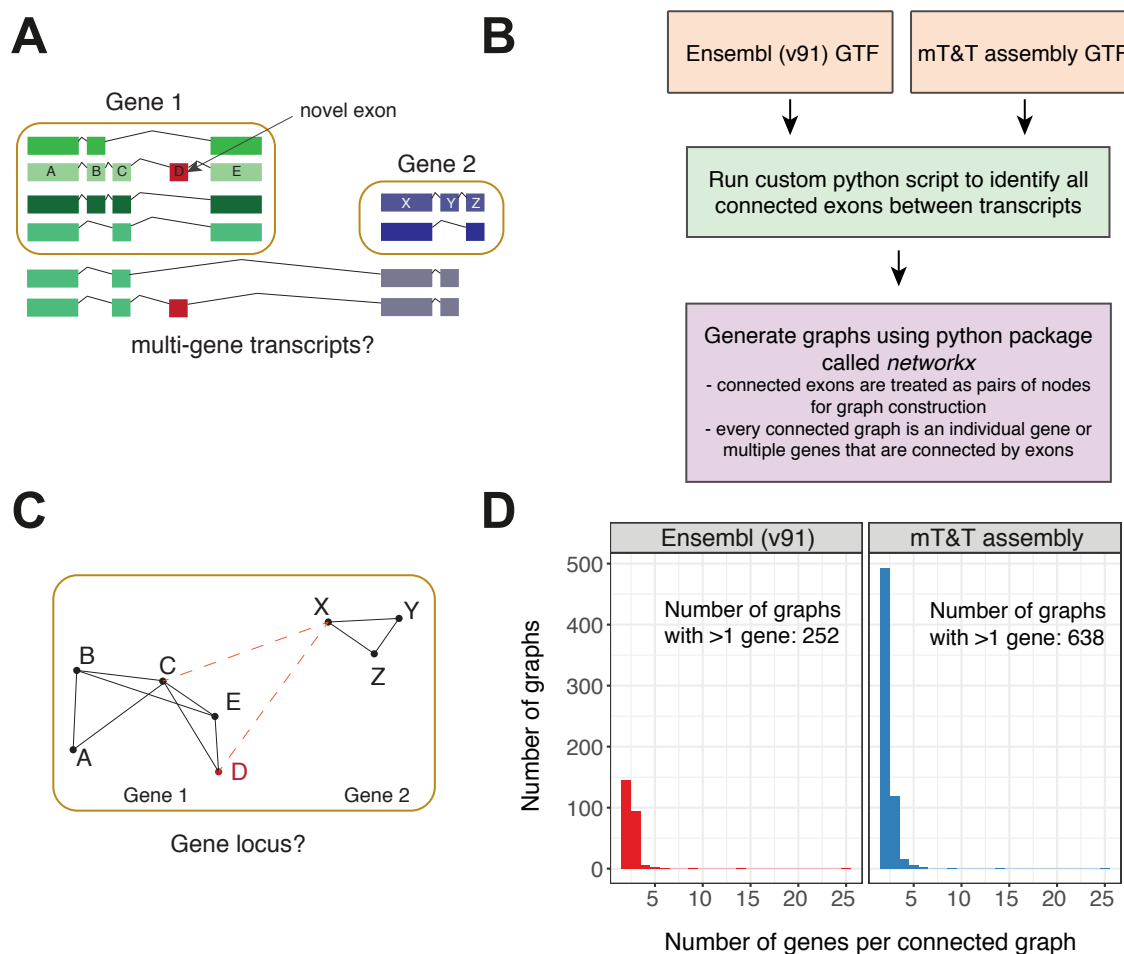
Given the reported issues of such multi-gene transcripts with *stringTie* (Section 2.1.3.2) and to understand if the assembly contained problematic artefactual multi-gene transcripts, an approach using exon graphs was chosen to assess the presence of such multi-gene transcripts. Simple graphs were used to generate networks, where individual exons are the nodes and each edge represents a connection between two exons. The number of genes within an independent graph then indicates an overlap between the exons of the contained genes. As hypothesized above, it is possible that gene boundaries are not strict and transcripts may span two genes suggesting that individual genes should be summarised into gene loci (Figure 3.6A). To assess such gene loci, I built separate network graphs of all exons in the Ensembl annotation and the mT&T assembly (analysis outline in Figure 3.6B). Each node in a graph is an exon and every edge a connection between two exons. For this approach, all connected exons in the Ensembl annotation and in the mT&T assembly were identified. The pairs of exons were subsequently used to generate non-directional graphs using the python package *networkx*. If transcripts were present, which connected two genes, then the resulting graph would contain both genes (Figure 3.6C). Such a connection could be caused by either existing transcripts or novel transcripts.

Running the graph-based analysis using exons in protein-coding genes, I identified 32.1 k individual graphs for the Ensembl annotation and 37.7 k for the mT&T assembly (Table 3.3). These exons could either be in different genes or in the same gene, but not connected by a splice junction previously. The majority of graphs only contained one

**Table 3.3:** Number of individual graphs obtained when using pairs of connected exons from either all the transcripts in protein-coding genes for Ensembl annotation or the mT&T assembly.

| Number of individual graphs within protein-coding genes |               |
|---------------------------------------------------------|---------------|
| Ensembl (v91)                                           | mT&T assembly |
| 32,064                                                  | 37,698        |

gene, while up to 25 genes were included in an individual graph in both the Ensembl annotation and the mT&T assembly (Figure 3.6D). The number of individual graphs with more than one gene was increased from 252 to 638 when comparing the Ensembl annotation to the mT&T assembly. This difference was mainly due to the 3-fold increase



**Figure 3.6:** Graph-based analysis to assess multi-gene transcripts. **A:** Scheme of two genes in a genomic locus with known and novel transcripts. **B:** Analysis outline to identify multi-gene transcripts using graphs. **C:** Graph representation of the example gene locus with potential connections between nodes (=exons). **D:** Histogram of the number of genes within each connected graph for the Ensembl annotation (v91, left panel) and the mT&T assembly (right panel), respectively. Only individual graphs with more than one gene were included in D.

in the number of graphs containing two genes.

Inspecting the graphs containing more than 5 genes in either of the assemblies, revealed one more graph in the mT&T assembly compared to the Ensembl annotation (Tables 3.4, 3.5). One of the graphs containing multiple genes in both assemblies was a complex locus with multiple genes from the family of UDP glucuronosyltransferase 1 *Ugt1* genes (Appendix Figure B.3). A closer look revealed that 9 genes from the *Ugt1* family such as *Ugt1a5*, *Ugt1a7c* had overlapping final exons. These overlaps are likely to happen due to similarity within gene families and would explain the existence of multi-gene transcripts

**Table 3.4:** Graphs with more than five genes in the Ensembl annotation (v91) using the exon graph approach.

| Graph ID   | Gene names                                                                                                                                                                                                 |
|------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| edge-25353 | Ctla2a,Ctla2b,4930486L24Rik,AC160962.1,CT010498.2,CT010498.1                                                                                                                                               |
| edge-30712 | Pcdha11,Pcdha4,Pcdha6,Pcdha1,Pcdha2,Pcdha7,Pcdha3,Pcdha9,Pcdha5,Pcdha8,Pcdha12,Pcdhac2,Pcdhac1                                                                                                             |
| edge-30713 | Gm42416,Pcdhgc5,Pcdhgb2,Gm37013,Gm37388,Pcdhga4,Pcdhgc3,Pcdhga10,Pcdhgb6,Pcdhgb4,Pcdhga6,Pcdhga1,Pcdhga2,Pcdhgb8,Pcdhga3,Pcdhga8,Pcdhga7,Pcdhgc4,Pcdhgb7,Pcdhga9,Pcdhgb5,Pcdhga5,Pcdhga12,Pcdhga11,Pcdhgb1 |
| edge-724   | Ugt1a10,Ugt1a5,Ugt1a6b,Ugt1a6a,Ugt1a7c,Ugt1a8,Ugt1a2,Ugt1a1,Ugt1a9                                                                                                                                         |

**Table 3.5:** Graphs with more than five genes in the mT&T assembly using the exon graph approach.

| Graph ID   | Gene names                                                                                                                                                                                                 |
|------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| edge-13994 | H2-Q4,H2-Q10,H2-D1,H2-Q6,H2-Q7,H2-Q2                                                                                                                                                                       |
| edge-15403 | Pcdha4,Pcdha5,Pcdhac1,Pcdha7,Pcdha3,Pcdha9,Pcdha12,Pcdha6,Pcdha2,Pcdha11,Pcdhac2,Pcdha1,Pcdha8                                                                                                             |
| edge-15404 | Pcdhga9,Pcdhga12,Pcdhgb8,Pcdhgc5,Pcdhga8,Pcdhga11,Gm37013,Pcdhgc3,Pcdhgb2,Pcdhga4,Pcdhgb4,Pcdhgb5,Pcdhgc4,Pcdhgb7,Pcdhgb6,Pcdhga6,Pcdhga5,Gm42416,Pcdhga1,Pcdhga10,Pcdhgb1,Pcdhga3,Pcdhga2,Pcdhga7,Gm37388 |
| edge-36203 | mt-Cytb,mt-Nd4,mt-Nd2,mt-Nd1,mt-Co1,mt-Nd4l                                                                                                                                                                |
| edge-413   | Ugt1a7c,Ugt1a2,Ugt1a6a,Ugt1a10,Ugt1a6b,Ugt1a9,Ugt1a8,Ugt1a1,Ugt1a5                                                                                                                                         |
| edge-8956  | CT010498.1,CT010498.2,Ctla2a,AC160962.1,Ctla2b,4930486L24Rik                                                                                                                                               |

found in the Ensembl annotation. A report by Zhang *et al.* has described the locus of the *Ugt1* family as a locus with a 'constant region', which is shared amongst 14 *Ugt1* genes in the mouse [157]. Interestingly, two other graphs summarised the (clustered) protocadherin  $\alpha$  and  $\gamma$  families known to occur in close proximity on the genome and often referred to as clusters of Pcdha@ and Pcdhg@ (reviewed in [158]). Furthermore, Zhang *et al.* have described them as having a similar structure to *Ugt1* with a constant (shared) and variable region.

The two individual graphs occurring only in the mT&T assembly included one graph with five mitochondrial genes and another graph with six genes from the histocompatibility 2 Q and D regions. In both cases these genes are likely either highly similar or highly complex and therefore the assembly in these regions might be challenging. The 252 individual graphs with more than two genes generated from the Ensembl annotation did not include the genes found in these mT&T graphs.

In conclusion, the graph approach to identify multi-gene transcripts revealed known examples, where the overlap between the genes has been previously described as being

biologically relevant. An increased number of graphs with multiple genes was identified in the mT&T assembly indicating the presence of novel multi-gene transcripts. The analysis performed in this section suggests that for some genes the wider locus should be considered as one unit rather than considering every gene in the locus individually. Due to the variety of mouse tissues used for the mT&T assembly, it could potentially aid the discovery of higher complexity in known fusion transcripts such as *Ugt1a* as further tissue-specific variants could be present. To assess if novel multi-gene transcripts affected the overall gene expression, all novel transcripts in mT&T-based graphs with more than one gene in the graph were classified as multi-gene. However, the gene quantification using the mT&T assembly was not strongly affected by the 948 genes containing novel multi-gene transcripts (Appendix Figure B.4), thus I assumed the presence of these in the mT&T assembly will likely not interfere with the comparison between mTEC and peripheral tissues.

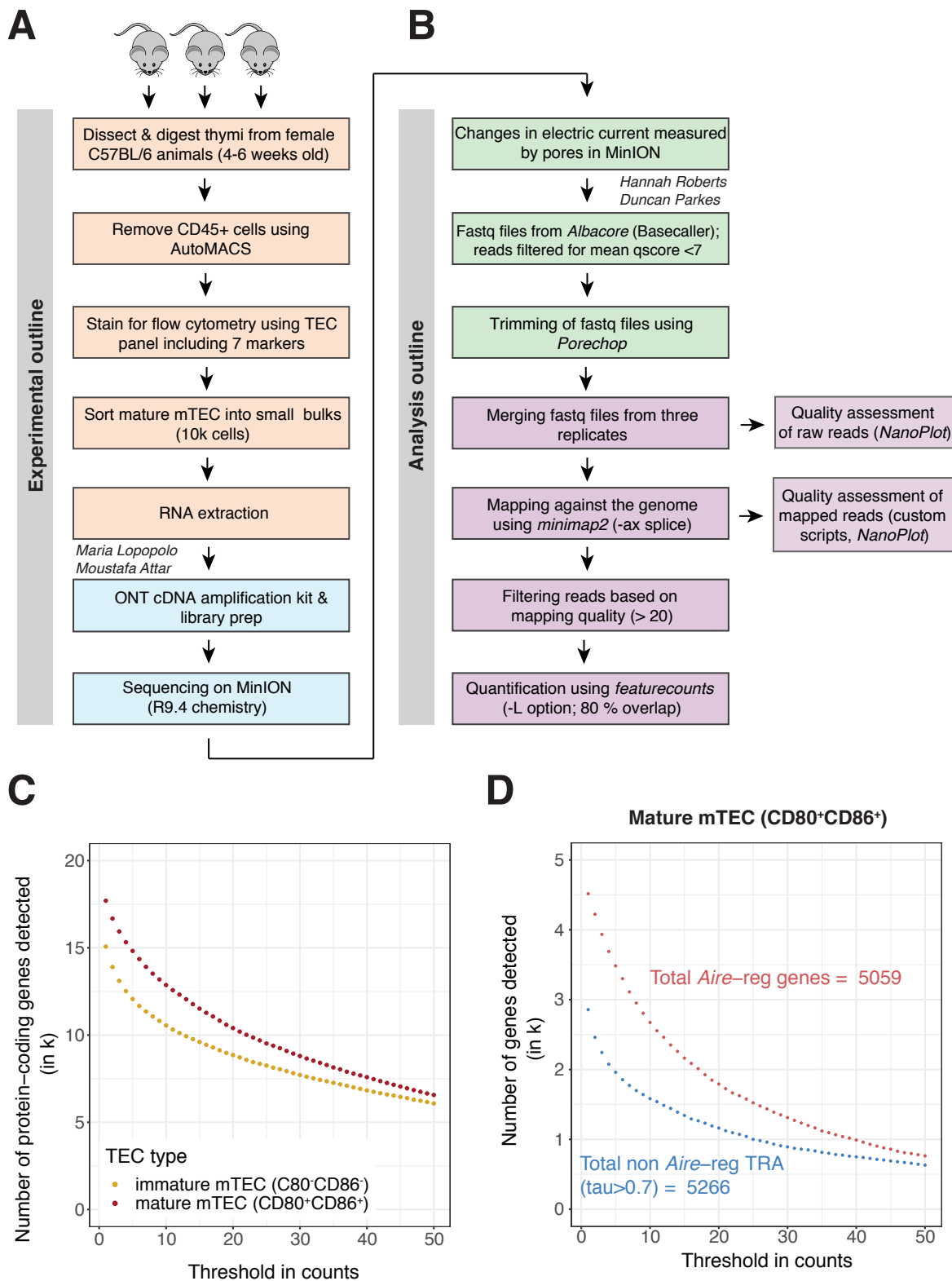


### 3.3 Validation of structures using low-input long-read RNA-sequencing of medullary thymic epithelial cells

Given the small cell numbers of mTEC subpopulations per mouse, a low-input method for long-read sequencing was needed to validate the assembled transcript structures. So far, few studies using low-input for long-read RNA-sequencing have been published. Motivated by a recently published protocol of single-cell ONT sequencing [84], I established a collaboration with Dr David Buck's group at the Wellcome Trust Centre for Human Genetics to generate long-read sequencing data from mature mTEC using ONT sequencing. The steps after the RNA extraction (library preparation and MinION loading) were performed by Moustafa Attar and Maria Lopopolo (Wellcome Centre for Human Genetics). An initial pilot run was performed on 6 k mature mTEC using the modified SmartSeq2 protocol as described by Byrne *et al.*[84]. However, the read length was not as long as expected (Appendix Figure B.7A), therefore the protocol was changed to the ONT cDNA-PCR sequencing protocol. Data was generated from 10 k immature (CD80<sup>-</sup>, CD86<sup>-</sup>) and mature (CD80<sup>+</sup>, CD86<sup>+</sup>) mTEC of two biological replicates and one pooled sample (pool of n=3 mice) (overview in Figure 3.7A, Appendix Table B.4). Using the ONT protocol, I obtained a mean read length of approximately 1,000 bp across all six samples (Figure B.7A).

#### 3.3.1 Quality assessment of long-read RNA-sequencing data

Reads generated by the ONT sequencing are typically long but have a relatively low base accuracy. In previous studies of ONT cDNA sequencing, median identities of 85 % were achieved [87]. Due to the nature of the reads, specific computational tools have to be used, which were combined into an ONT-specific analysis pipeline (Figure 3.7B). The first two computational steps namely base calling (*Albacore*) as well as the trimming of reads (*Porechop*) were performed by Hannah Roberts and Duncan Parkes (Wellcome Centre for Human Genetics). The three trimmed fastq files for each of the immature (CD80<sup>-</sup>, CD86<sup>-</sup>) and mature (CD80<sup>+</sup>, CD86<sup>+</sup>) TEC bio-replicates were joined for downstream analysis. The recently published aligner *minimap2* was used in the pipeline as it was shown to



**Figure 3.7:** Long-read sequencing of mTEC using Oxford Nanopore Technology. **A:** Outline of experimental work from primary cells to sequencing. **B:** Outline of analysis pipeline used for long-read sequencing data. **C:** Number of protein-coding genes detected in mature and immature mTEC (downsampled to same count number between the samples). **D:** Number of *Aire*-regulated genes or tissue-specific genes ( $\tau > 0.7$ ) detected in mature mTEC.

perform well on noisy long read data including ONT reads [90]. The tool *Nanoplot* was designed to assess the quality of ONT reads with respect to for instance read length or mapping quality [140]. Both merged datasets for immature and mature mTEC showed a robust correlation between the percentage identity and the average base quality (Pearson's R: 0.83 and 0.84, respectively; Appendix Figure B.7C). The percent identity is a measure of accuracy, which combines the edit (Levenshtein) distance between the alignment and the reference (stored in the 'NM' tag of bam file) normalised by the length of the alignment (excluding clipped bases). The median identities for immature and mature mTEC reported by *Nanoplot* were 89.6 % and 87.2 %, respectively. Another measure of error rate between reference and read would be the 'gap-compressed identity', which only penalises each gap in the alignment with 1 instead of the length of the respective gap (Appendix Figure B.7D). Consistent with the median identity, the median error for immature mTEC (7.7 %) was lower than for mature mTEC (9.5 %).

Apart from read length and mapping quality, another metric of interest was the number of splice junctions covered by individual reads. The majority of reads spanned less than 10 junctions, while around 20 % of reads spanned more than 5 splice junctions in both immature and mature mTEC (Appendix Figure B.7B). The number of protein-coding genes detected from the merged ONT datasets were comparable to the Illumina dataset with around 80 % of protein-coding genes detected in mature mTEC (Figure 3.7C). Examining the number of *Aire*-regulated and TRA ( $\tau > 0.7$ ; defined in Appendix Figures B.8 to B.6) revealed almost 4.5 k of *Aire*-regulated genes (90 %) and 2.8 k non *Aire*-regulated TRA (54 %) in mature mTEC. Taken together, the assessment of gene expression suggested a good coverage in the ONT datasets.

### 3.3.2 Building a transcriptome assembly from long-read RNA-sequencing data

A transcriptome assembly from ONT data was desired for two main reasons: Firstly, such an assembly could be used to validate the novel transcript structures identified in the short-read (Illumina) based mT&T assembly. Secondly, I wished to determine,

whether RNA-sequencing using the ONT platform is able to identify novel structures not seen with Illumina short-read sequencing. Given the possibility of 'promiscuous splicing' in TEC, it was plausible that structures exist in these cells that might not be easily discoverable with a short-read, reference-guided assembly strategy.

For both objectives outlined above, an assembly from the long-read ONT datasets is indispensable. Due to the high error rate in ONT data, a combined assembly tool for ONT and Illumina data would be the ideal solution. However, at the time of analysis no ONT-specific tools were available to perform a transcriptome assembly or combined transcriptome assembly. Tools for genome assembly such as *Canu* or *RNAspades* were tested but did either not work or not give meaningful results [91, 159]. Instead of a long-read specific tool, the general assembler *stringTie* was therefore used to generate a transcriptome assembly from the ONT dataset.

### 3.3.2.1 Assessment of reference-guided assembly revealed frequent UTR extensions in novel transcripts

The reference-guided *stringTie* assembly from the merged ONT datasets (immature and mature mTEC separately) was guided by the Ensembl annotation (v91) to match the approach taken for the generation of the mT&T assembly (Section 3.2.1). Subsequently, the resulting assembly was compared to the annotation using *gffcompare*. The class code selection and renaming ('\_O' for novel transcripts) was performed consistent with the approach in Section 3.2.2. The guided ONT assembly identified 10.2 k novel transcripts in mature mTEC, while 63.5 k known transcripts could be re-assembled (Table 3.6). A similar number of transcripts with unknown origin (10.3 k) could be assembled using the ONT data from mature mTEC (CD80<sup>+</sup>, CD86<sup>+</sup>).

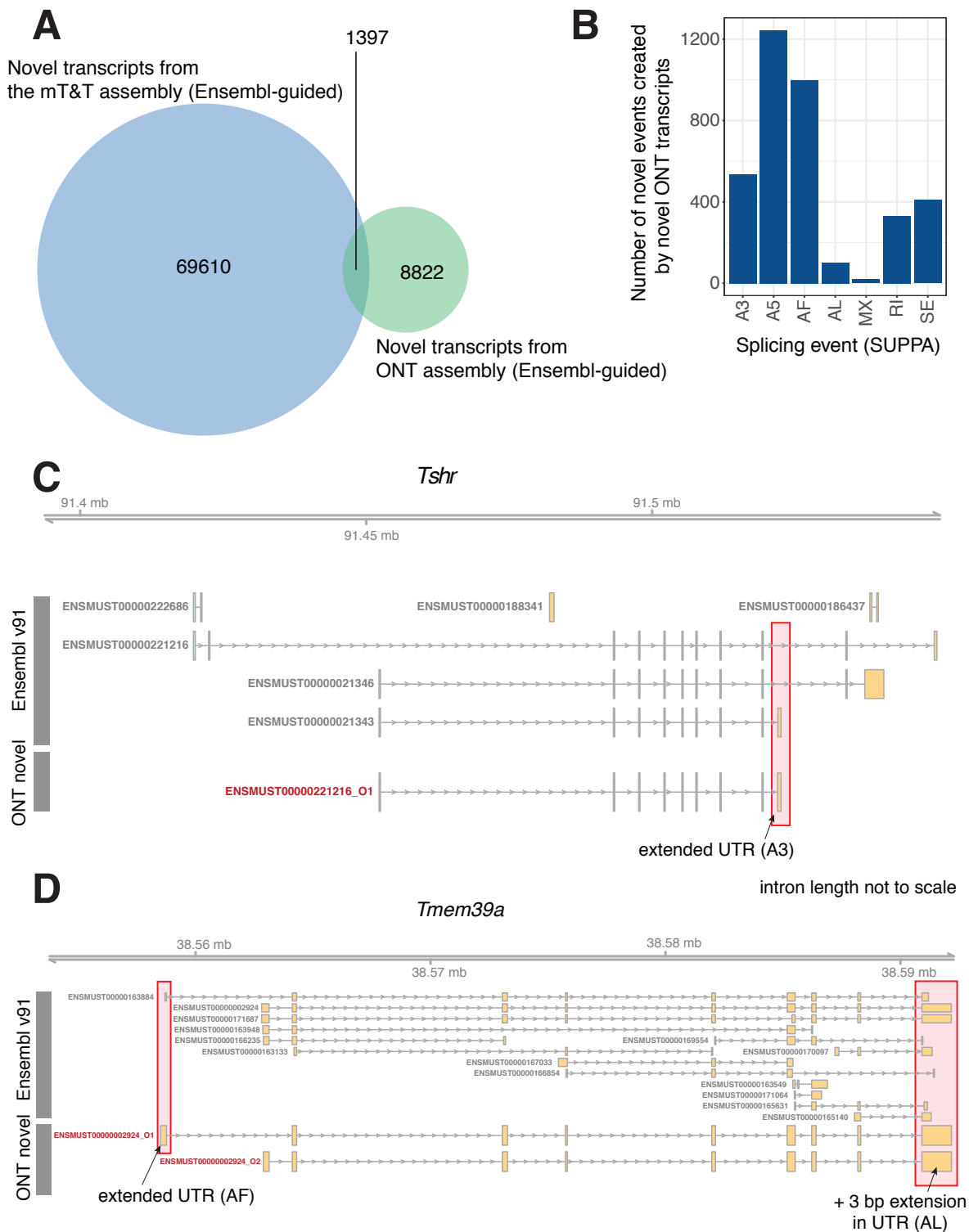
In order to compare the novel transcripts in the assemblies from the two sequencing technologies, the novel transcripts showing full intron-chain match between the mT&T assembly and the ONT assembly were evaluated. Surprisingly, only 13.7 % (1.4 k novel transcripts) of the novel transcripts identified from the ONT-reads showed full-intron chain matches with the mT&T novel transcripts (Figure 3.8A). The number of novel

transcripts in the mT&T assembly is as expected much larger due to the presence of novel transcripts assembled in other mouse tissues.

Inspection of examples of novel transcripts from the ONT assembly revealed frequent occurrences of extended UTR regions or splice sites being moved by a small number of base pairs. To get an overview of the types of novel transcripts generated in the ONT assembly, SUPPA was applied to assess the event types added by novel transcripts (Figure 3.8B). The majority of novel transcripts occurred in the categories alternative 5' splice site, alternative first exon and alternative 3' splice site confirming the trend observed in Integrative Genome Browser (IGV). Two genes containing novel transcripts falling into these categories are shown in Figure 3.8C, D. A novel transcript in *Tshr* resulted in a novel 3' splice sites in the final exon, thus producing a larger final exon. The two novel transcripts in *Tmem39a* contained extended first and last exons, which are likely UTR regions. The examples of novel ONT transcripts shown here could be likely produced by polymerase run-on (in case of extended UTR regions) or by mapping error (in case of splice sites). Furthermore, a high fraction of assembled transcripts did not represent interesting novel transcripts, that could potentially produce novel epitopes. To focus on novel transcripts with potential to produce novel epitopes and to exclude ONT mapping artefacts, only ONT reads containing novel splice junctions that were supported by Illumina data, were considered for the assembly.

**Table 3.6:** Summary of transcripts and their class codes assigned by *gffcompare* in the guided *stringTie* assembly from the mature mTEC (CD80<sup>+</sup>, CD86<sup>+</sup>) ONT dataset using Ensembl (v91) as a reference.

| Category | Class code        | Number of transcripts<br>Ensembl-guided |
|----------|-------------------|-----------------------------------------|
| Known    | =                 | 63,527                                  |
| Novel    | j + k + m + n + o | 10,219                                  |
| Unknown  | u + y             | 10,350                                  |
| Total    |                   | 84,096                                  |



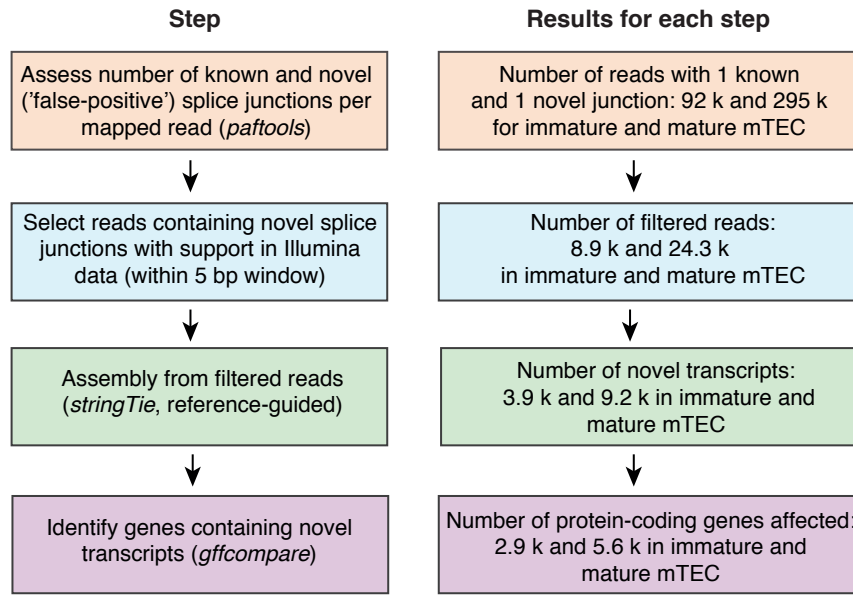
**Figure 3.8:** Stand-alone assembly of ONT sequencing reads. **A:** Venn diagram showing novel transcripts from the ONT assembly with full-intron chain match in the mT&T assembly. **B:** Splicing event types generated by novel ONT transcripts (guided by Ensembl). Each event was analysed separately, therefore one novel transcript can occur 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. In **C** and **D**, examples of novel transcripts assembled from ONT reads with stand-alone *stringTie* assembly are shown. The parts that are identified as novel are highlighted in red.

### 3.3.2.2 Identification of novel transcripts with potential novel epitopes in an assembly from pre-filtered ONT reads with Illumina support

To generate an assembly from ONT reads with high-confidence skipped exon events, the ONT reads were initially filtered for their potential to represent novel, potentially coding transcripts. The tool *paftools* was employed to assess whether junctions in a read were consistent with respect to a given annotation or were considered 'false-positives' (Figure 3.9). Such 'false-positive' splice junctions were subsequently treated as novel junctions (after filtering for proximity to known junctions) and 295 k reads contained at least one known and one novel junction for mature mTEC (92 k for immature mTEC). In a next analysis step, each novel junction was assessed for support in the Illumina dataset resulting in 8.9 k and 24.3 k reads for immature (CD80<sup>-</sup>, CD86<sup>-</sup>) and mature mTEC (CD80<sup>+</sup>, CD86<sup>+</sup>), respectively. A 5 bp window around the splice junction was allowed due to read errors in the ONT data. Finally, the filtered reads were subjected to a reference-guided *stringTie* assembly revealing 9.2 k novel transcript fragments from 5.6 k protein-coding genes in mature mTEC (Figure 3.9A). In the assembly from immature mTEC, novel transcripts were only found in 2.9 k protein-coding genes.

Comparison of the event types created by novel transcripts in this assembly from the filtered ONT reads to the previous stand-alone ONT assembly revealed a larger fraction of novel transcripts with skipped exon events indicating that more transcripts with potential novel epitopes could be assembled (Figures 3.8B and 3.10A). The novel transcripts in this approach were not a subset of the previous ONT assembly approach as *stringTie* was run with a lower coverage filter (0.001) due to the requirement for junctions to also be present in the Illumina dataset.

In order to understand what types of genes are affected by novel transcripts in the assembly from selected ONT reads, geneset over-enrichment analysis was performed on the 5.6 k genes with novel transcripts in mature mTEC. Overrepresented categories included 'alternative splicing', 'ER to cytosol transport' and 'antigen processing and presentation of endogenous antigen' (Figure 3.10B). Assessing the *tau* values of the

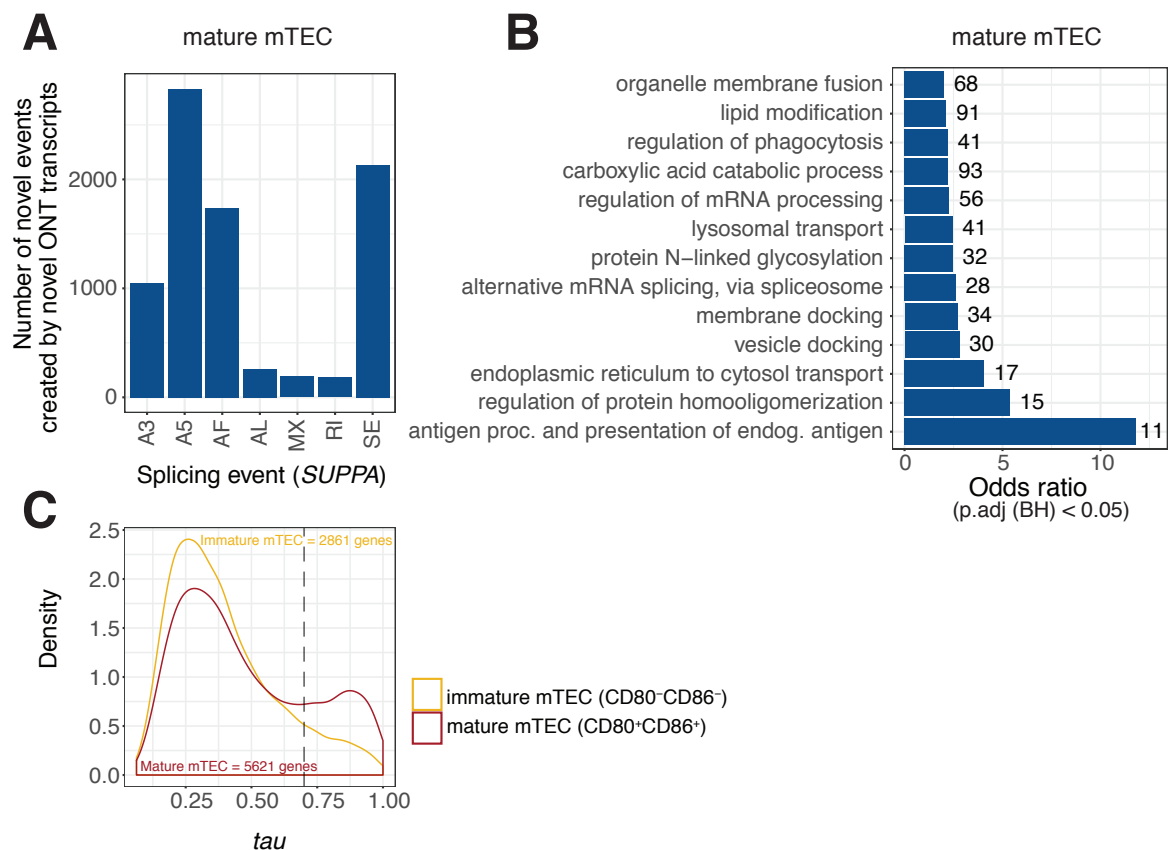


**Figure 3.9:** Outline to generate ONT assembly from filtered reads. Schematic overview of *paftools* assembly approach (left side) including the number of reads or transcripts retained at each step of the assembly process (right side).

affected genes in assembly from mature mTEC revealed that 1.3 k (23 %) out of 5.6 k protein-coding genes were tissue-restricted ( $\tau > 0.7$ ). In contrast, only 10 % of genes containing novel transcripts were tissue-restricted in the assembly from immature mTEC. The investigation of the distribution of  $\tau$  values between the two samples also showed a higher density of genes with higher  $\tau$  values in mature mTEC (Figure 3.10C).

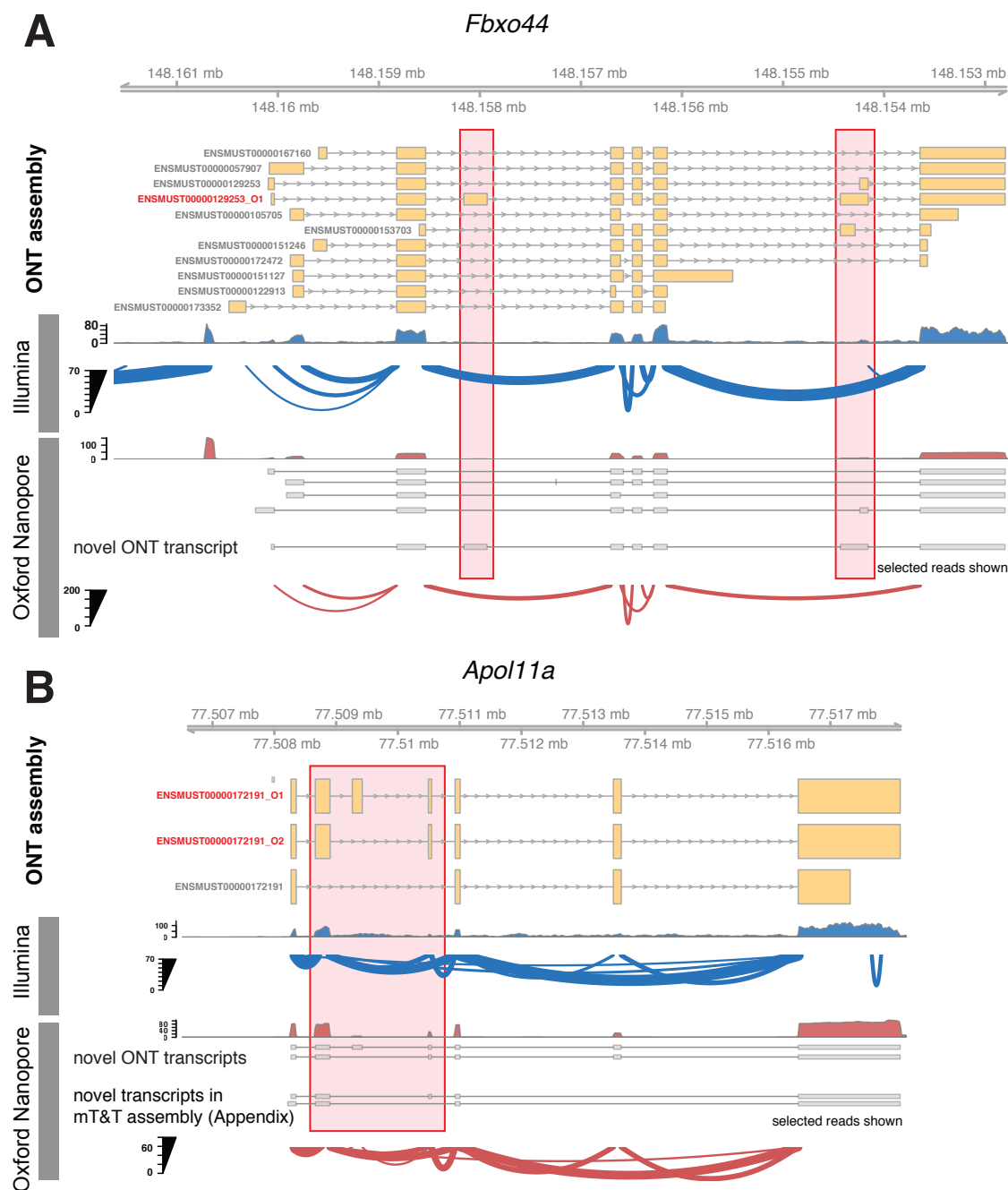
The ONT assembly from filtered reads revealed more meaningful novel transcripts with novel exons or skipped exon events than the stand-alone assembly from all ONT reads. The novel ONT transcript in *Fbxo44* contained a novel exon and a novel retained intron (Figure 3.11A). *Fbxo44* is part of the FBA family of ubiquitin ligase subunits (F-box proteins). The expected mechanism of these proteins involves the targeting of glycoproteins for proteasomal degradation. However, FBXO44 was the only family member that was reported to fail to bind to glycans [160]. All family members were able to bind to canonical components of the protein degradation complex (SCF). The family member FBXO3 was shown to promote the ubiquitinylation of AIRE *in vitro* [161]. The addition of ubiquitin to AIRE resulted in its degradation, but also in an





**Figure 3.10:** Results of the ONT assembly from filtered reads. **A:** Analysis of splicing event types within novel transcripts in the ONT assembly from mature mTEC (CD80<sup>+</sup>, CD86<sup>+</sup>). Each event was analysed separately, hence one novel transcript can occur in multiple categories. **B:** Odds ratios of over-enrichment testing of genesets from biological process gene ontologies. Fisher's Exact tests were performed using protein-coding genes harbouring novel transcripts in the ONT assembly from mature mTEC data. Representative categories are shown. **C:** Distribution of *tau* values for genes containing novel transcripts in immature and mature mTEC. The y-axis shows the density for each sample across the *tau* range from 0 to 1. Genes with *tau* > 0.7 were defined as tissue-restricted genes. The number of protein-coding genes with novel transcripts in each assembly is indicated in the plot. SE = skipped exon, RI = retained intron, MX = mutually exclusive exon, A3/A5 = alternative 3'/5' splice site, AF/AL = alternative first/last exon.

increased interaction between AIRE and p-TEFb, which were then more frequently recruited to *Aire*-regulated gene promoters. Three novel exons were identified in *Apol11a* producing two novel transcript structures (Figure 3.11B). One of these novel transcripts was independently found in the mT&T assembly (Appendix Figure C.4A). The family of Apolipoproteins includes 12 genes and 1 pseudogene on mouse chromosome 15 [162]. The locus containing all apolipoproteins in the mouse genome has very likely been subject to multiple gene duplication events [162]. Some apolipoproteins are secreted and have



**Figure 3.11:** Examples of novel transcripts from ONT assembly. In **A** and **B** examples for novel transcripts in *Fbxo44* and *Apol11a* are depicted. The novel transcripts were assembled from filtered ONT reads and are highlighted in red. Sashimi plots were generated from RNA-sequencing data from mature mTEC at 200 M reads (blue) or from merged ONT datasets for mature mTEC (CD8<sup>+</sup>, CD8<sup>+</sup>; 12 M reads total; red). A minimum of 5 reads is required for a splice junction to be plotted in the Sashimi plot.

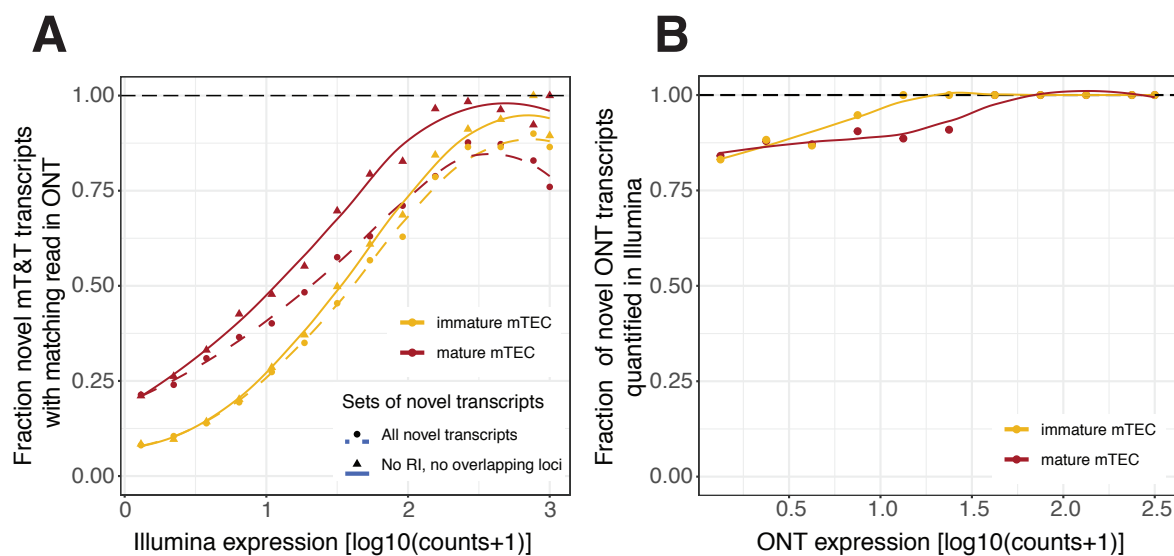
functions in lipid metabolism, while others are likely involved in antiviral responses. In addition to the validation of two novel mT&T transcripts with full-length ONT reads, two novel ONT transcripts were identified.

### 3.3.3 Validation of novel transcripts between short and long-read sequencing technologies

In order to compare novel transcripts between short-read Illumina sequencing data and long-read ONT data, an assembly was generated from one technology and quantified using the other technology (*featureCounts*, 80 % read overlap required). If a transcript could be quantified using the other technology, it is assumed to be validated (Figure 3.12). The ONT dataset was generated from (CD80<sup>+</sup>, CD86<sup>+</sup>) mTEC, which was a subset of the MHCII<sup>high</sup> mTEC in the Illumina RNA-sequencing dataset. Similarly, the ONT samples of immature mTEC represented CD80<sup>-</sup>, CD86<sup>-</sup> mTEC.

In the first part, the count data for novel transcripts from Illumina was binned by expression level and the fraction of transcripts detected by ONT were calculated per bin (Figure 3.12A). Initially, all novel transcripts were included, resulting in 82 % validated transcripts with more than 100 counts in the Illumina quantification of mature mTEC (dashed lines in Figure 3.12A). However, the non-validated transcripts in the highest expression bin mainly overlapped loci on the same or opposite strand or transcripts containing retained introns. Removing such novel transcripts from the analysis increased the fraction validated to 97 % in mature mTEC.

The equivalent analysis using the Illumina data to quantify the ONT assembly from filtered reads revealed a higher fraction of validated novel transcripts in the lower expression bin. The validation curve started from 88 % of novel ONT transcripts validated in the lowest expression bin and quickly plateaued at 100 % validated (Figure 3.12B). This analysis indicated that the read depth of the short-read data was more adequate for validation of the ONT transcripts, but that further long-read sequencing may be needed to validate the mT&T structures.



**Figure 3.12:** Validation of novel transcripts between technologies. **A:** Validation of novel transcript from the mT&T assembly with ONT reads (binned by expression in Illumina quantification). Only mT&T novel transcripts were included, that pass the TPM thresholds from the npIDR approach, determined for mature mTEC and immature mTEC, respectively. **B:** Equivalent analysis to A, where novel transcripts from the ONT assembly were quantified using Illumina data (200 M reads). No filtering of novel ONT transcripts was included. Quantification for both technologies was performed using *featureCounts* and requiring 80 % of the reads to overlap with transcripts (allowing for error in the ONT data).

## 3.4 Discussion

Medullary thymic epithelial cells have previously been shown to express a high fraction of protein-coding genes due to the process termed promiscuous gene expression. The gene expression patterns have been analysed in detail, while less is known about the splicing in TEC. In order to compare the splicing landscape in a variety of peripheral tissues and mTEC, a common reference annotation including splice variants from all samples was needed. The work reported in this chapter aimed to provide such a transcriptome assembly based on 22 deeply-sequenced mouse ENCODE tissues and deep RNA-sequencing datasets from mTEC.

The assembly from mouse mTEC and peripheral tissues presented here aimed to generate an assembly with high-confidence transcripts rather than a complete reference. The limitations of the mT&T assembly include short read length of 76 bp and downsampling to 60 M reads (30 M read pairs) per sample for quantification of the per tissue assemblies, which might not have captured all the information available in the mTEC RNA-sequencing data and the five deeply-sequenced mouse ENCODE tissues (Appendix Table B.3). However, I accepted this limitations in order to increase the number of tissues included (Appendix Table B.2). In a recent publication by Danan-Gotthold *et al.*, the authors used RNA-sequencing data subsampled to 50 M reads for the analysis of TRA junctions, which is comparable to the depth used here.

As part of the initial quality assessment of the mT&T assembly, multi-gene transcripts were assessed as these were shown as a possible artefact in *stringTie*-based assemblies (Section 2.1.3.2, Section 3.2.3.3, [125]). Furthermore, it is unclear how many multi-gene transcripts exist in Ensembl annotation and how likely it is that such transcripts are assembly artefacts or biologically relevant. To identify gene loci with multiple overlapping genes, graphs of overlapping genes were built. The identification of known multi-gene transcripts suggested that the graph-based approach was suitable to examine transcripts overlapping multiple genes. The gene locus around *Ugt1a6a* is one example of multi-gene

transcripts found in the Ensembl annotation (Appendix Figure B.3), which confirms that these exist in published annotation sets. Some publications have highlighted the possible function of such complex loci as being a gene cluster with a 'constant region' and tissue-specific first or last exons [157]. Zhang *et al.* further found that the constant and variable regions in the *Ugt1* gene locus were conserved between mouse, human and rat. The locus for UDP glucuronosyltransferases is arranged as an array with tissue-specific expression pattern in all three species [157]. Using the graph-based approach, a higher number of multi-gene transcripts was found in the mT&T assembly compared to Ensembl annotation. However, only 28 % of multi-gene transcripts could be validated with full-length reads from long-read sequencing data, thus some of these structures may be artefacts of the Illumina assembly. Importantly, transcriptome quantification was not strongly affected by the assembled multi-gene transcripts (Appendix Figure B.4).

In the second part, long-read sequencing data was used to validate novel transcripts discovered in the mT&T assembly as well as to identify potentially novel transcripts that might not be discoverable from short-read sequencing. The generated data from three merged MinION runs showed good gene coverage despite the limited number of reads (12 M reads, Appendix Table B.4). However, the known issue of relatively high base-calling error rate that is associated with such ONT data was often observed to result in the mismapping of splice sites. The overall error rate was less than 10 % as defined by gap-compressed identity. In comparison to previous versions of ONT chemistry the error rate has improved (up to 15 % in earlier chemistries). The median aligned base identity of both merged ONT datasets (>87 %) was higher than in another previous study (85 %) [87]. In general, it should be noted that ONT data is unstranded and the mapping direction has to be inferred by the mapper, thus resolving transcripts overlapping on the opposite strand could be problematic using ONT data.

A *stringTie* assembly from all ONT reads was not successful in clearly answering the question about novel transcripts in ONT data. Therefore, I focussed on novel skipped exon events and aimed to filter ONT reads with Illumina support for the novel splice junctions

prior to assembly. Although the pre-filtered reads resulted in a more reliable assembly with interesting novel transcripts, the pre-filtering of reads limited the repertoire of ONT reads to those that could produce multi-exonic novel transcripts with a novel splice site. Moreover, some options such as allowed window size around the splice site or number of novel splice sites in a read had to be selected and could potentially influence the analysis. Despite these limitations, the assembly of the filtered ONT reads resulted in an assembly that contained novel exons, which could potentially produce novel epitopes. The overall assembly from the filtered ONT dataset revealed 9.2 k novel transcripts with the majority including events associated with alternative 5' splice sites or skipped exons (Figures 3.9 and 3.10). The geneset over-enrichment analysis revealed interesting categories such as 'antigen processing and presentation', which was also amongst the categories enriched in the novel TEC-specific mT&T transcripts (Section 4.4 and Figure 3.10A). An example of a gene with novel ONT transcripts (*Apol11a*) was also found to have multiple novel transcripts in the mT&T assembly (Appendix Figure C.4A). Two of the TEC-specific mT&T novel transcripts for this gene were confirmed with ONT data. However, one of the two novel ONT transcripts had not been identified with an Illumina assembly suggesting that ONT reads can reveal additional information about transcript structures (Figure 3.11B). In comparison to a previously published study using ONT RNA-sequencing (from cDNA), I identified a higher number of novel transcripts in this study. Byrne *et al.* identified around 1,000 events with either alternative transcriptional start/end sites or alternative splicing events in single B cells using the SmartSeq2+MinION approach and a custom pipeline called *Mandalorion* [84]. The number of novel transcripts identified here is around 10-fold higher. However, the other report used material from single cells and had a lower number of ONT reads (8.5 k to 100 k reads per cell). In a further publication based on the human cell line GM12878, 30 k novel isoforms were identified using direct RNA-sequencing on the MinION (10 M reads in total) [87].

The last section of this chapter showed that most of the highly expressed novel mT&T transcripts can be validated with matching ONT reads (Figure 3.12). However, only 80 % of the ONT read was required to overlap with a given mT&T transcript, thus the novel

transcripts might not be validated with full-length ONT reads. Interestingly, the removal of transcripts containing retained introns or overlapping loci (on the same or opposite strand) resulted in an increased validation rate of mT&T novel transcripts with ONT reads (Figure 3.12). Although the validation rate of ONT novel transcripts with short Illumina reads was very high, the validation with Illumina reads may involve only partial coverage of the novel transcripts. Furthermore, a high validation rate would be expected as the ONT reads were filtered for partial Illumina coverage prior to assembly.

## Conclusion

Taken together, the work in this chapter resulted in a conservative, best-practise transcriptome assembly from short-read RNA-sequencing data from peripheral mouse tissues and mTEC. The generated mT&T assembly is an important pre-requisite for the comparison of transcripts across mouse tissues (Chapter 4). Further analysis of the loci with multi-gene transcripts could extend the number of known loci and the understanding of the function of such loci in the mouse genome. The generation of high-quality long-read sequencing data will also be an additional resource for the community working on TEC. The ONT assembly added to the available information on the TEC transcriptome and - where coverage allowed - a high fraction of novel transcripts from the mT&T assembly could be validated using long-read sequencing data. The comparison of the assemblies from long- versus short-read RNA-sequencing data suggested that even more novel transcripts could potentially be discovered with long-read RNA-sequencing datasets.



# 4

## Description of the splicing landscape in thymic epithelial cells in comparison to peripheral tissues

### Contents

---

|            |                                                                                                                   |            |
|------------|-------------------------------------------------------------------------------------------------------------------|------------|
| <b>4.1</b> | <b>Background . . . . .</b>                                                                                       | <b>92</b>  |
| <b>4.2</b> | <b>Assessment of transcriptome complexity in TEC . . . . .</b>                                                    | <b>94</b>  |
| 4.2.1      | High number of unique splice junctions detected in mature mTEC                                                    | 94         |
| 4.2.2      | Highest fraction of isoforms per protein-coding gene produced by mature mTEC . . . . .                            | 95         |
| 4.2.3      | More extensive annotated splicing in TEC compared to peripheral tissues . . . . .                                 | 97         |
| 4.2.4      | Description of annotated known splicing events in TEC at the population and single-cell level . . . . .           | 98         |
| <b>4.3</b> | <b>Analysis of peripheral representation of splicing in TEC . .</b>                                               | <b>102</b> |
| 4.3.1      | Highest fraction of tissue-restricted transcripts detected in TEC compared to the testis and cerebellum . . . . . | 102        |
| 4.3.2      | Transcripts specific to tissues with frequent autoimmunity in APS-1 patients underrepresented in TEC . . . . .    | 103        |
| 4.3.3      | Incomplete representation of neuronal-specific microexons in mTEC . . . . .                                       | 105        |
| <b>4.4</b> | <b>Characterisation of novel transcripts in TEC . . . . .</b>                                                     | <b>107</b> |
| 4.4.1      | Description of novel transcript examples in TEC . . . . .                                                         | 109        |
| 4.4.2      | Prediction of the impact on protein structure for TEC-specific novel transcripts . . . . .                        | 111        |
| <b>4.5</b> | <b>Discussion . . . . .</b>                                                                                       | <b>113</b> |

---

## 4.1 Background

Alternative splicing is the process by which different exons are spliced into a transcript in a context-specific manner (Section 1.3). The production of such alternative isoforms results in an increase in transcriptome complexity, which might as a consequence increase the proteome diversity. In the context of thymic epithelial cells, alternative splicing is an interesting subject to study because the degree to which peripheral splice variants must be reproduced in order to sufficiently select thymocytes and achieve self-tolerance is not fully understood. TEC are known to express a large set of tissue-restricted antigens (TRA) via promiscuous gene expression, which is largely regulated by the transcriptional regulator Aire (Section 1.2.3). Such TRA are presented to developing thymocytes to avoid the escape of auto-reactive T cells into the periphery. Epitopes for T cell receptors could potentially contain splice junctions or alternatively spliced regions. Thus, it might be essential for TEC to express not only the whole protein-coding gene repertoire but also protein-coding transcript repertoire for the efficient negative selection of T cells. However, research into the details of alternative splicing in TEC has been limited so far (Section 1.4).

Keane *et al.* suggested that splicing is a 'promiscuous process' resulting in the complex transcriptome observed in mature mTEC compared to other epithelial cell types [101]. The authors used entropy to assess splicing complexity in epithelial cells and nine other mouse tissues. Medullary TEC did not only have the highest splicing entropy, but also the highest number of tissue-specific isoforms. These tissue-restricted isoforms originated, to a varying degree, from all of the peripheral tissues examined. Taken together, the authors concluded that splicing is a promiscuous process with a likely role for AIRE (Section 1.4.1). With respect to the transcriptome complexity in TEC, Danan-Gotthold *et al.* further commented that in mature mTEC (MHCII high) the highest fraction of alternatively-spliced genes was detected [103]. In addition, in this study the fraction of splice junctions per detected gene was found to be higher in mature mTEC compared to all tissues (even brain and testis). Danan-Gotthold *et al.* identified one example, namely

*Fyn*, where both tissue-restricted isoforms were detected in mTEC.

The importance of alternative splicing in mTEC was shown in an inducible multiple sclerosis mouse model, where the absence of an exon in *Plp* in the thymus resulted in T cells in the periphery with reactivity to the specific exon (Section 1.4.2, [108]). Upon challenge of isolated T cells with a peptide derived from a region within the missing exon (PLP<sub>139-151</sub>), proliferation of the T cells was observed for SJL/J mice but not C57BL/6 mice. The authors further mapped a strain-specific MHCII-restricted epitope of PLP, which was only present in SJL/J mice.

From previous publications, the completeness of the splicing landscape in TEC compared to peripheral tissues was not comprehensively described. One hypothesis would be that the splicing landscape in TEC resembles a combination of all transcript structures observed in peripheral tissues. An alternative hypothesis would be that splicing in TEC is a stochastic process and that many transcript structures in TEC are hence likely to be novel and not found in the periphery. Previous studies have focussed on known splicing in mTEC, thus novel transcripts in mTEC have not been studied on a genome-wide scale in detail. Novel transcripts in mTEC could either occur in genes encoding for TRA or for other proteins. While novel transcripts in TRA would contribute to the variety of peptides presented to developing T cells, the novel transcripts in other genes could have a potential role in TEC biology.

## **Aim**

In this chapter, I aimed to improve the description and understanding of the splicing landscape in TEC using deep, stranded RNA-sequencing data. Firstly, I aimed to assess the hypothesis that splicing is a promiscuous process in mTEC. The second specific goal was to establish, how complete the TEC transcriptome is in terms of the extent to which it recapitulates peripheral transcript structures and exon usage. In addition to the comparison of the splicing landscape in mTEC with the periphery, I

aimed to characterise the potential function of novel transcripts identified using the mT&T assembly in TEC (Section 3.2).

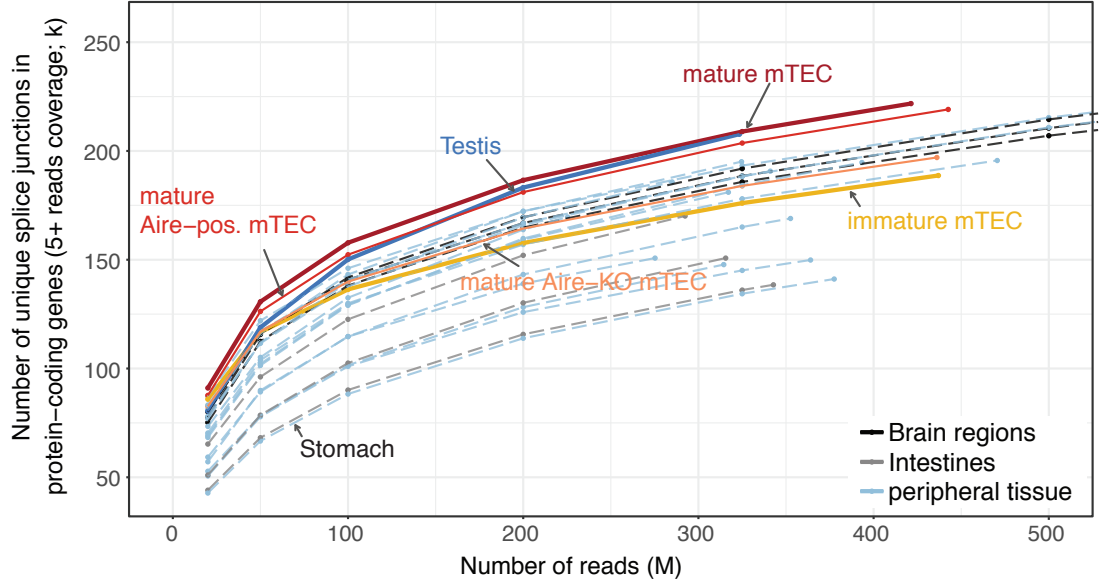
## 4.2 Assessment of transcriptome complexity in TEC

To understand the transcriptome of thymic epithelial cells compared to other peripheral tissues, I aimed to assess the transcriptome complexity using splice junctions and isoforms, the degree of annotated splicing as well as the presence of examples of tissue-specific splicing events. The dataset used here should be able to provide an improved insight over previous publications due to the availability of a greater number (n=22) of peripheral mouse tissues and of deeper RNA-sequencing data.

### 4.2.1 High number of unique splice junctions detected in mature mTEC

The initial analysis of the transcriptome complexity had two technical motivations: On the one hand, my goal was to understand whether the results were strongly affected by read depth. On the other hand, it was an important analysis to assess if the depth of sequencing was sufficient. The main biological aim was to compare the transcriptome complexity in mTEC to that in the peripheral tissues.

As a first step, I generated a single merged dataset of reads for each tissue by combining the biological replicate samples, which were then subset to a range of read depths and the number of splice junctions present was counted at each depth (Figure 4.1A). The analysis revealed that the number of unique splice junctions in protein-coding genes increased with read depth. I observed a minimal change in the slopes after 200 M reads indicating that the data were suitable for comparing splicing patterns between the tissues. The slope was similar for all tissues except the testis, in which the number of splice junctions increased more steeply at the higher read depths. The plot also revealed interesting biology: The number of unique splice junctions in protein-coding genes was higher in mature mTEC and *Aire*-positive compared to the other tissues surveyed except the testis. At 200 M reads, mature mTEC had the highest number of unique splice junctions followed by testis

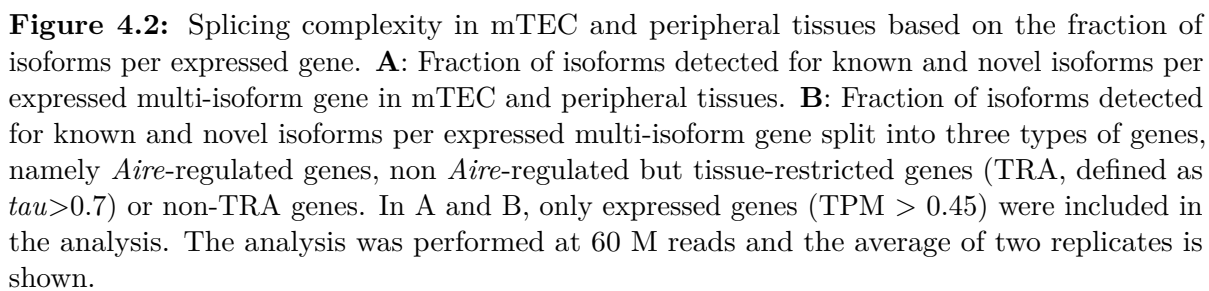


**Figure 4.1:** Splicing complexity in mTEC and peripheral tissues based on splice junctions. Saturation analysis of splice-junction detection in protein-coding genes. TEC populations are indicated with solid lines and the peripheral tissue groups with dashed lines.

and mature *Aire*-positive mTEC (Appendix Figure C.2). The high complexity in the testis could be in part due to the presence of *Aire* in this tissue [163, 44]. Promiscuously expressed genes can introduce a bias in the number of splice junctions detected because a higher number of expressed genes will likely result in a higher number of splice junctions detected. Hence, a better metric is the mean fraction of isoforms per expressed gene (Figure 4.2) as this metric is independent of the number of genes expressed.

#### 4.2.2 Highest fraction of isoforms per protein-coding gene produced by mature mTEC

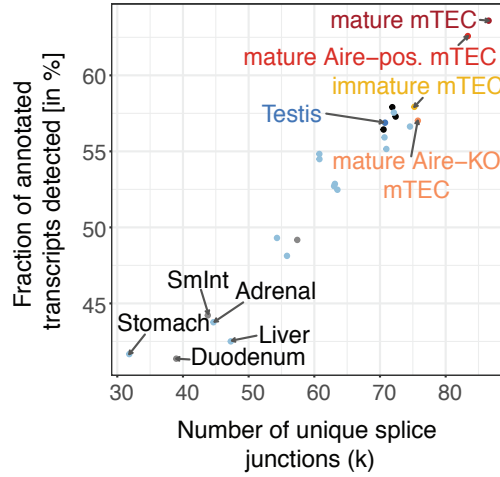
Assessing the fraction of known isoforms detected per expressed gene revealed the highest fraction for mature mTEC and *Aire*-positive mTEC followed by the three brain regions (Figure 4.2A). However, the fraction of novel transcripts detected was highest in the testis (consistent with Figure 4.7A). As previously observed, the stomach and duodenum showed a low complexity also with respect to the fraction of isoforms detected. Another tissue showing a low fraction of isoforms per gene was the adrenal.



If the genes were split into genesets depending on *Aire*-regulation and tissue-restriction, mature mTEC and other mTEC populations were amongst the top five tissues for non-*Aire* TRA or non-TRA (Figure 4.2B). However, for the most interesting category, namely *Aire*-regulated TRA, the fraction of isoforms expressed by mature mTEC was lower than for the other two gene categories, which is consistent with the previous observation of lower splicing complexity in *Aire*-induced TRA [50]. In addition, the fraction of isoforms per *Aire*-regulated TRA in mature mTEC was similar to the fractions observed for other peripheral tissues suggesting that not all possible isoforms for *Aire*-regulated TRA were produced in mature mTEC. The fractions for other mTEC populations were even amongst the lowest fractions observed in the analysis. These observations could be due to (i) the low expression level of *Aire*-regulated TRA in mature mTEC and (ii) the absence or low expression of AIRE in the other mTEC populations. Further assessments revealed that the number of genes included for mature mTEC and *Aire*-positive mTEC in the comparison for *Aire*-regulated TRA was almost twice as high as the number of genes for other tissues. This suggests that a possible issue with the presented analysis was that genes are not shared between the tissues. Differences in the distributions of isoform numbers per gene could potentially bias the quantification of the fraction of isoforms, therefore another strategy using only commonly expressed genes was applied in Section 4.3.

### 4.2.3 More extensive annotated splicing in TEC compared to peripheral tissues

While comparisons using only splice junctions or transcripts were performed in previous sections, here the number of detected transcripts was compared to the number of splice junctions for annotated transcripts. The analysis of annotated splicing in mTEC and peripheral tissues revealed the highest fraction of transcripts and highest number of splice junctions in mature mTEC followed by mature *Aire*-positive mTEC (Figure 4.3). Immature mTEC and *Aire*-KO mTEC as well as testis showed similarly high numbers of transcripts and splice junctions to the brain regions and other peripheral tissues at this read depth. Similar to previous analyses such as Figure 4.1, the stomach, duodenum and adrenal showed very low amount of annotated splicing.



**Figure 4.3:** Annotated splice variants in mTEC and peripheral tissues. The fraction of detected transcripts in annotated protein-coding genes compared to the number of annotated splice junctions in TEC and peripheral tissues are shown. The analysis was performed at 60 M reads and the average of two replicates is shown.

#### 4.2.4 Description of annotated known splicing events in TEC at the population and single-cell level

The number of well-characterised tissue-specific splicing events is extensive (for instance the review by Chen & Manley [53]). Such splicing events often involve mutually exclusive exon pairs or skipped exons, where one exon is only included in a certain tissue, while the alternative exon or the skipping event is only found in another tissue. Here, I aimed to analyse the presence of such tissue-specific splicing events in mature mTEC to assess whether this cell type would be able to produce both tissue-restricted splice variants. Using the high-depth (200 M) RNA-sequencing dataset, I confirmed the presence of both isoforms for selected examples such as Calcitonin (*Calca*), *Tjp1* and *Actn1* (Figure 4.4). In addition, I observed both isoforms for previously reported examples, namely *Oxr1* and *Fyn*, in mature mTEC (Data not shown) [101, 103].

In Tight Junction protein 1 (*Tjp1*), I found both skipping and inclusion isoform of exon 20 in mature mTEC (Figure 4.4B). The inclusion of exon 20, which encodes a phosphorylation site, in TJP1 is strongly tissue-specific [65]. In addition, I identified both the muscular and non-muscular isoforms of  $\alpha$ -Actinin (*Actn1*) in mature mTEC



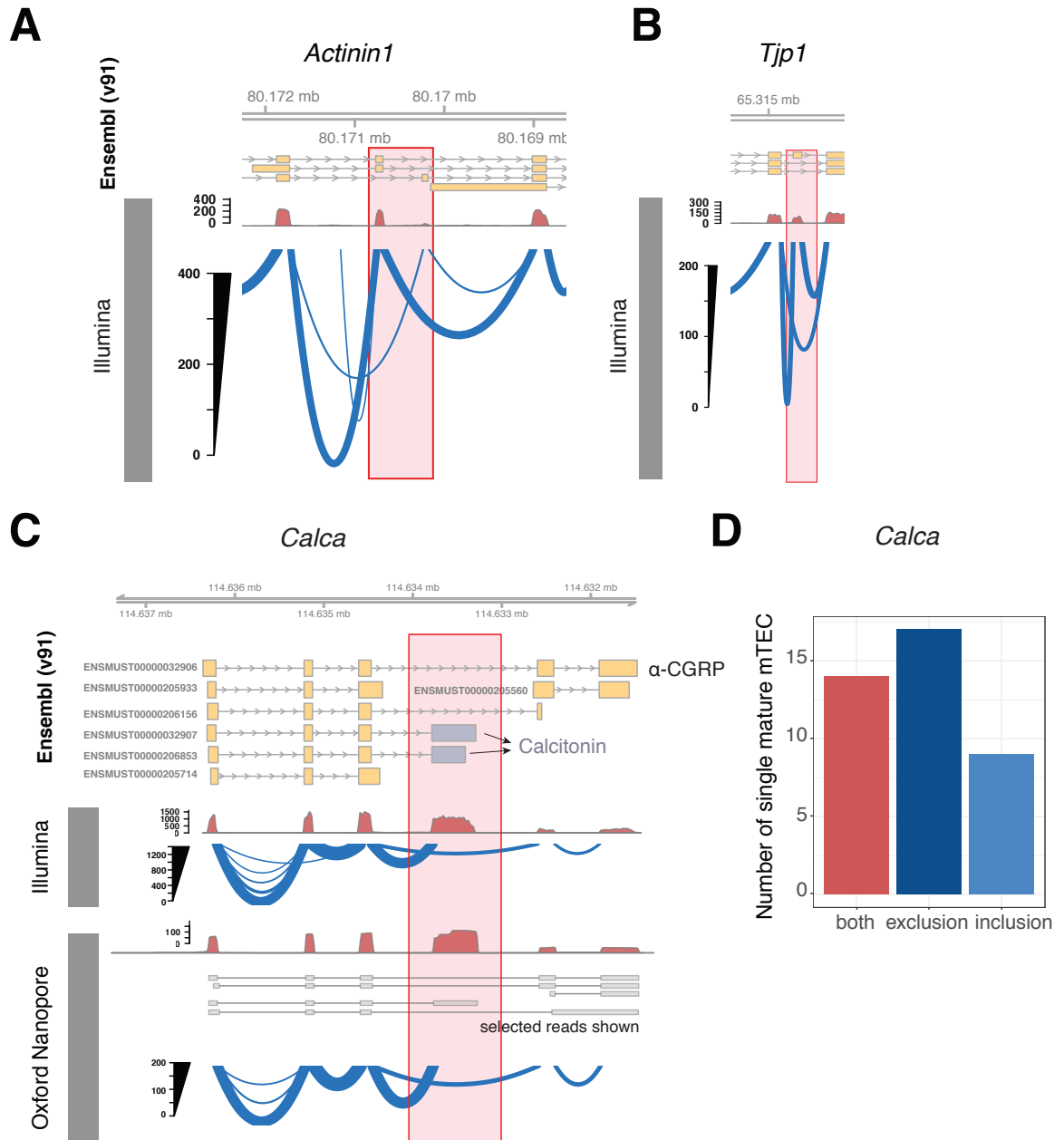
[164]. The exons creating the two isoforms of *Actn1* are a mutually exclusive exon pair that are normally expressed in different tissues (Figure 4.4C). Due to limited coverage for *Actn1* and *Tjp1* in the ONT data, these variants could not be clearly validated.

In the case of the gene *Calca*, the inclusion or skipping of an alternatively spliced exon was described to result in two different protein products (Figure 4.4A, [165]). The alternative splicing event is strongly tissue-specific and not usually found in the same tissue. In two isoforms of *Calca*, exon 4 is included resulting in a premature stop-codon and a shorter peptide product in both cases. This isoform is specific to the thyroid and produces the shorter peptide named CT [53, 165]. In contrast, the skipping of exon 4 in *Calca* forms a longer nervous system-specific peptide ( $\alpha$ -CGRP). Interestingly, isoforms skipping or including exon 4 were detected in mature mTEC with a higher proportion of reads supporting the inclusion isoform. The long-read ONT sequencing data from mature mTEC were used as a validation. The shorter isoform, a skipping isoform and one of the inclusion isoforms all had full-length reads confirming the finding from the short-read data. The presence of both *Calca* isoforms in mature mTEC could indicate that these variants are expressed for the negative selection of thymocytes.

Given the known differences in TRA expression between individual mTEC, it would be interesting to know, if examples of known isoforms that are normally restricted in expression to different tissues can be found co-expressed in single TEC. The alternative hypothesis is the expression of two isoforms in distinct subsets of mTEC. Both scenarios would result in the observed co-expression of both tissue-restricted isoforms on the population level. To assess this, I utilised deeply sequenced single-cell RNA-sequencing data from mature mTEC (average read depth of 7.4 M reads per cell; Appendix Figures D.7 and D.8). The data was generated using the SmartSeq2 protocol, which can provide reads across the whole transcript. Due to the limited coverage in single-cell data, only one example from the above could be sufficiently analysed. The gene *Calca* was detected in 40 single cells, which corresponds to 20 % of cells. In the majority of the cells only the exclusion isoform was detected (n=17, Figure 4.4D). Interestingly, however, in 35 % of

*Calca*-expressing single cells, both splice variants were found (n=14).

In summary, these observations indicate that mature mTEC have the ability to produce both versions of normally tissue-specific splice variants as has been previously reported [101, 103]. The observation that splice variants, which normally showed restricted expression to different tissues, can be detected in the same single mTEC, was particularly striking.



**Figure 4.4:** Examples of known tissue-specific alternative splicing events, which are recapitulated by TEC. **A:** Transcript models, read coverage and sashimi plots for the gene *Actinin1*. **B:** Transcript models, read coverage and sashimi plots for the gene *Tjp1*. **C:** Transcript models, read coverage and sashimi plots for the gene *Calca*. Sashimi plots in A-C were generated from mature mTEC (200 M reads) and only splice junctions with coverage of  $\geq 5$  reads are shown. The alternative splicing event is highlighted in red. **D:** Expression of tissue-restricted isoforms of *Calca* in single mature mTEC. In 35 % of single cells expressing *Calca*, both inclusion and exclusion transcripts were detected.

### 4.3 Analysis of peripheral representation of splicing in TEC

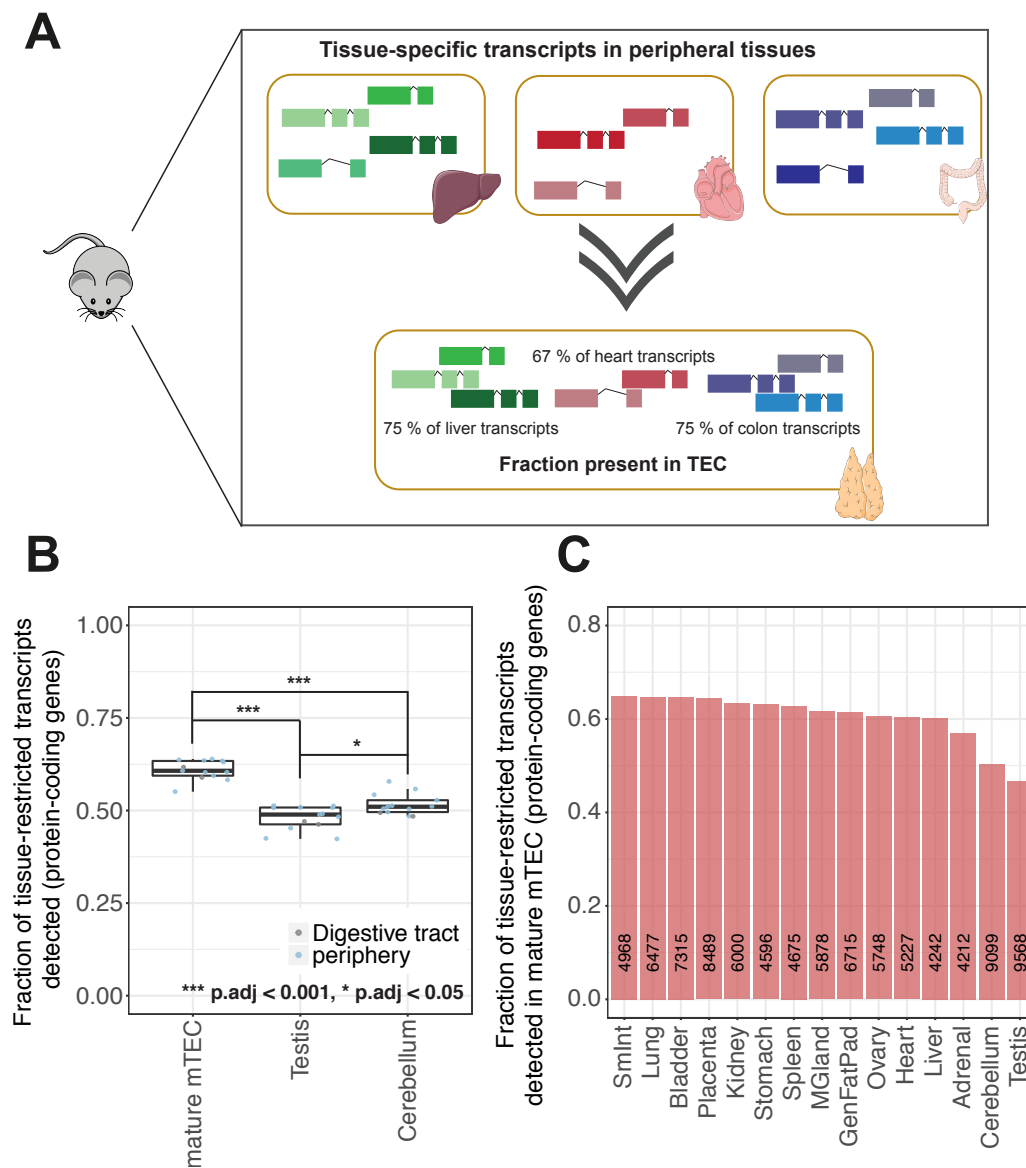
Previously (Section 4.2.2), I found a 3 % higher percentage of isoforms expressed per gene in mature mTEC (49.2 %) compared to the next highest peripheral tissue (cerebellum, 46.1 %). However, this analysis included a different number of genes in each tissue. A possible bias is that the included genes might differ in the number of isoforms present per gene. Overall, it remained unclear how complete the peripheral transcript repertoire is recapitulated in TEC. An alternative approach to the fraction of detected isoforms is to examine the representation of tissue-restricted transcripts from specific peripheral tissues in TEC.

