



Universidade do Minho
Escola de Engenharia

Ricardo Júdice Miragaia

Immune Tolerance Dissected by Single-Cell Transcriptomics

Ricardo Júdice Miragaia Immune Tolerance Dissected by Single-Cell Transcriptomics

UMinho | 2017

MIT Portugal

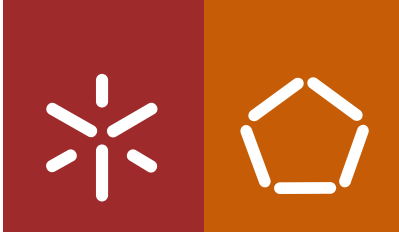
FCT

Fundação para a Ciência e a Tecnologia

MINISTÉRIO DA EDUCAÇÃO E CIÊNCIA



September 2017



Universidade do Minho
Escola de Engenharia

Ricardo Júdice Miragaia

Immune Tolerance Dissected by Single-Cell Transcriptomics

Doctoral Thesis
Doctoral Degree in Bioengineering

This work was executed under the supervision of
Professor Sarah A. Teichmann
Professor Eugénio M. F. Campos Ferreira

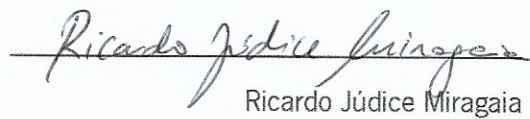
September 2017

STATEMENT OF INTEGRITY

I hereby declare having conducted my thesis with integrity. I confirm that I have not used plagiarism or any form of falsification of results in the process of the thesis elaboration.

I further declare that I have fully acknowledged the Code of Ethical Conduct of the University of Minho.

University of Minho, 11th September 2017


Ricardo Júdice Miragaia

Que Saudades, Pai.

ACKNOWLEDGEMENTS

Muitas pessoas, mais do que as que posso contar, foram importantes nesta etapa marcante e tão exigente. Aos Amigos que ficaram em Portugal, um sincero agradecimento por torcerem por mim, por me apoiarem e por, apesar da distância, me terem feito sempre sentir incluído.

À família que nasceu em Cambridge, que acompanharam mais de perto ainda todas as felicidades e dificuldades destes anos, obrigado por me guiarem e aturarem com tanta paciência, e obrigado por me terem proporcionado uma nova Casa.

Thank you Sarah, for the opportunity, for pushing me forward and for all the support throughout these years. Caro Professor Eugénio, obrigado pela paciência e apoio com que me brindou, mesmo à distância. Fico imensamente agradecido pela sua Amizade. I also have to thank my PhD program, MIT Portugal, and to Fundação para a Ciência e Tecnologia (Portugal) for the PhD Fellowship (SFRH/BD/51950/2012) that allowed this work to be developed.

Um agradecimento especial ao meu caro amigo, irmão e padrinho Kiko. Mais uma vez, foste uma peça fundamental no meu sucesso e felicidade. Obrigado por tudo!

Aos meus Pais, os melhores do Mundo, pela força, carinho e apoio incondicionais. Sempre me deram tudo deles, quando podiam e quando não podiam. Muito obrigado aos dois por me terem feito o que sou hoje. Gosto tanto de vocês. Pai, consegui acabar! Agora sim, tenho uma história completa para te contar... ias gostar! Tens feito e vais sempre fa zer muita falta. Mãe, muita força nesta nova fase. Sei que vais vencer tudo isto, e estarei aqui para ti, como sempre estiveste para mim.

Vitamina, sem ti não teria chegado ao fim do PhD. E sem o PhD, não teria chegado a ti. Foste sem dúvida o melhor que estes anos me trouxeram. Fizeste-me feliz novamente, muito feliz. Obrigado por cuidares de mim, obrigado por tudo o que puseste de parte e tudo o que fizeste, especialmente neste último ano. Obrigado por aturares todos os maus feitos. Obrigado por me ouvires. Estou orgulhoso por nós, mas especialmente, por ti. Admiro-te e Amo-Te. Casemo-nos já!

RESUMO

O conjunto de mecanismos que protege o organismo de agentes patogénicos tem de ser agressivo o suficiente para destruir microorganismos e parasitas. Consequentemente, isto torna-o potencialmente perigoso para os tecidos do hospedeiro. Assim sendo, têm de existir estratégias e pontos de controlo que evitem que o sistema imunitário reconheça antígenos do organismo ou antígenos externos mas inofensivos. Esta capacidade de não responder a determinados antígenos é designada por tolerância imunitária, e tem de ser assegurada especialmente a nível células T, uma vez que são elas as que coordenam grande parte das respostas imunitárias. Parte deste controlo é exercido durante o desenvolvimento das células T no timo, por um processo de selecção negativa conduzido por células tímicas epiteliais medulares (mTEC). As mTEC são especialmente intrigantes pela sua capacidade de expressar ectopicamente genes próprios de outros tecidos (antígenos restritos a tecidos, TRAs) para o exibirem às células T em desenvolvimento. A intensidade com que as células T reconhecem os péptidos apresentados pelas mTEC condiciona o destino das primeiras: se a ligação for fraca, podem sobreviver uma vez que não são auto-reactivas; se a ligação for muito forte, são eliminadas; ligações com força intermédia podem ser encaminhadas para um tipo celular particular, as células T reguladoras (Treg). Estas últimas actuam posteriormente nos tecidos periféricos, exercendo uma função anti-inflamatória, impedindo nomeadamente que células T auto-reactivas que possam ter escapado à selecção negativa montem uma reacção imunitária contra antígenos próprios.

Ambos estes tipos celulares, mTEC e Treg, têm sido estudados exaustivamente com o objectivo geral de compreender e potencialmente manipular o processo de tolerância imunitária. Além de ambas terem em comum um papel relevante no mesmo processo, e de defeitos em qualquer uma delas originarem problemas de auto-imunidade e inflamações crónicas, ambas as populações carecem de uma caracterização exaustiva com resolução célula-a-célula que permita identificar potenciais subpopulações e seus fenótipos. Assim sendo, usei métodos recentes para analisar o transcriptoma de células individuais (*single-cell RNA sequencing*, scRNA-seq) de forma a esclarecer a estrutura, desenvolvimento e funções de populações de mTECs e Tregs.

Em relação às mTEC, isolei células de ratinho sem enriquecer para nenhuma subpopulação conhecida, e obtive dados de scRNA-seq usando uma plataforma microfluídica comercial. Com estes dados, por agrupamento hierárquico, identifiquei três subpopulações dentro do compartimento das mTEC - jTEC (*Pdpr*-positivas), mTEChi (níveis de *Aire* elevados) e mTEClo (níveis de *Aire* reduzidos) -

que aparentam corresponder a fases consecutivas do desenvolvimento deste tipo celular, de acordo com a expressão do gene regulador *Aire* e a capacidade das mTEC em expressar TRAs. Foi a primeira vez que estas subpopulações foram identificadas e caracterizadas sem recurso a marcadores adicionais, como MHCII, durante a citometria de fluxo. Esta estratégia dá uma nova perspectiva nomeadamente do estado pós-*Aire* (mTEC_{lo}), que aparece como um mediador activo e potente de tolerância imunitária, capaz de expressar TRAs tão ou mais eficientemente que as paradigmáticas mTEC_{hi}. A análise de correlações entre genes, associada a análise de enriquecimento de locais de ligação de factores de transcrição, permitiu ainda a identificação de potenciais novos reguladores do desenvolvimento das mTEC (*Vdr*, *Plagl1*, *Zbtb7a*, *Hnf4g*).

As Treg, por seu lado, representam um desafio bastante diferente. A importância das populações de Treg em tecidos não-linfóides (NLTs) é agora clara, quer na manutenção da tolerância imunitária face a auto-antígenos e outros antígenos inofensivos, quer em funções não-imunitárias, como regulação de metabolismo e regeneração de tecidos. Assim sendo, foquei-me em caracterizar as populações presentes em NLTs de ratinho, nomeadamente na *lamina propria* do cólon e pele, e compará-las com as suas equivalentes em tecidos linfóides (nódulos linfáticos que drenam os tecidos e o baço). Usando células T de memória (T_{mem}) como referência, defini quais os genes que caracterizam exclusivamente NLT Treg ou que são comuns a outros tipos de células T nos NLT. Também pude definir quais as diferenças e semelhanças entre as Treg na pele e no cólon. Modelando os dados computacionalmente, estabeleci ainda a ordem pela qual estas adaptações aos NLT são adquiridas pelas Treg aquando do seu recrutamento para os tecidos, quer em homeostase, quer durante uma perturbação imunitária (neste caso, indução de melanoma). As semelhanças entre os vários tecidos e condições são consideráveis, sugerindo um programa de recrutamento e adaptação geral das Treg ao ambiente não-linfóide, que é ligeiramente ajustável ao NLT específico. Verifiquei ainda que estas adaptações começam a surgir ainda no nódulo linfático que drena o tecido. Por último, confirmei que vários dos marcadores identificados para NLT Treg podem ser usados em humanos.

Em resumo, dados de scRNA-seq foram usados de forma bastante distinta para esclarecer dois processos e tipos celulares fundamentais na tolerância imunitária, que de outra forma permaneceriam confundidos em análises a nível de populações.

ABSTRACT

The set of mechanisms that protects the organism against pathogens needs to be efficient to destroy invading microorganisms and parasites. Consequently, this makes it potentially dangerous for the host's tissue. Therefore, strategies and checkpoints have to be in place to avoid the production of immune responses against self-antigens or harmless external antigens. This ability, named immune tolerance, has to be enforced mainly at the level of T cells, as they coordinate most immune responses. An important part of this control is exerted during T cell development in the thymus, through a remarkable process of negative selection conducted by medullary thymic epithelial cells (mTECs), which ectopically express tissue-restricted antigens (TRAs) to present to the developing T cells. The strength of the signal elicited upon recognition of self-peptides conditions the fate of T cells: a weak signal allows T cells to survive, as they are not self-reactive; a strong signal leads to their elimination; a moderate signal can lead to the development of a particular T cell subtype, the regulatory T cells (Treg). These cells later act on the peripheral tissues by exerting anti-inflammatory influence, blocking self-reactive conventional T cells that might have escaped negative selection.

Both cell types, mTEC and Treg, have thus been extensively studied in order to understand and eventually control and fine-tune immune tolerance. However, both populations have been missing a comprehensive characterization, namely with single-cell resolution, which would allow the identification of subpopulations and their phenotypes. Therefore, I employed state-of-the-art methods to analyse the transcriptome of individual cells (single-cell RNA sequencing, scRNA-seq), so that the structure, development and functions of mTEC and Treg populations could be clarified.

Murine mTECs were sorted without enrichment for any known maturation stage, and single-cell data was generated using a commercial microfluidic platform. By hierarchical clustering, I identified three subpopulations within the mTEC compartment - jTEC (*Pdpr*-positive), mTEChi (*Aire*-high) e mTEClo (*Aire*-low) - that appear to correspond to consecutive phases of mTEC development, according to the expression of the *Aire* transcription factor and each subpopulation's capacity to express TRAs. For the first time, these subpopulations were identified and characterized using an unbiased approach, *i.e.* without using additional flow cytometry markers. This approach gives a new perspective of the post-*Aire* state (mTEClo), which emerges as an active and potent driver of immune tolerance, capable of expressing TRAs at least as efficiently as the paradigmatic mTEChi. Gene-gene correlations, coupled

with transcription factor-binding motif enrichment, I identified potential new regulators of mTEC development (*Vdr*, *Plagl1*, *Zbtb7a*, *Hnf4g*).

Treg, on the other hand, represented quite a different challenge. Their importance in non-lymphoid tissues (NLTs) is now clear, both in maintaining immune tolerance to self and other harmless antigens, and in eliciting non-immune functions, as metabolism regulation and tissue regeneration. Therefore, I focused in characterizing the NLT Treg populations in mouse, namely in the colonic lamina propria and skin, and comparing them to their lymphoid tissue counterparts (draining-lymph nodes and spleen). Using memory T cells (Tmem) as a reference, I define the gene signatures that exclusively characterize NLT Treg, or that are shared with other T cell types in the NLTs. I also define the differences and similarities between skin and colon Treg. Computational modelling allowed me to establish the order by which NLT adaptations are acquired during their recruitment to the tissues, both in steady-state and during an immune challenge (namely, during melanoma induction). The similarity between tissues and conditions is quite significant, suggesting the existence of a general programme of recruitment and adaptation to the non-lymphoid environment that can be fine-tuned to each specific NLT. I also verify that such adaptations start to arise still in the respective draining-lymph nodes. Finally, I confirm that several of the identified NLT Treg markers can be used in the human context.