#### 4.3.1 Highest fraction of tissue-restricted transcripts detected in TEC compared to the testis and cerebellum

To analyse the transcriptomic representation of peripheral tissues in TEC, a set of tissue-restricted transcripts was first defined for each peripheral tissue group (tissue groups as described in Appendix 2.1.9.1, scheme in Figure 4.5A,  $\tau \geq 0.9$ ). Overall, 46.8 % of these tissue-restricted transcripts were part of a tissue-restricted protein-coding gene ( $\tau \geq 0.7$ ). Secondly, the peripheral tissues with next highest splicing complexity namely testis and cerebellum (as a representative tissue for the brain) were chosen for comparison to mature mTEC. Thirdly, the presence of the tissue-specific transcripts sets was assessed in mature mTEC, testis and cerebellum for genes that were expressed in all three tissues (TPM > 0.45 defined as mean of npIDR thresholds in Figure 3.2; Figure 4.5B and Appendix Figure C.3A). Mature mTEC showed a significantly higher representation of these tissue-specific transcript sets than did either the testis or cerebellum (paired Wilcoxon rank test, Benjamini-Hochberg (BH) adjusted p-value < 0.001). Surprisingly, the cerebellum also showed an increase in tissue-restricted transcripts compared to the testis (BH adjusted p-value = 0.02).

### 4.3.2 Transcripts specific to tissues with frequent auto-immunity in APS-1 patients underrepresented in TEC

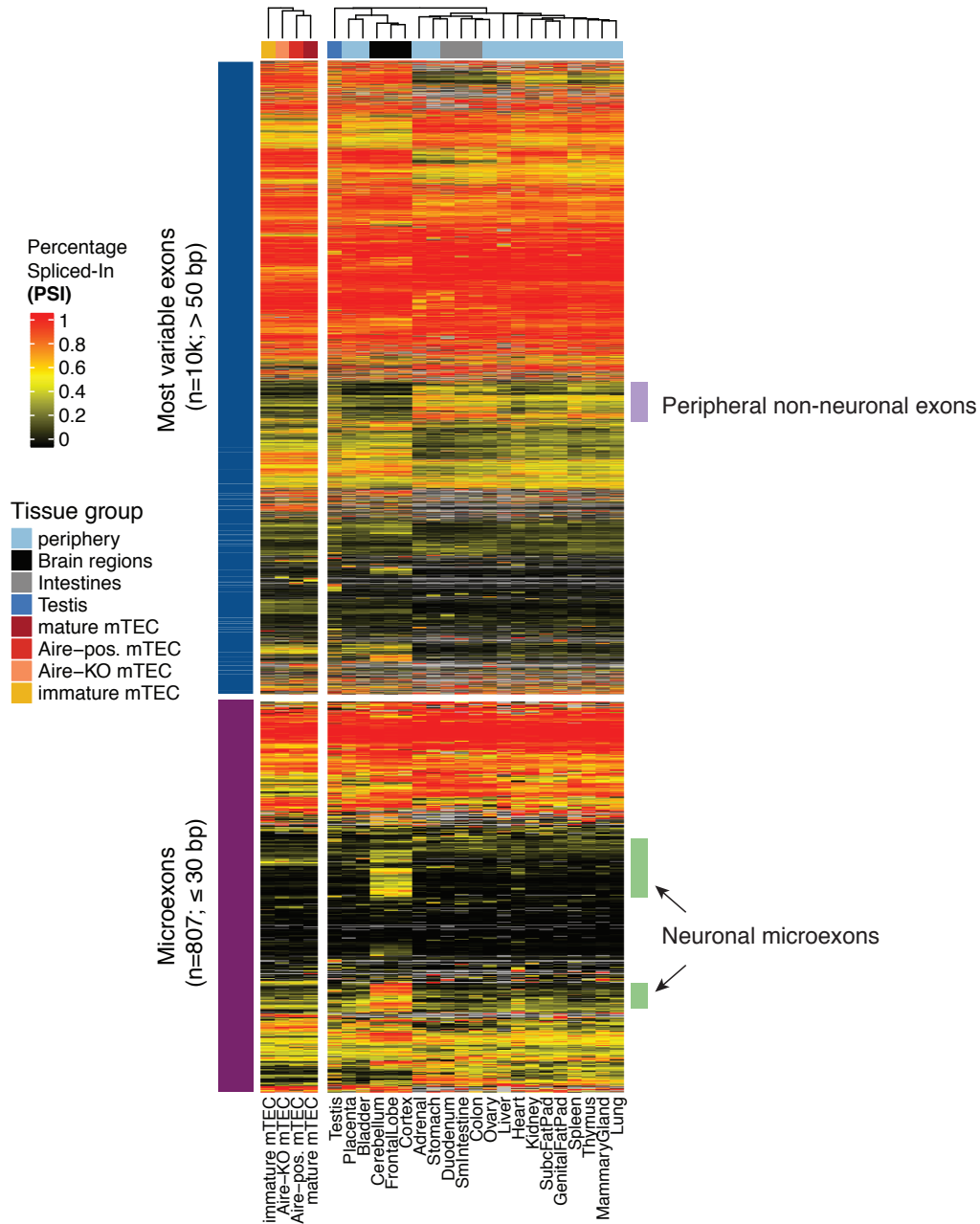
In the next step, I analysed which tissues might be under-represented in TEC with respect to splice-isoform coverage. For this analysis, all genes expressed in mature mTEC were included (TPM > 0.45). Interestingly, amongst the six most under-represented tissues in TEC (Figure 4.5C), I found the ovary, liver, adrenal and testis. These tissue are all frequent sites of autoimmunity in APS-1 patients [166, 24]. The underrepresentation of tissue-restricted transcripts from these tissues could indicate that developing T cells might not be exposed to epitopes arising from these transcripts and therefore not negatively selected for these epitopes. To ensure that the effect seen on the transcript level is not a result of differences on the gene level, the fraction of protein-coding genes detected in mature mTEC was assessed (TPM > 0.45). The genes had to contain at least one tissue-restricted transcript ( $\tau \geq 0.9$ ). The fraction of genes detected was 100 % for all evaluated peripheral tissues (Appendix Figure C.3B) indicating that the differences were present on the transcript rather than the gene level.



**Figure 4.5:** Representation of tissue-restricted isoforms in mTEC. **A:** Schematic summary of the analysis for representation of tissue-restricted transcripts in TEC. **B:** Mature mTEC expressed significantly more tissue-restricted transcripts ( $\tau \geq 0.9$ ) than does either the testis or brain (adjusted p-value  $< 0.001$  (Benjamini-Hochberg adjusted), Wilcox test). Only protein-coding genes that were commonly expressed ( $> 0.45$  TPM) in all three tissues were considered in this analysis (novel transcripts not included). **C:** Fraction of tissue-restricted transcripts ( $\tau \geq 0.9$ ) detected in mature mTEC. Only protein-coding genes with at  $> 0.45$  TPM in mature mTEC were included. The numbers in the bars indicate the number of tissue-restricted transcripts found in the respective tissues (novel transcripts not included). One representative tissue is shown per tissue group (Figure C.1).

### 4.3.3 Incomplete representation of neuronal-specific micro-exons in mTEC

Given the high but incomplete and variable recapitulation of splice isoforms from peripheral tissues in mTEC, my goal was to understand whether any parts of the transcriptome were underrepresented or absent. Such absent or underrepresented transcripts/exons could potentially result in auto-reactive T cells that could not be negatively selected for a potential epitope. As shown in Figure 4.5C, underrepresented tissues at the transcript level included for instance adrenal and ovary. A specific feature of the transcriptome, that could potentially cause epitopes to be absent, are exons, that are not spliced into transcripts. To assess the usage of individual exons, the percentage spliced-in (PSI) for all protein-coding exons was calculated and the exons were split into categories depending on their length (Figure 4.6). Examining the PSI for exons longer than 50 bp revealed a set of peripheral but non-neuronal exons, that were infrequently included in TEC. Around 800 exons were shorter than 30 bp and classified as microexons. Microexons have recently been identified in neuronal tissues and can be as short as 3 bp [58]. Interestingly, a set of neuronal-specific microexons was not spliced-in in any type of mTEC. In contrast, for longer exons ( $> 50$  bp) such a large set of brain-specific exons could not be identified. In summary, these observations indicate that at the exon level, specific subsets of peripheral exons are absent from the transcriptome of mTEC.



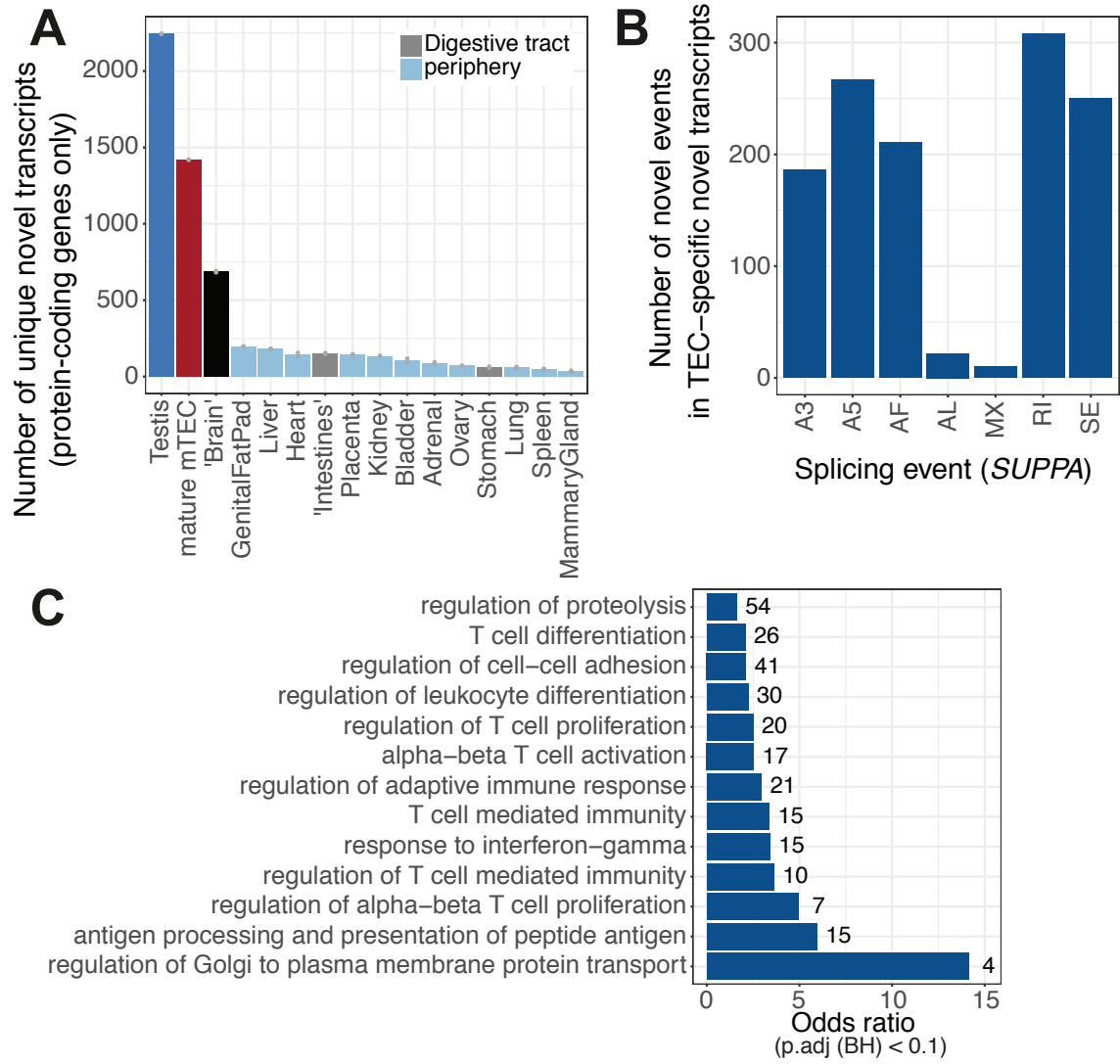
**Figure 4.6:** Heatmap of the percentage spliced-in (PSI) values for exons in protein-coding genes. The 10 k most highly variable exons with normal length ( $> 50$  bp) and all microexons ( $\leq 30$  bp) are shown. PSI values were calculated by SUPPA using the mT&T assembly as annotation.



## 4.4 Characterisation of novel transcripts in TEC

In addition to being a useful tool for comparison of TEC and the periphery, the mT&T assembly also offers the possibility to assess novel transcripts. To assess the expression of tissue-specific novel transcripts, one representative from each tissue group (Appendix 2.1.9.1) was used to calculate the *tau* value (Equation 2.2) for each novel transcript. To find novel transcripts that were unique to each tissue, the novel transcripts were filtered for  $\tau > 0.99$  and at least 2-fold higher expression compared to the tissue with the next highest expression level. In contrast to annotated splice isoforms (Figure 4.3), this assessment revealed that the testis expressed the highest number of such novel transcripts (Figure 4.7A). Mature mTEC expressed 1,572 novel TEC-specific transcripts from 1,159 protein-coding genes. Amongst these genes, 18 % were *Aire*-regulated genes and 30 % were tissue-restricted genes ( $\tau > 0.5$ ). The genes containing novel transcripts were enriched in gene categories including 'T cell differentiation' or 'antigen processing and presentation of peptide antigen' (Figure 4.7C) indicating that these novel transcripts could be functionally relevant in mTEC. In addition, the gene ontology category involving protein transport to the membrane could be interesting in the context of mTEC as it is not well understood, how self-peptides are trafficked to the cell surface of mTEC.

The generation of new splice junctions or exon combinations does not necessarily lead to new epitopes, therefore it is important to understand what types of splice events the novel transcripts include. The assessment of such event types can be performed using scripts, which were developed to generate annotations for well-established splicing tools. Using the annotation script from SUPPA [167], the most frequent event appearing across TEC-specific novel transcripts were retained intron (RI) events (approximately 300) followed by alternative 5' splice sites and skipped exons (Figure 4.7B). In contrast, only ten mutually exclusive exon events occurred in novel transcripts. Following this approach, a novel transcript could occur in multiple categories because it is an 'event-centric' approach.

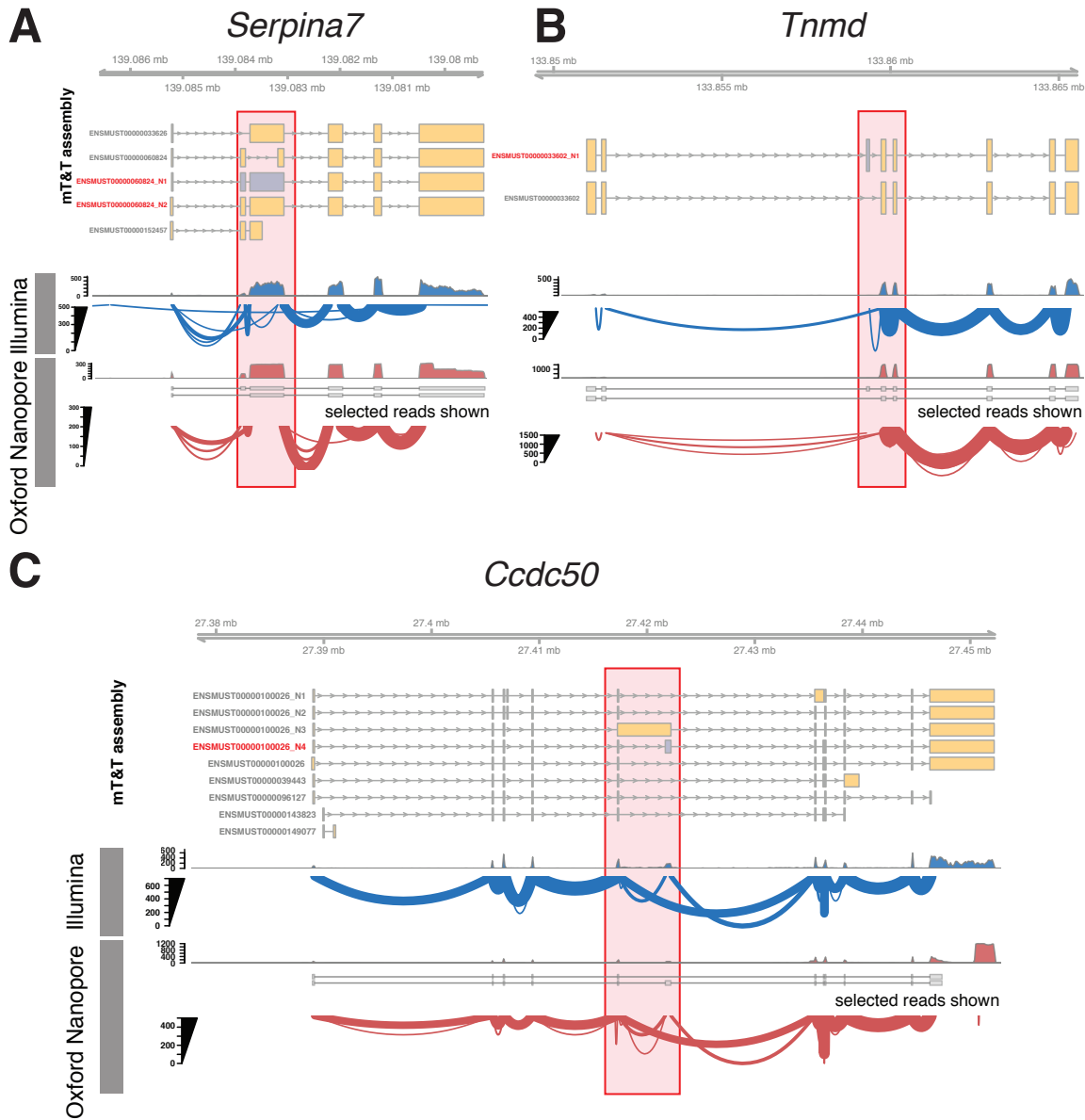


**Figure 4.7:** Assessment of novel transcripts in the mT&T assembly. **A:** The number of novel transcripts 'uniquely' detected ( $\tau > 0.99$ ; 2-fold higher expression than the next tissue) in each tissue or tissue group.  $\tau$  values for each novel transcript were calculated based on tissue groups (Section 2.1.9.1). **B:** Analysis of splicing event types within TEC-specific novel transcripts. Each event was analysed separately, hence one novel transcript can occur 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. **C:** Odds ratios of over-enrichment testing of genesets from biological process gene ontologies. Fisher's Exact tests were performed using protein-coding genes harbouring novel transcripts in mature mTEC (BH adjusted p-value < 0.1). The numbers next to each bar indicate the number of genes with novel transcripts in each category.

#### 4.4.1 Description of novel transcript examples in TEC

To assess the transcript structures of some TEC-specific novel transcripts from the mT&T assembly, examples of different types of novel structures were investigated. A change in the first exon is less likely to introduce novel epitops as these structures often represent extension of the 5' untranslated region (UTR). Therefore the focus was on novel transcript examples with novel internal exons or exon combinations in protein-coding genes (Figure 4.8). As a further computational validation of novel transcripts, the ONT long-read sequencing data from mature mTEC were used to confirm a given structure.

As a first example, the novel transcript in *Serpina7* generated a new combination of known exons by inclusion of the second and third exon together. SERPINA7 is a member of the protein family of Serpins or Serine proteases. The Serpin family consists of 60 functional genes in the mouse and the resulting proteins have diverse functions including proteinase inhibition, involvement in apoptosis and hormone transport [168]. SERPINA7 belongs to the Serpin family members, which do not show protease activity, but instead it is known to be involved in the transport of thyroid hormones. Thus, it is also called thyroxine-binding protein. Another interesting novel transcript, that was supported by full-length ONT reads, was identified in the gene encoding for Tenomodulin (*Tnmd*), which has been identified as a conserved marker and modulator of tendons and ligaments (reviewed in Shukunami *et al.* or Dex *et al.* [169, 170]). Suggested pathways regulated by *Tnmd* are for instance remodelling processes involving Matrix Metalloproteinases (MMPs) or vascularisation (via VEGF-A). A further example of a novel internal exon was in a gene called 'Coiled-Coil Domain Containing 50' (*Ccdc50*) encoding for a protein called YMER. Following studies of a transgenic mouse model of YMER, Tsukiyama *et al.* suggested that YMER positively regulates Fas signalling via NF- $\kappa$ B towards apoptosis in liver cells, but negatively regulates TNF receptor signalling via NF- $\kappa$ B resulting in impaired immune responses of the transgenic mice [171]. Further examples of novel transcripts were for instance a novel second exon in Apolipoprotein L11a (*Apol11a*) and novel first exons in the cytokine *Il27* and the zinc finger protein *Zfp54* (Appendix Figure C.4).



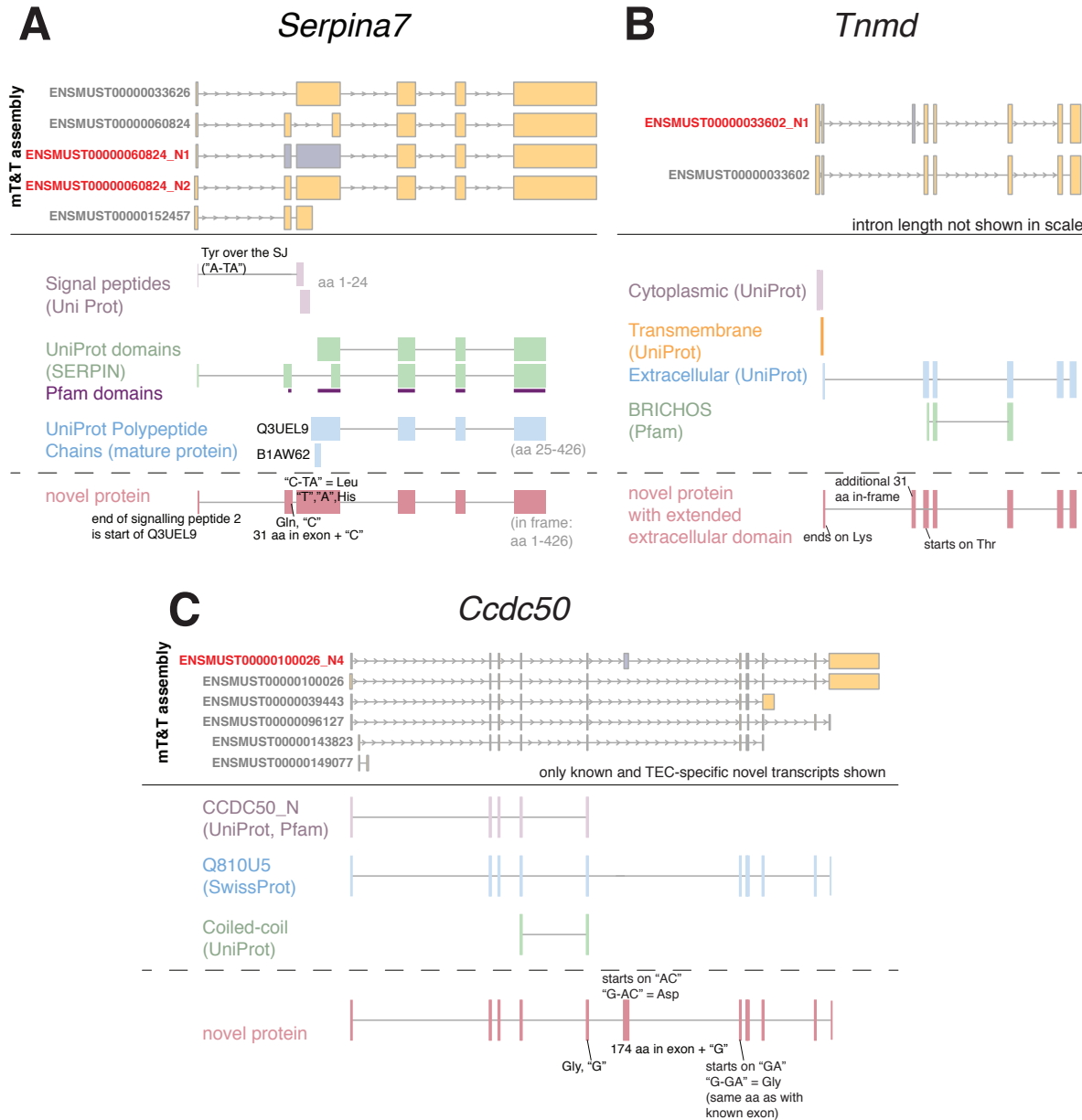
**Figure 4.8:** Examples of novel TEC-specific transcripts. In **A**, **B**, **C** the gene loci of *Serpina7*, *Tnmd* and *Ccdc50* with identified novel transcripts (highlighted in red) are shown. Sashimi plots were generated from RNA-sequencing data from mature mTEC at 200 M reads (blue) or from merged ONT datasets for mature mTEC (12 M reads total; red). Only reads for relevant isoforms are shown for the ONT data. Novel transcripts specific to TEC are highlighted in red.

#### 4.4.2 Prediction of the impact on protein structure for TEC-specific novel transcripts

In order to understand the possible effects of a novel transcript structure on the resulting protein, the potential protein structures produced by the three novel examples highlighted in Figure 4.8 were assessed. The known protein sequences were inspected in UCSC Genome Browser [172] and potential protein structures were then constructed using the known protein start sites (Figure 4.9). The novel transcript in *Serpina7* creates a novel splice junction from exon 2 to exon 3 potentially resulting in a novel Leucin (splice junction sequence: 'C'-'TA'; Figure 4.9A). Translation would then continue in-frame for the remaining part of the mature protein Q3UEL9. Thus, the hypothesis would be that the novel transcript could result in a novel protein including the signalling domain as well as the SERPIN domain.

The protein structure of TNMD consists of a short cytoplasmic domain, a transmembrane domain and a long extracellular domain [173, 170]. The BRICHOS domain located in the extracellular part of TNMD was described as a potential chaperone-like domain [174]. The potential novel protein structure could produce an extended extracellular domain (31 amino acids in-frame), which I predicted to not effect the structure of the BRICHOS domain (Figure 4.9B). However, the novel exon and its splice junctions did not show a very high conservation using the 'Placental Conservation' and '60-Vertebrate Conservation' in UCSC (Appendix Figure C.5).

The novel exon in *Ccdc50* could result in the addition of 174 amino acids downstream of the coiled-coil region. The reading frame, the nucleotide sequence and amino acid sequence across the second novel junction would remain unchanged (splice junction sequence: 'G'-'GA'). Furthermore, the region of the novel internal exon and the two novel splice sites showed high conservation supporting the suggested novel transcript structure (Appendix Figure C.6).



**Figure 4.9:** Peptide prediction for TEC-specific novel transcripts. The genomic loci with the described and predicted peptide structures for *Serpina7* in **A**, *Tnmd* in **B** and *Ccdc50* in **C** (protein information from UCSC Genome Browser, mm10). The bottom row in each panel indicates a novel peptide that could be produced by the novel transcript in TEC. The genomic location including information about conservation for the examples shown in B and C are given in Appendix Figures C.5 and C.6. aa=amino acids, SJ=splice junction

## 4.5 Discussion

In this chapter, the landscape of splicing in mTEC was explored and compared to peripheral mouse tissues. This work was motivated by previous observations of increased splicing complexity in mature mTEC [101]. Furthermore, a hypothesis was that the high splicing diversity in mTEC could give rise to a more diverse set of epitopes, which would be presented to developing T cells.

An issue faced by splicing research in mTEC is the special gene expression program in mature mTEC termed 'promiscuous gene expression' (PGE) (see Section 1.2.3). This process involves expression of TRA under the control of AIRE, FEZF2 and possibly other unknown factors. While a high number of protein-coding genes are expressed in these cells, a lot of these TRA are lowly expressed at the population level. For splicing analysis, the challenge is to assess the transcriptome and its complexity irrespective of the differences in the number of genes expressed. In a recent publication, the authors pointed out that the increase in splicing entropy is likely a consequence of three processes in mTEC: more alternatively spliced genes are expressed, the number of splice variants per gene is higher and the isoform expression profile is flattened [103]. In this work, I confirmed the high number of splice junctions (Section 4.2.1) and tissue-restricted transcripts (Section 4.3). For the first time, I assessed the transcriptome complexity using the fraction of isoforms per gene and showed a slightly higher fraction of isoforms per gene in mature mTEC compared to peripheral tissues (Section 4.2.2). The higher splicing entropy observed in previous publications [175, 50] did not necessarily mean higher complexity with respect to the number of isoforms detected per gene. Although the overall isoform fraction was higher in mature mTEC, the analysis of fraction of isoforms also revealed that not all isoforms were produced by mature mTEC for certain genesets such as *Aire*-regulated genes.

With respect to known splice variants, I also confirmed additional examples of genes, where both transcripts, that are normally found in different tissues, were detected in mature mTEC (Section 4.2.4) [103]. It has so far not been shown whether, in such

cases, both isoforms are expressed in the same single mature mTEC or whether they are expressed in different subsets of single cells. At the population level, an average across all single cells would be obtained and therefore the above mentioned scenarios cannot be distinguished. Here, I detected both isoforms of *Calca* in individual mature mTEC (Figure 4.4D). However, such a co-expression of transcripts could not be confirmed for other examples due to limited coverage and small number of the single cells surveyed.

Furthermore, I found that the representation of tissue-restricted transcripts is significantly increased in mature mTEC compared to cerebellum or testis. Here, I detected 40 to 60 % of tissue-restricted transcripts from peripheral tissues in TEC (Figure 4.5B), which is consistent with previous findings of TRA junction representation ranging from 35-60 % [103]. In addition, Danan-Gotthold *et al.* have found the lowest representation of TRA junctions for the testis, which is also what I observed for the tissue-restricted transcripts in mature mTEC (Figure 4.5C).

So far, the focus of splicing analysis in TEC has been on how much of the transcriptome is included and how complex it is. However, the absence of certain features of the splicing landscape could also give important indications about the biology of TEC. Therefore, I assessed microexons as a distinct feature of the splicing landscape. These short (3-30 bp) exons have been identified in the brain and their inclusion leads to additional amino acids being incorporated in proteins [58]. Here, I identified a set of neuronal-specific microexons, which were absent in mTEC (Figure 4.6). Such a set of neuronal-specific exons was not as obvious for longer exons, but there was also a set of non-neuronal peripheral exons, which was infrequently included in mTEC. The observed underrepresentation of testis- and brain-specific transcripts or exons could be due to these tissues (i) being immunoprivileged sites [106, 105] or (ii) having very specialised and complex splicing profiles (e.g. microexons in neuronal tissues) that are not recapitulated in TEC.

Given the known diversity and the high level of annotated splicing in mature mTEC, the number of novel transcripts specific to mature mTEC was unexpectedly low compared



to the testis. This could be an indication that splicing in mature mTEC is more strictly regulated than in the testis. The testis is also known to express *Aire* [163] and have a special transcriptional program possibly similar to PGE [176, 163, 44]. Alternatively, the number of novel isoforms found in the testis may reflect the fact that the testis is a complex tissue with a variety of cell types, that may not be well covered by existing annotations [177].

Some examples of TEC-specific novel transcripts including transcripts in *Serpina7* and *Il27* were validated with full-length reads from the ONT platform (Figure 4.7, Appendix Figure C.4). The novel variant in *Il27* might be involved in T cell maturation. *Il27* is part of the *Il12* family, which has controversial roles as pro- and anti-inflammatory cytokines and was therefore described as a family of 'immunoregulatory cytokines' (reviewed by Vignali & Kuchroo [178]). In T cell development, however, *Il27* expressed by CD11c<sup>+</sup> thymic dendritic cells was required for the final stages of CD4<sup>+</sup> T cell maturation [179]. In that study, the expression of *Il27* by mTEC was not studied. In our gene expression data, the level of *Il27* was high in populations of mature and *Aire*-positive mTEC and robust in single mTEC (Data not shown).

## Conclusion

In summary, this chapter resulted in a detailed description of the splicing landscape in TEC compared to peripheral mouse tissues. The higher splicing complexity that has previously been reported for mTEC was shown to involve the expression of a higher number of isoforms per gene in mature mTEC. For the first time, novel transcripts in mature mTEC were assessed and examples investigated. The validation of these structures with full-length reads from ONT data confirmed their existence in TEC. However, further lab-based experiments would be required to confirm that they are translated and hence give rise to novel protein structures. An important finding was also the significantly higher representation of tissue-restricted transcripts in mature mTEC compared to brain or testis indicating that expression of tissue-restricted splice variants is a distinct feature

of these cells that potentially increases epitope diversity. Despite the high complexity of splicing in mature mTEC, the absence of a set of neuronal-specific microexons in mTEC and the lower fraction of isoforms in *Aire*-regulated TRA showed that TEC do not express all tissue-restricted splice isoforms and that their splicing landscape is not complete. This evidence would support the hypothesis that splicing is not a promiscuous process in mature mTEC. Further evidence from this chapter includes that mature mTEC did not have an extensive number of novel transcripts and did not produce promiscuous splicing structures. Together, the observations in this chapter suggested that splicing is not a promiscuous process in mature mTEC, but likely regulated to produce a specific repertoire of tissue-restricted transcripts for T cell selection.

# 5

## A computational investigation of possible mechanisms of alternative splicing in thymic epithelial cells

### Contents

---

|            |                                                                                                                |            |
|------------|----------------------------------------------------------------------------------------------------------------|------------|
| <b>5.1</b> | <b>Background</b>                                                                                              | <b>118</b> |
| <b>5.2</b> | <b>Assessment of the influence of AIRE and maturity on splicing in TEC</b>                                     | <b>121</b> |
| 5.2.1      | A limited role for AIRE in splicing in mature mTEC                                                             | 121        |
| 5.2.2      | Decreased frequency of retained intron events upon TEC maturation                                              | 122        |
| 5.2.3      | Validation of retained introns and discovery of novel retained introns using long-read sequencing              | 124        |
| 5.2.4      | Regulation of transcript expression by AIRE                                                                    | 128        |
| <b>5.3</b> | <b>Definition of tissue-restricted splicing factors and their expression across mouse tissues and TEC</b>      | <b>131</b> |
| 5.3.1      | Distinct splicing factor expression in TEC and brain compared to other peripheral tissues                      | 131        |
| 5.3.2      | Expression of a subset of peripheral tissue-restricted splicing factors in TEC                                 | 134        |
| <b>5.4</b> | <b>Prioritisation of a candidate for functional follow-up</b>                                                  | <b>139</b> |
| 5.4.1      | Differentially expressed splicing factors between TEC populations                                              | 139        |
| 5.4.2      | Identification of tissue-restricted splicing factors with frequent and robust expression in single mature mTEC | 141        |
| 5.4.3      | Assessment of splicing in single-cell pools stratified by splicing factor expression                           | 146        |
| 5.4.4      | <i>Rbfox1</i> as a candidate splicing factor for <i>in vivo</i> follow-up                                      | 149        |
| <b>5.5</b> | <b>Discussion</b>                                                                                              | <b>151</b> |

---

## 5.1 Background

Splicing is known to be of importance to the immune system. For example, in the context of T cell activation a positive feedback loop was described to involve the splicing factor CELF2 and the JNK signalling pathway [180, 181, 182]. Furthermore, alternative splicing is of potential relevance to self-tolerance, as peptides of exons and splice junctions of different tissue-restricted antigen (TRA) isoforms might need to be presented to developing T cells to achieve normal self-tolerance. Two publications described the importance of effects of thymic alternative splicing for two autoimmune diseases (type I diabetes; multiple sclerosis; Section 1.4.2) [108, 116]. The report by Klein *et al.*, which involved use of a mouse model of multiple sclerosis, described a difference in splicing of the gene *Plp* between the thymus and the brain, with an exon being absent in the shorter thymic transcript DM20 [108]. In case of the induction of Experimental Autoimmune Encephalomyelitis (EAE) with a peptide specific to the missing exon in the thymic transcript (PLP<sub>139-151</sub>), differences in susceptibility to EAE were observed between SJL/J mice and C57BL/6 mice. The mapping of each strain's MHCII-restricted epitopes of PLP revealed that SJL/J mice had such an epitope located in the exon that was not expressed in the thymus, and thus potentially auto-reactive T cells could be generated. As a consequence, SJL/J mice were highly susceptible to induction of EAE using PLP<sub>139-151</sub>, as discussed in Section 1.4.2.

In work detailed in the previous chapter, I showed that the highest fraction of known splice variants was found in mature mTEC compared to other peripheral mouse tissues (Section 4.2.3). Furthermore, a significantly higher number of tissue-restricted transcripts was found in mature mTEC compared to other tissues known for their high splicing complexity (4.3). The mechanisms by which the variety of splice variants is produced in TEC and the factors that regulate splicing in these cells remain unknown. One possible mechanism involves AIRE, which was shown to regulate a set of almost 4,000 genes [27]. The large majority (86 %) of these genes showed an expression pattern with restriction to certain peripheral tissues, which are known sites of autoimmunity in patients carrying

mutations in AIRE [27]. This disease termed APS-1 manifests with a wide range of autoimmune pathologies (Section 1.2.3). In a report by Keane *et al.*, the authors described a role for AIRE in splicing regulation due to a lower average number of exons per gene in *Aire*-knockout (*Aire*-KO) mTEC based on microarray data. In another recent report, the differences in exon usage between wildtype and *Aire*-KO mTEC were analysed using RNA-sequencing data [41]. This study revealed that for certain cases exon usage changed while the gene expression level was unchanged in the comparison between wildtype and *Aire*-KO mTEC. Moreover, another study has classified a large fraction of AIRE binding partners as pre-mRNA processing machinery [30]. In contrast to these observations, Danan-Gotthold *et al.* found no significant difference in the percentage of TRA junctions detected in *Aire*-KO and wildtype mTEC for six of the eight tissues that they surveyed (the exceptions being the kidney and testis) [103].

Aside from AIRE, mechanisms that regulate splicing in mTEC could involve targeted expression of a set of splicing factors. For tissues such as the brain, a high transcriptome complexity [64, 65] and the involvement of alternative splicing in a variety of essential processes such as neuronal differentiation, morphogenesis and the establishment of a neuron network have been described (reviewed by Raj & Blencowe, [74]). In neuronal cells, a variety of tissue-restricted splicing factors (TRSF) such as Nova, Rbfox or Ptpb, which frequently have redundant family members, function together to regulate these essential processes in a complex interplay. A hypothesis is that a high level of alternative splicing in neuronal tissue supports the complex function of the brain. While a large set of ubiquitously expressed splicing factors exists, the focus of the analysis here was on TRSF expression in mature mTEC as these TRSF are likely involved in creating the observed transcriptome diversity in mature mTEC. St. Pierre *et al.* observed the presence of a range of *Aire*-independent splicing factors in mTEC using deeply-sequenced RNA-sequencing data [50]. The splicing factors were obtained from the database of RNA-binding proteins [107] and the expression subsequently assessed in wildtype and *Aire*-KO mTEC. Seven *Aire*-independent splicing factors with at least 1 RPKM and upregulated in mature mTEC (relative to cTEC and immature mTEC) were shown. The

factors *Dppa5a*, *Pabpn1l* and *Snrpn* showed the largest difference in expression between cTEC and immature mTEC compared to mature mTEC. The highest expression in mature mTEC was observed for *Msi1* and *Synj2*.

The hypothesis was that the expression of these TRSF would be specifically regulated in order to reproduce peripheral, tissue-restricted splice variants in TEC. So far, the set of *Aire*-independent TRSF expressed in mature mTEC has not been well established. Here, I focussed on *Aire*-independent TRSF as (i) the role of AIRE in splicing was controversial and (ii) promiscuously-expressed TRSF are likely to be infrequently expressed in single mTEC, and hence only able to affect the splicing of the small number of self-antigens with which they are co-expressed. *Aire*-dependent TRSF only become functionally mature in the thymic medulla if the purpose of their expression is for presentation as TRA on the surface of mTEC.

Functional investigations of splicing in TEC were lacking. Publicly-available RNA-sequencing datasets from peripheral tissues together with deeply sequenced datasets from mTEC can aid the identification of TRSF regulating splicing in TEC. These could either be peripheral splicing factors, that are reused in TEC, or factors with so far unknown activity in splicing regulation. Single-cell RNA-sequencing datasets can further be used to distinguish promiscuously-expressed splicing factors from factors with robust expression across single cells.

An aspect of AIRE's role in splicing that had not been investigated was the regulation of transcript usage by AIRE. In such cases the gene expression of a gene might not be affected by AIRE, but AIRE could modulate the choice of transcript from these genes. Previous analyses used single-end sequencing with limited depth and were therefore lacking power to detect such differential transcript usage. Using new paired-end, deeply sequenced RNA-sequencing data from wildtype and *Aire*-KO mice the role of AIRE in the regulation of transcript usage or splicing in general could be re-assessed.

## Aim

Although a high transcriptome complexity in mTEC and a potential role for alternative splicing in self-tolerance has been reported, functional studies to connect individual splicing factors to the TEC transcriptome were lacking. In this chapter, I aimed to computationally assess possible mechanism and regulators of splicing in TEC. Therefore I examined (i) the role of AIRE in the regulation of splicing, (ii) the expression of TRSF in TEC and (iii) a possible role of the TRSF in single mature mTEC. Finally, I aimed to prioritise a TRSF for an *in vivo* follow-up study (iv).

## 5.2 Assessment of the influence of AIRE and maturity on splicing in TEC

Splicing in medullary TEC could be influenced by two essential processes in this cell type, namely by maturation or by the transcriptional regulator AIRE. The effect of maturation on splicing has not been described in detail for this cell type, while the role of AIRE has been controversial (Section 1.4.1). Therefore, I aimed to utilise the RNA-sequencing data of wildtype mTEC and the *Aire*-KO model to evaluate splicing regulation in mTEC (Appendix Table B.1). The tool *rMATS* was used for the analysis of differential splicing between two conditions. This widely used tool is based on a likelihood-ratio test to identify differential splicing events [183]. In comparison to other splicing tools such as *MISO*, its advantage is that it can model information from replicate samples (paired testing scheme only possible for >3 replicates). Differential splicing can be analysed for different event types such as skipped exons, mutually exclusive exons, retained introns (RI) or alternative splice sites.

### 5.2.1 A limited role for AIRE in splicing in mature mTEC

The *rMATS* analysis between *Aire*-KO and *Aire*-positive mTEC revealed a total of 562 differential splicing events across all event types in protein-coding genes (Figure 5.1A; Appendix Table D.1). The inclusion difference (delta percentage spliced-in, delta PSI) was filtered for a minimum absolute delta PSI of 0.2 due to the observation of a high number of RI events with low delta PSI values (Appendix Figure D.1A). The number of

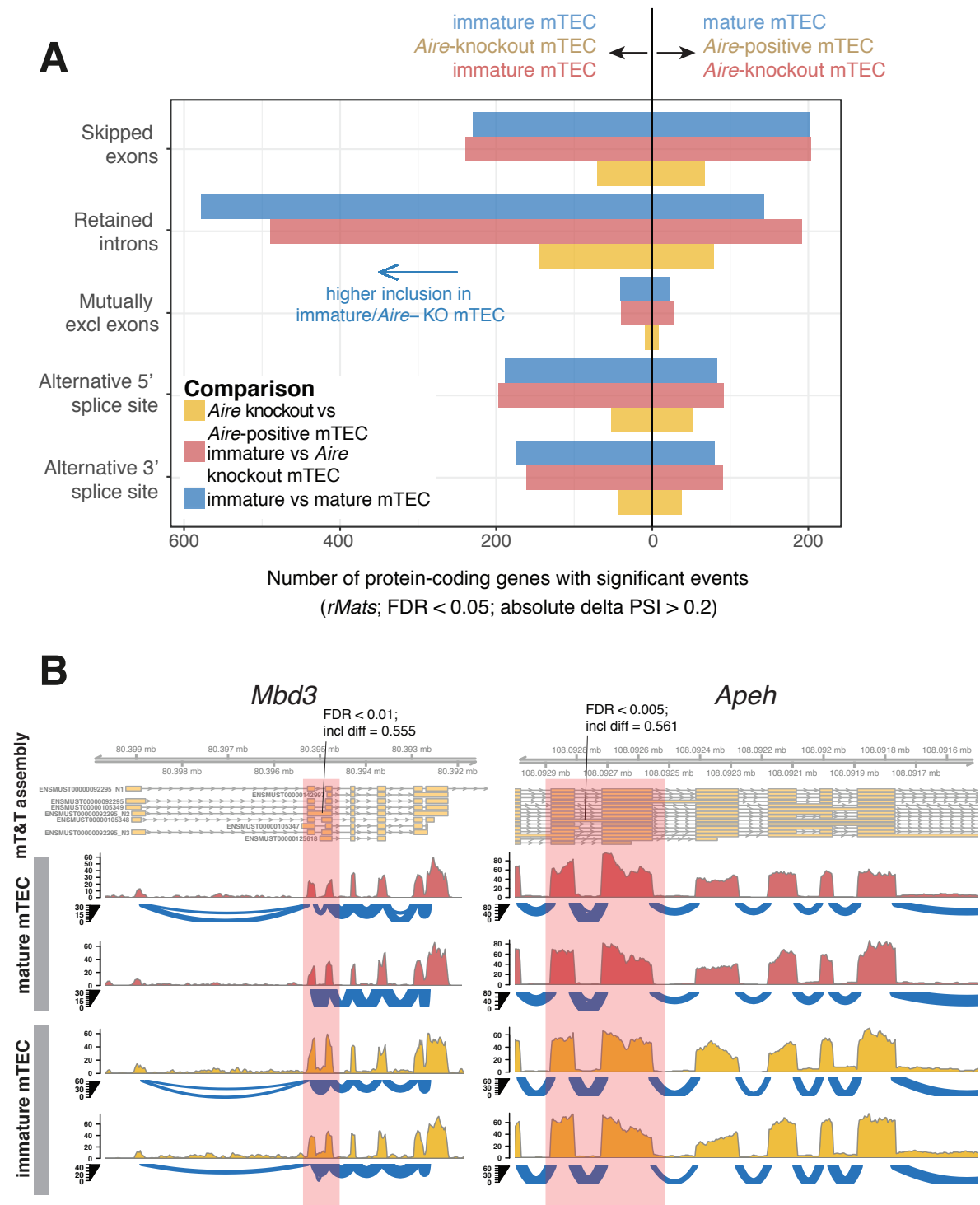
events with higher inclusion in *Aire*-positive mTEC was similar to the number of events in *Aire*-KO mTEC. Furthermore, the total number of differential splicing events detected in this comparison was only one third of the number of events observed between the immature and mature wildtype samples indicating that expression of AIRE has a lower impact on splicing in mTEC than does maturation status. In total, more than 1.7 k differentially spliced events occurred in the comparison between mature and immature mTEC. Additionally, for all events the number of significant inclusion events was higher in immature mTEC than in mature mTEC. Most noticeably, the highest number of significant inclusion events was obtained for RI in immature mTEC.

### 5.2.2 Decreased frequency of retained intron events upon TEC maturation

In order to demonstrate that the effect of maturity was independent of the effect of *Aire* - which is more highly expressed by mature cells - I compared immature mTEC with mature *Aire*-KO mTEC. These two samples should allow to compare the effect of maturity state irrespective of *Aire* expression. The number of events in this comparison (1,729) was similar to the comparison between immature and mature mTEC (1,741), and the trend towards more significant events in immature mTEC held true as well.

A geneset over-representation analysis of genes containing RI events with significantly higher inclusion level in immature mTEC (absolute delta PSI > 0.2, n=727 genes) revealed enrichment for categories such as 'DNA repair', 'protein methylation' and 'histone modification' (Appendix Figure D.1B). Two example genes from the *rMATS* analysis, where retained introns were more frequently spliced-in in immature mTEC compared to mature mTEC, were *Mbd3* and *Apeh* (Figure 5.1B). *Mbd3* is a Methyl-CpG Binding Domain (MBD) Protein, which - in contrast to other MBD family members - binds to methylated, unmethylated and 5-hydroxymethylated DNA equally well. The MBD family members are involved in various processes of epigenetic regulation (reviewed by Du *et al.* [184]) and another family member, MBD1, was shown to interact with AIRE [33]. Acylpeptide hydrolase (*Apeh*) was described to play a role in protein modification by





**Figure 5.1:** Differential splicing analysis using *rMATS*. **A:** Comparison of splicing between (i) immature vs mature mTEC, (ii) *Aire*-KO versus *Aire*-positive mTEC and (iii) immature vs *Aire*-KO mTEC. Significant differences in splicing (5% FDR) were identified using the *rMATS* algorithm (two replicates at 60 M reads) and only filtered events (absolute delta percentage spliced-in (PSI) > 0.2) in protein-coding genes are shown. The mT&T assembly was used as annotation for *rMATS*. **B:** Two examples of retained introns in *Mbd3* and *Apeh* with significantly higher inclusion in immature mTEC compared to mature mTEC are shown. The retained intron region is highlighted in red and delta PSI values as well as p-values from *rMATS* are indicated. The Sashimi plot only includes junctions with 5 or more reads coverage.

removing acetyl groups from the amino-terminal amino acid of proteins [185]. Moreover, Palmieri *et al.* suggested that acylpeptide hydrolase negatively regulates proteasome activity in cell lines [186].

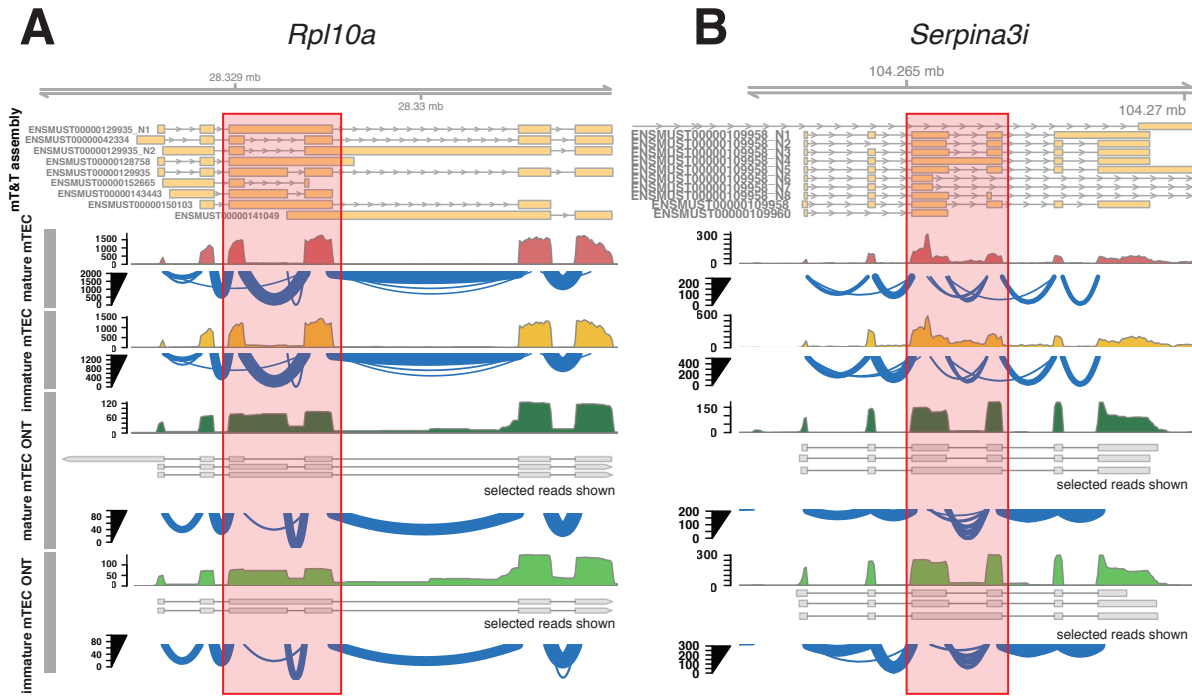
Intron retention has been reported as a measure to regulate gene expression [187]. From the example genes observed to have increased intron retention here, I hypothesised that mTEC might use intron retention to prevent translation of certain transcripts until the cells have reached maturity. This might be important for controlling process such as epigenetic regulation and proteasomal degradation, as they might otherwise interfere with the maturation process itself. An additional example of a differentially included RI between immature and mature mTEC was in a gene known to be expressed in mature mTEC, namely *thymus, brain and testis associated* (*Tbata*; also known as *Stromal protein associated with thymi and lymph node* [*Spatial*] [188]; Appendix Figure D.2). *Spatial* was shown to be expressed during thymus organogenesis and is known to be an *Aire*-regulated gene in mature mTEC [189].

### 5.2.3 Validation of retained introns and discovery of novel retained introns using long-read sequencing

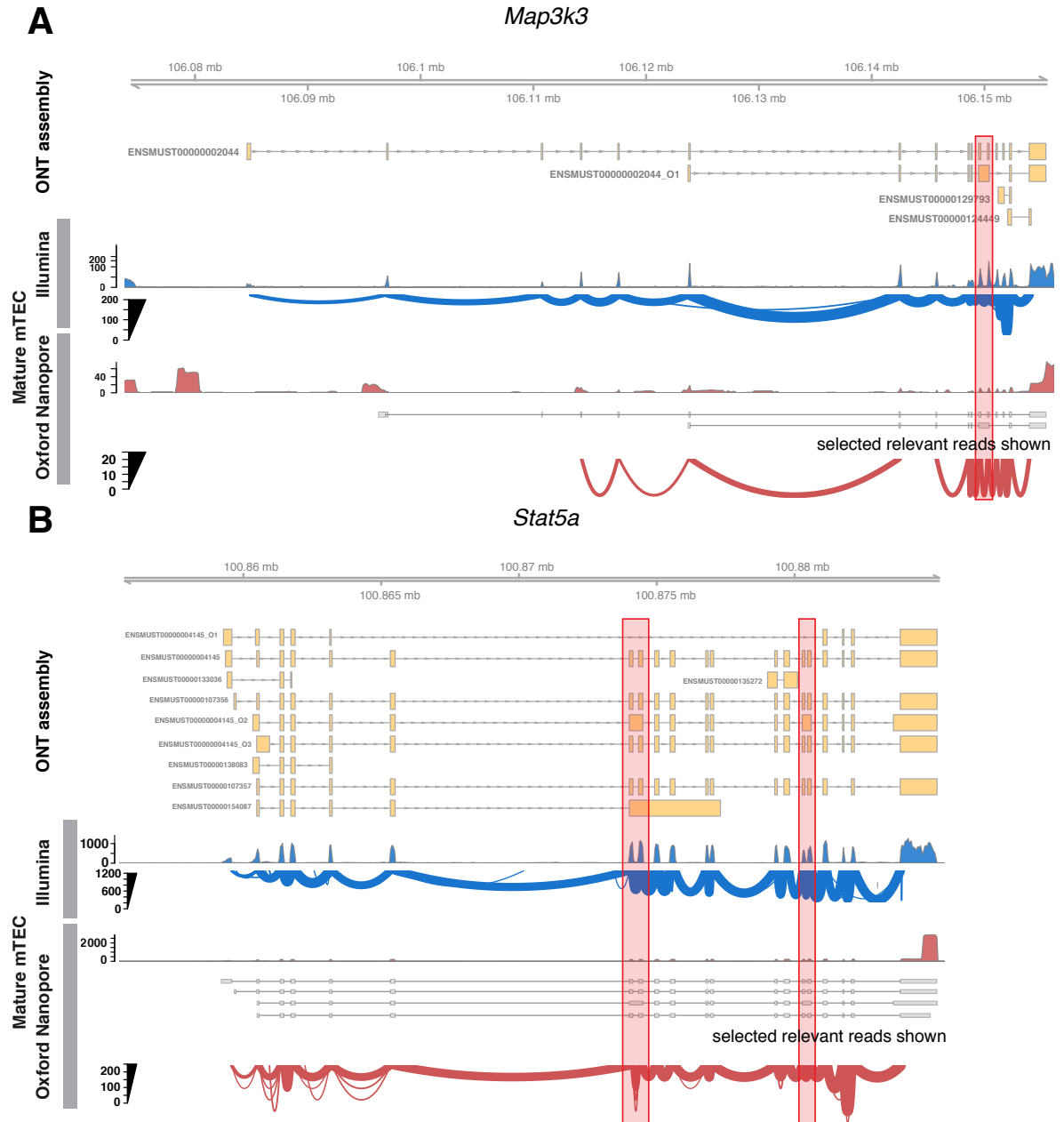
In Section 3.3.3, I observed that novel RI-containing and multi-gene transcripts were generally detected in a lower fraction in ONT data than all other transcripts. The fraction of novel mT&T transcripts containing RI that were detected in Illumina data and also found in ONT data was 45 % in mature mTEC (Appendix Table D.2). The phenomenon was also observed by inspecting individual examples in Integrative Genome Browser (IGV), where reads frequently started or ended in long retained introns, indicating that the library construction from the 3' polyA tail in the ONT cDNA protocol was not as efficiently amplifying RI-containing transcripts as the Illumina random-priming library construction. However, highly expressed (>100 counts) RI-containing novel mT&T transcripts were more frequently validated in ONT data (90 %, Appendix Table D.2).

Although the validation of retained introns with full-length ONT reads was challenging, some examples of known and novel retained introns (from the mT&T assembly) were validated by full-length reads (Figure 5.2). One example of a known transcript with retained introns was amongst the transcripts from the gene *Rpl10a*, where a partial or full retention of an intron was observed (Figure 5.2A). RPL10a is a highly conserved ribosomal protein required for ribosome assembly. It was described as downregulated by the immunosuppressive drug cyclosporin in the thymus and was previously named CsA-19 and NEDD-6 [190]. Riboseq of selected RPL10a-containing ribosomes revealed that certain transcripts might be enriched or depleted from ribosomes depending on the proteins included in the ribosome complex [191]. The second example of a validated RI was found in a novel mT&T transcript in the gene *Serpina3i*. *Serpina3i* is one of nine members of the Serpina3 family of genes within Serpin Clade A [168].

Despite the challenges of selecting novel ONT transcripts with RI events from the assembly described in Section 3.3.2.2, I identified two novel RI transcripts with low coverage ( $< 5$  reads) in the ONT assembly (Figure 5.3). The first novel ONT transcript containing full-length reads across a novel RI was in *Map3k3* (or *Mekk3*), which encodes for 'mitogen-activated protein kinase kinase kinase 3' and is part of the complex MAPK pathway (reviewed by Wagner & Nebreda [192]). The second example gene with a novel RI-containing transcript in the ONT datasets was *Stat5a*, which upon phosphorylation forms complexes and acts as an important transcriptional activator (reviewed by Shuai & Liu, [193]). The two novel retained introns connected two exons coding for the *alpha*-helix coiled-coiled domain and the Src homology 2 (SH2) domain, respectively [194]. Another STAT family member, namely *Stat3*, was found to be essential for mTEC maintenance as shown in a *Keratin5*-cre conditional knockout mouse model [195], demonstrating that the STAT family has an important role in TEC biology.



**Figure 5.2:** Validation of a retained intron in a known (A) and a mT&T novel (B) transcript using long-read sequencing data. In **A** and **B** the gene loci of *Rpl10a* and *Serpina3i* with retained introns (highlighted in red) are shown, respectively. Sashimi plots were generated from Illumina data from mature (MHCII high) and immature (MHCII low) mTEC or from merged ONT datasets for mature (CD80<sup>+</sup>, CD86<sup>+</sup>) and immature (CD80<sup>-</sup>, CD86<sup>-</sup>) mTEC. Selected reads are shown for ONT data. The Sashimi plot only includes junctions with 5 or more reads coverage.



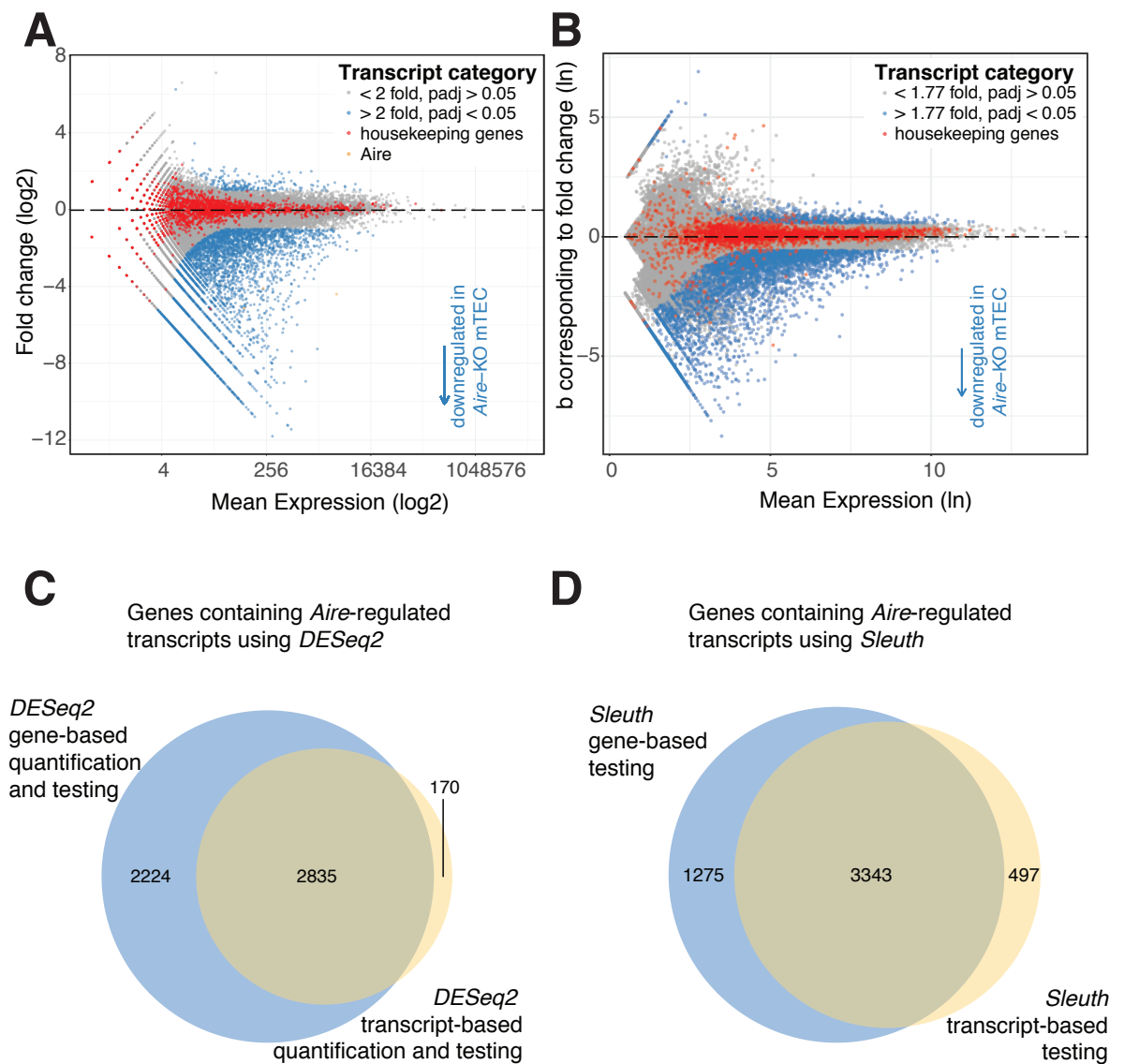
**Figure 5.3:** ONT novel transcripts with retained introns in mature mTEC. In **A** and **B**, two examples of novel ONT transcripts containing retained introns, namely *Map3k3* and *Stat5a* are shown. Sashimi plots were generated from Illumina data from mature mTEC (MHCII high) or from merged ONT datasets for mature mTEC (CD80<sup>+</sup>, CD86<sup>+</sup>). Selected and relevant reads are shown for ONT data. The retained intron regions are highlighted in red. The Sashimi plot only includes junctions with 5 or more reads coverage.

### 5.2.4 Regulation of transcript expression by AIRE

The well-described regulator of promiscuous gene expression, AIRE, and its target genes have been extensively described. However, the transcripts which are regulated by AIRE have not been defined yet. It is possible that a gene might not be differentially regulated by AIRE, while one of its transcripts is regulated by AIRE. To understand AIRE's contribution to transcript abundances, I defined *Aire*-regulated transcripts using two methods and compared them to the set of *Aire*-regulated genes identified by each method.

A classical method to define differentially expressed genes is *DESeq2*. Here, I used the tool to find differentially expressed transcripts. In total, around 3 k transcripts in protein-coding genes were identified as significantly, more than 2-fold downregulated in the absence of AIRE (Figure 5.4A; Benjamini-Hochberg adjusted p-value < 0.05). As a control, all transcripts arising from the mouse 1:1 orthologues of the human house keeping genes identified by de Jonge *et al.* were highlighted [133]. None of the 3 k transcripts in housekeeping genes were differentially expressed between the two conditions (Appendix Table D.3).

The second approach involved the recently published pair of tools, *Kallisto* for transcript quantification and *Sleuth* for differential expression analysis [130, 131] (Figure 5.4B). *Sleuth* does not calculate the fold change but a metric termed 'b', which the authors described as analogous to a fold change (on the natural log scale). In order to compare the two approaches, I used a linear model to estimate the 'b' that would correspond to the log2-fold change (Section 2.1.5.3, Figure D.4). In short, the fold change between the two samples was calculated and plotted against the 'b' metric determined by *Sleuth*. A simple linear model was fitted to determine that a log2(2) corresponded to b=0.569. Using this threshold for the *Sleuth* results, 4,618 transcripts in protein-coding genes were significantly downregulated in *Aire*-KO samples compared to 3,021 transcripts in the *DESeq2* approach (Appendix Table D.3), suggesting that *Sleuth* was more sensitive in finding differentially expressed transcripts. Comparing the transcripts identified by both approaches revealed approximately 2,600 transcripts in common, while *Sleuth* identified



**Figure 5.4:** Definition of *Aire*-regulated transcripts. **A:** Differentially expressed transcripts in *Aire*-positive versus *Aire*-KO samples using *featureCounts* and *DESeq2* (Ensembl annotation v91). **B:** Differentially expressed transcripts in *Aire*-positive versus *Aire*-KO samples using *Kallisto* and *Sleuth* (Ensembl annotation v91). In A and B, transcripts in house-keeping genes (red) were plotted on top of other transcripts. The list of 468 house-keeping genes was obtained by mapping (1:1 orthologues only) the human house keeping genes identified by de Jonge *et al.* to their mouse counterparts [133]. **C:** Venn diagram to compare protein-coding genes identified either as *Aire*-regulated genes or containing *Aire*-regulated transcripts by two independent *DESeq2* analyses. **D:** Same as in C for the *Sleuth* approach. As a threshold for *Aire*-regulated genes or transcripts,  $b > \log(1.77)$  was chosen as it was determined to be a comparable threshold to 2 fold, which was used for the *DESeq2* analysis (Appendix Figure D.4)

an additional 2,000 transcripts as *Aire*-regulated in contrast to 609 found in the *DESeq2* analysis (Appendix Figure D.5A).

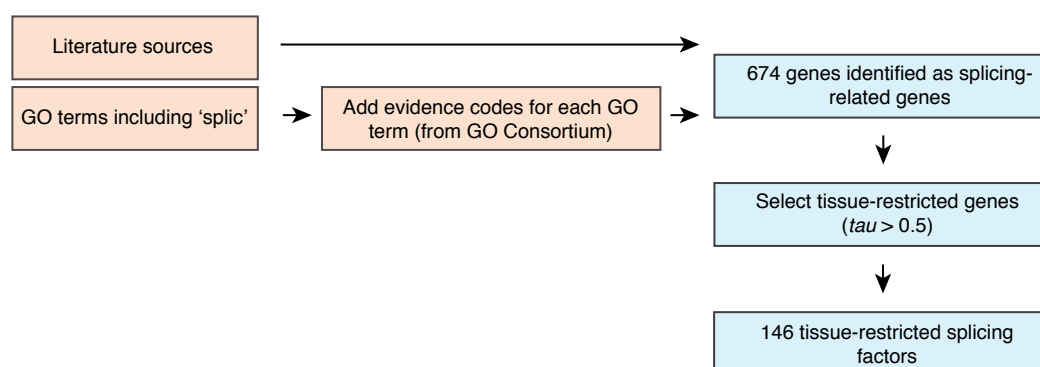
Next, I assessed the intersection between *Aire*-regulated transcripts and *Aire*-regulated genes identified by each approach. Overall, *Sleuth* identified fewer *Aire*-regulated genes than did *DESeq2* (5.1 k for *DESeq2* (Appendix Figure B.5) and 4.6 k for *Sleuth* Appendix Figure D.3). However, the *DESeq2* transcript-based approach only identified an additional 170 protein-coding genes containing *Aire*-regulated transcripts, whereas the *Sleuth* approach identified almost 500 additional protein-coding genes as affected by AIRE at the transcript level (Figure 5.4).

Geneset over-enrichment analysis of the genes containing *Aire*-regulated transcripts with either approach but not defined as *Aire*-regulated at the gene-level in the respective approach, revealed categories involved in muscle function or development as well as cell adhesion in both approaches (Appendix Figure D.6A, B). In addition, the genes containing *Aire*-regulated transcripts, that were only identified in the *Sleuth* transcript-based approach, were enriched for categories such as 'positive regulation of interleukin-5 production', 'skeletal muscle adaptation' and 'peptide secretion' (Appendix Figure D.6B). The *Aire*-regulated transcripts in non-*Aire*-regulated genes with highest fold changes between *Aire*-KO and *Aire*-positive mTEC are shown in Appendix Figure D.6C. These transcripts arise, for instance, from Myosin (*Myo18a*, *Myo7a*) or immune-related genes such as the chemokine receptor *Ccr4*. Furthermore, I could identify *Aire*-regulated transcripts (in non *Aire*-regulated genes) in the trophoblast glycoprotein *Tpbp* or the gene *Gpr179*, which is expressed in the retina and human mutations in *GPR179* lead to night blindness [196]. To assess whether differentially spliced events regulated by *Aire* were affecting the same genes as those showing transcript modulation by *Aire*, I compared the genes containing skipped exon events defined by *rMATS* (Figure 5.1) and the genes containing *Aire*-regulated transcripts defined by *Sleuth* (Figure 5.4). I found a small set of only 8 protein-coding genes in common between the two gene lists (Appendix Figure D.5B).



### 5.3 Definition of tissue-restricted splicing factors and their expression across mouse tissues and TEC

To assess whether alternative splicing regimes from the periphery might be used in TEC to reproduce peripheral splice variants, I focussed on the expression of TRSF in TEC and the periphery. In order to identify a set of TRSF with a potential role in TEC, I defined a list of splicing-related genes from multiple literature sources and gene ontology (GO) terms (Figure 5.5) [53, 197, 79, 77, 78, 50]. Further information about the GO terms was obtained from the GO Consortium, which annotated every GO category with an 'Experimental Evidence Code' ranging from experimental evidence to only electronic annotation. The initial list of 674 genes was further filtered to TRSF with  $\tau < 0.5$  ( $\tau$  defined in Section 2.1.9.3). Finally, a set of 146 TRSF was obtained.



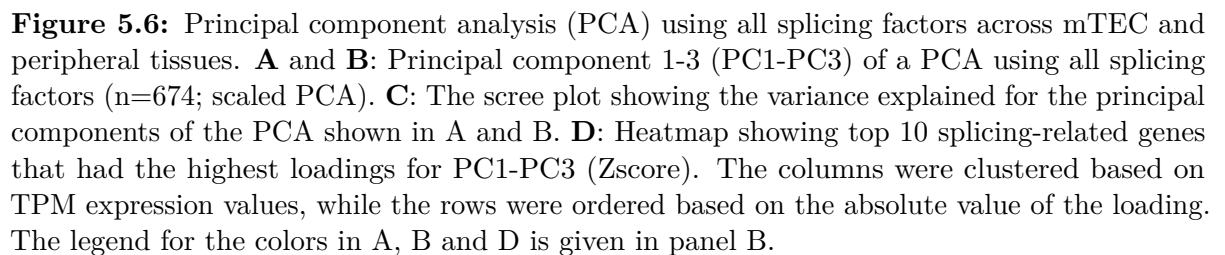
**Figure 5.5:** Identification of a tissue-restricted splicing factor set. Outline of steps to define tissue-restricted splicing factors. The tau values for all protein-coding genes were defined in Appendix 2.1.9.3.

#### 5.3.1 Distinct splicing factor expression in TEC and brain compared to other peripheral tissues

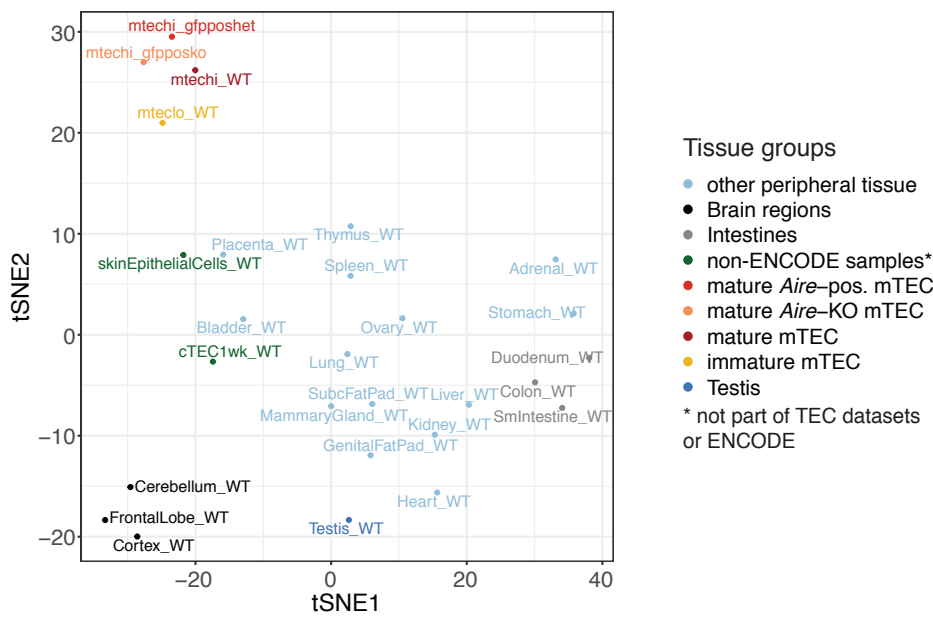
To understand whether mTEC have a globally different expression pattern of splicing factors compared to other peripheral tissues, I used the expression of all splicing-related factors ( $n=674$ ) for a Principal Component Analysis (PCA). PCA is a linear dimensionality reduction method, which can be applied to summarise high-dimensional data into fewer dimensions for instance for visualisation purposes [198, 199]. Inspecting the principal components (PC) revealed that the brain regions, testis and TEC had a very different

expression profile compared to other peripheral tissues including the digestive tract (PC1) (Figure 5.6A). PC2 explained the variation found between the TEC types and the brain regions plus testis. Although on PC2 the brain regions and testis appeared to have a similar profile, they were separated clearly along PC3 (Figure 5.6B). The scree plot revealed that PC1 to PC15 explained the majority of the variance between the samples (Figure 5.6C). To assess which splicing-related genes were involved in the observed separation of the tissues across PC1, PC2 and PC3, the top 5 genes were selected from the loadings for each PC. The scaled expression showed that PC1 was influenced by genes such as the known splicing factors *Srsf1* [200] and *Prpf38a* (Figure 5.6D). The latter was recently identified in a siRNA screen in breast cancer cell lines, where a role in splicing was observed in subsequent knockdown experiments [201]. The separation between the TEC samples versus the testis and brain regions observed across PC2 was driven by for instance *Nova2* and *Celf5* expression, which were described as primarily expressed in neuronal tissues [202, 203]. PC3 showed a distinct profile of splicing-related genes for the testis confirming previously observed testis signatures [79]. One of the genes detected at high levels in the testis (otherwise only detected at low levels in mTEC) was *Rbm44*. The protein contains a RNA-binding motif, but its role in splicing or other forms of RNA metabolism are unknown. *Rbm44* knockout mice did not show a disruption in testis morphology but increased fertility for unknown reasons [204]. Another gene with high expression in the testis was *Rbmy*, a RNA-binding protein specifically expressed during spermatogenesis (reviewed by Elliott [205]).

A summary of the meaningful PCs into a t-distributed stochastic neighbour embedding (t-SNE; [206]) was used to obtain tissues with most similar profile with respect to splicing factor expression. The first 15 PCs were used as input for the t-SNE as they explained the majority of the variation (Figure 5.6C). Similar to the PCA, all mTEC types clustered together, while the brain regions showed a distinct profile from mTEC (Figure 5.7). The digestive tract including the stomach and adrenal were also observed in a separate location to the other peripheral tissues. In general, a high similarity between the epithelial cell types in the dataset, namely skin epithelial cells and TEC, was expected. However, skin



epithelial cells and cTEC did not cluster with mTEC, suggesting that the splicing factor profile in other epithelial cell types and cTEC differed from mTEC. For cTEC, a similar profile to mTEC was observed for PC1 and PC2, but not if the PCs were combined into the t-SNE. Taken together, the PCA and t-SNE revealed five groups of tissues with respect to splicing factor expression, which were neuronal tissues, intestinal tissues, mTEC, testis and 'other peripheral tissues'. However, these results for TEC samples may be confounded by the presence of promiscuously-expressed splicing factors.



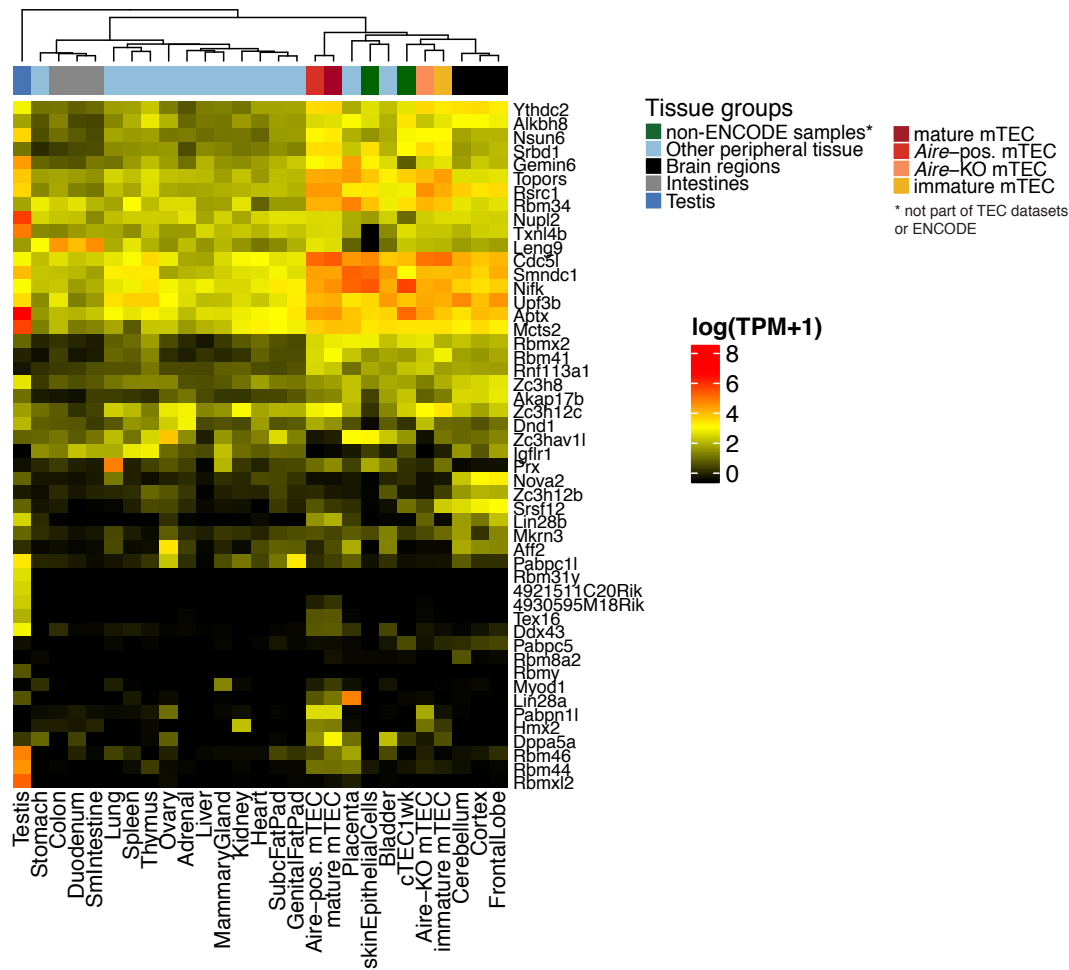
**Figure 5.7:** t-SNE analysis of splicing factor expression profile across mTEC and peripheral tissues. For the t-SNE the first 15 principal components from the PCA based on all splicing factors and a perplexity of 5 were used. 674 genes were included in the PCA (Figure 5.6).

### 5.3.2 Expression of a subset of peripheral tissue-restricted splicing factors in TEC

To identify subsets of TRSF that might contribute to the diverse splicing in mTEC, the 50 TRSF with the most variable expression levels between the TEC and peripheral tissues (out of 146) were plotted (Figure 5.8). As observed in the PCA and t-SNE here as well as other studies [79], the testis had a very distinct profile of TRSF expression (Figure 5.8) with high expression of *Rbm44* and *Rbmy* as described above (Figure 5.6D). The hierarchical clustering also revealed high similarity between the intestinal tissues

plus stomach, which showed a different profile to other peripheral tissues. The observed clusters of tissues from the hierarchical clustering of TRSF were similar to the t-SNE, where the intestinal samples (plus stomach and adrenal) also clustered together (Figure 5.7). The other large group of tissues included the brain regions, all TEC, skin epithelial cells, placenta and bladder. Some splicing-related genes with restriction to very few tissues could be observed. For instance *Periaxin* (*Prx*) was highly detected in the lung, but only lowly in other tissues. Based on the phylogenetic relationship with splicing factor AHNK and *in vitro* co-expression with the splicing factor SRSF2 in nuclear speckles, a role for *Prx* in splicing has been suggested [207]. Similar to *Prx*, *Lin28a* was only found in the placenta, mature mTEC and *Aire*-positive mTEC. LIN28A is known for its negative regulation of micro-RNAs such as the let-7 micro-RNA family, and Piskounova *et al.* found it to be highly expressed in human breast tumors by quantitative RT-PCR [208]. The observation that LIN28A was localised in the cytoplasm in comparison to nuclear localisation of LIN28B suggested that it might not act as a direct splicing regulator.

As indicated above, a large number of genes categorised as TRSF had not been clearly connected to splicing, therefore I next focussed on TRSF with literature evidence for a direct functional role in splicing (Table 5.1). The selected splicing factors were compared across the available mouse tissues and mTEC (Figure 5.9). As expected, known brain-specific splicing factors such as *Nova1*, *Nova2* and *Celf4* showed high expression in all three brain regions, but were not expressed in TEC. *Rbm20* and *Rbm24* were observed in the heart, which was consistent with its described expression in the heart and muscle [62, 209]. Furthermore, *Srpk2* and *Srpk3* were highly expressed in the testis and spleen, respectively [79, 210]. In addition, another testis-specific splicing factor was *Brdt* [211]. The epithelial splicing factor named *Epithelial splicing related protein 1* (*Esrp1*) was highly expressed in the digestive tract tissues as well as in cTEC, skin epithelial cells and all mTEC types. A range of factors showed *Aire*-dependent expression such as *Rbfox3*, *Elavl2* or *Elavl3* (indicated by '\*' between the respective mTEC samples in Figure 5.9). In contrast, other factors such as *Rbm20* and *Rbm24* were differentially expressed between immature mTEC and mature *Aire*-KO mTEC, indicating that they



**Figure 5.8:** Heatmap of 50 most highly variable tissue-restricted splicing factors. Upper-quartile normalised TPM are shown and samples were clustered (Manhattan distance, complete linkage, optimised leaf order).

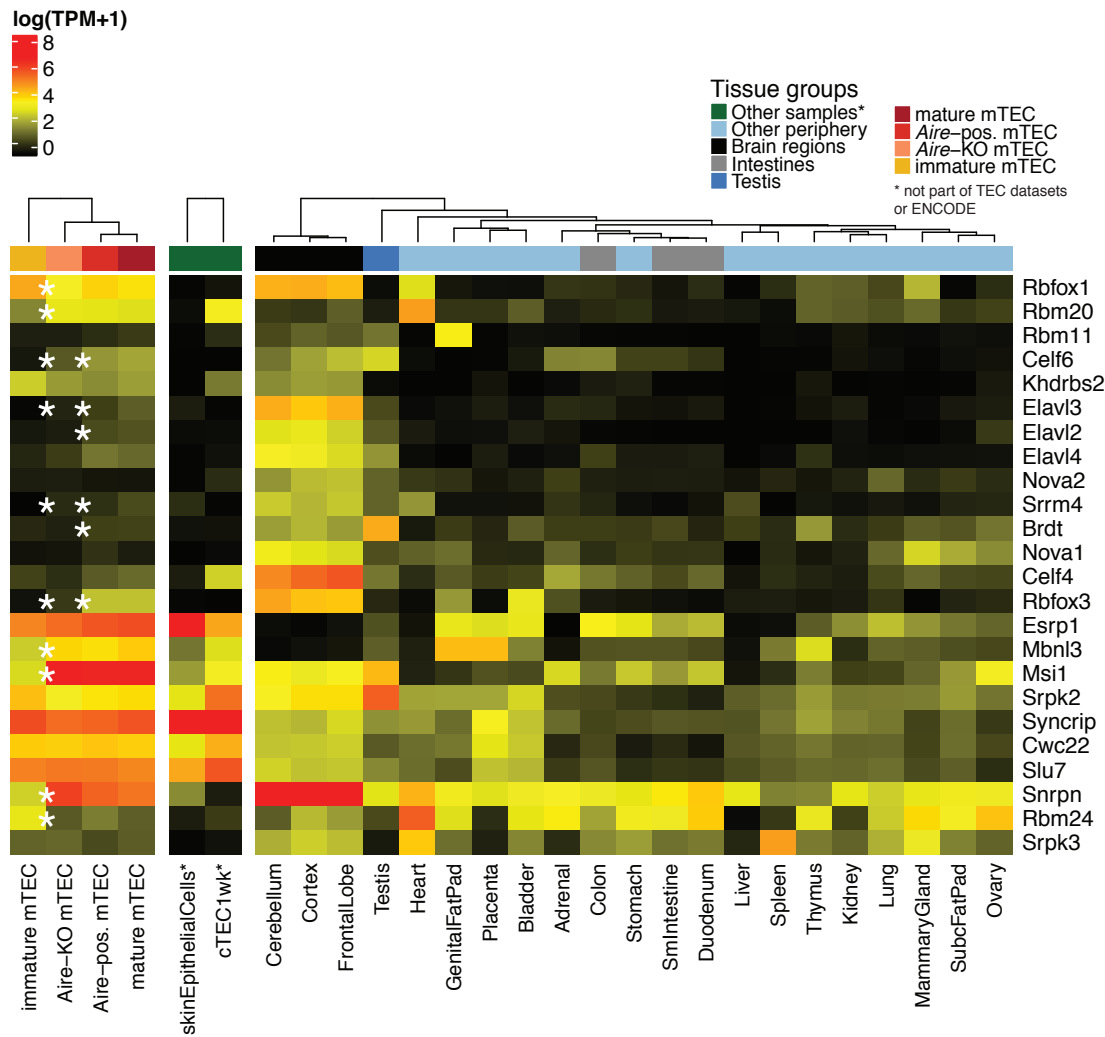
might be maturity-regulated in mTEC. Differential expression analysis between TEC types is detailed in Section 5.4.1.

Interestingly, *Rbfox1* showed very restricted expression in the heart and brain regions, but was also found to be robustly expressed in all mTEC types. However, it was not detected in cTEC and skin epithelial cells, suggesting that the factor might be expressed for a specific purpose, such as splicing of self-antigens, in mTEC. Another factor, that was highly expressed in mTEC, was *Musashi1* (*Msi1*), but it additionally showed expression in cTEC, skin epithelial cells and low levels in digestive tract tissues, indicating that it is likely to have a general role in splicing in epithelial cells.

**Table 5.1:** List of manually curated TRSF. Splicing factors were selected based on  $\tau > 0.5$  and literature supporting a role in alternative splicing. NSC = neuronal stem cells, EMT = epithelial to mesenchymal transition.

| Splicing factor        | Tissue described in                 | Reference  |
|------------------------|-------------------------------------|------------|
| Brdt                   | Testis                              | [211]      |
| Celf4, Celf6           | Neurons                             | [203]      |
| Cwc22                  | involved in spliceosome             | [212]      |
| Elavl2, Elavl3, Elavl4 | Neurons                             | [213]      |
| Esrp1                  | Epithelium                          | [214]      |
| Khdrbs2                | Neurons                             | [215]      |
| Mbnl3                  | Muscle                              | [216]      |
| Msi1                   | NSC, cancer EMT, photoreceptors     | [217, 218] |
| Nova1, Nova2           | Neurons                             | [72]       |
| Rbfox1, Rbfox3         | Neurons                             | [68]       |
| Rbm11                  | Neurons, Testis                     | [219]      |
| Rbm20, Rbm24           | Muscle, Heart                       | [62, 209]  |
| Slu7                   | Liver                               | [220]      |
| Snrpn                  | embryo, neuronal (imprinted region) | [221]      |
| Srpk2                  | Testis                              | [79]       |
| Srpk3                  | Muscle, Heart, Spleen               | [210]      |
| Srrm4                  | Neurons                             | [58, 222]  |
| Syncip                 | Neurons                             | [223]      |

One of the splicing factors with very low, *Aire*-dependent expression in mTEC was *Srrm4*, which is also named nSR100, and was described as an important factor controlling microexon inclusion in neuronal tissues [58, 222]. Given that this factor has been shown to direct inclusion of neuronal microexons [58, 222], its low expression provides a likely explanation for the observed lack of neuronal microexon expression in TEC (Figure 4.6). Taken together, this analysis revealed that mTEC expressed a specific subset of peripheral splicing factors including epithelial splicing factors and other TRSF such as *Rbfox1* and *Msi1*. Thus, splicing in mTEC might be a highly regulated process to ensure expression of a particular set of TRSF in order to produce the wide range of splice variants observed in mTEC.



**Figure 5.9:** Heatmap of selected tissue-restricted splicing factors with a described role in the literature. Upper-quartile normalised TPM (*Salmon*) are shown and samples were clustered within groups (Manhattan distance, complete linkage, optimised leaf order). \* between two TEC types indicates that the factor was > 2-fold and significantly differentially expressed in the respective comparison (details in Figure 5.10, significant corresponded to Benjamini-Hochberg adjusted p-value < 0.05).



## 5.4 Prioritisation of a candidate for functional follow-up

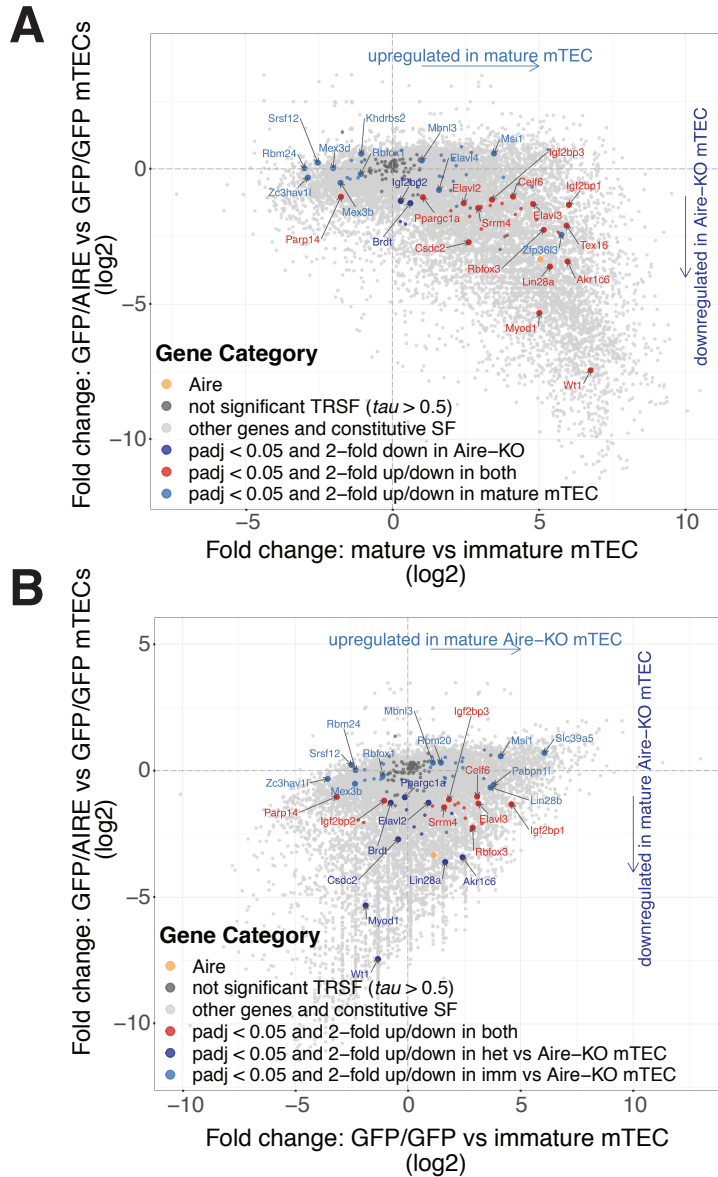
Different types of datasets were investigated in order to prioritise a candidate TRSF for functional follow-up: Population RNA-sequencing was used to assess the dependency of splicing factors on maturity status of mTEC and on AIRE. Single-cell RNA-sequencing data from mature mTEC (defined by high expression of MHCII) was used to examine, how frequently a factor was expressed at the single-cell level.

### 5.4.1 Differentially expressed splicing factors between TEC populations

To assess how splicing factors in mTEC were regulated, differential expression analyses between the mTEC types were performed using *DESeq2*. Genes differentially expressed between *Aire*-KO and *Aire*-positive mTEC were defined as promiscuously-expressed genes. Due to their known infrequent expression at the single-cell level [27, 41], promiscuously-expressed genes are less likely to have a functional role in the overall mTEC population and were therefore not prioritised for functional follow-up. While the differentially-expressed genes between immature and mature mTEC would contain some promiscuously-expressed genes, other differentially expressed genes had a functional role related to T cell selection. These 'maturity genes' are exemplified, for instance, by proteins involved in antigen presentation (MHCII or CD86). The regulation of splicing factor expression could therefore be directed by AIRE or independent mechanisms related to mTEC maturity status.

The comparison between *Aire*-KO and *Aire*-positive mTEC revealed a set of *Aire*-regulated splicing factors (y-axes in Figure 5.10). In total, 31 factors were identified as *Aire*-regulated with 14 regulated by AIRE but not by maturity status. No genes were upregulated upon the loss of *Aire*, and *Aire* itself was strongly downregulated in the *Aire*-KO mTEC. Amongst the *Aire*-regulated and known alternative splicing factors were a range of neuronal splicing factors, for instance *Elavl2*, *Elavl3* and *Rbfox3*. Most strongly regulated by *Aire* was the tumor suppressor gene *Wilms tumor 1 gene* (*Wt1*), which

itself has a range of isoforms but no clear record as a splicing factor. An example of a gene downregulated in *Aire*-KO mTEC but not in mature mTEC was the testis-specific factor *Brdt* [211].



**Figure 5.10:** Differential expression analysis of splicing factors between different TEC types. **A:** Fold changes for differential expression between immature and mature mTEC (x-axis) as well as differential expression between mature *Aire*-KO and *Aire*-positive mTEC (y-axis). **B:** Fold changes for differential expression between immature and mature *Aire*-KO mTEC (x-axis) as well as differential expression between mature *Aire*-KO and *Aire*-positive mTEC (y-axis). In both panels, tissue-restricted and significantly differentially expressed splicing factors are highlighted (>2 fold, Benjamini-Hochberg adjusted p-value < 0.05). Labelled splicing factors were either selected for their known role in alternative splicing (Table 5.1) or were the top 5 differentially expressed factors for the respective comparison.

The comparison between immature and mature mTEC revealed 57 splicing factors as differentially expressed (Figure 5.10A, x-axis). *Rbm24* and *Srsf12* were two well-known examples downregulated in mature mTEC, while *Msi1* and *Zfp36l3* were upregulated in mature mTEC. *Zfp36l3* is a rodent-specific placental and yolk sac gene involved in the regulation of mRNA stability [224], while *Msi1* was involved in photoreceptor splicing [218]. Two known alternative splicing factors, *Khdrbs2* and *Rbfox1*, were 2-fold downregulated in mature mTEC. Both factors were described using *in vivo* studies for their alternative splicing behaviour in neuronal tissues [145, 215].

In order to understand which factors could be up-regulated in mature mTEC irrespective of AIRE, immature mTEC were compared to mature mTEC from homozygous *Aire*-KO animals (Figure 5.10B, x-axis). *Rbm20* and *Pabpn1l* showed upregulation in mature *Aire*-KO mTEC, suggesting that they were regulated by *Aire*-independent mTEC maturity factors.

#### 5.4.2 Identification of tissue-restricted splicing factors with frequent and robust expression in single mature mTEC

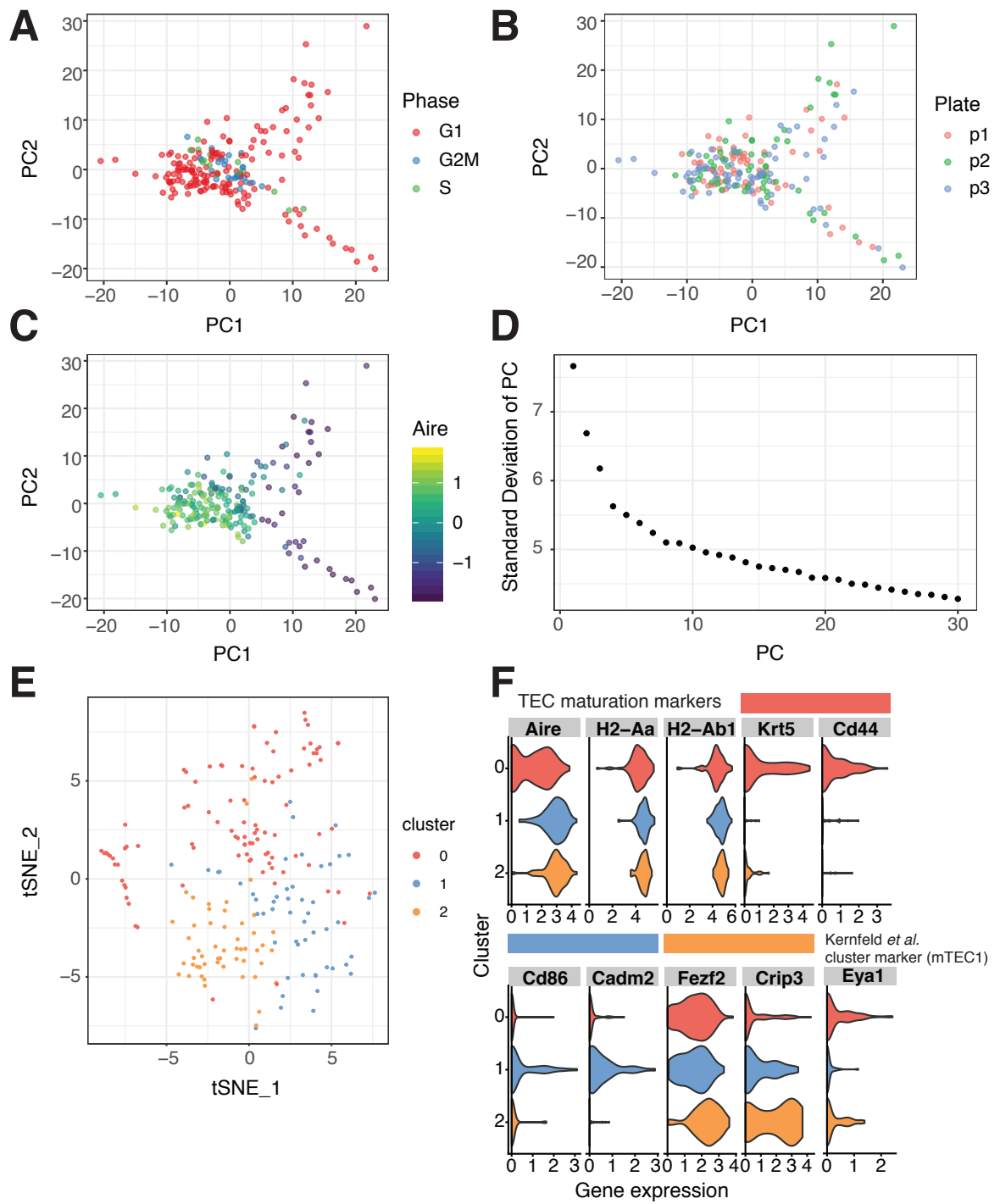
To prioritise TRSF showing robust and frequent expression in mature mTEC, single-cell RNA-sequencing data from mature mTEC were examined for the frequency and expression level of each factor. The initial quality assessment of the single-cell dataset based on for instance number of genes expressed or alignment rate, revealed 201 single cells with high-quality data (average 3.9 M read pairs per cell). The quality assessment and filtering are detailed in Section 2.1.7, Appendix Figures D.7, D.8 and Appendix Table D.4.

After quality assessment, the data was normalised using the R package *Seurat* (Section 2.1.7). The effects of number of counts and cell cycle phase were regressed out. As a first step, the data structure and potential subsets within the single mature mTEC were identified using PCA and t-SNE. Inspecting PC1 and PC2 from the PCA based on the most highly variable genes (n=2,101) revealed no obvious structure for the plate or cell cycle phase (Figure 5.11A, B). I noted that the expression level of *Aire*

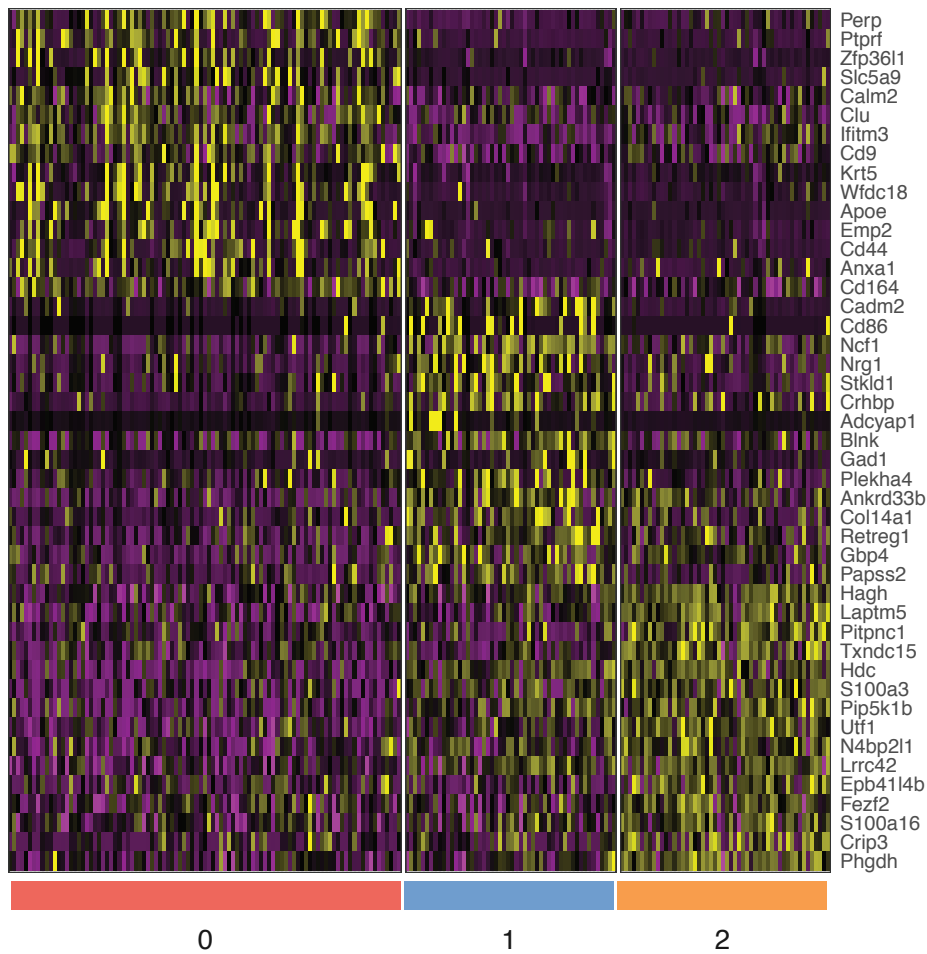
appeared to be positively correlated with PC1 (Figure 5.11C). The scree plot revealed that the first 20 PCs explained the majority of the variance in the data, thus these were used for the t-SNE and cluster analysis (Figure 5.11D). The clustering approach using a shared nearest neighbour graph and Louvain clustering implemented in *Seurat*, identified three clusters of mTEC. A small difference in *Aire* expression between the clusters was observed with cluster 0 showing lower expression levels of *Aire* (Figure 5.11F).

Amongst the identified cluster markers for cluster 0 were *Keratin5* (*Krt5*) and *Cd44* (Figure 5.11F, 5.12). While KRT5 is a widely used marker for mTEC, CD44 is an important molecule for T cell development. The block of CD44 was shown to result in reduced size of thymus lobes in fetal thymic organ culture (FTOC) experiments [225]. Cluster 1 was marked by for instance *Cd86* expression, which is an important molecule in antigen presentation. Furthermore, *Cadm2* was highly expressed in cluster 1. *Cadm2* contains an Ig-like domain but has not been described in the thymus, however, its family member *Cadm1* was reported as an important interaction partner for CRTAM on thymocytes [226]. Thymi from FTOC cultures treated with recombinant CADM1-Fc protein were significantly smaller than the ones treated with control Ig, suggesting that this family of proteins might have a role in TEC. The final cluster was characterised by a high expression level of *Fezf2*, which was described as a regulator of PGE [18], and *Crip3*. Kirchner *et al.* detected the thymus LIM protein (TLP) encoded by *Crip3* specifically in the stromal compartment of the thymus [227]. However, in the knockout mouse model, no effects on T cell development were observed.

In a recent single-cell study of thymus organogenesis, Kernfeld *et al.* described three clusters of mTEC across the development from embryonic day 12.5 to the day of birth [228]. One cluster was marked by high expression of *Fezf2*, while the other two cluster showed high level of *Krt5* and *Eya1*, respectively. Two of the mTEC clusters identified in adult mTEC here showed consistent markers with the clusters identified by Kernfeld *et al.* (Figure 5.11F).



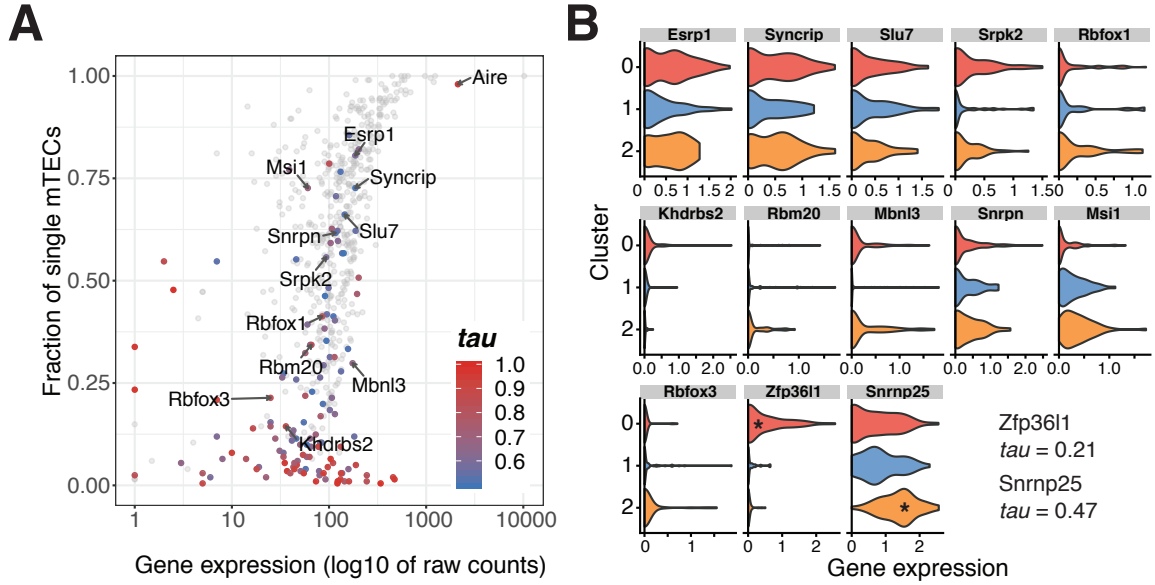
**Figure 5.11:** Principal component analysis (PCA), t-SNE and clustering of single mature mTEC. In **A**, **B** and **C** Principal Components 1 (PC1) and PC2 from a PCA based on all highly variable genes ( $n=2,101$ ) are shown colored for the cell cycle phase (**A**), the sorting plate (**B**) and *Aire* expression (**C**). **D**: Scree plot of the PCA. **E**: t-SNE based on the PCA in **A-C** using 20 PCs (perplexity = 30) colored by cluster assignments. Clusters were defined by the shared nearest neighbour graph plus Lovain clustering approach implemented in *Seurat* (resolution = 1). **F**: Violin plots of mTEC maturation markers and selected cluster markers for the identified clusters. A heatmap with additional cluster markers can be found in Figure 5.12.



**Figure 5.12:** Heatmap of identified cluster markers using *Seurat*. The top 15 cluster markers are shown for the 3 identified clusters. The clusters were identified using 20 PC and a resolution of 1, while the markers were identified using the Wilcoxon Rank Test in *Seurat*.

Given that promiscuously expressed genes were described to be infrequently expressed at the single-cell level, I focused on frequently expressed TRSF on the basis that they were more likely to have a functional role in mTEC (Figure 5.13A). As expected, robust and frequent expression was observed for the epithelial splicing factor *Esrp1*. Moreover, the majority of highly tissue-restricted splicing factors was expressed infrequently, consistent with the expectation for promiscuously expressed genes. Amongst frequently expressed, selected TRSF (Table 5.1) in single mature mTEC were *Msi1* (73 %), *Srpk2* (56 %), *Rbfox1* (41 %), *Rbm20* (34 %) and *Mbnl3* (30 %).

To understand if subsets of single mTEC with a specific peripheral splicing program existed, I investigated whether any marker genes for the identified mTEC clusters were



**Figure 5.13:** Splicing factor expression in single mature mTEC. **A:** Gene expression in mature mTEC and fraction of single cells positive for a splicing factor. Expression level was calculated as log10-normalised of raw counts, where the gene was detected. The known tissue-restricted splicing factors (TRSF, defined in Table 5.1), that showed expression in more than 10 % of single mature mTEC, are labelled. The color scale indicates the  $\tau$  value of the respective factor for TRSF ( $\tau > 0.5$ ). TRSF were plotted on top of ubiquitously expressed splicing factors (grey). *Aire* was detected in 97 % of single cells and added to the panel. **B:** Violin plots of selected TRSF with robust and frequent ( $> 10\%$ ) expression in single cells. In addition, cluster marker genes categorised as splicing-related genes were included.

TRSF. Given that none of the marker genes were TRSF, I examined the expression pattern of the robustly ( $> 10\%$  of single cells) expressed, selected TRSF across the identified mTEC clusters (Figure 5.13B). The violin plots revealed higher expression of *Rbfox1*, *Snrpn* and *Msi1* in cluster 2, while some TRSF including *Khdrbs2* and *Mbnl3* showed low expression in cluster 1 compared to their expression in other clusters. The intersection of marker genes for the identified mTEC cluster and all splicing factors obtained two marker genes, namely *Zfp3611* and *Snrnp25* (Figure 5.13B, bottom row). *Zfp3611* was a marker gene for cluster 0 and was also amongst the top 15 markers (Figure 5.12). *Snrnp25* had a  $\tau$  value close to the threshold for TRSF definition ( $\tau=0.47$ ) and was a marker gene for cluster 2. Consistent with the absence of TRSF amongst cluster marker genes, a hierarchical clustering of the single cells based on the expression of all TRSF did not reveal a grouping of cells that matched the clusters identified above (Data not shown). Overall, no clear patterns of expression indicative of a subset of mTEC

reusing a specific splicing program were observed.

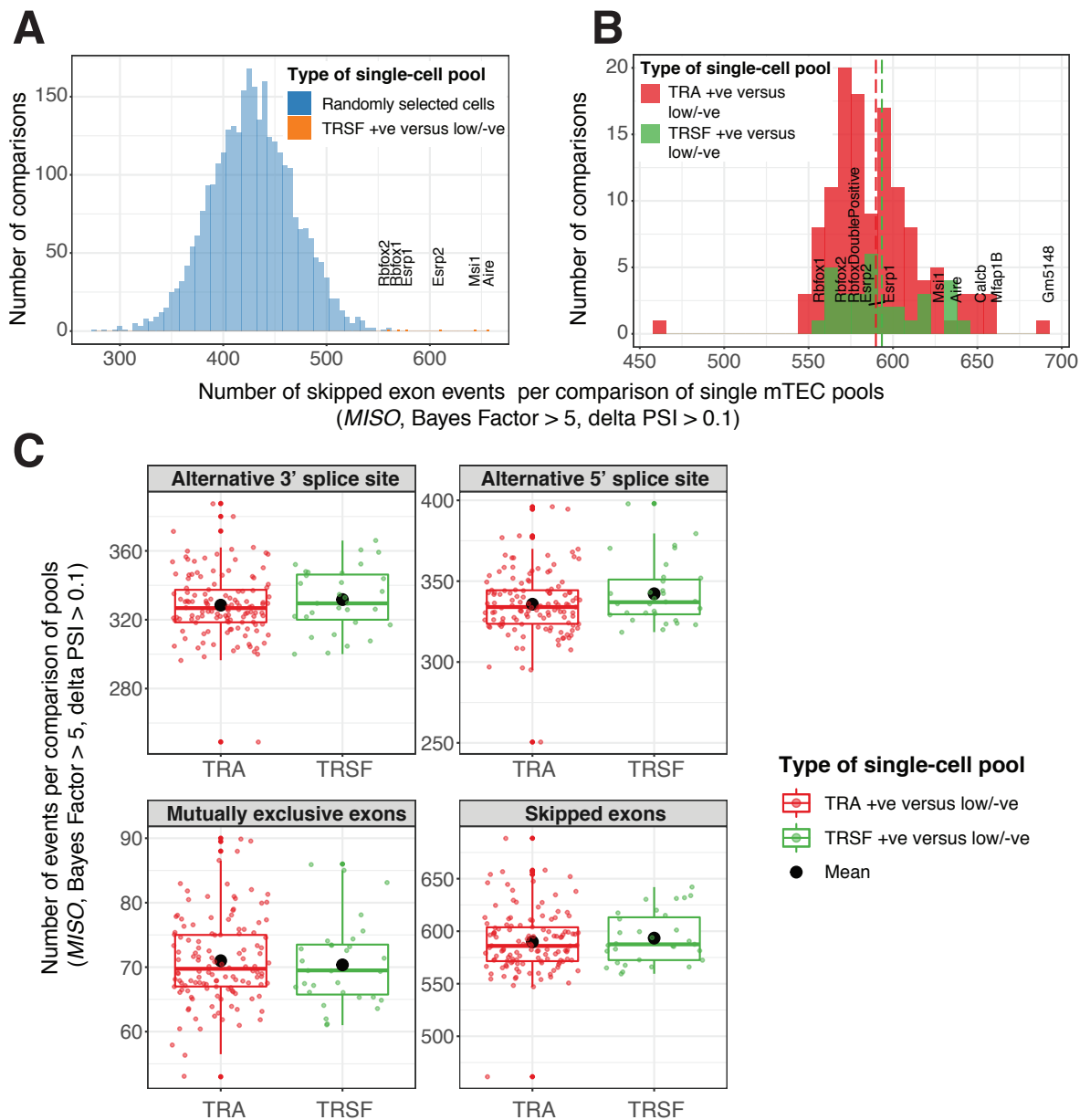
### 5.4.3 Assessment of splicing in single-cell pools stratified by splicing factor expression

To assess, if TEC expressing candidate splicing factors harboured an increased number of splice variants, such as for instance skipped exon events, I aimed to stratify the single cells by TRSF expression. Due to the unavailability of methods to assess splicing at the single-cell level directly, an approach using *in silico* single-cell pools was chosen.

#### 5.4.3.1 Comparison of pools stratified by selected tissue-restricted splicing factors with random pools

To investigate the influence of individual TRSF, a pipeline was developed including subsampling of reads to common depth between the single cells, generating pools (n=50 cells) and running the widely-used tool 'Mixture of Isoforms' (*MISO*) to assess differential splicing (Appendix Figure D.9A, Section 2.1.7.3, [139]). *MISO* applies a Bayesian model to determine the inclusion and exclusion levels based on the read evidence. On the one hand, single cells were stratified by the expression of TRSF or *Aire* into positive (TRSF<sup>+ve</sup>) and negative/low (TRSF<sup>-ve</sup>) single-cell pools. On the other hand, 50 random cells were merged into 2 k unique control pools. The *MISO* analysis for skipped exons (SE) events revealed a normal distribution with a mean of 427 for skipped exon (SE) events between randomly generated pools (Figures 5.14A, Appendix Figure D.9B). In contrast, the comparison between TRSF<sup>-ve</sup> pools and TRSF<sup>+ve</sup> pools showed a minimum of 550 SE events and the highest number for the *Aire* comparison. Surprisingly, the majority of comparisons had a higher number of SE events than the random pools. Amongst the TRSF, the highest difference in SE events was observed for *Msi1* followed by the two epithelial splicing factors *Esrp1* and *Esrp2*. The stratification by *Rbfox1* and *Rbfox2* resulted in a lower number of SE events.





**Figure 5.14:** Differential splicing events in *in silico* single-cell pools. **A:** Histograms showing the differential skipped exon events for comparisons of two types of single-cell pools. One type included comparisons between pools containing randomly selected cells (blue), while the other type of pools were based on single cells with high (+ve) or low (-ve) expression of a tissue-restricted splicing factor (TRSF; orange). In total, 2 k random pools were used to run 3 k random, unique comparisons between these. **B:** Histograms like in A showing pools based on single cells with high (+ve) or low (-ve) expression of a TRSF (green) or a tissue-restricted antigen (TRA; red). The dashed lines indicate the mean for each group. Median of 10 separate comparisons with different 10 -ve pools were plotted for each gene (Appendix Figure D.9C). **C:** Boxplots showing the number of differentially spliced events for four different splicing event types. The jittered dots represent individual comparisons and large black circles indicate the mean. Solid circles above or below the boxplot represent the outliers of the boxplot and these data points are therefore duplicated. PSI = percentage spliced-in, TRA = tissue-restricted antigen, TRSF = tissue-restricted splicing factor.

### 5.4.3.2 Comparison of pools stratified by expression of tissue-restricted splicing factors or tissue-restricted genes

As an additional control, the distribution of differential splicing events for pools of 126 randomly-selected tissue-restricted genes (referred to as tissue-restricted antigens, TRA) and additional TRSF (total  $n=32$  TRSF) was assessed. The number of SE events between TRA<sup>-ve</sup> and TRA<sup>+ve</sup> pools was comparable to the numbers observed for the TRSF comparisons (Figure 5.14B, Appendix Figure D.9B) suggesting that random pools were not an appropriate control for this analysis. In addition, I found differences in quality metrics between the initial TRSF<sup>-ve</sup> and TRSF<sup>+ve</sup> pools with for instance fewer genes detected in the TRSF<sup>-ve</sup> pools (Appendix Figure D.10). To produce a more robust estimate of the differential splicing events, I generated in total 10 gene<sup>-ve</sup> pools for each gene to compare to the one gene<sup>+ve</sup> pool and used the median of the 10 iterations (Appendix Figure D.9C). Assessing the comparisons between TRA<sup>-ve</sup> and TRA<sup>+ve</sup> pools revealed the second highest number of SE events for a gene called *Mfap1B* (Figure 5.14B). Although it was not included in the list of splicing-related genes defined here, *Mfap1B* is predicted to have a role in splicing according to *UniProtKB* and studies of gene ontology [229]. Inspecting in total 10 TRA comparisons with the highest number of SE events revealed no other TRA with a predicted role in splicing (Appendix Table D.6).

The comparison between TRSF and TRA-stratified pools revealed a small increase in the number of alternative 3' splice sites and alternative 5' splice sites (Figure 5.14C, not significant using Wilcoxon Rank Test). The number of SE events was comparable between the two types of *in silico* pools, while the number of mutually exclusive exons was slightly lower in the TRSF comparisons. The numbers of events for each TRSF and the TRA with highest number of SE events are given in Appendix Table D.5 and D.6, respectively.

In summary, the *in silico* pooling approach using *MISO* did not reveal significant differences in the numbers of differential splicing events that were detected between pools of (i) TRSF<sup>+ve</sup> vs TRSF<sup>-ve</sup> and (ii) TRA<sup>+ve</sup> vs TRA<sup>-ve</sup> mTEC. Possible reasons could

be either limited sensitivity from low numbers of cells or a bias in transcript coverage at the 3' end in the single-cell SmartSeq2 dataset, thus the results were not used for prioritisation. Given the observation of a high number of SE events in *Mfap1B*<sup>+ve</sup> pools compared to *Mfap1B*<sup>-ve</sup> pools suggested that the analysis could be meaningful. Moreover, it might suggest that splicing in TEC could be regulated by previously unknown or less well-described splicing factors.

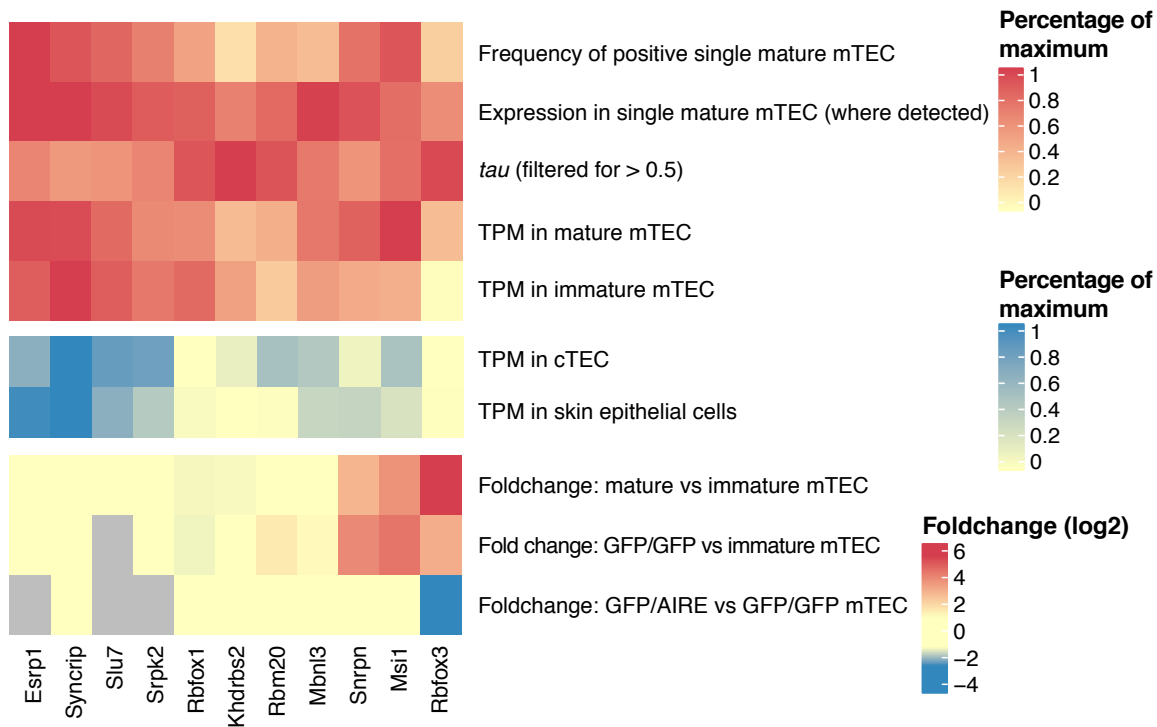
#### 5.4.4 *Rbfox1* as a candidate splicing factor for *in vivo* follow-up

To further investigate the potential mechanisms of splicing in TEC, I next sought to select candidate splicing factor(s) for follow-up experiments such as analysis of protein expression in TEC and an *in vivo* functional analysis (Chapter 6).

The selection criteria were (i) *Aire*-independent in mature mTEC, (ii) tissue-restricted expression in peripheral tissues, (iii) specific expression in mature mTEC, (iv) robust and frequent expression in single mTEC, (v) non-redundancy and (vi) a well-described functional role in alternative splicing. To assess these criteria, the results from the previous sections were summarised in a heatmap including all manually curated TRSF, that were detected in at least 10 % of single mature mTEC (Figure 5.15, Appendix Tables D.7 and D.8).

Following inspection of the heatmap, *Mbnl3*, *Khdrbs2* and *Rbfox3* were excluded due to their low frequency in single cells suggesting that they were promiscuously expressed. Moreover, *Rbfox3* was strongly downregulated in the absence of AIRE and therefore classified as an *Aire*-regulated gene. The factors *Syncrin*, *Slu7*, *Esrp1* and *Srpk2* were the least tissue-restricted as defined by *tau* and showed expression across all mTEC types, cTEC and skin epithelial cells. Thus, *Esrp1* was considered to be a general epithelial-restricted splicing factor.

After excluding the above factors, *Msi1*, *Snrpn*, *Rbfox1* and *Rbm20* were left. The factor *Msi1* was excluded due to its controversial role in alternative splicing regulation. It was implicated in photoreceptor splicing by Murphy *et al.* [218], but also detected in the



**Figure 5.15:** Heatmap to prioritise tissue-restricted splicing factors (TRSF) for functional follow-up study. The top and middle heatmap was row-scaled for the percentage of maximum value in the respective row, while the bottom heatmap shows fold changes on the log2-scale from the *DESeq2* analyses (names as in Figure 5.10). The non-significant (Benjamini-Hochberg adjusted p-value > 0.05) fold changes are colored in grey, while changes of < 2 fold are represented in light yellow. TPM values were generated with *Salmon* (as used in Figure 5.9). The TRSF were pre-filtered for  $\tau > 0.5$ , frequency detected in single cells > 0.1 (Figure 5.13) and manually curated for the role in alternative splicing (Table 5.1).

cytoplasm rather than the nucleus in a human luminal breast cancer, suggesting indirect effects on splicing [217]. Moreover, no loxP-targeted animal models were available for this factor at the time. The evidence for SNRPN's role in splicing was limited and thus it was not selected for the follow-up study. Besides mTEC, *Rbm20* was also expressed in cTEC, while *Rbfox1* was not detected in cTEC or skin epithelial cells suggesting that *Rbfox1* was specifically expressed in mTEC and hence might have a role in regulating splicing in mTEC. Furthermore, the literature on *Rbfox1* as an alternative splicing factor is very comprehensive. The Rbfox family and its potential target exons have been extensively described using its known binding motif and HITS-CLIP data [73]. Although RBFOX1 and RBFOX2 are highly similar, the possible redundancy of RBFOX1 and RBFOX2 could be mitigated by using a double conditional knockout mouse model. Such mice with

loxP sites in both *Rbfox1* and *Rbfox2* were available for the generation of a conditional knockout using the loxP/cre recombinase system [145, 146]. Taken together, *Rbfox1* was chosen as a candidate for functional follow-up experiments in Chapter 6.

## 5.5 Discussion

The first part of this computational evaluation of possible splicing mechanism in mTEC included the assessment of AIRE's role in splicing and the effect of TEC maturation on splicing. Previously, the role of AIRE in splicing had been controversial, thus I aimed to assess it using the deep-sequenced, paired-end RNA-sequencing data. The analysis revealed a limited effect of AIRE on splicing in mature mTEC (Figure 5.1), which is consistent with the recent publication by Danan-Gotthold *et al.* [103].

Following the finding of a relatively small role for *Aire* in alternative splicing, I next investigated whether *Aire* might have a larger role in shaping transcript usage. The deep-sequenced TEC population data offered a new opportunity to assess the transcript regulation due to increased power to detect differentially used transcripts. Using *DESeq2* and *Sleuth*, I described an effect of AIRE on transcripts in an additional 170 and 497 genes relative to gene level differential expression analysis, respectively. Examples of such genes were *Gpr179* or *Tpbp*, which were not regulated by AIRE at the gene-level but in which AIRE might modulate the choice of transcript produced from the locus. Previously, Meredith *et al.* had observed differentially expressed exons in genes not regulated by AIRE, when comparing wildtype to *Aire*-KO mTEC. The authors further concluded that AIRE enhances exon inclusion. Although I showed a limited role for AIRE in regulation of alternative splicing, which is consistent with a previous report by Danan-Gotthold *et al.* [103], the potential role of AIRE in regulating transcript usage was compatible with the conclusion by Meredith *et al.* [41].

Interestingly, I observed a significantly higher level of intron retention in immature mTEC compared to mature mTEC. Intron retention has only been studied more

extensively in recent years. Multiple roles for RI in gene expression regulation have been described [230]. These include nuclear retention plus degradation, RNA stabilisation and translation initiation. In general, an increase in intron retention has been described in terminally differentiated cells relative to immature cells, as is found, for instance, in mature neurons [231]. Braunschweig *et al.* detected an increase in RI upon maturation and an association of stalled RNA-Polymerase II with RI regions, suggesting that increased RI is used to downregulate gene expression during differentiation. An increase of RI events was also observed across erythroid differentiation in isolated human cells [232]. In contrast, the level of intron retention was observed to be higher before T cell activation [233], where RI could function as a 'priming' signal to have transcripts available for rapid increases in gene expression. Here, I observed a higher number of RI in the less mature cells, which would not fit with the general trend observed for neuronal or erythroid differentiation. However, it has recently been shown that terminally differentiated mTEC downregulated the expression of MHCII [15] and therefore have been included in the 'immature' MHCII low population that was analysed in this chapter (10-15 % of immature mTEC [Georg Holländer, Personal Communication]). As a consequence, it is possible that the increase in RI is either a novel biological feature of immature mTEC or could be caused by a mixture of immature and terminally differentiated mTEC in the RNA-sequencing sample. An example of a more frequently retained intron in immature mTEC was in the gene *Apeh*. Recently, *Apeh* was described as a regulator of the proteasome [186]. The high inclusion level for a RI in immature mTEC could indicate the storage of unprocessed transcripts to be upregulated rapidly upon maturation to mature mTEC. Furthermore, retained introns may be used by immature mTEC to prevent for instance the PGE machinery to interfere with mTEC differentiation. To identify and validate novel RI events, novel RI-containing mT&T transcripts were examined for individual ONT reads covering the whole transcript structure. One identified example was the ribosomal protein RPL10A, which is important for ribosomal assembly [191]. A novel transcript in mTEC could potentially play a role in the regulation of translation, which is an important process to produce the wide range of TRA for presentation on the cell surface.

In the second part of this chapter, I showed that peripheral tissues and TEC are distinguishable by their expressed splicing factor profile (Figure 5.7) suggesting that splicing factors are part of the tissue-specific gene expression profile (reviewed by Baralle & Giudice [57]). It should be noted, that some tissues are more heterogeneous in cell type composition, while others such as mTEC and cerebellum are composed of one cell type. Although a clear cell type definition could result in a more distinct splicing factor expression profile, even heterogeneous tissues, such as for instance the digestive tract clustered together in the PCA as well as hierarchical clustering (Figures 5.6 and 5.9), indicating that they either shared the splicing factor profile or/and had shared cell types within the sample.

Assessing the overall expression of TRSF in mTEC revealed the absence of some peripheral splicing factors such as the brain-specific splicing factors *Nova1*, *Nova2* and *Celf4* (Figure 5.9). In addition, the levels of *Rbm11*, *Brd4* and *Srp3* were low in mTEC. Interestingly, the absence of the splicing factor *Srrm4* provides a likely explanation for the absence of a brain-specific microexon set in mTEC (Figure 4.6) as this factor is a main regulator of microexon inclusion [58]. In contrast, factors such as *Msi1*, *Rbfox1* and *Snrpn* were highly expressed in mature mTEC, which was also reflected by frequent and robust expression of these at the single-cell level (Figure 5.13). Furthermore, two splicing factors described as *Aire*-independent and upregulated in mature mTEC by St. Pierre *et al.*, were frequently detected in single mature mTEC: *Larp6* was found in 50 % of single cells, while 80 % were positive for *Synj2*. These two factors were not included in the prioritisation for follow-up experiments due to the lack of literature evidence for a functional role in alternative splicing.

The observation of a specific set of highly expressed TRSF in mTEC suggested that splicing in mTEC is more likely a ‘directed’ process that depends on a particular subset of TRSF than a consequence of promiscuously-expressed splicing factors. The presence of a set of highly and frequently expressed splicing factors at the single-cell level supported this hypothesis. To perform functional follow-up experiments (Chapter 6), I aimed to

computationally prioritise a TRSF taking several criteria into account (Section 5.4): the criteria included that the factor should be unaffected by AIRE, should be robust and frequently expressed in single mature mTEC, be found in mature mTEC but not other epithelial cells and have a well-described role in alternative splicing.

One approach was to assess TRSF expression in single cells to examine potential subgroups of mature mTEC with distinct TRSF profiles. However, no clear subgroups of single cells with respect to TRSF expression were observed. As an alternative approach, single cells were stratified by TRSF expression, *in silico* pooled and differential splicing assessed (Figure 5.14). Due to the absence of published splicing tools for the single-cell level at the time, the pooling approach was necessary. The approach identified a high number of SE events between pools, that were low and high in expression for a TRA, namely *MfapB1*, that has been predicted to play a role in splicing [229]. However, no significant difference in the number of alternative splicing (AS) events for known alternative splicing categories could be observed between TRSF and other tissue-restricted genes. This suggested a potential methodological issue with the analysis or - less likely - that the TRSF surveyed do not play a major role in splicing in these cells. To rule out a methodological issue analyses of positive and negative control datasets would be required, which are currently unavailable. Methodological issues could be a low sensitivity to detect splicing events (given the small number of cells or pools used). Furthermore, the coverage of splicing events could be insufficient. Library preparation using the SmartSeq2 protocol should result in reads covering the whole transcript and inspecting example coverage plots for single cells did not reveal a marked 3' bias. Given that the *in silico* pooling strategy did not reveal differences between the numbers of differential splicing events that were found in comparisons of (i) TRA<sup>+ve</sup> vs TRA<sup>-ve</sup> and (ii) TRSF<sup>+ve</sup> vs TRSF<sup>-ve</sup> pools, it was not used for further prioritisation. Recently, methods have been developed to analyse splicing at the single-cell level, which could be used to repeat the analysis with cells stratified by TRSF expression [234, 235].



As discussed in Section 5.4.4, *Rbfox1* was selected as a candidate factor for experimental follow-up as it showed robust and frequent expression at the single-cell level and specific expression in mature mTEC but not in cTEC or skin epithelial cells (Figure 5.15). Further investigations would be needed to confirm the presence of protein in mouse thymus sections, before *Rbfox1* could be probed in a TEC-targeted knockout mouse model. The members of the highly conserved Rbfox splicing factor family have a well-established role as alternative splicing regulators during development and in neuronal and muscle cell types (reviewed by Conboy [68]). The characteristic RNA recognition motif (RRM) of the Rbfox family is identical between RBFOX1 and RBFOX2 indicating a possible redundancy between the two factors [236]. However, separate conditional knockout mouse models of *Rbfox1* or *Rbfox2* in the central nervous system stem and neural progenitor cells showed severe phenotypes in the brain [145, 146]. Although RBFOX2 was upregulated in the conditional *Nestin*-cre knockout of *Rbfox1*, a phenotype was observed, suggesting that they are only partially redundant [145]. Brain-specific knockout mouse models of both *Rbfox1* and *Rbfox2* died perinatally suggesting that the two factors were essential for brain development and function [146]. Apart from neuronal tissues, RBFOX1 was also shown to regulate splicing in muscle cells. A conditional knockout in adult muscle cells using an inducible *Acta1*-cre had no obvious myopathy but reduced force generation and altered calcium signalling in muscles [237]. RNA-sequencing from this mouse model revealed around 200 differentially spliced events, which were enriched in GO categories relating to cytoskeleton organization and myofibrillar structure.

## Conclusion

In this chapter, I assessed potential mechanisms of splicing in mTEC finding a lower number of differentially spliced events in the *Aire*-KO versus *Aire*-positive mTEC than was observed for immature versus mature mTEC, indicating a limited role for *Aire* in regulating splicing in this cell type. Further analyses on the transcript level suggest that AIRE may have a more important role in transcript usage. In addition, analysis of the expression of TRSF showed the presence of specific set of TRSF in mTEC suggesting the 'directed' expression of such genes could contribute to the high transcriptome diversity of

these cells, as observed in Chapter 4. Finally, I identified *Rbfox1* as a suitable candidate for further investigation and possible *in vivo* study of splicing in mTEC. *Rbfox1* showed robust and frequent expression at the single-cell level as well as specific expression in mature mTEC but not in other epithelial cells such as cTEC or skin epithelial cells. *Rbfox1*'s role in mTEC will be further investigated in Chapter 6.

# 6

## *In vivo* function of *Rbfox1* and *Rbfox2* in thymic epithelial cells

### Contents

---

|            |                                                                                                                                                                   |            |
|------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------|
| <b>6.1</b> | <b>Background . . . . .</b>                                                                                                                                       | <b>158</b> |
| <b>6.2</b> | <b>Validation of <i>Rbfox1</i> and <i>Rbfox2</i> as candidates for TEC-targeted knockout mouse models . . . . .</b>                                               | <b>160</b> |
| 6.2.1      | Validation of proteins from the Rbfox family in TEC . . . . .                                                                                                     | 161        |
| 6.2.2      | Assessment of <i>Rbfox1</i> transcript structures in TEC . . . . .                                                                                                | 162        |
| 6.2.3      | Summary of available mouse models for loxP-cre system . . . . .                                                                                                   | 165        |
| 6.2.4      | Targeted <i>Rbfox1</i> <sup>lox/lox</sup> : <i>Rbfox2</i> <sup>lox/lox</sup> mice showed no difference in thymus architecture compared to wildtype mice . . . . . | 167        |
| <b>6.3</b> | <b>TEC-targeted loss of <i>Rbfox1</i> expression . . . . .</b>                                                                                                    | <b>170</b> |
| 6.3.1      | Normal thymus architecture in mice lacking <i>Rbfox1</i> expression in TEC . . . . .                                                                              | 170        |
| 6.3.2      | <i>Rbfox1</i> expression in TEC is dispensable for TEC maturation and the expression of AIRE . . . . .                                                            | 170        |
| <b>6.4</b> | <b>TEC-targeted loss of <i>Rbfox2</i> expression . . . . .</b>                                                                                                    | <b>174</b> |
| 6.4.1      | Unaffected thymus cellularity in TEC with a loss of <i>Rbfox2</i> expression . . . . .                                                                            | 174        |
| 6.4.2      | Reduced number of mTEC upon loss of <i>Rbfox2</i> expression . . . . .                                                                                            | 176        |
| <b>6.5</b> | <b>Transcriptome analysis of mTEC with a loss of <i>Rbfox1</i> or <i>Rbfox2</i> . . . . .</b>                                                                     | <b>180</b> |
| 6.5.1      | Small number of differentially expressed genes upon loss of <i>Rbfox1</i> and <i>Rbfox2</i> in mature mTEC . . . . .                                              | 184        |
| 6.5.2      | Assessing the effects of <i>Rbfox1</i> or <i>Rbfox2</i> loss on splicing in mature mTEC . . . . .                                                                 | 188        |
| <b>6.6</b> | <b>Breeding of the conditional double-knockout mice . . . . .</b>                                                                                                 | <b>196</b> |
| <b>6.7</b> | <b>Discussion . . . . .</b>                                                                                                                                       | <b>196</b> |

---

## 6.1 Background

Conditional knockout mouse models are the 'gold-standard' to assess a genes' function in the *in vivo* context of development, homeostasis or disease. In thymic epithelial cells, such models can reveal the specific contributions of a gene to thymus development and the selection of thymocytes. To date, no TEC-specific conditional knockout mouse model of a splicing factor has been published. Thus, the functional contribution of specific splicing factors to self-antigen splicing in TEC is unknown.

One aim of this DPhil project was the identification of a candidate splicing factor for functional follow-up in a conditional mouse model. In the previous chapter, the tissue-restricted splicing factor (TRSF) *Rbfox1* was selected due to (i) its robust expression in TEC at the population and single-cell level, (ii) its specific expression in mTEC but not in other epithelial cells and (iii) its known contribution to tissue-specific splicing.

The Rbfox (Fox) family of TRSF includes three well-described splicing factors, namely Rbfox1 (Fox1, A2bp1), Rbfox2 (Rbm9) and Rbfox3 (NeuN). The characteristic, shared RNA recognition motif (RRM) domain is identical between RBFOX1 and RBFOX2 in human and mouse at the protein level, while 94 % of sequence is shared with RBFOX3 [238, 236]. The RNA element recognised by the Rbfox family is (U)GCAUG [239, 240, 241]. Depending on the localisation of the (U)GCAUG element in the downstream or upstream intron, Rbfox regulates the inclusion or exclusion of an alternative exon, respectively [236, 242](Figure 1.4). Based on this observation and data from high-throughput sequencing of RNA isolated by crosslinking immunoprecipitation (HITS-CLIP), which identifies binding sites of proteins to RNA, a splicing network for Rbfox was generated, suggesting that Rbfox regulates more than 1,000 exons [73]. In 54 % of these splicing events Rbfox was predicted to regulate inclusion, while in only 14 % of the events exclusion was predicted. In 32 % the direction of regulation was unknown. The events considered in the splicing network were cassette exons, two consecutive cassette exons - termed tandem cassette exons - and mutually exclusive exons. Although predictions of splicing

regulation and target exons have been published, the complexity of splicing regulation by various combinations of splicing factors and locations of binding sites is only partially understood (reviewed in Fu & Ares [56]). Hence, it is not trivial to assign one splicing factor to one specific splicing event.

The importance of the Rbfox family has been strengthened by reports of mutations in Rbfox genes leading to autism spectrum disorder, schizophrenia or epilepsy [243, 59, 61, 60]. To gain better understanding of the role of Rbfox in the brain, *Rbfox1* and *Rbfox2* conditional knockout animals using a *Nestin*-cre driver for tissue-targeted deletion were generated. Mice with a conditional loss of *Rbfox1* showed epileptic behaviour and increased neuronal hyperexcitation [145], whereas the loss of *Rbfox2* resulted in defects in cerebellar development and in the adult brain architecture [146]. A conditional knockout of *Rbfox1* in skeletal muscle also demonstrated a requirement for *Rbfox1* in muscle function [237].

The studies of RBFOX1's role in the brain showed that in the absence of RBFOX1, RBFOX2 was upregulated indicating that *Rbfox2* might partially restore the splicing defects occurring in a single-knockout of *Rbfox1* [145, 146]. However, some RBFOX1 target transcripts were not spliced by RBFOX2 as measured by Affymetrix Exon Junctions Arrays, hence the authors concluded that the two proteins were not fully redundant. The conditional knockout of both *Rbfox1* and *Rbfox2* using *Nestin*-cre showed a severe phenotype with mice dying perinatally indicating that either of these proteins was required for brain development and function [146]. In addition, the authors found that *Rbfox1*<sup>lox/WT</sup>:*Rbfox2*<sup>lox/lox</sup>:*Nestin*-cre<sup>+/-</sup> animals were small, suffered from severe ataxia and 93 % died or required euthanasia by four weeks of age. In summary, the loss of both *Rbfox1* and *Rbfox2* had severe consequences for brain development and function. One copy of *Rbfox1* was not able to fully restore the phenotype confirming the importance of functional splicing factors in the brain.

## Aim

Given the possible redundancy of *Rbfox1* and *Rbfox2*, the single conditional knockout of *Rbfox1* might not be sufficient to understand its contribution to TEC development and function with respect to thymocyte selection. Therefore, I chose a strategy, that interrogated a conditional double- as well as individual single-knockout mice.

The first aim was (i) to validate the protein expression of RBFOX1 in TEC. Second, I aimed (ii) to cross mice with conditional alleles for either or both *Rbfox* family members (*Rbfox1*<sup>lox/lox</sup>;*Rbfox2*<sup>lox/lox</sup>) with mice carrying the cre recombinase under the control of the *FoxN1* promoter (*FoxN1-cre*<sup>+/-</sup>; Section 2.2.1.2 [144]). As a next step, I sought (iii) to assess the thymi of conditional knockout mice for size and architecture. Moreover, (iv) a comprehensive phenotyping of TEC using flow cytometry was needed to assess the effects of *Rbfox1* and/or *Rbfox2* on TEC development and maturation. Finally, to understand the changes in splicing and promiscuous gene expression in the absence of *Rbfox1* and/or *Rbfox2*, I aimed (v) to generate RNA-sequencing data and to interrogate this data for differences in alternative splicing.

## 6.2 Validation of *Rbfox1* and *Rbfox2* as candidates for TEC-targeted knockout mouse models

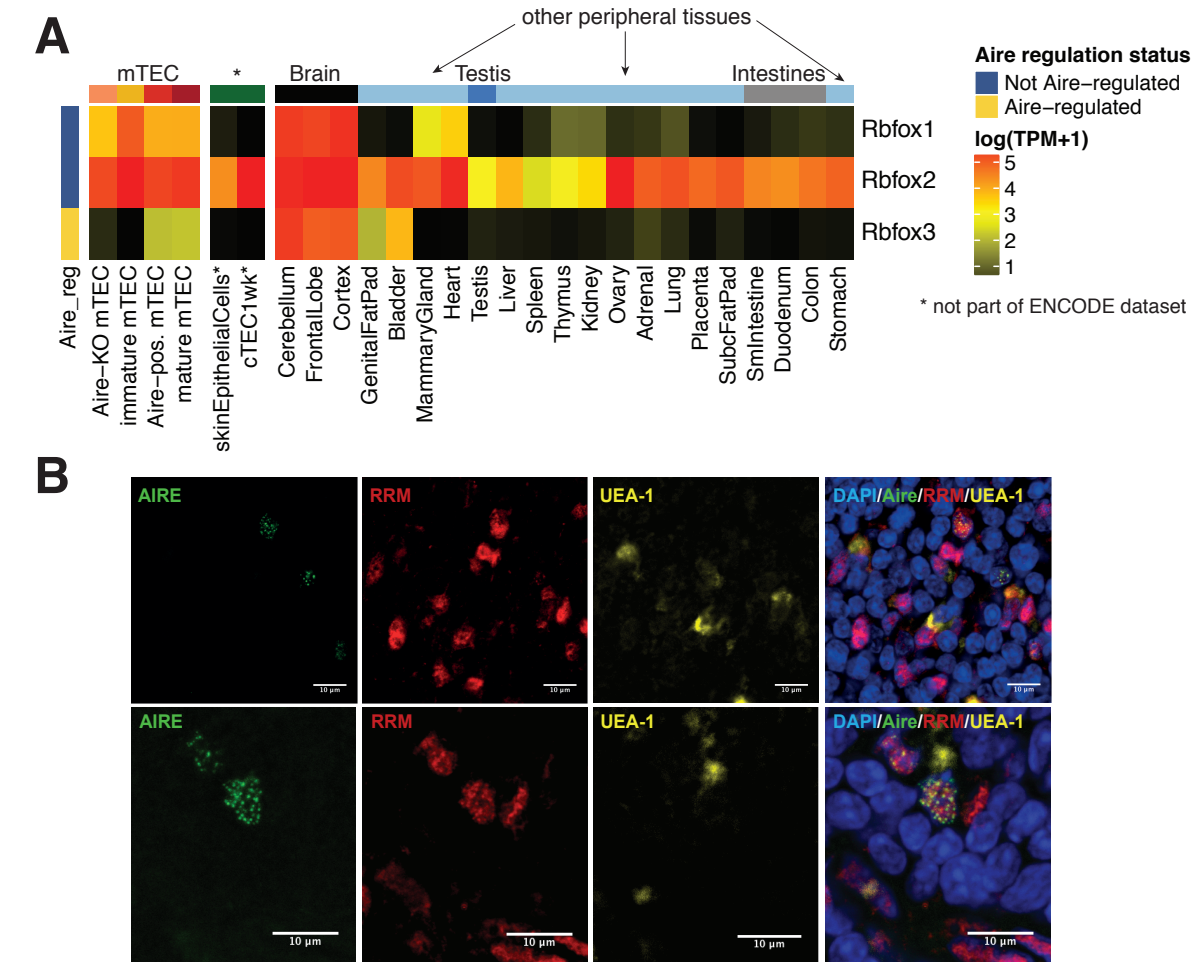
To assess the gene expression of individual *Rbfox* family members in mouse peripheral tissues including TEC, a heatmap of expression levels was generated, which showed the expression pattern for the *Rbfox* family members described in the literature: *Rbfox2* was found to be ubiquitously expressed in all mouse tissues tested (Figure 6.1A) [236, 244, 242], whereas *Rbfox1* and *Rbfox3* were more restricted in their expression pattern. In addition to reports of *Rbfox1* expression in neuronal tissues including the retina, *Rbfox1* was also detected in the heart in other reports and the data presented here (Figure 6.1A) [245, 246, 236]. This finding is consistent with previous reports of *Rbfox1* in muscle [239, 247]. In this dataset the heart was the only muscle tissues included, thus other muscle types could not be assessed. As discussed in Chapter 5, *Rbfox1* was also found in all mTEC types, but not in cTEC or other epithelial tissues including the skin (Figure

6.1A). *Rbfox3* showed expression in the genital fatpad and the bladder. Within the thymus, *Rbfox3* was only detected in mature and mature *Aire*-positive mTEC, suggesting that its expression was *Aire*-regulated. At the single-cell level, at least one *Rbfox* family member was detected in 72 % of single mTEC, while only 16 % showed gene expression of both *Rbfox1* and *Rbfox2* (n=201 single cells, Data not shown). Taken together, this transcriptomics data suggests that *Rbfox1* might be an interesting candidate for which to test a possible role in TEC function as its expression is detected in all mTEC stages, but not other epithelial cell types assessed.

### 6.2.1 Validation of proteins from the *Rbfox* family in TEC

Prior to proceeding with a conditional knockout model, I aimed to examine the RBFOX1 and RBFOX2 protein expression in TEC. Due to the lack of signals with RBFOX1- and RBFOX2-specific antibodies in thymic sections, an antibody directed against the common RRM domain of the *Rbfox* family was used. The antibody was kindly provided by Prof. Douglas Black from the University of California. Using immunohistochemistry, the nuclear localisation of the *Rbfox* family was demonstrated in TEC, whereas thymocytes did not express this family of proteins (Figure 6.1B). Additionally, the RRM signal frequently, but not exclusively, overlapped in TEC with their characteristic speckled pattern of AIRE expression, but was not limited to AIRE-positive nuclei (Figure 6.1B). As a control, a staining was performed without secondary antibody showing no signals for AIRE or RRM (Appendix Figure E.2).