In summary, scRNA-seq data was used in quite distinct ways to clarify processes and cell types fundamental for immune tolerance, which would otherwise remain confounded in analyses at the population level.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS.....	vii
RESUMO.....	ix
ABSTRACT	xi
TABLE OF CONTENTS	xiii
ABBREVIATIONS.....	xvii
LIST OF FIGURES.....	xviii
LIST OF TABLES	xx
CONTRIBUTIONS OF OTHER CO-WORKERS	xxi
CHAPTER I - General Introduction.....	3
I.1 The Immune System	5
I.1.1 Innate immunity	5
I.1.2 Adaptive immunity.....	6
I.1.2.1 General features of the adaptive immune response	6
I.1.2.2 Recirculation pattern by the adaptive immune system	7
I.1.2.3 Phases of adaptive immune responses	8
I.2 T cells	11
I.2.1 Development of T cells.....	11
I.2.2 Subpopulations of TCR $\alpha\beta$ CD4+ T cells.....	12
I.2.3 Prolonged persistence in non-lymphoid tissues.....	13
I.3 Immune tolerance	14
I.3.1 mTECs: main drivers of central tolerance	14
I.3.1.1 Negative selection of thymocytes	14
I.3.1.2 Promiscuous gene expression by mTEC.....	15
I.3.1.3 mTEC development.....	16
I.3.2 Regulatory T cells	17
I.3.2.1 Phenotype, heterogeneity and function.....	17
I.3.2.2 Non-lymphoid tissue Treg	18
I.3.2.2.1 Phenotype and functions	18
I.3.2.2.2 Origin and establishment of NLT Treg populations.....	19
I.4 Single-cell mRNA sequencing technologies	20

I.4.1.1 Experimental protocols for scRNA-seq	21
I.4.1.2 Computational analysis of scRNA-seq data	25
I.4.1.3 Applications of scRNA-seq to Immunology	28
I.5 Objectives of this thesis	31
CHAPTER II - Material and Methods	35
II.1 scRNA-seq of the mTEC population.....	35
II.1.1 Experimental procedures	35
II.1.1.1 Mice.....	35
II.1.1.2 Tissue processing.....	35
II.1.1.3 FACS sorting	36
II.1.1.4 Single-cell RNA-sequencing using C1 Single-Cell Auto Prep System.....	36
II.1.1.5 Nextera XT Illumina Library prep – manual	38
II.1.1.6 Illumina High-throughput Sequencing.....	39
II.1.2 Computational analyses	39
II.1.2.1 Processing and quality control of single-cell mRNA-seq data	39
II.1.2.2 Differential expression	39
II.1.2.3 Consensus Clustering.....	40
II.1.2.4 Genomic clustering.....	40
II.1.2.5 Monocle	40
II.1.2.6 Binding motif enrichment analysis	41
II.2 scTreg of Treg populations.....	42
II.2.1 Experimental procedures	42
II.2.1.1 Mice.....	42
II.2.1.2 Human samples.....	42
II.2.1.3 Isolation of murine leukocytes for steady-state skin dataset	42
II.2.1.4 Isolation of murine leukocytes for steady-state colon dataset	43
II.2.1.5 Melanoma induction and cell isolation.....	43
II.2.1.6 Isolation of leukocytes from Human skin	44
II.2.1.7 Isolation of leukocytes from Human colon	44
II.2.1.8 Peripheral blood mononuclear cell isolation	44
II.2.1.9 Flow cytometry and single-cell RNA sequencing.....	45
II.2.2 Computational analyses	46
II.2.2.1 RNA expression quantification.....	46
II.2.2.2 scRNA-seq quality control	46
II.2.2.3 Dimensionality reduction methods	47
II.2.2.4 Cell cycle analysis	47
II.2.2.5 Differential expression analysis	48

II.2.2.6 Differential co-expression analysis	48
II.2.2.7 Obtaining a migration latent variable for steady-state Tregs.....	48
II.2.2.8 Identifying a common tissue migration trajectory in control and melanoma.....	49
II.2.2.9 Switch-like genes in the migration latent variable.....	50
II.2.2.10 Detection of expanded clonotypes.....	50
II.2.2.11 GO Term enrichment.....	51
II.2.2.12 Clustering analysis	51

CHAPTER III - Single-cell RNA-seq resolves self-antigen expression during mTEC

development	55
III.1 Abstract	55
III.2 Results	56
III.2.1 Single-cell mRNA sequencing of mTEC identifies global characteristics of TRA expression.....	56
III.2.2 scRNA-seq resolves three major subpopulations along mTEC differentiation	60
III.2.3 Binding motif and coexpression analysis highlights potential drivers of mTEC maturation	68
III.2.4 TRA expression during mTEC development.....	69
III.2.5 Expression of Aire-dependent TRAs is associated with strong genomic enrichment at single-cell level and is induced during jTEC-mTEC transition	73
III.3 Discussion	76
III.3.1 Unbiased analysis resolves mTEC development in adult	76
III.3.2 Relevance of post-Aire states in central tolerance.....	77
III.3.3 Lack of order optimizes the negative selection process	77

CHAPTER IV - Single cell transcriptomics of regulatory T cells reveals trajectories

of tissue adaptation	81
IV.1 Abstract	81
IV.2 Results	82
IV.2.1 Transcriptomes of CD4+ Treg and Tmem cells cluster by cell type and tissue	82
IV.2.2 Gene expression signatures of NLT regulatory and memory CD4 ⁺ T cells.....	89
IV.2.3 Tregs in skin and colon present contrasting biases	92
IV.2.4 Pseudospace alignment of Tregs from LNs to NLTs reveals stages of tissue adaptation	93
IV.2.5 Recruitment of Tregs to steady-state and challenged non-lymphoid tissues rely on shared mechanisms.....	102
IV.2.6 Evolutionary conservation between mouse and human NLT Treg cells.....	110
IV.3 Discussion.....	114

IV.3.1 Phenotype of NLT CD4 ⁺ T cells as a set of tissue and cell type-specific modules.....	114
IV.3.2 Consistency in the recruitment and adaptation of Tregs to NLTs	115
IV.3.3 Tumour Treg adaptation follows the path of steady-state NLT Tregs	116
IV.3.4 NLT adaptation from mouse to human	117
CHAPTER V - Conclusions and Future Work	121
References.....	123
Annexes.....	139

ABBREVIATIONS

APC – antigen-presenting cell
BCR – B cell receptor
bp – base pair
CTL – cytotoxic T cell
FACS – Fluorescence Activated Cell Sorting
IVT – *in vitro* transcription
jTEC – junctional thymic epithelial cell
LN – lymph node
mg – milligram
MHC – Major Histocompatibility Complex
mTEC – medullary thymic epithelial cell
mTEChi – mTEC-Aire^{hi}
mTEClo – mTEC-Aire^{lo}
ng – nanogram
nL – nanoliter
NLT – non-lymphoid tissue
PCA – principal component analysis
PCR – polymerase chain reaction
PGE – promiscuous gene expression
scRNA-seq - single-cell RNA-sequencing
Tcm – central memory T cell
TCR – T cell receptor
Tem – effector memory T cell
Treg – regulatory T cell
Trm – tissue-resident memory T cell
TF – transcription factor
TRA – tissue-restricted antigen
µg – microgram
µL – microliter

LIST OF FIGURES

FIGURE 1 – T LYMPHOCYTE RECIRCULATION.....	8
FIGURE 2 – PHASES OF ADAPTIVE IMMUNE RESPONSES.....	10
FIGURE 3 - CD4 ⁺ T CELL SUBTYPES.....	13
FIGURE 4 – SINGLE-CELL RNA-SEQUENCING METHODS.....	24
FIGURE 5 - scRNA-SEQ ANALYSES: FROM A GENE EXPRESSION MATRIX DATASETS TO COMPLETE RECONSTRUCTION OF A DIFFERENTIATION PROCESS.....	27
FIGURE 6 - TWO CONTRASTING STRATEGIES FOR DISSECTING T CELL DIFFERENTIATION USING scRNA-SEQ.	29
FIGURE 7 - EXPERIMENTAL WORKFLOW AND EXPRESSION OF TISSUE-RESTRICTED ANTIGENS ON POPULATION LEVEL.	56
FIGURE 8 - FLOW CYTOMETRY CELL SORTING STRATEGY FOR ISOLATION OF MTECs.	57
FIGURE 9 – MTEC scRNA-SEQ QUALITY CONTROL.	58
FIGURE 10 - CORRELATION BETWEEN INDIVIDUAL MTECs AND TISSUES.	59
FIGURE 11 - EXPRESSION OF FEZF2-REGULATED TRAs AT THE SINGLE-CELL LEVEL.	60
FIGURE 12 - PRINCIPAL COMPONENT ANALYSIS OF C1 DATASET.	61
FIGURE 13 - ANALYSES OF MTEC SUBPOPULATION STRUCTURE.	63
FIGURE 14 - SUBPOPULATION COMPARISONS FOR IN-HOUSE C1 DATASET.	64
FIGURE 15 - PCA AND CONSENSUS CLUSTERING ON PUBLICLY AVAILABLE AND IN-HOUSE SMARTSEQ2 MTEC DATASETS.....	65
FIGURE 16 - DIFFERENTIAL EXPRESSION OF GENES ACROSS MTEC SUBPOPULATIONS IN PUBLICLY AVAILABLE AND IN-HOUSE SMARTSEQ2 MTEC DATASETS.	66
FIGURE 17 - DIFFERENTIAL EXPRESSION OF GENES ACROSS MTEC SUBPOPULATIONS.	68
FIGURE 18 - EXPRESSION OF TRAs ACROSS MTEC SUBPOPULATIONS.	71
FIGURE 19 - TRA EXPRESSION SECTIONED BY <i>AIRE</i> AND <i>FEZF2</i> DEPENDENCY AND MTEC SUBPOPULATIONS IN C1 DATASET.	72
FIGURE 20 - CHARACTERIZATION OF PUBLICLY AVAILABLE AND IN-HOUSE SMARTSEQ2 MTEC DATASETS.	73
FIGURE 21 – GENOMIC DISTRIBUTION OF EXPRESSED TRAs ACROSS MTEC SUBPOPULATIONS.....	74
FIGURE 22 - STEADY-STATE scRNA-SEQ DATASETS OF CD4 ⁺ T CELLS FROM LT AND NLT.....	82
FIGURE 23 - SORTING SCHEME FOR MURINE TREG AND Tmem.	83
FIGURE 24 - QUALITY-CONTROL OF STEADY-STATE COLON AND SKIN scRNA-SEQ DATASETS.	84
FIGURE 25 - ADDITIONAL CHARACTERIZATION OF SORTED CD4 ⁺ T CELL POPULATIONS FROM STEADY-STATE MOUSE TISSUES.....	85
FIGURE 26 - ASSESSMENT OF INTRA-POPULATION HETEROGENEITY BY CONSENSUS CLUSTERING IN COLONIC CELL POPULATIONS.	87
FIGURE 27 - ASSESSMENT OF INTRA-POPULATION HETEROGENEITY BY CONSENSUS CLUSTERING IN SKIN CELL POPULATIONS.....	88
FIGURE 28 - CHARACTERIZATION OF TREG AND Tmem NLT ADAPTATIONS.	90
FIGURE 29 - PERCENTAGE OF CELLS EXPRESSING NLT MARKERS FROM MOUSE COLON (A) AND SKIN (B).	91
FIGURE 30 - SPECIFIC TREG ADAPTATIONS TO SKIN AND COLON ENVIRONMENTS.....	93
FIGURE 31 - RECONSTRUCTION OF TREG RECRUITMENT FROM LYMPHOID TO NON-LYMPHOID TISSUES IN STEADY-STATE.	95
FIGURE 32 - BGPLVM LATENT VARIABLE RELEVANCE FOR LN-TO-NLT TRANSITION IN THE COLON AND SKIN DATASETS.....	97
FIGURE 33 - THE EFFECT OF THE “PSEUDOSPACE” LATENT VARIABLE CAN BE IDENTIFIED IN INDIVIDUAL TISSUES.	98
FIGURE 34 - COMPARISON OF BGPLVM IN Tmem AND TREG FROM THE COLON AND SKIN DATASETS.....	99
FIGURE 35 - CHARACTERIZING CORE GENES IN STEADY-STATE ADAPTATION TO COLON AND SKIN.....	100

FIGURE 36 - EXAMINING THE MIGRATION PROGRAMME SHARED BETWEEN THE COLON AND SKIN DATASETS.	101
FIGURE 37 - QUALITY-CONTROL OF MOUSE MELANOMA SCRNA-SEQ DATASET.	103
FIGURE 38 - RECRUITMENT AND ADAPTATION OF TREG TO THE TUMOUR ENVIRONMENT RECAPITULATES STEADY-STATE MIGRATION.	105
FIGURE 39 - CONTROL AND MELANOMA RESPONSE IN THE TUMOUR MIGRATION TRAJECTORY.	106
FIGURE 40 - THE MIGRATION EFFECT DETECTED USING MRD IS PRESENT IN BOTH CONDITIONS INDIVIDUALLY.	107
FIGURE 41 - EXAMPLES OF INDIVIDUAL GENES VARYING IN THE CONTROL AND MELANOMA MIGRATION TRAJECTORY.	108
FIGURE 42 - HUMAN-MOUSE COMPARISON OF NON-LYMPHOID TISSUE TREG MARKER GENES.	111
FIGURE 43 - QUALITY-CONTROL OF STEADY-STATE HUMAN COLON AND SKIN SCRNA-SEQ.	112
FIGURE 44 - NLT CD4 ⁺ T CELL MARKERS IN HUMAN COLON AND SKIN SCRNA-SEQ DATASETS.	113
SUPPLEMENTARY FIGURE 1 - CELL CYCLE PHASE DETECTION BY CYCLONE.	139
SUPPLEMENTARY FIGURE 2 - CD4 ⁺ T CELLS TCR CLONOTYPES IN DIFFERENT TISSUES.	140

LIST OF TABLES

TABLE 1 - MAIN FEATURES OF THE ADAPTIVE IMMUNE SYSTEM	7
TABLE 2 – SINGLE-CELL CAPTURE ALTERNATIVES	23
TABLE 3 – MIXES FOR SCRNA-SEQ WITH C1 SINGLE-CELL AUTO PREP SYSTEM.	37
TABLE 4 – C1 THERMAL CYCLING PROTOCOLS.	37
TABLE 5 - NEXTERA XT PCR PROTOCOL.	38
TABLE 6 - ANTIBODY PANELS USED FOR FLOW CYTOMETRY.	45
TABLE 7 - QUALITY CONTROL CRITERIA FOR FILTERING SINGLE CELL TRANSCRIPTOMES IN EACH DATASET, AND PARAMETERS FOR THEIR VISUALIZATION IN tSNE (LAST TWO COLUMNS). CELLS WERE KEPT IF THEY PASSED ALL THESE FILTERS.	47

CONTRIBUTIONS OF OTHER CO-WORKERS

mTEC section (Chapter III): Xiuwei Zhang performed scLVM correction, the correlation between tissues and mTECs, and the network analysis.