The expression in mTEC of a protein containing the RRM domain was confirmed using flow cytometry (Figure 6.2A). Within mature mTEC (MHCII high) 58 % of cells were AIRE<sup>+</sup> and RRM<sup>+</sup>, while 36 % showed low expression of AIRE but high levels of RRM. To compare the RRM expression in AIRE<sup>-</sup> to AIRE<sup>+</sup> mature mTEC (MHCII high), the mature mTEC population was split by their AIRE expression (Figure 6.2A (v), left panel). Subsequently, the RRM expression in the two subpopulations of mature mTEC revealed that amongst AIRE-negative mTEC 83 % also showed signals for RRM (Figure 6.2A (v), right panel). In the AIRE<sup>+</sup> subpopulations the frequency of RRM<sup>+</sup> cells



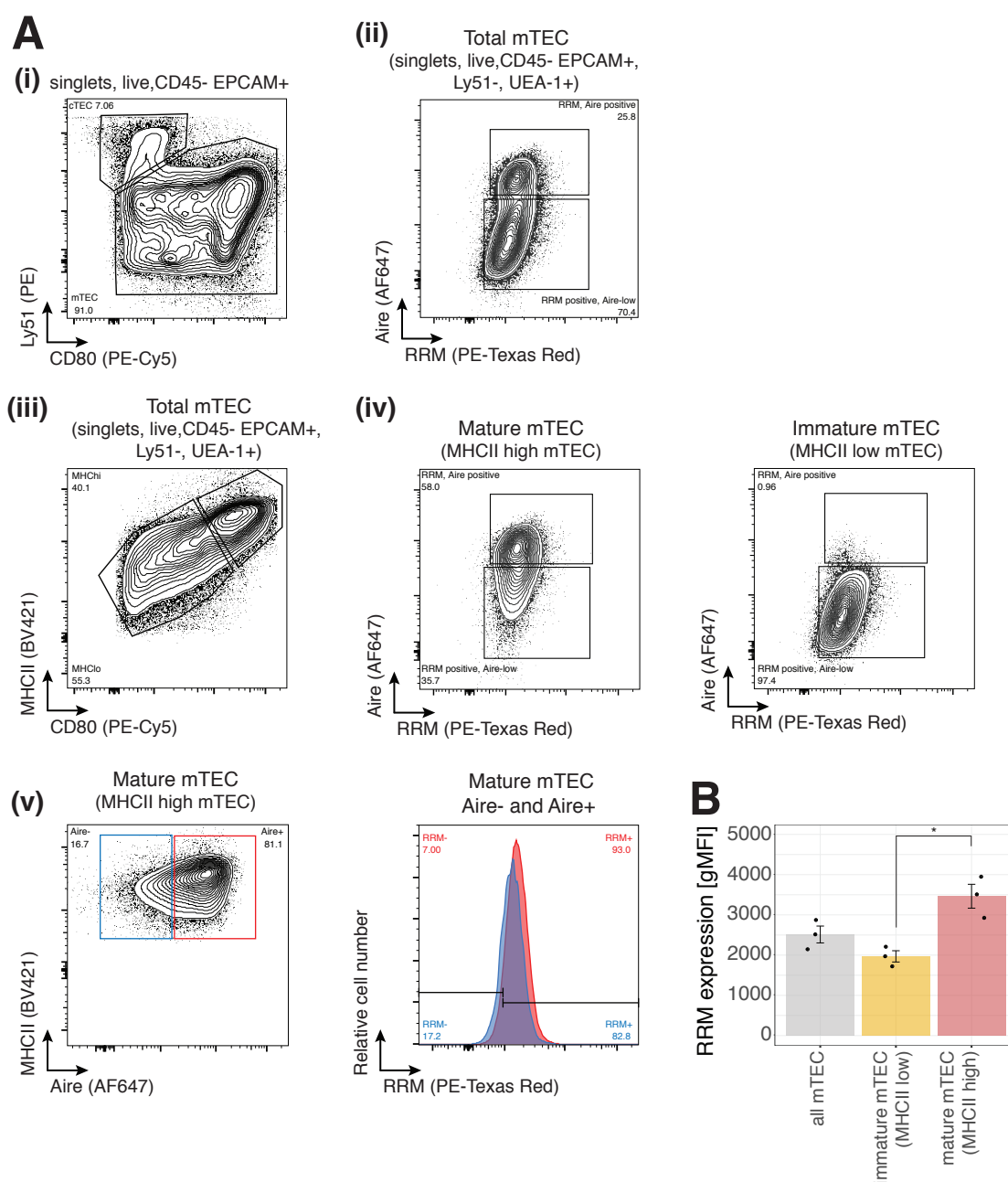
**Figure 6.1:** Expression of *Rbfox* family members at the transcript and protein level. **A:** Heatmap of the expression of the *Rbfox* family across different mouse tissues including TEC (clustered within group, Manhattan distance, complete linkage). **B:** Immunofluorescence staining for nuclear RRM domain containing proteins in sections from a mouse thymus (8  $\mu\text{m}$  sections). A control slide without secondary antibody is shown in Appendix Figure E.2. Images were taken with 63x magnification.

was 93 %. The expression level as measured by geometric mean fluorescence intensity (gMFI) was significantly, almost 2-fold higher in mature mTEC (MHCII high) compared to immature mTEC (MHCII low) (Figure 6.2B).

### 6.2.2 Assessment of *Rbfox1* transcript structures in TEC

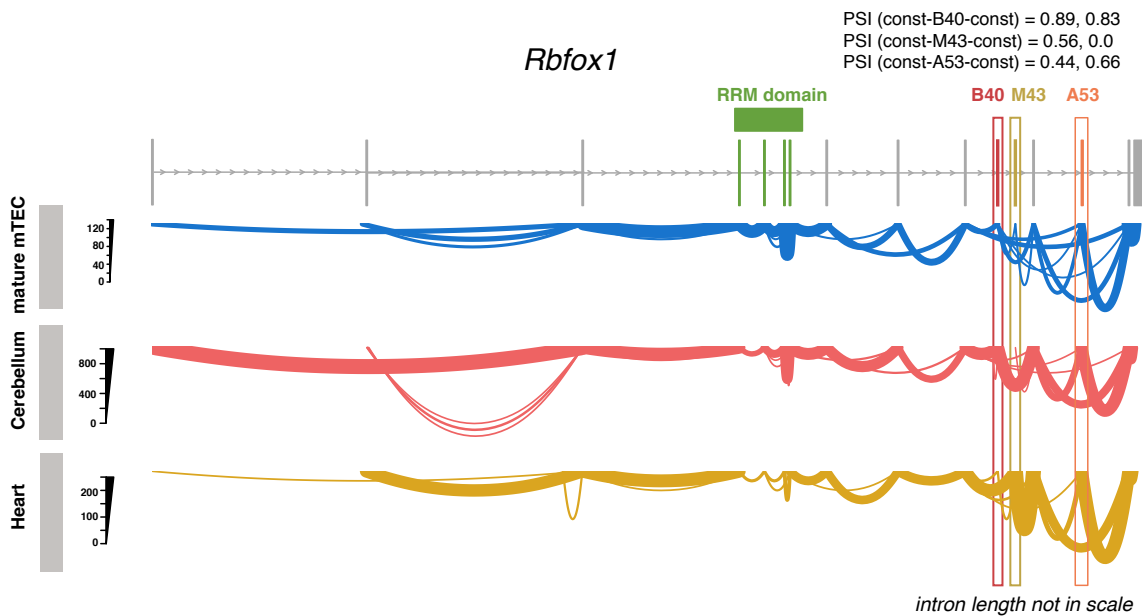
Given that nuclear localisation was shown for RRM-containing proteins but not RBFOX1 specifically, I aimed to establish that nuclear isoforms of *Rbfox1* are present in TEC. Both *Rbfox1* and *Rbfox2* have been reported to generate tissue-specific isoforms, with the main





**Figure 6.2:** Flow cytometric analysis of RBFOX family expression in wildtype mTEC. **A:** (i) Representative graph demonstrating gating for mTEC and cTEC from enriched cells (EPCAM<sup>+</sup>, CD45<sup>-</sup>) isolated from wildtype mice. Single, live cells were used as input (Appendix Figure E.1B). (ii) Representative graph demonstrating gating for AIRE and RRM in total mTEC as defined in (i). (iii) Representative graph demonstrating gating for mTEC subtypes defined by the expression of MHCII and CD80 and the AIRE and RRM expression in the two subtypes (iv). (v) Representative graph demonstrating gating for AIRE<sup>+</sup> and AIRE<sup>-</sup> mTEC amongst mature mTEC on the left (defined in (iii)). On the right, histograms of RRM expression in AIRE<sup>+</sup> and AIRE<sup>-</sup> gates are shown. **B:** Expression level of RRM measured by gMFI for all mTEC and mTEC subtypes as defined in (i) and (iii), respectively. Data was generated in one experiment with 3 female mice (4 weeks old). The significance between RRM expression levels was tested using a Welch Two Sample t-test (mean  $\pm$  SE is shown, \* p-value < 0.05).

distinction observed when comparing muscle and brain tissues [239]. In the muscle-specific isoform, an exon termed M43 was included, whereas another exon, B40, was found to be included in the cerebellar isoform. Furthermore, Nakahata *et al.* described an exon termed A53, which if included resulted in RBFOX1 being localised predominantly in the cytoplasm. To assess whether nuclear isoforms of *Rbfox1* are present in TEC, the transcript structures of *Rbfox1* in mature mTEC were examined (Figure 6.3). Moreover, the transcript structures in myocardium (heart) and cerebellum (brain) were compared to mature mTEC to investigate which peripheral forms were present in mTEC. Both isoforms of *Rbfox1* were detected in mature mTEC, with a higher proportion of transcripts characteristic of the cerebellar isoform. The percentage spliced-in (PSI) values for the cerebellar-specific exon were 0.89 and 0.83 in two mature mTEC replicates (60 M reads), while the PSI values for the myocardium-specific isoform were 0.56 and 0.0. Furthermore, the exon A53 promoting cytoplasmic location was infrequently included (PSI of 0.44, 0.66). Taken together, the assessment of transcript structures in *Rbfox1* suggested that the majority of *Rbfox1* transcripts represented the cerebellar, nuclear transcript version.

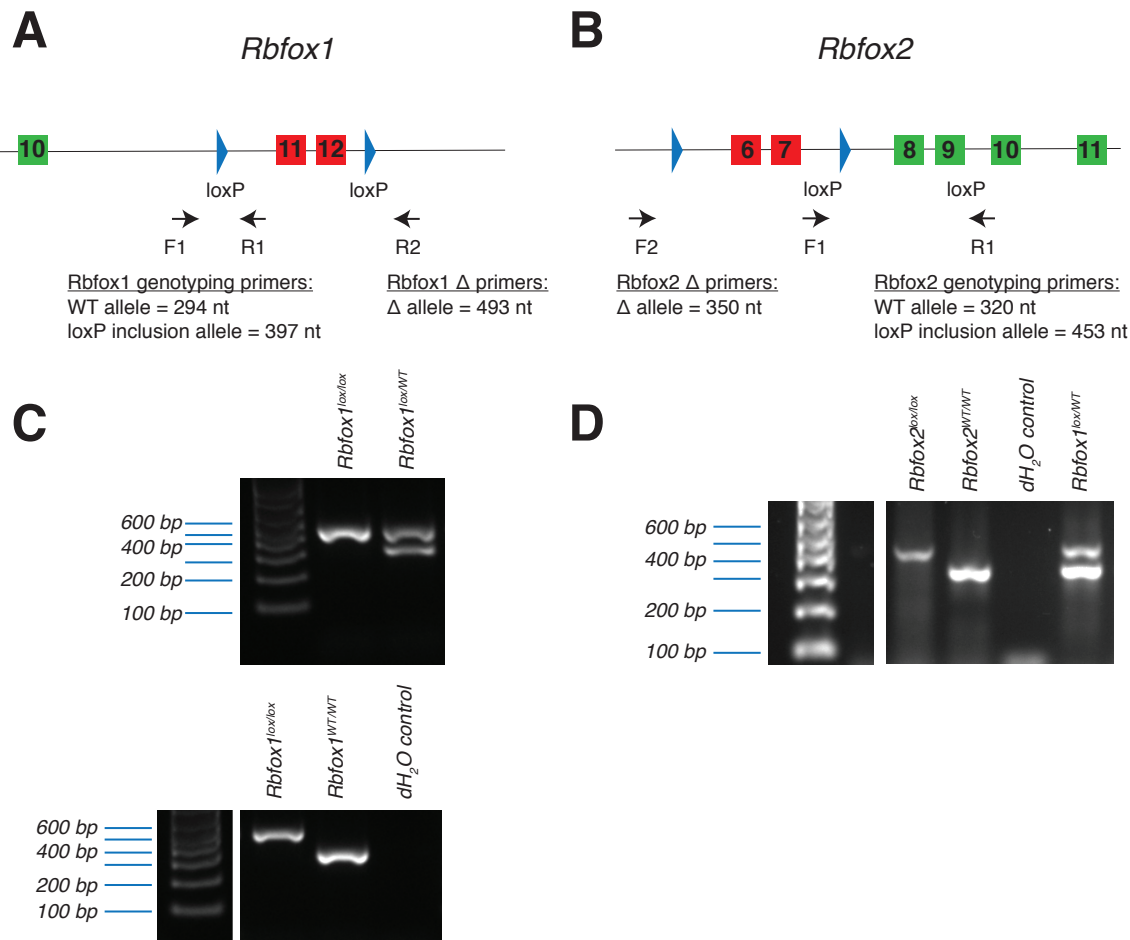


**Figure 6.3:** Sashimi plot and coverage track for *Rbfox1* in heart, cerebellum and mature mTEC. The cerebellar- (B40) and myocardium-specific (M43) exons as well as the RRM domain are highlighted at the top. All Sashimi plots were generated from Illumina datasets at 200 M reads. The indicated percentage spliced-in (PSI) values were generated by *rMATS* in two replicates of mature mTEC (60 M reads). The exon sequences for each PSI value are indicated and adapted from Nakahata *et al.* [239]. Const = constitutive exon.

These observations together with the protein validation confirmed *Rbfox1* as a suitable candidate for a TEC-targeted knockout mouse model.

### 6.2.3 Summary of available mouse models for loxP-cre system

To investigate the *in vivo* role of *Rbfox1* in TEC, I established a colony of mice with a TEC-targeted loss of *Rbfox1* and/or *Rbfox2* alleles using the expression of the Cre recombinase under the transcriptional control of the *FoxN1* locus, designated  $Rbfox1^{lox/lox}:Rbfox2^{lox/lox}:cre^{+/-}$  (DKO) and  $Rbfox1^{lox/lox}:cre^{+/-}$  (SKO1) or  $Rbfox2^{lox/lox}:cre^{+/-}$  (SKO2). Mice with the targeted alleles but lacking the Cre transgene are referred to as  $Rbfox1^{lox/lox}:Rbfox2^{lox/lox}:cre^{-/-}$  (WT). A breeding pair of  $Rbfox1^{lox/lox}:Rbfox2^{lox/lox}$  was kindly provided by Prof. Douglas Black. However, the received breeding pair was infertile for unknown reasons, which required a change of the original breeding strategy to a scheme mating  $Rbfox1^{lox/WT}:Rbfox2^{lox/WT}:cre^{+/-}$  with  $Rbfox1^{lox/WT}:Rbfox2^{lox/WT}:cre^{-/-}$ . This necessity resulted in a six months delay in establishing a colony. In parallel, mice with a loss of either *Rbfox1* or *Rbfox2* expression in TEC were generated. The targeted locus and representative analyses of genotypes are shown in Figure 6.4 [145, 146].

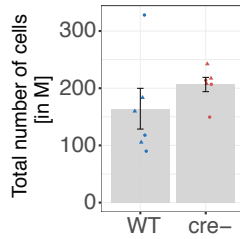


**Figure 6.4:** Information on *Rbfox1*- and *Rbfox2*-targeted mouse models. **A** and **B**: Genomic loci and location of loxP sites for *Rbfox1* (A) and *Rbfox2* (B) loci. The loci were modified from Gehman *et al.* [145, 146]. **C** and **D** are showing the respective genotyping PCR products in DNA isolated from mice carrying homozygous or heterozygous loxP sites in *Rbfox1* or *Rbfox2*, respectively.

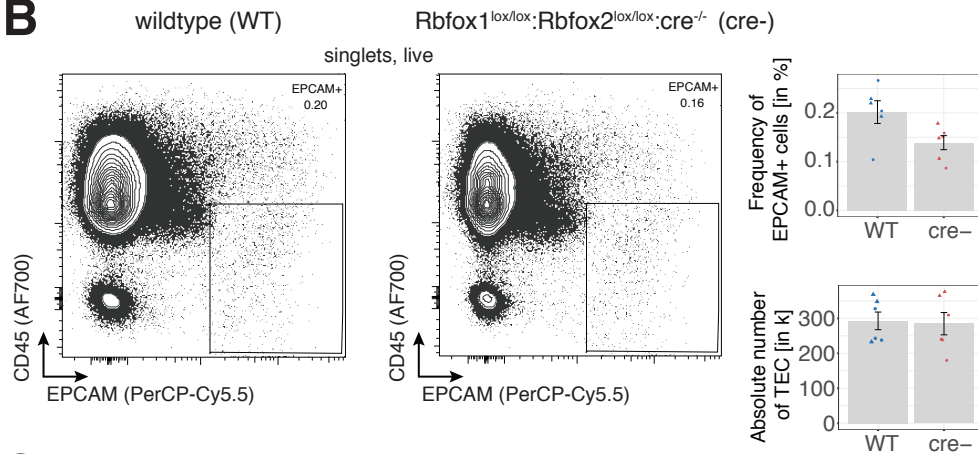
#### 6.2.4 Targeted Rbfox1<sup>lox/lox</sup>:Rbfox2<sup>lox/lox</sup> mice showed no difference in thymus architecture compared to wildtype mice

To ensure that TEC were not affected by the presence of the loxP sites, TEC from Rbfox1<sup>lox/lox</sup>:Rbfox2<sup>lox/lox</sup>:cre<sup>-/-</sup> were compared to TEC from wildtype animals (Figure 6.5). The overall thymic cellularity was not significantly different comparing these two groups of animals (Figure 6.5A). To further establish, whether the introduction of loxP sites had an impact on TEC differentiation, I next determined the frequency and phenotypes of EPCAM<sup>+</sup> cells. Well-defined TEC cell surface markers and the intracellular expression of AIRE were used to identify the separate TEC subpopulations. Although the total number of TEC was comparable between the two mouse strains, the frequency of EPCAM<sup>+</sup> cells was slightly higher in wildtype mice in two independent experiments (0.14 % versus 0.20 %,  $p=0.05$  using Welch Two Sample t-test; Figure 6.5B). Importantly and relevant to this study, the frequency of cTEC and mTEC as well as their different subpopulations were similar (Figures 6.5C and 6.6A). The fraction of AIRE<sup>+</sup> cells and the expression level of AIRE were not significantly different within the mTEC compartment (Figure 6.6B, C). Overall, the cellularity and phenotype of the distinct TEC compartments and the expression level of AIRE were comparable for targeted Rbfox1<sup>lox/lox</sup>:Rbfox2<sup>lox/lox</sup> and wildtype animals.

**A**

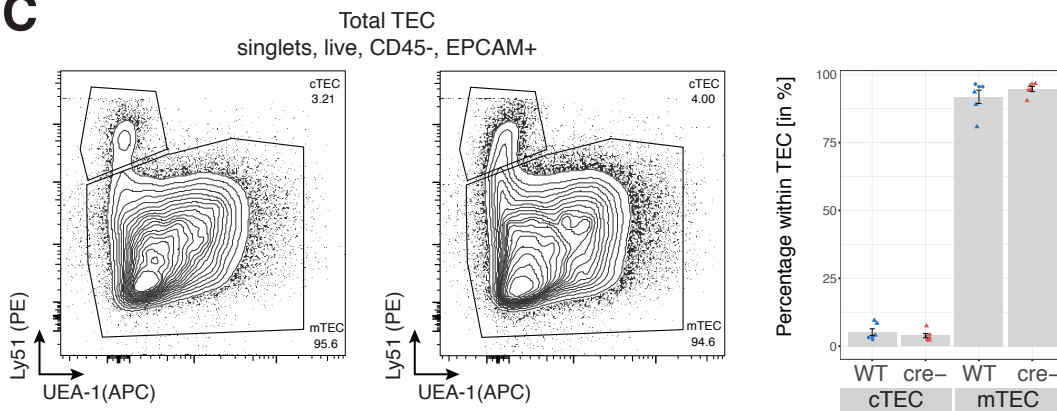


**B**

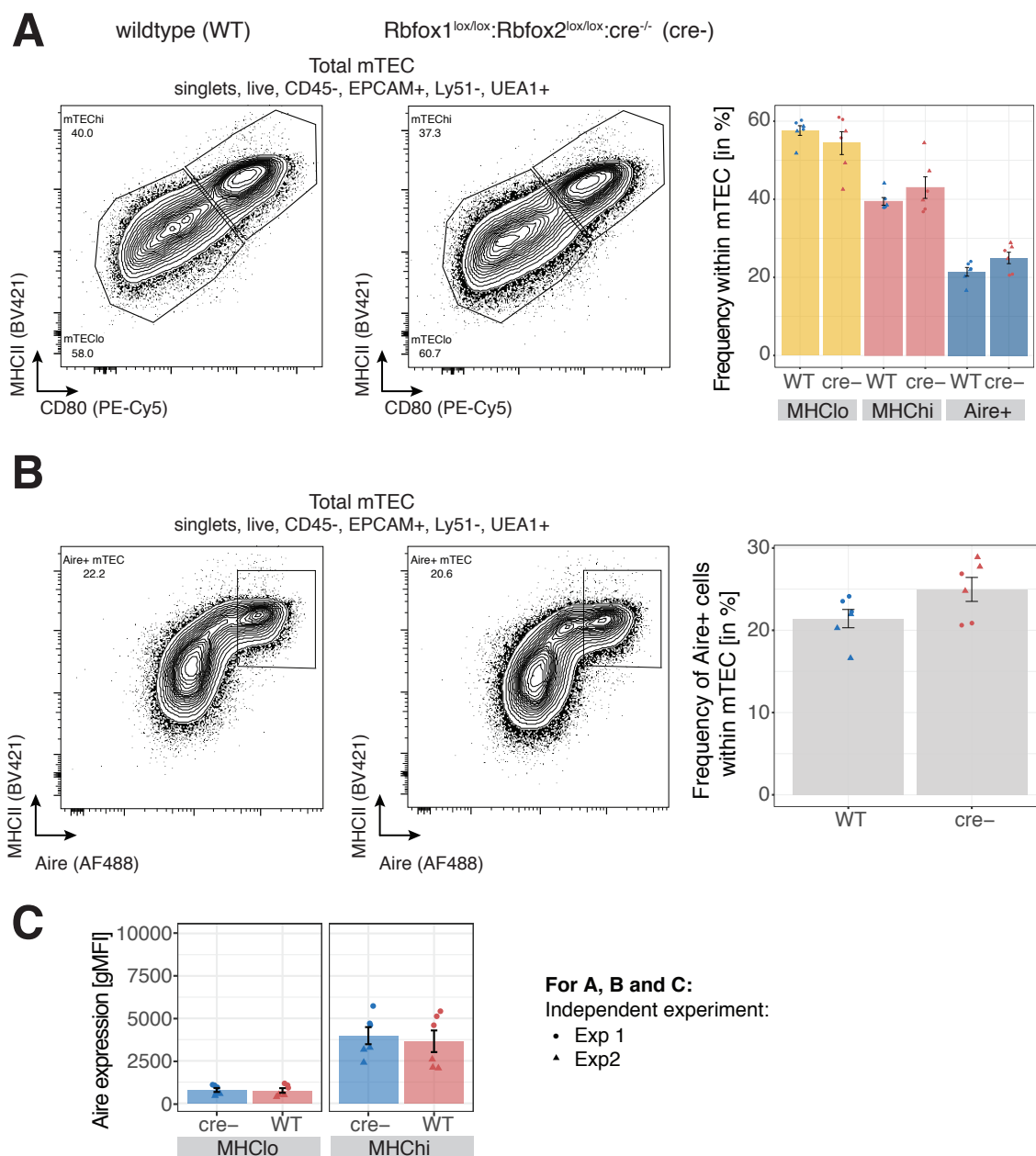


**For B and C:**  
Independent experiment  
• Exp1  
▲ Exp2

**C**



**Figure 6.5:** Comparative analysis of the thymus cellularity and the TEC compartment of wildtype (WT) and  $Rbfox1^{lox/lox}:Rbfox2^{lox/lox}$  (cre-) mice. **A:** Total cellularity of WT thymi compared to  $Rbfox1^{lox/lox}:Rbfox2^{lox/lox}$  thymi. **B:** Representative graph demonstrating gating for TEC (EPCAM<sup>+</sup>, CD45<sup>-</sup>) among non-enriched cells isolated from the indicated mouse strains (left). Their frequencies and absolute TEC cellularity are shown in bar graphs (right). A graph showing the gating for single, live cells is given in Appendix Figure E.1A. **C:** Representative graph demonstrating gating for TEC subtypes from enriched cells (EPCAM<sup>+</sup>, CD45<sup>-</sup>; left) and their frequencies (right). A graph showing the gating for single, live cells is given in Appendix Figure E.1B. Data was generated in two independent experiments with 6 female mice in each experiment (4-6 weeks old). The significance between genotypes was tested using a Welch Two Sample t-test and no significant differences of  $p < 0.05$  were found (mean  $\pm$  SE is shown in bar graphs).



**Figure 6.6:** Analysis of mTEC subpopulations and AIRE expression between wildtype animals (WT) and  $Rbfox1^{lox/lox};Rbfox2^{lox/lox};cre^{-/-}$  mice. **A:** Representative graph demonstrating gating for mTEC subtypes defined by MHCII and CD80 status (left) and their frequencies (right). Medullary mTEC were defined in Figure 6.5C. **B:** Representative graph demonstrating gating for AIRE<sup>+</sup> mTEC amongst all mTEC (EPCAM<sup>+</sup>, CD45<sup>-</sup>, Ly<sup>-</sup>, UEA-1<sup>+</sup>) isolated from the indicated mouse strains (left) and frequencies (right). **C:** AIRE expression in mTEC subpopulations defined in A given as gMFI. Data presented in bar graphs represents the combined results from two independent experiments with six female mice in each experiment (4-6 weeks old). The two independent experiments are indicated by different symbols in the bar graphs. No significant differences were observed for the comparison between  $Rbfox1^{lox/lox};Rbfox2^{lox/lox};cre^{-/-}$  and wildtype samples using a Welch Two Sample t-test (mean  $\pm$  SE shown in bar graphs).

## 6.3 TEC-targeted loss of *Rbfox1* expression

Given its role in splicing and its robust expression in mTEC, *Rbfox1* was a good candidate with a potential function in regulating splicing in mTEC. To probe the role of *Rbfox1* in TEC, TEC-targeted single knockouts of *Rbfox1* were generated (SKO1 hereafter). The *Rbfox1*<sup>lox/lox</sup>cre<sup>-/-</sup> mice are referred to as WT.

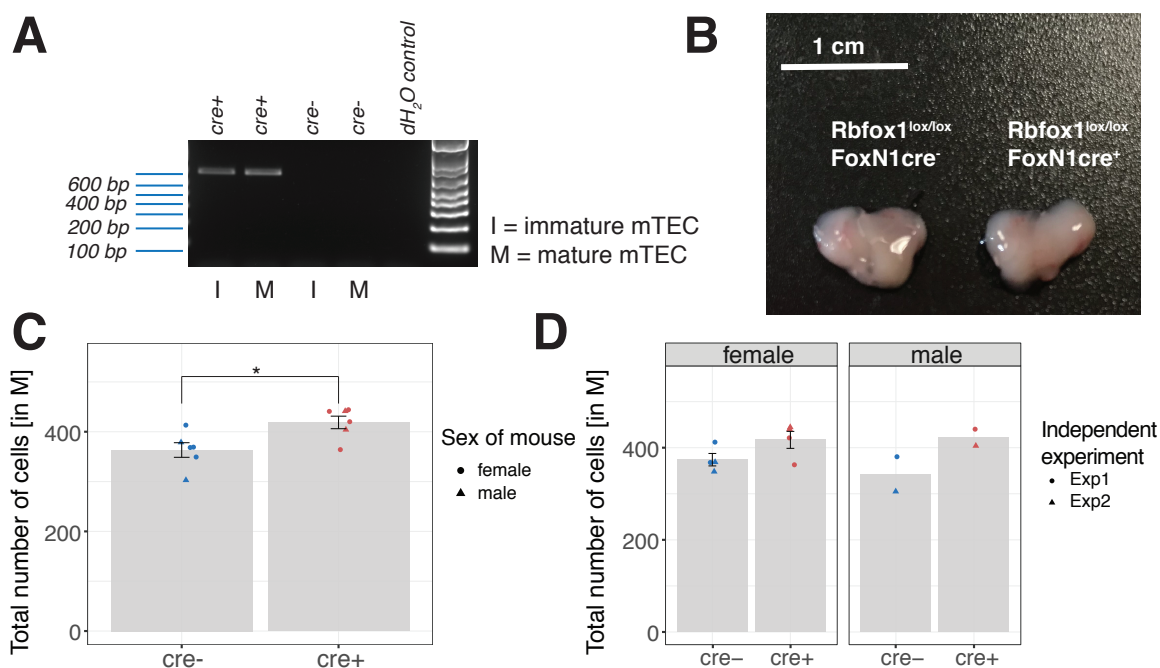
### 6.3.1 Normal thymus architecture in mice lacking *Rbfox1* expression in TEC

The excision of *Rbfox1* exons 11 and 12 in SKO1 mice was confirmed by PCR analysis of genomic DNA from FACS-sorted immature and mature mTEC (Figure 6.7A). The overall thymus size was comparable between SKO1 and WT mice, but an increase in total thymus cellularity was observed in the absence of *Rbfox1* (491 M in SKO1 versus 363 M in WT,  $p=0.02$  using Welch Two Sample t-test; Figure 6.7B, C). Using only the female mice for the statistical test revealed no significance (Figure 6.7D, 417 M versus 374 M,  $p=0.11$ ,  $n=4$  female mice per independent experiment). Due to the time constraints, it was not possible to perform the experiments with littermates of one sex, but to avoid biases the experiments were performed with sex-matched littermates.

### 6.3.2 *Rbfox1* expression in TEC is dispensable for TEC maturation and the expression of AIRE

To understand the TEC composition and differentiation in SKO1 mice, TEC phenotyping experiments were performed using (i) a sample without depletion of CD45<sup>+</sup> cells to determine the total number of epithelial cells (EPCAM<sup>+</sup>) and (ii) a sample enriched for TEC to determine TEC subtypes and AIRE expression by intracellular AIRE staining. For each experiment six mice were analysed and two independent experiments were performed. Parallel to an increased total cellularity (Figure 6.7), the absolute number of EPCAM<sup>+</sup> cells was also non-significantly increased in SKO1 mice relative to WT mice (674 k in SKO1 versus 561 k in WT,  $p=0.29$  using Welch Two Sample t-test). The frequencies of EPCAM<sup>+</sup> cells, however, remained unaffected (Figure 6.8A). A difference in the frequencies of cTEC and mTEC was not observed in SKO1 thymi (Figure 6.8B).

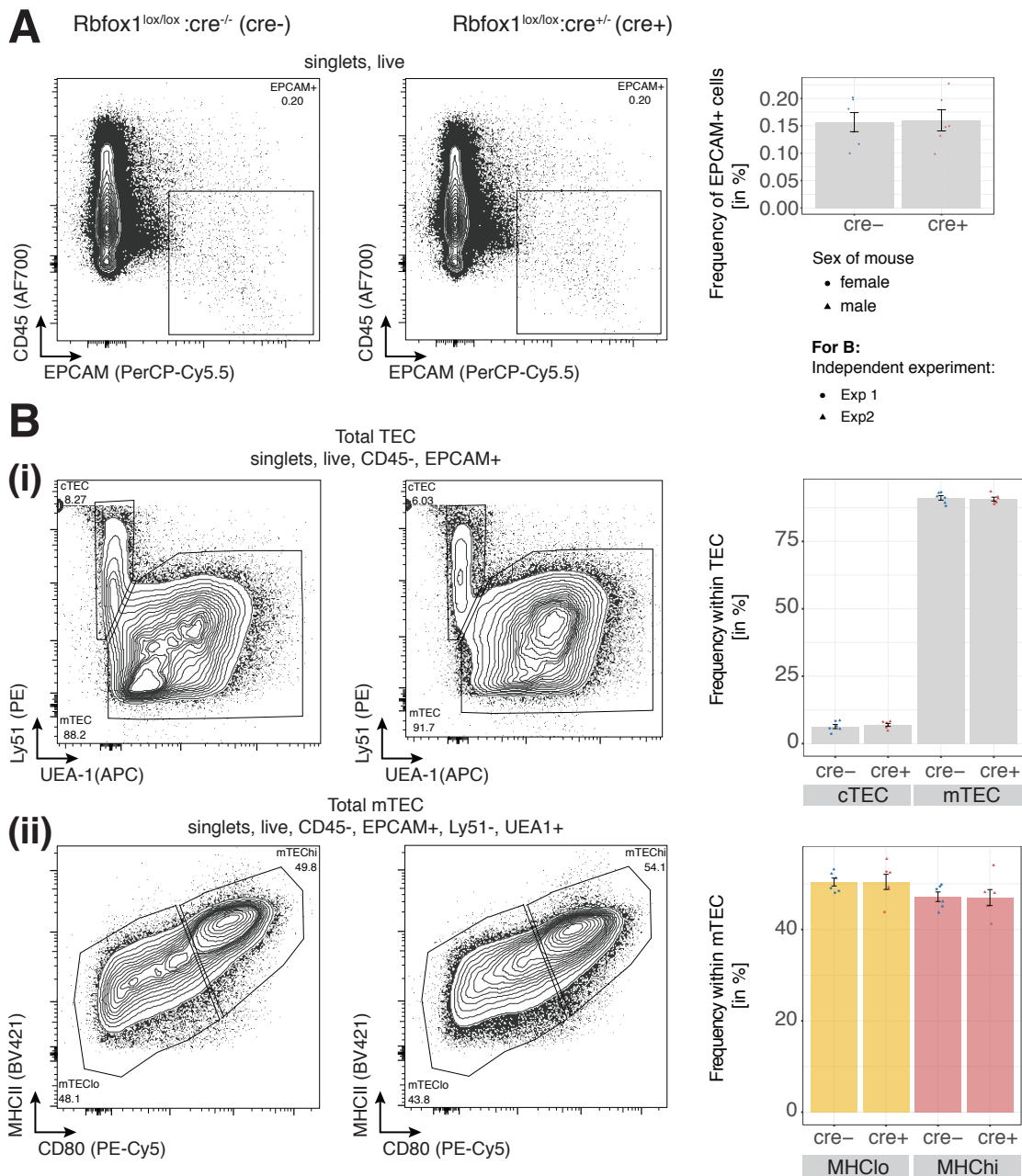




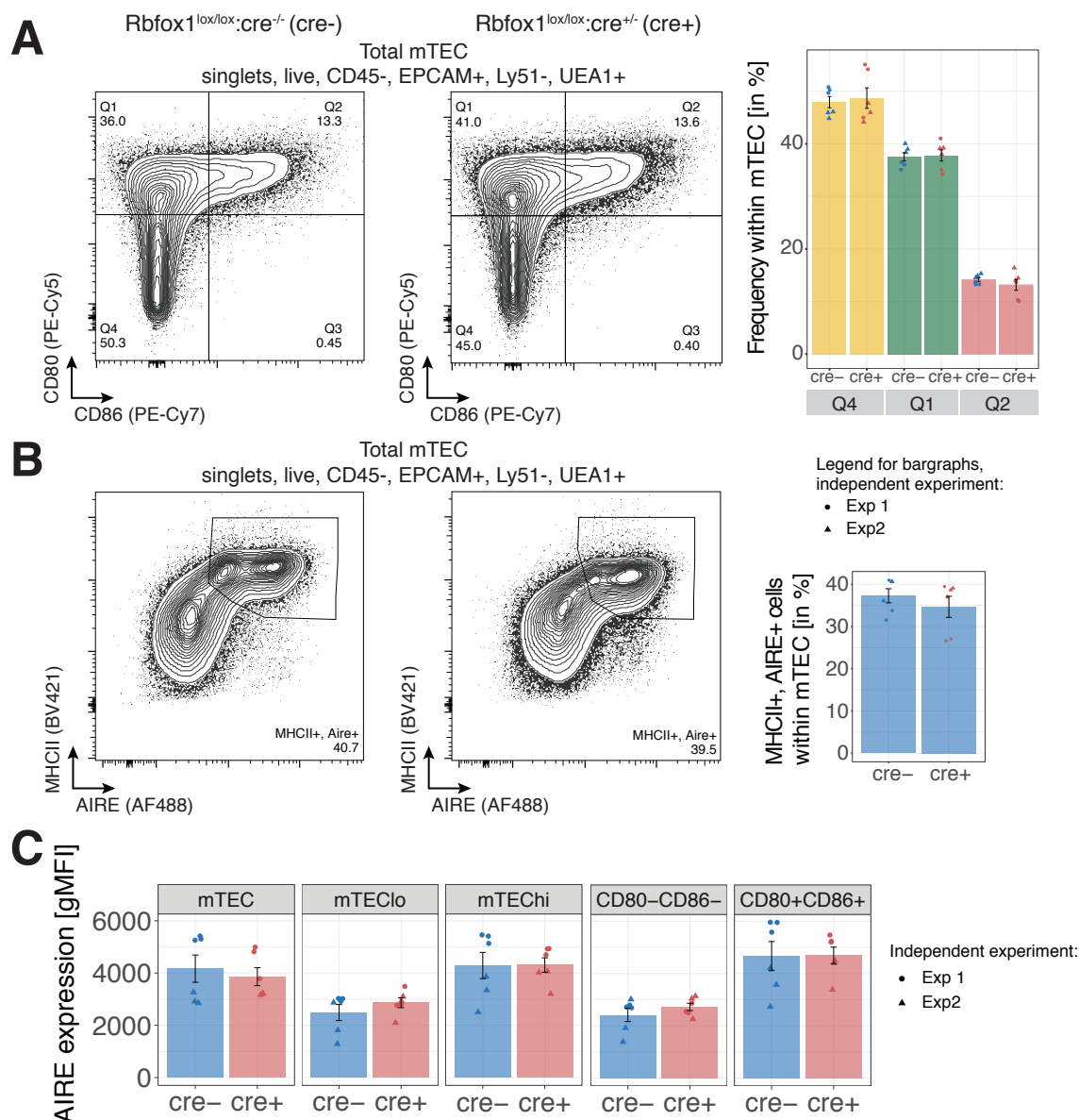
**Figure 6.7:** Assessment of thymus anatomy and cellularity in *Rbfox1*<sup>lox/lox</sup>:*cre*<sup>-/-</sup> (cre-) and *Rbfox1*<sup>lox/lox</sup>:*cre*<sup>+/+</sup> (cre+). **A:** PCR analysis of genomic exons 11 and 12 of *Rbfox1*. Primer sets were described in Figure 6.4A. **B:** Representative image of gross anatomy of the thymus in WT and SKO1 mice. **C:** Total thymus cellularity is shown with symbols indicating the sex of each mouse. **D:** Total thymus cellularity split by sex of the mice. Symbols indicate the independent experiment. Data was generated in two independent experiments with 6 mice in each experiment (4-6 weeks old). The significance between genotypes was tested using a Welch Two Sample t-test (mean ± SE is shown in bar graphs; \* *p*-value < 0.05).

Moreover, the frequencies of mature and immature mTEC defined by the level of MHCII or CD80 cell surface expression and further differentiated by CD86 positivity did not change in the absence of *Rbfox1* expression (Figures 6.8B and 6.9A).

A further analysis of distinct mTEC maturation stages using the expression of CD80 and CD86 revealed comparable AIRE expression in these cells (Figure 6.9B, C) demonstrating that the targeted absence of *Rbfox1* in these cells did not comprise their differentiation.



**Figure 6.8:** Phenotyping of thymic epithelial cells from thymi of *Rbfox1*<sup>lox/lox</sup>:cre<sup>-/-</sup> (cre-) and *Rbfox1*<sup>lox/lox</sup>:cre<sup>+/-</sup> (cre+) mice. **A:** Representative graph demonstrating gating for TEC (EPCAM<sup>+</sup>, CD45<sup>-</sup>) among non-enriched cells isolated from the indicated mouse strains (left) and their frequencies (right). A graph showing the gating for single, live cells is given in Appendix Figure E.1A. **B:** (i) Representative graph demonstrating gating for TEC subtypes from enriched cells (EPCAM<sup>+</sup>, CD45<sup>-</sup>; left) and (ii) subsequent gating for mTEC subtypes defined by MHCII and CD80 status (left). A graph showing the gating for single, live cells is given in Appendix Figure E.1B. Data presented in bar graphs (right) represents the combined results from two independent experiments with six mice in each experiment (4-6 weeks old). The two independent experiments are indicated by different symbols in the bar graphs. No significant differences were observed for the comparison between *Rbfox1*<sup>lox/lox</sup>:cre<sup>-/-</sup> and *Rbfox1*<sup>lox/lox</sup>:cre<sup>+/-</sup> using a Welch Two Sample t-test (mean  $\pm$  SE shown in bar graphs).



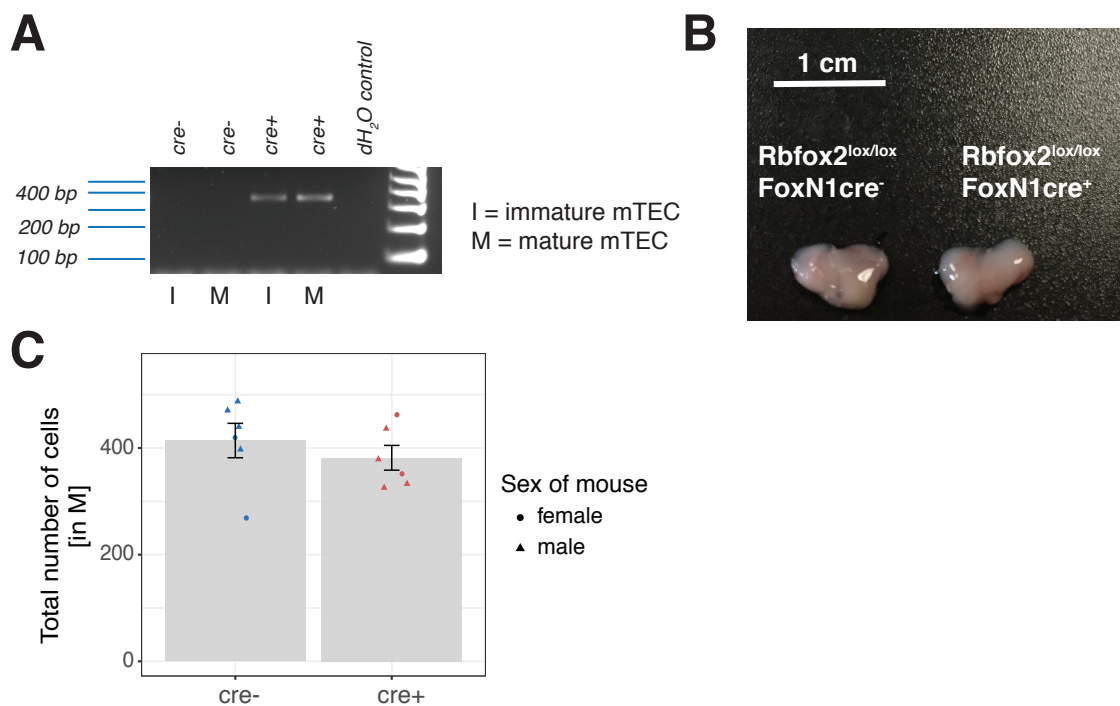
**Figure 6.9:** Analysis of TEC subpopulations and AIRE expression across these in Rbfox1<sup>lox/lox</sup>:cre<sup>-/-</sup> (cre-) and Rbfox1<sup>lox/lox</sup>:cre<sup>+/-</sup> (cre+) mice. **A:** Representative graph demonstrating gating for mTEC subtypes based on CD80 and CD86 surface expression amongst all mTEC (EPCAM<sup>+</sup>, CD45<sup>-</sup>) isolated from the indicated mouse strains (left) and frequencies (right). Medullary mTEC were defined in Figure 6.8B. **B:** Representative graph demonstrating gating for AIRE<sup>+</sup> mTEC amongst all mTEC (EPCAM<sup>+</sup>, CD45<sup>-</sup>, LY51<sup>-</sup>, UEA-1<sup>+</sup>, defined in Figure 6.8B) isolated from the indicated mouse strains (left). Data presented in bar graphs (right) represents the combined results from two independent experiments with six mice in each experiment (4-6 weeks old). **C:** Bar graphs showing AIRE expression (gMFI) across mTEC subpopulations as defined in in Figure 6.8B (ii) and in A. The two independent experiments are indicated by different symbols in the bar graphs. No significant differences were observed for the comparison between Rbfox1<sup>lox/lox</sup>:cre<sup>-/-</sup> and Rbfox1<sup>lox/lox</sup>:cre<sup>+/-</sup> using a Welch Two Sample t-test (mean  $\pm$  SE shown in bar graphs).

## 6.4 TEC-targeted loss of *Rbfox2* expression

*Rbfox2* is the more constitutively expressed member of the *Rbfox* family (Figure 6.1) and is indispensable for brain development [146]. Furthermore, *Rbfox2* was able to partially restore the loss of *Rbfox1*, whereas *Rbfox1* could not restore the developmental brain defects observed in the absence of *Rbfox2*. Thus, mice with a TEC-specific loss of *Rbfox2* expression were generated and analysed. The *Rbfox2*<sup>lox/lox</sup>:cre<sup>+/-</sup> mice are referred to as SKO2, while the *Rbfox2*<sup>lox/lox</sup>:cre<sup>-/-</sup> are referred to as WT.

### 6.4.1 Unaffected thymus cellularity in TEC with a loss of *Rbfox2* expression

The excision of exons 6 and 7 in *Rbfox2* was confirmed by PCR based on genomic DNA from FACS-sorted immature and mature mTEC (Figure 6.10A). The gross anatomical appearance of the thymus including its size was not affected by the loss of *Rbfox2* expression in TEC (Figure 6.10B). There was, however, a small, non-significant decrease in the thymus cellularity in SKO2 mice (382 M in SKO2 versus 414 M in WT, p=0.44 using Welch Two Sample t-test; Figure 6.10B, C).

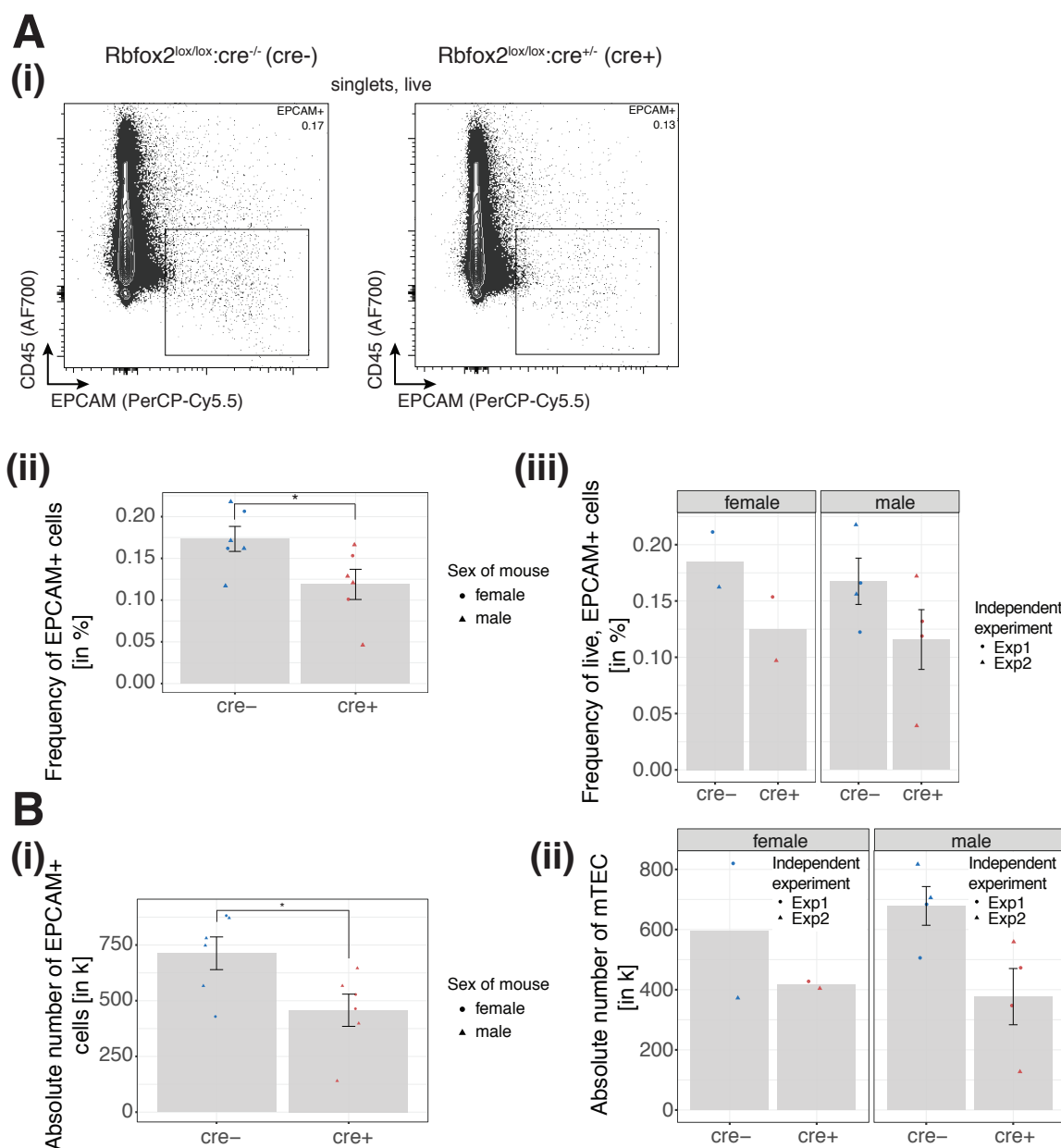


**Figure 6.10:** Assessment of thymus anatomy and cellularity in *Rbfox2*<sup>lox/lox:cre<sup>-/-</sup></sup> (cre-) and *Rbfox2*<sup>lox/lox:cre<sup>+/-</sup></sup> (cre+). **A:** PCR analysis of genomic exons 6 and 7 of *Rbfox2*. Primer locations and product sizes are given in Figure 6.4. **B:** Representative image of gross anatomy of the thymus in *Rbfox2*<sup>lox/lox:cre<sup>-/-</sup></sup> and *Rbfox2*<sup>lox/lox:cre<sup>+/-</sup></sup> mice. **C:** Total thymus cellularity is shown with symbols indicating the sex of each mouse. Data was generated in two independent experiments with 6 mice in each experiment (4-6 weeks old). No significant difference was observed between genotypes using a Welch Two Sample t-test (mean  $\pm$  SE is shown in bar graph).

### 6.4.2 Reduced number of mTEC upon loss of *Rbfox2* expression

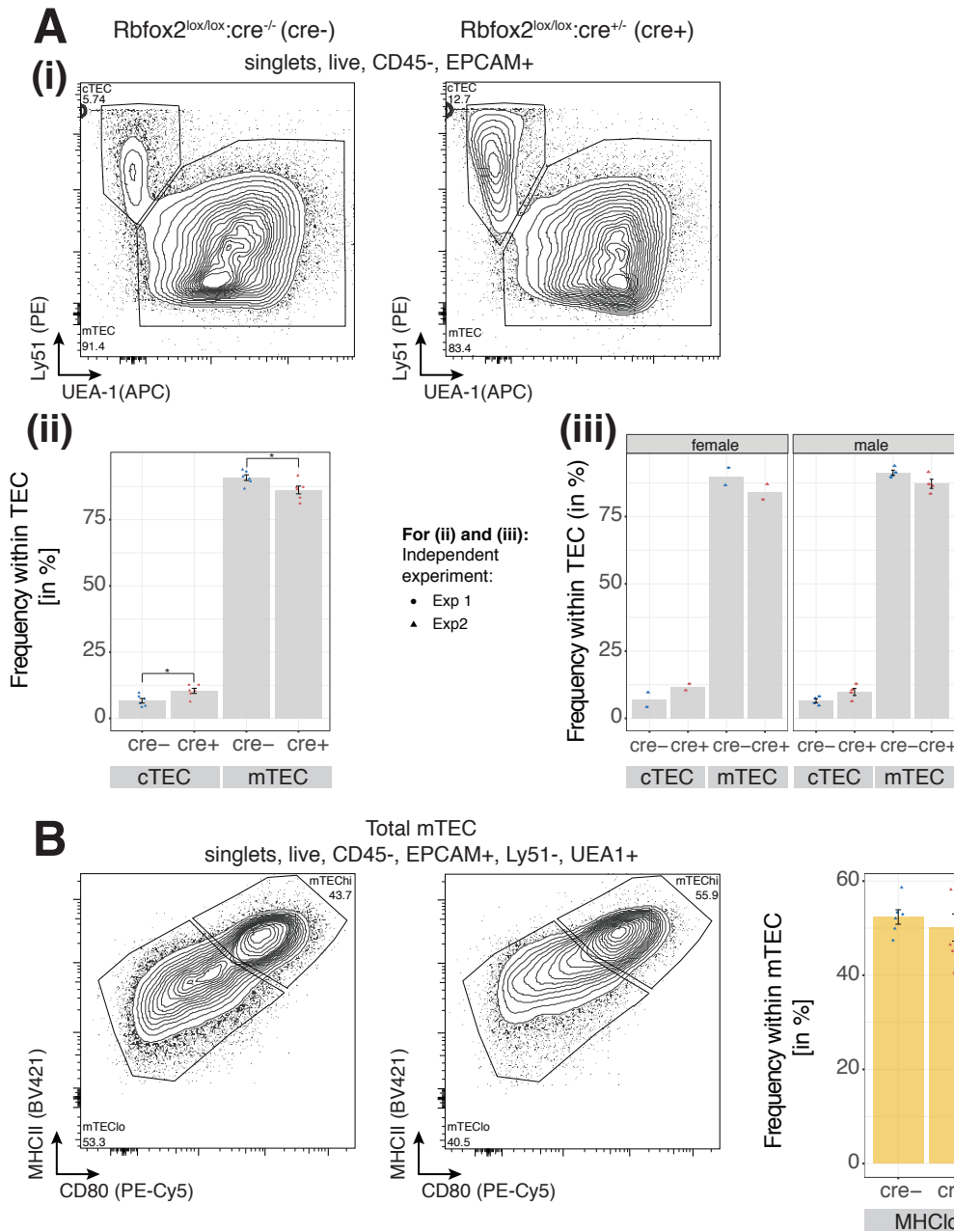
As shown in Figure 6.10C, the total cellularity of the thymus was reduced by 8 % in SKO2 animals ( $p=0.44$ ). A significant reduction in EPCAM<sup>+</sup> cells was observed upon loss of *Rbfox2* in TEC (0.12 % in SKO2 versus 0.17 % in WT,  $p=0.044$  using Welch Two Sample t-test, Figure 6.11A (ii)). Consequently, the total TEC cellularity was lower in SKO2 mice (Figure 6.11B (i)). For both the frequency and absolute number of EPCAM<sup>+</sup> cells, the significant difference could not be confirmed if only male mice were tested (Figure 6.11A (iii), B (ii)). Although the reduction in EPCAM<sup>+</sup> cells was observed for female mice, the number of female mice was insufficient to test for significance. Further experiments using female mice will be required to confirm the reduction in TEC cellularity.

The frequencies of cTEC and mTEC revealed a difference between the cTEC and mTEC lineage with a significant increase of the former and reduction of the latter cell type in SKO2 mice (Figure 6.12A). Similar to the observation for the total TEC cellularity, the difference in cTEC and mTEC frequencies was not significant anymore when testing between male mice only, respectively ( $p=0.09$  for both cTEC and mTEC using Welch Two Sample t-test, for mTEC: 87.2 % in SKO2 versus 91.3 % in WT; Figure 6.12A (iii)). The mTEC subpopulations as defined by MHCII, CD80 and CD86 expression did, however, not show differences in frequencies when comparing WT and SKO2 mice (Figure 6.12B, Figure 6.13A), implying that mTEC maturation was not defective upon loss of *Rbfox2*. The expression level of AIRE in total mTEC and mTEC subpopulations was not affected by the loss of *Rbfox2* (Figure 6.13B, C). In addition, the frequency of AIRE-positive mTEC and the AIRE expression in total mTEC or mTEC subpopulations was unchanged (Figures 6.13B).



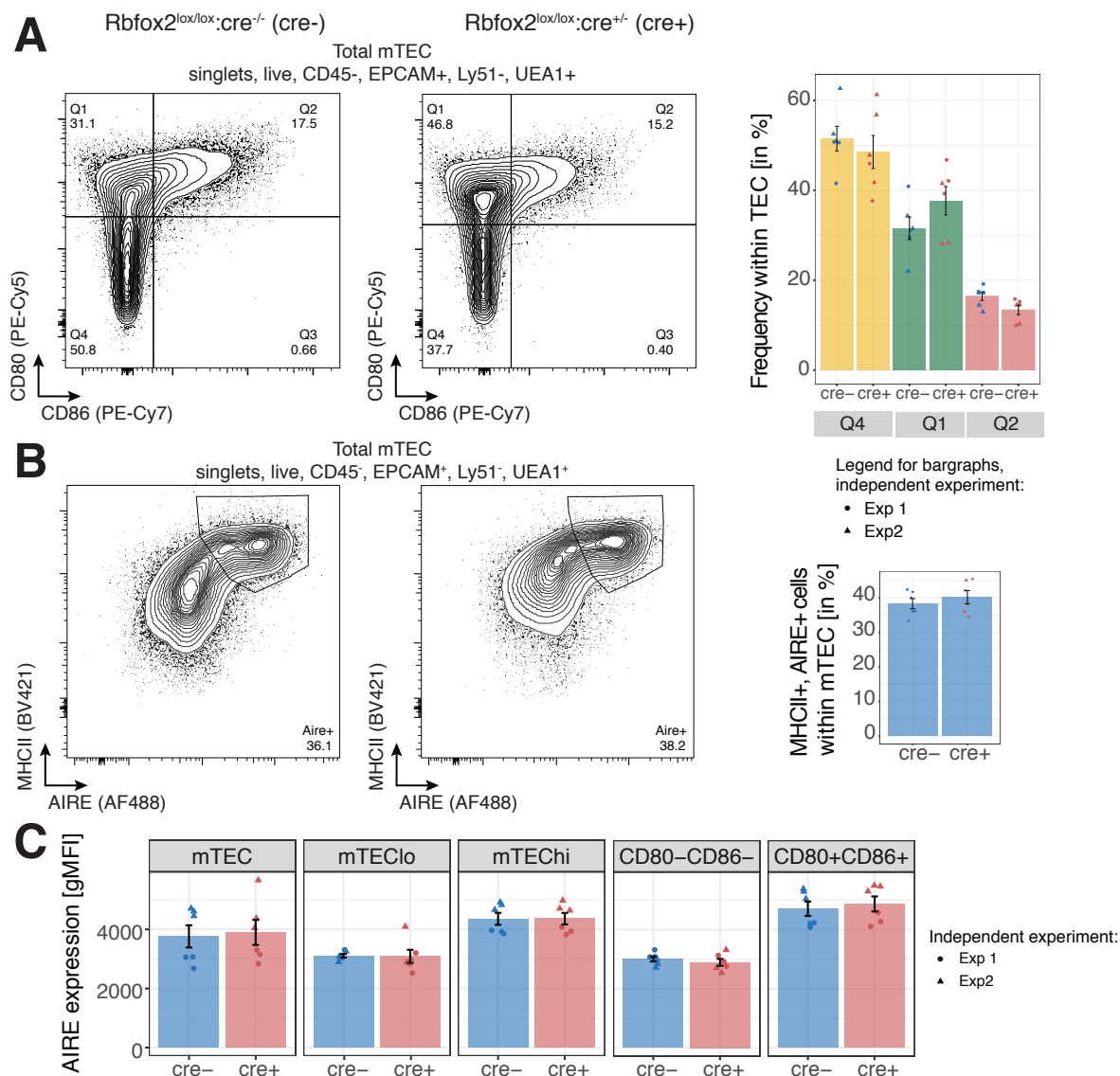
**Figure 6.11:** Phenotyping of thymic epithelial cells from thymi of  $Rbfox2^{lox/lox}:cre^{-/-}$  (cre-) compared to  $Rbfox2^{lox/lox}:cre^{+/-}$  (cre+) mice. **A:** Representative graph demonstrating gating for TEC (EPCAM<sup>+</sup>, CD45<sup>-</sup>) among non-enriched cells isolated from the indicated mouse strains (i) and their frequencies (ii). The frequency was further split by sex of the mice with the independent experiment indicated with different symbols (iii). A Welch Two Sample t-test between the results for the male mice did not reveal a significant difference ( $p=0.17$ ). The female results could not be tested for significance due to the lack of replicates. A graph showing the gating for single, live cells is given in Appendix Figure E.1A. **B:** (i) Absolute number of EPCAM<sup>+</sup> cells combined from both independent experiments ( $p=0.03$ ). The sex of the mice is indicated by shape. Similar to A, the results for the male mice did not reveal a significant difference ( $p=0.07$ ) and the female results could not be tested for significance due to the lack of replicates. Data presented was generated in two independent experiments and combined for statistical analysis where indicated (each experiment  $n=6$  mice; 4-6 weeks old; mean  $\pm$  SE is shown in bar graphs; \*  $p$ -value < 0.05 for Welch Two Sample t-test).





**Figure 6.12:** Analysis of thymic epithelial cell subtypes in thymi from Rbfox2<sup>lox/lox</sup>:cre<sup>-/-</sup> (cre-) compared to Rbfox2<sup>lox/lox</sup>:cre<sup>+/-</sup> (cre+). **A:** (i) Representative graph demonstrating gating for TEC subtypes from enriched cells (EPCAM<sup>+</sup>, CD45<sup>-</sup>) and their frequencies (ii). The frequency was further split by sex of the mice with the independent experiment indicated with different symbols (iii). A Welch Two Sample t-test between the results for the male mice did not reveal a significant difference ( $p=0.09$  for cTEC and mTEC). The female results could not be tested for significance due to the lack of replicates. A graph showing the gating for single, live cells is given in Appendix Figure E.1B. **B:** The subsequent gating for mTEC subtypes defined by MHCII and CD80 status (left) and their frequencies (right) are shown. Data presented was generated in two independent experiments and combined for statistical analysis as indicated (each experiment  $n=6$  mice; 4-6 weeks old; mean  $\pm$  SE is shown in bar graphs; \*  $p$ -value  $< 0.05$  for Welch Two Sample t-test). If no significance is indicated, the Welch Two Sample t-test did not reveal a significant difference.





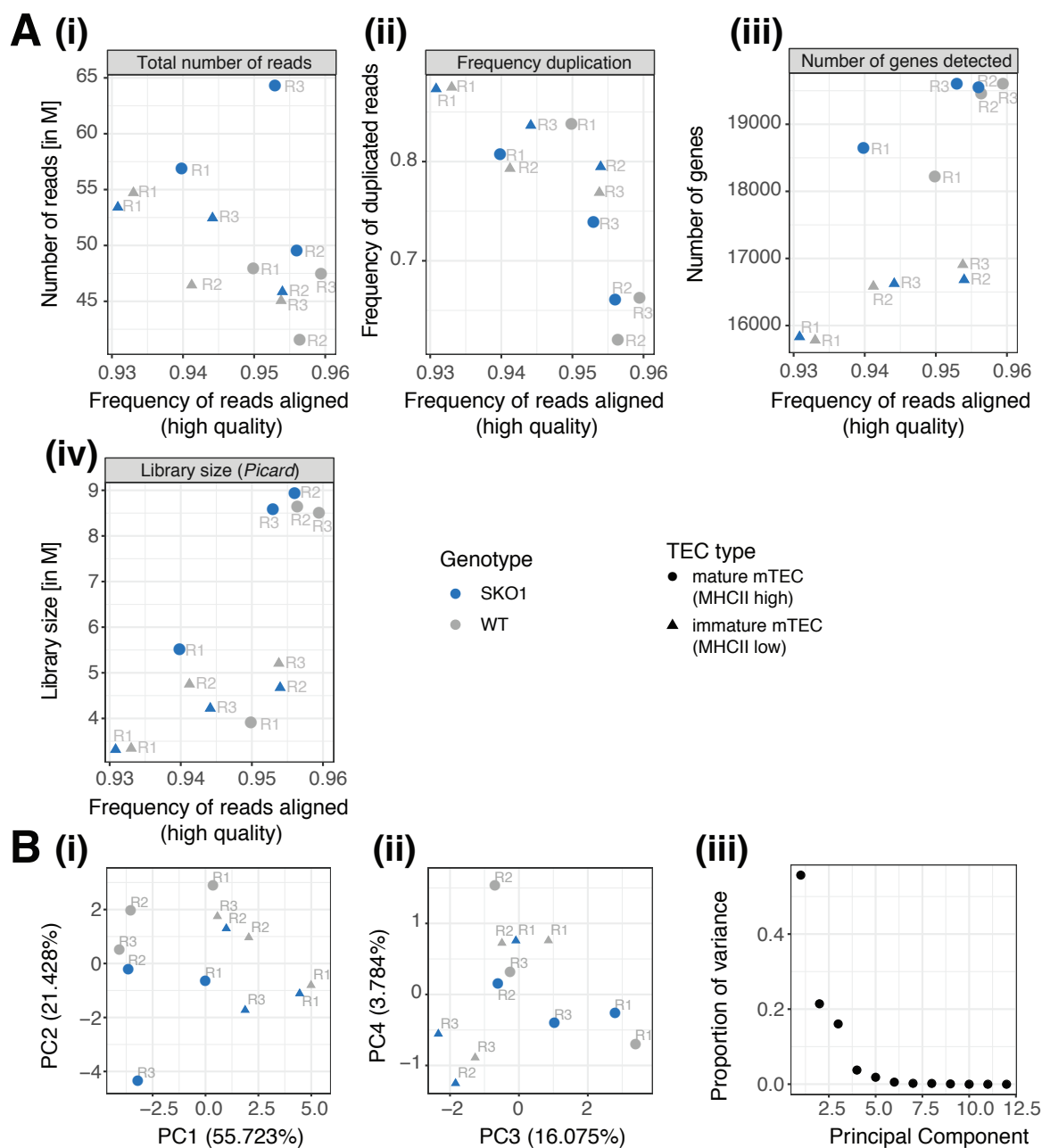
**Figure 6.13:** Analysis of TEC subpopulations and AIRE expression across these in Rbfox2<sup>lox/lox</sup>:cre<sup>-/-</sup> (cre-) and Rbfox2<sup>lox/lox</sup>:cre<sup>+/-</sup> (cre+) mice. **A:** Representative graph demonstrating gating for mTEC subtypes based on CD80 and CD86 surface expression amongst all mTEC (EPCAM<sup>+</sup>, CD45<sup>-</sup>, LY51<sup>-</sup>, UEA-1<sup>+</sup>) isolated from the indicated mouse strains (left) and frequencies (right). The two independent experiments are indicated by different symbols in the bar graph. **B:** Representative graph demonstrating gating for AIRE<sup>+</sup> mTEC amongst all mTEC (EPCAM<sup>+</sup>, CD45<sup>-</sup>, LY51<sup>-</sup>, UEA-1<sup>+</sup>) isolated from the indicated mouse strains (left). Data presented in bar graphs (right) represents the combined results from two independent experiments with six mice in each experiment (4-6 weeks old). **C:** Bar graphs showing AIRE expression (gMFI) across mTEC subpopulations as defined in Figure 6.12A (i), B and in A. The two independent experiments are indicated by different symbols in the bar graphs. No significant differences were observed for the comparison between Rbfox2<sup>lox/lox</sup>:cre<sup>-/-</sup> and Rbfox2<sup>lox/lox</sup>:cre<sup>+/-</sup> using a Welch Two Sample t-test (mean  $\pm$  SE shown in bar graphs).

## 6.5 Transcriptome analysis of mTEC with a loss of *Rbfox1* or *Rbfox2*

To globally assess changes in gene expression as well as with changes in splicing in mTEC with a loss of *Rbfox1* or *Rbfox2*, stranded, paired-end RNA-sequencing was generated from low input samples each comprised of approximately 10k cells. Immature mTEC with low cell surface expression of MHCII and mature mTEC with high cell surface expression of MHCII were sorted and RNA purified. The sequenced samples for SKO1 and SKO2 mice comprised of gender-matched samples with two female and one male from each genotype (n=12 samples per genotype; Appendix Table E.1). The library preparation and sequencing was performed at the Oxford Genomics Centre (Wellcome Trust Center for Human Genetics), yielding on average 24.5 M stranded, 150 bp read pairs per sample (Appendix Table E.1).

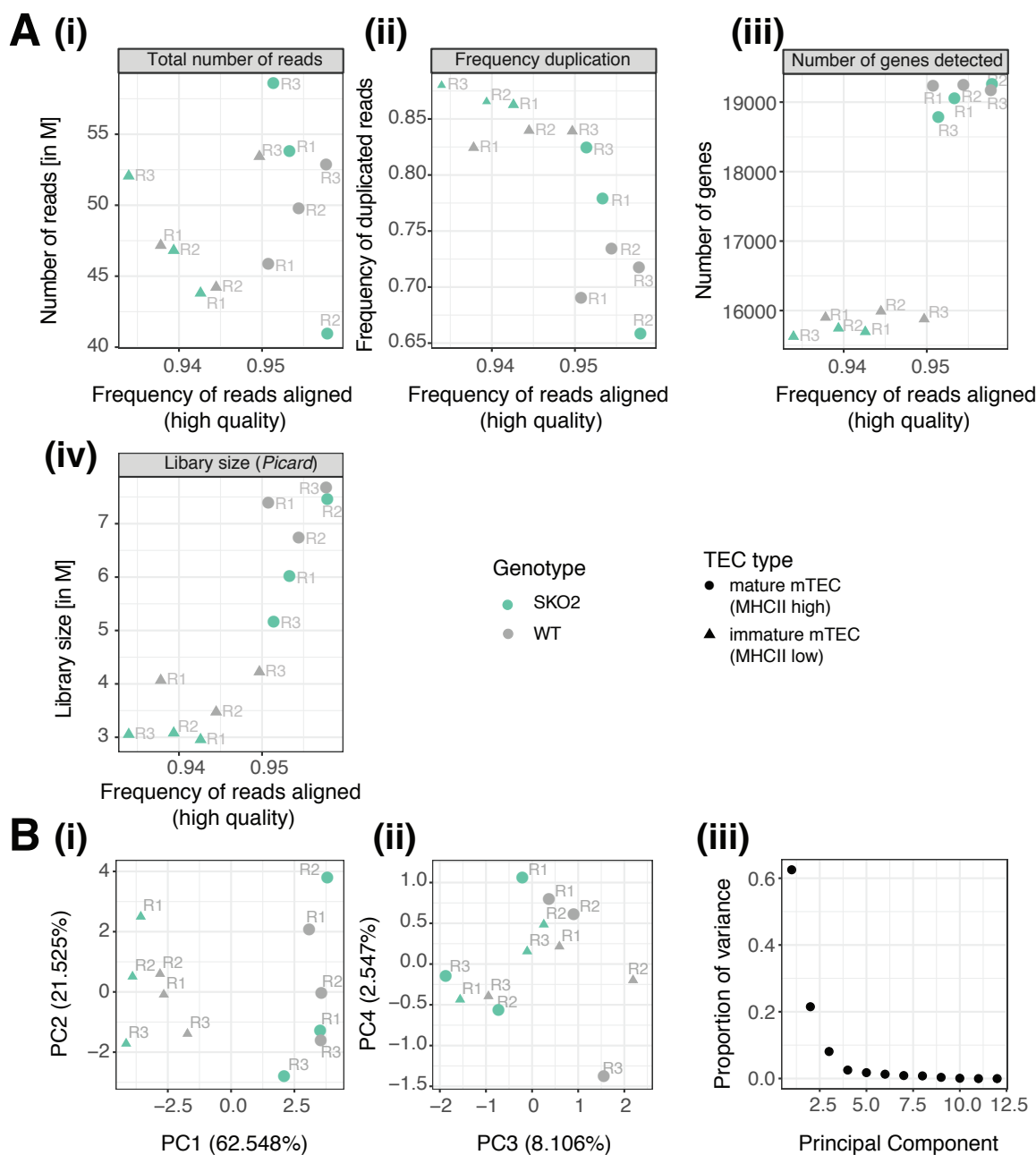
Initially, the read quality was assessed and the reads mapped to the mouse genome using *HISAT2* (mm10). Both the read and post-mapping quality assessment revealed a relatively high level of duplication in all samples (Figures 6.14A (ii) and 6.15A (ii)). The duplication rate ranged from 60 % to almost 90 %. As a consequence, the library complexities were smaller than expected with fewer than 10 M unique DNA fragments (defined by *EstimateLibraryComplexity* from *Picard*) being detected per sample (Figures 6.14A (iv) and 6.15A (iv)). In both datasets, a clear difference in number of protein-coding genes detected was observed for immature and mature mTEC as expected from the biology of these cells. However, in the SKO1 dataset two replicates from immature populations, namely replicate 1 for SKO1 and WT, had a lower number of genes detected compared to all other replicates (Figure 6.14A (iii)). Both immature mTEC as well as one mature mTEC sample from replicate 1 was further amongst the samples with the lowest library size (Figure 6.14A (iv)).

To identify samples with lower quality, a Principal Component Analysis (PCA) was performed on all quality metrics reported by the *scseq* pipeline (Figures 6.14B and 6.15B). The scree plots revealed that most variation was explained by PC1 to PC4, hence these



**Figure 6.14:** Quality assessment of RNA-sequencing data for immature and mature mTEC from  $Rbfox1^{lox/lox};cre^{-/-}$  (WT) and  $Rbfox1^{lox/lox};cre^{+/-}$  mice (SKO1). **A:** Selected quality metrics plotted against the frequency of reads aligned with high quality. In (i) to (iv), the total number of reads, the frequency duplicated (from *Picard*), the number of protein-coding genes detected (by *Salmon*) and the library size (from *Picard*) are plotted, respectively. The genotype of mice is indicated by color, while the TEC subpopulation is indicated by symbols. The replicate number is given in each plot. **B:** Principal Component Analysis (PCA) of quality metrics calculated by the *scseq* pipeline. In (i) and (ii), Principal Component 1 (PC1) versus PC2 and PC3 versus PC4 are shown, respectively. The loadings for PC1 to PC4 are given in Appendix Table E.2. (iii) Scree plot of the PCA showing the variance explained by each PC. The legend for all panels is given in A (iv) (right).

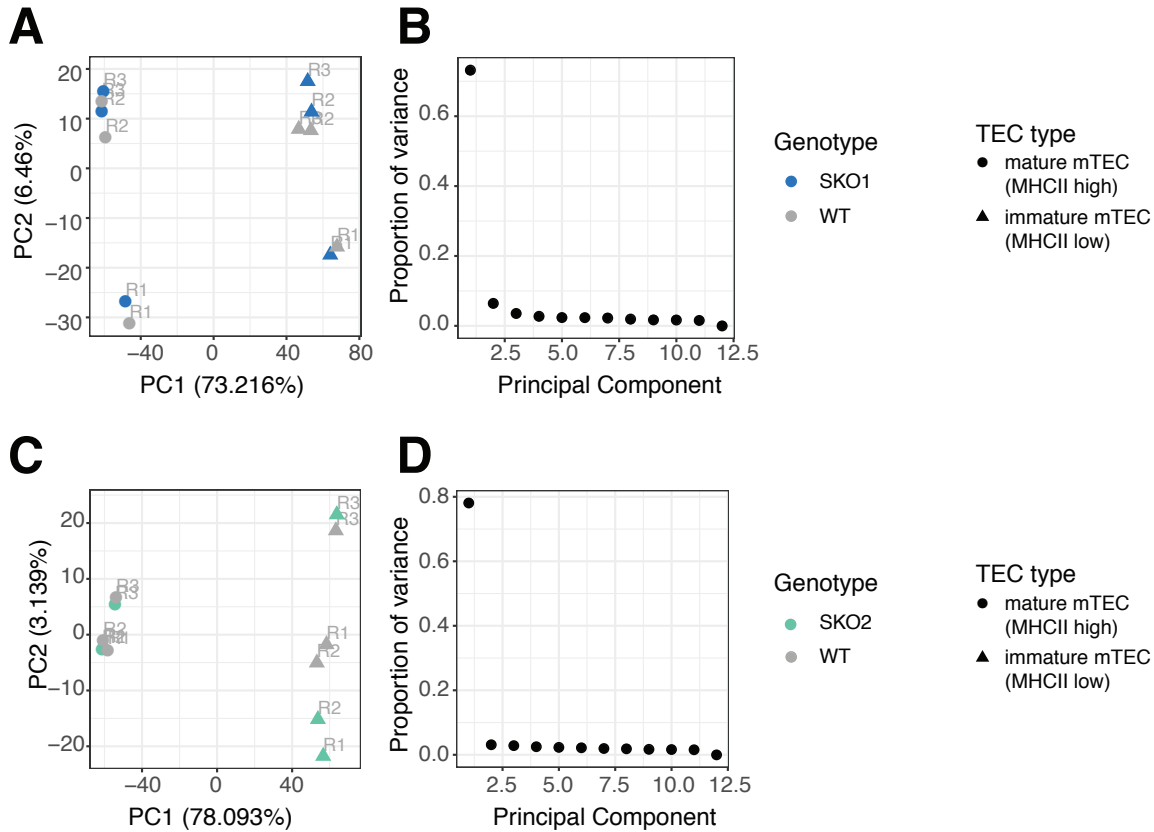
were plotted individually and the top loadings for these PCs are given in Appendix Tables E.2 and E.3. The separation along PC1 was mainly between immature and mature mTEC, which represented a separation based on the number of genes detected. One sample, namely mature mTEC from SKO1 replicate 3, was separated along PC2. Examining the loadings for this PC suggested that the high number of total reads in this sample was driving PC2. To further investigate if replicate 1 from the SKO1 dataset should be excluded, the examination of the PCA revealed that the samples from this replicate were also separated on PC1 (immature samples) and PC3 (mature samples), respectively (Figure 6.14B). These were due to the lower number of genes expressed in the immature mTEC samples from replicate 1 (along PC1) and the low median insert size of both samples from replicate 1 (along PC3). For the samples from SKO2, the quality assessment did not reveal any outliers, but a slightly lower number of total reads (40.8-58.4 M) than the SKO1 dataset (Figure 6.15A (i)). Similar to the SKO1 dataset, the duplication rate was very high ranging from 66 to 88 % (Figure 6.15A (ii)). Consistent with the SKO1 dataset, PC1 mainly described the number of genes detected in each sample, which clearly separated the immature from mature mTEC samples (Figure 6.15B). Two WT samples were disjoint from other samples along PC3, which was due to the higher median insert size in these samples. Given the observations in the SKO1 dataset above, all samples from the SKO1 and SKO2 dataset were subsampled to a common read depth for all downstream analyses (40 M reads).



**Figure 6.15:** Quality assessment of RNA-sequencing data for immature and mature mTEC from  $Rbfox2^{lox/lox:cre^{-/-}}$  (WT) and  $Rbfox2^{lox/lox:cre^{+/-}}$  mice (SKO2). **A** In (i) to (iv), selected quality metrics plotted against the frequency of reads aligned with high quality. The total number of reads, the frequency duplicated (from *Picard*), the number of protein-coding genes detected (by *Salmon*) and the library size (from *Picard*) are plotted, respectively. The genotype of mice is indicated by color, while the TEC subpopulation is indicated by the symbols. The replicate number is given in each plot. **B:** Principal Component Analysis (PCA) of quality metrics calculated by the *scseq* pipeline. In (i) and (ii), Principal Component 1 (PC1) versus PC2 and PC3 versus PC4 are shown, respectively. The loadings for PC1 to PC4 are given in Appendix Table E.3. (iii) Scree plot of the PCA showing the variance explained by each PC. The legend for all panels is given in A (iv) (right).

### 6.5.1 Small number of differentially expressed genes upon loss of *Rbfox1* and *Rbfox2* in mature mTEC

To examine the effect of the loss of *Rbfox1* or *Rbfox2* on gene expression, first a PCA of gene expression values was performed (Figure 6.16). The gene expression differed most between immature and mature mTEC from each single-knockout model, which explained 81 % or 78 % of the variance in the SKO1 and SKO2 dataset, respectively. No clear distinction between mTEC samples with loss of *Rbfox1* or *Rbfox2* were observed in the PCAs.

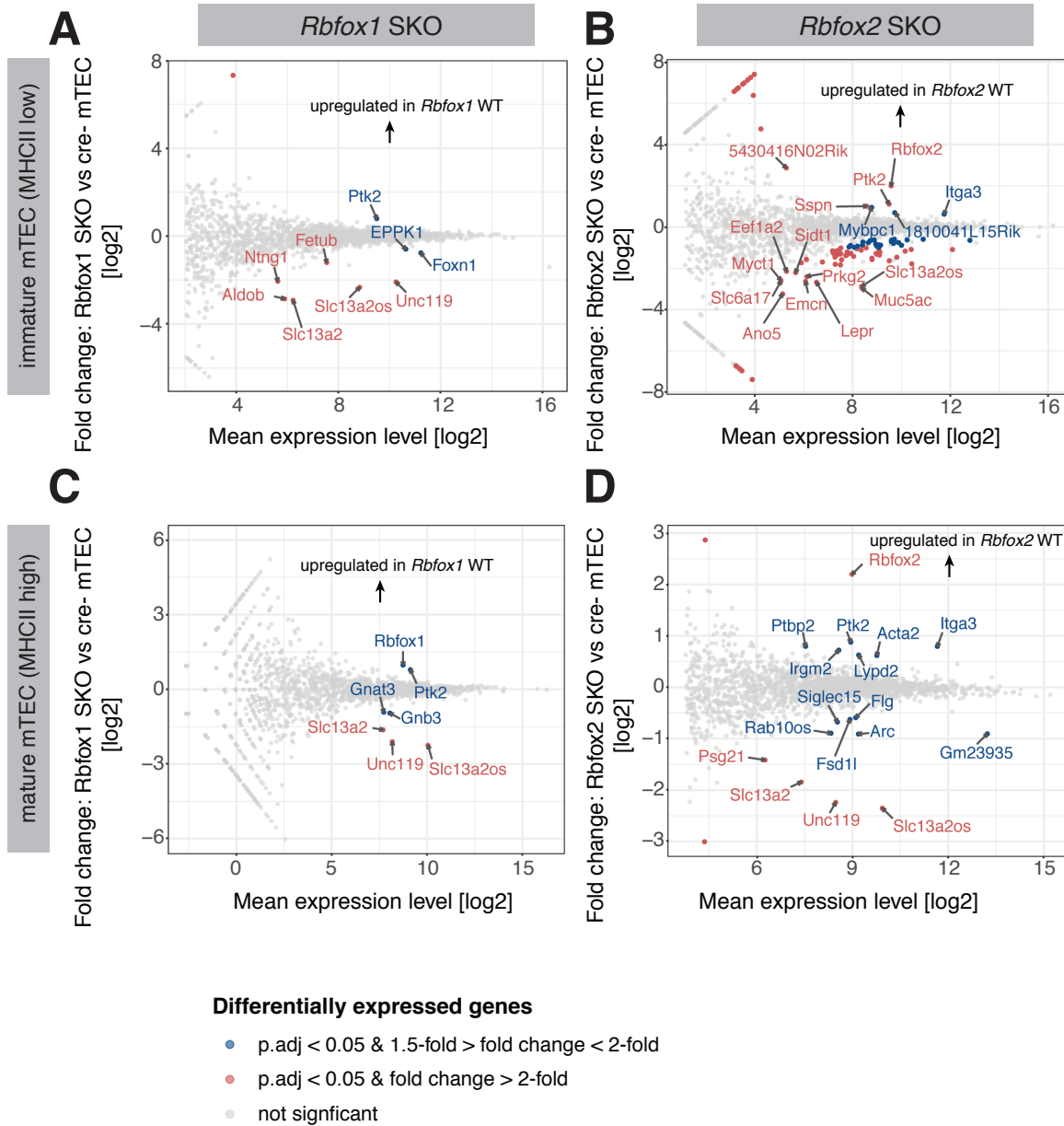


**Figure 6.16:** Principal Component Analysis using gene expression data for immature and mature mTEC from both single-knockout models. **A:** PC1 and PC2 of the PCA based on gene expression for *Rbfox1*<sup>lox/lox</sup>:cre<sup>-/-</sup> (WT) and *Rbfox1*<sup>lox/lox</sup>:cre<sup>+/-</sup> mice (SKO1). **B:** Scree plot for PCA shown in A. **C:** PC1 and PC2 of the PCA based on gene expression for *Rbfox2*<sup>lox/lox</sup>:cre<sup>-/-</sup> (WT) and *Rbfox2*<sup>lox/lox</sup>:cre<sup>+/-</sup> mice (SKO2). **D:** Scree plot for PCA shown in C. The color indicates the genotype of the samples, while the symbols represent the mTEC subpopulation. The PCA was based on variance-stabilised counts above the 25<sup>th</sup> quantile detected in at least 3 samples.

In addition to the PCA, differentially expressed genes between each of the single-knockout mice and the corresponding cre<sup>-</sup> samples were identified using *DESeq2* (Figure 6.17, Appendix Tables E.4 and E.5). A higher number of genes was differentially expressed between immature single-knockout versus wildtype mTEC samples than between mature mTEC samples for both SKO1 and SKO2 datasets. In Table 6.1, the differentially expressed genes were characterised by their tissue-restriction ( $\tau > 0.7$ ) and whether they were categorised as a splicing factor (Section 5.3) or as an *Aire*-regulated gene (Appendix Figures B.5 and B.6). The number of differentially expressed splicing factors was low with only *Rbfox1* downregulated in mature mTEC from SKO1 mice. In SKO2 mice, *Rbfox2* was also downregulated in both mTEC types. Additionally the splicing factor *Ptbp2* was downregulated in mature mTEC from SKO2 mice. In the *Nestin*-cre model, the authors observed a 60 % increase in RBFOX2 expression upon loss of *Rbfox1*. In the single-knockout models generated here, the expression of the other *Rbfox* family members was not strongly affected in either SKO1 or SKO2 mice. One TRSF, namely *Rbm20*, showed upregulation in immature mTEC from SKO2 mice. As shown in Section 5.3.2, *Rbm20* was described as a TRSF acting in muscle tissues including the heart.

In the SKO1 dataset, genes such as *Aldob* were upregulated in immature mTEC with a loss of *Rbfox1* (Figure 6.17A). *Aldob* encodes for Aldolase B (ALDOB), which is an enzyme involved in glycolysis (reviewed by Chang *et al.* [248]). In humans it is specific to the kidney and liver, while other members of the aldolase family are found in other tissues.

Amongst the significantly downregulated genes in immature mTEC from SKO2 mice was *Rbfox2* itself (Figure 6.17B), suggesting that the knockout decreased the gene expression of *Rbfox2*. Two examples of downregulated genes were *Sarcospan* (*Sspn*) and *Integrin $\alpha$ 3* (*Itga3*). In skeletal muscle *Sspn* was shown to be part of the dystrophin–glycoprotein complex, which was suggested to be involved in connecting the extracellular matrix with the intracellular actin cytoskeleton in muscle cells (reviewed by Barresi & Campbell [249]). *Itga3* is part of the integrin family and involved in cell adhesion and associated signalling pathways (reviewed by Moreno-Layseca *et al.* [250]).



**Figure 6.17:** Differentially expressed genes between *Rbfox1*<sup>lox/lox:cre<sup>±</sup></sup> (SKO1) or *Rbfox2*<sup>lox/lox:cre<sup>±</sup></sup> (SKO2) and the corresponding cre<sup>-/-</sup> mice (WT). **A** and **B**: Differentially expressed genes between immature mTEC in SKO1 and SKO2 compared to WT immature mTEC, respectively. **C** and **D**: Differentially expressed genes between mature mTEC in SKO1 and SKO2 compared to WT mature mTEC, respectively. Counts for all genes were obtained using *featureCounts* and subsequently differentially expressed genes were identified using *DESeq2*. Significantly (Benjamini-Hochberg adjusted p-value < 0.05) and genes with a fold change of > 1.5 or > 2 are colored. In all panels, only 2,000 non-significantly differently expressed genes were randomly selected and are shown in the plot. The top 10 differentially expressed genes in each direction with more than 25 counts were labelled.



A gene upregulated in immature and mature mTEC with a loss of either *Rbfox1* or *Rbfox2* was *Slc13a2* (significant in all comparisons except immature mTEC from the SKO2 dataset, where Benjamini-Hochberg adjusted p-value = 0.053, Figure 6.17). In addition, *Slc6a17* was upregulated in immature mTEC from SKO2 mice (Figure 6.17B). The *Slc6* family comprises of neurotransmitter transporters [251], which regulate the uptake of neurotransmitters into neurons, while sodium-coupled transport is regulated by members of the *Slc13* gene family [252]. *Slc13a2* is a member specifically expressed in the kidney and small intestine. In all four comparisons another gene named *Slc13a2os* was downregulated, which overlaps *Slc13a2* on the opposite strand. The upregulation of *Slc13a2* and *Slc13a2os* might be due to the *FoxN1*-cre construct used. The gene *Unc119* encoding for retinal photoreceptors was also upregulated in mature mTEC samples from both SKO1 and SKO2 mice and immature samples from SKO2 mice. Inspecting the genomic locus of the genes mentioned above revealed that *Slc13a2* and *Slc13a2os* are located directly downstream of *FoxN1*, while *Unc119* is located directly upstream of *FoxN1* on chromosome 11.

A commonly downregulated gene in immature mTEC and mature mTEC from SKO1 and SKO2 mice was *Ptk2*. The *FoxN1*-cre mouse strain used here, is known to have a disruption of the *Ptk2* gene due to integration of the PAC construct containing *FoxN1*-cre. Although the integration would only affect one allele of *Ptk2*, the downregulation observed

**Table 6.1:** Summary of differentially expressed genes in mature and immature with targeted loss of *Rbfox1* or *Rbfox2*. The columns indicate the total number of differentially expressed genes, the number of tissue-restricted antigens (TRA,  $\tau > 0.7$ ), splicing factors (SF) and *Aire*-regulated genes (*Aire*). All genes were included in the analysis, but only protein-coding TRA, SF and *Aire*-regulated genes were counted. Genes can occur in multiple categories.

|                    | <i>Rbfox1</i> SKO |     |    |      | <i>Rbfox2</i> SKO |     |    |      |
|--------------------|-------------------|-----|----|------|-------------------|-----|----|------|
|                    | Total             | TRA | SF | Aire | Total             | TRA | SF | Aire |
| immature, SKO up   | 8                 | 5   | 0  | 2    | 82                | 26  | 1  | 47   |
| immature, SKO down | 2                 | 1   | 0  | 1    | 17                | 9   | 1  | 7    |
| mature, SKO up     | 5                 | 3   | 0  | 0    | 11                | 5   | 0  | 1    |
| mature, SKO down   | 2                 | 1   | 1  | 0    | 8                 | 1   | 2  | 2    |

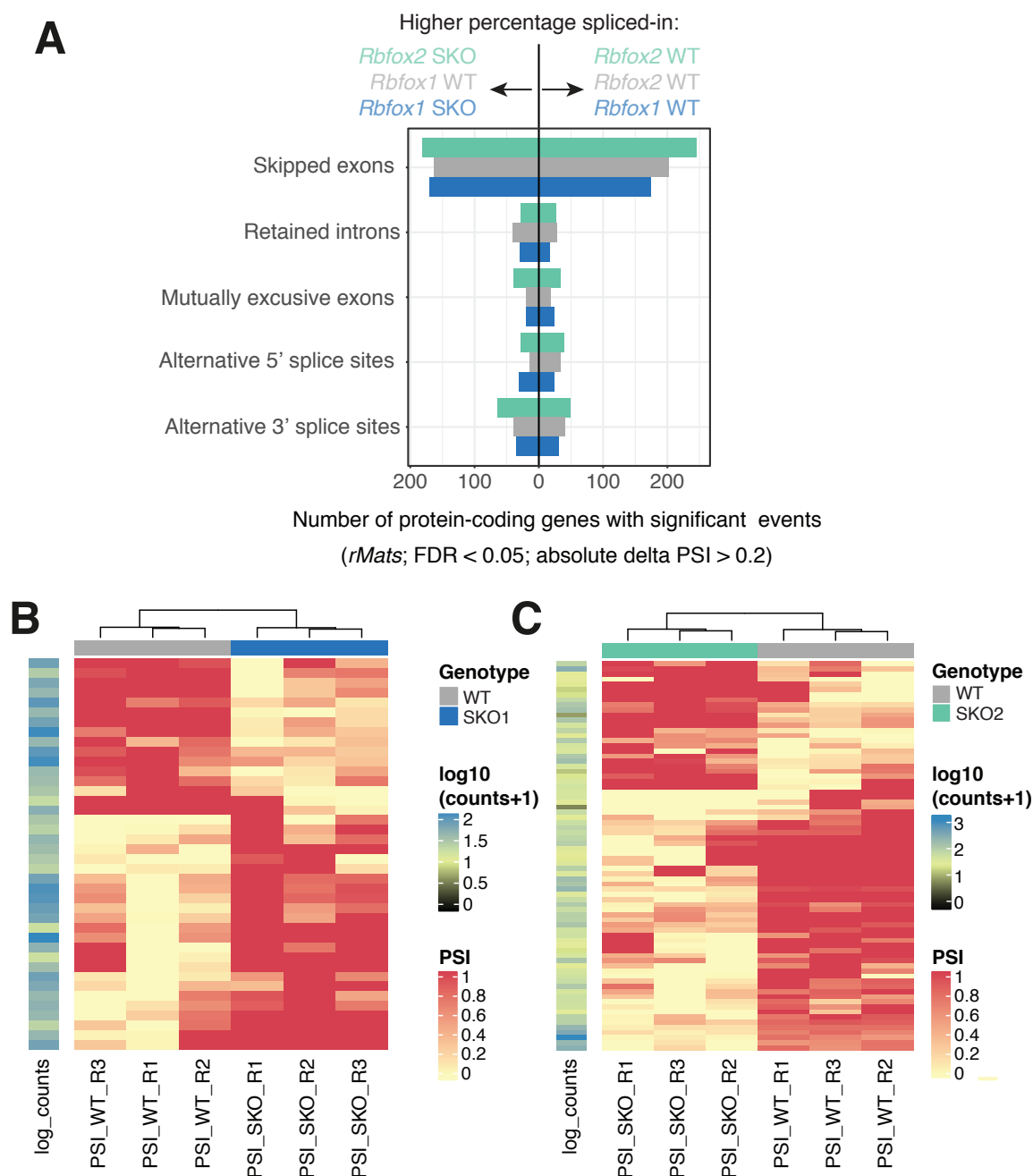
here was likely a consequence of the integration rather than biologically meaningful.

The number of differentially expressed genes in mature mTEC upon loss of *Rbfox1* was the smallest in all comparisons performed here. Amongst two upregulated genes in mature mTEC from SKO1 mice was *FoxN1*, which might be identified as upregulated due the reads originating from the *FoxN1*-cre construct that are mapping to the endogenous *FoxN1* locus. Besides *Rbfox1*, the only other upregulated genes in the comparison of mature mTEC from SKO1 compared to their wildtype counterparts were *Gnb3* and *Gnat3*. *Gnb3* was recently reported in the context of thymic tuft cells, where *Gnb3* was upregulated in thymic tuft cells compared to intestinal tuft cells based on single-cell RNA-sequencing data [19]. The intestinal tuft cells were sorted and sequenced based on the expression of tdTomato under the control of the *Hpgds* promoter. Overall, few genes were differentially expressed in mature mTEC of the SKO1 and SKO2 mice compared to the respective cre<sup>-</sup> littermates. However, given that splicing factors were knocked out in these models, a large difference in gene expression might not be expected as splicing differences might not be reflected at the gene level.

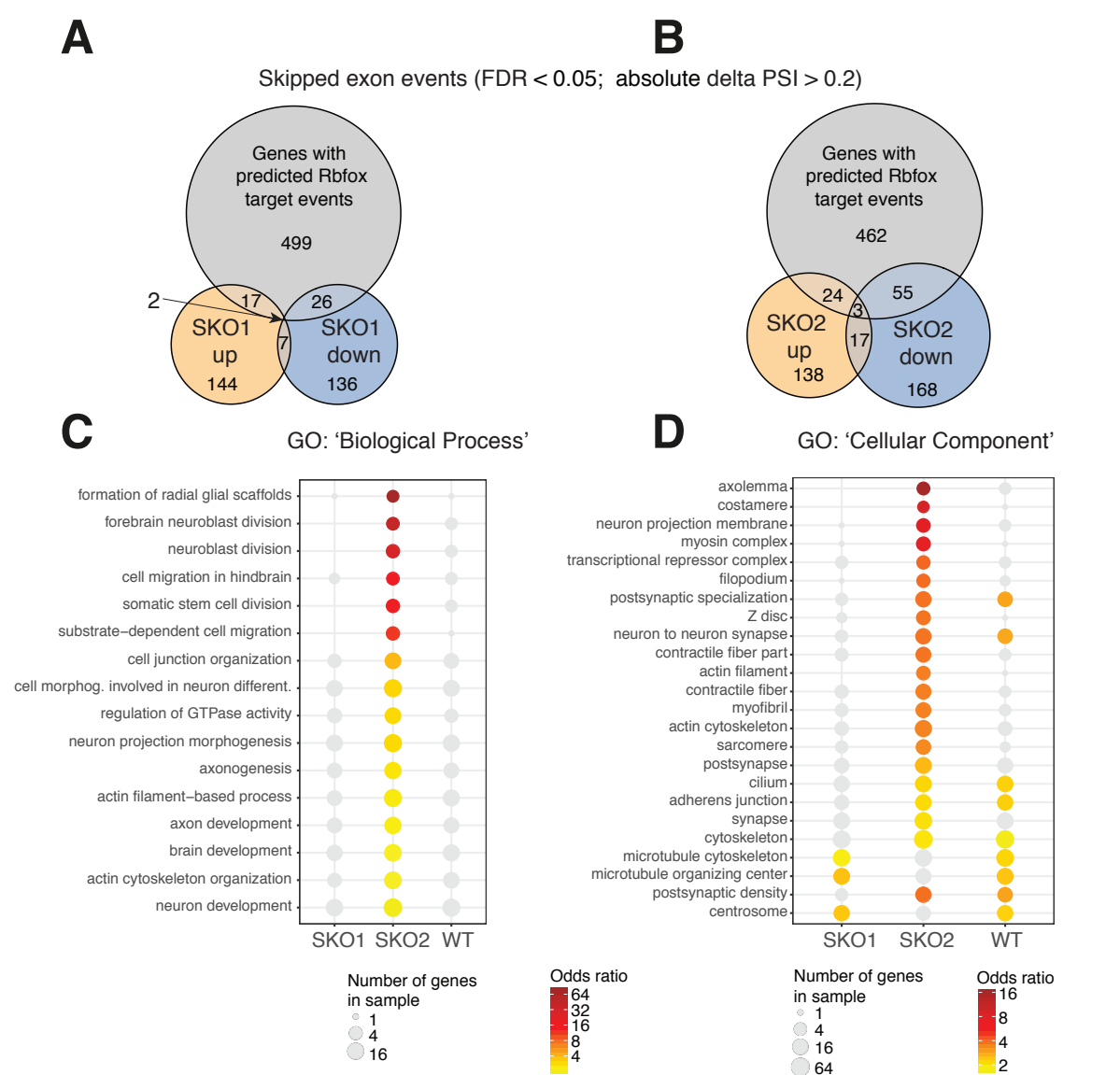
### 6.5.2 Assessing the effects of *Rbfox1* or *Rbfox2* loss on splicing in mature mTEC

To identify differentially spliced events between mTEC with a loss of either *Rbfox1* or *Rbfox2* and the respective cre<sup>-</sup> samples (WT), a comparison in mature mTEC using *rMATS* was performed (Figure 6.18A). The number of genes with significantly different skipped exon (SE) events was highest in the SKO2 versus WT comparison (Figure 6.18A). The number of genes with significantly higher SE inclusion was balanced between mature mTEC with a loss of *Rbfox1* and mature mTEC from wildtype counterparts.

The number of differentially spliced SE events in the WT comparison was lower than the SKO2 comparison, but higher than the number obtained from the SKO1 comparison (Figure 6.18A). This observation suggests a high level of biological variation in splicing in mTEC, which is expected as these cells are a heterogeneous cell type with respect to



**Figure 6.18:** Differential splicing events between mature mTEC from *Rbfox1*<sup>lox/lox</sup>:cre<sup>+/-</sup> (SKO1) or *Rbfox2*<sup>lox/lox</sup>:cre<sup>+/-</sup> (SKO2) and the corresponding cre<sup>-/-</sup> mice (WT) or from a comparison between both cre<sup>-/-</sup> samples. **A:** Number of differential splicing events determined by *rMATS* and filtered for events in protein-coding genes with FDR < 0.05 and a difference in percentage spliced-in (absolute delta PSI) > 0.2. **B** and **C:** Heatmaps of PSI values for skipped exon events in each replicate from the SKO1 and SKO2 datasets, respectively. The skipped exon events were filtered for a minimum absolute delta PSI of 0.5 (n=40 events for SKO1 and n=76 events for SKO2 datasets). The left column indicates the combined counts for inclusion and exclusion of the respective event as determined by *rMATS*.



**Figure 6.19:** Assessment of genes affected by differentially spliced events in mature mTEC with a loss of *Rbfox1* (SKO1) or *Rbfox2* (SKO2) or mature mTEC from both *cre*<sup>-/-</sup> (WT) samples. **A** and **B**: Venn diagrams showing the overlap between protein-coding genes with differential skipped exon events (Figure 6.18A) and predicted Rbfox targets genes [73] for SKO1 and SKO2 comparisons, respectively. The differential skipped exon events from Figure 6.18A were split into upregulated or downregulated events with respect to the single-knockout samples. Only genes with at least 0.5 TPM in the SKO sample were included. **C** and **D**: Geneset over-enrichment analysis for genes containing differentially spliced events in mature mTEC for the comparisons shown in Figure 6.18A. In C and D, the over-enrichment within Gene Ontology (GO) categories 'Biological Process' and 'Cellular Component' are shown, respectively. SKO1 corresponds to the comparison between mature mTEC with a loss of *Rbfox1* and WT littermates, while SKO2 shows the genes containing events between *Rbfox2* and WT mice. WT refers to the comparison of the *cre*<sup>-/-</sup> littermates from both datasets. In C and D, genesets were filtered for adjusted p-value < 0.05 (Benjamini-Hochberg correction), minimum two genes from a geneset in the sample and odds ratio > 2. Insignificant differences are colored in grey and representative categories from each comparison are shown.

gene expression. As a result, a high number of cells might have to be sampled to observe consistent splicing patterns, especially if the effect size (the impact of *Rbfox1/Rbfox2*) on splicing in these cells is relatively small. More importantly, the observation of few splicing alterations could be explained by the known redundancy of RBFOX1 and RBFOX2 [145, 146]. In previous studies few alternatively spliced events were observed upon the loss of *Rbfox1* or *Rbfox2* in neurons [145, 146]. For other splicing events such as mutually exclusive exons the number of significant events with higher inclusion were comparable between both single-knockouts and the wildtype counterparts (Figure 6.18A). For alternative 5' splice sites fewer events with higher inclusion were observed in mature mTEC with a loss of *Rbfox1* compared to mature mTEC from wildtype littermates.

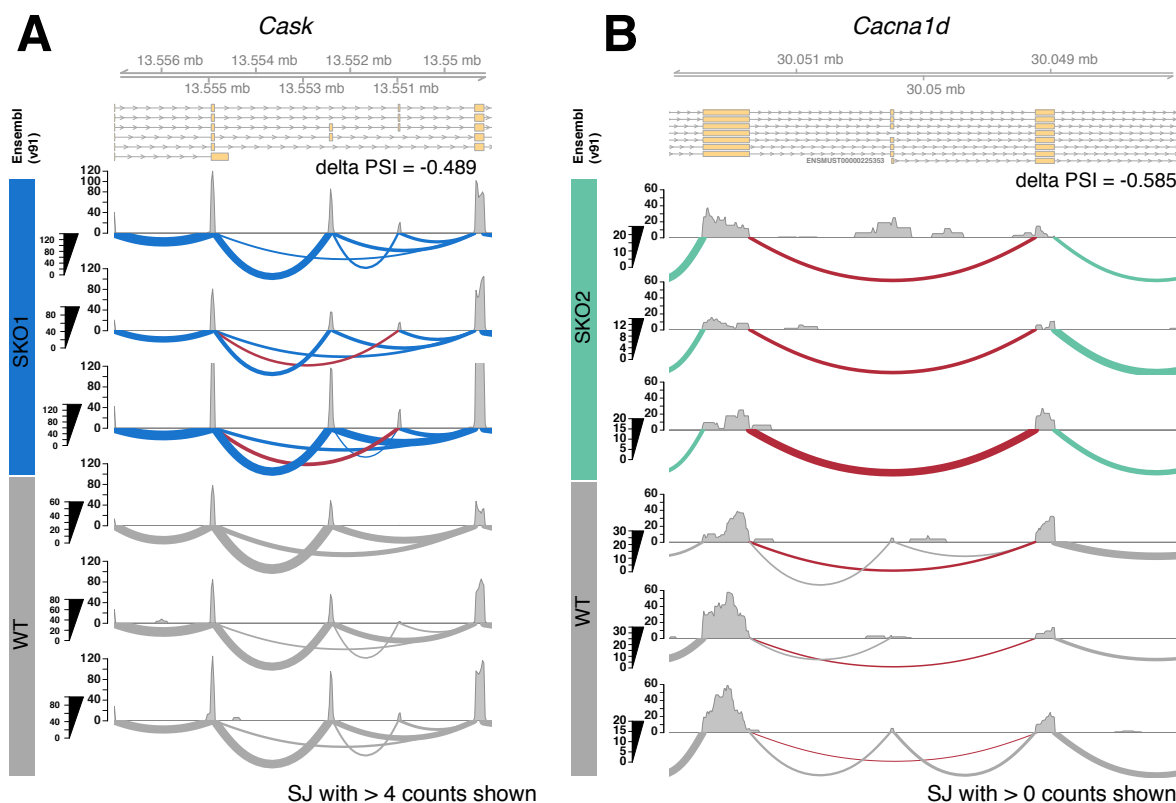
To further examine, how the PSI differed between the single-knockout and corresponding *cre*<sup>-</sup> samples, a heatmap of events with the highest delta PSI values was generated (absolute delta PSI > 0.5, Figure 6.18B, C, Appendix Figure E.3). The heatmaps revealed a more consistent pattern of inclusion and exclusion between the replicates of the SKO1 and SKO2 samples than observed for the WT samples. However, for some events inconsistencies between the replicates were observed. For the WT and SKO1 comparison, only 38 and 40 events had an absolute delta PSI > 0.5, respectively. Approximately twice as many events with absolute delta PSI > 0.5 were observed for the comparison of mature mTEC from SKO2 mice compared to corresponding *cre*<sup>-</sup> littermates (n=76 events).

Despite the limitations with respect to read depth and input cell numbers discussed above, the genes containing SE events from Figure 6.18A were compared to predicted *Rbfox* target events [73] to assess, whether *Rbfox* targets were enriched amongst genes with differentially spliced events in mature mTEC (Figure 6.19A, B). Some overlap was found between genes affected by predicted *Rbfox* splicing and genes affected by differential splicing in mature mTEC of SKO1 or SKO2 mice. To assess whether the genes containing differentially spliced (irrespective of direction) were over-enriched in genes containing predicted *Rbfox* targets, a Fisher Exact tests were performed using a list of expressed genes defined to contain SE events as the background (Section 2.1.6.3). For

genes differentially spliced in SKO1, an odds ratio of 3.8 and a significant over-enrichment was observed ( $p < 3e-12$ ). The over-enrichment of genes with differentially spliced events in the SKO2 dataset within the *Rbfox* target geneset revealed an even higher odds ratio of 6.7 ( $p < 2.2e-16$ ). In the *Rbfox* splicing network by Weyn-Vanhentenryck *et al.*, *Rbfox* was predicted to regulate a larger number of inclusion events than exclusion events (for a mixture of SE, tandem cassette exons and mutually exclusive exons) [73]. In the analysis performed here, a larger number of SE events showed lower inclusion levels in mature mTEC from SKO2 mice compared to the WT samples, which would be consistent with the trend predicted by Weyn-Vanhentenryck *et al.*.

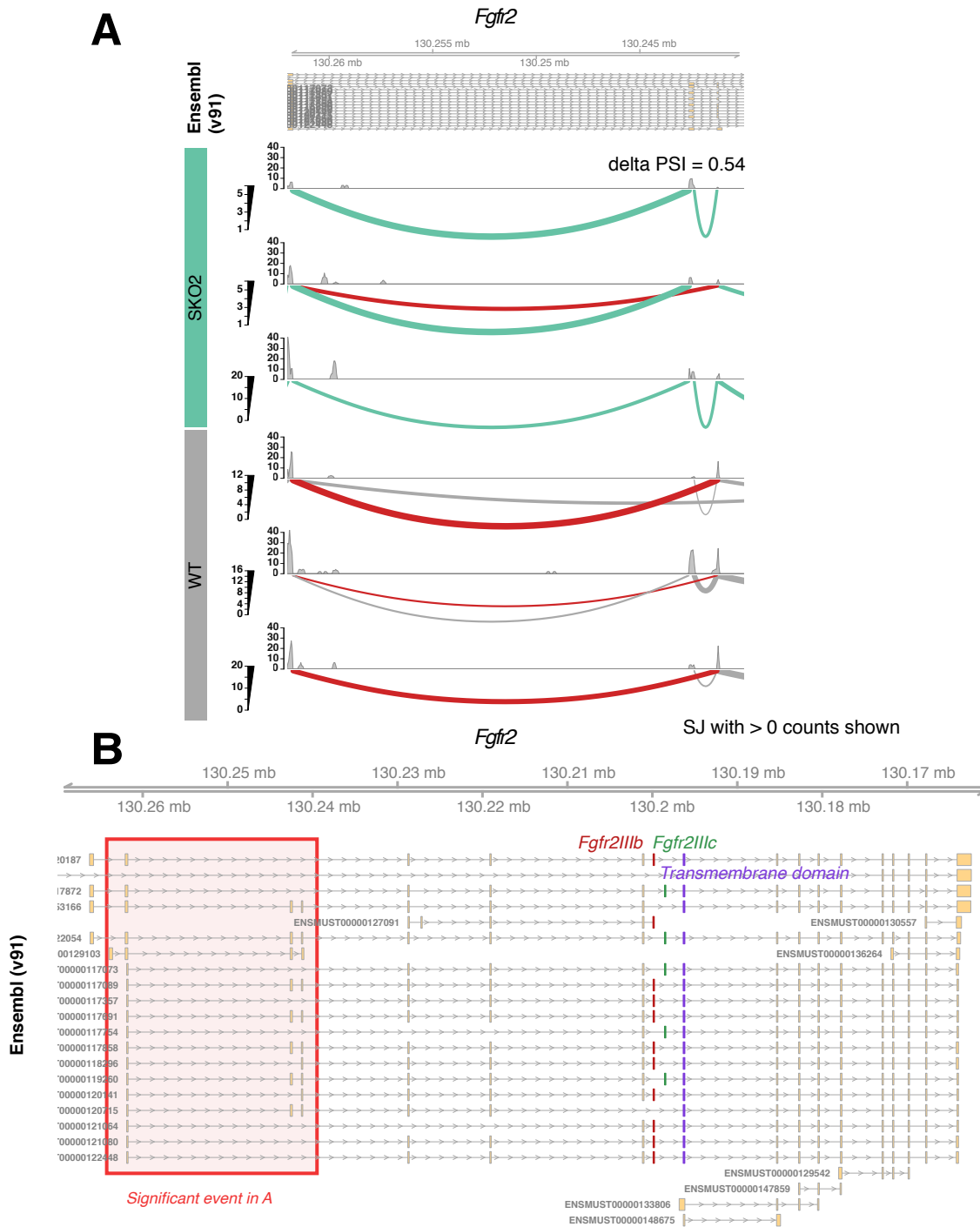
To investigate the types of genes affected by differential splicing in SKO1 and SKO2 samples, a geneset over-representation analysis was performed. The SKO1 and WT comparisons revealed no over-enriched genesets in the Gene Ontology (GO) categories classified as 'Biological Process' after filtering on adjusted p-value (Benjamini-Hochberg correction, Figure 6.19C). In the GO categories classified as 'Cellular Component' only few categories such as 'microtubule organizing center' or 'centrosome' were significantly over-enriched in these comparisons (Figure 6.19C). The category 'synapse', which is compatible with the types of gene known to be affected by RBFOX1 splicing activity, was over-enriched with an odds ratio of 1.7 (unadjusted p-value = 0.01, Figure 6.19D). The geneset over-enrichment for the events affected by the loss of *Rbfox2* contained categories related to muscle function and development ('myosin complex', 'contractile fiber') and to neuronal development ('brain development', 'neuroblast division'), which are both processes affected by the *Rbfox* family. While some of these categories such as 'neuron development' also had a high number of genes affected in the SKO1 and WT comparison, they were not significantly over-enriched in these comparisons. However, two categories related to neuronal activity ('postsynaptic specialisation', 'neuron to neuron synapse') were significantly over-enriched in the WT comparison.

To examine whether the differentially splicing events observed in mature mTEC were consistent with events observed in previous *Rbfox1* or *Rbfox2* conditional knockout



**Figure 6.20:** Examples of Rbfox target genes with differentially spliced events in mature mTEC with a loss of *Rbfox1* (SKO1) or *Rbfox2* (SKO2). **A** and **B**: *Cask* and *Cacna1d* as examples with differentially spliced events in SKO1 and SKO2 mice, respectively. *Cacna1d* showed differentially spliced exons in *Rbfox1:Nestin-cre* and *Rbfox2:Nestin-cre* models using Exon Junction Arrays [145, 146]. However, these events showed different directions in the two models. Sashimi plots were generated from RNA-sequencing data of all replicates and only splice junction reads with at least 5 counts are shown. The delta percentage spliced-in (PSI) as determined by *rMATS* is shown. The splice junctions skipping the respective exon are indicated in red.

mouse models, such as the *Nestin-cre* driven model [145, 146], the genes containing such differentially spliced events were compared. Gehman *et al.* had observed a small number of genes harbouring Rbfox-regulated splicing events in these *Nestin-cre* targeted models using junction microarrays (validated by RT-PCR). Moreover, only a small number of 11 overlapping genes with splicing events showing the same direction of splicing (for instance exon inclusion) was observed between the mouse models with a loss *Rbfox1* or *Rbfox2* [145, 146]. In SKO1 mice, five genes affected by differential splicing were found to be in common with these previously published models. A differentially spliced exon in the synapse-associated gene *Cask* showed skipping of an exon in SKO1 mice, which was not observed in the corresponding *cre*<sup>-</sup> littermates (Figure 6.20A). For SKO2 mice, only one



**Figure 6.21:** Differential splicing event in *Fgfr2* in mature mTEC with a loss of *Rbfox2* (SKO2). **A:** Sashimi plots to show coverage and splice junctions in each replicate of mature mTEC from SKO2 mice and respective cre<sup>-</sup> mice. Sashimi plots were generated from RNA-sequencing data of each replicate and all splice junction counts are included. The delta percentage spliced-in (PSI) as determined by *rMATS* is shown. The splice junctions skipping the respective exon are indicated in red. **B:** Genomic locus of the *Fgfr2* gene with exons highlighted, which define *Fgfr2IIIb* and *Fgfr2IIIc*, where the former was shown as an important signalling molecule in TEC development [253]. The exons were assigned to isoforms and domains according to Hajihosseini *et al.* [254].



gene, *Cacna1d*, with differentially spliced exons was in common with the validated genes from the *Nestin*-cre driven loss of *Rbfox2* (Figure 6.20B). A limitation of the comparison between events identified here and previously by Gehman *et al.* was the absence of information on genomic coordinates of the events, thus matching between events was performed based on the gene containing the respective event.

Another gene containing an event with high delta PSI between mature mTEC from SKO2 mice and wildtype counterparts was *Fgfr2* (Figure 6.21). *Fgfr2* encodes for fibroblast growth factor receptor 2 (FGFR2) and has an established role in TEC development, where the isoform *Fgfr2IIIb* was shown to be important for proliferation of TEC during thymus development [253]. The alternatively spliced exon identified here was located in the extracellular domain of FGFR2, which is responsible for ligand binding. A differential splicing event in the extracellular domain could therefore have an effect on ligand binding affinity and potentially signalling through the receptor.

Although the RNA-sequencing dataset obtained from the SKO1 and SKO2 models appeared to be of limited suitability for differential splicing analysis due to the low library size, a number of examples with differentially spliced events in genes with known Rbfox targets were found. Furthermore, the identified events were enriched in categories with known relevance to Rbfox splicing targets.

## 6.6 Breeding of the conditional double-knockout mice

Initially, a TEC-targeted loss of both *Rbfox1* and *Rbfox2* was the aim. However, I experienced a number of complications resulting in significant delays. First, the received breeding pair was infertile. As a consequence, the mating scheme was adjusted to mating  $Rbfox1^{lox/WT}:Rbfox2^{lox/WT}:cre^{+/-}$  with  $Rbfox1^{lox/WT}:Rbfox2^{lox/WT}:cre^{-/-}$ , resulting in a significant extension of breeding time to achieve  $cre^{+/-}$  double-knockout mice. In addition, issues occurred with the heritability of the  $cre^{+}$  allele together with the  $Rbfox2^{lox}$  allele, where matings including mice with the genotype  $Rbfox1^{lox/lox}:Rbfox2^{lox/WT}:cre^{+/-}$  did not result in  $Rbfox1^{lox/lox}:Rbfox2^{lox/lox}:cre^{+/-}$  offspring. Due to these issues breeding of these animals and subsequent experiments could not be completed prior to submission of this thesis.

## 6.7 Discussion

A TEC-targeted knockout of a TRSF was generated to examine its impact on mTEC development and self-antigen splicing in mTEC. *Rbfox1* was computationally identified as a suitable candidate for functional follow-up experiments, because it showed robust and frequent expression at the single-cell level, and specific expression in populations of mature mTEC but not in cTEC or skin epithelial cells (Section 5.4). In this chapter, first, the presence of the Rbfox family based on their common RRM domain was validated in mTEC (Section 6.2). After the successful validation of a protein containing the RRM domain by immunofluorescence imaging as well as by flow cytometry, I generated *in vivo* models with a TEC-targeted loss of either *Rbfox1* and *Rbfox2* alone or combined. Analysis of the double-knockout mice could not be included in this thesis due to time constraints (Section 6.6).

The loss of *Rbfox1* expression in TEC resulted in a mild increase (15 %) in overall thymus cellularity without compromising that of the epithelial compartment (Table 6.2). Although the TEC cellularity in the SKO1 mice was unaffected, alterations in thymocyte selection or maturation stages are possible. An increase in thymocyte numbers

in thymi of SKO1 mice could account for the observed increase in thymus cellularity with a comparable fraction of EPCAM<sup>+</sup> cells in these mice. Possible reasons for a larger number of thymocytes could be a block at the double-positive (DP) stage, where the DP compartment might be provided with new cells continuously, with only few cells progressing to single-positive stages. However, this would only increase the thymus cellularity if the accumulated DP cells would not undergo apoptosis. Surface molecules on TEC could be targets of RBFOX1-mediated splicing and alterations due to differential splicing upon the loss of RBFOX1 could account for thymocyte developmental blocks. To analyse different thymocyte developmental and selection stages, flow cytometry could be used.

In contrast, mice deficient for the expression of *Rbfox2* in TEC displayed a relative decrease in medullary in lieu of an increase of cortical thymic epithelia. However, the differences in TEC frequencies were not significantly different in male mice. In SKO2 mice, the maturation of mTEC or the expression of maturation markers such as AIRE was not altered. The decrease in TEC cellularity and decrease in mTEC frequency in SKO2 mice could be due to alterations in TEC development. During embryonic TEC development, different signals are important for patterning and differentiation, including those mediated by the family of fibroblast growth factors (Fgf) [253]. The Fgf receptor important in thymic development is FGFR2, which I found to be differentially spliced in mature mTEC from mice with a TEC-targeted loss of *Rbfox2* (Figure 6.21). FGFR2IIIb encoded for by

**Table 6.2:** Summary of results of thymus cellularity and TEC phenotyping for mice with TEC-targeted loss of *Rbfox1* and *Rbfox2* expression.

|                                            | <b>SKO1</b> | <b>WT</b> | <b>p-value</b> |
|--------------------------------------------|-------------|-----------|----------------|
| Total cellularity [in M]                   | 491         | 363       | 0.02           |
| Frequency of EPCAM <sup>+</sup> [in %]     | 0.16        | 0.16      | 0.90           |
| Frequency of mTEC [in %]                   | 90.68       | 91.13     | 0.69           |
| Frequency of AIRE <sup>+</sup> mTEC [in %] | 34.67       | 37.28     | 0.41           |
|                                            | <b>SKO2</b> | <b>WT</b> | <b>p-value</b> |
| Total cellularity [in M]                   | 382         | 414       | 0.44           |
| Frequency of EPCAM <sup>+</sup> [in %]     | 0.12        | 0.17      | 0.04           |
| Frequency of mTEC [in %]                   | 86.18       | 90.83     | 0.03           |
| Frequency of AIRE <sup>+</sup> mTEC [in %] | 40.21       | 38.32     | 0.45           |

on *Fgfr2* isoform was shown to be present on the surface of developing TEC [253]. The thymi of mice that lack this receptor isoform developed until embryonic day 12 (E12), but did not continue to proliferate as much as wildtype thymi after E12 [253]. As a consequence, *FGFR2IIIb*<sup>-/-</sup> mice were shown to have a hypoplastic thymus. I hypothesize that the differential splicing in the extracellular domain of *Fgfr2* receptor could have an impact on TEC development and result in the reduced TEC cellularity. However, from the data presented here it cannot be assessed, (i) whether the same splice variant and RBFOX2 itself were present during embryonic development and (ii) if all spliced variants are translated into protein. However, previous reports of RBFOX2 expression in human embryonic stem cells [255] and its requirement during brain development [146] would suggest that RBFOX2 protein is produced at least during neuronal development. The genes involved in the regulation of differentiation from a bipotent TEC progenitor into mTEC and cTEC are still under investigation. Therefore an assessment of such genes, that might be affected by differential splicing in TEC with a loss of *Rbfox2*, was not possible.

To further assess the thymic architecture with respect to the cTEC and mTEC compartment of the SKO1 and SKO2 mice, hematoxylin & eosin (H&E) staining of tissue sections and immunofluorescent staining of Keratin 5 (K5) and Keratin 8 (K8) would be required. Furthermore, an assessment of peripheral T cells and regulatory T cells could give indications of a reduced or defective T cell repertoire in these mice. In knockout mouse models of proteins important for PGE, such as *Aire*, signs of autoimmune disease have been observed by assessing lymphocyte infiltrates or auto-antibodies in mice aged up to 35 weeks [26]. To assess whether autoimmune phenotypes occur in the mouse models generated here, lymphocyte infiltration and auto-antibodies in peripheral organs of aged mice would need to be assessed.

The TEC-targeted loss of *Rbfox1* had limited effects on the transcriptome with a low number of differentially expressed genes in mature mTEC from SKO1 mice. This observation was expected given that *Rbfox1* regulated splicing, which might not

necessarily effect gene expression. However, major splicing changes could affect overall gene expression, especially if alternatively spliced transcripts would undergo non-sense mediated decay resulting in reduced transcript levels and consequently reduced gene expression. In a *Nestin*-cre driven knockout of *Rbfox1*, an upregulation of RBFOX2 was observed in nuclear extracts from the brain indicating that RBFOX2 compensated for the loss of RBFOX1 [145]. In contrast, the upregulation of RBFOX1 found in the *Rbfox2*<sup>lox/lox</sup>:*Nestin*-cre<sup>+</sup> mice was only modest [146]. In immature and mature mTEC of SKO1 or SKO2, the gene expression of other Rbfox family members was not altered.

Similar to the SKO1 mice, in mature mTEC from SKO2 mice a low number of differentially expressed genes was observed. Amongst the downregulated genes was another splicing factor, namely *Ptbp2*. *Ptbp2* is a neuronal splicing factor [256], but has recently also been implicated to regulate splicing during spermatogenesis using RNA-sequencing and HITS-CLIP from a conditional *Ptbp2*-knockout model [257]. *Ptbp2* was previously identified as a splicing target of RBFOX2 in human embryonic stem cells *in vitro* [255]. The read density from a cross-linking immunoprecipitation plus sequencing (CLIP-seq) experiment using RBFOX2 was observed to be unexpectedly high in a region of *Ptbp2*. In addition, increased exon skipping was observed upon siRNA knockdown of *Rbfox2* in these cells. From these reports and the decrease in *Ptbp2* observed here, it could be speculated that RBFOX2 binding to *Ptbp2* has a stabilising effect on *Ptbp2* transcripts. Upon the loss of *Rbfox2*, such an effect would be lost resulting in decreased expression of *Ptbp2*. In contrast, Gehman *et al.* could not validate differential splicing of *Ptbp2* in the *Rbfox2*:*Nestin*-cre mouse model [146].

With respect to splicing alterations in mature mTEC deficient of either *Rbfox1* or *Rbfox2*, the dataset generated here (i) was possibly not sampled from enough cells and (ii) did likely not have a high enough read depth to sufficiently assess differential splicing. The overall read numbers were similar to the ENCODE mouse tissue datasets used in previous chapters, but the libraries were found to be of low complexity, as shown by their high duplication rate (Figure 6.14, 6.15). The assessment of differential splicing

in lowly expressed TRA might therefore only be partially possible. Another likely explanation for the low number of splicing alterations in single-knockout animals could be the redundancy between RBFOX1 and RBFOX2. Gehman *et al.* have previously shown that the cerebellum of *Rbfox1*<sup>+/-</sup>:*Rbfox2*<sup>-/-</sup> mice had a similar phenotype to the one observed in *Rbfox2*<sup>-/-</sup> cerebella [146]. The authors suggested that *Rbfox2* was likely required during cerebellum development and in the adult cerebellum. In the latter, it is partially redundant with *Rbfox1*. A double-knockout of both *Rbfox1* and *Rbfox2* in TEC could reveal the full repertoire of splice variants mediated by these factors.

Despite the limited read depth, the differentially spliced events showed overlap with known *Rbfox* target genes and GO categories representing processes regulated by *Rbfox*. These observations suggest that the observed differential splicing events might be biologically relevant.

For the *Rbfox1* conditional knockout using *Nestin*-cre, a limited number of differential splicing events were previously observed [145]. Gehman *et al.* reported only 20 significantly differentially spliced events with support from Affymetrix Exon Junction Arrays and RT-PCR data. This number was substantially lower than for instance observed in the neocortex of *Nova1* null mice [202]. The *Nestin*-cre driven *Rbfox2* knockout resulted in 29 significantly differential inclusion events for cassette exons and mutually exclusive exons using the same methodology as used for the analysis of the *Rbfox1*<sup>lox/lox</sup>:*Nestin*-cre model [146]. Between the two *Nestin*-cre knockout models, eleven common changes with the same direction of the change occurred [146]. Gehman *et al.* further reported that not all splicing changes were predictable by the location of the binding sequence. In mature mTEC from SKO1 mice, five genes with significantly differentially spliced events were overlapping the 36 genes identified in *Nestin*-cre driven knockouts of *Rbfox1* and *Rbfox2*. In SKO2 mice, only one common gene was found with previous models. These observations could either be suggestive of a cell-type specific set of splicing events regulated by RBFOX1 and RBFOX2 or it could be a consequence of different technologies (exon junction microarrays versus RNA-sequencing) used to detect differential splicing.

## Conclusion

In this chapter, TEC-specific conditional knockout mouse models of the TRSF *Rbfox1* and another, more ubiquitously-expressed *Rbfox* family member, *Rbfox2*, were generated. While the loss of *Rbfox1* in TEC revealed a comparable thymus cellularity, TEC cellularity and mTEC maturation, the loss of *Rbfox2* resulted in a reduced number of mTEC. The investigation of the transcriptome showed few differentially expressed genes in immature and mature mTEC with a loss of either *Rbfox1* or *Rbfox2*. Differential splicing analysis revealed enrichment of categories relevant to *Rbfox* function and known examples of *Rbfox* targets in mature TEC with a conditional loss of either *Rbfox1* or *Rbfox2*. However, further transcriptomic analysis of more deeply sequenced samples comprising of larger numbers of cells - including from double-knockout animals for both factors - is needed to confirm a general role for *Rbfox1* and *Rbfox2* in splicing in this cell type.





# 7

## General Discussion

The thymus is an essential organ for T cell development and selection, in which T cells with randomly-generated receptors get selected based on the reactivity between the T cell receptor (TCR) on their surface and MHC-peptide complexes on antigen-presenting cells, which are mainly comprised of thymic epithelial cells (TEC). Via a process termed promiscuous gene expression (PGE), medullary TEC (mTEC) can express 89 % of tissue-restricted antigens (TRA) for presentation to the developing T cells [27]. The role of alternative splicing of TRA in the process of self-tolerance is not fully understood. A first insight was provided by a mechanistic study showing that the absence of an alternatively-spliced exon in the thymus resulted in the escape of potentially auto-reactive T cells. These T cells showed auto-reactivity upon stimulation with a peptide derived from the missing exon in the *Plp* gene [108]. Advances in RNA-sequencing revealed genome-wide information on splicing in mTEC in recent years. Observations of high splicing entropy [101, 50] and a high number of TRA junctions in mTEC [103] revealed that these cells have a high transcriptome complexity. While the reports regarding splicing complexity in mature mTEC were consistent, the contribution of AIRE to splicing in mTEC has been controversial. One report suggested that ‘promiscuous splicing’ in mature mTEC was regulated by AIRE [101]. A further study showed differentially expressed exons without changes in gene expression in *Aire*-knockout samples [41], while another report found no significant difference in the detected fraction of TRA junctions for the majority of assessed tissues between wildtype mature mTEC and *Aire*-knockout mTEC [103]. In this thesis I extend the understanding of splicing in TEC by computationally and experimentally assessing the splicing landscape and the role of individual splicing factors in TEC.

## Splicing landscape in TEC

The first aim of this thesis was to assess the splicing landscape in TEC, which I addressed by generating a common reference transcriptome (mT&T assembly) from publically available datasets in combination with unpublished data from our groups (Chapter 3, Appendix Tables B.1 and B.2). The transcriptome assembly was then validated by long-read sequencing from the Oxford Nanopore Technology (ONT) platform (Section 3.3). To my knowledge, this is the first time a long-read RNA-sequencing dataset has

been generated from TEC. Using the ONT dataset I validated 97 % of highly expressed novel mT&T transcripts (Section 3.3.3). The ONT dataset with average read length of 1,000 bp was additionally important for the validation of transcript structures. Using these reads I could also confirm the presence of both isoforms in a known alternatively spliced gene in mature mTEC, where the isoforms are normally found in two different peripheral tissues. A remaining issue with the ONT technology is the high error rate and the inaccuracies in mapping the splice sites unambiguously. The limited number of ONT reads, absence of strand information and the high number of required amplification cycles were further limitations of the dataset. Consistent with a previous publication, where Workman *et al.* identified a large set of unannotated transcripts using ONT data, I also extended the TEC transcriptome by an additional 9,200 novel ONT transcripts found in a transcriptome assembly from filtered ONT reads (Section 3.3.2.2). The transcript coverage by individual ONT reads confirms that the structures originated from one individual read rather than from a combination of short reads, which could have been derived from different transcript structures that did not necessarily originate from the same cell. The long-read sequencing therefore extended the information on the mTEC transcriptome and will serve as an important resource for future studies of transcript structures in mTEC. Such studies could for example include the study of lincRNA and their role in TEC.

The high splicing complexity in mature mTEC observed by quantifying the total number of splice junctions was consistent with previous reports of splicing in mTEC (Section 4.2.1) [101, 103]. For the first time, the fraction of isoforms was assessed here, revealing the highest fraction of isoforms per gene in mature mTEC in comparison to 22 peripheral tissues examined here (50 %, Section 4.2.2). Furthermore, I showed that the presence of tissue-restricted transcripts was significantly higher in mature mTEC than in testis and brain, which also have highly complex transcriptomes (Section 4.3). The results of examining annotated and novel transcripts revealed the highest fraction of annotated transcripts (65 %, Section 4.2.3), but fewer novel transcripts than for instance observed for the testis (1,572 novel transcripts in mature mTEC, Section 4.4). In addition, the assessment of percentage spliced-in values for individual exons including short microexons

showed the absence of a set of peripheral tissue-restricted exons and a set of neuronal microexons in mTEC (Section 4.3.3). These observations suggest that alternative splicing in mTEC might not be a promiscuous process, which does not agree with the report by Keane *et al.* [101] (further discussed below).

## Alternative splicing in single mTEC

Promiscuous gene expression within the mTEC compartment was shown to be heterogeneous with individual *Aire*-regulated TRA only found in a small frequency of individual mTEC [27, 41]. Therefore, alternative splicing might also differ between individual mTEC and would ideally be assessed at the single-cell level. For one example gene, *Calca*, I showed that two transcripts normally restricted to two different peripheral tissues were simultaneously detected in the same single mature mTEC (Section 4.2.4). Following this observation I speculated that there might be heterogeneity in splicing at the single-cell level. Such a heterogeneity could be assessed using SmartSeq2 datasets, which were shown to have good 5' to 3' coverage of transcripts [136, 258]. Although the transcript coverage was not strongly biased, the read depth was not sufficient to assess further examples at the single-cell level. The low input for current single-cell technologies requires multiple amplification steps. Such amplification can lead to over-representation of highly expressed genes and, even more so, to over-representation of major isoforms of a gene. This over-representation would result in loss of information about the transcript variety within each gene. To disentangle the transcript structures from short-read datasets with low read depth is therefore challenging.

## AIRE's contribution to alternative splicing in TEC

One aim was to re-evaluate the role of AIRE in splicing using deep RNA-sequencing data from wildtype and *Aire*-knockout mTEC. Here I observed a limited direct role of AIRE in splicing regulation, but instead found that AIRE likely regulated transcript usage (Sections 5.2.1 and 5.2.4). This observation can explain previous reports by Meredith *et*

*al.* of differential exon expression between *Aire*-knockout and wildtype mTEC irrespective of gene expression [41]. Furthermore, this agrees with the hypothesis proposed by Danan-Gotthold *et al.* of a minor role for AIRE in the regulation of alternative splicing.

## Regulation of alternative splicing in TEC

The spliceosome has previously been described as a large, dynamic complex, involving RNA as well as a variety of proteins [259]. During the process of splicing, multiple steps require energy in the form of ATP. Given the energy expense to generate the variety of splice isoforms in mTEC, one can hypothesise that the process of alternative splicing in mTEC would have a specific purpose in this cell type. In peripheral tissues such processes affected by alternative splicing were for instance in development (*Dab1* splicing by NOVA1 and NOVA2 in brain development [57]), cell recognition (*Pcdh* splicing in inter-neuronal cell recognition [74]) and tissue function (*Titin* splicing in by RBM20 in the heart [62]). Given the special function of mature mTEC as antigen-presenting cells, a purpose for the extensive alternative splicing would be the requirement for presentation of alternative splice isoforms on the cell surface of mature mTEC to mediate T cell selection. Although the transcriptome complexity in TEC has been reported previously, it remained unclear, how the tissue-restricted splicing was achieved. One of the aims of this work was to evaluate the expression of tissue-restricted splicing factors (TRSF) and interrogate whether these are underlying the tissue-restricted splicing observed in mTEC. The expression of a specific set of TRSF in TEC could indicate that splicing is a 'directed' process rather than as suggest by Keane *et al.* a 'promiscuous' process. A survey of the peripheral mouse tissues and TEC for the expression of splicing factors revealed the grouping of tissues based on splicing factor expression (Section 5.3) confirming that TRSF expression was important for tissue identity. Furthermore, I found a set of TRSF such as *Esrp1*, *Msi1*, *Rbfox1* and *Rbm24* to be highly expressed in mTEC. However, not all peripheral TRSF were detected in mTEC, where the expression of for instance *Nova1*, *Nova2* and *Srrm4* was absent. The TRSF *Srrm4* is a major regulator of neuronal microexon inclusion [222]. Consistent with the absence of *Srrm4*, I also observed the

absence of a set of neuronal microexons (Section 4.3.3), suggesting that the transcriptome in TEC was incomplete with respect to neuronal microexons. A hypothesis for the absence of other splicing factors such as *Nova1* and *Nova2* could be that these factors are co-regulating splicing events with other splicing factors. In the case of NOVA, it has been shown that it regulated inclusion or exclusion of specific exons synergistically with RBFOX [75]. Therefore, it could be speculated that the presence of RBFOX in TEC is sufficient for the splicing of neuronal antigens usually spliced by NOVA. An alternative hypothesis would be the absence of NOVA target exons in TEC. The under-representation of brain-specific transcripts as shown for the cerebellum (Figure 4.5) could indicate the absence of such neuronal transcripts.

In general, alternative splicing is a complex process with a range of splicing factors functioning in a regulatory network (Section 1.3.3). Moreover, splicing factors frequently occur in families with high redundancy such as RBFOX1 and RBFOX2 [145]. These two observations on the one hand indicate that a range of splicing factors could be responsible for the transcriptome complexity in mature mTEC. On the other hand, it is a challenging task to attribute one alternative splicing event to one splicing factor and vice versa, if a redundant regulatory network is present. Another approach utilising the single-cell dataset to interrogate TRSF function in TEC was to investigate differential splicing events via *in silico* single-cell pools based on TRSF expression in single cells. However, this approach did not show higher levels of differential splicing events in single-cell pools with high versus low TRSF expression compared to pools stratified by high versus low TRA expression (Section 5.4.3). Possible explanations might be (i) redundancy between splicing factors or (ii) that splicing in TEC was regulated by other factors, possibly not characterised as TRSF yet. The analysis could be improved by assessing appropriate positive and negative controls to ensure the dataset is suitable for the approach.

## Investigation of the *in vivo* effects of a tissue-restricted splicing factor in TEC

For a functional study of splicing in mTEC, a mouse model remains the most suitable approach. However, a knockout of parts of the core spliceosome was not an option to assess the impact of alternative splicing in mTEC as constitutive splicing would be affected, which was not of direct interest here. Moreover, the core spliceosome, for instance the component CWC27, is required for development and normal cell function [260]. The interest here was specifically in the production of tissue-restricted splice variants, thus a TEC-specific conditional knockout of a TRSF could lead to a (partial) loss of these splice variants in TEC (depending on redundancy between splicing factors). Furthermore, investigating the binding sites of splicing factors or RNA-binding proteins could give important indications about their function and targets in TEC. The binding sites could also be compared to those identified in the periphery. Experiments such as HITS-CLIP require a large number of cells as input (for instance 100-150 mg of tissue lysate from the mouse brain [73]). While not achievable in mouse due to the low number of mTEC, application of HITS-CLIP to human could give insights into which RNA binding proteins interact with RNA in mouse TEC.

The computational prioritisation revealed *Rbfox1* as a suitable candidate (Chapter 5) and I therefore hypothesised that it would regulate splicing in mTEC. The final aim of the thesis was to generate, for the first time, a TEC-specific conditional knockout mouse model of a splicing factor (Chapter 6). In mice with a TEC-targeted loss of *Rbfox1* (SKO1), I observed a mild increase in thymus cellularity with unaffected TEC cellularity. These observations could indicate an accumulation of double-positive thymocytes in the thymic cortex of SKO1 mice. Experiments investigating the phenotypes of thymocytes could support this hypothesis. The double-positive thymocytes could possibly not be able to mature further into single-positive T cells, due to the absence of positive or negative selection signals from cTEC. In contrast to SKO1 mice, a significant decrease of the fraction of EPCAM<sup>+</sup> cells and the relative frequency of mTEC was observed in mice with a TEC-targeted loss of *Rbfox2* (SKO2). An exon in the FGF receptor *Fgfr2* was

found to have higher inclusion levels in the SKO2 mice and it is possible that such an alteration could affect TEC development. *Rbfox2* is, however, the non tissue-restricted member of the Rbfox family and is likely to have a general role in epithelial cells, which is not directly connected to producing tissue-restricted splice variants.

Although the RNA-sequencing data from the single-knockout models had a low library complexity and possibly not the power to detect all differential splicing events, the genes with observed splicing events upon loss of *Rbfox2* were enriched in categories related to peripheral tissues with Rbfox activity such as muscle-related or neuron-related categories. For *Rbfox1* such enrichments in the genes containing differentially spliced events was not observed. These observation indicated that the loss of *Rbfox2* had an impact on the TEC transcriptome and as a consequence potentially on T cell selection.

Overall, the SKO1 and SKO2 mice can only partially answer whether the absence of RBFOX-mediated splice variants has an effect on central tolerance induction, due to the known redundancy of RBFOX1 and RBFOX2 [145, 146]. To account for this redundancy and to fully answer the hypothesis of splicing regulation by *Rbfox1* in mTEC, a double-knockout of both *Rbfox1* and *Rbfox2* was desired, but due to time constraints only investigations of the single-knockout mice could be performed here. The analysis of the double-knockout would be required to investigate the effects of a larger set of RBFOX-regulated splice variants in TEC. A complete absence of all peripheral splice variants in TEC could potentially result in T cells not being appropriately selected. T cells with reactivity against spliced epitopes would not be negatively selected and thus potentially auto-reactive T cells could be present in the periphery. As shown by Klein *et al.* this was the case for *Plp* exons. However, an external stimulus such as injection of PLP peptides produced from the missing exon was required to induce disease. Given these observations, it is possible that in SKO1/SKO2 animals auto-reactive T cells would escape the thymus in the absence of expression of tissue-restricted splice variants in TEC. As a consequence, auto-reactive T cells would be present in the periphery, however, it is possible that these might not cause disease due to suppression by regulatory T cells.



The repertoire of T cells exiting the thymus could be assessed using single-cell TCR sequencing (reviewed by De Simone *et al.* [261]). However, the current limitation with this method is the scalability to investigate the large T cell repertoire, but future methods will likely address this limitations.

### **Splicing is likely a 'directed' process rather than a 'promiscuous' process in TEC**

Following the observation of high splicing complexity (Chapter 4), the expression of a subset of peripheral TRSF in mTEC (Section 5.3.2) and the limited role of AIRE (Section 5.2.1) I hypothesised that alternative splicing in mTEC was a 'directed' process rather than the result of promiscuously-expressed TRSF. The lack of a set of peripheral exons such as neuronal microexons (Section 4.3.3) was further in disagreement with the hypothesis of promiscuous splicing in mTEC. If alternative splicing in mTEC is a 'directed' process, it could be important for achieving self-tolerance. Future work will be needed to rule out the option of promiscuous splicing by for instance examining splicing patterns in a larger number of single cells to assess if co-expression of specific transcript structures are present. Further studies could investigate whether promiscuously-expressed splicing factors are properly spliced, translated and then localised to the nucleus in mTEC.

### **Future studies**

An open question regarding splicing in achieving central tolerance was the variety of peptides bound to MHCII in mTEC. It is further unclear, whether these peptides included epitopes from tissue-restricted splice junctions or transcripts. Over the past two decades it has been shown that the peptides bound to MHCII are around 14 amino acids in length with a core of 9 amino acids (reviewed by Unanue [262]). The MHCII peptidome of mTEC has, however, not been investigated on a global scale due to limitations in the technologies and the limited cell number of mTEC. So far, only studies of MHCI- and MHCII-bound peptides on dendritic cells from human thymus tissue have been assessed [263]. Using 30 M cells, Adamopoulou *et al.* identified up to 48 peptides from MHCII

complexes per human donor. The authors found differences between individual donors with respect to protein composition that MHCI and MHCII ligands had originated from. Another study by Collado *et al.* analysed the MHCII peptidome, more specifically the HLA-DR-associated peptidome, in human thymus samples and found an average peptide length of 15 amino acids [264]. However, Collado *et al.* commented that only a small number of peptides were derived from TRA and needed to be further investigated. Due to the low cell numbers, the analysis could not be performed using only mTEC or cTEC as input. Taken together, recent approaches have not yet been successful in answering questions about the global MHCII peptidome in mTEC or cTEC.

With respect to peptides that include splice junctions, an even higher resolution of the MHCII peptidome would be required as spliced epitopes are likely rare. Irrespective of the MHC complex, the identification of peptides produced from alternatively-spliced transcripts remains challenging [96]. Current mass spectrometry approaches have been shown to have limited sensitivity - even with a high input - to detect peptides derived from alternatively spliced transcripts [96], which makes a validation on the peptide level challenging. Future advances in mass spectrometry methods could open avenues to understanding the existence of spliced epitopes and whether these are likely to be bound by MHCII complexes. In the context of TEC, especially method advancements in MHCII peptidomics for instance from lower input would produce valuable data to address the hypothesis that a range of spliced epitopes is indeed presented on MHCII complexes.

Another question around RNA biology in mTEC that could be addressed in the future is the presence of RNA modifications in mTEC. The recently published direct RNA-sequencing protocol using the ONT platform can provide information on such RNA modifications. In TEC biology, RNA modifications could extend the understanding of how transcripts from promiscuously-expressed genes are potentially 'targeted' for surface presentation. It could be hypothesised that specific RNA modifications would ensure the translation of such transcripts and subsequent transport of the peptides to the proteasome. For instance, the methylation of adenosin at the N<sup>1</sup> position, termed m<sup>1</sup>A, was frequently

found near translational start sites and was correlated with a higher level of translation (reviewed by Roundtree *et al.* [265]). Moreover, mRNA transcripts that had acetylated cytosins (ac<sup>4</sup>C) were translated more efficiently [266]. With current technologies direct RNA-sequencing, however, requires a high amount of input RNA, which is challenging to achieve from mouse TEC.

The regulatory network of splicing factors was shown to be complex and redundant, therefore deciphering which splicing factor regulates which splicing event is not trivial. Hence, experiments have frequently been performed in *in vitro* system with over-expression of a minigene with splice sites of interest [267, 268]. In such a system, the components can be controlled, while in an *in vivo* model a variety of factors remains unknown. If a splicing factor of interest is expressed in a mTEC cell line, experiments using minigenes containing splice variants detected in RNA-sequencing from mature mTEC could give indications about the activity of the respective splicing factor in mTEC. An over-expression system with a defined splicing factor set could also be examined. Although culturing primary mTEC *in vitro* is challenging, recent method developments might facilitate future mTEC *in vitro* studies [269].

## Summary

In summary, this thesis has generated novel insights into self-antigen splicing in mTEC using short-read and long-read RNA-sequencing data. These insights include the confirmation of a high splicing complexity in mTEC, the accurate but incomplete representation of peripheral isoforms in mTEC and the discovery of novel transcripts in mTEC. Furthermore, the primary effect of AIRE in alternative splicing was observed to be in transcript usage, which reconciled the results from previous studies. As additional mechanisms of alternative splicing regulation in mTEC, TRSF were defined and their expression examined in TEC and peripheral tissues. Finally, the investigation of the *in vivo* function of a TRSF, namely *Rbfox1*, revealed that *Rbfox1* expression was dispensable

for TEC development and maturation. However, genes with differential splicing events in mature mTEC were found to be over-enriched in Rbfox target genes.

# Bibliography

- [1] G. Gaud, R. Lesourne, and P. E. Love, “Regulatory mechanisms in T cell receptor signalling”, *Nature Reviews Immunology*, vol. 18, p. 485, Aug. 2018.
- [2] O. Boyman and J. Sprent, “The role of interleukin-2 during homeostasis and activation of the immune system”, *Nature Reviews Immunology*, vol. 12, pp. 180–190, Mar. 2012.
- [3] Y. Y. Wan and R. A. Flavell, “How Diverse—CD4 Effector T Cells and their Functions”, *Journal of Molecular Cell Biology*, vol. 1, pp. 20–36, Oct. 2009.
- [4] L. Klein, B. Kyewski, P. M. Allen, and K. A. Hogquist, “Positive and negative selection of the T cell repertoire: what thymocytes see (and don’t see)”, *Nature Reviews Immunology*, vol. 14, pp. 377–391, May 2014.
- [5] H. T. Petrie and J. C. Zúñiga-Pflücker, “Zoned Out: Functional Mapping of Stromal Signaling Microenvironments in the Thymus”, *Annual Review of Immunology*, vol. 25, no. 1, pp. 649–679, 2007.
- [6] Y. Takahama, “Journey through the thymus: stromal guides for T-cell development and selection”, *Nature Reviews Immunology*, vol. 6, pp. 127–135, Feb. 2006.
- [7] S. Murata, K. Sasaki, T. Kishimoto, S.-i. Niwa, H. Hayashi, Y. Takahama, and K. Tanaka, “Regulation of CD8+ T Cell Development by Thymus-Specific Proteasomes”, *Science*, vol. 316, pp. 1349–1353, June 2007.
- [8] S. R. Daley, D. Y. Hu, and C. C. Goodnow, “Helios marks strongly autoreactive CD4+ T cells in two major waves of thymic deletion distinguished by induction of PD-1 or NF-kB”, *Journal of Experimental Medicine*, vol. 210, pp. 269–285, Feb. 2013.

- [9] G. L. Stritesky, Y. Xing, J. R. Erickson, L. A. Kalekar, X. Wang, D. L. Mueller, S. C. Jameson, and K. A. Hogquist, “Murine thymic selection quantified using a unique method to capture deleted T cells”, *Proceedings of the National Academy of Sciences*, vol. 110, pp. 4679–4684, Mar. 2013.
- [10] K. A. Hogquist, Y. Xing, F.-C. Hsu, and V. S. Shapiro, “T cell adolescence: maturation events beyond positive selection”, *Journal of immunology (Baltimore, Md. : 1950)*, vol. 195, pp. 1351–1357, Aug. 2015.
- [11] M. Matloubian, C. G. Lo, G. Cinamon, M. J. Lesneski, Y. Xu, V. Brinkmann, M. L. Allende, R. L. Proia, and J. G. Cyster, “Lymphocyte egress from thymus and peripheral lymphoid organs is dependent on S1p receptor 1”, *Nature*, vol. 427, pp. 355–360, Jan. 2004.
- [12] J. Gordon, V. A. Wilson, N. F. Blair, J. Sheridan, A. Farley, L. Wilson, N. R. Manley, and C. C. Blackburn, “Functional evidence for a single endodermal origin for the thymic epithelium”, *Nature Immunology*, vol. 5, pp. 546–553, May 2004.
- [13] C. C. Bleul, T. Corbeaux, A. Reuter, P. Fisch, J. S. Mönting, and T. Boehm, “Formation of a functional thymus initiated by a postnatal epithelial progenitor cell”, *Nature*, vol. 441, pp. 992–996, June 2006.
- [14] T. Corbeaux, I. Hess, J. B. Swann, B. Kanzler, A. Haas-Assenbaum, and T. Boehm, “Thymopoiesis in mice depends on a Foxn1-positive thymic epithelial cell lineage”, *Proceedings of the National Academy of Sciences*, vol. 107, pp. 16613–16618, Sept. 2010.
- [15] T. C. Metzger, I. S. Khan, J. M. Gardner, M. L. Mouchess, K. P. Johannes, A. K. Krawisz, K. M. Skrzypczynska, and M. S. Anderson, “Lineage Tracing and Cell Ablation Identify a Post-Aire-Expressing Thymic Epithelial Cell Population”, *Cell Reports*, vol. 5, pp. 166–179, Oct. 2013.
- [16] J. Abramson and G. Anderson, “Thymic Epithelial Cells”, *Annual Review of Immunology*, vol. 35, no. 1, p. null, 2017.

- [17] M. S. Anderson and M. A. Su, “AIRE expands: new roles in immune tolerance and beyond”, *Nature Reviews Immunology*, vol. 16, pp. 247–258, Apr. 2016.
- [18] H. Takaba, Y. Morishita, Y. Tomofuji, L. Danks, T. Nitta, N. Komatsu, T. Kodama, and H. Takayanagi, “Fezf2 Orchestrates a Thymic Program of Self-Antigen Expression for Immune Tolerance”, *Cell*, vol. 163, pp. 975–987, Nov. 2015.
- [19] C. Bornstein, S. Nevo, A. Giladi, N. Kadouri, M. Pouzolles, F. Gerbe, E. David, A. Machado, A. Chuprin, B. Tóth, O. Goldberg, S. Itzkovitz, N. Taylor, P. Jay, V. S. Zimmermann, J. Abramson, and I. Amit, “Single-cell mapping of the thymic stroma identifies IL-25-producing tuft epithelial cells”, *Nature*, vol. 559, pp. 622–626, July 2018.
- [20] C. N. Miller, I. Proekt, J. v. Moltke, K. L. Wells, A. R. Rajpurkar, H. Wang, K. Rattay, I. S. Khan, T. C. Metzger, J. L. Pollack, A. C. Fries, W. W. Lwin, E. J. Wigton, A. V. Parent, B. Kyewski, D. J. Erle, K. A. Hogquist, L. M. Steinmetz, R. M. Locksley, and M. S. Anderson, “Thymic tuft cells promote an IL-4-enriched medulla and shape thymocyte development”, *Nature*, vol. 559, pp. 627–631, July 2018.
- [21] A. J. White, K. Nakamura, W. E. Jenkinson, M. Saini, C. Sinclair, B. Seddon, P. Narendran, K. Pfeffer, T. Nitta, Y. Takahama, J. H. Caamano, P. J. L. Lane, E. J. Jenkinson, and G. Anderson, “Lymphotoxin Signals from Positively Selected Thymocytes Regulate the Terminal Differentiation of Medullary Thymic Epithelial Cells”, *The Journal of Immunology*, vol. 185, pp. 4769–4776, Oct. 2010.
- [22] E. Lkhagvasuren, M. Sakata, I. Ohigashi, and Y. Takahama, “Lymphotoxin  $\beta$  Receptor Regulates the Development of CCL21-Expressing Subset of Postnatal Medullary Thymic Epithelial Cells”, *The Journal of Immunology*, vol. 190, pp. 5110–5117, May 2013.
- [23] X. Wang, M. Laan, R. Bichele, K. Kisand, H. S. Scott, and P. Peterson, “Post-Aire Maturation of Thymic Medullary Epithelial Cells Involves Selective Expression of Keratinocyte-Specific Autoantigens”, *Frontiers in Immunology*, vol. 3, Mar. 2012.