Treg section (Chapter IV): Agnieszka Chomka and Ahmed Hegazy collected and sorted the steady-state cells from mouse; Ahmed Hegazy collected and sorted the human colonic cells; Laura Jardine collected and sorted the human skin cells; Angela Riedel induced the melanoma challenge and collected the tissues; Tomás Gomes was deeply involved in all the steps of the analyses; final s were created by Tomás Gomes.

CHAPTER I

General Introduction

CHAPTER I - General Introduction

All multicellular organisms are regularly exposed to pathogenic agents that aim at invading and colonizing them. In response to this threat, they have developed the immune system: a set of barriers, cell types and molecules that can fight the replication and propagation of pathogens. Although indispensable for the organism's survival, this system can cause severe complications when deregulated or misdirected. Namely, the immune system has to be able to prevent immune reactions against its own host as well as against other innocuous antigens, a property known as immune tolerance. Understanding how tolerance is induced and established will be fundamental in the development of efficient immunotherapies against autoimmune diseases, chronic inflammations and allergies.

From all the cell types in the immune system, T lymphocytes, or T cells, are arguably the most important lineage to control due to their capacity to recognize antigens and unparalleled ability to orchestrate the immune responses against them. Therefore, the specificity of T cells is carefully supervised during their development, resulting in the deletion of self-reactive T cells or their repurposing into regulatory T cells.

In this thesis I have addressed two cell types that play pivotal and complementary roles in immune tolerance: medullary thymic epithelial cells (mTEC) and CD4⁺ regulatory T cells (Treg). While mTECs are responsible for the selection process that depletes self-reactive T cells, CD4⁺ regulatory T cells (Treg) control self-reactive T cells that escape the selection process, as well as all other immune responses triggered in the periphery.

Although both cell types have been extensively studied in recent years, further progress has been largely impaired by lack of unbiased and comprehensive single-cell resolution. Therefore, here I am using state-of-the-art single-cell RNA sequencing in order to investigate the development, heterogeneity and phenotype of mTECs and Tregs, one cell at a time.

In Chapter I, I describe broad aspects of the immune system, with focus on adaptive immunity. Immune tolerance is then further addressed, with particular attention to mTEC and Treg functions and development. As the work described in this thesis relies largely on single-cell RNA sequencing

approaches, this chapter also includes a section on the experimental protocols and computational tools available, as well as specific examples of their use in the study of the immune system.

I finish this Chapter with a description of the motivations for this study, the specific questions that I propose to address, and the relevance of this work in the context of the mTEC and Treg fields.

In Chapter II, I describe the experimental procedures and computational analyses performed. They have been separated in two sections, which correspond to the study of mTECs (section II.1) and Tregs (section II.2).

In section Chapter III, I describe how I performed scRNA-seq of murine medullary thymic epithelial cells using the C1 platform, and how I used these data to find and characterize mTEC subpopulations. I characterize them in terms of simple gene expression and, most importantly, in terms of their ability to induce immune tolerance. scRNA-seq allows me to divide mTECs in an unbiased way into three subpopulations, which differ in their maturation state and their capacity to express tissue restricted antigens. Despite these differences, all three subpopulations appear to be able to express most of those genes. Importantly, I show that the most differentiated mTEC stage is still actively taking part in the induction of immune tolerance.

In Chapter IV, I describe the phenotype of murine CD4⁺ T cells, both regulatory and memory T cells (Treg and Tmem), found in non-lymphoid tissues. I focus mostly on Tregs and I describe computationally the trajectory of recruitment and adaptation to non-lymphoid tissues, skin and colon, in steady state. Using a murine melanoma model, I compare the steady-state with the challenged condition. The reconstruction of pseudotime trajectories reveals a core set of genes that is upregulated in Tregs during adaptation to colonic and skin environments, which is also largely shared by cells undergoing recruitment to tumours. The same perturbation also shows that adaptation starts in the lymph nodes and not upon arrival of Tregs to the non-lymphoid tissues.

Finally, I conclude this section performing a human-mouse comparison of non-lymphoid tissue markers.

Lastly, in Chapter V I collect the general and most relevant findings described in this thesis.

I.1 The Immune System¹

I.1.1 Innate immunity

The first line of protection of the organism against foreign agents consists of the innate immune system, which is present under some form in all multicellular organisms.

The response elicited is rapid, targeting structures shared across several families of microbes, and it remains unchanged in subsequent contacts with the same agent.

The innate branch of immunity is composed of physical barriers, cell intrinsic defense programs, multiple immune cell types and biochemical defenses. Intact skin and mucosal surfaces of the gastrointestinal and respiratory tracts constitute the three main epithelia that separate the organism from pathogens in the surrounding environment. Besides the physical separation they impose, cells in these barriers are responsible for the production of antimicrobial peptides, defensins and cathelicidins, which exert their immune function directly through microbicidal effects and indirectly through activation of immune cells. The cellular component of innate immunity includes cells from the myeloid lineage (neutrophils, macrophages and dendritic cells), which are able to express pattern recognition receptors that respond to foreign molecules, *e.g.* lipoproteins, peptoglycans, nucleic acids, present in either the extra- or intracellular environment. Activation of receptors such as Toll-like receptors (TLR), C-type lectins and NLRs, ultimately leads to increased expression of inflammatory cytokines, chemokines, costimulatory molecules (CD80, CD86) and antiviral cytokines.

In parallel, dendritic cells and macrophages function as professional antigen-presenting cells (APCs). They are able to efficiently sample, process, and exhibit antigens to T cells via Major Histocompatibility Complex (MHC) ultimately leading to T cell activation, and thus linking innate immunity and the activation of the adaptive branch.

Overall, innate immunity responses are fundamental in the defense of the organism, granting prompt response to invasions. However, microbes often circumvent these general responses and their elimination then relies on the more powerful mechanisms of the adaptive immune system.

¹ Unless specific references are mentioned, the information on this section is reviewed in [1]

I.1.2 Adaptive immunity

Vertebrates have developed an additional line of defense that overcomes some of the limitations of the innate immune system. This set of mechanisms can be divided into humoral and cellular responses, respectively driven by B and T lymphocytes, which work together to eliminate pathogens following three complementary strategies. First, T lymphocytes, namely T helper cells, boost the microbicidal abilities of phagocytes and stimulate antibody production by B cells. Secondly, the antibodies secreted by B cells bind to extracellular pathogens, impairing their ability to infect the host and facilitating their phagocytosis. Thirdly, cytotoxic T cells (CTLs) destroy cells infected by microbes that are inaccessible to antibodies.

I.1.2.1 General features of the adaptive immune response

In contrast with the innate branch, the repertoire of molecules that can be recognized by the adaptive immune system is extremely diverse, and the lymphocytes stimulated are specific for each antigen (Table 1). Furthermore, the response of the adaptive immune system against any antigen becomes more proficient with every exposure to the same antigen. As a result of a primary response against a given antigen, lymphocytes are able to establish long-lived memory populations that change the dynamics of subsequent contacts. Secondary responses will thus be prompter and more powerful than the primary one. Remarkably, the response given by this branch is specialized, varying in accordance with the specific pathogen that triggers it, thus enhancing its control over each infection.

To keep pace with the rapid replication rate of infecting pathogens, responding lymphocytes are able to divide at high rates, thus increasing the pool of specific lymphocytes in a process known as clonal expansion. Finally, it is important to consider that such a competent and effective system has to be tightly regulated. After the threat has been eliminated, several mechanisms, namely deprivation of survival factors, induce the contraction of the adaptive immune response and the restoration of homeostasis.

It is equally important that responses are not triggered altogether against self-antigens, as well as other harmless antigens that might be recognized by the adaptive immune system. This feature is defined as immune tolerance, and will be further addressed in a later section of the Introduction (section I.3).

Table 1 - Main features of the adaptive immune system

Feature	Functional significance
Specificity	Ensures that distinct antigens elicit specific responses
Diversity	Enables immune system to respond to a large variety of antigens
Memory	Leads to enhanced responses to repeated exposures to the same antigen
Clonal expansion	Increases number of antigen-specific lymphocytes to keep pace with microbes
Specialization	Generates responses that are optimal for defense against different types of microbes
Contraction and homeostasis	Allows immune system to respond to newly encountered antigens
Non-reactivity to self	Prevents injury to the host during responses to foreign antigens

I.1.2.2 Recirculation pattern by the adaptive immune system

All circulating blood cells in the adult, including lymphocytes, are generated in the bone marrow from the division and differentiation of a population of hematopoietic stem cells (HSCs). While the maturation of B lymphocytes unfolds completely within the bone marrow, immature T lymphocytes have to migrate to the thymus to finish their development.

When fully mature, lymphocytes leave the primary lymphoid organs and adopt a traffic pattern known as lymphocyte recirculation (Figure 1): they continuously circulate between blood stream, secondary lymphoid organs (lymph nodes and spleen), the lymphatic system, and back to the blood stream until they find the antigen they are specific to.

Such encounters are potentiated specifically in the lymph nodes, due to high concentration of antigens and APCs drained from the surrounding non-lymphoid tissues. Upon antigen recognition, lymphocytes can go back to the blood stream and enter sites of infection/inflammation in non-lymphoid tissues (NLT) to fight the ongoing immune challenge. All these steps are governed by the regulation of expression of adhesion molecules (integrins and selectins) and chemokine receptors that promote preferential migration and/or retention in specific tissues, in a process called homing.

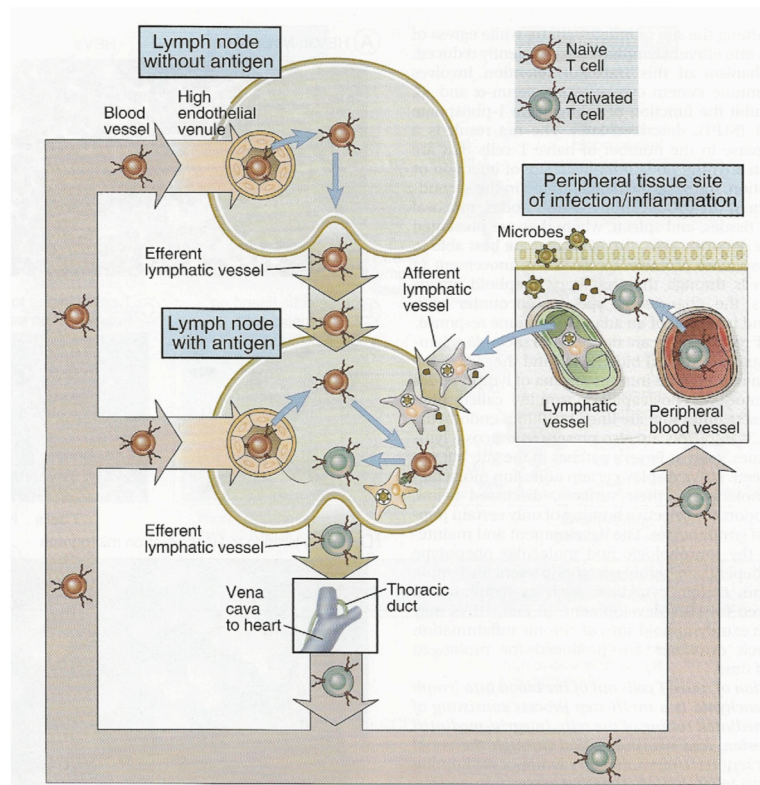


Figure 1 – T lymphocyte recirculation.

Naïve T cells usually enter the lymph nodes coming from the blood stream through high endothelial venules. Dendritic cells coming from non-lymphoid tissues and presenting antigens found there enter the lymph nodes via afferent lymphatic vessels. If T cells are able to recognize these peptides, they get activated and return to circulation through the lymphatic system, and then to the bloodstream, finally reaching the sites of inflammation in peripheral tissues through venules. From [1].

1.1.2.3 Phases of adaptive immune responses

The number of naïve lymphocytes with specificity to a given antigen is in the order of 1 in 10^5 or 10^6 . In addition, the amount of the antigen itself is usually small. Therefore, to increase the chances of lymphocytes encountering the right antigen, there is a need for capturing microbes, concentrating them in certain locations, and efficiently presenting them to lymphocytes. Mainly dendritic cells, the paradigm of APCs, play the important role of sampling antigens in the epithelia and connective tissues, and migrating to the tissue-draining lymph nodes where they present them to recirculating lymphocytes.