- [24] O. Bruserud, B. E. Oftedal, A. B. Wolff, and E. S. Husebye, “AIRE-mutations and autoimmune disease”, *Current Opinion in Immunology*, vol. 43, pp. 8–15, Dec. 2016.
- [25] C. Ramsey, O. Winqvist, L. Puhakka, M. Halonen, A. Moro, O. Kämpe, P. Eskelin, M. Peltö-Huikko, and L. Peltonen, “Aire deficient mice develop multiple features of APECED phenotype and show altered immune response”, *Human Molecular Genetics*, vol. 11, pp. 397–409, Feb. 2002.
- [26] M. S. Anderson, E. S. Venanzi, L. Klein, Z. Chen, S. P. Berzins, S. J. Turley, H. v. Boehmer, R. Bronson, A. Dierich, C. Benoist, and D. Mathis, “Projection of an Immunological Self Shadow Within the Thymus by the Aire Protein”, *Science*, vol. 298, pp. 1395–1401, Nov. 2002.
- [27] S. N. Sansom, N. Shikama-Dorn, S. Zhanybekova, G. Nusspaumer, I. C. Macaulay, M. E. Deadman, A. Heger, C. P. Ponting, and G. A. Holländer, “Population and single-cell genomics reveal the *Aire* dependency, relief from Polycomb silencing, and distribution of self-antigen expression in thymic epithelia”, *Genome Research*, vol. 24, pp. 1918–1931, Dec. 2014.
- [28] T. Org, A. Rebane, K. Kisand, M. Laan, U. Haljasorg, R. Andreson, and P. Peterson, “AIRE activated tissue specific genes have histone modifications associated with inactive chromatin”, *Human Molecular Genetics*, vol. 18, pp. 4699–4710, Dec. 2009.
- [29] A. S. Koh, R. E. Kingston, C. Benoist, and D. Mathis, “Global relevance of Aire binding to hypomethylated lysine-4 of histone-3”, *Proceedings of the National Academy of Sciences*, vol. 107, pp. 13016–13021, July 2010.
- [30] J. Abramson, M. Giraud, C. Benoist, and D. Mathis, “Aire’s Partners in the Molecular Control of Immunological Tolerance”, *Cell*, vol. 140, pp. 123–135, Jan. 2010.
- [31] M. Gaetani, V. Matafora, M. Saare, D. Spiliotopoulos, L. Mollica, G. Quilici, F. Chignola, V. Mannella, C. Zucchelli, P. Peterson, A. Bachi, and G. Musco,



- “AIRE-PHD fingers are structural hubs to maintain the integrity of chromatin-associated interactome”, *Nucleic Acids Research*, vol. 40, pp. 11756–11768, Dec. 2012.
- [32] S. Yang, K. Bansal, J. Lopes, C. Benoist, and D. Mathis, “Aire’s plant homeodomain(PHD)-2 is critical for induction of immunological tolerance”, *Proceedings of the National Academy of Sciences*, vol. 110, pp. 1833–1838, Jan. 2013.
- [33] M. Waterfield, I. S. Khan, J. T. Cortez, U. Fan, T. Metzger, A. Greer, K. Fasano, M. Martinez-Llordella, J. L. Pollack, D. J. Erle, M. Su, and M. S. Anderson, “The transcriptional regulator Aire coopts the repressive ATF7ip-MBD1 complex for the induction of immunotolerance”, *Nature Immunology*, vol. 15, pp. 258–265, Mar. 2014.
- [34] I. Oven, N. Brdicková, J. Kohoutek, T. Vaupotic, M. Narat, and B. M. Peterlin, “AIRE recruits P-TEFb for transcriptional elongation of target genes in medullary thymic epithelial cells”, *Molecular and Cellular Biology*, vol. 27, pp. 8815–8823, Dec. 2007.
- [35] K. Žumer, A. Plemenitaš, K. Saksela, and B. M. Peterlin, “Patient mutation in AIRE disrupts P-TEFb binding and target gene transcription”, *Nucleic Acids Research*, vol. 39, pp. 7908–7919, Oct. 2011.
- [36] M. Giraud, N. Jmari, L. Du, F. Carallis, T. J. F. Nieland, F. M. Perez-Campo, O. Bensaude, D. E. Root, N. Hacohen, D. Mathis, and C. Benoist, “An RNAi screen for Aire cofactors reveals a role for Hnrnp1 in polymerase release and Aire-activated ectopic transcription”, *Proceedings of the National Academy of Sciences*, vol. 111, pp. 1491–1496, Jan. 2014.
- [37] B. Kyewski and J. Derbinski, “Self-representation in the thymus: an extended view”, *Nature Reviews Immunology*, vol. 4, pp. 688–698, Sept. 2004.
- [38] J. Derbinski, S. Pinto, S. Rösch, K. Hexel, and B. Kyewski, “Promiscuous gene expression patterns in single medullary thymic epithelial cells argue for a stochastic

- mechanism”, *Proceedings of the National Academy of Sciences*, vol. 105, no. 2, pp. 657–662, 2008.
- [39] J. Villasenor, W. Besse, C. Benoist, and D. Mathis, “Ectopic expression of peripheral-tissue antigens in the thymic epithelium: Probabilistic, monoallelic, misinitiated”, *Proceedings of the National Academy of Sciences*, vol. 105, pp. 15854–15859, Oct. 2008.
- [40] P. Brennecke, A. Reyes, S. Pinto, K. Rattay, M. Nguyen, R. K  chler, W. Huber, B. Kyewski, and L. M. Steinmetz, “Single-cell transcriptome analysis reveals coordinated ectopic gene-expression patterns in medullary thymic epithelial cells”, *Nature Immunology*, vol. 16, pp. 933–941, Aug. 2015.
- [41] M. Meredith, D. Zemmour, D. Mathis, and C. Benoist, “Aire controls gene expression in the thymic epithelium with ordered stochasticity”, *Nature Immunology*, vol. 16, pp. 942–949, Aug. 2015.
- [42] T. N. LaFlam, G. Seumois, C. N. Miller, W. Lwin, K. J. Fasano, M. Waterfield, I. Proekt, P. Vijayanand, and M. S. Anderson, “Identification of a novel cis-regulatory element essential for immune tolerance”, *The Journal of Experimental Medicine*, vol. 212, pp. 1993–2002, Nov. 2015.
- [43] U. Haljasorg, R. Bichele, M. Saare, M. Guha, J. Maslovskaja, K. K  nd, A. Remm, M. Pihlap, L. Tomson, K. Kisand, M. Laan, and P. Peterson, “A highly conserved NF- $\kappa$ B-responsive enhancer is critical for thymic expression of Aire in mice”, *European Journal of Immunology*, vol. 45, pp. 3246–3256, Dec. 2015.
- [44] C. E. Schaller, C. L. Wang, G. Beck-Engeser, L. Goss, H. S. Scott, M. S. Anderson, and M. Wabl, “Expression of Aire and the Early Wave of Apoptosis in Spermatogenesis”, *The Journal of Immunology*, vol. 180, pp. 1338–1343, Feb. 2008.
- [45] T. Yamano, J. Nedjic, M. Hinterberger, M. Steinert, S. Koser, S. Pinto, N. Gerdes, E. Lutgens, N. Ishimaru, M. Busslinger, B. Brors, B. Kyewski, and L. Klein,

- “Thymic B Cells Are Licensed to Present Self Antigens for Central T Cell Tolerance Induction”, *Immunity*, vol. 42, pp. 1048–1061, June 2015.
- [46] G. Bin, Z. Jiarong, W. Shihao, S. Xiuli, X. Cheng, C. Liangbiao, and Z. Ming, “Aire Promotes the Self-Renewal of Embryonic Stem Cells Through Lin28”, *Stem Cells and Development*, vol. 21, pp. 2878–2890, Apr. 2012.
- [47] J. M. Gardner, J. J. DeVoss, R. S. Friedman, D. J. Wong, Y. X. Tan, X. Zhou, K. P. Johannes, M. A. Su, H. Y. Chang, M. F. Krummel, and M. S. Anderson, “Deletional Tolerance Mediated by Extrathymic Aire-Expressing Cells”, *Science*, vol. 321, pp. 843–847, Aug. 2008.
- [48] J. M. Gardner, T. C. Metzger, E. J. McMahon, B. B. Au-Yeung, A. K. Krawisz, W. Lu, J. D. Price, K. P. Johannes, A. T. Satpathy, K. M. Murphy, K. V. Tarbell, A. Weiss, and M. S. Anderson, “Extrathymic Aire-Expressing Cells Are a Distinct Bone Marrow-Derived Population that Induce Functional Inactivation of CD4+ T Cells”, *Immunity*, vol. 39, pp. 560–572, Sept. 2013.
- [49] A. C. Anderson, L. B. Nicholson, K. L. Legge, V. Turchin, H. Zaghoulani, and V. K. Kuchroo, “High Frequency of Autoreactive Myelin Proteolipid Protein-Specific T Cells in the Periphery of Naive Mice”, *The Journal of Experimental Medicine*, vol. 191, pp. 761–770, Mar. 2000.
- [50] C. St-Pierre, A. Trofimov, S. Brochu, S. Lemieux, and C. Perreault, “Differential Features of AIRE-Induced and AIRE-Independent Promiscuous Gene Expression in Thymic Epithelial Cells”, *The Journal of Immunology*, vol. 195, pp. 498–506, July 2015.
- [51] S. M. Berget, C. Moore, and P. A. Sharp, “Spliced segments at the 5’ terminus of adenovirus 2 late mRNA.”, *Proceedings of the National Academy of Sciences of the United States of America*, vol. 74, pp. 3171–3175, Aug. 1977.

- [52] L. T. Chow, R. E. Gelinas, T. R. Broker, and R. J. Roberts, “An amazing sequence arrangement at the 5’ ends of adenovirus 2 messenger RNA”, *Cell*, vol. 12, pp. 1–8, Sept. 1977.
- [53] M. Chen and J. L. Manley, “Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches”, *Nature Reviews Molecular Cell Biology*, vol. 10, pp. 741–54, 2009.
- [54] M. Shapiro and P. Senapathy, “RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression.”, *Nucleic Acids Research*, vol. 15, pp. 7155 – 7174, Nov. 1987.
- [55] C. B. Burge, T. Tuschl, P. A. Sharp, C. B. Burge, T. Tuschl, and P. A. Sharp, “Splicing of Precursors to mRNAs by the Spliceosomes”, *2354 Nucleic Acids Research*, 1999.
- [56] X.-D. Fu and M. Ares, “Context-dependent control of alternative splicing by RNA-binding proteins”, *Nature Reviews Genetics*, vol. 15, pp. 689–701, Aug. 2014.
- [57] F. E. Baralle and J. Giudice, “Alternative splicing as a regulator of development and tissue identity”, *Nature Reviews Molecular Cell Biology*, vol. 18, pp. 437–451, July 2017.
- [58] M. Irimia, R. Weatheritt, J. D. Ellis, N. Parikshak, T. Gonatopoulos-Pournatzis, M. Babor, M. Quesnel-Vallières, J. Tapial, B. Raj, D. O’Hanlon, M. Barrios-Rodiles, M. E. Sternberg, S. Cordes, F. Roth, J. Wrana, D. Geschwind, and B. Blencowe, “A Highly Conserved Program of Neuronal Microexons Is Misregulated in Autistic Brains”, *Cell*, vol. 159, pp. 1511–1523, Dec. 2014.
- [59] K. Bhalla, H. A. Phillips, J. Crawford, O. L. D. McKenzie, J. C. Mulley, H. Eyre, A. E. Gardner, G. Kremmidiotis, and D. F. Callen, “The de novo chromosome 16 translocations of two patients with abnormal phenotypes (mental retardation and epilepsy) disrupt the *A2BP1* gene”, *Journal of Human Genetics*, vol. 49, pp. 308–311, June 2004.

- [60] C. L. Martin, J. A. Duvall, Y. Ilkin, J. S. Simon, M. G. Arreaza, K. Wilkes, A. Alvarez-Retuerto, A. Whichello, C. M. Powell, K. Rao, E. Cook, and D. H. Geschwind, “Cytogenetic and molecular characterization of A2bp1/FOX1 as a candidate gene for autism”, *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, vol. 144B, no. 7, pp. 869–876, 2007.
- [61] G. Barnby, A. Abbott, N. Sykes, A. Morris, D. E. Weeks, R. Mott, J. Lamb, A. J. Bailey, and A. P. Monaco, “Candidate-Gene Screening and Association Analysis at the Autism-Susceptibility Locus on Chromosome 16p: Evidence of Association at GRIN2a and ABAT”, *The American Journal of Human Genetics*, vol. 76, pp. 950–966, June 2005.
- [62] W. Guo, S. Schafer, M. L. Greaser, M. H. Radke, M. Liss, T. Govindarajan, H. Maatz, H. Schulz, S. Li, A. M. Parrish, V. Dauksaite, P. Vakeel, S. Klaassen, B. Gerull, L. Thierfelder, V. Regitz-Zagrosek, T. A. Hacker, K. W. Saupe, G. W. Dec, P. T. Ellinor, C. A. MacRae, B. Spallek, R. Fischer, A. Perrot, C. Özcelik, K. Saar, N. Hubner, and M. Gotthardt, “*RBM20*, a gene for hereditary cardiomyopathy, regulates titin splicing”, *Nature Medicine*, vol. 18, pp. 766–773, May 2012.
- [63] M. M. Refaat, S. A. Lubitz, S. Makino, Z. Islam, J. M. Frangiskakis, H. Mehdi, R. Gutmann, M. L. Zhang, H. L. Bloom, C. A. MacRae, S. C. Dudley, A. A. Shalaby, R. Weiss, D. M. McNamara, B. London, and P. T. Ellinor, “Genetic variation in the alternative splicing regulator RBM20 is associated with dilated cardiomyopathy”, *Heart Rhythm*, vol. 9, pp. 390–396, Mar. 2012.
- [64] N. L. Barbosa-Morais, M. Irimia, Q. Pan, H. Y. Xiong, S. Gueroussov, L. J. Lee, V. Slobodeniuc, C. Kutter, S. Watt, R. Colak, T. Kim, C. M. Misquitta-Ali, M. D. Wilson, P. M. Kim, D. T. Odom, B. J. Frey, and B. J. Blencowe, “The evolutionary landscape of alternative splicing in vertebrate species”, *Science (New York, N.Y.)*, vol. 338, pp. 1587–1593, Dec. 2012.

- [65] J. Merkin, C. Russell, P. Chen, and C. B. Burge, “Evolutionary Dynamics of Gene and Isoform Regulation in Mammalian Tissues”, *Science*, vol. 338, pp. 1593–1599, Dec. 2012.
- [66] Y. Wang, X. Xiao, J. Zhang, R. Choudhury, A. Robertson, K. Li, M. Ma, C. B. Burge, and Z. Wang, “A complex network of factors with overlapping affinities represses splicing through intronic elements”, *Nature Structural & Molecular Biology*, vol. 20, pp. 36–45, Jan. 2013.
- [67] M. Pascual, M. Vicente, L. Monferrer, and R. Artero, “The Muscleblind family of proteins: an emerging class of regulators of developmentally programmed alternative splicing”, *Differentiation*, vol. 74, pp. 65–80, Mar. 2006.
- [68] J. G. Conboy, “Developmental regulation of RNA processing by Rbfox proteins: Rbfox regulation of RNA processing”, *Wiley Interdisciplinary Reviews: RNA*, vol. 8, p. e1398, Mar. 2017.
- [69] F. Yue and Mouse ENCODE Consortium, “A comparative encyclopedia of DNA elements in the mouse genome”, *Nature*, vol. 515, pp. 355–364, Nov. 2014.
- [70] D. D. Pervouchine, S. Djebali, A. Breschi, C. A. Davis, P. P. Barja, A. Dobin, A. Tanzer, J. Lagarde, C. Zaleski, L.-H. See, M. Fastuca, J. Drenkow, H. Wang, G. Bussotti, B. Pei, S. Balasubramanian, J. Monlong, A. Harmanici, M. Gerstein, M. A. Beer, C. Notredame, R. Guigó, and T. R. Gingeras, “Enhanced transcriptome maps from multiple mouse tissues reveal evolutionary constraint in gene expression”, *Nature Communications*, vol. 6, p. 5903, Jan. 2015.
- [71] D. D. Licatalosi, A. Mele, J. J. Fak, J. Ule, M. Kayikci, S. W. Chi, T. A. Clark, A. C. Schweitzer, J. E. Blume, X. Wang, J. C. Darnell, and R. B. Darnell, “HITS-CLIP yields genome-wide insights into brain alternative RNA processing”, *Nature*, vol. 456, pp. 464–469, Nov. 2008.

- [72] J. Ule, G. Stefani, A. Mele, M. Ruggiu, X. Wang, B. Taneri, T. Gaasterland, B. J. Blencowe, and R. B. Darnell, “An RNA map predicting Nova-dependent splicing regulation”, *Nature*, vol. 444, pp. 580–586, Nov. 2006.
- [73] S. Weyn-Vanhentenryck, A. Mele, Q. Yan, S. Sun, N. Farny, Z. Zhang, C. Xue, M. Herre, P. Silver, M. Zhang, A. Krainer, R. Darnell, and C. Zhang, “HITS-CLIP and Integrative Modeling Define the Rbfox Splicing-Regulatory Network Linked to Brain Development and Autism”, *Cell Reports*, vol. 6, pp. 1139–1152, Mar. 2014.
- [74] B. Raj and B. Blencowe, “Alternative Splicing in the Mammalian Nervous System: Recent Insights into Mechanisms and Functional Roles”, *Neuron*, vol. 87, pp. 14–27, July 2015.
- [75] C. Zhang, M. A. Frias, A. Mele, M. Ruggiu, T. Eom, C. B. Marney, H. Wang, D. D. Licatalosi, J. J. Fak, and R. B. Darnell, “Integrative Modeling Defines the Nova Splicing-Regulatory Network and Its Combinatorial Controls”, *Science*, vol. 329, pp. 439–443, July 2010.
- [76] Y. I. Li, L. Sanchez-Pulido, W. Haerty, and C. P. Ponting, “RBFOX and PTBP1 proteins regulate the alternative splicing of micro-exons in human brain transcripts”, *Genome Research*, vol. 25, pp. 1–13, Jan. 2015.
- [77] M. Jangi and P. Sharp, “Building Robust Transcriptomes with Master Splicing Factors”, *Cell*, vol. 159, pp. 487–498, Oct. 2014.
- [78] N. L. Barbosa-Morais, “Systematic genome-wide annotation of spliceosomal proteins reveals differential gene family expansion”, *Genome Research*, vol. 16, pp. 66–77, Dec. 2006.
- [79] A. R. Grosso, A. Q. Gomes, N. L. Barbosa-Morais, S. Caldeira, N. P. Thorne, G. Grech, M. von Lindern, and M. Carmo-Fonseca, “Tissue-specific splicing factor gene expression signatures”, *Nucleic Acids Research*, vol. 36, pp. 4823–4832, July 2008.

- [80] M. Jain, H. E. Olsen, B. Paten, and M. Akeson, “The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community”, *Genome Biology*, vol. 17, p. 239, Nov. 2016.
- [81] R. J. Roberts, M. O. Carneiro, and M. C. Schatz, “The advantages of SMRT sequencing”, *Genome Biology*, vol. 14, p. 405, July 2013.
- [82] H. Tilgner, F. Jahanbani, T. Blauwkamp, A. Moshrefi, E. Jaeger, F. Chen, I. Harel, C. D. Bustamante, M. Rasmussen, and M. P. Snyder, “Comprehensive transcriptome analysis using synthetic long-read sequencing reveals molecular co-association of distant splicing events”, *Nature Biotechnology*, vol. 33, pp. 736–742, May 2015.
- [83] J. L. Weirather, M. de Cesare, Y. Wang, P. Piazza, V. Sebastiano, X.-J. Wang, D. Buck, and K. F. Au, “Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis”, *F1000Research*, vol. 6, p. 100, June 2017.
- [84] A. Byrne, A. E. Beaudin, H. E. Olsen, M. Jain, C. Cole, T. Palmer, R. M. DuBois, E. C. Forsberg, M. Akeson, and C. Vollmers, “Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells”, *Nature Communications*, vol. 8, p. ncomms16027, July 2017.
- [85] M. Bolisetty, G. Rajadinakaran, and B. Graveley, “Determining Exon Connectivity in Complex mRNAs by Nanopore Sequencing”, *bioRxiv*, p. 019752, 2015.
- [86] D. R. Garalde, E. A. Snell, D. Jachimowicz, B. Sipos, J. H. Lloyd, M. Bruce, N. Pantic, T. Admassu, P. James, A. Warland, M. Jordan, J. Ciccone, S. Serra, J. Keenan, S. Martin, L. McNeill, E. J. Wallace, L. Jayasinghe, C. Wright, J. Blasco, S. Young, D. Brocklebank, S. Juul, J. Clarke, A. J. Heron, and D. J. Turner, “Highly parallel direct RNA sequencing on an array of nanopores”, *Nature Methods*, Jan. 2018.
- [87] R. E. Workman, A. Tang, P. S. Tang, M. Jain, J. R. Tyson, P. C. Zuzarte, T. Gilpatrick, R. Razaghi, J. Quick, N. Sadowski, N. Holmes, J. G. d. Jesus,



- K. Jones, T. P. Snutch, N. J. Loman, B. Paten, M. W. Loose, J. T. Simpson, H. E. Olsen, A. N. Brooks, M. Akeson, and W. Timp, “Nanopore native RNA sequencing of a human poly(A) transcriptome”, *bioRxiv*, p. 459529, Nov. 2018.
- [88] S. Oikonomopoulos, Y. C. Wang, H. Djambazian, D. Badescu, and J. Ragoussis, “Benchmarking of the Oxford Nanopore MinION sequencing for quantitative and qualitative assessment of cDNA populations”, *Scientific Reports*, vol. 6, p. srep31602, Aug. 2016.
- [89] A. D. Tang, C. M. Soulette, M. J. v. Baren, K. Hart, E. Hrabeta-Robinson, C. J. Wu, and A. N. Brooks, “Full-length transcript characterization of SF3b1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns”, *bioRxiv*, p. 410183, Sept. 2018.
- [90] H. Li, “Minimap2: pairwise alignment for nucleotide sequences”, *Bioinformatics*, vol. 34, pp. 3094–3100, Sept. 2018.
- [91] S. Koren, B. P. Walenz, K. Berlin, J. R. Miller, N. H. Bergman, and A. M. Phillippy, “Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation”, *Genome Research*, vol. 27, pp. 722–736, May 2017.
- [92] M. Tardaguila, L. d. l. Fuente, C. Marti, C. Pereira, F. J. Pardo-Palacios, H. d. Risco, M. Ferrell, M. Mellado, M. Macchietto, K. Verheggen, M. Edelmann, I. Ezkurdia, J. Vazquez, M. Tress, A. Mortazavi, L. Martens, S. Rodriguez-Navarro, V. Moreno-Manzano, and A. Conesa, “SQANTI: extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification”, *Genome Research*, vol. 28, pp. 396–411, Mar. 2018.
- [93] E. T. Wang, R. Sandberg, S. Luo, I. Khrebtkova, L. Zhang, C. Mayr, S. F. Kingsmore, G. P. Schroth, and C. B. Burge, “Alternative isoform regulation in human tissue transcriptomes”, *Nature*, vol. 456, pp. 470–476, Nov. 2008.

- [94] Q. Pan, O. Shai, L. J. Lee, B. J. Frey, and B. J. Blencowe, “Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing”, *Nature Genetics*, vol. 40, pp. 1413–1415, Dec. 2008.
- [95] M. L. Tress, F. Abascal, and A. Valencia, “Alternative Splicing May Not Be the Key to Proteome Complexity”, *Trends in Biochemical Sciences*, vol. 42, pp. 98–110, Feb. 2017.
- [96] B. J. Blencowe, “The Relationship between Alternative Splicing and Proteomic Complexity”, *Trends in Biochemical Sciences*, vol. 0, May 2017.
- [97] R. J. Weatheritt, T. Sterne-Weiler, and B. J. Blencowe, “The ribosome-engaged landscape of alternative splicing”, *Nature Structural & Molecular Biology*, vol. advance online publication, Nov. 2016.
- [98] C. St-Pierre, S. Brochu, J. R. Vanegas, M. Dumont-Lagacé, S. Lemieux, and C. Perreault, “Transcriptome sequencing of neonatal thymic epithelial cells”, *Scientific Reports*, vol. 3, May 2013.
- [99] M. Giraud, H. Yoshida, J. Abramson, P. B. Rahl, R. A. Young, D. Mathis, and C. Benoist, “Aire unleashes stalled RNA polymerase to induce ectopic gene expression in thymic epithelial cells”, *Proceedings of the National Academy of Sciences*, vol. 109, pp. 535–540, Jan. 2012.
- [100] A. Chuprin, A. Avin, Y. Goldfarb, Y. Herzig, B. Levi, A. Jacob, A. Sela, S. Katz, M. Grossman, C. Guyon, M. Rathaus, H. Y. Cohen, I. Sagi, M. Giraud, M. W. McBurney, E. S. Husebye, and J. Abramson, “The deacetylase Sirt1 is an essential regulator of Aire-mediated induction of central immunological tolerance”, *Nature Immunology*, vol. 16, pp. 737–745, July 2015.
- [101] P. Keane, R. Ceredig, and C. Seoighe, “Promiscuous mRNA splicing under the control of AIRE in medullary thymic epithelial cells”, *Bioinformatics*, vol. 31, no. 7, pp. 986–990, 2014.

- [102] W. Ritchie, S. Granjeaud, D. Puthier, and D. Gautheret, “Entropy Measures Quantify Global Splicing Disorders in Cancer”, *PLoS Computational Biology*, vol. 4, p. e1000011, Mar. 2008.
- [103] M. Danan-Gotthold, C. Guyon, M. Giraud, E. Y. Levanon, and J. Abramson, “Extensive RNA editing and splicing increase immune self-representation diversity in medullary thymic epithelial cells”, *Genome Biology*, vol. 17, p. 219, 2016.
- [104] M. Soumillon, A. Necsulea, M. Weier, D. Brawand, X. Zhang, H. Gu, P. Barthès, M. Kokkinaki, S. Nef, A. Gnirke, M. Dym, B. de Massy, T. Mikkelsen, and H. Kaessmann, “Cellular Source and Mechanisms of High Transcriptome Complexity in the Mammalian Testis”, *Cell Reports*, vol. 3, pp. 2179–2190, June 2013.
- [105] M. Fijak, S. Bhushan, and A. Meinhardt, “Immunoprivileged Sites: The Testis”, in *Suppression and Regulation of Immune Responses: Methods and Protocols* (M. C. Cuturi and I. Anegón, eds.), Methods in Molecular Biology, pp. 459–470, Totowa, NJ: Humana Press, 2011.
- [106] B. Engelhardt, P. Vajkoczy, and R. O. Weller, “The movers and shapers in immune privilege of the CNS”, *Nature Immunology*, vol. 18, pp. 123–131, Feb. 2017.
- [107] K. B. Cook, H. Kazan, K. Zuberi, Q. Morris, and T. R. Hughes, “RBPDB: a database of RNA-binding specificities”, *Nucleic Acids Research*, vol. 39, pp. D301–D308, Jan. 2011.
- [108] L. Klein, M. Klugmann, K.-A. Nave, V. K. Tuohy, and B. Kyewski, “Shaping of the autoreactive T-cell repertoire by a splice variant of self protein expressed in thymic epithelial cells”, *Nature medicine*, vol. 6, no. 1, pp. 56–61, 2000.
- [109] R. L. Terry, I. Ifergan, and S. D. Miller, “Experimental Autoimmune Encephalomyelitis in Mice”, in *Multiple Sclerosis: Methods and Protocols* (R. Weissert, ed.), Methods in Molecular Biology, pp. 145–160, New York, NY: Springer New York, 2016.

- [110] V. K. Tuohy, R. A. Sobel, and M. B. Lees, “Myelin proteolipid protein-induced experimental allergic encephalomyelitis. Variations of disease expression in different strains of mice.”, *The Journal of Immunology*, vol. 140, pp. 1868–1873, Mar. 1988.
- [111] L. Wang, J. Winnewisser, C. Federle, G. Jessberger, K.-A. Nave, H. B. Werner, B. Kyewski, L. Klein, and M. Hinterberger, “Epitope-Specific Tolerance Modes Differentially Specify Susceptibility to Proteolipid Protein-Induced Experimental Autoimmune Encephalomyelitis”, *Frontiers in Immunology*, vol. 8, 2017.
- [112] J. T. Chang, E. M. Shevach, and B. M. Segal, “Regulation of Interleukin (IL)-12 Receptor  $\beta 2$  Subunit Expression by Endogenous IL-12: A Critical Step in the Differentiation of Pathogenic Autoreactive T Cells”, *Journal of Experimental Medicine*, vol. 189, pp. 969–978, Mar. 1999.
- [113] B. M. Segal and E. M. Shevach, “IL-12 unmasks latent autoimmune disease in resistant mice.”, *Journal of Experimental Medicine*, vol. 184, pp. 771–775, Aug. 1996.
- [114] J. Reddy, Z. Illes, X. Zhang, J. Encinas, J. Pyrdol, L. Nicholson, R. A. Sobel, K. W. Wucherpfennig, and V. K. Kuchroo, “Myelin proteolipid protein-specific CD4<sup>+</sup>CD25<sup>+</sup> regulatory cells mediate genetic resistance to experimental autoimmune encephalomyelitis”, *Proceedings of the National Academy of Sciences*, vol. 101, pp. 15434–15439, Oct. 2004.
- [115] R. S. Dogra, P. Vaidyanathan, K. R. Prabakar, K. E. Marshall, J. C. Hutton, and A. Pugliese, “Alternative splicing of G6pc2, the gene coding for the islet-specific glucose-6-phosphatase catalytic subunit-related protein (IGRP), results in differential expression in human thymus and spleen compared with pancreas”, *Diabetologia*, vol. 49, pp. 953–957, May 2006.
- [116] V. M. de Jong, J. R. F. Abreu, A. A. Verrijn Stuart, A. R. van der Slik, K. Verhaeghen, M. A. Engelse, B. Blom, F. J. T. Staal, F. K. Gorus, and B. O. Roep, “Alternative splicing and differential expression of the islet autoantigen IGRP

- between pancreas and thymus contributes to immunogenicity of pancreatic islets but not diabetogenicity in humans”, *Diabetologia*, vol. 56, pp. 2651–2658, Dec. 2013.
- [117] J. Diez, Y. Park, M. Zeller, D. Brown, D. Garza, C. Ricordi, J. Hutton, G. S. Eisenbarth, and A. Pugliese, “Differential Splicing of the IA-2 mRNA in Pancreas and Lymphoid Organs as a Permissive Genetic Mechanism for Autoimmunity Against the IA-2 Type 1 Diabetes Autoantigen”, *Diabetes*, vol. 50, pp. 895–900, Apr. 2001.
- [118] Y. S. Park, E. Kawasaki, K. Kelemen, L. Yu, M. R. Schiller, M. Rewers, M. Mizuta, G. S. Eisenbarth, and J. C. Hutton, “Humoral autoreactivity to an alternatively spliced variant of ICA512/IA-2 in Type I diabetes”, *Diabetologia*, vol. 43, pp. 1293–1301, Oct. 2000.
- [119] D. Sims, N. E. Ilott, S. N. Sansom, I. M. Sudbery, J. S. Johnson, K. A. Fawcett, A. J. Berlanga-Taylor, S. Luna-Valero, C. P. Ponting, and A. Heger, “CGAT: computational genomics analysis toolkit”, *Bioinformatics*, vol. 30, pp. 1290–1291, May 2014.
- [120] L. Goodstadt, “Ruffus: a lightweight Python library for computational pipelines”, *Bioinformatics*, vol. 26, pp. 2778–2779, Nov. 2010.
- [121] D. Kim, B. Langmead, and S. L. Salzberg, “HISAT: a fast spliced aligner with low memory requirements”, *Nature Methods*, vol. 12, pp. 357–360, Mar. 2015.
- [122] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin, “The Sequence Alignment/Map format and SAMtools”, *Bioinformatics*, vol. 25, pp. 2078–2079, Aug. 2009.
- [123] R. Patro, G. Duggal, M. I. Love, R. A. Irizarry, and C. Kingsford, “Salmon provides fast and bias-aware quantification of transcript expression”, *Nature Methods*, Mar. 2017.

- [124] A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras, “STAR: ultrafast universal RNA-seq aligner”, *Bioinformatics*, vol. 29, pp. 15–21, Jan. 2013.
- [125] L. Venturini, S. Caim, G. G. Kaithakottil, D. L. Mapleson, and D. Swarbreck, “Leveraging multiple transcriptome assembly methods for improved gene structure annotation”, *GigaScience*, vol. 7, Aug. 2018.
- [126] A. A. Hagberg, D. A. Schult, and P. J. Swart, “Exploring Network Structure, Dynamics, and Function using NetworkX”, in *Proceedings of the 7th Python in Science Conference*, (Pasadena, CA USA), pp. 11 – 15, Gaël Varoquaux and Travis Vaught and Jarrod Millman, 2008.
- [127] M. I. Love, W. Huber, and S. Anders, “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2”, *Genome Biology*, vol. 15, no. 12, p. 550, 2014.
- [128] Y. Liao, G. K. Smyth, and W. Shi, “featureCounts: an efficient general purpose program for assigning sequence reads to genomic features”, *Bioinformatics*, vol. 30, pp. 923–930, Apr. 2014.
- [129] J. H. Bullard, E. Purdom, K. D. Hansen, and S. Dudoit, “Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments”, *BMC Bioinformatics*, vol. 11, p. 94, 2010.
- [130] N. L. Bray, H. Pimentel, P. Melsted, and L. Pachter, “Near-optimal probabilistic RNA-seq quantification”, *Nature Biotechnology*, vol. 34, pp. 525–527, May 2016.
- [131] H. Pimentel, N. L. Bray, S. Puente, P. Melsted, and L. Pachter, “Differential analysis of RNA-seq incorporating quantification uncertainty”, *Nature Methods*, vol. advance online publication, June 2017.
- [132] L. Yi, H. Pimentel, N. L. Bray, and L. Pachter, “Gene-level differential analysis at transcript-level resolution”, *Genome Biology*, vol. 19, p. 53, Apr. 2018.

- [133] H. J. M. d. Jonge, R. S. N. Fehrmann, E. S. J. M. d. Bont, R. M. W. Hofstra, F. Gerbens, W. A. Kamps, E. G. E. d. Vries, A. G. J. v. d. Zee, G. J. t. Meerman, and A. t. Elst, “Evidence Based Selection of Housekeeping Genes”, *PLOS ONE*, vol. 2, p. e898, Sept. 2007.
- [134] A. R. Quinlan and I. M. Hall, “BEDTools: a flexible suite of utilities for comparing genomic features”, *Bioinformatics*, vol. 26, pp. 841–842, Mar. 2010.
- [135] J. L. Trincado, J. C. Entizne, G. Hysenaj, B. Singh, M. Skalic, D. J. Elliott, and E. Eyraas, “SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions”, *Genome Biology*, vol. 19, p. 40, Mar. 2018.
- [136] S. Picelli, A. K. Björklund, O. R. Faridani, S. Sagasser, G. Winberg, and R. Sandberg, “Smart-seq2 for sensitive full-length transcriptome profiling in single cells”, *Nature Methods*, vol. 10, pp. 1096–1098, Sept. 2013.
- [137] A. E. Handel, N. Shikama-Dorn, S. Zhanybekova, S. Maio, A. N. Graedel, S. Zuklys, C. P. Ponting, and G. A. Holländer, “Comprehensively Profiling the Chromatin Architecture of Tissue Restricted Antigen Expression in Thymic Epithelial Cells Over Development”, *Frontiers in Immunology*, vol. 9, 2018.
- [138] A. Butler, P. Hoffman, P. Smibert, E. Papalexi, and R. Satija, “Integrating single-cell transcriptomic data across different conditions, technologies, and species”, *Nature Biotechnology*, vol. 36, pp. 411–420, May 2018.
- [139] Y. Katz, E. T. Wang, E. M. Airoidi, and C. B. Burge, “Analysis and design of RNA sequencing experiments for identifying isoform regulation”, *Nature Methods*, vol. 7, pp. 1009–1015, Nov. 2010.
- [140] W. De Coster, S. D’Hert, D. T. Schultz, M. Cruts, C. Van Broeckhoven, and B. Berger, “NanoPack: visualizing and processing long read sequencing data”, *Bioinformatics*, vol. 34, pp. 2666–2669, Aug. 2018.
- [141] R. Suzuki and H. Shimodaira, “Pvclust: an R package for assessing the uncertainty in hierarchical clustering”, *Bioinformatics*, vol. 22, pp. 1540–1542, June 2006.

- [142] N. Kryuchkova-Mostacci and M. Robinson-Rechavi, “A benchmark of gene expression tissue-specificity metrics”, *Briefings in Bioinformatics*, p. bbw008, Feb. 2016.
- [143] I. Yanai, H. Benjamin, M. Shmoish, V. Chalifa-Caspi, M. Shklar, R. Ophir, A. Bar-Even, S. Horn-Saban, M. Safran, E. Domany, D. Lancet, and O. Shmueli, “Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification”, *Bioinformatics*, vol. 21, pp. 650–659, Mar. 2005.
- [144] S. Zuklys, J. Gill, M. P. Keller, M. Hauri-Hohl, S. Zhanybekova, G. Balciunaite, K.-J. Na, L. T. Jeker, K. Hafen, N. Tsukamoto, T. Amagai, M. M. Taketo, W. Krenger, and G. A. Hollander, “Stabilized b-Catenin in Thymic Epithelial Cells Blocks Thymus Development and Function”, *The Journal of Immunology*, vol. 182, pp. 2997–3007, Mar. 2009.
- [145] L. T. Gehman, P. Stoilov, J. Maguire, A. Damianov, C.-H. Lin, L. Shiue, M. Ares, I. Mody, and D. L. Black, “The splicing regulator Rbfox1 (A2bp1) controls neuronal excitation in the mammalian brain”, *Nature Genetics*, vol. 43, pp. 706–711, May 2011.
- [146] L. T. Gehman, P. Meera, P. Stoilov, L. Shiue, J. E. O’Brien, M. H. Meisler, M. Ares, T. S. Otis, and D. L. Black, “The splicing regulator Rbfox2 is required for both cerebellar development and mature motor function”, *Genes & Development*, vol. 26, pp. 445–460, Mar. 2012.
- [147] J. Tapial, K. C. H. Ha, T. Sterne-Weiler, A. Gohr, U. Braunschweig, A. Hermoso-Pulido, M. Quesnel-Vallières, J. Permanyer, R. Sodaiei, Y. Marquez, L. Cozzuto, X. Wang, M. Gómez-Velázquez, T. Rayon, M. Manzanares, J. Ponomarenko, B. J. Blencowe, and M. Irimia, “An atlas of alternative splicing profiles and functional associations reveals new regulatory programs and genes that simultaneously express multiple major isoforms”, *Genome Research*, vol. 27, pp. 1759–1768, Oct. 2017.



- [148] Y. Liu, J. F. Ferguson, C. Xue, I. M. Silverman, B. Gregory, M. P. Reilly, and M. Li, “Evaluating the Impact of Sequencing Depth on Transcriptome Profiling in Human Adipose”, *PLOS ONE*, vol. 8, p. e66883, June 2013.
- [149] J. M. Mudge and J. Harrow, “Creating reference gene annotation for the mouse C57bl6/J genome assembly”, *Mammalian Genome*, vol. 26, no. 9-10, pp. 366–378, 2015.
- [150] A. R. R. Forrest and FANTOM Consortium, “A promoter-level mammalian expression atlas”, *Nature*, vol. 507, pp. 462–470, Mar. 2014.
- [151] S. Djebali, C. A. Davis, A. Merkel, A. Dobin, T. Lassmann, A. Mortazavi, A. Tanzer, J. Lagarde, W. Lin, F. Schlesinger, C. Xue, G. K. Marinov, J. Khatun, B. A. Williams, C. Zaleski, J. Rozowsky, M. Röder, F. Kokocinski, R. F. Abdelhamid, T. Alioto, I. Antoshechkin, M. T. Baer, N. S. Bar, P. Batut, K. Bell, I. Bell, S. Chakraborty, X. Chen, J. Chrast, J. Curado, T. Derrien, J. Drenkow, E. Dumais, J. Dumais, R. Duttagupta, E. Falconnet, M. Fastuca, K. Fejes-Toth, P. Ferreira, S. Foissac, M. J. Fullwood, H. Gao, D. Gonzalez, A. Gordon, H. Gunawardena, C. Howald, S. Jha, R. Johnson, P. Kapranov, B. King, C. Kingswood, O. J. Luo, E. Park, K. Persaud, J. B. Preall, P. Ribeca, B. Risk, D. Robyr, M. Sammeth, L. Schaffer, L.-H. See, A. Shahab, J. Skancke, A. M. Suzuki, H. Takahashi, H. Tilgner, D. Trout, N. Walters, H. Wang, J. Wrobel, Y. Yu, X. Ruan, Y. Hayashizaki, J. Harrow, M. Gerstein, T. Hubbard, A. Reymond, S. E. Antonarakis, G. Hannon, M. C. Giddings, Y. Ruan, B. Wold, P. Carninci, R. Guigó, and T. R. Gingeras, “Landscape of transcription in human cells”, *Nature*, vol. 489, pp. 101–108, Sept. 2012.
- [152] Q. Li, J. B. Brown, H. Huang, and P. J. Bickel, “Measuring reproducibility of high-throughput experiments”, *The Annals of Applied Statistics*, vol. 5, pp. 1752–1779, Sept. 2011.
- [153] J. E. Lattin, K. Schroder, A. I. Su, J. R. Walker, J. Zhang, T. Wiltshire, K. Saijo, C. K. Glass, D. A. Hume, S. Kellie, and M. J. Sweet, “Expression analysis of G

- Protein-Coupled Receptors in mouse macrophages”, *Immunome Research*, vol. 4, p. 5, Apr. 2008.
- [154] M. Pertea, D. Kim, G. M. Pertea, J. T. Leek, and S. L. Salzberg, “Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown”, *Nature Protocols*, vol. 11, pp. 1650–1667, Sept. 2016.
- [155] M. Pertea, G. M. Pertea, C. M. Antonescu, T.-C. Chang, J. T. Mendell, and S. L. Salzberg, “StringTie enables improved reconstruction of a transcriptome from RNA-seq reads”, *Nature Biotechnology*, vol. 33, pp. 290–295, Mar. 2015.
- [156] H. Storrval, D. Ramsköld, and R. Sandberg, “Efficient and Comprehensive Representation of Uniqueness for Next-Generation Sequencing by Minimum Unique Length Analyses”, *PLoS ONE*, vol. 8, p. e53822, Jan. 2013.
- [157] T. Zhang, P. Haws, and Q. Wu, “Multiple variable first exons: a mechanism for cell- and tissue-specific gene regulation”, *Genome Research*, vol. 14, pp. 79–89, Jan. 2004.
- [158] I. S. Gul, P. Hulpiau, Y. Saeys, and F. van Roy, “Evolution and diversity of cadherins and catenins”, *Experimental Cell Research*, vol. 358, pp. 3–9, Sept. 2017.
- [159] E. Bushmanova, D. Antipov, and A. Lapidus, “rnaSPAdes: a de novo transcriptome assembler and its application to RNA-Seq data”, p. 15.
- [160] K. A. Glenn, R. F. Nelson, H. M. Wen, A. J. Mallinger, and H. L. Paulson, “Diversity in Tissue Expression, Substrate Binding, and SCF Complex Formation for a Lectin Family of Ubiquitin Ligases”, *Journal of Biological Chemistry*, vol. 283, pp. 12717–12729, May 2008.
- [161] W. Shao, K. Zumer, K. Fujinaga, and B. M. Peterlin, “FBXO3 Promotes Ubiquitylation and Transcriptional Activity of AIRE”, *Journal of Biological Chemistry*, p. jbc.M116.724401, June 2016.

- [162] M. Kreit, D. Vertommen, L. Gillet, and T. Michiels, “The Interferon-Inducible Mouse Apolipoprotein L9 and Prohibitins Cooperate to Restrict Theiler’s Virus Replication”, *PLOS ONE*, vol. 10, p. e0133190, July 2015.
- [163] M. Halonen, M. Peltö-Huikko, P. Eskelin, L. Peltonen, I. Ulmanen, and M. Kolmer, “Subcellular Location and Expression Pattern of Autoimmune Regulator (Aire), the Mouse Orthologue for Human Gene Defective in Autoimmune Polyendocrinopathy Candidiasis Ectodermal Dystrophy (APECED)”, *Journal of Histochemistry & Cytochemistry*, vol. 49, pp. 197–208, Feb. 2001.
- [164] N. Gromak, A. J. Matlin, T. A. Cooper, and C. W. J. Smith, “Antagonistic regulation of alpha-actinin alternative splicing by CELF proteins and polypyrimidine tract binding protein”, *RNA*, vol. 9, pp. 443–456, Apr. 2003.
- [165] S. L. Chew, “Alternative Splicing of mRNA as a Mode of Endocrine Regulation”, *Trends in Endocrinology & Metabolism*, vol. 8, pp. 405–413, Dec. 1997.
- [166] N. Shikama, G. Nusspaumer, and G. A. Holländer, “Clearing the AIRE: On the Pathophysiological Basis of the Autoimmune Polyendocrinopathy Syndrome Type-1”, *Endocrinology and Metabolism Clinics of North America*, vol. 38, pp. 273–288, June 2009.
- [167] G. P. Alamancos, A. Pagès, J. L. Trincado, N. Bellora, and E. Eyraç, “Leveraging transcript quantification for fast computation of alternative splicing profiles”, *RNA*, vol. 21, pp. 1521–1531, Sept. 2015.
- [168] C. Heit, B. C. Jackson, M. McAndrews, M. W. Wright, D. C. Thompson, G. A. Silverman, D. W. Nebert, and V. Vasiliou, “Update of the human and mouse SERPIN gene superfamily”, *Human Genomics*, vol. 7, no. 1, p. 22, 2013.
- [169] C. Shukunami, Y. Yoshimoto, A. Takimoto, H. Yamashita, and Y. Hiraki, “Molecular characterization and function of tenomodulin, a marker of tendons and ligaments that integrate musculoskeletal components”, *Japanese Dental Science Review*, vol. 52, pp. 84–92, Nov. 2016.

- [170] S. Dex, D. Lin, C. Shukunami, and D. Docheva, “TENogenic MODULating INsider factor: systematic assessment on the functions of tenomodulin gene”, *Gene*, vol. 587, pp. 1–17, Aug. 2016.
- [171] T. Tsukiyama, M. Matsuda-Tsukiyama, M. Bohgaki, S. Terai, S. Tanaka, and S. Hatakeyama, “Ymer Acts as a Multifunctional Regulator in Nuclear Factor-kappa-B and Fas Signaling Pathways”, *Molecular Medicine*, vol. 18, pp. 587–597, 2012.
- [172] W. J. Kent, C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, and a. D. Haussler, “The Human Genome Browser at UCSC”, *Genome Research*, vol. 12, pp. 996–1006, June 2002.
- [173] V. P. Alexandrov and S. I. Naimov, “A Prospectus of Tenomodulin”, *Folia Medica*, vol. 58, pp. 19–27, Mar. 2016.
- [174] S. D. Knight, J. Presto, S. Linse, and J. Johansson, “The BRICHOS Domain, Amyloid Fibril Formation, and Their Relationship”, *Biochemistry*, vol. 52, pp. 7523–7531, Oct. 2013.
- [175] P. A. Keane and C. Seoighe, “Intron length coevolution across mammalian genomes”, *Molecular Biology and Evolution*, p. msw151, Aug. 2016.
- [176] K. C. Kleene, “A possible meiotic function of the peculiar patterns of gene expression in mammalian spermatogenic cells”, *Mechanisms of Development*, vol. 106, pp. 3–23, Aug. 2001.
- [177] D. Ramsköld, E. T. Wang, C. B. Burge, and R. Sandberg, “An Abundance of Ubiquitously Expressed Genes Revealed by Tissue Transcriptome Sequence Data”, *PLoS Computational Biology*, vol. 5, p. e1000598, Dec. 2009.
- [178] D. A. A. Vignali and V. K. Kuchroo, “IL-12 family cytokines: immunological playmakers”, *Nature Immunology*, vol. 13, pp. 722–728, Aug. 2012.

- [179] F. Tang, C. Barbacioru, Y. Wang, E. Nordman, C. Lee, N. Xu, X. Wang, J. Bodeau, B. B. Tuch, A. Siddiqui, K. Lao, and M. A. Surani, “mRNA-Seq whole-transcriptome analysis of a single cell”, *Nature Methods*, vol. 6, pp. 377–382, May 2009.
- [180] N. M. Martinez, L. Agosto, J. Qiu, M. J. Mallory, M. R. Gazzara, Y. Barash, X.-d. Fu, and K. W. Lynch, “Widespread JNK-dependent alternative splicing induces a positive feedback loop through CELF2-mediated regulation of MKK7 during T-cell activation”, *Genes & Development*, vol. 29, pp. 2054–2066, Oct. 2015.
- [181] M. J. Mallory, S. J. Allon, J. Qiu, M. R. Gazzara, I. Tapescu, N. M. Martinez, X.-D. Fu, and K. W. Lynch, “Induced transcription and stability of CELF2 mRNA drives widespread alternative splicing during T-cell signaling”, *Proceedings of the National Academy of Sciences*, vol. 112, pp. E2139–E2148, Apr. 2015.
- [182] S. Ajith, M. R. Gazzara, B. S. Cole, G. Shankarling, N. M. Martinez, M. J. Mallory, and K. W. Lynch, “Position-dependent activity of CELF2 in the regulation of splicing and implications for signal-responsive regulation in T cells”, *RNA Biology*, vol. 13, pp. 569–581, Apr. 2016.
- [183] S. Shen, J. W. Park, Z.-x. Lu, L. Lin, M. D. Henry, Y. N. Wu, Q. Zhou, and Y. Xing, “rMATS: Robust and flexible detection of differential alternative splicing from replicate RNA-Seq data”, *Proceedings of the National Academy of Sciences*, vol. 111, pp. E5593–E5601, Dec. 2014.
- [184] Q. Du, P.-L. Luu, C. Stirzaker, and S. J. Clark, “Methyl-CpG-binding domain proteins: readers of the epigenome”, *Epigenomics*, vol. 7, pp. 1051–1073, Apr. 2015.
- [185] A. Adibekian, B. R. Martin, C. Wang, K.-L. Hsu, D. A. Bachovchin, S. Niessen, H. Hoover, and B. F. Cravatt, “Click-generated triazole ureas as ultrapotent *in vivo*-active serine hydrolase inhibitors”, *Nature Chemical Biology*, vol. 7, pp. 469–478, July 2011.
- [186] G. Palmieri, P. Bergamo, A. Luini, M. Ruvo, M. Gogliettino, E. Langella, M. Saviano, R. N. Hegde, A. Sandomenico, and M. Rossi, “Acylpeptide Hydrolase

- Inhibition as Targeted Strategy to Induce Proteasomal Down-Regulation”, *PLOS ONE*, vol. 6, p. e25888, Oct. 2011.
- [187] J. J.-L. Wong, A. Y. M. Au, W. Ritchie, and J. E. J. Rasko, “Intron retention in mRNA: No longer nonsense”, *BioEssays*, vol. 38, pp. 41–49, Jan. 2016.
- [188] F. A. Flomerfelt, M. G. Kim, and R. H. Schwartz, “Spatial, a gene expressed in thymic stromal cells, depends on three-dimensional thymus organization for its expression”, *Genes and Immunity*, vol. 1, pp. 391–401, Aug. 2000.
- [189] M. Saade, M. Irla, M. Yammine, N. Boulanger, G. Victorero, R. Vincentelli, J. M. Penninger, G. A. Holländer, S. Chauvet, and C. Nguyen, “Spatial (Tbata) expression in mature medullary thymic epithelial cells”, *European Journal of Immunology*, vol. 40, no. 2, pp. 530–538, 2010.
- [190] N. Fisicaro, M. Katerelos, J. Williams, D. Power, A. D’Apice, and M. Pearce, “Identification of genes downregulated in the thymus by cyclosporin-A: Preliminary characterization of clone CSA-19”, *Molecular Immunology*, vol. 32, pp. 565–572, June 1995.
- [191] Z. Shi, K. Fujii, K. M. Kovary, N. R. Genuth, H. L. Röst, M. N. Teruel, and M. Barna, “Heterogeneous Ribosomes Preferentially Translate Distinct Subpools of mRNAs Genome-wide”, *Molecular Cell*, vol. 67, pp. 71–83.e7, July 2017.
- [192] E. F. Wagner and A. R. Nebreda, “Signal integration by JNK and p38 MAPK pathways in cancer development”, *Nature Reviews Cancer*, vol. 9, pp. 537–549, Aug. 2009.
- [193] K. Shuai and B. Liu, “Regulation of JAK–STAT signalling in the immune system”, *Nature Reviews Immunology*, vol. 3, pp. 900–911, Nov. 2003.
- [194] Y. Watanabe, M. Ikegawa, Y. Naruse, and M. Tanaka, “A novel splicing variant form suppresses the activity of full-length signal transducer and activator of transcription 5a”, *The FEBS Journal*, vol. 276, no. 21, pp. 6312–6323, 2009.

- [195] S. Sano, Y. Takahama, T. Sugawara, H. Kosaka, S. Itami, K. Yoshikawa, J.-i. Miyazaki, W. van Ewijk, and J. Takeda, “Stat3 in Thymic Epithelial Cells Is Essential for Postnatal Maintenance of Thymic Architecture and Thymocyte Survival”, *Immunity*, vol. 15, pp. 261–273, Aug. 2001.
- [196] E. Orhan, L. Pr  zeau, S. El Shamieh, K. M. Bujakowska, C. Michiels, Y. Zagar, C. Vol, S. S. Bhattacharya, J.-A. Sahel, F. Sennlaub, I. Audo, and C. Zeitz, “Further insights into GPR179: expression, localization, and associated pathogenic mechanisms leading to complete congenital stationary night blindness”, *Investigative Ophthalmology & Visual Science*, vol. 54, pp. 8041–8050, Dec. 2013.
- [197] H. Han, M. Irimia, P. J. Ross, H.-K. Sung, B. Alipanahi, L. David, A. Golipour, M. Gabut, I. P. Michael, E. N. Nachman, E. Wang, D. Trcka, T. Thompson, D. O’Hanlon, V. Slobodeniuc, N. L. Barbosa-Morais, C. B. Burge, J. Moffat, B. J. Frey, A. Nagy, J. Ellis, J. L. Wrana, and B. J. Blencowe, “MBNL proteins repress ES-cell-specific alternative splicing and reprogramming”, *Nature*, vol. 498, pp. 241–245, June 2013.
- [198] K. Pearson, “LIII. On lines and planes of closest fit to systems of points in space”, *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, pp. 559–572, Nov. 1901.
- [199] H. Hotelling, “Analysis of a complex of statistical variables into principal components.”, *Journal of Educational Psychology*, vol. 24, pp. 417–441, Jan. 1933.
- [200] J. Long and J. Caceres, “The SR protein family of splicing factors: master regulators of gene expression”, *Biochemical Journal*, vol. 417, pp. 15–27, Jan. 2009.
- [201] S. Chan, P. Sridhar, R. Kirchner, Y. J. Lock, Z. Herbert, S. Buonamici, P. Smith, J. Lieberman, and F. Petrocca, “Basal-A Triple-Negative Breast Cancer Cells Selectively Rely on RNA Splicing for Survival”, *Molecular Cancer Therapeutics*, vol. 16, pp. 2849–2861, Dec. 2017.

- [202] J. Ule, A. Ule, J. Spencer, A. Williams, J.-S. Hu, M. Cline, H. Wang, T. Clark, C. Fraser, M. Ruggiu, B. R. Zeeberg, D. Kane, J. N. Weinstein, J. Blume, and R. B. Darnell, “Nova regulates brain-specific splicing to shape the synapse”, *Nature Genetics*, vol. 37, pp. 844–852, Aug. 2005.
- [203] T. Dasgupta and A. N. Ladd, “The importance of CELF control: molecular and biological roles of the CUG-BP, Elav-like family of RNA-binding proteins”, *Wiley Interdisciplinary Reviews: RNA*, vol. 3, no. 1, pp. 104–121, 2012.
- [204] T. Iwamori, Y.-N. Lin, L. Ma, N. Iwamori, and M. M. Matzuk, “Identification and Characterization of RBM44 as a Novel Intercellular Bridge Protein”, *PLOS ONE*, vol. 6, p. e17066, Feb. 2011.
- [205] D. J. Elliott, “The role of potential splicing factors including RBMY, RBMX, hnRNPG-T and STAR proteins in spermatogenesis\*”, *International Journal of Andrology*, vol. 27, no. 6, pp. 328–334, 2004.
- [206] L. van der Maaten and G. Hinton, “Visualizing High-Dimensional Data Using t-SNE”, *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.
- [207] A. de Morrée, M. Droog, L. Grand Moursel, I. J. M. Bisschop, A. Impagliazzo, R. R. Frants, R. Klooster, and S. M. van der Maarel, “Self-regulated alternative splicing at the AHNAK locus”, *The FASEB Journal*, vol. 26, pp. 93–103, Sept. 2011.
- [208] E. Piskounova, C. Polytarchou, J. Thornton, R. LaPierre, C. Pothoulakis, J. Hagan, D. Iliopoulos, and R. Gregory, “Lin28a and Lin28b Inhibit let-7 MicroRNA Biogenesis by Distinct Mechanisms”, *Cell*, vol. 147, pp. 1066–1079, Nov. 2011.
- [209] J. Yang, L.-H. Hung, T. Licht, S. Kostin, M. Looso, E. Khrameeva, A. Bindereif, A. Schneider, and T. Braun, “RBM24 Is a Major Regulator of Muscle-Specific Alternative Splicing”, *Developmental Cell*, vol. 31, pp. 87–99, Oct. 2014.
- [210] O. Nakagawa, M. Arnold, M. Nakagawa, H. Hamada, J. M. Shelton, H. Kusano, T. M. Harris, G. Childs, K. P. Campbell, J. A. Richardson, I. Nishino, and E. N. Olson, “Centronuclear myopathy in mice lacking a novel muscle-specific protein



- kinase transcriptionally regulated by MEF2”, *Genes & Development*, vol. 19, pp. 2066–2077, Sept. 2005.
- [211] E. Shang, G. Salazar, T. E. Crowley, X. Wang, R. A. Lopez, X. Wang, and D. J. Wolgemuth, “Identification of unique, differentiation stage-specific patterns of expression of the bromodomain-containing genes Brd2, Brd3, Brd4, and Brdt in the mouse testis”, *Gene Expression Patterns*, vol. 4, pp. 513–519, Sept. 2004.
- [212] A. Alexandrov, D. Colognori, M.-D. Shu, and J. A. Steitz, “Human spliceosomal protein CWC22 plays a role in coupling splicing to exon junction complex deposition and nonsense-mediated decay”, *Proceedings of the National Academy of Sciences*, vol. 109, pp. 21313–21318, Dec. 2012.
- [213] H. J. Okano and R. B. Darnell, “A Hierarchy of Hu RNA Binding Proteins in Developing and Adult Neurons”, *Journal of Neuroscience*, vol. 17, pp. 3024–3037, May 1997.
- [214] T. W. Bebee, J. W. Park, K. I. Sheridan, C. C. Warzecha, B. W. Cieply, A. M. Rohacek, Y. Xing, and R. P. Carstens, “The splicing regulators Esrp1 and Esrp2 direct an epithelial splicing program essential for mammalian development”, *eLife*, vol. 4, p. e08954, 2015.
- [215] T. Iijima, Y. Iijima, H. Witte, and P. Scheiffele, “Neuronal cell type-specific alternative splicing is regulated by the KH domain protein SLM1”, *J Cell Biol*, vol. 204, pp. 331–342, Feb. 2014.
- [216] P. Konieczny, E. Stepniak-Konieczna, and K. Sobczak, “MBNL proteins and their target RNAs, interaction and splicing regulation”, *Nucleic Acids Research*, vol. 42, pp. 10873–10887, Sept. 2014.
- [217] Y. Katz, F. Li, N. J. Lambert, E. S. Sokol, W.-L. Tam, A. W. Cheng, E. M. Airoidi, C. J. Lengner, P. B. Gupta, Z. Yu, R. Jaenisch, and C. B. Burge, “Musashi proteins are post-transcriptional regulators of the epithelial-luminal cell state”, *eLife*, vol. 3, p. e03915, Nov. 2014.

- [218] D. Murphy, B. Cieply, R. Carstens, V. Ramamurthy, and P. Stoilov, “The Musashi 1 Controls the Splicing of Photoreceptor-Specific Exons in the Vertebrate Retina”, *PLOS Genet*, vol. 12, p. e1006256, Aug. 2016.
- [219] S. Pedrotti, R. Busà, C. Compagnucci, and C. Sette, “The RNA recognition motif protein RBM11 is a novel tissue-specific splicing regulator”, *Nucleic Acids Research*, vol. 40, pp. 1021–1032, Feb. 2012.
- [220] M. Elizalde, R. Urtasun, M. Azkona, M. U. Latasa, S. Goñi, O. García-Irigoyen, I. Uriarte, V. Segura, M. Collantes, M. D. Scala, A. Lujambio, J. Prieto, M. A. Ávila, and C. Berasain, “Splicing regulator SLU7 is essential for maintaining liver homeostasis”, *The Journal of Clinical Investigation*, vol. 124, pp. 2909–2920, July 2014.
- [221] M.-S. Lee, Y.-S. Lin, Y.-F. Deng, W.-T. Hsu, C.-C. Shen, Y.-H. Cheng, Y.-T. Huang, and C. Li, “Modulation of alternative splicing by expression of small nuclear ribonucleoprotein polypeptide N”, *FEBS Journal*, vol. 281, pp. 5194–5207, Dec. 2014.
- [222] M. Quesnel-Vallières, M. Irimia, S. P. Cordes, and B. J. Blencowe, “Essential roles for the splicing regulator nSR100/SRRM4 during nervous system development”, *Genes & Development*, vol. 29, pp. 746–759, Apr. 2015.
- [223] F. Rizzo, M. Nizzardo, S. Vashisht, E. Molteni, V. Melzi, M. Taiana, S. Salani, P. Santonicola, E. Di Schiavi, M. Bucchia, A. Bordoni, I. Faravelli, N. Bresolin, G. P. Comi, U. Pozzoli, and S. Corti, “Key role of SMN/SYNCRIP and RNA-Motif 7 in spinal muscular atrophy: RNA-Seq and motif analysis of human motor neurons”, *Brain*, vol. 142, pp. 276–294, Feb. 2019.
- [224] T. J. Gingerich, D. J. Stumpo, W. S. Lai, T. A. Randall, S. J. Steppan, and P. J. Blackshear, “Emergence and evolution of Zfp36l3”, *Molecular Phylogenetics and Evolution*, vol. 94, pp. 518–530, Jan. 2016.

- [225] C. Schwärzler, S. Oliferenko, and U. Günthert, “Variant isoforms of CD44 are required in early thymocyte development”, *European Journal of Immunology*, vol. 31, no. 10, pp. 2997–3005, 2001.
- [226] O. Medina-Contreras, G. Soldevila, G. Patiño-Lopez, E. Canche-Pool, R. Vallero-Rios, and V. Ortiz-Navarrete, “Role of CRTAM during mouse early T lymphocytes development”, *Developmental & Comparative Immunology*, vol. 34, pp. 196–202, Feb. 2010.
- [227] J. Kirchner, K. A. Forbush, and M. J. Bevan, “Identification and Characterization of Thymus LIM Protein: Targeted Disruption Reduces Thymus Cellularity”, *Molecular and Cellular Biology*, vol. 21, pp. 8592–8604, Dec. 2001.
- [228] E. M. Kernfeld, R. M. J. Genga, K. Neherin, M. E. Magaletta, P. Xu, and R. Maehr, “A Single-Cell Transcriptomic Atlas of Thymus Organogenesis Resolves Cell Types and Developmental Maturation”, *Immunity*, June 2018.
- [229] A. K. C. Ulrich and M. C. Wahl, “Human MFAP1 is a cryptic ortholog of the *Saccharomyces cerevisiae* Spp381 splicing factor”, *BMC Evolutionary Biology*, vol. 17, p. 91, Mar. 2017.
- [230] A. G. Jacob and C. W. J. Smith, “Intron retention as a component of regulated gene expression programs”, *Human Genetics*, vol. 136, pp. 1043–1057, Sept. 2017.
- [231] U. Braunschweig, N. L. Barbosa-Morais, Q. Pan, E. N. Nachman, B. Alipanahi, T. Gonatopoulos-Pournatzis, B. Frey, M. Irimia, and B. J. Blencowe, “Widespread intron retention in mammals functionally tunes transcriptomes”, *Genome Research*, vol. 24, pp. 1774–1786, Nov. 2014.
- [232] H. Pimentel, M. Parra, S. L. Gee, N. Mohandas, L. Pachter, and J. G. Conboy, “A dynamic intron retention program enriched in RNA processing genes regulates gene expression during terminal erythropoiesis”, *Nucleic Acids Research*, vol. 44, pp. 838–851, Jan. 2016.

- [233] T. Ni, W. Yang, M. Han, Y. Zhang, T. Shen, H. Nie, Z. Zhou, Y. Dai, Y. Yang, P. Liu, K. Cui, Z. Zeng, Y. Tian, B. Zhou, G. Wei, K. Zhao, W. Peng, and J. Zhu, “Global intron retention mediated gene regulation during CD4<sup>+</sup> T cell activation”, *Nucleic Acids Research*, vol. 44, pp. 6817–6829, Aug. 2016.
- [234] Y. Song, O. B. Botvinnik, M. T. Lovci, B. Kakaradov, P. Liu, J. L. Xu, and G. W. Yeo, “Single-Cell Alternative Splicing Analysis with Expedition Reveals Splicing Dynamics during Neuron Differentiation”, *Molecular Cell*, vol. 67, pp. 148–161.e5, July 2017.
- [235] Y. Huang and G. Sanguinetti, “BRIE: transcriptome-wide splicing quantification in single cells”, *Genome Biology*, vol. 18, Dec. 2017.
- [236] J. G. Underwood, P. L. Boutz, J. D. Dougherty, P. Stoilov, and D. L. Black, “Homologues of the *Caenorhabditis elegans* Fox-1 Protein Are Neuronal Splicing Regulators in Mammals”, *Molecular and Cellular Biology*, vol. 25, pp. 10005–10016, Nov. 2005.
- [237] S. Pedrotti, J. Giudice, A. Dagnino-Acosta, M. Knoblauch, R. K. Singh, A. Hanna, Q. Mo, J. Hicks, S. Hamilton, and T. A. Cooper, “The RNA-binding protein Rbfox1 regulates splicing required for skeletal muscle structure and function”, *Human Molecular Genetics*, vol. 24, pp. 2360–2374, Apr. 2015.
- [238] K. K. Kim, Y. C. Kim, R. S. Adelstein, and S. Kawamoto, “Fox-3 and PSF interact to activate neural cell-specific alternative splicing”, *Nucleic Acids Research*, vol. 39, pp. 3064–3078, Apr. 2011.
- [239] S. Nakahata and S. Kawamoto, “Tissue-dependent isoforms of mammalian Fox-1 homologs are associated with tissue-specific splicing activities”, *Nucleic Acids Research*, vol. 33, pp. 2078–2089, Jan. 2005.
- [240] Y. Jin, H. Suzuki, S. Maegawa, H. Endo, S. Sugano, K. Hashimoto, K. Yasuda, and K. Inoue, “A vertebrate RNA-binding protein Fox-1 regulates tissue-specific splicing

- via the pentanucleotide GCAUG”, *The EMBO Journal*, vol. 22, pp. 905–912, Feb. 2003.
- [241] S. D. Auweter, R. Fasan, L. Reymond, J. G. Underwood, D. L. Black, S. Pitsch, and F. H.-T. Allain, “Molecular basis of RNA recognition by the human alternative splicing factor Fox-1”, *The EMBO Journal*, vol. 25, pp. 163–173, Jan. 2006.
- [242] J. L. Ponthier, C. Schluepen, W. Chen, R. A. Lersch, S. L. Gee, V. C. Hou, A. J. Lo, S. A. Short, J. A. Chasis, J. C. Winkelmann, and J. G. Conboy, “Fox-2 Splicing Factor Binds to a Conserved Intron Motif to Promote Inclusion of Protein 4.1r Alternative Exon 16”, *Journal of Biological Chemistry*, vol. 281, pp. 12468–12474, May 2006.
- [243] J. Sebat, B. Lakshmi, D. Malhotra, J. Troge, C. Lese-Martin, T. Walsh, B. Yamrom, S. Yoon, A. Krasnitz, J. Kendall, A. Leotta, D. Pai, R. Zhang, Y.-H. Lee, J. Hicks, S. J. Spence, A. T. Lee, K. Puura, T. Lehtimäki, D. Ledbetter, P. K. Gregersen, J. Bregman, J. S. Sutcliffe, V. Jobanputra, W. Chung, D. Warburton, M.-C. King, D. Skuse, D. H. Geschwind, T. C. Gilliam, K. Ye, and M. Wigler, “Strong Association of De Novo Copy Number Mutations with Autism”, *Science*, vol. 316, pp. 445–449, Apr. 2007.
- [244] G. W. Yeo, X. Xu, T. Y. Liang, A. R. Muotri, C. T. Carson, N. G. Coufal, and F. H. Gage, “Alternative splicing events identified in human embryonic stem cells and neural progenitors”, *PLoS Comput Biol*, vol. 3, 2007.
- [245] A. E. McKee, E. Minet, C. Stern, S. Riahi, C. D. Stiles, and P. A. Silver, “A genome-wide in situ hybridization map of RNA-binding proteins reveals anatomically restricted expression in the developing mouse brain”, *BMC Developmental Biology*, vol. 5, p. 14, July 2005.
- [246] T.-R. Kiehl, H. Shibata, T. Vo, D. P. Huynh, and S.-M. Pulst, “Identification and expression of a mouse ortholog of A2bp1”, *Mammalian Genome*, vol. 12, pp. 595–601, Aug. 2001.

- [247] C. Zhang, Z. Zhang, J. Castle, S. Sun, J. Johnson, A. R. Krainer, and M. Q. Zhang, “Defining the regulatory network of the tissue-specific splicing factors Fox-1 and Fox-2”, *Genes & Development*, vol. 22, pp. 2550–2563, Sept. 2008.
- [248] Y.-C. Chang, Y.-C. Yang, C.-P. Tien, C.-J. Yang, and M. Hsiao, “Roles of Aldolase Family Genes in Human Cancers and Diseases”, *Trends in Endocrinology & Metabolism*, vol. 29, pp. 549–559, Aug. 2018.
- [249] R. Barresi and K. P. Campbell, “Dystroglycan: from biosynthesis to pathogenesis of human disease”, *Journal of Cell Science*, vol. 119, pp. 199–207, Jan. 2006.
- [250] P. Moreno-Layseca, J. Icha, H. Hamidi, and J. Ivaska, “Integrin trafficking in cells and tissues”, *Nature Cell Biology*, vol. 21, p. 122, Feb. 2019.
- [251] A. S. Kristensen, J. Andersen, T. N. Jørgensen, L. Sørensen, J. Eriksen, C. J. Loland, K. Strømgaard, and U. Gether, “SLC6 Neurotransmitter Transporters: Structure, Function, and Regulation”, *Pharmacological Reviews*, vol. 63, pp. 585–640, Sept. 2011.
- [252] A. M. Pajor, “Molecular properties of the SLC13 family of dicarboxylate and sulfate transporters”, *Pflügers Archiv*, vol. 451, pp. 597–605, Feb. 2006.
- [253] J.-M. Revest, R. K. Suniara, K. Kerr, J. J. T. Owen, and C. Dickson, “Development of the Thymus Requires Signaling Through the Fibroblast Growth Factor Receptor R2-IIIb”, *The Journal of Immunology*, vol. 167, pp. 1954–1961, Aug. 2001.
- [254] M. K. Hajihosseini, S. Wilson, L. De Moerlooze, and C. Dickson, “A splicing switch and gain-of-function mutation in FgfR2-IIIC hemizygotes causes Apert/Pfeiffer-syndrome-like phenotypes”, *Proceedings of the National Academy of Sciences*, vol. 98, pp. 3855–3860, Mar. 2001.
- [255] G. W. Yeo, N. G. Coufal, T. Y. Liang, G. E. Peng, X. . D. Fu, and F. H. Gage, “An RNA code for the FOX2 splicing regulator revealed by mapping RNA-protein interactions in stem cells”, *Nat Struct Mol Biol*, vol. 16, 2009.

- [256] D. D. Licatalosi, M. Yano, J. J. Fak, A. Mele, S. E. Grabinski, C. Zhang, and R. B. Darnell, “Ptbp2 represses adult-specific splicing to regulate the generation of neuronal precursors in the embryonic brain”, *Genes & Development*, vol. 26, pp. 1626–1642, July 2012.
- [257] M. M. Hannigan, L. L. Zagore, and D. D. Licatalosi, “Ptbp2 Controls an Alternative Splicing Network Required for Cell Communication during Spermatogenesis”, *Cell Reports*, vol. 19, pp. 2598–2612, June 2017.
- [258] A. Kolodziejczyk, J. K. Kim, V. Svensson, J. Marioni, and S. Teichmann, “The Technology and Biology of Single-Cell RNA Sequencing”, *Molecular Cell*, vol. 58, pp. 610–620, May 2015.
- [259] M. C. Wahl, C. L. Will, and R. Lührmann, “The Spliceosome: Design Principles of a Dynamic RNP Machine”, *Cell*, vol. 136, pp. 701–718, Feb. 2009.
- [260] M. Xu, Y. A. Xie, H. Abouzeid, C. T. Gordon, A. Fiorentino, Z. Sun, A. Lehman, I. S. Osman, R. Dharmat, R. Riveiro-Alvarez, L. Bapst-Wicht, D. Babino, G. Arno, V. Busetto, L. Zhao, H. Li, M. A. Lopez-Martinez, L. F. Azevedo, L. Hubert, N. Pontikos, A. Eblimit, I. Lorda-Sanchez, V. Kheir, V. Plagnol, M. Oufadem, Z. T. Soens, L. Yang, C. Bole-Feysot, R. Pfundt, N. Allaman-Pillet, P. Nitschké, M. E. Cheetham, S. Lyonnet, S. A. Agrawal, H. Li, G. Pinton, M. Michaelides, C. Besmond, Y. Li, Z. Yuan, J. von Lintig, A. R. Webster, H. Le Hir, P. Stoilov, J. Amiel, A. J. Hardcastle, C. Ayuso, R. Sui, R. Chen, R. Allikmets, D. F. Schorderet, G. Black, G. Hall, R. Gillespie, S. Ramsden, F. Manson, P. Sergouniotis, C. Inglehearn, C. Toomes, M. Ali, M. McKibbin, J. Poulter, E. Lord, A. Nemeth, S. Halford, S. Downes, and J. Yu, “Mutations in the Spliceosome Component CWC27 Cause Retinal Degeneration with or without Additional Developmental Anomalies”, *The American Journal of Human Genetics*, vol. 100, pp. 592–604, Apr. 2017.
- [261] M. De Simone, G. Rossetti, and M. Pagani, “Single Cell T Cell Receptor Sequencing: Techniques and Future Challenges”, *Frontiers in Immunology*, vol. 9, 2018.

- [262] E. R. Unanue, “The Secrets of the Class II MHC Peptidome Start To Be Revealed”, *The Journal of Immunology*, vol. 196, pp. 939–940, Feb. 2016.
- [263] E. Adamopoulou, S. Tenzer, N. Hillen, P. Klug, I. A. Rota, S. Tietz, M. Gebhardt, S. Stevanovic, H. Schild, E. Tolosa, A. Melms, and C. Stoeckle, “Exploring the MHC-peptide matrix of central tolerance in the human thymus”, *Nature Communications*, vol. 4, p. 2039, June 2013.
- [264] J. A. Collado, I. Alvarez, M. T. Ciudad, G. Espinosa, F. Canals, R. Pujol-Borrell, M. Carrascal, J. Abian, and D. Jaraquemada, “Composition of the HLA-DR-associated human thymus peptidome”, *European Journal of Immunology*, vol. 43, no. 9, pp. 2273–2282, 2013.
- [265] I. A. Roundtree, M. E. Evans, T. Pan, and C. He, “Dynamic RNA Modifications in Gene Expression Regulation”, *Cell*, vol. 169, pp. 1187–1200, June 2017.
- [266] S. H. Choi and K. D. Meyer, “Acetylation takes aim at mRNA”, *Nature Structural & Molecular Biology*, vol. 25, p. 1067, Dec. 2018.
- [267] A. Damianov, Y. Ying, C.-H. Lin, J.-A. Lee, D. Tran, A. A. Vashisht, E. Bahrami-Samani, Y. Xing, K. C. Martin, J. A. Wohlschlegel, and D. L. Black, “Rbfox Proteins Regulate Splicing as Part of a Large Multiprotein Complex LASR”, *Cell*, vol. 165, pp. 606–619, Apr. 2016.
- [268] B. Amir-Ahmady, P. L. Boutz, V. Markovtsov, M. L. Phillips, and D. L. Black, “Exon repression by polypyrimidine tract binding protein”, *RNA*, vol. 11, pp. 699–716, May 2005.
- [269] M. Sekai, J. Wang, N. Minato, and Y. Hamazaki, “An improved clonogenic culture method for thymic epithelial cells”, *Journal of Immunological Methods*, Feb. 2019.
- [270] M. Albertsen, M. Teperek, G. Elholm, E.-M. Füchtbauer, and K. Lykke-Hartmann, “Localization and Differential Expression of the Krüppel-Associated Box Zinc Finger Proteins 1 and 54 in Early Mouse Development”, *DNA and Cell Biology*, vol. 29, pp. 589–601, July 2010.



# Appendices





## Appendix - Computational tools & Experimental materials

All software packages and all R packages including their version numbers and references are given in Table A.1 and Table A.3, respectively.

**Table A.1:** List of software packages with corresponding version numbers.

| <b>Software</b> | <b>Version</b>         |
|-----------------|------------------------|
| Albacore        | v2.3.1                 |
| bedtools        | v2.25.0                |
| Cufflinks       | v2.2.1                 |
| DESeq2          | v1.18.1                |
| fastx_trimmer   | v0.0.14                |
| featureCounts   | v1.6.3                 |
| gffcompare      | v.0.10.6               |
| HISAT2          | v2.1.1                 |
| jupyter         | v4.2.3                 |
| Kallisto        | v0.43.1                |
| maptools        | v3.2                   |
| Minimap2        | v2.14-r883             |
| MISO            | 0.5.3                  |
| Nanopack        | v1.12.0                |
| networkx        | v2.1                   |
| pandas          | v0.17.1                |
| Porechop        | v0.2.3                 |
| python          | 2.7.11 or 3.6.2        |
| R               | 3.4.0 or 3.4.2         |
| rMATS           | v3.2.5                 |
| rpy2            | v2.8.6                 |
| salmon          | v0.11.3                |
| samtools        | v1.3.1                 |
| Sjcounts        | v3.1                   |
| Sleuth          | v0.29                  |
| stringTie       | v1.3.3b or 1.3.5 (ONT) |
| SUPPA           | v2.3                   |

**Table A.2:** Overview of quality metrics included in the *scseq* pipeline.

| Metric                     | Information                                                                                                                                                                                                          |
|----------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| cufflinks-no-genes         | Number of genes detected using <i>Cufflinks</i>                                                                                                                                                                      |
| cufflinks-no-genes-pc      | Number of protein-coding genes detected using <i>Cufflinks</i>                                                                                                                                                       |
| featurecounts-no-genes     | Number of genes detected using <i>featureCounts</i>                                                                                                                                                                  |
| featurecounts-no-genes-pc  | Number of protein-coding genes detected using <i>featureCounts</i>                                                                                                                                                   |
| fraction-spike             | Fraction of reads mapping to spike-ins amongst uniquely mapped reads (if present)                                                                                                                                    |
| fraction-spliced           | Fraction of spliced reads amongst uniquely mapped reads                                                                                                                                                              |
| library-size               | Library size computed using <i>EstimateLibraryComplexity</i> ( <i>Picard</i> )                                                                                                                                       |
| median-insert-size         | Median insert size computed using <i>InsertSizeMetrics</i> ( <i>Picard</i> )                                                                                                                                         |
| no-pairs                   | Number of read pairs examined reported by <i>EstimateLibraryComplexity</i>                                                                                                                                           |
| nreads-uniq-map-genome     | Number of reads uniquely mapped to the genome                                                                                                                                                                        |
| nreads-uniq-map-spike      | Number of reads uniquely mapped to spike-in sequences (if present)                                                                                                                                                   |
| pct-adapter                | Computed by <i>AlignmentSummaryMetrics</i> ( <i>Picard</i> ) and defined as: Frequency of unaligned PF reads (passing Illumina's filters) 'that match to a known adapter sequence right from the start of the read.' |
| pct-coding                 | Frequency of PF reads mapping to protein-coding regions, computed by <i>Picard</i> function <i>RnaSeqMetrics</i>                                                                                                     |
| pct-duplication            | Frequency of duplicated reads computed by <i>EstimateLibraryComplexity</i>                                                                                                                                           |
| pct-mrna                   | Frequency of bases mapped to UTRs and coding regions, defined by <i>RnaSeqMetrics</i>                                                                                                                                |
| pct-reads-aligned          | Frequency of PF reads aligned, computed by <i>AlignmentSummaryMetrics</i>                                                                                                                                            |
| pct-pf-reads-aligned-hq    | Frequency of PF reads aligned with high quality (MAPQ > 20), computed by <i>AlignmentSummaryMetrics</i>                                                                                                              |
| pct-reads-aligned-in-pairs | Frequency of reads aligned in pairs, computed by <i>AlignmentSummaryMetrics</i>                                                                                                                                      |
| three-prime-bias           | Three prime bias calculated as part of the <i>scseq</i> pipeline from <i>Picard</i> coverage histogram                                                                                                               |
| total-reads                | Total number of reads reported by <i>AlignmentSummaryMetrics</i>                                                                                                                                                     |

**Table A.3:** List of R packages with corresponding version numbers.

| R package      | Version               |
|----------------|-----------------------|
| AnnotationDbi  | 1.40.0                |
| biomaRt        | 2.34.2                |
| cba            | 0.2.19                |
| circlize       | 0.4.3                 |
| ComplexHeatmap | 1.17.1                |
| DESeq2         | 1.18.1                |
| dplyr          | 0.7.6                 |
| GenomicRanges  | 1.30.3                |
| ggplot2        | 2.2.1/3.0.0 (RStudio) |
| ggrepel        | 0.8.0                 |
| gridExtra      | 2.3                   |
| gsfisher       | 0.2                   |
| Gviz           | 1.22.3                |
| Iranges        | 2.12.0                |
| Matrix         | 1.2.14                |
| openxlsx       | 4.1.0                 |
| plyr           | 1.8.4                 |
| prcomp         |                       |
| pvclust        | 2.0.0                 |
| RColorBrewer   | 1.1.2                 |
| reshape2       | 1.4.3                 |
| Rsamtools      | 1.30.0                |
| RSQLite        | 2.1.1                 |
| Rtsne          | 0.13                  |
| Seurat         | 2.3.4                 |
| Sleuth         | 0.29.0                |
| tidyr          | 0.8.1                 |
| viridis        | 0.5.1                 |

**Table A.4:** Filters applied to *gsfisher* output before manual filtering for the respective figures. The genes containing transcripts from the sets indicated in the first columns were used for geneset enrichment testing.

| Geneset filters                                                                                    |                  |         |            |
|----------------------------------------------------------------------------------------------------|------------------|---------|------------|
| Geneset based on:                                                                                  | adjusted p-value | N genes | Odds ratio |
| mT&T novel transcripts                                                                             | 0.1              | 2       | 1          |
| Retained introns                                                                                   | (p-value) 0.01   | 2       | 1.5        |
| ONT novel transcripts                                                                              | 0.05             | 3       | 2          |
| <i>Aire</i> -regulated transcripts                                                                 | 0.05             | 2       | 2          |
| Genes affected by differential splicing in mature mTEC with loss of <i>Rbfox1</i> or <i>Rbfox2</i> | 0.05             | 2       | 2          |

Dilutions and detailed information on all antibodies used for immunofluorescence experiments are given in Tables A.5 and A.6. Please note that UEA-1 is a lection and was in-house labelled with Cy5.

**Table A.5:** List of primary antibodies used for immunofluorescent imaging.

| Marker             | Clone   | Company                  | Catalog No | Dilution |
|--------------------|---------|--------------------------|------------|----------|
| Aire               | 5H12    | eBioscience              | 14-5934    | 1:500    |
| cytokeratin 5 (K5) | EP1601Y | AbCAM                    | ab52635    | 1:300    |
| RRM                | -       | provided by D.Black      | -          | 1:500    |
| UEA-1              | -       | in-house labelled lectin | -          | 1:150    |

**Table A.6:** List of secondary antibodies used for immunofluorescent imaging. All secondary antibodies were used in a 1:500 dilution. AF = Alexa Fluorophore.

| Species              | Fluorophore | Company    | Catalog No |
|----------------------|-------------|------------|------------|
| goat $\alpha$ rat    | AF488       | Invitrogen | A-11006    |
| goat $\alpha$ rabbit | AF555       | Invitrogen | A27039     |

All antibodies used in flow cytometry experiments are listed in Table A.7. Please note that UEA-1 is a lection and was in-house labelled with Cy5.

All primers used are listed in Table A.8.

The buffers or reagents prepared in-house and the commercially available reagents and kits are listed in Tables A.9 and A.10.

**Table A.7:** List of FACS antibodies used for experiments. AF = Alexa Fluorophore.

| Marker               | Clone       | Fluorochrome  | Company       | Catalog No | Dilution |
|----------------------|-------------|---------------|---------------|------------|----------|
| CD45                 | 30-F11      | AF700         | BioLegend     | 103128     | 1:1000   |
| EpCAM                | G8.8        | PerCp-Cy5.5   | BioLegend     | 118220     | 1:1000   |
| Ly51                 | 6C3         | PE            | BioLegend     | 108308     | 1:200    |
| UEA1                 | -           | Cy5           | in-house made | -          | 1:500    |
| MHC Class II (IA/IE) | M5/114.15.2 | APC-Cy7       | BioLegend     | 107628     | 1:1000   |
| MHC Class II (IA/IE) | M5/114.15.2 | BV421         | BioLegend     | 107631     | 1:1000   |
| CD80                 | 16-10A1     | PE-Cy5        | BioLegend     | 104712     | 1:1000   |
| CD86                 | GL-1        | PE-Cy7        | BioLegend     | B238554    | 1:1000   |
| AIRE                 | 5H12        | AF488         | Invitrogen    | 53-5934-82 | 1:1000   |
| AIRE                 | 5H12        | eFluor 660    | eBioscience   | 50-5934    | 1:1000   |
| AIRE                 | 5H12        | AF647         | eBioscience   | 51-5934-82 | 1:1000   |
| CD8a                 | 53-6.7      | AF700         | BioLegend     | 100730     | 1:1000   |
| CD3e                 | 145-2C11    | PerCp-Cy5.5   | eBioscience   | 45-0031    | 1:500    |
| TCRb                 | H57-597     | PE            | Invitrogen    | 12-5961-83 | 1:1000   |
| TCRb                 | H57-597     | PE-Cy5        | BioLegend     | 109210     | 1:1000   |
| CD8                  | 53-6.7      | PE-Cy7        | eBioscience   | 25-0081    | 1:1000   |
| CD8a                 | 53-6.7      | AF647         | BioLegend     | 100724     | 1:1000   |
| CD4                  | ??          | PE Fluor 6.10 | Invitrogen    | 61-0042-82 | 1:1000   |



**Table A.8:** List of primers used for genotyping and genomic excision PCR. All primers were diluted to a stock concentration of 10  $\mu$ M. M = molar, R = reverse, F = forward.

| Gene name         | Primer sequence              | Comment                             |
|-------------------|------------------------------|-------------------------------------|
| Rbfox1_loxP_exc_F | ATGCCCATGCGAGTGAAAAAT        | genotyping; genomic excision        |
| Rbfox1_loxP_R     | TGCAGCACATTGAAACCTTC         | genotyping (product: 294 or 397 bp) |
| Rbfox1_exc_R      | AGCCAGTCAGCTGGAGTGAT         | genomic excision (product: 493 bp)  |
| Rbfox2_loxP_F     | CAGAAA CAAGAAAGGCGCTCACTTCAG | genotyping (product: 320 or 453 bp) |
| Rbfox2_loxP_exc_R | CTC TGACTTATACATGCACCTC      | genotyping; genomic excision        |
| Rbfox2_exc_F      | GATGGGGTTTTATATG TGGAG       | genomic excision (product: 350 bp)  |
| FoxN1_iCre_F      | TCTGATGAAGTCAGGAAGAACC       | genotyping (product: 500 bp)        |
| FoxN1_iCre_R      | GAGA TGTCCTTCACTCTGA TTG     | genotyping (product: 500 bp)        |

**Table A.9:** List of in-house buffers. M = molar.

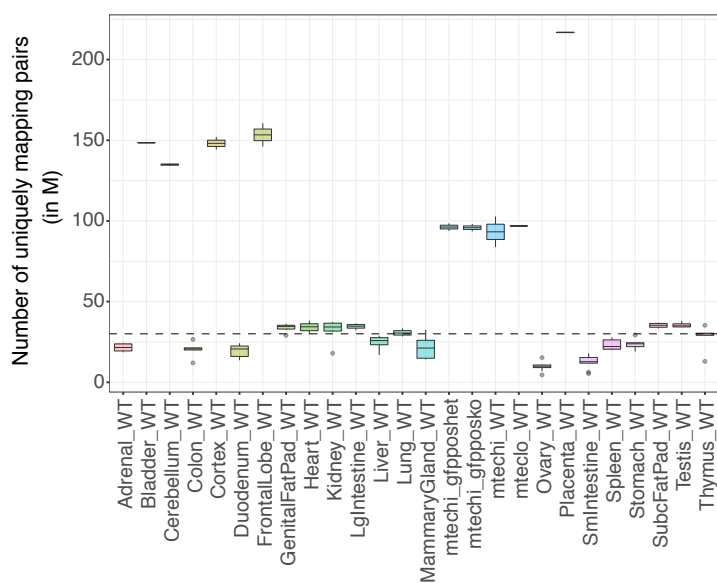
| Buffer                      | Component                               | Concentration | Company |
|-----------------------------|-----------------------------------------|---------------|---------|
| Lysis buffer                | Tris                                    | 100 mM        | Sigma   |
|                             | Ethylenediaminetetraacetic acid (EDTA)  | 5mM           | Sigma   |
|                             | Sodium Chloride (NaCl)                  | 200 mM        | VWR     |
|                             | Sodium Dodecyl Sulfate (SDS)            | 0.2 %         | Sigma   |
|                             | Proteinase K                            | 20 mg/ml      | Sigma   |
| Tris-EDTA (TE) buffer       | Tris                                    | 10 mM         | Sigma   |
|                             | EDTA                                    | 0.01 mM       | Sigma   |
| Tissue digestion buffer     | PBS                                     | 1x            | Sigma   |
|                             | Liberase                                | 2.5 mg/ml     | Roche   |
|                             | DNAse                                   | 10 mg/ml      | Roche   |
| MACS buffer                 | PBS                                     | 1x            | Sigma   |
|                             | Bovine Serum Albumin (BSA)              | 0.5 %         | Sigma   |
|                             | EDTA                                    | 5mM           | Sigma   |
| FACS buffer                 | PBS                                     | 1x            | Sigma   |
|                             | fetal calve serum (FCS) or BSA          | 2 %           | Sigma   |
| Red blood cell lysis buffer | Ammonium chloride<br>NH <sup>4</sup> Cl | 1.66 %        | Sigma   |

**Table A.10:** List of commercially available kits or reagents.

| Name of reagent or kit                 | Company                                   |
|----------------------------------------|-------------------------------------------|
| 1 kb plus ladder                       | NEB                                       |
| 1 ml syringes                          | Terumo                                    |
| 100 µm filter                          | Greiner bio-one                           |
| 30 µm filter                           | Sysmex                                    |
| 40 µm filter paper                     | Sefar                                     |
| Agarose                                | Invitrogen                                |
| anti-CD45 beads                        | Miltenyl Biotec                           |
| Bioanalyzer RNA 6000 Pico              | Agilent                                   |
| cDNA-PCR Sequencing Kit                | Oxford Nanopore Technologies (SQK-PCS108) |
| DAPI                                   | Sigma                                     |
| dNTPs                                  | Sigma                                     |
| ethanol                                | VWR                                       |
| Fix-permeabilisation kit (eBioscience) | Invitrogen                                |
| Fixable AQUA stain kit                 | ThermoFisher Scientific                   |
| GelRed                                 | Biotium                                   |
| Isopropanol                            | VWR                                       |
| Methanol                               | VWR                                       |
| ProLong Gold antifade reagent          | Invitrogen                                |
| OCT                                    | Tissue Trek                               |
| PAP pen                                | Sigma                                     |
| PFA                                    | Sigma                                     |
| Phire Animal Tissue Direct PCR kit     | ThermoFisher Scientific                   |
| RNEasy Micro Plus kit                  | Qiagen                                    |
| SequalPrep Long PCR kit                | ThermoFisher Scientific                   |
| SuperFrost Plus slides                 | VWR                                       |
| Tapestation High Sensitivity RNA       | Agilent                                   |
| Taq polymerase PCR kit                 | Sigma                                     |
| TBE                                    | Sigma                                     |
| Triton-x                               | Sigma                                     |

# B

## Appendix - Chapter 3



**Figure B.1:** Number of uniquely mapping reads per tissue before combining and subsampling replicates from each tissue (trimmed to 76 bp before mapping).

**Table B.1:** Overview of unpublished datasets used for analyses. CGAT = Computational Genomics Analysis and Training programme, KO = knockout, ONT = Oxford Nanopore Technology, pos. = positive, Seq. = sequencing, WT = Wellcome Trust.

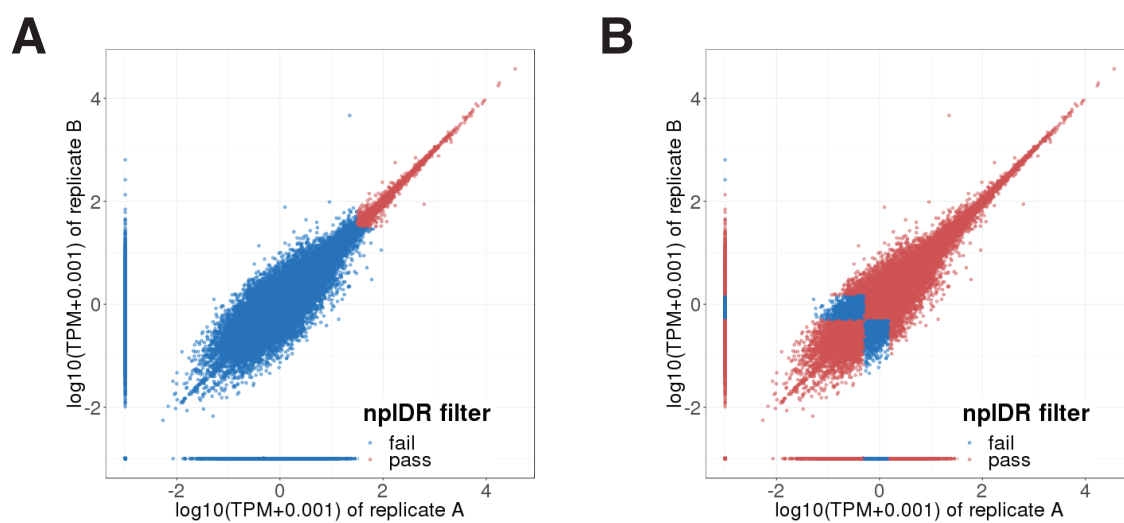
| TEC population                       | Age (weeks) | Genetic modification                         | Cellular phenotype                            | Seq. technology                  | Data generated by                                                  |
|--------------------------------------|-------------|----------------------------------------------|-----------------------------------------------|----------------------------------|--------------------------------------------------------------------|
| immature mTEC (2 replicates)         | 4-6         | wildtype                                     | CD45- EpCAM+<br>Ly51- UEA+<br>MHCII low       | Illumina TruSeq                  | N. Shikama-Dorn, S. Zuklys<br>K. Hafen; grant provided by CGAT/MCR |
| mature mTEC (2 replicates)           | 4-6         | wildtype                                     | CD45- EpCAM+<br>Ly51- UEA+<br>MHCII high      | Illumina TruSeq                  | N. Shikama-Dorn, S. Zuklys<br>K. Hafen; grant provided by CGAT/MCR |
| mature Aire-pos. mTEC (2 replicates) | 4-6         | heterozygous Aire-knockout (GFP/AIRE)        | CD45- EpCAM+<br>Ly51- UEA+<br>MHCII high GFP+ | Illumina TruSeq                  | N. Shikama-Dorn, S. Zuklys<br>K. Hafen; grant provided by CGAT/MCR |
| mature Aire-KO mTEC (2 replicates)   | 4-6         | homozygous Aire-knockout (GFP/GFP)           | CD45- EpCAM+<br>Ly51- UEA+<br>MHCII high GFP+ | Illumina TruSeq                  | N. Shikama-Dorn, S. Zuklys<br>K. Hafen; grant provided by CGAT/MCR |
| immature mTEC                        | 4           | wildtype                                     | CD45- EpCAM+<br>Ly51- UEA+<br>MHCII low       | SmartSeq2 + ONT                  | K.Jansen<br>grant provided by WT                                   |
| mature mTEC                          | 4           | wildtype                                     | CD45- EpCAM+<br>Ly51- UEA+<br>MHCII high      | SmartSeq2 + ONT                  | K.Jansen<br>grant provided by WT                                   |
| immature mTEC (3 replicates)         | 4           | wildtype                                     | CD45- EpCAM+<br>Ly51- UEA+<br>CD80- CD86-     | cDNA ONT library protocol + ONT  | K.Jansen<br>grant provided by WT                                   |
| mature mTEC (3 replicates)           | 4           | wildtype                                     | CD45- EpCAM+<br>Ly51- UEA+<br>CD80+ CD86+     | cDNA ONT library protocol + ONT  | K.Jansen<br>grant provided by WT                                   |
| mature mTEC (3 replicates)           | 4-5         | Rbfox1 <sup>lox/lox</sup> cre <sup>+/-</sup> | CD45- EpCAM+<br>Ly51- UEA+<br>MHCII high      | NEB stranded protocol + Illumina | K.Jansen<br>grant provided by WT                                   |
| mature mTEC (3 replicates)           | 4-5         | Rbfox1 <sup>lox/lox</sup> cre <sup>-/-</sup> | CD45- EpCAM+<br>Ly51- UEA+<br>MHCII high      | NEB stranded protocol + Illumina | K.Jansen<br>grant provided by WT                                   |
| immature mTEC (3 replicates)         | 4-5         | Rbfox1 <sup>lox/lox</sup> cre <sup>+/-</sup> | CD45- EpCAM+<br>Ly51- UEA+<br>MHCII low       | NEB stranded protocol + Illumina | K.Jansen<br>grant provided by WT                                   |
| immature mTEC (3 replicates)         | 4-5         | Rbfox1 <sup>lox/lox</sup> cre <sup>-/-</sup> | CD45- EpCAM+<br>Ly51- UEA+<br>MHCII low       | NEB stranded protocol + Illumina | K.Jansen<br>grant provided by WT                                   |
| mature mTEC (3 replicates)           | 4-5         | Rbfox2 <sup>lox/lox</sup> cre <sup>+/-</sup> | CD45- EpCAM+<br>Ly51- UEA+<br>MHCII high      | NEB stranded protocol + Illumina | K.Jansen<br>grant provided by WT                                   |
| mature mTEC (3 replicates)           | 4-5         | Rbfox2 <sup>lox/lox</sup> cre <sup>-/-</sup> | CD45- EpCAM+<br>Ly51- UEA+<br>MHCII high      | NEB stranded protocol + Illumina | K.Jansen<br>grant provided by WT                                   |
| immature mTEC (3 replicates)         | 4-5         | Rbfox2 <sup>lox/lox</sup> cre <sup>+/-</sup> | CD45- EpCAM+<br>Ly51- UEA+<br>MHCII low       | NEB stranded protocol + Illumina | K.Jansen<br>grant provided by WT                                   |
| immature mTEC (3 replicates)         | 4-5         | Rbfox2 <sup>lox/lox</sup> cre <sup>-/-</sup> | CD45- EpCAM+<br>Ly51- UEA+<br>MHCII low       | NEB stranded protocol + Illumina | K.Jansen<br>grant provided by WT                                   |

**Table B.2:** Overview of population RNA-sequencing samples before subsampling or merging. Averages across all replicates per tissue are shown. M = million, N = Number, pairs = read pairs, repl. = replicates, uniq. = uniquely.

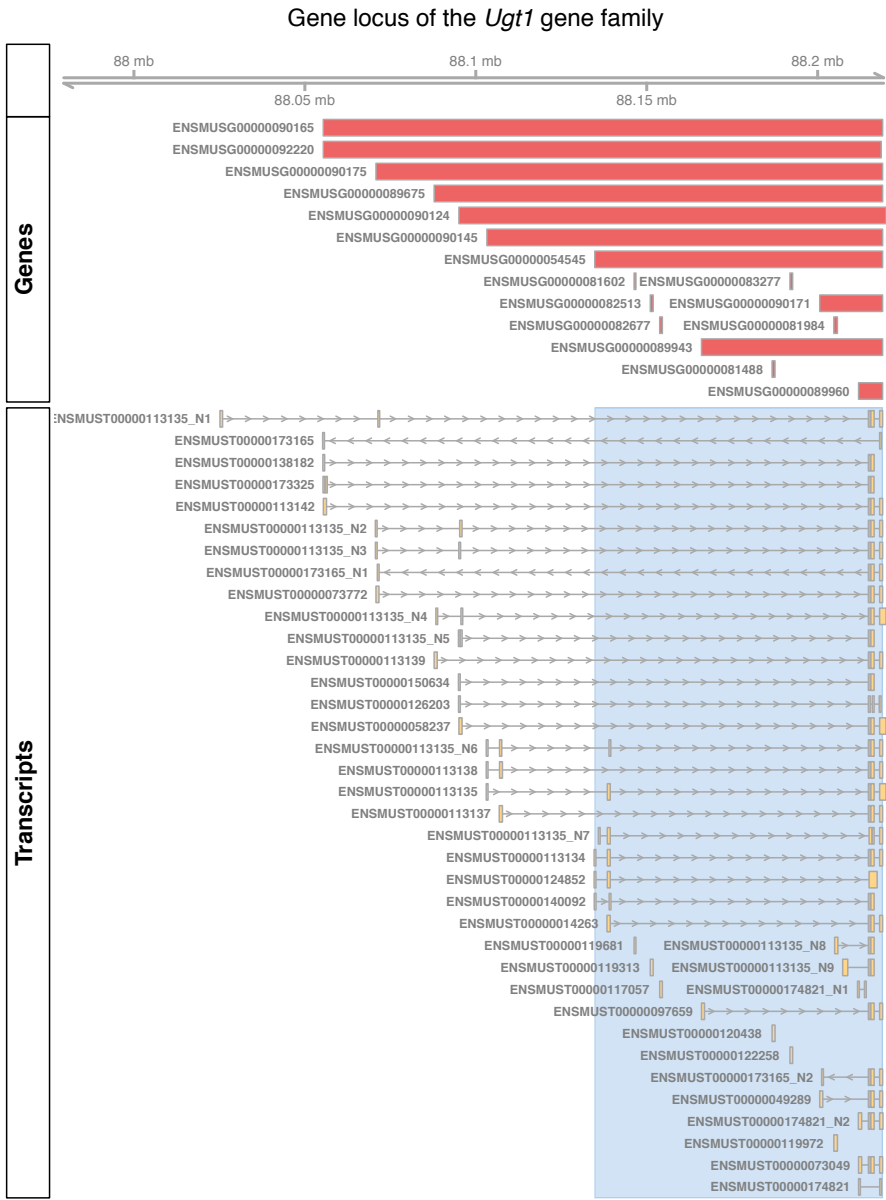
| Tissue           | N pairs     | N uniq. mapping reads | Fastq pairs | N repl. | Possible pairs for repl.(M) |
|------------------|-------------|-----------------------|-------------|---------|-----------------------------|
| Adrenal-WT       | 21558033    | 39341003.83           | 24715354.17 | 6       | 64.67                       |
| Bladder-WT       | 148454775.5 | 263009162.5           | 161561813.5 | 2       | 148.45                      |
| Cerebellum-WT    | 134941016.5 | 239616613             | 147839921   | 2       | 134.94                      |
| Colon-WT         | 20218738.17 | 35835365.33           | 21834292.17 | 6       | 60.66                       |
| Cortex-WT        | 148104847.5 | 262917550             | 160934451   | 2       | 148.1                       |
| Duodenum-WT      | 19362465.14 | 30760462.71           | 22272536.43 | 7       | 58.09                       |
| FrontalLobe-WT   | 153354175   | 275638700             | 172533290   | 2       | 153.35                      |
| GenitalFatPad-WT | 33658515.5  | 61799569              | 36991680.5  | 4       | 67.32                       |
| Heart-WT         | 34052900    | 58375745.67           | 38684199.67 | 3       | 34.05                       |
| Kidney-WT        | 32054940.67 | 56921143.17           | 35179850    | 6       | 96.16                       |
| LgIntestine-WT   | 34562411.75 | 59867196.5            | 37154036.75 | 4       | 69.12                       |
| Liver-WT         | 24693180.5  | 44743773.17           | 27114695.17 | 6       | 74.08                       |
| Lung-WT          | 30713771    | 55780881              | 33266539.75 | 4       | 61.43                       |
| MammaryGland-WT  | 21592801.67 | 38857911.33           | 24500436.33 | 6       | 64.78                       |
| mtechi-gfpposhet | 96196277    | 170388010             | 98023337    | 2       | 96.2                        |
| mtechi-gfpposko  | 95817533.5  | 171084886             | 97294394.5  | 2       | 95.82                       |
| mtechi-WT        | 93256090.5  | 166614895             | 94557085.5  | 2       | 93.26                       |
| mtcilo-WT        | 96901144.5  | 174051847.5           | 98219963.5  | 2       | 96.9                        |
| Ovary-WT         | 9758176.9   | 18218526.7            | 10567705.7  | 10      | 48.79                       |
| Placenta-WT      | 216837397   | 381184211.5           | 235579360.5 | 2       | 216.84                      |
| SmIntestine-WT   | 12668839.6  | 21748013.8            | 13805692.8  | 10      | 63.34                       |
| Spleen-WT        | 23293764.67 | 40780143              | 25432382.67 | 6       | 69.88                       |
| Stomach-WT       | 23726362.5  | 41627148.67           | 28091760.83 | 6       | 71.18                       |
| SubcFatPad-WT    | 35241194.5  | 65805048              | 39291786    | 4       | 70.48                       |
| Testis-WT        | 35567725.75 | 65319494.75           | 37690909.5  | 4       | 71.14                       |
| Thymus-WT        | 27963920.5  | 49591529.5            | 30262828.5  | 6       | 83.89                       |

**Table B.3:** Overview of population RNA-sequencing samples including which replicates were merged per sample (for 2 replicates at 60 M reads). M = million, N = Number, repl. = replicates.

| Tissue             | N repl. | Average depth | Read length | Data source        | Replicate A | Replicate B    |
|--------------------|---------|---------------|-------------|--------------------|-------------|----------------|
| Adrenal            | 6       | 50-70 M reads | 76 bp       | ENCODE             | R1; R5      | R2; R4         |
| Colon              | 6       | 50-70 M reads | 76 bp       | ENCODE             | R1; R4      | R2; R5         |
| Duodenum           | 7       | 50-70 M reads | 76 bp       | ENCODE             | R1; R5      | R2; R3         |
| Genital Fat Pad    | 4       | 50-70 M reads | 76 bp       | ENCODE             | R3          | R4             |
| Heart              | 3       | 50-70 M reads | 76 bp       | ENCODE             | R1          | R3             |
| Kidney             | 6       | 50-70 M reads | 76 bp       | ENCODE             | R1          | R4             |
| Large Intestine    | 4       | 50-70 M reads | 76 bp       | ENCODE             | R3          | R4             |
| Liver              | 6       | 50-70 M reads | 76 bp       | ENCODE             | R1; R6      | R3; R4         |
| Lung               | 4       | 50-70 M reads | 76 bp       | ENCODE             | R1          | R2; R3         |
| Mammary Gland      | 6       | 50-70 M reads | 76 bp       | ENCODE             | R3          | R2; R6         |
| Ovary              | 9       | 50-70 M reads | 76 bp       | ENCODE             | R1; R10; R2 | R4; R5; R6; R7 |
| Small Intestine    | 9       | 50-70 M reads | 76 bp       | ENCODE             | R1; R2      | R5; R6; R7     |
| Spleen             | 6       | 50-70 M reads | 76 bp       | ENCODE             | R1; R2      | R3; R4         |
| Stomach            | 6       | 50-70 M reads | 76 bp       | ENCODE             | R4; R5      | R1; R2         |
| Subc Fat Pad       | 4       | 50-70 M reads | 76 bp       | ENCODE             | R1          | R4             |
| Testis             | 4       | 50-70 M reads | 76 bp       | ENCODE             | R1          | R3             |
| Thymus             | 6       | 50-70 M reads | 76 bp       | ENCODE             | R4          | R2; R3         |
| Frontal Lobe       | 2       | 200 M reads   | 100 bp      | ENCODE             | R1          | R2             |
| Bladder            | 2       | 200 M reads   | 100 bp      | ENCODE             | R1          | R2             |
| Cortex             | 2       | 200 M reads   | 100 bp      | ENCODE             | R1          | R2             |
| Placenta           | 2       | 200 M reads   | 100 bp      | ENCODE             | R1          | R2             |
| Cerebellum         | 2       | 200 M reads   | 100 bp      | ENCODE             | R1          | R2             |
| Mature mTEC        | 2       | 170 M reads   | 100 bp      | Sansom/ Hollaender | R1          | R2             |
| Immature mTEC      | 2       | 170 M reads   | 100 bp      | Sansom/ Hollaender | R1          | R2             |
| Aire-positive mTEC | 2       | 170 M reads   | 100 bp      | Sansom/ Hollaender | R1          | R2             |
| Aire-KO mTEC       | 2       | 170 M reads   | 100 bp      | Sansom/ Hollaender | R1          | R2             |

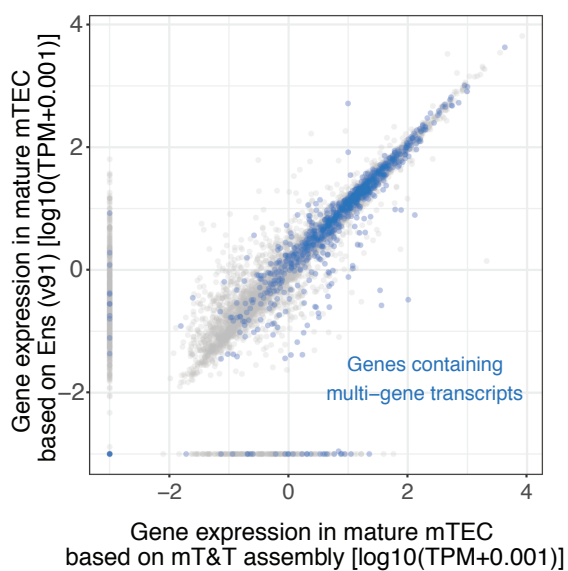


**Figure B.2:** IDR results from other approaches. **A:** Transcripts passing npIDR filter if R package *IDR* is used. **B:** Transcripts passing npIDR filter if scripts published by Pervouchine *et al.* are used.

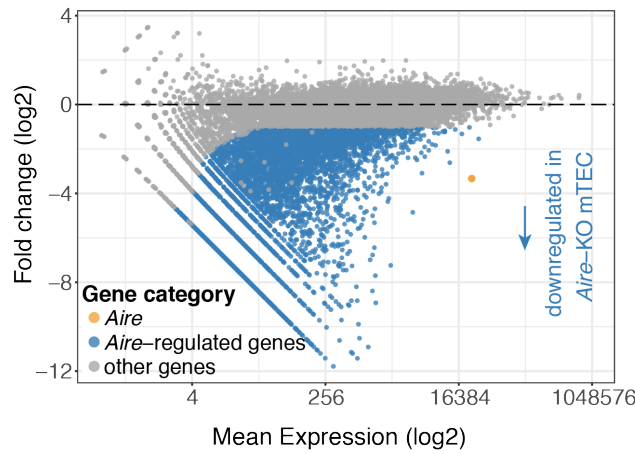


**Figure B.3:** Example of multi-gene transcripts occurring in both Ensembl 91 annotation and mT&T assembly. The gene locus shown contains the *Ugt1* family. The top bar indicates the genomic coordinates, followed by one panel indicating the gene ranges (red). The transcripts panel (bottom) depicts all transcripts in this region.

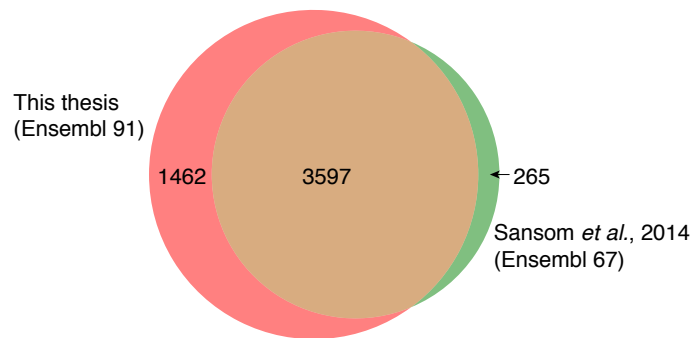




**Figure B.4:** Correlation of gene expression quantified using Ensembl 91 annotation or mT&T assembly for one example tissue (mature mTEC). 10,000 genes were selected randomly. All genes affected by multi-gene transcripts according to the exon graphs in the mT&T assembly but not in the Ensembl annotation were plotted in blue on top of other genes (n=948 genes).



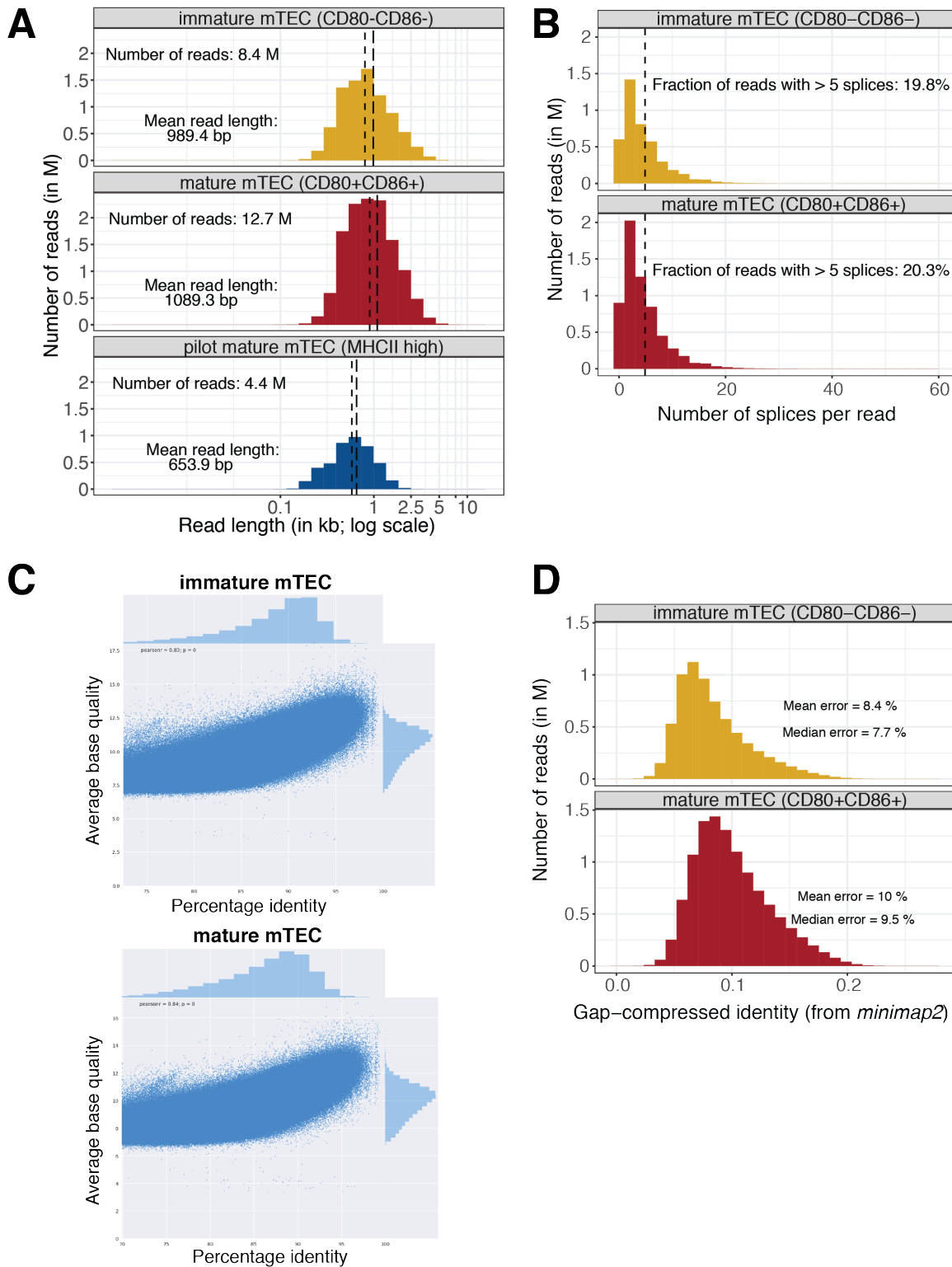
**Figure B.5:** Identification of *Aire*-regulated genes from *DESeq2* analysis. Differential expression of protein-coding genes was analysed between *Aire*-knockout and *Aire*-positive mTEC. Genes with a 2-fold and significant downregulation were defined as *Aire*-regulated genes (Benjamini-Hochberg adjusted p-value < 0.05). The quantification with *featureCounts* was based on Ensembl annotation (v91).



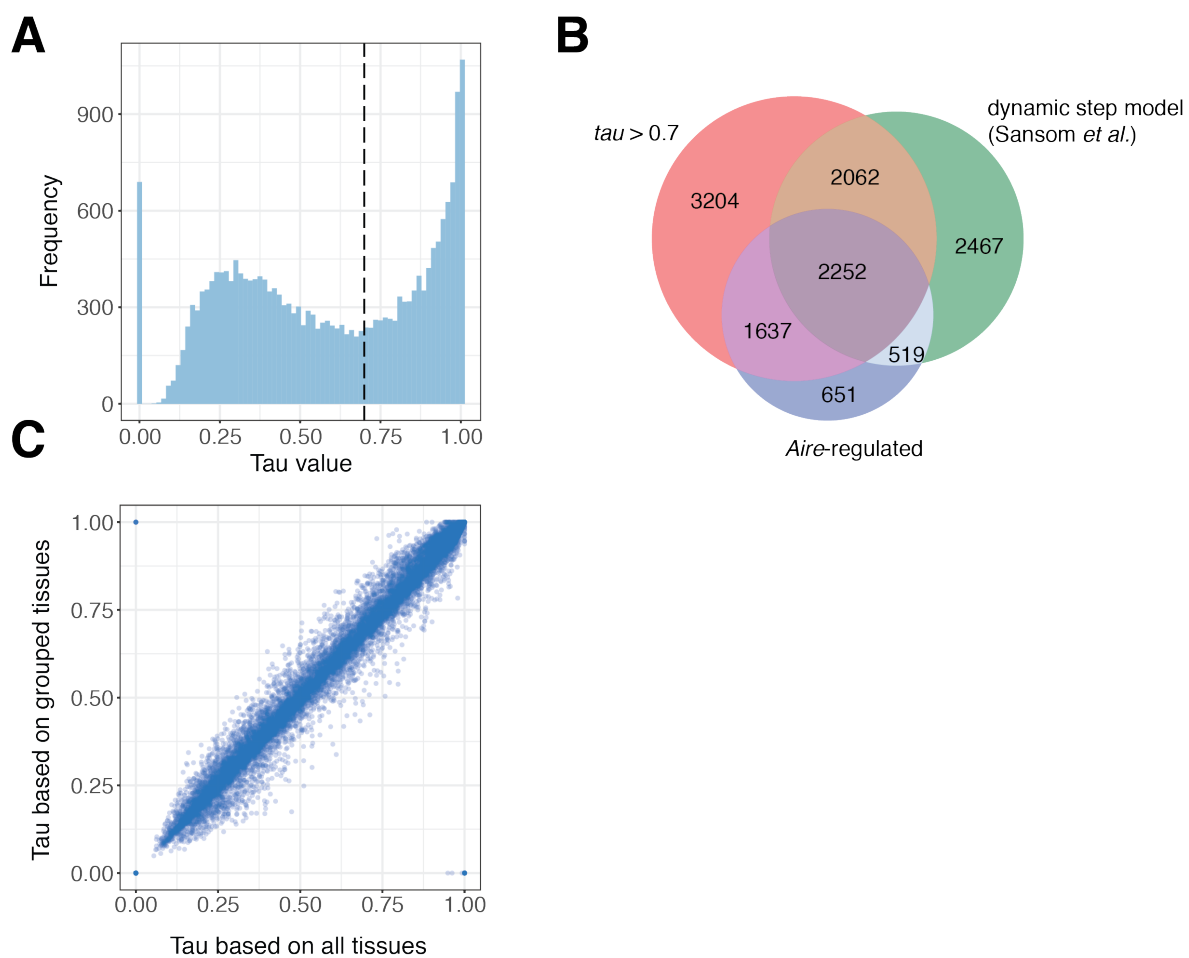
**Figure B.6:** Comparison of *Aire*-regulated genes with a previously defined set. The defined genes here were compared to the ones defined in Sansom *et al.* [27]. Only protein-coding genes according to Ensembl v91 were included.

**Table B.4:** Overview table of all ONT samples. Read length is given for reads that pass quality filters (quality score > 7) only.

| Dataset                   | Cell number; sorting markers | Avg read length | N pass reads | N mapped reads |
|---------------------------|------------------------------|-----------------|--------------|----------------|
| Pilot run (mature)        | 6000; MHChi                  | 706 bp          | 4.4 M        | 3.4 M          |
| Mature (pool of 3 mice)   | 10000; CD80+CD86+            | 1096 bp         | 4.6 M        | 4 M            |
| Immature (pool of 3 mice) | 10000; CD80-CD86-            | 1052 bp         | 3 M          | 2.5 M          |
| Mature WT1                | 10000; CD80+CD86+            | 1185 bp         | 4.8 M        | 4.1 M          |
| Immature WT1              | 10000; CD80-CD86-            | 1065 bp         | 2.7 M        | 2.2 M          |
| Mature WT2                | 10000; CD80+CD86+            | 1130 bp         | 3.3 M        | 2.9 M          |
| Immature WT2              | 10000; CD80-CD86-            | 1006 bp         | 2.6 M        | 2.2 M          |



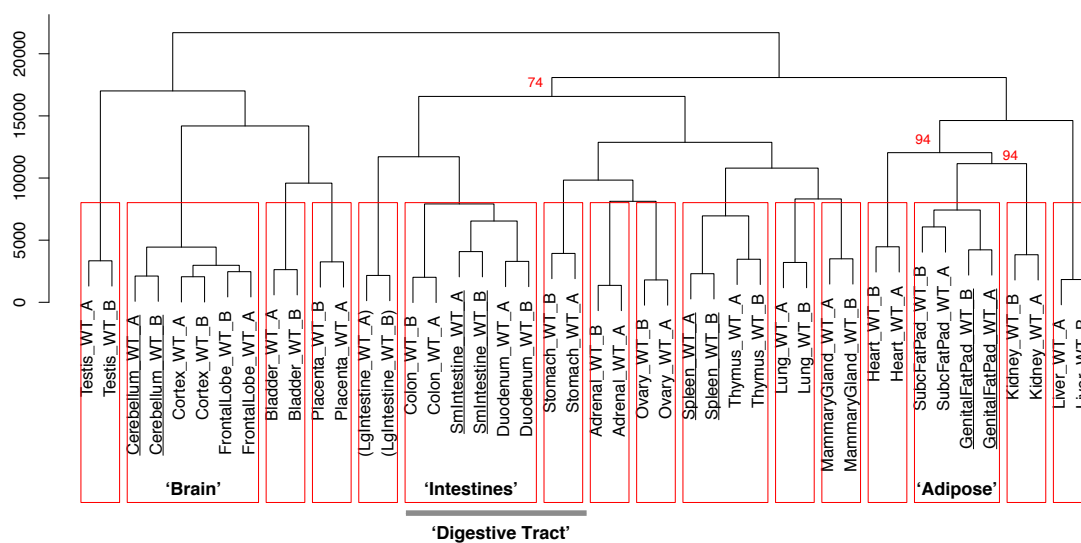
**Figure B.7:** Quality assessment of ONT data. **A:** Histograms of read length for Oxford Nanopore experiments. Immature and mature mTEC show the results of 3 biological replicates, while the pilot data represent one replicate. **B:** Histograms of number of splices per read from merged immature and mature mTEC. **C:** Correlation between percentage identity (as a measure of accuracy) and average base quality for pooled immature (top panel) and mature (bottom panel) ONT samples are shown. Plots were generated by *Nanoplot*. **D:** Gap-compressed divergence as a measure of error (0 representing no errors in alignment) from *Minimap2* for immature and mature mTEC.



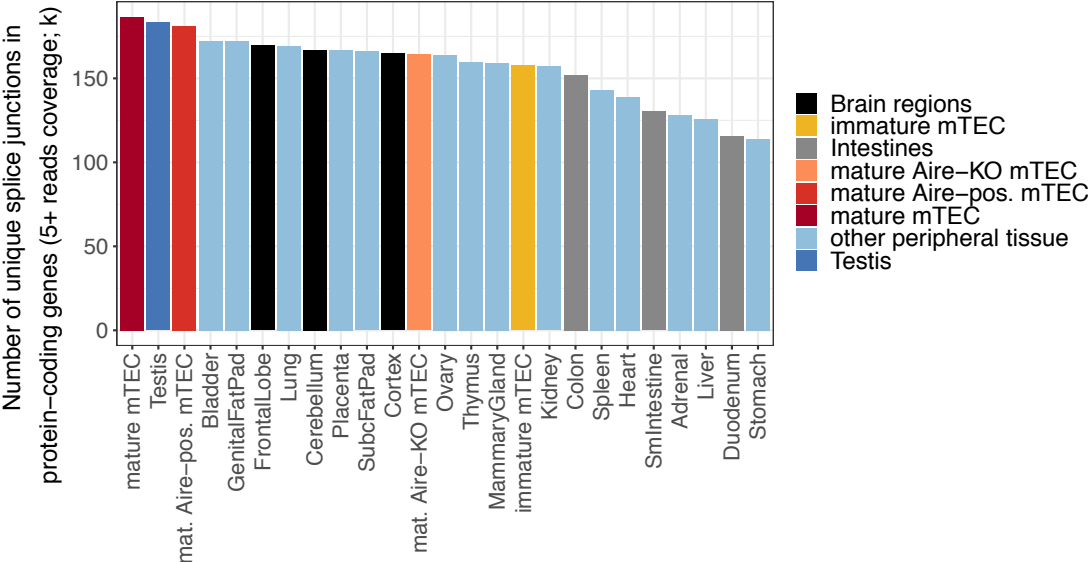
**Figure B.8:** Definition of tissue-restricted genes using  $\tau$ . **A:** Identification of genes with tissue-restricted expression for grouped tissues; Formula to calculate  $\tau$  in Equation 2.2. **B:** Comparison of tissue-restricted genes with a previous published method for tissue-restriction as well as *Aire*-regulated genes. **C:** Comparison of  $\tau$  value calculated based on all tissues or based on tissue groups (for definition of tissue groups, see Section 2.1.9.1).



## Appendix - Chapter 4

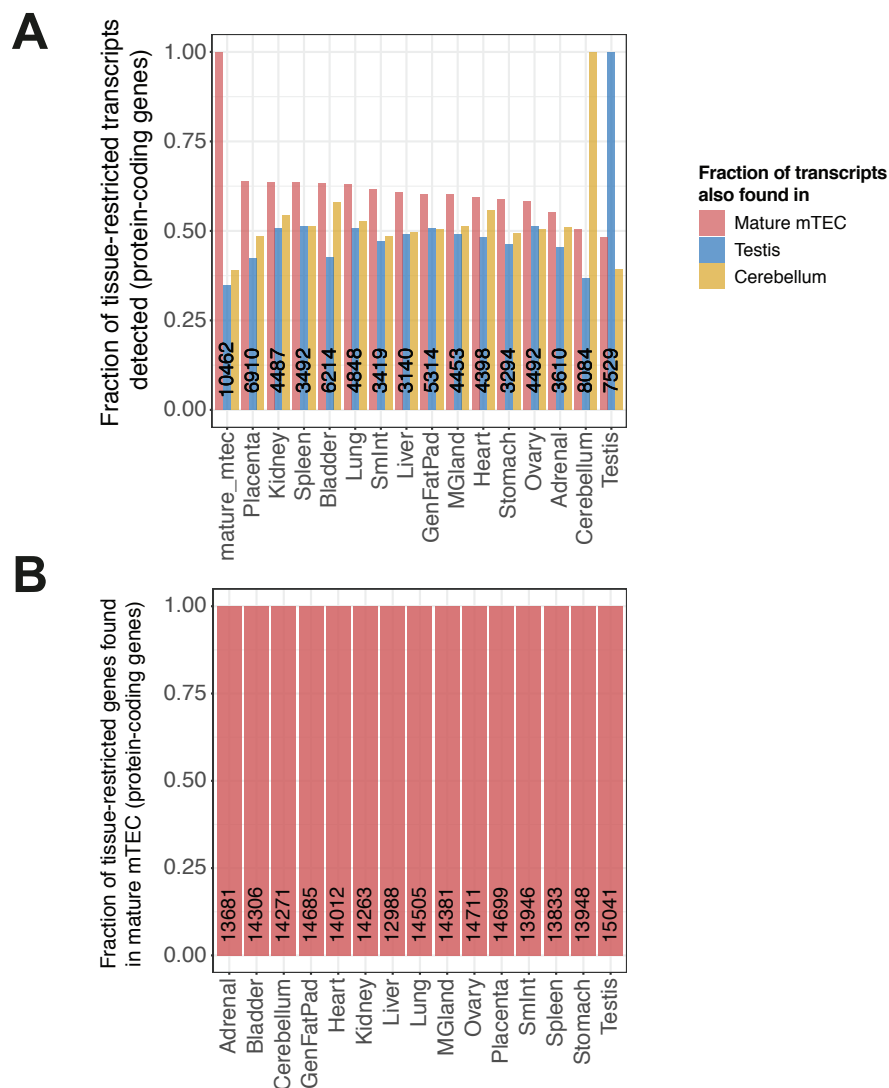


**Figure C.1:** Definition of tissue groups. Hierarchical clustering of most highly variable transcripts in protein-coding genes (Manhattan distance, complete linkage). All clusters had approximate unbiased (AU) p-value from *pvcust* of  $\leq 95$  if not indicated in the figure.



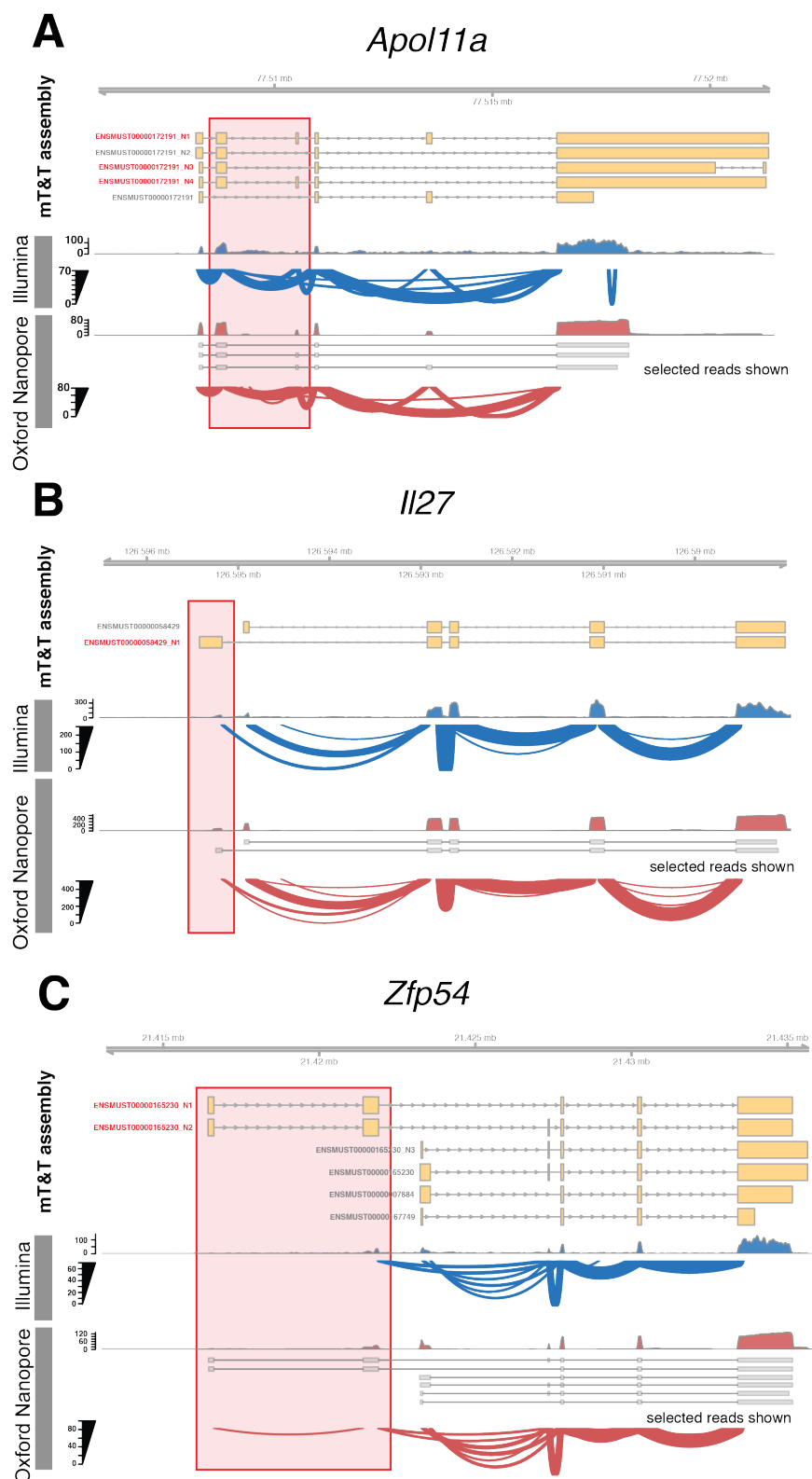
**Figure C.2:** Transcriptome complexity in TEC (corresponds to Figure 4.1). Number of unique splice junction at 200 M reads in mTEC types and peripheral tissues.



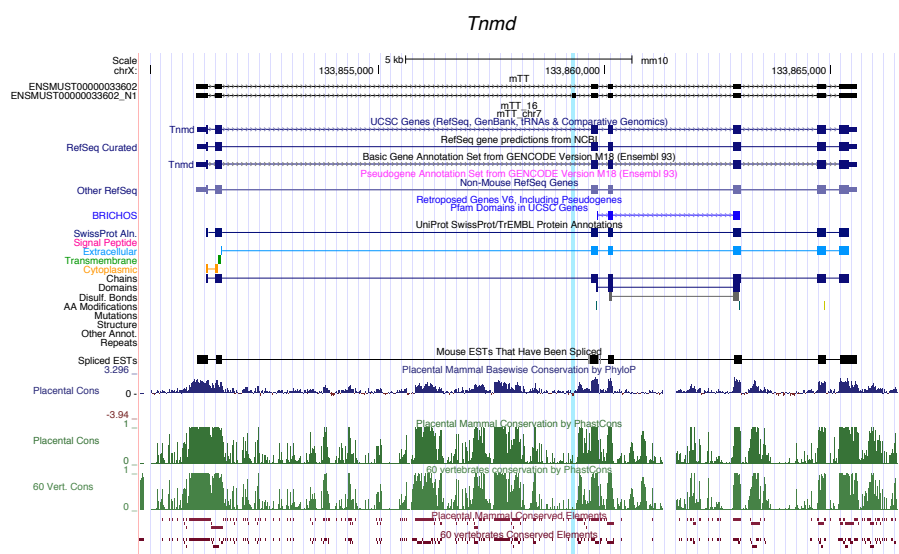


**Figure C.3:** Peripheral representation of tissue-restricted transcripts. **A:** Fraction of peripheral tissue-restricted transcripts ( $\tau \geq 0.9$ ) found in the tested tissues (TEC, Testis, Cerebellum). This figure includes the values used for the boxplot in Figure 4.5B. However, the values for mature mTEC, Testis and Cerebellum were not included in the boxplot. Only protein-coding genes that were commonly expressed ( $> 0.45$  TPM) in all three tissues were considered in this analysis. Novel transcripts were not included. **B:** Fraction of genes with at least one transcript with  $\tau \geq 0.9$  expressed in mature mTEC for each peripheral tissue group. Only protein-coding genes with expression level of  $> 0.45$  TPM in mature mTEC were included.

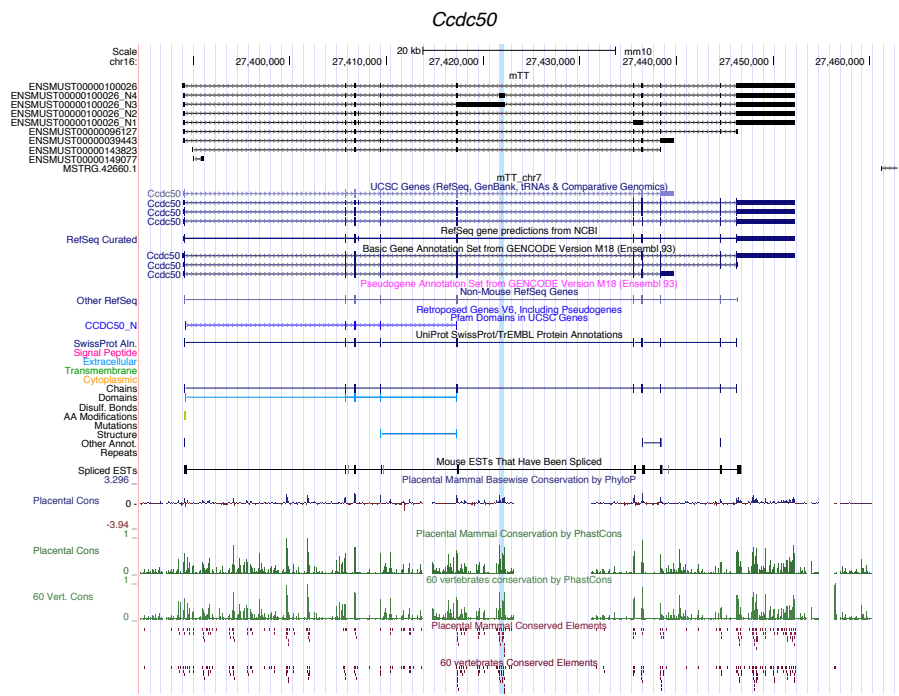
The novel structure in Apolipoprotein L1a (*Apol11a*) shown in Figure C.4A contained a novel second exon, which was frequently included. Further novel transcripts from the ONT assembly and details about *Apol11a* can be found in Section 3.3.2.2. An example of a gene with a novel first exon was in the cytokine *Il27*, which is part of the Il12-cytokine family. It was described in a pro- and anti-inflammatory context [178]. Finally, two novel first exons were identified in the zinc finger domain protein *Zfp54*. ZFP54 contains a Krueppel-associated box (KRAB) domain and multiple zinc finger motifs [270]. Moreover, it was shown to bind to DNA via the KRAB and is a potential transcriptional repressors in early development.



**Figure C.4:** Examples of TEC-specific novel transcripts showing *Apol11a* in **A**, *Il27* in **B** and *Zfp54* in **C**. The novel TEC-specific transcripts are highlighted in red. Sashimi plots were generated from RNA-sequencing data from mature mTEC at 200 M reads (blue) or from merged ONT datasets for mature mTEC (red). Selected reads supporting the transcript structures are shown.



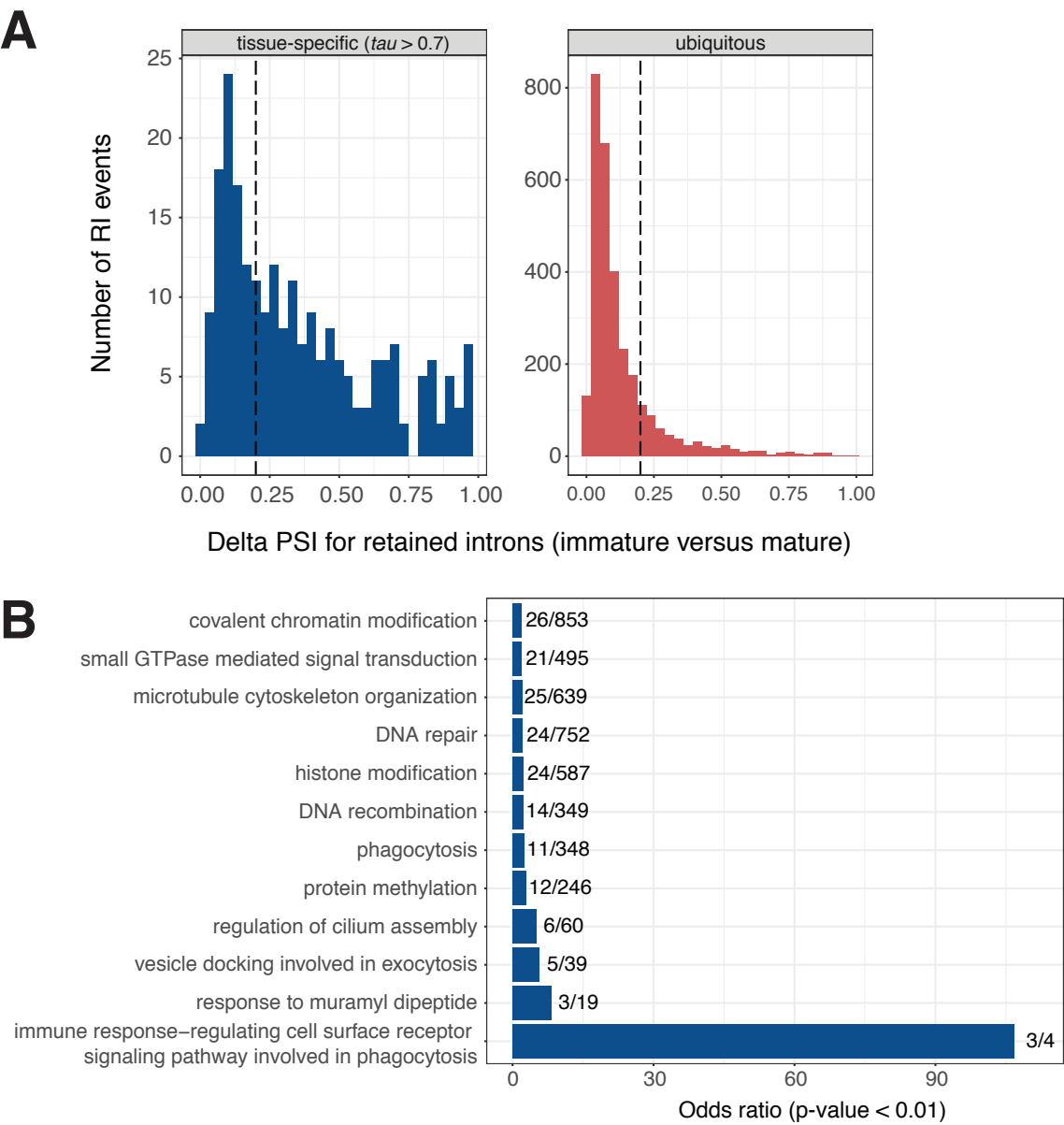
**Figure C.5:** Protein prediction for *Tnmd*. Overview of *Tnmd* locus including mT&T assembly track, known proteins and PhastCon scores. (downloaded from *USCS Genome Browser* on 11/01/19)



**Figure C.6:** Protein prediction for *Ccdc50*. Overview of *Ccdc50* locus including mT&T assembly track, known proteins and PhastCon scores. (downloaded from *USCS Genome Browser* on 11/01/19)

# D

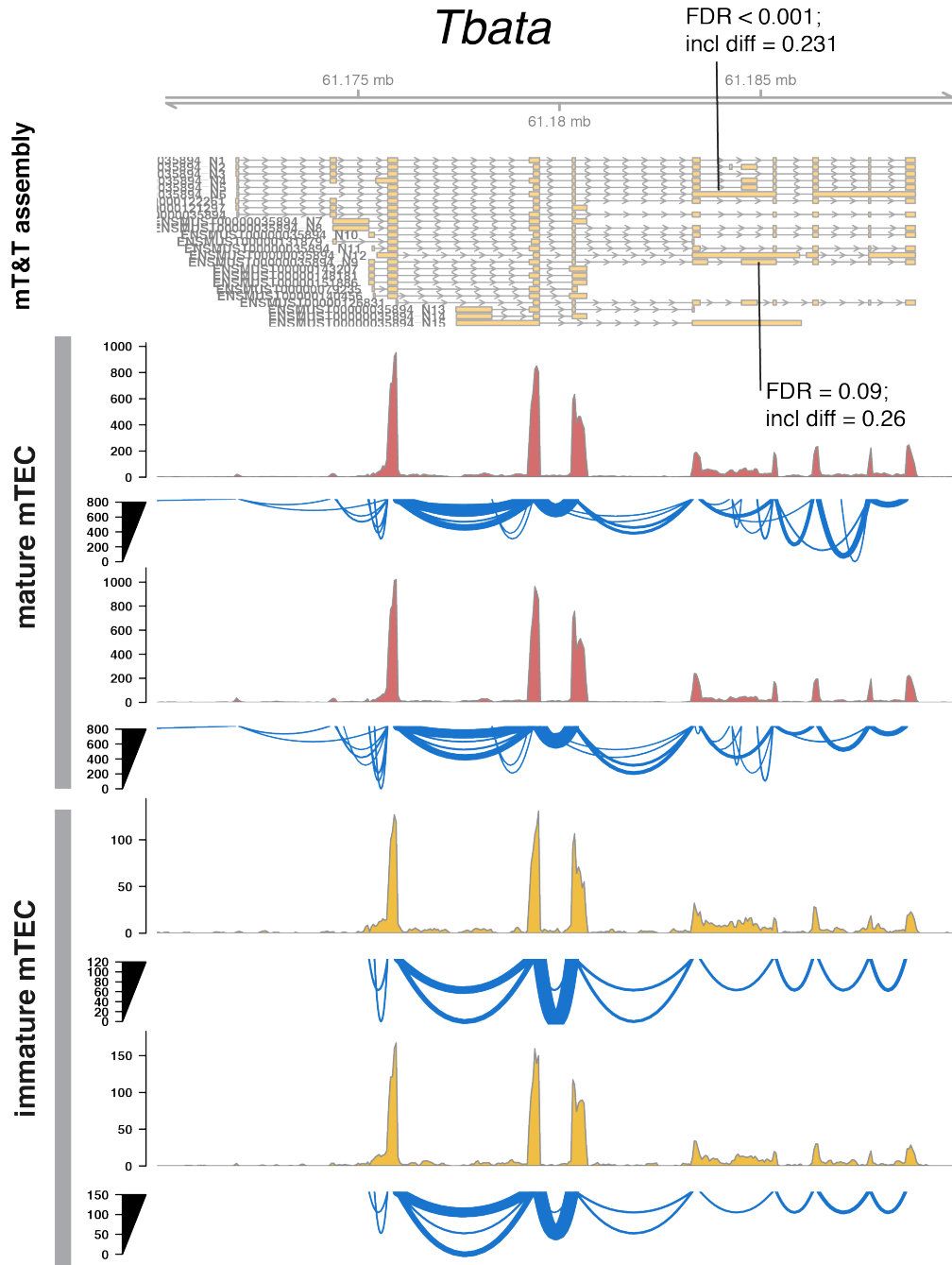
## Appendix - Chapter 5



**Figure D.1:** Characterisation of genes containing significantly differently spliced retained intron (RI) events in immature mTEC compared to mature mTEC. **A:** The protein-coding genes are divided into ubiquitously expressed and tissue-specific using the  $\tau$  metric. **B:** Geneset over-representation analysis using GO categories of Biological Processes (BP) in Fisher's Exact tests. The genes included contained RI events differentially included between immature and mature mTEC with a minimum delta PSI of 0.2. The categories were filtered by the number of genes per category more than 2, the minimum odds ratio higher than 1 and the p-value < 0.01. Representative categories are shown.

**Table D.1:** Number of events per splice event type in *rMATS* analysis of mTEC. The filters applied to the events were absolute delta PSI > 0.2 and FDR < 0.05. SE = skipped exon, RI = retained intron, MXE = mutually exclusive exon, A3SS/A5SS = alternative 3'/5' splice site.

| Comparison          | Event | Sample   | Events filtered | Events unfiltered |
|---------------------|-------|----------|-----------------|-------------------|
| Aire KO vs Aire pos | A3SS  | Aire KO  | 43              | 56                |
| Aire KO vs Aire pos | A5SS  | Aire KO  | 52              | 61                |
| Aire KO vs Aire pos | MXE   | Aire KO  | 9               | 11                |
| Aire KO vs Aire pos | RI    | Aire KO  | 145             | 195               |
| Aire KO vs Aire pos | SE    | Aire KO  | 70              | 82                |
| immature vs Aire KO | A3SS  | immature | 161             | 397               |
| immature vs Aire KO | A5SS  | immature | 197             | 502               |
| immature vs Aire KO | MXE   | immature | 40              | 61                |
| immature vs Aire KO | RI    | immature | 489             | 1888              |
| immature vs Aire KO | SE    | immature | 239             | 417               |
| immature vs mature  | A3SS  | immature | 174             | 438               |
| immature vs mature  | A5SS  | immature | 188             | 542               |
| immature vs mature  | MXE   | immature | 41              | 63                |
| immature vs mature  | RI    | immature | 578             | 2356              |
| immature vs mature  | SE    | immature | 230             | 384               |
| Aire KO vs Aire pos | A3SS  | Aire pos | 37              | 47                |
| Aire KO vs Aire pos | A5SS  | Aire pos | 52              | 66                |
| Aire KO vs Aire pos | MXE   | Aire pos | 8               | 10                |
| Aire KO vs Aire pos | RI    | Aire pos | 79              | 143               |
| Aire KO vs Aire pos | SE    | Aire pos | 67              | 82                |
| immature vs Aire KO | A3SS  | Aire KO  | 90              | 147               |
| immature vs Aire KO | A5SS  | Aire KO  | 91              | 147               |
| immature vs Aire KO | MXE   | Aire KO  | 27              | 43                |
| immature vs Aire KO | RI    | Aire KO  | 192             | 315               |
| immature vs Aire KO | SE    | Aire KO  | 203             | 307               |
| immature vs mature  | A3SS  | mature   | 80              | 130               |
| immature vs mature  | A5SS  | mature   | 83              | 125               |
| immature vs mature  | MXE   | mature   | 23              | 46                |
| immature vs mature  | RI    | mature   | 143             | 282               |
| immature vs mature  | SE    | mature   | 201             | 295               |



**Figure D.2:** *Tbata* as an additional example of retained introns (RI) with significantly higher inclusion in immature mTEC compared to mature mTEC. The delta PSI values and FDR from *rMATS* are indicated. Sashimi plots were generated from Illumina data at 200 M reads and splice junctions with 5+ reads coverage are shown.