Lymphocyte activation triggered upon antigen recognition starts with clonal expansion of the reactive clones, followed by their differentiation into more mature states. After inducing antigen elimination and

as homeostasis of the system is reestablished, the vast majority of the activated lymphocytes is then routed towards apoptosis, or to a lesser extent, towards a memory phenotype. Despite following the same general trajectory, T and B cell reactions and functions upon activation are quite distinct.

Activation of CD4⁺ and CD8⁺ T cells lineages happens upon engagement of the T cell receptor (TCR) by antigens being presented by APCs, respectively via class II and class I MHC molecules. CD4⁺ T helper cells differentiate into effector cells, which contribute to the immune response mainly via cytokine production. They egress from the lymph node and migrate towards the sites of inflammation, where they stimulate phagocytes that have captured microbes to destroy them. They are also capable of stimulating eosinophils that, in turn, target parasites that cannot be phagocytosed. In the lymph nodes, CD4⁺ T helper cells also play a fundamental role in activating B cells against protein antigens. CD8⁺ T cells are activated and differentiate into CTLs capable of killing infected host cells, either by viruses or by bacteria that survive in the phagocytic vacuoles. The function and characteristics of the T cell lineage, especially CD4⁺ T helper cells are of major relevance for this thesis and will be described in detail in section I.2.

Activation of B cells ultimately leads to the secretion of antibodies. In contrast to T cells, B cell receptors (BCRs) can be activated by free lipid or polysaccharide antigens that possess repetitive structures and can thus engage with a high number of BCRs simultaneously on the same cell. However, for protein antigens, activation of B cells depends on activating signals transmitted by T helper cells. For this, B cells work as APCs to interact with reactive T cells that will in turn activate them.

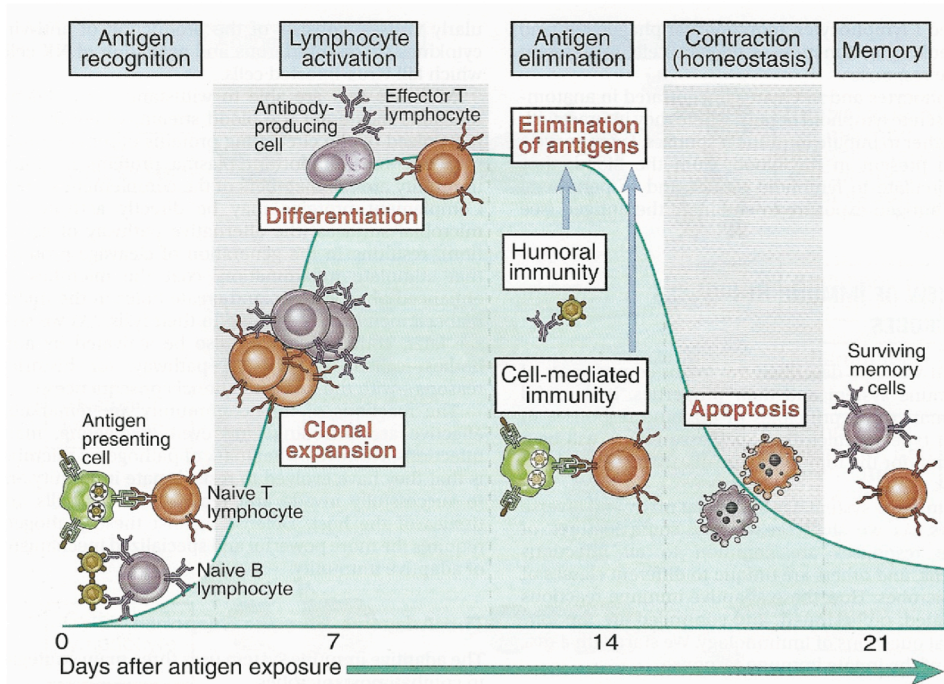


Figure 2 – Phases of adaptive immune responses.

Adaptive immunity responses consist of five phases. It starts with antigen recognition, then lymphocyte activation (concomitant clonal expansion and differentiation), and antigen elimination, after which responding lymphocyte population contracts by cell death, finally returning to homeostasis leaving memory lymphocytes behind. The time on the x-axis can vary depending on the immune challenge. The y-axis is an arbitrary measure of magnitude of response. From [1].

I.2 T cells

I.2.1 Development of T cells

Typically, the most immature T cells arrive to the thymic cortex via blood vessels, migrating through cortex and then towards the thymic medulla as they go through several stages of maturation and undergo two steps of selection, one positive and one negative (negative selection is further explored in section I.3.1.1). Fully mature T cells emerge from the thymus expressing fully functional TCR on the cell surface. During maturation, the genes encoding the TCR chains undergo somatic rearrangement of gene segments that are spatially separated in the germline loci. Although the TCR can be formed by either α and β or γ and δ chains, the majority of the mature T cell population is composed of $\alpha\beta$ T cells. The α locus is composed of multiple V and J segments, while the β locus has V, D and J segments. The combinatorial associations between these segments, plus additional inclusion and exclusion of nucleotides at their junctions, gives rise to the diversity of the TCR repertoire.

The development of T cells in the thymus (thymocytes) happens as they migrate from the cortex to the medulla. The thymocytes that arrive to the thymic cortex from the bone marrow proliferate and undergo apoptosis at extremely high rates, being estimated that 95% of them die before getting to the medulla. They constitute the most immature cell state, which is known as pro-T cell stage (or double-negative thymocytes), and is characterized by having the TCR locus in their germline configuration, and for not expressing TCR, CD3, or the coreceptors CD4 and CD8. After rearrangement, 90% of these double-negative cells will eventually give rise to cells expressing α and β chains. Then, they pass through the pre-T stage, during which V-D-J recombination of the β locus is completed. A pre-TCR, composed of this β chain and an invariant α chain, transduces a signal that inhibits the rearrangement of the other β locus and promotes the expression of CD4 and CD8. The double-positive stage, cells are both CD4 and CD8-positive, and they start expressing low levels of the TCR receptor after the α chain is rearranged. The resulting TCR $\alpha\beta$ CD4⁺CD8⁺ T cells are then positively selected, *i.e.* rescued from programmed cell death, if the TCR is able to recognize with low avidity the peptide-MHC complexes on the surface of thymic epithelial cells. In the medulla, already as single-positive T cells committed to either CD4 or CD8 lineage, T cells are negatively selected based on high-avidity recognition of self-antigens (clonal deletion) (see section I.3.1). By the end of their maturation, the resulting naïve T cells will then enter the bloodstream and start exerting their function.

1.2.2 Subpopulations of TCR $\alpha\beta$ CD4⁺ T cells

As described above, naïve CD4⁺ T cells that arise from the thymus adopt a pattern of recirculation between secondary lymphoid organs (SLOs) and the bloodstream until they find the correct antigen (Figure 1). Upon TCR engagement, an activation and differentiation process ensues leading naïve CD4⁺ T cells down one of several lineages of effector T cells. The result of this differentiation will depend on the cues given by the surrounding inflammatory environment: TCR signaling, co-stimulation via CD28, and activation of distinct cytokine receptors [2], which will affect transcription, chromatin structure, translation and signaling transduction within activated T cells.

Ultimately, several differentiated and functionally distinct T helper (Th) cell subsets can develop and be classified according to the primary cytokines expressed and their master transcription factors (Th1, Th2, Th9, Tfh, Th17, Treg). Each subset of T helper cells has been associated with the control of different immune challenges or particular functions within the adaptive immune response, as described in Figure 3. In contrast with the pro-inflammatory functions of most CD4⁺ T cell lineages, regulatory T cells (Tregs) play a critical role in the suppression of the immune response, conveying peripheral tolerance to self- and non-self-antigens. Treg can arise directly from the thymus (thymic Treg, tTreg) or, as T helper lineages, from differentiation of naïve T cells in the periphery (peripheral Treg, pTreg). As one of the main subjects explored in this thesis, the Treg lineage will be further addressed in section 1.3.2.

When the immune challenge that triggered the activation and differentiation of CD4⁺ T cells is eliminated, approximately 90% of effector cells are expected to die during the contraction phase of the adaptive immune response (1-2 weeks). The remaining cells set-up long-lived populations of quiescent memory cells, capable of sporadic self-renewal and survival in the absence of their antigen [3]. Central memory T cells (T_{cm}) recirculate mainly through lymphoid organs due to the expression of CD62L (L-selectin) and the chemokine receptor CCR7, and are able to produce IL2 and replicate at high rates in case of a recurrent infection. Effector memory T cells (T_{em}), on the other hand, can circulate through the non-lymphoid tissues and will produce high amounts of cytokines upon TCR activation. More recently, it has been shown that tissue-resident memory T cells (T_{rm}) can be formed and preferentially reside in the non-lymphoid tissues, playing an important role in the immediate control of immune challenges and, to some extent, intervening in non-immune processes [4–6].

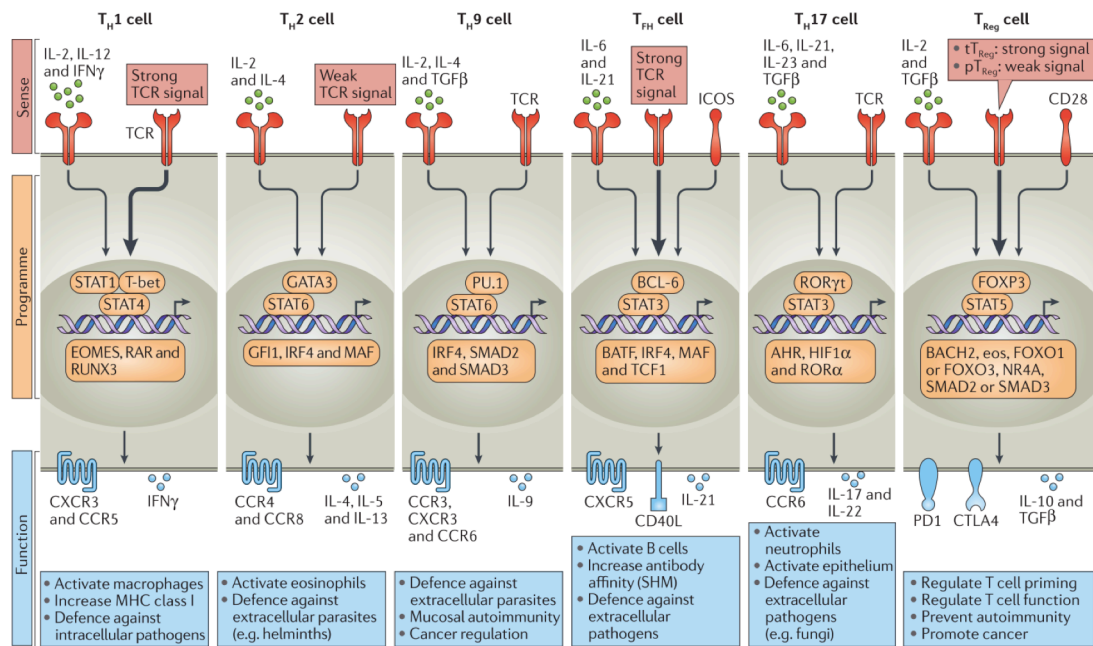


Figure 3 - CD4⁺ T cell subtypes.

Each CD4⁺ T cell subset can be defined by the signals they sense (red), the transcription factors that govern their differentiation (orange) and the effector molecules and chemokine receptors that contribute to their function (blue) in the control of specific pathogens or immune pathologies. Adapted from [7].

1.2.3 Prolonged persistence in non-lymphoid tissues

In recent years, it has become evident that memory T cells can be divided into circulating cells, T_{em} and T_{cm}, and non-circulating, residing cells, the T_{rm}. T_{rm} dwell in non-lymphoid tissues, such as the colon, lung and skin, and they respond locally and rapidly, constituting an early line of response to immune challenges [4]. They exhibit a phenotype that is distinct from the circulating cells, and appears to be heavily influenced by their location. These are likely to be adaptations to new functions, different migratory patterns and to the different survival signals in non-lymphoid environments. Such compartmentalization is also seen at the level of the TCR repertoire of the T_{rm}, which does not overlap with the repertoire from circulating cells.

In spite of T_{rm}'s demonstrable relevance, the functions, development and maintenance of T_{rm}, and in particular of CD4⁺ T_{rm}, are still poorly understood. The importance in understanding the differences and similarities between circulating and tissue-residing T cells increased with the expansion of T cell-targeting and T cell-based approaches for treatment of cancer autoimmunity and inflammatory diseases [6].

1.3 Immune tolerance²

One of the most enigmatic questions in Immunology is the ability of the immune system to effectively react against pathogenic agents while remaining unresponsive to self and other harmless antigens. This is explained by the induction of tolerance to the former but not to the latter group of antigens. Problems in establishing tolerance are thus frequently linked to the development of autoimmune disorders, chronic inflammations and allergies.

There are two main categories of tolerance: central and peripheral. The former is exerted at the level of immature lymphocytes as they undergo maturation in a primary lymphoid organ, namely the thymus for T cells and the fetal liver or bone marrow for B cells. The latter, on the other hand, happens at the level of mature lymphocytes while they are recirculating through secondary lymphoid organs.