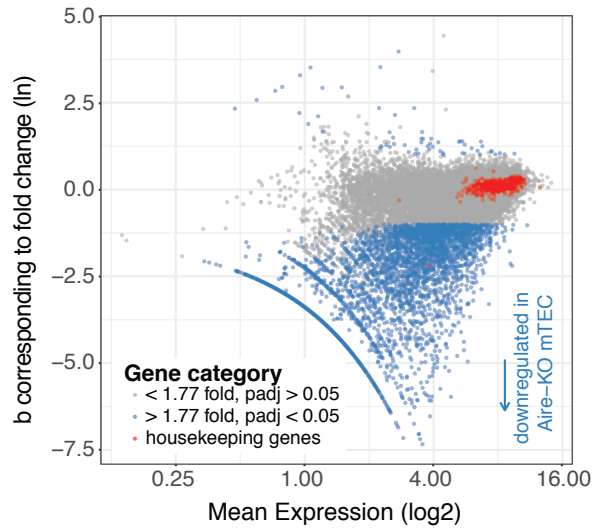


**Table D.2:** Validation of novel transcripts containing retained introns between Illumina and ONT data. The Illumina data was generated from mature (MHCII high) and immature (MHCII low) mTEC, while the ONT data was from CD80<sup>+</sup>CD86<sup>+</sup> mature mTEC and CD80<sup>-</sup>CD86<sup>-</sup> immature mTEC. N=Number

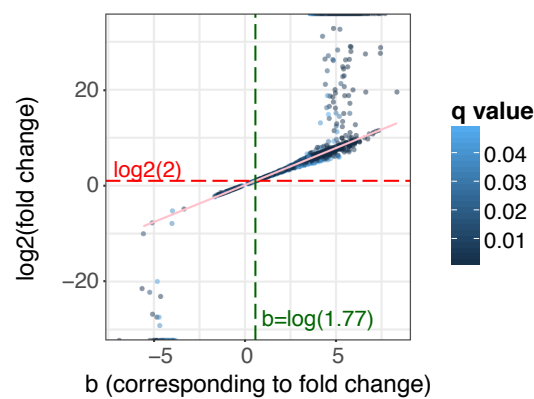
| mTEC type | Threshold counts (Illumina) | N of RI-transcripts | N validated in ONT | Fraction validated |
|-----------|-----------------------------|---------------------|--------------------|--------------------|
| immature  | 0                           | 6395                | 1922               | 30 %               |
| immature  | 100                         | 416                 | 349                | 84 %               |
| mature    | 0                           | 3998                | 1817               | 45 %               |
| mature    | 100                         | 242                 | 217                | 90 %               |

**Table D.3:** Number of *Aire*-regulated transcripts for *DESeq2* and *Sleuth*. In addition, the number of differentially expressed transcripts in house-keeping genes is indicated. The threshold for b was determined using a linear model in Figure D.4

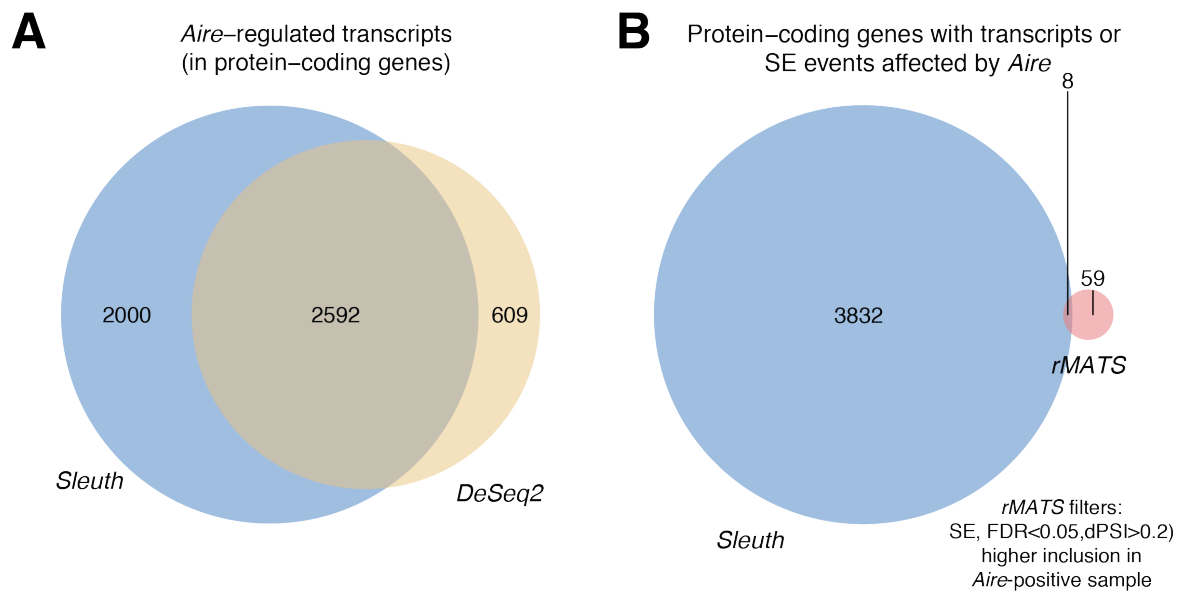
|                                                                | <i>DESeq2</i><br>( $< 2$ -fold) | <i>Sleuth</i><br>( $b < \log(1.77)$ ) |
|----------------------------------------------------------------|---------------------------------|---------------------------------------|
| <i>Aire</i> -regulated transcripts                             | 4,099                           | 4,726                                 |
| <i>Aire</i> -regulated transcripts<br>in protein-coding genes  | 3,021                           | 4,618                                 |
| differentially expressed transcripts<br>in house-keeping genes | 0                               | 4                                     |



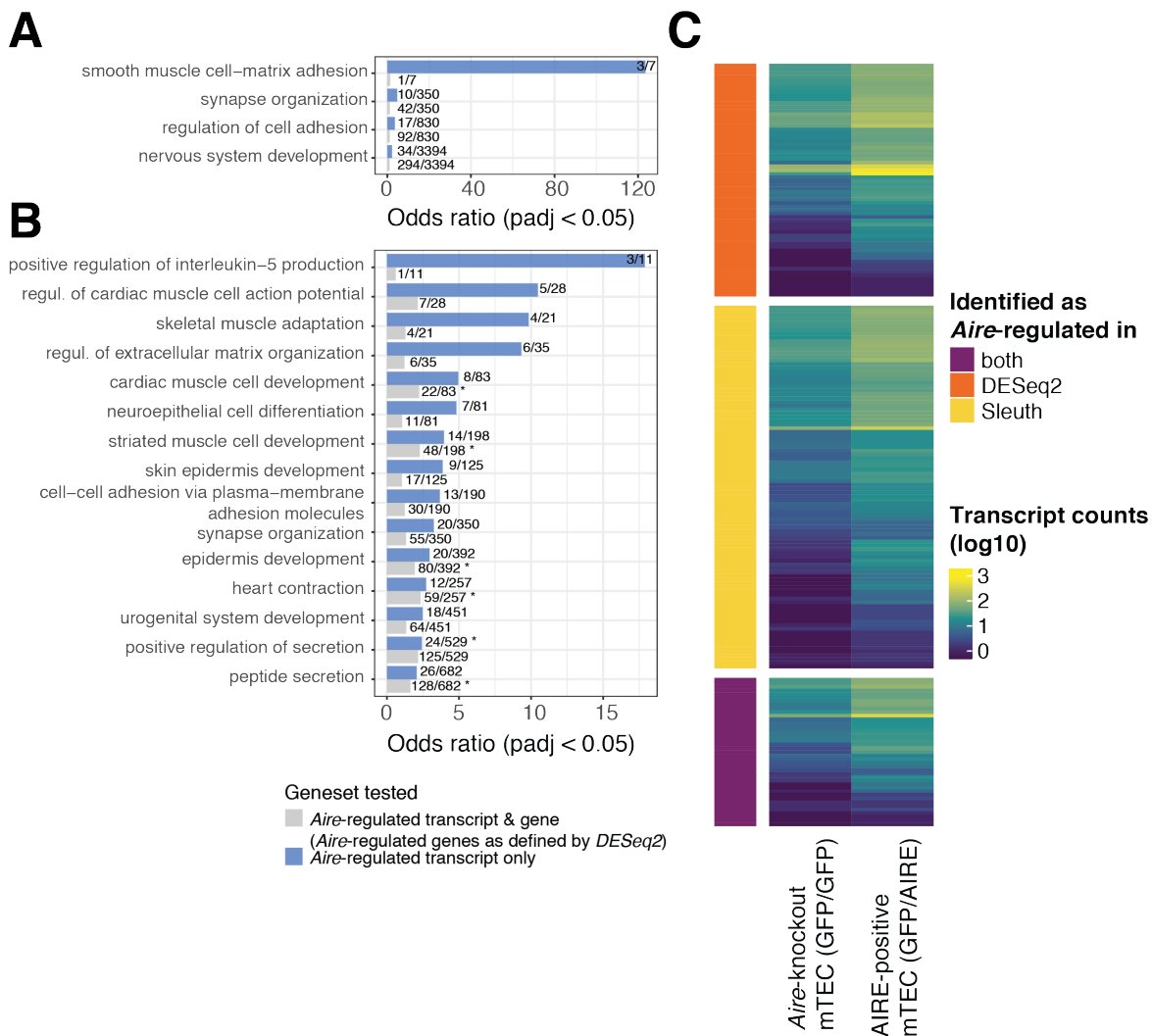
**Figure D.3:** MA plot of *Sleuth*-based identification of *Aire*-regulated genes. Differentially expressed genes between *Aire*-positive and *Aire*-knockout mTEC based on *Sleuth* (Wald test). *Aire*-regulated genes were defined as  $b < 1.77$  (corresponds to fold change on natural log scale) and Benjamini-Hochberg adjusted p-value  $< 0.05$ . Housekeeping genes are plotted in red.



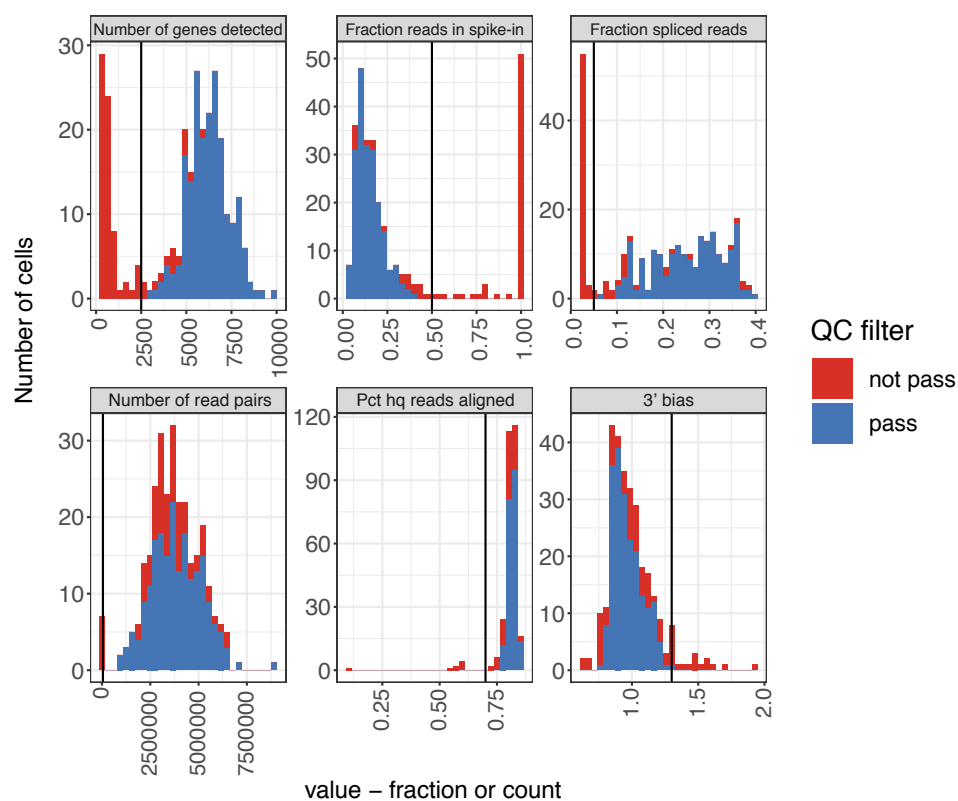
**Figure D.4:** Determination of the fold change corresponding  $b$  from *Sleuth*. The fold changes ( $\log_2$ ) were compared to  $b$  (natural log) from *Sleuth* to determine the appropriate  $b$  for an equivalent definition to *DESeq2*. The threshold for  $b$  (green line) was determined using the model (pink line).



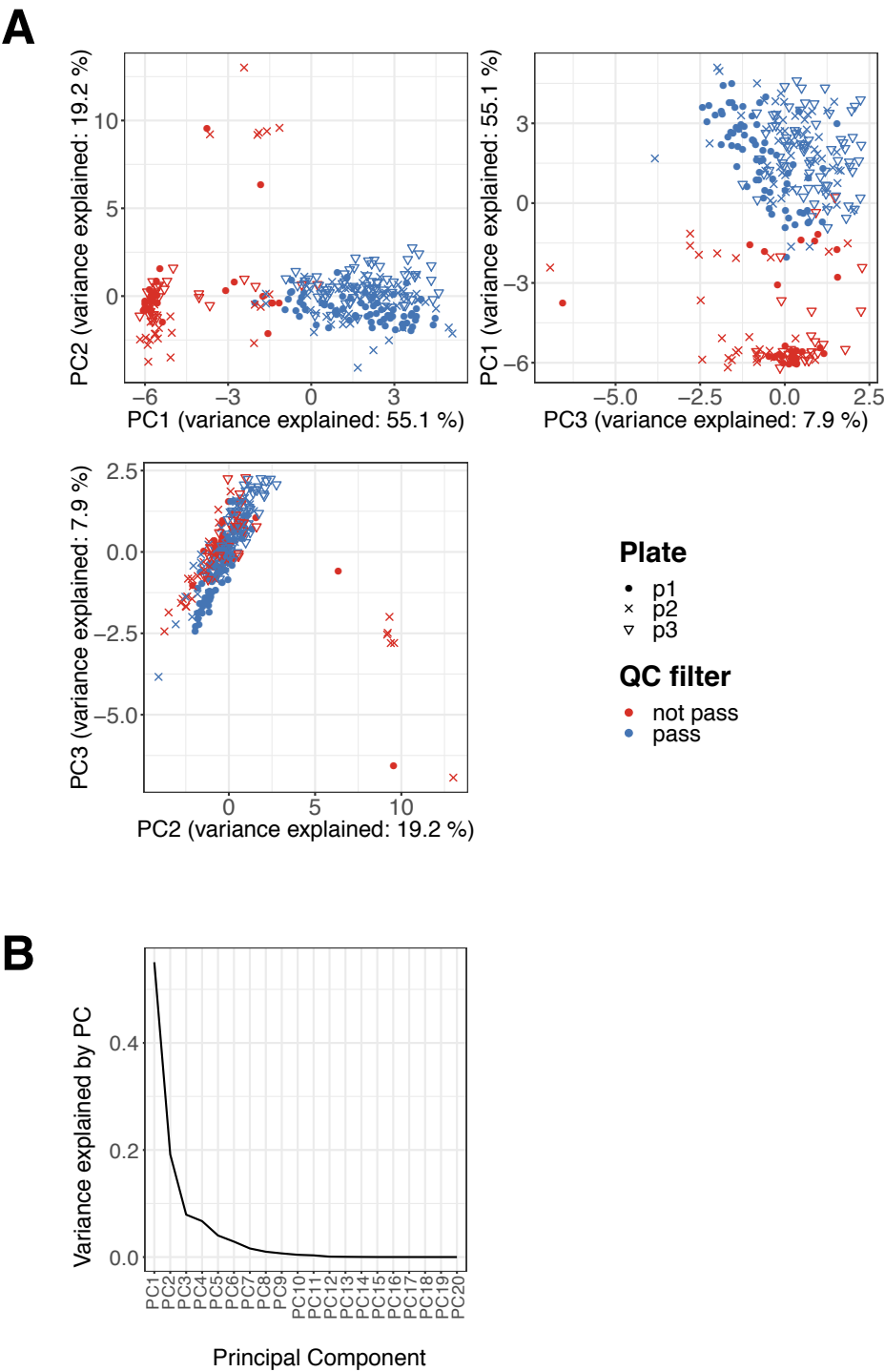
**Figure D.5:** Comparison of *Aire*-regulated transcripts and genes with splicing changes affected by *Aire* between different methods. **A:** Venn diagram showing overlap of *Aire*-regulated transcripts in protein-coding genes between *Sleuth* and *DESeq2* approaches as defined in Figure 5.4. **B:** Venn diagram showing overlap between protein-coding genes containing *Aire*-regulated transcripts as defined by *Sleuth* (defined in Figure 5.4) and skipped exon (SE) events with higher inclusion in *Aire*-positive mTEC defined by *rMATs* (Figure 5.1, Appendix Table D.1). The *rMATs* output was filtered for FDR < 0.05, absolute delta percentage spliced-in > 0.2.



**Figure D.6:** Assessment of genes containing *Aire*-regulated transcripts but not *Aire*-regulated genes and *Aire*-regulated transcripts not part of *Aire*-regulated genes. **A** and **B**: Genesets over-representation using Fisher’s Exact tests amongst genes containing *Aire*-regulated transcripts but not identified as *Aire*-regulated at the gene level (blue bars). Representative categories are shown for the *DESeq2* approach in **A**, while **B** shows categories over-enriched by the *Sleuth*-based approach. For both analyses the categories were filtered for more than 2 genes per category, odds ratio > 2 and adjusted p-value < 0.05 (Benjamini Hochberg correction). The grey bars indicate the over-enrichment for the set of genes defined as *Aire*-regulated and containing *Aire*-regulated transcripts by the respective method. Significant categories are highlighted with a ‘\*’ (adjusted p-value < 0.05). **B**: Heatmap of transcript counts (log10-scale) for transcripts identified as *Aire*-regulated transcripts but not included in *Aire*-regulated genes. The combination of transcripts identified in either or both approaches are shown. 200 transcripts with the highest fold change between *Aire*-positive and *Aire*-knockout mTEC were included.



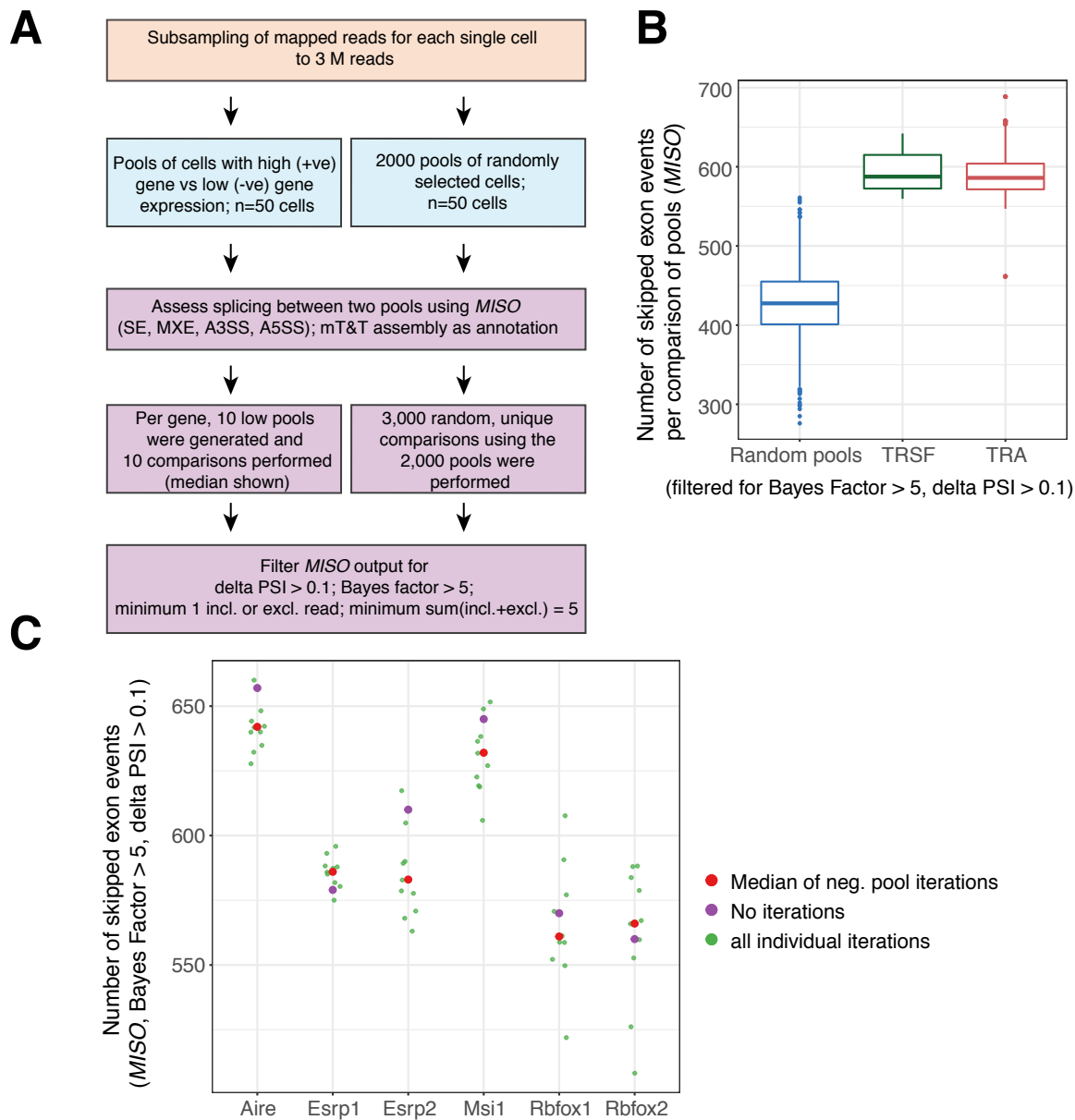
**Figure D.7:** Histograms of selected quality metrics for 285 single mature mTEC. Threshold for the respective quality metrics are indicated in each panel. An additional filter was percentage duplication  $< 0.5$ . The number of genes detected was determined using *Cufflinks*. Details on all filters are described in Section 2.1.7 and full description of metrics is given in Appendix Table A.2.



**Figure D.8:** Principal Component Analysis of all quality metrics in single mature mTEC. **A:** Three panels showing the combinations of PC1-PC3. Color indicates, whether a cell passed the quality filters as defined in Figure D.7. **B:** Scree plot for PCA of quality metrics in single mature mTEC. All metrics included in the PCA are given in Table D.4.

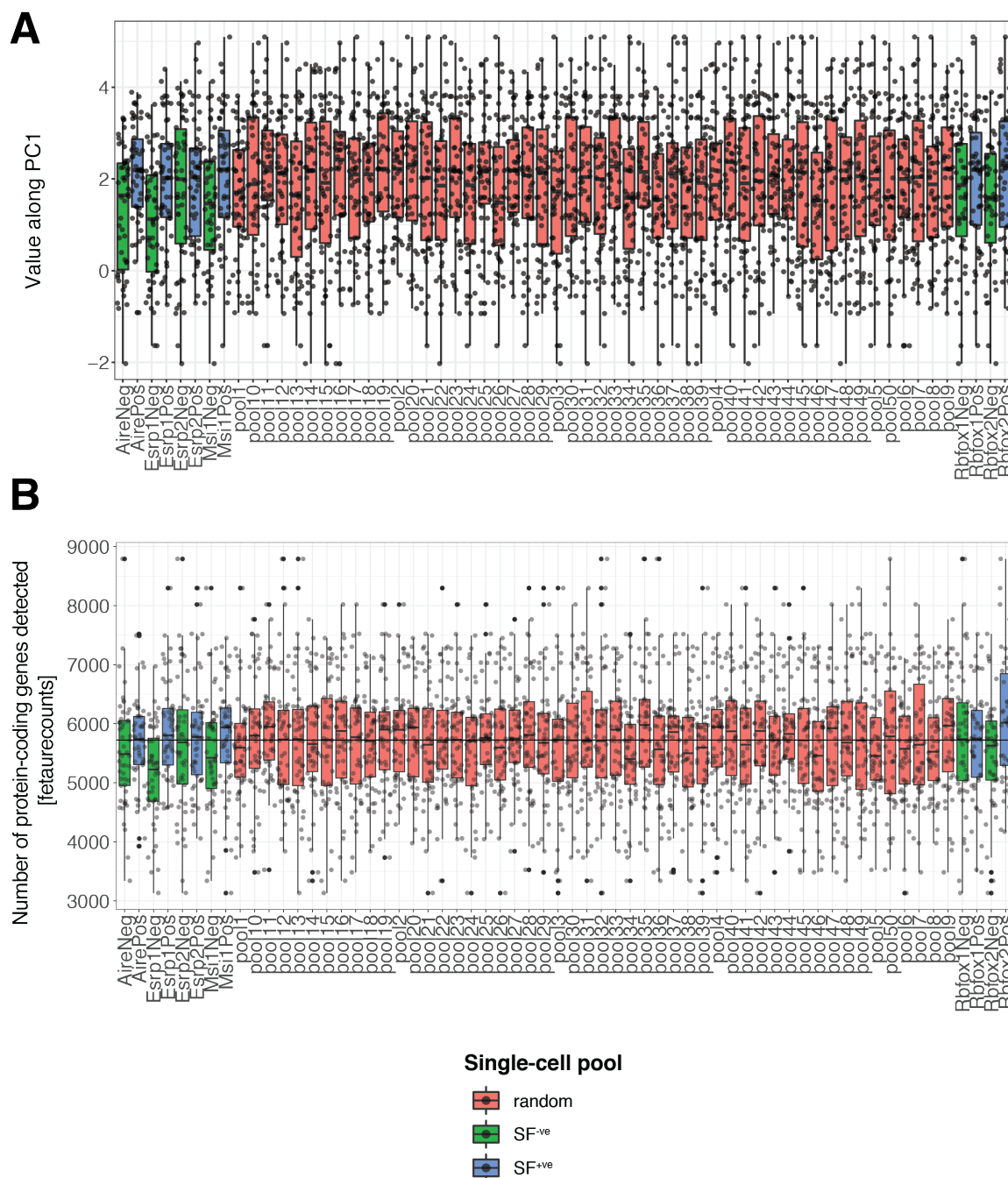
**Table D.4:** Loadings for the first 5 PCs for the PCA of all quality metrics in single mTEC. The explanation and source of each quality metric is described in Table A.2.

| Quality metric             | PC1    | PC2    | PC3    | PC4    | PC5    |
|----------------------------|--------|--------|--------|--------|--------|
| cufflinks-no-genes         | 0.290  | -0.052 | 0.076  | -0.104 | 0.022  |
| pct-mrna                   | 0.289  | 0.120  | -0.052 | -0.005 | -0.003 |
| featurecounts-no-genes     | 0.289  | -0.055 | 0.084  | -0.108 | 0.022  |
| cufflinks-no-genes-pc      | 0.289  | -0.048 | 0.088  | -0.105 | 0.021  |
| featurecounts-no-genes-pc  | 0.287  | -0.050 | 0.097  | -0.109 | 0.022  |
| pct-coding                 | 0.279  | 0.123  | -0.133 | 0.149  | 0.021  |
| fraction-spliced           | 0.276  | 0.115  | -0.146 | 0.170  | 0.035  |
| library-size               | 0.264  | -0.138 | -0.148 | -0.026 | 0.016  |
| nreads-uniq-map-genome     | 0.247  | -0.184 | -0.207 | -0.216 | -0.052 |
| median-insert-size         | 0.218  | 0.122  | -0.010 | 0.407  | -0.014 |
| pct-reads-aligned          | 0.144  | -0.316 | 0.321  | 0.301  | -0.275 |
| pct-pf-reads-aligned-hq    | 0.068  | -0.356 | 0.358  | 0.313  | -0.314 |
| no-pairs                   | 0.046  | -0.416 | -0.420 | -0.117 | -0.019 |
| pct-reads-aligned-in-pairs | 0.041  | -0.393 | 0.272  | -0.113 | 0.040  |
| total-reads                | 0.026  | -0.422 | -0.419 | -0.102 | -0.019 |
| pct-adapter                | -0.033 | 0.221  | -0.284 | -0.015 | -0.869 |
| three-prime-bias           | -0.101 | 0.006  | 0.317  | -0.634 | -0.244 |
| nreads-uniq-map-spike      | -0.248 | -0.209 | -0.147 | 0.192  | 0.004  |
| pct-duplication            | -0.263 | -0.200 | -0.053 | 0.092  | 0.042  |
| fraction-spike             | -0.282 | -0.118 | -0.023 | 0.133  | 0.031  |



**Figure D.9:** Pipeline scheme and additional data for *in silico* single-cell pooling analysis. **A:** Outline of the analysis pipeline to generate single-cell pools and assess the differential splicing events using *MISO* (details in Section 2.1.7.3). **B:** Boxplots to summarise number of skipped exon events for three different types of pools (combination of Figure 5.14A and B) **C:** Dotplots showing the effects of TRSF<sup>-ve</sup> on differential splicing. All individual iterations with negative pools are shown in green, while the median for the iterations and the initial result without iterations are highlighted in red and purple, respectively. SE = skipped exon, RI = retained intron, MXE = mutually exclusive exon, A3/A5 = alternative 3'/5' splice sites.





**Figure D.10:** Quality metrics of *in silico* single-cell pools. **A:** Boxplots of the PC1 value (from Figure D.8) for each cell in a pool. **B:** Boxplots of the number of protein-coding genes expressed (*featurecounts*) for each cell in a pool. In both panels, the horizontal line indicates the median across the random pools. The color legend for both panels is given below. The jittered dots represent individual cells, while the box plots shows a summary of the data. Solid, black circles above or below the boxplot represent the outliers of the boxplot and these data points are therefore duplicated.

**Table D.5:** Overview of events for all tissue-restricted splicing factor (TRSF) comparisons per event type using single-cell pooling approach. A3SS/A5SS = alternative 3'/5' splice site, MXE = mutually exclusive exon, SE = skipped exon.

| Gene id             | A3SS  | A5SS  | MXE  | SE    |
|---------------------|-------|-------|------|-------|
| Aire                | 326   | 336   | 78   | 642   |
| Aptx                | 336.5 | 318.5 | 73.5 | 611.5 |
| Cir1                | 348   | 360.5 | 63.5 | 601   |
| Cwc22               | 300   | 325.5 | 61   | 567.5 |
| Ddx20               | 304.5 | 326   | 65.5 | 585.5 |
| Esrp1               | 347   | 372   | 66   | 586   |
| Esrp2               | 327   | 333   | 83   | 583   |
| Gemin6              | 300.5 | 320   | 66   | 559.5 |
| Larp6               | 341.5 | 337   | 73   | 594   |
| Lsm5                | 333   | 370.5 | 71   | 630.5 |
| Magohb              | 329.5 | 352.5 | 67.5 | 631   |
| Mbnl3               | 352.5 | 339.5 | 69.5 | 587.5 |
| Msi1                | 344   | 344   | 61   | 632   |
| Ppwd1               | 307.5 | 324   | 62   | 565   |
| Raver2              | 311.5 | 323.5 | 69.5 | 572.5 |
| Rbfox1              | 324   | 323   | 67   | 561   |
| Rbfox2              | 317   | 330   | 65   | 566   |
| RbfoxDoublePositive | 309.5 | 331.5 | 75.5 | 581.5 |
| Rbm20               | 321.5 | 330.5 | 72.5 | 566   |
| Rbm34               | 326   | 343   | 67.5 | 586.5 |
| RbmX2               | 329.5 | 329   | 64   | 572.5 |
| Rsrc1               | 334.5 | 352   | 85   | 599.5 |
| Setx                | 341   | 342.5 | 73   | 570   |
| Slu7                | 345.5 | 342.5 | 65.5 | 588   |
| Snrpb2              | 359   | 350   | 71.5 | 598.5 |
| Snrpd1              | 366   | 398   | 73.5 | 617.5 |
| SrpK2               | 318.5 | 336   | 74.5 | 581   |
| Ssb                 | 347.5 | 370.5 | 69.5 | 615   |
| Syncrip             | 360.5 | 379.5 | 74   | 634   |
| Synj2               | 322   | 337.5 | 67   | 590.5 |
| Thoc1               | 352   | 330.5 | 86   | 620   |

**Table D.6:** Overview of events for other tissue-restricted gene (TRA) comparisons per event type using single-cell pooling approach. 10 TRA with the highest number of skipped exon events are shown. A3SS/A5SS = alternative 3'/5' splice site, MXE = mutually exclusive exon, SE = skipped exon.

| Gene id | A3SS  | A5SS  | MXE  | SE    |
|---------|-------|-------|------|-------|
| Gm5148  | 387.5 | 396   | 69   | 688.5 |
| Mfap1B  | 361   | 358   | 86.5 | 658   |
| Calcb   | 330.5 | 364.5 | 80   | 656.5 |
| Arl6    | 358   | 327.5 | 73.5 | 654   |
| Gem     | 337   | 358.5 | 73   | 652   |
| Ap1m2   | 360   | 364   | 75   | 647.5 |
| Dsp     | 354   | 356.5 | 71   | 646   |
| Cd52    | 309   | 356   | 72   | 644.5 |
| Dio1    | 356   | 355.5 | 80.5 | 639   |

**Table D.7:** Overview table summarising the characteristics of selected tissue-restricted splicing factors ( $\tau > 0.5$ ) used for prioritisation, part 1. The columns 'sc-freq' and 'sc-expr' indicate the frequency of single mTEC expressing a factor and the expression level (where detected), respectively. The remaining columns indicate the TPM values from *Salmon* for cTEC, mature mTEC (mtechi), immature mTEC (mtecto), skin epithelial cell (SEC).

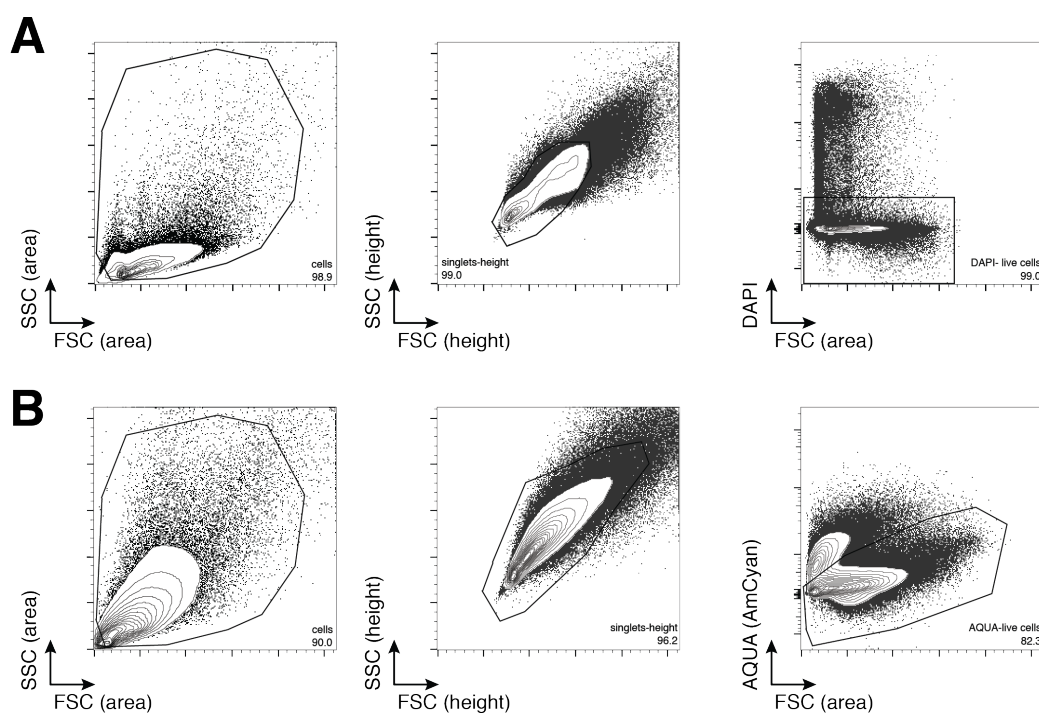
| gene id            | gene name | tau  | sc-freq | sc-expr | cTEC | mtechi | mtecto | SEC  |
|--------------------|-----------|------|---------|---------|------|--------|--------|------|
| ENSMUSG00000002007 | Srpk3     | 0.67 | 0.01    | 1.97    | 0.17 | 1      | 0.59   | 0.06 |
| ENSMUSG00000003410 | Elavl3    | 0.94 | 0.03    | 1.58    | 0.01 | 1.24   | 0.05   | 0.31 |
| ENSMUSG00000008489 | Elavl2    | 0.89 | 0.05    | 2.28    | 0.05 | 1      | 0.32   | 0.11 |
| ENSMUSG00000008658 | Rbfox1    | 0.81 | 0.41    | 1.93    | 0    | 3.87   | 4.81   | 0.29 |
| ENSMUSG00000020409 | Slu7      | 0.53 | 0.66    | 2.16    | 5.38 | 4.99   | 5.16   | 4.44 |
| ENSMUSG00000021047 | Nova1     | 0.65 | 0.06    | 0.95    | 0.02 | 0.29   | 0.2    | 0    |
| ENSMUSG00000024268 | Celf4     | 0.8  | 0.07    | 1.81    | 2.7  | 1.52   | 1.05   | 0.46 |
| ENSMUSG00000025576 | Rbfox3    | 0.86 | 0.21    | 1.41    | 0    | 2.29   | 0.08   | 0.04 |
| ENSMUSG00000026058 | Khdrbs2   | 0.9  | 0.14    | 1.57    | 0.83 | 2.3    | 3.09   | 0    |
| ENSMUSG00000028546 | Elavl4    | 0.92 | 0.03    | 2.13    | 0.04 | 1.2    | 0.38   | 0.04 |
| ENSMUSG00000029279 | Brd1      | 0.8  | 0.03    | 1.61    | 0.02 | 0.92   | 0.64   | 0    |
| ENSMUSG00000030411 | Nova2     | 0.78 | 0.07    | 1.54    | 0.57 | 0.17   | 0.28   | 0.03 |
| ENSMUSG00000032297 | Celf6     | 0.88 | 0.04    | 1.28    | 0.01 | 2.07   | 0.13   | 0    |
| ENSMUSG00000032423 | Syncrin   | 0.51 | 0.73    | 2.27    | 6.52 | 5.83   | 5.96   | 6.77 |
| ENSMUSG00000032940 | Rbm11     | 0.9  | 0.01    | 2.12    | 0.57 | 1.02   | 0.51   | 0.05 |
| ENSMUSG00000036109 | Mbnl3     | 0.65 | 0.3     | 2.25    | 2.91 | 4.55   | 3.21   | 2.25 |
| ENSMUSG00000038132 | Rbm24     | 0.65 | 0.02    | 1.91    | 0.61 | 1.14   | 3.25   | 0.36 |
| ENSMUSG00000040728 | Esrp1     | 0.6  | 0.81    | 2.27    | 4.28 | 5.87   | 5.15   | 6.54 |
| ENSMUSG00000043639 | Rbm20     | 0.82 | 0.34    | 1.83    | 3.3  | 2.76   | 1.75   | 0.08 |
| ENSMUSG00000054256 | Msi1      | 0.7  | 0.73    | 1.79    | 3.18 | 6.18   | 2.62   | 1.55 |
| ENSMUSG00000062604 | Srpk2     | 0.6  | 0.56    | 1.98    | 5.14 | 3.96   | 4.37   | 2.89 |
| ENSMUSG00000063919 | Srrm4     | 0.86 | 0.07    | 1.74    | 0.01 | 0.23   | 0.03   | 0.65 |
| ENSMUSG00000102252 | Snrpn     | 0.53 | 0.62    | 2.07    | 0.62 | 5.21   | 2.78   | 2.32 |

**Table D.8:** Overview table summarising the characteristics of selected tissue-restricted splicing factors ( $\tau > 0.5$ ) used for prioritisation, part 2. This part of the table summarises the *DESeq2* fold changes (FC) and the adjusted p-values (p) (Benjamini-Hochberg correction). The abbreviations for the *DESeq2* comparisons are as follows FC-WT/p-WT = immature versus mature mTEC, FC-Aire/p-Aire = *Aire*-knockout versus *Aire*-positive mTEC, FC-imm-Aire/p-imm-Aire = immature versus mature *Aire*-positive mTEC.

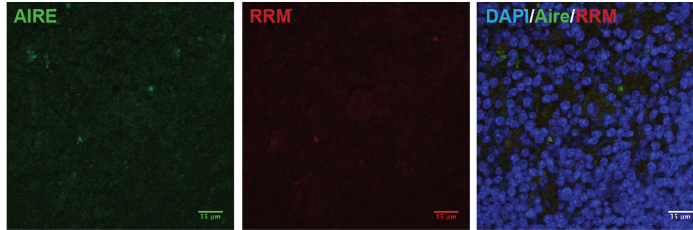
| gene id             | gene name | FC-WT | FC-Aire | FC-imm-Aire | p-WT     | p-Aire   | p-imm-Aire |
|---------------------|-----------|-------|---------|-------------|----------|----------|------------|
| ENSMUSG00000002007  | Srp3      | 0.57  | -0.41   | -0.04       | 2.13E-01 | 4.70E-01 | 9.48E-01   |
| ENSMUSG00000003410  | Elavl3    | 4.8   | -1.29   | 3.13        | 8.64E-33 | 8.17E-07 | 3.25E-10   |
| ENSMUSG00000008489  | Elavl2    | 2.43  | -1.26   | 0.92        | 3.21E-09 | 6.70E-04 | 1.02E-01   |
| ENSMUSG00000008658  | Rbfox1    | -1.07 | -0.19   | -1.11       | 3.57E-52 | 2.32E-02 | 5.86E-56   |
| ENSMUSG000000020409 | Slu7      | -0.23 | 0.1     | 0.02        | 6.36E-04 | 1.88E-01 | 7.23E-01   |
| ENSMUSG000000021047 | Nova1     | 0.03  | 0.23    | 0.34        | 9.47E-01 | 6.50E-01 | 4.46E-01   |
| ENSMUSG000000024268 | Celf4     | 0.87  | -0.95   | -0.46       | 8.93E-05 | 4.36E-04 | 1.19E-01   |
| ENSMUSG000000025576 | Rbfox3    | 5.16  | -2.25   | 2.88        | 2.57E-39 | 3.10E-18 | 8.68E-08   |
| ENSMUSG000000026058 | Khdrbs2   | -1.05 | 0.57    | -0.82       | 7.37E-16 | 6.00E-03 | 2.15E-07   |
| ENSMUSG000000028546 | Elavl4    | 1.59  | -0.77   | 0.42        | 3.19E-08 | 1.05E-02 | 2.69E-01   |
| ENSMUSG000000029279 | Brd1      | 0.62  | -1.26   | -0.76       | 4.27E-02 | 6.01E-04 | 5.85E-02   |
| ENSMUSG000000030411 | Nova2     | -0.57 | -0.1    | -0.84       | 5.11E-02 | 8.27E-01 | 7.90E-03   |
| ENSMUSG000000032297 | Celf6     | 4.11  | -1.02   | 3.08        | 3.20E-40 | 1.83E-07 | 3.23E-18   |
| ENSMUSG000000032423 | Syncip    | -0.5  | 0.2     | -0.5        | 4.37E-12 | 2.23E-03 | 1.38E-17   |
| ENSMUSG000000032940 | Rbm11     | 0.93  | -0.4    | 0.01        | 9.83E-03 | 4.09E-01 | 9.88E-01   |
| ENSMUSG000000036109 | Mbnl3     | 1.01  | 0.32    | 1.11        | 4.38E-38 | 8.50E-06 | 2.73E-48   |
| ENSMUSG000000038132 | Rbm24     | -2.99 | 0.02    | -2.32       | 4.38E-84 | 9.29E-01 | 2.89E-59   |
| ENSMUSG000000040728 | Esrp1     | 0.46  | 0.02    | 0.41        | 2.71E-15 | 8.09E-01 | 5.34E-15   |
| ENSMUSG000000043639 | Rbm20     | 0.92  | 0.32    | 1.47        | 1.71E-18 | 1.57E-03 | 4.23E-50   |
| ENSMUSG000000054256 | Msi1      | 3.47  | 0.57    | 4.12        | 0.00E+00 | 3.35E-16 | 0.00E+00   |
| ENSMUSG000000062604 | Srp2      | -0.74 | -0.15   | -0.88       | 1.09E-26 | 6.09E-02 | 1.18E-45   |
| ENSMUSG000000063919 | Srrm4     | 2.93  | -1.44   | 1.61        | 5.75E-06 | 5.66E-03 | 4.91E-02   |
| ENSMUSG000000102252 | Snrpn     | 2.69  | 0.61    | 3.68        | 4.19E-99 | 9.19E-17 | 1.09E-250  |

# E

## Appendix - Chapter 6



**Figure E.1:** Example gating schemes for live cells. **A:** Gating example for definition of live cells using DAPI. **B:** Example gating for definition of live cells in intracellular stains using AQUA as a live-dead marker.



**Figure E.2:** Immunofluorescence staining without primary antibody in mouse thymus sections (8  $\mu\text{m}$  sections). Images were taken with 63x lens.

**Table E.1:** Overview table of all RNA-sequencing samples from conditional knockout mouse model. The markers used for FACS sorting are given in Table B.1. The replicate name given in the Table was used throughout the analysis. The column 'no-pairs' represent the number of read pairs in M reported by *EstimateLibraryComplexity* from *Picard*.

| Targeted gene | Cre status | Replicate | Samples from TEC | Gender | Litter | no-pairs   |
|---------------|------------|-----------|------------------|--------|--------|------------|
| Rbfox1        | cre+       | R1        | immature; mature | female | A      | 26 M; 28 M |
| Rbfox1        | cre+       | R2        | immature; mature | female | B      | 23 M; 24 M |
| Rbfox1        | cre+       | R3        | immature; mature | male   | B      | 26 M; 32 M |
| Rbfox1        | cre-       | R1        | immature; mature | female | A      | 27 M; 23 M |
| Rbfox1        | cre-       | R2        | immature; mature | female | B      | 23 M; 20 M |
| Rbfox1        | cre-       | R3        | immature; mature | male   | B      | 22 M; 23 M |
| Rbfox2        | cre+       | R1        | immature; mature | male   | C      | 22 M; 26 M |
| Rbfox2        | cre+       | R2        | immature; mature | female | C      | 23 M; 20 M |
| Rbfox2        | cre+       | R3        | immature; mature | female | D      | 25 M; 29 M |
| Rbfox2        | cre-       | R1        | immature; mature | male   | C      | 23 M; 22 M |
| Rbfox2        | cre-       | R2        | immature; mature | female | C      | 22 M; 24 M |
| Rbfox2        | cre-       | R3        | immature; mature | female | D      | 26 M; 26 M |

**Table E.2:** Loadings for the first 4 PCs for the PCA of all quality metrics in RNA-sequencing data from immature and mature mTEC of  $Rbfox1^{lox/lox};cre^{+/-}$  (SKO1) and  $Rbfox1^{lox/lox};cre^{-/-}$  (WT) mice. The explanation and source of each quality metric is described in Table A.2.

| Quality metric             | PC1    | PC2    | PC3    | PC4    |
|----------------------------|--------|--------|--------|--------|
| fraction-spliced           | -0.217 | 0.217  | 0.312  | -0.205 |
| salmon-no-genes-pc         | -0.303 | -0.037 | 0.148  | 0.071  |
| salmon-no-genes            | -0.308 | -0.079 | 0.068  | 0.127  |
| featurecounts-no-genes-pc  | -0.303 | -0.025 | 0.151  | 0.053  |
| featurecounts-no-genes     | -0.308 | -0.069 | 0.085  | 0.115  |
| three-prime-bias           | 0.237  | -0.002 | -0.363 | -0.261 |
| nreads-uniq-map-genome     | -0.023 | -0.456 | 0.230  | -0.227 |
| no-pairs                   | 0.030  | -0.457 | 0.233  | -0.193 |
| pct-duplication            | 0.289  | -0.069 | 0.140  | -0.325 |
| library-size               | -0.299 | -0.085 | -0.068 | 0.242  |
| pct-reads-aligned-in-pairs | 0.026  | 0.252  | 0.447  | 0.040  |
| median-insert-size         | -0.215 | -0.041 | -0.363 | -0.400 |
| pct-mrna                   | 0.237  | 0.262  | 0.188  | -0.123 |
| pct-coding                 | -0.023 | 0.376  | 0.339  | -0.350 |
| pct-reads-aligned          | -0.270 | 0.142  | -0.154 | -0.339 |
| total-reads                | 0.038  | -0.457 | 0.232  | -0.181 |
| pct-adapter                | 0.279  | -0.043 | 0.151  | 0.362  |
| pct-pf-reads-aligned-hq    | -0.307 | 0.066  | -0.035 | -0.145 |

**Table E.3:** Loadings for the first 4 PCs for the PCA of all quality metrics in RNA-sequencing data from immature and mature mTEC of  $Rbfox2^{lox/lox};cre^{+/-}$  (SKO2) and  $Rbfox2^{lox/lox};cre^{-/-}$  (WT) mice. The explanation and source of each quality metric is described in Table A.2.

| Quality metric             | PC1    | PC2    | PC3    | PC4    |
|----------------------------|--------|--------|--------|--------|
| fraction-spliced           | 0.271  | -0.070 | 0.170  | -0.006 |
| salmon-no-genes-pc         | 0.294  | 0.005  | -0.002 | 0.059  |
| salmon-no-genes            | 0.295  | 0.022  | 0.020  | 0.066  |
| featurecounts-no-genes-pc  | 0.294  | 0.007  | 0.003  | 0.050  |
| featurecounts-no-genes     | 0.295  | 0.025  | 0.022  | 0.049  |
| three-prime-bias           | -0.261 | 0.150  | 0.020  | -0.346 |
| nreads-uniq-map-genome     | 0.113  | -0.453 | -0.180 | 0.023  |
| no-pairs                   | 0.062  | -0.478 | -0.207 | 0.034  |
| pct-duplication            | -0.262 | -0.190 | -0.132 | 0.138  |
| library-size               | 0.285  | 0.060  | 0.112  | -0.148 |
| pct-reads-aligned-in-pairs | 0.205  | 0.071  | -0.515 | -0.055 |
| median-insert-size         | 0.091  | -0.174 | 0.679  | 0.275  |
| pct-mrna                   | -0.258 | -0.133 | 0.106  | -0.500 |
| pct-coding                 | 0.013  | -0.430 | 0.288  | -0.443 |
| pct-reads-aligned          | 0.271  | 0.031  | -0.024 | -0.386 |
| total-reads                | 0.054  | -0.481 | -0.201 | 0.041  |
| pct-adapter                | -0.252 | -0.166 | 0.048  | 0.381  |
| pct-pf-reads-aligned-hq    | 0.295  | 0.009  | 0.032  | -0.065 |

**Table E.4:** Summary of differentially expressed genes between  $Rbfox1^{lox/lox}:cre^{+/-}$  mice and  $Rbfox1^{lox/lox}:cre^{-/-}$  mice (including all genes with fold change  $> 1.5$ ). The log2 fold change (log2FC), mean expression level (expr) and adjusted p-value (p.adj) were calculated using *featureCounts* and *DESeq2*. The column 'mTEC' indicates which TEC type the comparison was performed on, while the direction represents whether the gene was upregulated (up-SKO) or downregulated (down-SKO) in the SKO1 sample. The different gene categories such as *Aire*-regulated genes (Aire) or splicing factor (SF) were included. The tissue-restriction was defined by  $\tau > 0.7$  (column 'tau'). Finally, the gene biotype from Ensembl was indicated.

| gene name | gene id            | log2FC | expr   | p.adj | mTEC          | direction | Aire reg | SF    | tau            | biotype        |
|-----------|--------------------|--------|--------|-------|---------------|-----------|----------|-------|----------------|----------------|
| Slc38a5   | ENSMUSG00000031170 | 7.317  | 3.876  | 0.005 | immature-mTEC | down-SKO  | TRUE     | FALSE | 0.751          | protein-coding |
| Ptk2      | ENSMUSG00000022607 | 0.799  | 9.499  | 0.001 | immature-mTEC | down-SKO  | FALSE    | FALSE | 0.21           | protein-coding |
| EPPK1     | ENSMUSG00000115388 | -0.593 | 10.636 | 0.004 | immature-mTEC | up-SKO    | TRUE     | FALSE | 0.612          | protein-coding |
| Foxn1     | ENSMUSG00000002057 | -0.767 | 11.226 | 0     | immature-mTEC | up-SKO    | FALSE    | FALSE | 0.977          | protein-coding |
| Fetub     | ENSMUSG00000022871 | -1.201 | 7.54   | 0.013 | immature-mTEC | up-SKO    | FALSE    | FALSE | 0.776          | protein-coding |
| Ntng1     | ENSMUSG00000059857 | -2.054 | 5.627  | 0.029 | immature-mTEC | up-SKO    | FALSE    | FALSE | 0.893          | protein-coding |
| Unc119    | ENSMUSG00000002058 | -2.09  | 10.255 | 0     | immature-mTEC | up-SKO    | FALSE    | FALSE | 0.315          | protein-coding |
| Slc13a2os | ENSMUSG00000085747 | -2.335 | 8.827  | 0     | immature-mTEC | up-SKO    | FALSE    | FALSE | not determined | antisense-RNA  |
| Aldob     | ENSMUSG00000028307 | -2.856 | 5.889  | 0.005 | immature-mTEC | up-SKO    | TRUE     | FALSE | 0.742          | protein-coding |
| Slc13a2   | ENSMUSG00000001095 | -2.924 | 6.226  | 0     | immature-mTEC | up-SKO    | FALSE    | FALSE | 0.819          | protein-coding |
| Rbfox1    | ENSMUSG00000008658 | 0.974  | 8.739  | 0     | mature-mTEC   | down-SKO  | FALSE    | TRUE  | 0.814          | protein-coding |
| Ptk2      | ENSMUSG00000022607 | 0.785  | 9.12   | 0.001 | mature-mTEC   | down-SKO  | FALSE    | FALSE | 0.21           | protein-coding |
| Gnat3     | ENSMUSG00000028777 | -0.919 | 7.725  | 0.025 | mature-mTEC   | up-SKO    | FALSE    | FALSE | 0.987          | protein-coding |
| Gnb3      | ENSMUSG00000023439 | -0.957 | 8.047  | 0.001 | mature-mTEC   | up-SKO    | FALSE    | FALSE | 0.874          | protein-coding |
| Slc13a2   | ENSMUSG00000001095 | -1.629 | 7.685  | 0     | mature-mTEC   | up-SKO    | FALSE    | FALSE | 0.819          | protein-coding |
| Unc119    | ENSMUSG00000002058 | -2.109 | 8.173  | 0     | mature-mTEC   | up-SKO    | FALSE    | FALSE | 0.315          | protein-coding |
| Slc13a2os | ENSMUSG00000085747 | -2.25  | 10.045 | 0.001 | mature-mTEC   | up-SKO    | FALSE    | FALSE | not determined | antisense-RNA  |

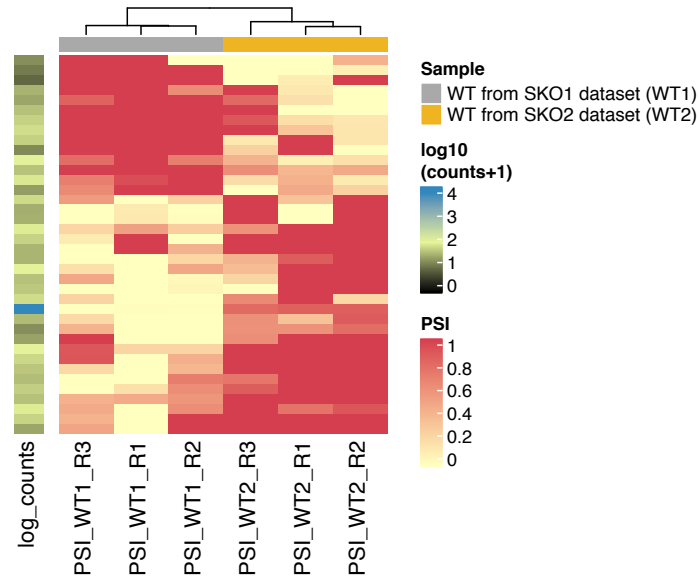


**Table E.5:** Summary of differentially expressed genes between  $Rbfox2^{lox/lox};cre^{+/-}$  mice and  $Rbfox2^{lox/lox};cre^{-/-}$  mice (including all genes with fold change  $> 1.5$ ). The log2 fold change (log2FC), mean expression level (expr) and adjusted p-value (p.adj) were calculated using *featureCounts* and *DESeq2*. The column 'mTEC' indicates which TEC type the comparison was performed on, while the direction represents whether the gene was upregulated (up-SKO) or downregulated (down-SKO) in the SKO2 sample. The different gene categories such as *Aire*-regulated genes (Aire) or splicing factor (SF) were included. The tissue-restriction was defined by  $\tau > 0.7$  (column 'tau'). Finally, the gene biotype from Ensembl was indicated.

| gene name     | gene id             | log2FC | expr   | p.adj | mTEC          | direction | Aire reg | SF    | tau            | biotype          |
|---------------|---------------------|--------|--------|-------|---------------|-----------|----------|-------|----------------|------------------|
| Ghrl          | ENSMUSG000000064177 | 7.396  | 3.979  | 0.004 | immature-mTEC | down-SKO  | TRUE     | FALSE | 0.87           | protein-coding   |
| Iqca          | ENSMUSG000000026301 | 7.244  | 3.827  | 0.002 | immature-mTEC | down-SKO  | TRUE     | FALSE | 0.88           | protein-coding   |
| Clec4a1       | ENSMUSG000000049037 | 7.123  | 3.706  | 0.008 | immature-mTEC | down-SKO  | FALSE    | FALSE | 0.601          | protein-coding   |
| Dppa3         | ENSMUSG000000046323 | 6.982  | 3.565  | 0.011 | immature-mTEC | down-SKO  | TRUE     | FALSE | 0.985          | protein-coding   |
| Dnah7a        | ENSMUSG000000096141 | 6.92   | 3.504  | 0.014 | immature-mTEC | down-SKO  | TRUE     | FALSE | 0.92           | protein-coding   |
| Emilin3       | ENSMUSG000000050700 | 6.744  | 3.326  | 0.037 | immature-mTEC | down-SKO  | FALSE    | FALSE | 0.809          | protein-coding   |
| Gm13785       | ENSMUSG000000081730 | 6.673  | 3.256  | 0.034 | immature-mTEC | down-SKO  | FALSE    | FALSE | not determined | proc.-pseudogene |
| Pitx2         | ENSMUSG000000028023 | 6.566  | 3.149  | 0.045 | immature-mTEC | down-SKO  | TRUE     | FALSE | 0.771          | protein-coding   |
| Gldn          | ENSMUSG000000046167 | 6.377  | 3.937  | 0.008 | immature-mTEC | down-SKO  | TRUE     | FALSE | 0.804          | protein-coding   |
| Lrrc10b       | ENSMUSG000000090291 | 4.751  | 4.247  | 0.025 | immature-mTEC | down-SKO  | TRUE     | FALSE | 0.706          | protein-coding   |
| 5430416N02Rik | ENSMUSG000000097772 | 2.866  | 5.286  | 0.005 | immature-mTEC | down-SKO  | FALSE    | FALSE | not determined | lincRNA          |
| Rbfox2        | ENSMUSG000000033565 | 1.997  | 9.573  | 0     | immature-mTEC | down-SKO  | FALSE    | TRUE  | 0.25           | protein-coding   |
| Ptk2          | ENSMUSG000000022607 | 1.112  | 9.488  | 0     | immature-mTEC | down-SKO  | FALSE    | FALSE | 0.21           | protein-coding   |
| Sspn          | ENSMUSG000000030255 | 1.01   | 8.577  | 0     | immature-mTEC | down-SKO  | FALSE    | FALSE | 0.565          | protein-coding   |
| Mybpc1        | ENSMUSG000000020061 | 0.949  | 8.779  | 0     | immature-mTEC | down-SKO  | FALSE    | FALSE | 0.661          | protein-coding   |
| 1810041L15Rik | ENSMUSG000000062760 | 0.689  | 9.723  | 0     | immature-mTEC | down-SKO  | FALSE    | FALSE | 0.748          | protein-coding   |
| Itga3         | ENSMUSG000000001507 | 0.626  | 11.745 | 0     | immature-mTEC | down-SKO  | FALSE    | FALSE | 0.35           | protein-coding   |
| Mgll          | ENSMUSG000000033174 | -0.591 | 9.707  | 0.006 | immature-mTEC | up-SKO    | FALSE    | FALSE | 0.341          | protein-coding   |
| Ctsc          | ENSMUSG000000030560 | -0.592 | 10.886 | 0.008 | immature-mTEC | up-SKO    | FALSE    | FALSE | 0.326          | protein-coding   |
| Cdkn1a        | ENSMUSG000000023067 | -0.626 | 10.228 | 0.017 | immature-mTEC | up-SKO    | FALSE    | FALSE | 0.369          | protein-coding   |
| Efemp1        | ENSMUSG000000020467 | -0.629 | 8.883  | 0.017 | immature-mTEC | up-SKO    | TRUE     | FALSE | 0.345          | protein-coding   |
| Nrp1          | ENSMUSG000000025810 | -0.641 | 9.591  | 0.013 | immature-mTEC | up-SKO    | FALSE    | FALSE | 0.39           | protein-coding   |
| Lars2         | ENSMUSG000000035202 | -0.644 | 12.797 | 0     | immature-mTEC | up-SKO    | FALSE    | FALSE | 0.221          | protein-coding   |
| Hpse2         | ENSMUSG000000074852 | -0.66  | 8.717  | 0.037 | immature-mTEC | up-SKO    | FALSE    | FALSE | 0.865          | protein-coding   |
| Tspyl4        | ENSMUSG000000039485 | -0.666 | 8.716  | 0.029 | immature-mTEC | up-SKO    | FALSE    | FALSE | 0.669          | protein-coding   |
| Sulf2         | ENSMUSG00000006800  | -0.707 | 9.585  | 0     | immature-mTEC | up-SKO    | FALSE    | FALSE | 0.38           | protein-coding   |
| Mylk          | ENSMUSG000000022836 | -0.727 | 9.616  | 0.014 | immature-mTEC | up-SKO    | TRUE     | FALSE | 0.48           | protein-coding   |
| Prrx1         | ENSMUSG000000026586 | -0.73  | 8.424  | 0.047 | immature-mTEC | up-SKO    | TRUE     | FALSE | 0.519          | protein-coding   |
| Pcolce        | ENSMUSG000000029718 | -0.751 | 9.874  | 0.016 | immature-mTEC | up-SKO    | TRUE     | FALSE | 0.399          | protein-coding   |
| Nupr1         | ENSMUSG000000030717 | -0.758 | 8.641  | 0.023 | immature-mTEC | up-SKO    | FALSE    | FALSE | 0.441          | protein-coding   |
| Cav1          | ENSMUSG000000007655 | -0.796 | 9.67   | 0     | immature-mTEC | up-SKO    | TRUE     | FALSE | 0.429          | protein-coding   |
| Pmp22         | ENSMUSG000000018217 | -0.824 | 8.944  | 0.04  | immature-mTEC | up-SKO    | TRUE     | FALSE | 0.398          | protein-coding   |
| Nid1          | ENSMUSG000000005397 | -0.864 | 9.05   | 0.045 | immature-mTEC | up-SKO    | TRUE     | FALSE | 0.302          | protein-coding   |
| Frk           | ENSMUSG000000019779 | -0.872 | 8.404  | 0.03  | immature-mTEC | up-SKO    | FALSE    | FALSE | 0.606          | protein-coding   |
| Lama2         | ENSMUSG000000019899 | -0.878 | 9.143  | 0.001 | immature-mTEC | up-SKO    | FALSE    | FALSE | 0.404          | protein-coding   |
| Fabp4         | ENSMUSG000000062515 | -0.885 | 9.999  | 0     | immature-mTEC | up-SKO    | FALSE    | FALSE | 0.486          | protein-coding   |
| Pcdh15        | ENSMUSG000000052613 | -0.895 | 7.917  | 0.008 | immature-mTEC | up-SKO    | FALSE    | FALSE | 0.859          | protein-coding   |
| Aadac         | ENSMUSG000000027761 | -0.911 | 7.911  | 0.011 | immature-mTEC | up-SKO    | TRUE     | FALSE | 0.779          | protein-coding   |
| Nrp2          | ENSMUSG000000025969 | -0.916 | 8.979  | 0.001 | immature-mTEC | up-SKO    | FALSE    | FALSE | 0.458          | protein-coding   |
| Flt1          | ENSMUSG000000029648 | -0.928 | 8.198  | 0.019 | immature-mTEC | up-SKO    | TRUE     | FALSE | 0.593          | protein-coding   |
| mt-Nd6        | ENSMUSG000000064368 | -0.964 | 7.854  | 0.016 | immature-mTEC | up-SKO    | FALSE    | FALSE | 0.348          | protein-coding   |
| Ddr2          | ENSMUSG000000026674 | -0.985 | 9.068  | 0.01  | immature-mTEC | up-SKO    | TRUE     | FALSE | 0.522          | protein-coding   |
| Mest          | ENSMUSG000000051855 | -0.987 | 8.072  | 0.047 | immature-mTEC | up-SKO    | TRUE     | FALSE | 0.538          | protein-coding   |
| Heg1          | ENSMUSG000000075254 | -1.013 | 8.553  | 0.043 | immature-mTEC | up-SKO    | TRUE     | FALSE | 0.376          | protein-coding   |
| Hdac9         | ENSMUSG000000004698 | -1.017 | 7.518  | 0.021 | immature-mTEC | up-SKO    | FALSE    | FALSE | 0.695          | protein-coding   |
| 5330417C22Rik | ENSMUSG000000040412 | -1.047 | 9.066  | 0.007 | immature-mTEC | up-SKO    | FALSE    | FALSE | 0.73           | protein-coding   |
| Apoc1         | ENSMUSG000000040564 | -1.052 | 8.179  | 0.016 | immature-mTEC | up-SKO    | TRUE     | FALSE | 0.686          | protein-coding   |

|           |                      |        |        |       |               |          |       |       |                |                |
|-----------|----------------------|--------|--------|-------|---------------|----------|-------|-------|----------------|----------------|
| Jak3      | ENSMUSG000000031805  | -1.086 | 10.395 | 0     | immature-mTEC | up-SKO   | FALSE | FALSE | 0.362          | protein-coding |
| Gm26917   | ENSMUSG000000097971  | -1.09  | 12.083 | 0     | immature-mTEC | up-SKO   | FALSE | FALSE | not determined | lincRNA        |
| Col8a1    | ENSMUSG000000068196  | -1.118 | 8.349  | 0.008 | immature-mTEC | up-SKO   | TRUE  | FALSE | 0.554          | protein-coding |
| Sftpd     | ENSMUSG000000021795  | -1.127 | 7.425  | 0.041 | immature-mTEC | up-SKO   | TRUE  | FALSE | 0.911          | protein-coding |
| Cacna1c   | ENSMUSG000000051331  | -1.155 | 7.178  | 0.009 | immature-mTEC | up-SKO   | TRUE  | FALSE | 0.598          | protein-coding |
| Rbm20     | ENSMUSG000000043639  | -1.16  | 7.346  | 0.013 | immature-mTEC | up-SKO   | FALSE | TRUE  | 0.817          | protein-coding |
| Aqp1      | ENSMUSG000000004655  | -1.167 | 10.149 | 0     | immature-mTEC | up-SKO   | TRUE  | FALSE | 0.355          | protein-coding |
| Tmem132c  | ENSMUSG000000034324  | -1.208 | 7.249  | 0.027 | immature-mTEC | up-SKO   | TRUE  | FALSE | 0.689          | protein-coding |
| Adgra2    | ENSMUSG000000031486  | -1.213 | 7.776  | 0.04  | immature-mTEC | up-SKO   | TRUE  | FALSE | 0.397          | protein-coding |
| Krt15     | ENSMUSG000000054146  | -1.217 | 9.112  | 0     | immature-mTEC | up-SKO   | TRUE  | FALSE | 0.773          | protein-coding |
| Adgre5    | ENSMUSG000000002885  | -1.229 | 7.386  | 0.009 | immature-mTEC | up-SKO   | FALSE | FALSE | 0.433          | protein-coding |
| Pi15      | ENSMUSG000000067780  | -1.269 | 7.216  | 0.05  | immature-mTEC | up-SKO   | TRUE  | FALSE | 0.575          | protein-coding |
| Atp2a1    | ENSMUSG000000030730  | -1.275 | 7.928  | 0.016 | immature-mTEC | up-SKO   | TRUE  | FALSE | 0.83           | protein-coding |
| Cdh5      | ENSMUSG000000031871  | -1.28  | 8.96   | 0     | immature-mTEC | up-SKO   | TRUE  | FALSE | 0.495          | protein-coding |
| Scara5    | ENSMUSG000000022032  | -1.282 | 7.353  | 0.01  | immature-mTEC | up-SKO   | TRUE  | FALSE | 0.587          | protein-coding |
| Dchs2     | ENSMUSG000000102692  | -1.286 | 7.513  | 0.02  | immature-mTEC | up-SKO   | TRUE  | FALSE | 0.845          | protein-coding |
| Kdr       | ENSMUSG0000000062960 | -1.303 | 7.502  | 0.025 | immature-mTEC | up-SKO   | TRUE  | FALSE | 0.495          | protein-coding |
| Myog      | ENSMUSG000000026459  | -1.316 | 8.987  | 0     | immature-mTEC | up-SKO   | FALSE | FALSE | 0.955          | protein-coding |
| Cd300lg   | ENSMUSG000000017309  | -1.325 | 7.78   | 0.019 | immature-mTEC | up-SKO   | TRUE  | FALSE | 0.475          | protein-coding |
| Dusp3     | ENSMUSG000000003518  | -1.329 | 7.381  | 0.002 | immature-mTEC | up-SKO   | FALSE | FALSE | 0.266          | protein-coding |
| Dsg1b     | ENSMUSG0000000061928 | -1.356 | 9.095  | 0     | immature-mTEC | up-SKO   | FALSE | FALSE | 0.923          | protein-coding |
| Epha3     | ENSMUSG000000052504  | -1.384 | 7.72   | 0     | immature-mTEC | up-SKO   | FALSE | FALSE | 0.689          | protein-coding |
| Actc1     | ENSMUSG0000000068614 | -1.423 | 8.041  | 0.008 | immature-mTEC | up-SKO   | TRUE  | FALSE | 0.815          | protein-coding |
| Cavin2    | ENSMUSG000000045954  | -1.453 | 7.279  | 0.025 | immature-mTEC | up-SKO   | TRUE  | FALSE | 0.474          | protein-coding |
| Cd36      | ENSMUSG000000002944  | -1.463 | 9.808  | 0     | immature-mTEC | up-SKO   | TRUE  | FALSE | 0.463          | protein-coding |
| Plvap     | ENSMUSG000000034845  | -1.509 | 8.79   | 0     | immature-mTEC | up-SKO   | TRUE  | FALSE | 0.372          | protein-coding |
| Sema3e    | ENSMUSG000000063531  | -1.566 | 6.108  | 0.047 | immature-mTEC | up-SKO   | TRUE  | FALSE | 0.786          | protein-coding |
| Scarb1    | ENSMUSG000000037936  | -1.576 | 8.793  | 0     | immature-mTEC | up-SKO   | TRUE  | FALSE | 0.428          | protein-coding |
| Klhl41    | ENSMUSG000000075307  | -1.584 | 7.548  | 0.01  | immature-mTEC | up-SKO   | FALSE | FALSE | 0.764          | protein-coding |
| Muc5b     | ENSMUSG0000000066108 | -1.633 | 9.499  | 0.008 | immature-mTEC | up-SKO   | TRUE  | FALSE | 0.896          | protein-coding |
| Olfml2b   | ENSMUSG000000038463  | -1.696 | 6.764  | 0.009 | immature-mTEC | up-SKO   | TRUE  | FALSE | 0.435          | protein-coding |
| Kcnj16    | ENSMUSG000000051497  | -1.729 | 5.899  | 0.02  | immature-mTEC | up-SKO   | TRUE  | FALSE | 0.829          | protein-coding |
| Unc119    | ENSMUSG000000002058  | -1.786 | 10.419 | 0     | immature-mTEC | up-SKO   | FALSE | FALSE | 0.315          | protein-coding |
| Ltb4r2    | ENSMUSG000000040432  | -1.803 | 7.288  | 0     | immature-mTEC | up-SKO   | FALSE | FALSE | 0.912          | protein-coding |
| Svep1     | ENSMUSG000000028369  | -1.836 | 7.502  | 0.001 | immature-mTEC | up-SKO   | TRUE  | FALSE | 0.485          | protein-coding |
| Eef1a2    | ENSMUSG000000016349  | -2.142 | 5.295  | 0.022 | immature-mTEC | up-SKO   | FALSE | FALSE | 0.778          | protein-coding |
| Sidtl     | ENSMUSG000000022696  | -2.205 | 5.669  | 0.04  | immature-mTEC | up-SKO   | TRUE  | FALSE | 0.591          | protein-coding |
| Prkg2     | ENSMUSG0000000029334 | -2.401 | 6.131  | 0.008 | immature-mTEC | up-SKO   | FALSE | FALSE | 0.825          | protein-coding |
| Myct1     | ENSMUSG000000046916  | -2.58  | 5.065  | 0.029 | immature-mTEC | up-SKO   | TRUE  | FALSE | 0.685          | protein-coding |
| Slc6a17   | ENSMUSG000000027894  | -2.629 | 5.064  | 0.01  | immature-mTEC | up-SKO   | TRUE  | FALSE | 0.766          | protein-coding |
| Emcn      | ENSMUSG000000054690  | -2.643 | 6.075  | 0     | immature-mTEC | up-SKO   | TRUE  | FALSE | 0.456          | protein-coding |
| Lepr      | ENSMUSG000000057722  | -2.676 | 6.524  | 0     | immature-mTEC | up-SKO   | TRUE  | FALSE | 0.516          | protein-coding |
| Slc13a2os | ENSMUSG000000085747  | -2.882 | 8.365  | 0     | immature-mTEC | up-SKO   | FALSE | FALSE | not determined | antisense-RNA  |
| Muc5ac    | ENSMUSG000000037974  | -2.968 | 8.385  | 0.047 | immature-mTEC | up-SKO   | TRUE  | FALSE | 0.921          | protein-coding |
| Ano5      | ENSMUSG000000055489  | -3.242 | 5.137  | 0.019 | immature-mTEC | up-SKO   | TRUE  | FALSE | 0.793          | protein-coding |
| Egf       | ENSMUSG000000028017  | -6.72  | 3.226  | 0.036 | immature-mTEC | up-SKO   | FALSE | FALSE | 0.799          | protein-coding |
| Trp53cor1 | ENSMUSG000000085912  | -6.743 | 3.249  | 0.041 | immature-mTEC | up-SKO   | FALSE | FALSE | not determined | lincRNA        |
| Il27      | ENSMUSG000000044701  | -6.757 | 3.262  | 0.025 | immature-mTEC | up-SKO   | FALSE | FALSE | 0.966          | protein-coding |
| Gucylb2   | ENSMUSG000000021933  | -6.866 | 3.372  | 0.02  | immature-mTEC | up-SKO   | FALSE | FALSE | 0.928          | protein-coding |
| Pyroxd2   | ENSMUSG000000060224  | -6.884 | 3.389  | 0.023 | immature-mTEC | up-SKO   | TRUE  | FALSE | 0.494          | protein-coding |
| Fam83a    | ENSMUSG000000051225  | -6.965 | 3.47   | 0.01  | immature-mTEC | up-SKO   | TRUE  | FALSE | 0.671          | protein-coding |
| Krt90     | ENSMUSG000000048699  | -7.387 | 3.893  | 0.001 | immature-mTEC | up-SKO   | TRUE  | FALSE | 0.983          | protein-coding |
| Gm17250   | ENSMUSG000000097235  | 2.864  | 4.384  | 0.014 | mature-mTEC   | down-SKO | FALSE | FALSE | not determined | lincRNA        |
| Rbfox2    | ENSMUSG000000033565  | 2.197  | 8.984  | 0     | mature-mTEC   | down-SKO | FALSE | TRUE  | 0.25           | protein-coding |
| Ptk2      | ENSMUSG000000022607  | 0.878  | 8.946  | 0     | mature-mTEC   | down-SKO | FALSE | FALSE | 0.21           | protein-coding |
| Ptbp2     | ENSMUSG000000028134  | 0.797  | 7.531  | 0.027 | mature-mTEC   | down-SKO | FALSE | TRUE  | 0.364          | protein-coding |
| Itga3     | ENSMUSG000000001507  | 0.795  | 11.659 | 0     | mature-mTEC   | down-SKO | FALSE | FALSE | 0.35           | protein-coding |
| Irgm2     | ENSMUSG000000069874  | 0.721  | 8.57   | 0     | mature-mTEC   | down-SKO | FALSE | FALSE | 0.372          | protein-coding |
| Lypd2     | ENSMUSG000000022595  | 0.626  | 9.193  | 0.019 | mature-mTEC   | down-SKO | TRUE  | FALSE | 0.838          | protein-coding |

|           |                     |        |        |       |             |          |       |       |                |                             |
|-----------|---------------------|--------|--------|-------|-------------|----------|-------|-------|----------------|-----------------------------|
| Acta2     | ENSMUSG00000035783  | 0.617  | 9.762  | 0     | mature-mTEC | down-SKO | TRUE  | FALSE | 0.409          | protein-coding              |
| Flg       | ENSMUSG000000102439 | -0.591 | 9.102  | 0.03  | mature-mTEC | up-SKO   | FALSE | FALSE | 0.911          | protein-coding              |
| Fsd11     | ENSMUSG000000054752 | -0.626 | 8.919  | 0.007 | mature-mTEC | up-SKO   | FALSE | FALSE | 0.513          | protein-coding              |
| Siglec15  | ENSMUSG000000091055 | -0.676 | 8.528  | 0.004 | mature-mTEC | up-SKO   | FALSE | FALSE | 0.924          | protein-coding              |
| Rab10os   | ENSMUSG000000079179 | -0.894 | 8.324  | 0     | mature-mTEC | up-SKO   | FALSE | FALSE | not determined | antisense-RNA               |
| Gm23935   | ENSMUSG000000076258 | -0.906 | 13.224 | 0.033 | mature-mTEC | up-SKO   | FALSE | FALSE | not determined | miRNA                       |
| Arc       | ENSMUSG000000022602 | -0.91  | 9.186  | 0     | mature-mTEC | up-SKO   | FALSE | FALSE | 0.774          | protein-coding              |
| Psg21     | ENSMUSG000000070796 | -1.415 | 6.256  | 0.002 | mature-mTEC | up-SKO   | TRUE  | FALSE | 0.973          | protein-coding              |
| Slc13a2   | ENSMUSG000000001095 | -1.845 | 7.391  | 0     | mature-mTEC | up-SKO   | FALSE | FALSE | 0.819          | protein-coding              |
| Unc119    | ENSMUSG000000002058 | -2.242 | 8.472  | 0     | mature-mTEC | up-SKO   | FALSE | FALSE | 0.315          | protein-coding              |
| Slc13a2os | ENSMUSG000000085747 | -2.353 | 9.935  | 0     | mature-mTEC | up-SKO   | FALSE | FALSE | not determined | antisense-RNA               |
| Gm8995    | ENSMUSG000000063286 | -3.005 | 4.366  | 0.003 | mature-mTEC | up-SKO   | FALSE | FALSE | not determined | transcr.-unproc.-pseudogene |



**Figure E.3:** Heatmap of PSI values for skipped exon events in the  $cre^{-/-}$  replicates from the SKO1 and SKO2 datasets. The skipped exon events were filtered for a minimum absolute delta PSI of 0.5 (n=38 events). The left column indicates the combined counts for inclusion and exclusion of the respective event as determined by *rMATS*.