For T cells, central tolerance is enforced by medullary thymic epithelial cells (mTECs), which induce clonal deletion of highly reactive T cells and differentiation of mildly self-reactive cells into thymic regulatory T cells (tTreg). Interestingly, another branch of Tregs, peripherally-derived Tregs (pTreg), differentiates from naïve T cells throughout the organism, controlling immune reactions against innocuous foreign antigens, thus being part of the peripheral tolerance. The biology of mTECs and Tregs is addressed separately in the next sections.

1.3.1 mTECs: main drivers of central tolerance ³

1.3.1.1 Negative selection of thymocytes

Haematopoietic stem cell-derived T lymphoid progenitors migrate from the bone marrow to the thymus relying on chemokine receptors such as CCR7, CCR9 and CXCR4. In the thymus, developing T cells, *i.e.* thymocytes, are positively selected in the cortex (outer region) by cortical thymic epithelial cells (cTECs). The selection is based on the ability of their TCR to interact with MHC molecules. Then, CCR7 expression drives thymocytes to migrate to the thymic medulla (inner region) and to interact with APCs, namely with medullary thymic epithelial cells (mTECs) and dendritic cells.

² Unless specific references are mentioned, the information on this section is reviewed in [1]

³ Parts of this section were adapted from [170].

The medulla of the thymus plays a major role in tolerance induction via negative selection of thymocytes (maturing T cells), also designated as clonal deletion. In fact, disruption of the spatial structure of the medulla and other perturbations that block thymocytes from maturing in the medulla environment, give rise to autoimmune manifestations [8]. This function of the medulla has been mainly linked to the ability of mTECs to “ectopically” express a range of tissue-restricted antigens (TRAs) in a rather obscure process designated as promiscuous gene expression (PGE) (section 1.3.1.2). The expressed TRA proteins are broken down into peptides and presented to thymocytes¹ which will take different paths upon engagement depending on the binding affinity. According to the affinity model of thymocyte selection (reviewed in [8,9]), low affinity binding will promote thymocyte survival, while high affinity will induce thymocyte apoptosis. Within a window of medium affinity interactions, thymocytes can be differentiated towards a conventional T cell or a thymic regulatory T cell (tTreg) phenotype [8,10–13]. As a result, the conventional T cell repertoire is largely purged of clones that interact strongly with self peptide-MHC complexes, ensuring that they will not mount immune reactions against self-antigens.

1.3.1.2 Promiscuous gene expression by mTEC

At the population level, the characteristic promiscuous gene expression by mTEC allows them to virtually express all the genes in the genome. At the cellular level however, such expression is highly variable, with individual TRAs in many cases expressed only by a small fraction (1-3%) of the mTECs [14–17]. Despite being the focus of many studies, multiple dimensions of this remarkable process are still far from being understood.

The most well characterized factor contributing to PGE in mTECs is the AIRE protein encoded by the Autoimmune regulator gene (*Aire*). Its deletion leads to a polysymptomatic autoimmune disorder Autoimmune polyendocrinopathy-candidiasis-ectodermal dystrophy (APECED) [18], and its expression in thymus depends on p65 and RELB binding to conserved enhancer elements upstream of the *Aire* locus [19,20]. AIRE acts largely independently of DNA sequence, as indicated by its discrepant targets in different cell types [21,22], binding to hypomethylated Lys4 of histone 3 and other repressive epigenetic marks, promoting gene expression by inducing the release of stalled RNA polymerases (reviewed in [23]). In addition, it has recently been suggested that AIRE preferentially sits in super-enhancer regions, which can form intra- and inter-chromosome loops in a dynamic fashion, thus explaining the high cell-to-cell variability of TRA expression in mTECs [24].

Interestingly, a significant proportion of TRAs have been found to be expressed also in the absence of *Aire*, strongly suggesting the existence of unknown complementary mechanisms [14,25]. In line with this notion, the transcription factor Fez family zinc-finger 2, *Fezf2*, was recently identified to induce the expression of some *Aire*-independent TRAs [26].

At the single-cell level, TRAs expressed in each mTEC seem to be functionally unrelated [27], *i.e.* they do not recapitulate functional modules seen in the tissues that express them. Nonetheless, some co-expression modules appear to be present, and might be linked to sub-stages of mTEC development [28].

Understanding of PGE has been impaired by a combination of its stochastic nature, the incomplete understanding of the mechanisms and factors driving TRA expression, and the confounding effect of a yet unclear mTEC developmental process.

1.3.1.3 mTEC development

The development of mTECs has been proposed to originate from bipotent thymic epithelial progenitor cells (bTEP), also giving rise to cTECs [29,30]. These bTEP have been poorly characterized so far, especially in the adult thymus. In fact, a great majority of the cells that acquire mTEC fate originate in lineage-restricted progenitors [31], and they have been shown to progress through an early developmental stage termed junctional TEC (jTEC) [32], into a mature *Aire*-expressing stage, eventually progressing into a post-AIRE stage [33,34]. This process of development of functional mTECs is known to be driven by the histone deacetylase 3 (Hdac3) and NF- κ B signaling, the latter through RANKL and CD40L produced by thymocytes [35]. As mTECs progress along this trajectory, they migrate from the outer to the inner region of the medulla [33].

Mature postnatal mTECs can be divided into two main subpopulations based of their ability to present antigens, *i.e.* based on the expression levels of CD80 and MHC Class II molecules: CD80^{hi}MHCII^{hi} and CD80^{lo}MHCII^{lo}. However, these subpopulations are intrinsically heterogeneous and their developmental relationships are unclear²² [23]. While CD80^{hi}MHCII^{hi} are mostly *Aire*-expressing mTECs, and have been reported to be subdivided according to subsets of TRAs expressed [36], CD80^{lo}MHCII^{lo} typically express less *Aire*, including both pre- and post-Aire stages (reviewed in [23]).

Until recently, the mTEC population expressing high levels of AIRE (CD80^{hi}MHCII^{hi}) was considered the apex of mTEC maturation, after which cells would enter AIRE-driven apoptosis [37]. However, it

has recently been demonstrated by cell fate mapping that upregulation of *Aire* does not mark the last stage of mTEC life cycle. Instead, lymphotoxin- α (LT α)–LT β R signals coming from developing thymocytes [38] induce cells to proceed into a post-*Aire* stage characterized by loss of *Aire*, *Cd80* and *MHCII* expression and upregulation of epithelial differentiation markers, such as desmoglein-1 and 3, involucrin and Spink5 [33,34]. Although post-*Aire* mTECs are regarded as terminally differentiated, the exact function and progression through intermediate steps are not clear.

In summary, the developmental stages in the thymic medulla are still incompletely understood. Furthermore, the mechanisms by which TRA expression is gained, and to which extent it is maintained in the post-*Aire* state remain poorly understood. The cell-intrinsic and developmental heterogeneity of the epithelial cells, have made these mechanisms difficult to elucidate using population-level approaches.

1.3.2 Regulatory T cells

1.3.2.1 Phenotype, heterogeneity and function

Regulatory T (Treg) cells are responsible for the suppression of immune responses, thus limiting inflammation, allergies, autoimmune diseases and influencing the development of tumours. They exert their functions by killing or inhibiting T cells and APCs, *e.g.* through CTLA4 and GZMB, and by producing suppressive cytokines such as TGF- β , IL-10, IL-35, and galectin-1. The key role of Treg is apparent from the severe immune deregulations motivated by neonatal thymectomy and other impairments of Treg development (reviewed in [39]). *Foxp3* gene is the best marker to identify Treg, despite being transiently expressed in human T helper during activation. This transcription factor is part of the forkhead/winged-helix family of transcription factors, it is highly conserved between species and mutations in the gene were identified as the causative factor responsible for Scurfy, in mouse, and the Immunodysregulation, Polyendocrinopathy, and Enteropathy, X-linked syndrome (IPEX), in human.

Most Treg arise from the thymic medulla – thymic-derived Treg (tTreg) [40] – as a functionally mature T cell subpopulation able to migrate towards inflammation sites and suppressing effector lymphocytes upon recognition of the cognate antigen (reviewed in [41,42]). It is also known that, under the correct conditions (weak TCR signaling, TGF- β and IL-2), naïve CD4⁺ T cells can give rise to Foxp3⁺ suppressor cells in the periphery – peripherally derived Treg (pTreg) [40]. Although their generic

suppressor role seems the same, tTreg cells are believed to be in charge of controlling autoimmune reactions (tolerance to self-antigens) while pTreg mediate neo-antigen tolerance, *e.g.* antigens from commensal bacteria, environmental antigens, food (reviewed in [43]). Interestingly, functional differences between tTreg and pTreg have not been found.

The Treg lineage has been shown to exhibit remarkable plasticity. In fact, they seem to be able to enhance their ability to control inflammatory responses through co-expression of transcription factors typically associated with the differentiation program of effector T helper cells, such as *T-bet*, *Gata3*, *Irf4* and *Stat3* [43,44]. This ability of Treg cells to adapt to their surrounding inflammatory milieu is critical in preventing the development of context-dependent immunopathology.

In recent years, the diversity of Treg phenotypes has been reported in the context of different states of activation and location. Central Treg (cTreg) can be found recirculating between SLOs, expressing low amounts of CD25, and high levels of CD62L, CCR7 and CD45RA. Activation induces differentiation into effector Treg (eTreg), which upregulate CD25, CD44, KLRG1, CD103, the transcription factors BLIMP-1, BATF, IRF4 and the downregulate CD62L, CCR7 and CD45RA, which overall translates into a Treg phenotype that is more suppressive and spends less time in SLOs. Additionally, non-lymphoid tissue Treg (NLT Treg), which are phenotypically similar to eTreg, have also been identified.

1.3.2.2 Non-lymphoid tissue Treg

1.3.2.2.1 Phenotype and functions

As for T helper cells, compartmentalisation of regulatory T cell populations has been observed across several tissues. Multiple studies have described distinct adaptations that NLT Tregs develop in contrast to their lymphoid counterparts, which provide them with tools to control both immune and non-immune processes, most notably in the environment of the visceral adipose tissue (VAT) [45], skeletal muscle [46], and the colonic lamina propria [47]. These NLT Tregs exhibit a differential transcriptional profile with obvious enrichment of transcripts encoding effector molecules (*Ctla4*, *Gzmb*, *Klrg1*), chemokines and their receptors (*e.g.* *Ccr4*), and immunosuppressive cytokines such as *Il10* [48]. In addition, NLT Tregs also express a TCR repertoire distinct from their lymphoid counterparts, yet another indication of tissue specialisation.

Remarkably, the role of NLT Tregs can also extend beyond the immune regulatory function and has been implicated in multiple unique tissue-specific physiological processes. For instance, VAT Tregs have been shown to influence overall metabolic parameters, such as obesity, obesity-related

inflammation and glucose tolerance via IL33-signaling through ST2 (reviewed in [49]). Also in an ST2-dependent mechanism, Tregs in the skeletal muscle have been shown to promote muscle repair, partially by the production of amphiregulin, *Areg* [46].

1.3.2.2.2 Origin and establishment of NLT Treg populations

The exact origin of NLT Treg is diverse and largely dependent on the specific tissue being considered. It seems like VAT Treg are largely thymic-derived Treg, as suggested by observations such as the expression of tTreg markers Neuropilin-1 (*Nrp1*) and Helios (*Ikzf2*) and the non-overlapping TCR repertoire between Treg and effector T cells in the VAT. A similar situation characterizes skeletal muscle Treg. Colonic Tregs, however, appear to be heterogeneous and divided into a tTreg and a pTreg population, Gata3⁺ and Ror- γ t⁺ (Rorc⁺) respectively.

In terms of dynamics of retention in each NLT, Treg seem to also be fairly heterogeneous. While VAT Treg population remains stable, *i.e.* there is a low number of Treg being recruited and leaving the NLT, skeletal muscle and colonic Treg migration rates to and from NLTs are quite higher.

Recruitment seems to be driven both by general NLT receptors, such as *Ccr8*, *Ccr2*, *Ccr5*, in conjunction with tissue-specific ones, such as *Gpr15*, a homing factor to the colon. Interestingly, the attraction and survival in the NLTs seems to be motivated by local antigens, at least for VAT and colonic Ror- γ t⁺ Treg.

The timing and the location where Treg acquire NLT phenotypes is even more obscure than the phenotypes themselves. Although it seems more likely that such differences are established upon arrival of Tregs to the NLTs after induction by tissue-specific cues, the existence of biases imposed in the thymus or any mixture of these two alternatives have not been disproven.

1.4 Single-cell mRNA sequencing technologies⁴

In the last decade, Biology has been heavily influenced by the rise of omics approaches, which can provide information on large pools of molecules simultaneously. This effect is particularly relevant for techniques that use high-throughput sequencing platforms, such as whole-genome sequencing, sequencing of chromatin immunoprecipitation products (ChIP-seq), sequencing of accessible chromatin (ATAC-seq) and sequencing of RNA (RNA-seq). The comprehensive characterization made possible through these methods is remarkably powerful, and has been driving remarkable progress in multiple biological and biomedical fields.

The level of proteins, the final products that carry out most functions in the cell, constitutes the most accurate data to characterize and understand cell function. Unfortunately, the cost, scalability and complexity of proteomics are significantly behind nucleic acid sequencing techniques [50]. Consequently, and since RNA levels explain approximately 40% of variance in protein levels [51], gene expression is most often used as a proxy for protein levels.

For most studies, RNA-seq profiling had focused on populations of cells (bulk RNA-seq), which imposed two main limitations to this method. First, the number of cells needed for bulk RNA-seq is typically in the order of the hundred of thousands of cells, which can be problematic for rare populations. Second, the population-perspective implies that FACS-sorting, cell culture selecting for specific cell types, tissue dissection or any other strategy for cell enrichment has to be used prior to bulk RNA-seq in order to obtain a group of cells that is considered homogeneous. Naturally, homogeneity can only be as good as the knowledge of the population structure and its associated markers, as well as the efficiency and sensitivity of the purification step. Consequently, due to either lack of knowledge or imperfect enrichment, heterogeneity can easily be introduced in the populations analysed by bulk RNA-seq, meaning that results will be an average of the gene expression levels across all cells and all potential subpopulations captured.

In 2009 however, soon after bulk RNA-seq was introduced, the first single-cell RNA-sequencing (scRNA-seq) method was published [52], laying the ground for the exploration of cell populations in a comprehensive and unsupervised manner. In the following years, single-cell transcriptomics went through a speedy maturation period, both experimentally and analytically, while they have also been applied to an increasing number of biological questions and fields.

⁴ Parts of this section were adapted from [171].

In the next sections, the current state of the scRNA-seq field will be addressed, in terms of protocols and platforms available, their challenges, as well as the most common and useful analytical tools and strategies used when analyzing such datasets. This Chapter will end a few examples of scRNA-seq applied to the Immunology field will be explored.

1.4.1.1 Experimental protocols for scRNA-seq

All scRNA-seq protocols involve four major steps: cell isolation (and lysis), reverse transcription, amplification, sequencing library preparation (and sequencing). Each of these steps contains particular challenges, which have been addressed by diverse strategies (Figure 4).

Cell isolation is very time-consuming if not automated in any way. Hence, the initial scRNA-seq studies, which were based on manual isolation of single-cells, were limited to very low numbers (reviewed in [53]). However, the throughput of scRNA-seq methods quickly increased from tens of cells up to tens of thousands, as capture of single-cells moved to automated valve-based microfluidics systems, single-cell sorting into 96 and 384-well plates, and more recently, to droplet-based microfluidics and micro-well methods.

The Fluidigm C1 Autoprep System (C1) was the first automated alternative to perform scRNA-seq in a medium-throughput manner. In this method, cells in suspension are loaded into a disposable microfluidic chip where they are then captured in 96 individual compartments (capture sites). The capture efficiency, *i.e.* how many of the capture sites contain a single cell, can be assessed under a microscope by the user, and it will mainly depend on the cell type being used (size and physical properties), on the number of cells loaded, and the amount of debris present on the cell suspension. The reagents for cell lysis, reverse transcription and amplification are then added to the chip that is then loaded once again into the C1, which processes the single-cells up to cDNA within approximately 8 hours. This platform, being very user friendly, with little hands-on time needed, enabled the first generation of studies with considerable number of cells to be produced. The data generated not only provided the first glimpse on how biological systems are organised at the single-cell level, but were also used to develop fundamental analytical tools and strategies, some of which are mentioned in the next section.

Nonetheless, as time went by, several drawbacks of this system became apparent: the need to have a dedicated machine, the high cost of consumables and reagents (initially limited to the SMARTer kit (Clontech)), lack of flexibility to accommodate experimental designs that include multiple conditions or

controls, lack of ability to alter the in-built scripts. Although some of these issues have been addressed, namely by allowing the use of customised programs which opened C1 to scRNA-seq chemistries besides SMARTer, this platform was increasingly overshadowed by newer approaches in the last few years.

The publication of the Smart-seq2 protocol [54], an improved SMARTer chemistry that uses off-the-shelf reagents, has addressed several of the issues affecting C1. First and foremost, it made the procedure much more cost effective by not relying on commercial kits. Second, this protocol relied on FACS sorting into microtiter plates to capture single-cells. The lower costs achieved per cell allowed the developers of Smart-seq2 to use larger volumes, thus facilitating plate-based reactions. Third, a plate-based approach does not require specialized equipment such as the C1, and it dramatically increased the flexibility to include multiple conditions in scRNA-seq experiments, *e.g.* different cell types, different conditions, use of controls. Finally, after collecting cells in plates, the resulting lysates can be stored and processed later, which further increases flexibility and facilitates collaborative work between distantly located groups.

Manually, plate-based Smart-seq2 is simple enough to achieve medium-throughput single-cell processing (up to ten 96-well plates per day, from RNA to amplified cDNA). However, for increased throughput, Smart-seq2 can be adapted to use robots (*e.g.* Mantis (Formulatrix), Hamilton STAR (Hamilton Robotics) liquid handlers).

More recently, new approaches have been developed that further decrease the cost per cell while increasing the number of cells per run to the order of tens of thousands. So far, droplet-based microfluidics is the most widely-used alternative, with two publicly available protocols Drop-seq [55], InDrops [56] and its commercial variant, Chromium (10xGenomics). These methods entrap single-cells in aqueous droplets, together with either microparticles or hydrogel beads, which in turn deliver barcoded primers that tag the RNA from each cell uniquely.

Using nanoliter droplets as microreactors decreases the amount of reagents needed, driving the price dramatically down, allowing for the analysis of tens of thousands of cells. In addition, no cell-size biases are imposed by this strategy, in contrast with valve-based microfluidics. Capture efficiency in InDrops and Chromium is around 70%, in clear advantage over other approaches (valve-based microfluidics: 1-10%, Drop-seq: 2-4%). Despite all the advantages, some limitations of droplet-based approaches should be considered, namely the inability to link transcriptome data to other single-cell information acquired previously (*e.g.* fluorescence, morphology) and the decreased sensitivity for gene detection. The latter limitation stems from the fact that sequencing more cells up to saturation (*i.e.*

when further increase of sequencing depth does not translate to more genes detected) becomes limiting in terms of cost. Thus, when comparing populations with subtle transcriptomic differences, other approaches should be preferred. Finally, the number of starting cells should also be considered, as 10 000 cells or more are preferred for droplet-based experiments.

Table 2 – Single-cell capture alternatives.

	Micromanipulation	Valve-based microfluidics	Plate-based	Droplet-based microfluidics
No. of cells	Tens	Hundreds	Hundreds	Thousands/tens of thousands
Dedicated platform	Not needed	Needed	Not needed	Needed
Cell selection	Morphology; fluorescence	Pre-capture (e.g. FACS)	FACS	Pre-capture (e.g. FACS)
Single-cell information	Visual observation	Visual observation	FACS measurements (protein markers, size, granularity)	No
Reaction volume	Microliter	Nanoliter	Microliter	Nanoliter
Time	Very slow	Fast	Fast	Fast
Experimental design	More flexible	Less flexible	More flexible	Less flexible
Cost per cell	\$\$	\$\$\$	\$\$	\$
Percentage of initial cells	NA	1-10%	NA	Drop-seq: 2-4% InDrop: 60-90%

In parallel, different methods of reverse transcription and amplification were also introduced, granting access to either full-length transcripts, 5' or 3' fragments (reviewed in [57]), either by PCR or *in vitro* transcription (IVT). The large repertoire of currently existing protocols for scRNA-seq is an outcome of the combinatorial usage of the different strategies for cell isolation, reverse transcription and amplification, which were summarised in Figure 4.

In addition, an increasing number of protocols are now allowing the integration of scRNA-seq with single-cell genomics [58,59], epigenomics [60] and CRISPR screenings [61–63].

Systematic comparisons of scRNA-seq protocols have now been published [64,65]. On one hand, they show that most single-cell mRNA protocols are able to correctly assess the abundance of mRNA molecules present across cells (good accuracy). On the other hand, significant differences in lower detection limits were detected between protocols (variable sensitivity).

Nevertheless, read depth per cell rather than protocol seems like the main factor to produce reliable scRNA-seq data. Ensuring around 1 million reads/cell is recommended to maximize accuracy and, more importantly, sensitivity, while keeping costs to the minimum [64].

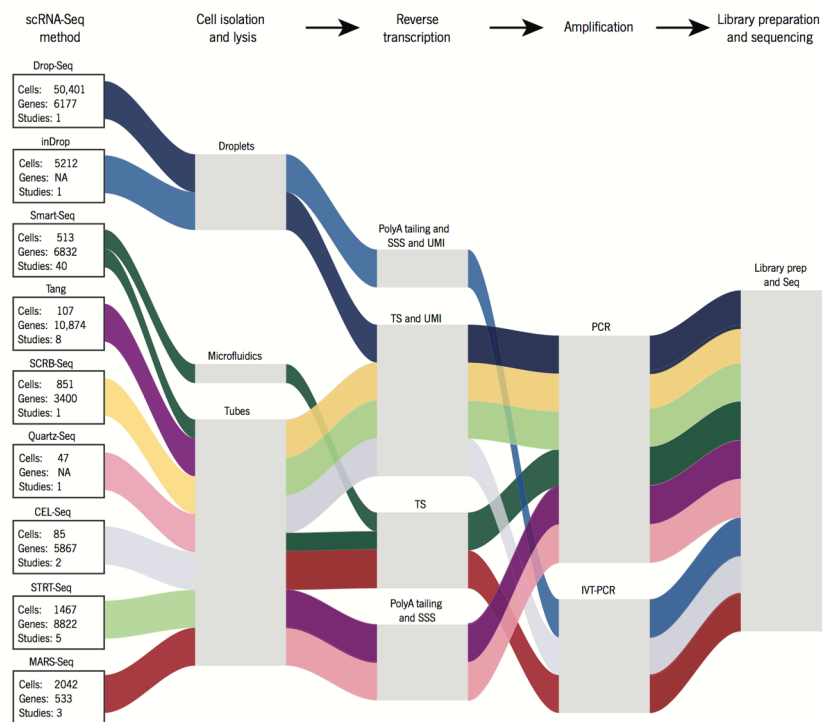


Figure 4 – Single-cell RNA-sequencing methods.

For each of the steps of scRNA-seq protocols (top row), several alternative strategies can be used. Combinations of all these alternatives give rise to the currently available methods. From [66].

Therefore, the choice of protocol will ultimately be influenced by practical factors [65]. It will depend on the number of starting cells available (droplet-based microfluidics usually require more input cells [65]), whether full transcripts are required or not (while 3' or 5'-end reads are suitable for gene quantification, study of isoform variants [67] or TCR reconstruction [68] require full-transcripts), whether index-sorting data for each cell is desirable (plate-based approaches can record fluorescence levels of important proteins [69]), number of conditions and controls (plate-based approaches are more amenable to complex experimental designs), how many cells and how well characterized they should be (for similar sequencing costs, droplet-based approaches will yield more cells but will detect

mostly their most highly expressed genes, while plate-based approaches yield fewer cells but have a larger spectrum of gene detection), the cost and the availability of dedicated platforms (microfluidic approaches require either commercial or home-made dedicated platforms, while plate-based approaches depend on FACS machines).

1.4.1.2 Computational analysis of scRNA-seq data

Computationally, scRNA-seq data has presented new challenges to the RNA-seq community [70]. These challenges include, among others: (1) Technical differences between cells stemming from library quality and batch effects: low quality of libraries can be detected by a preliminary quality-control stage where quality of each library is assessed based on various criteria (such as total number of mapped reads or number of detected genes) [71]. Batch effects can be addressed by batch correction methods (these include traditional and newly developed methods to remove unwanted factors[72]). (2) Variable or inefficient capture of mRNA (or “dropouts”) due to low amount of starting material, which can affect the ability to detect lowly-expressed genes. This can be addressed by methods that impute the data and restore its likely structure [73]. (3) Difficulties in discerning between technical noise and genuine biological heterogeneity.

In the past couple of years, these have been addressed by a combination of adaptations of bulk-RNA analysis methods, and the development of completely new tools, several of which have now been integrated into user-friendly packages [72].

In general terms, gene expression data of a differentiation process with single-cell resolution has the potential to provide answers to five main questions [53]:

1. Are there distinct clusters of cells within the general cell population (corresponding to either a variety of differentiation stages or to various, fully-differentiated, cell types)?
2. Which genes characterize each cell population?
3. Which gene modules regulate differentiation?
4. How do cells progress through the differentiation process?
5. How is the differentiation process spatially reflected in the relevant tissue?

These questions can be addressed by utilising different techniques (Figure 5):

1. Identifying cell clusters can be achieved by either applying dimensionality reduction techniques or hierarchical clustering algorithms.
2. The biological roles of the identified clusters can then be evaluated based on genes that differentially expressed (DE) between clusters, using available methods to infer differential expression (reviewed in [72,74]).
3. Measuring gene co-occurrence or gene-gene correlations across single-cells can be used to build co-expression networks that can yield insights into important gene modules and gene regulation involved in differentiation. With the recent integration of CRISPR screenings with scRNA-seq [61–63], there is now the potential to systematically perturb these networks and assess true regulatory relationships between their elements.
4. Although single-cell data of a differentiation path is a static snapshot, differentiation is a dynamic and continuous process, encompassing a continuous range of cell-states along this process. Multiple algorithms have been recently developed to interpret such cell-to-cell differences and reconstruct trajectories of cell differentiation. Ordering cells along this “pseudotime” enables studying gene kinetics during differentiation, inference of regulatory relationships and determination of branching points between cell-fates [75–80] (reviewed in [72,74]).
5. Finally, in systems such as embryonic development and organogenesis where differentiation and localization go in tandem, matching single-cells back to their original location within the tissue of interest is imperative to understanding the tissue physiology. This can be achieved computationally by mapping single-cells against a spatial reference [81,82], allowing the spatial visualization of cell subpopulations, markers and gene modules identified in previous stages.

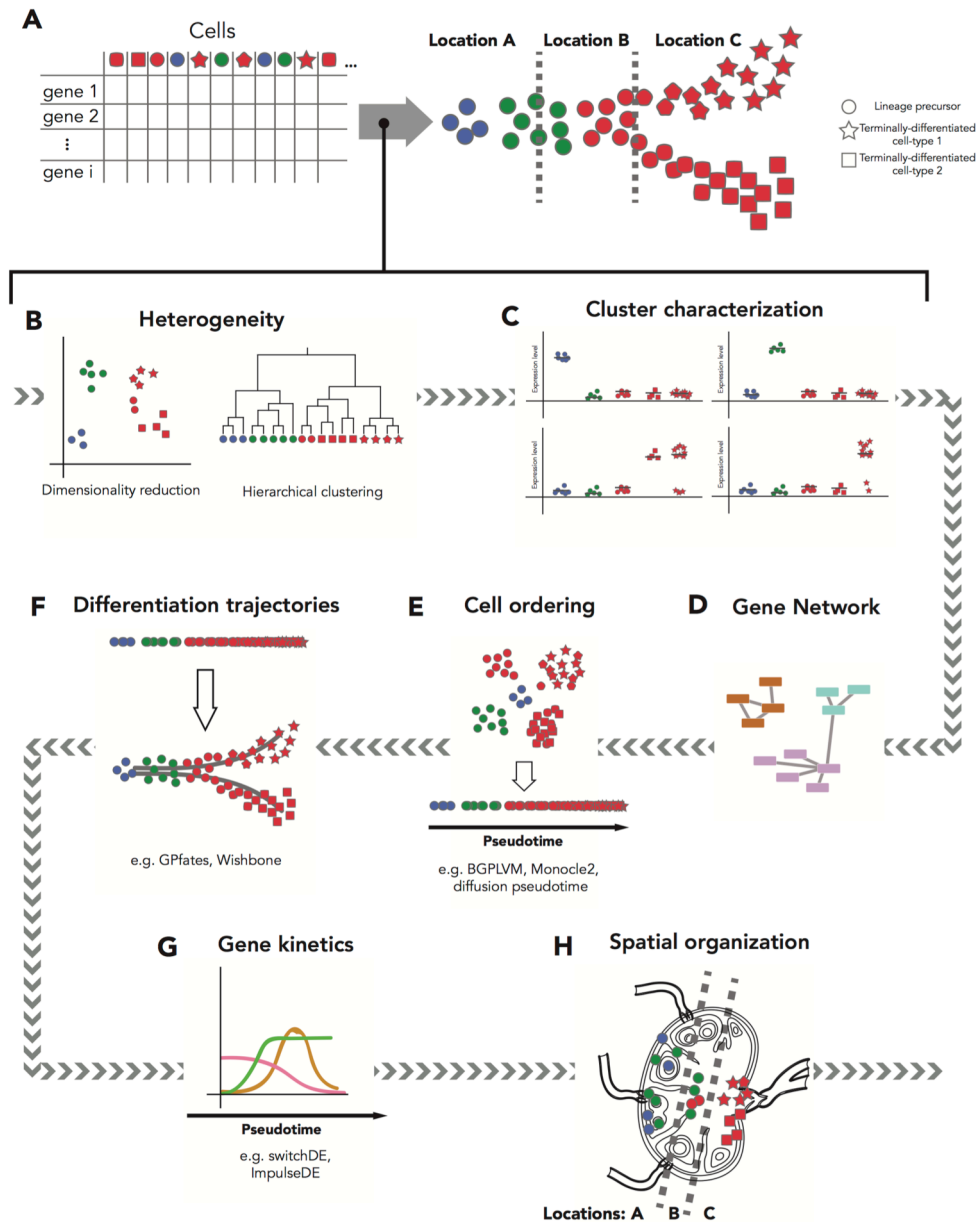


Figure 5 - scRNA-seq analyses: from a gene expression matrix datasets to complete reconstruction of a differentiation process.

(A) Analysis of scRNA-seq take as input a cell vs. gene matrix and ideally lead to an ordered set of cells, fully defined and structured at multiple levels. (B-H) Several steps have to be undertaken to conduct a meaningful scRNA-seq analysis. (B) Heterogeneity within cell populations can be addressed either by dimensionality reduction algorithms or hierarchical clustering (reviewed in [72,74]). (C) Multiple methods exist to determine which genes are differentially expressed by each cluster[83]. (D) Individual genes can then be grouped into regulatory networks[74,84] to predict functional associations. (E) At the cell level, inferring a pseudotime of high resolution differentiation is a powerful application of scRNA-seq. Naturally, numerous computational tools have now been developed to order cells based on their transcriptomic differences [77,79,80,85,86] (reviewed in [72,74]). (F) Similarly, tools to specifically separate more than one co-existing lineages are also emerging [75,79,80,85] (reviewed in [72,74]) (G) Once differentiation trajectories are established, one can investigate the kinetics of gene expression along them [87,88]. (H) Mapping single-cell information back to the tissue of origin, thus recovering spatial information, has been attempted in some systems[81,82].

I.4.1.3 Applications of scRNA-seq to Immunology

The rise of scRNA-seq has been influencing several fields across Biology. The potential impact that these technologies can have specifically in Immunology is very significant, as heterogeneity and plasticity within known populations is remarkably high (as discussed in section I.3.2), namely within the adaptive immune system.

To cope with the vast array of invading pathogens, T cells rely on naïve cells that can rapidly adapt and differentiate towards specific fates according to the any given immune challenge. Single-cell resolution of adaptive immune responses can capture in detail such processes, giving insights into how different challenges and tissues influence immune differentiation, which cell subtypes arise in different stages of infection, and how adaptive immune cells further differentiate to acquire memory phenotypes.

A recent work used scRNA-seq of in vivo and in vitro murine Th17 cells to explore their path to pathogenesis in autoimmune encephalomyelitis (EAE) [89] (Figure 6A). The authors collected Th17 cells from lymph node (LN) and the central nervous system (CNS), and assigned the cells into discrete stages of differentiation. Coupling this information with scRNA-seq of in vitro Th17, a subset of CNS-dwelling Th17 was shown to be particularly pathogenic, while most other subsets were non-pathogenic. Ultimately, multiple genes were shown to drive pathogenesis.

In another study, differentiation within the CD8⁺ lymphocyte branch was addressed in the context of lymphocytic choriomeningitis virus (LCMV) infection [90]. Here, several time-points rather than tissues were studied, spanning both early and late infection. Single-cell data shows that two subpopulations arise as soon as responder cells undergo their first cell division, adopting either T effector or T memory phenotypes. This surprisingly early separation of phenotypes was then used to classify intermediate differentiation states as “effector-like” or “memory-like”, identifying *Ezh2* as a new “effector” driver gene, associated with epigenetic repression of “memory” traits.

Differentiation of naïve CD4⁺ T cells into effector stages was tracked by scRNA-seq during malaria infection in a mouse model [80] (Figure 6B). The authors ordered hundreds of cells from multiple time-points based on their transcriptome similarities, establishing a continuous “pseudotime” using a novel approach based on Gaussian processes. As in many other differentiation events, at a specific differentiation stage along the pseudotime, two cell types arise: Th1 and Tfh. Therefore, the authors applied a second approach in order to disentangle one differentiation trajectory from the other. This framework allowed an in-depth study of the Th1-Tfh bifurcation point, which led to the discovery of key genes, such as *Lgals1*, driving differentiation along each lineage.

T cells, as a population, are able to recognize an immense variety of pathogenic molecules. This ability relies on the somatic recombination of the T-cell receptor (TCR) locus, which can be used as a barcode to follow development of T cells, such as in the above mentioned examples. scRNA-seq of T cells has provided a unique resource to study TCR recombination and clonality, and several methods have been developed to reconstruct TCR sequences in single-cell data [68,91–94]. Using one of these approaches [68], clonally-related T cells were shown to be able to adopt both Th1 and Tfh fates during malaria infection [80].

The combination of single-cell resolution and cell tagging can be valuable in tracking cellular fates during cell differentiation of numerous lineages unrelated to T cells [95]. It is thus likely that several synthetic barcoding strategies, mostly using genome-editing systems [96–101], would prove to be powerful tools in future studies of cell differentiation.

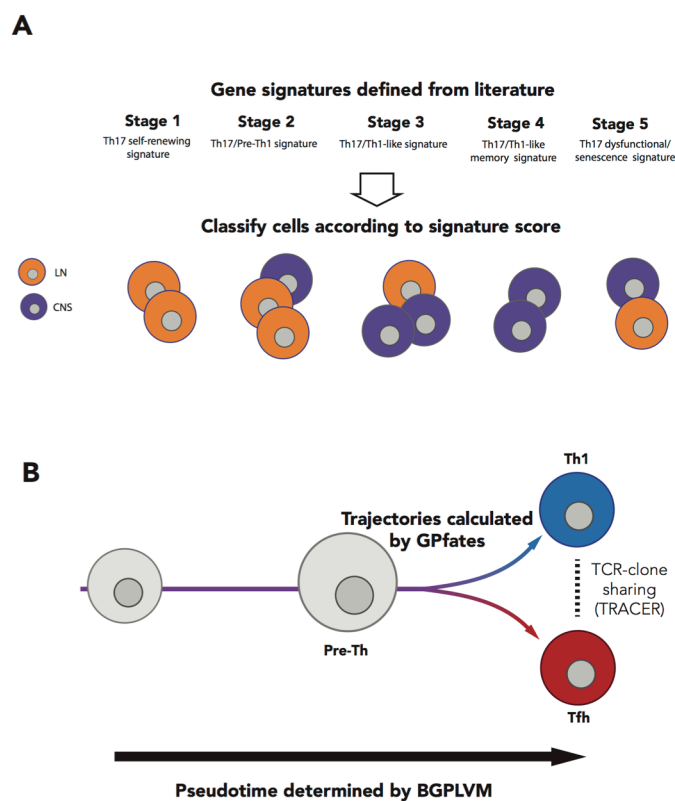


Figure 6 - Two contrasting strategies for dissecting T cell differentiation using scRNA-seq.

(A) Supervised functional annotation approach used to classify subpopulations of Th17 cells from lymph-node (LN) and central nervous system (CNS) in EAE [89]. First, based on the literature, gene signatures for five stages were defined. Cells were scored for each of the five stages and were assigned to the stage for which they score the highest. (B) Unsupervised dimensionality reduction approach based on Gaussian processes (Bayesian Gaussian Processes Latent Variable Modelling – BGPLVM) is used to order cells along a continuous variable of transcriptomic changes – pseudotime. At a second stage, two trends of cell differentiation were detected and separated using Overlapping Mixture of Gaussian Processes (OMGP/GPfates [80]).

I.5 Objectives of this thesis

Both mTECs and Tregs have attracted a great deal of attention in recent years. However, and as highlighted in the previous sections, the more has been uncovered about their functions, phenotypes, development and/or population structure, the more intriguing and challenging they have become. Further characterization of both systems has collided with the lack of appropriate marker genes, and biased by the limited range of the existing ones. Although the use of RNA-sequencing has brought a more comprehensive characterization of the known populations of cells, this “bulk” perspective has not been enough to efficiently dissect the complex heterogeneity within them. In this thesis, I thus decided to leverage the power of state-of-the-art scRNA-seq approaches to perform a comprehensive and unbiased characterization of mTECs and Tregs.

By applying scRNA-seq to the mTEC population (Chapter III), my main objective is to characterize their developmental process. Due to the constant turnover of mTECs in the adult medulla, it is likely that several stages can be captured in the snapshot that scRNA-seq data represents. My sub-aims include:

- Determination of potential drivers of mTEC differentiation
- Characterization of gene expression between subpopulations
- Defining the role played by each subpopulation and their ability to induce immune tolerance

By profiling Treg and Tmem cells from multiple lymphoid and non-lymphoid tissues (Chapter IV), I am aiming at determining Treg features that explain their recruitment, adaptation and retention in non-lymphoid tissues.

Sub-aims include:

- Determining gene signatures that characterize different cell types and/or tissues of origin
- Assessing the presence of subpopulations of Treg and Tmem within each tissue
- Determining how and where non-lymphoid tissue Treg phenotype is established
- Comparing recruitment and adaptation of Tregs in steady-state and disease
- Evaluating conservation of marker genes between mouse and human

CHAPTER II

Materials and Methods

CHAPTER II - Material and Methods

This Chapter is divided into two main sections, each concerning one of the two systems studied in this thesis: mTECs and Tregs.

II.1 scRNA-seq of the mTEC population⁵

II.1.1 Experimental procedures

II.1.1.1 Mice

C57BL/6 mice were maintained under specific pathogen-free conditions at the Wellcome Trust Genome Campus Research Support Facility (Cambridge, UK). These animal facilities are approved by and registered with the UK Home Office. All procedures were in accordance with the Animals (Scientific Procedures) Act 1986. The protocols were approved by the Animal Welfare and Ethical Review Body of the Wellcome Trust Genome Campus.

II.1.1.2 Tissue processing

Thymi were collected from 2 and 4 week-old wild-type C57BL/6 male and female mice. Epithelial cell isolation was performed based on [102]. Up to 3 thymi were cleaned of fat and connective tissue, finely minced and pooled together. Thymocytes were flushed by gentle agitation with a magnetic stirrer in RPMI-1640 for 30 min, at 4°C. Thymic fragments were recovered by settling, and the supernatant discarded. After further dispersion, three additional washes were performed. Fragments were then incubated in 5 mL of 0.125% (w/v) Collagenase D and 0.1% (w/v) DNase I (both from Roche) in RPMI-1640, at 37°C for 15 min, with gentle pipetting every 5 min. The supernatant was collected and kept on ice, while the thymic fragments were subject to two further incubations. The remaining fragments were finally resuspended in 5ml of 0.125%(w/v) Collagenase/Dispase (Roche) and 0.1%(w/v) DNaseI in RPMI-1640 for 30 min at 37°C, with gentle agitation every 15 min. All the collected fractions were pooled, centrifuged at 450x g for 5 min and incubated in 5 mM EDTA, 1%

⁵ Parts of this section were adapted from [170].

FCS, 0.02% (w/v) NaN_3 in PBS (EDTA/FACS buffer) for 10 min at 4°C. After filtering through a 100 μm -strainer, the resulting cell suspension was depleted of hematopoietic cells by Magnetic-Activated Cell Sorting (MACS) using CD45-MicroBeads (Miltenyi Biotec).

II.1.1.3 FACS sorting

For sorting, the recovered fraction was blocked using anti-CD16/CD32 (clone 2.4G2, Tonbo) and then stained using anti-CD45-PerCP-Cy5.5 (clone 30-F11, BioLegend), anti-Ly-51-FITC (clone 6C3, BioLegend), anti-CD326(Ep-CAM)-AF647 (clone G8.8, BioLegend), UEA-1-Biotin and Streptavidin-Pacific Blue. mTECs (CD45-Ly-51-UEA⁺) were sorted with a MoFlo™ XDP (Beckman Coulter, Inc.). Propidium iodide was used as a viability dye.

II.1.1.4 Single-cell RNA-sequencing using C1 Single-Cell Auto Prep System

The microfluidics platform C1 Single-Cell Auto Prep System (C1) from Fluidigm was used for single-cell capture, reverse transcription and cDNA amplification. The protocol (PN 100-5950 B1) and original scripts provided by the company were used for all runs. This protocol relies on the SMARTer chemistry (SMARTer Ultra Low Input RNA Kit for Sequencing - v3, Clontech, cat.#634852). In total, three C1 runs were performed: one with cells from 2-week old mice and, on a separate day, two parallel runs with cells from 4-week old mice. After priming of a “medium-sized” chip, ideal for 10-17 μm cells (C1™ Single-Cell Preamp IFC, Fluidigm, cat.#100-5480), 5000-10,000 recently sorted mTECs were loaded and captured. Visual inspection of all 96 capture-sites across the chip provided information on number of cells captured, presence of debris and overall appearance of cells. No live/dead staining was used, as live mTECs had previously been sorted based on PI. Lysis, reverse transcription and PCR mixes (Table 3) were all loaded into the appropriate inlets. Within the following 6h30min, C1 took the cells through lysis, reverse transcription and amplification (Table 4), then taking 2 hours to harvest the resulting cDNA. All the material (approximately 3 μL) was then manually transferred from the C1 chip to 96-well plates containing 10 μL of C1 DNA Dilution Reagent. At this point, samples could be frozen down. However, quality of the cDNA was typically assessed before storage, by running 1 μL of cDNA on an Agilent Bioanalyzer High-sensitivity chip. Ideally, cDNA size ranges between 400-10,000bp (according to SMARTer Ultra Low Input RNA Kit for Sequencing - v3 User Manual). cDNA was not detected between 0-400bp, which suggests there were not issues of RNA degradation, contamination or primer concatenation. A small dataset of mTECs was processed using the Smartseq2 protocol (section II.2.1.9).

Table 3 – Mixes for scRNA-seq with C1 Single-Cell Auto Prep System.

Cell lysis		Reverse transcription		Amplification (PCR)	
Reagents	Vol. (μL)	Reagents	Vol. (μL)	Reagents	Vol. (μL)
ERCC mix (1:2000) (Ambion)	1.0	C1 Loading reagent (Fluidigm)	1.2	PCR water (Clontech)	63.5
RNase Inhibitor (Clontech)	0.5	5x First-strand buffer (Clontech)	11.2	10x Advantage 2 PCR buffer (Clontech)	10.0
3' SMART CDS Primer II A (Clontech)	7.0	Dithiothreitol (Clontech)	1.4	50x dNTP mix (Clontech)	4.0
Dilution buffer (Clontech)	11.5	dNTP mix (10mM each) (Clontech)	5.6	IS PCR primer (Clontech)	4.0
		SMARTer II A Oligonucleotide (Clontech)	5.6	50x Advantage 2 Polymerase mix (Clontech)	4.0
		RNase inhibitor (Clontech)	1.4	C1 Loading reagent (Fluidigm)	4.5
		SMARTScribe™ reverse transcriptase (Clontech)	5.6		
TOTAL	20.0	TOTAL	32.0	TOTAL	90.0

Table 4 – C1 thermal cycling protocols.

Cell lysis		Amplification (PCR)		
Temperature	Time	Temperature	Time	Cycles
72°C	3min	95°C	1min	1
4°C	10min	95°C	20sec	
25°C	1min	58°C	4min	5
Reverse transcription		68°C	6min	
Temperature	Time	95°C	20sec	
42°C	90min	64°C	30sec	9
70°C	10min	68°C	6min	
		95°C	30sec	
		64°C	30sec	7
		68°C	7min	
		72°C	10min	1

II.1.1.5 Nextera XT Illumina Library prep – manual

The information gathered from visual inspection of C1 chips (see section II.1.1.4) was then used to select single-cells for downstream processing. The libraries for sequencing were prepared using Nextera XT DNA Sample Preparation Kit (Illumina), according to the protocol supplied by Fluidigm (PN 100-5950 B1). Based on the average concentration of samples assessed by Bioanalyzer, each plate was diluted to a final concentration in the 0.1-0.3ng/ μ L range. 1.25 μ L of each sample were mixed with 3.75 μ L of tagmentation mix (pre-mixed 2.5 μ L of Tagment DNA Buffer and 1.25 μ L of Amplification Tagment Mix) on a new 96-well plate. Plates were then vortexed at medium speed for 20sec and centrifuged (2000xg, 5min), to remove bubbles. They were then tagmented at 55°C for 10min, and then kept at 10°C until 1.25 μ L of NT buffer were added per sample to stop the reaction. After vortexing at medium speed and centrifuging (2000xg, 5min), 3.75 μ L of PCR Master Mix (NPM) were added to each sample. Indices from Illumina Nextera XT Index Kit were then added to each plate: 1.25 μ L of Index Primer 1 (N701-N712) were added to each row (from 1 to 12, respectively), and 1.25 μ L of Index Primer 2 (S501-S508) to each column (from A to H, respectively). After vortexing at medium speed and centrifuging (2000xg, 2min), plates went through PCR amplification (Table 5). Per plate, 1 μ L of all 96 samples were then pooled together in one 1.5mL eppendorf. Finally, AMPure XP beads (Beckman Coulter) were used to purify sample pools for sequencing (removal of excess dNTPs, primers, salts and enzymes). To each pool, 87 μ L of room-temperature beads were added (1:0.9) and mixed. Libraries from 96 single cells were pooled (1 μ L per sample) and subsequently purified using AMPure XP beads (Beckman Coulter).

Table 5 - Nextera XT PCR protocol.

Amplification (PCR)		
Temperature	Time	Cycles
72°C	3min	1
95°C	30sec	1
95°C	10sec	
55°C	30sec	12
72°C	60sec	
72°C	5min	1
10°C	hold	

II.1.1.6 Illumina High-throughput Sequencing

Pooled samples were sequenced on an Illumina HiSeq 2500 instrument, using paired-end 100-base pair reads.

II.1.2 Computational analyses

II.1.2.1 Processing and quality control of single-cell mRNA-seq data

Reads were mapped to the *Mus musculus* genome (Ensembl version 38.75) concatenated with the ERCC sequences, using GSNAP (version 2014-05-15_v2, [103]) with default parameters. The read counts for each gene were determined using HTseq (version 2.6.0, [104]), and TPM calculated. Only genes expressed with 5 or more TPM in at least 5 cells across all datasets were kept. As cell quality control measures, cells with fewer than 1000 genes, fewer than 500,000 reads mapping to exons or with more than 20% reads mapped to mitochondrial genes were excluded from further analyses.

Cyclone package [105] was used to determine the cell cycle phase of each cell. scLVM package [106] was run for all datasets, and the corrected matrices were used for the datasets showing relevant number of cycling cells, *i.e.* Sansom *et al.* and Brennecke *et al.* datasets [28,107]. For the in-house dataset, which presented a very limited number of cycling cells, I chose not to use scLVM as it introduced additional confounding factors to the analysis: the effect of cell size (number of genes detected) was spread across multiple PC1s (Figure 12A), while it exhibited strong negative correlation with PC1 in the original matrix (Spearman correlation -0.92). Nevertheless, cell consensus clustering (see below) was also performed with the corrected matrix to confirm that the three clusters were not affected by cell cycle.

II.1.2.2 Differential expression

For DE analysis, two linear models were fit to the expression levels of each gene separately: a full model containing the information for each mTEC subpopulation and a reduced model only including an intercept term. These were then compared by a likelihood-ratio test, and *p*-values were adjusted to account for the false discovery rate associated to multiple testing.

II.1.2.3 Consensus Clustering

For cell clustering, the genes contributing the most to PC2 and PC3 in the main dataset PCA were used ($|\text{gene loading}| > 0.02$). For each dataset, clusters were determined using the ConsensusClusterPlus package [108] with 70% cell and gene resampling, in 2000 resampling events. Although up to 6 clusters were explored per dataset, 3 clusters was the most stable option in all cases.

II.1.2.4 Genomic clustering

A nearest-neighbour method was used to assess genomic clustering of expressed TRAs. First, a set of genes with a similar distance distribution for each TRA subset was determined, minimizing the Kullback-Leibler (KL) divergence between them in an iterative manner. The residual divergence between these distributions (KL_{gen}) will later be taken into account. Then, for each cell, the distance between each expressed TRA gene and its closest expressed TRA was calculated. In parallel, a similar number of expressed control genes were sampled and their distances to the nearest expressed neighbour control gene were measured. This sampling step was repeated a thousand times per cell and these distances were used as background. To compare the mean distance in both distributions the Mann-Whitney-Wilcoxon test was used, and to quantify how similar the distributions were, I calculated their KL divergence and subtracted KL_{gen}. This analysis was conducted per mTEC sub-population and for each TRA subset, *i.e.* *Aire*-dependent, *Aire*-enhanced and *Aire*-unaffected TRAs.

II.1.2.5 Monocle

The Monocle algorithm was originally described in [79]. A new release (Monocle2), with improved algorithms was used. In brief, the algorithm estimates for each cell a degree of differentiation based on the expression of a defined set of marker genes. The analysis was performed using the normalized data (TPM) and all genes expressed. The direction of pseudo-time was inferred from expression patterns of known markers of jTECs and mTECs.

II.1.2.6 Binding motif enrichment analysis

Binding motif enrichment analysis was performed on the genes, which are correlated with PC2 or PC3 in Figure 17A. Genes which have loadings greater than 0.02 on PC2 or PC3 were selected and input to the `gprofile` function in the R package `gProfileR` [109], using its default settings. The function `gprofile` outputs the enriched TF families and corresponding target genes from the input gene set. For each pair of TF and target gene, Spearman correlation and Jaccard Index were calculated. The Jaccard Index was calculated based on binarized gene expression levels. Network visualisation was created using TF-target pairs with Spearman correlation $|r| > 0.3$ ($p\text{-value} < 0.005$) and Jaccard Index $j > 0.3$.

II.2 scTreg of Treg populations⁶

II.2.1 Experimental procedures

II.2.1.1 Mice

All mice were maintained under specific pathogen-free conditions at the Wellcome Trust Genome Campus Research Support Facility (Cambridge, UK) and at the Kennedy Institute for Rheumatology (Oxford, UK). All procedures were in accordance with the Animals Scientific Procedures Act 1986. For steady-state experiments, the Foxp3-GFP-KI mouse reporter line [110] was used. The melanoma challenge was performed in Foxp3-IRES-GFP knock-in reporter mice [111] purchased from The Jackson Laboratory (stock no. 006772). In both cases, 6-14 week-old females were used.

II.2.1.2 Human samples

Human skin and blood samples were obtained from patients undergoing breast reduction plastic surgeries (REC approval number: 08/H0906/95+5).

Surgical-resection specimens were obtained from patients attending the John Radcliffe Hospital Gastroenterology Unit (Oxford, UK). These specimens were obtained from normal regions of bowel adjacent to resected colorectal tumours from patients undergoing surgery. Informed, written consent was obtained from all donors. Human experimental protocols were approved by the NHS Research Ethics System (Reference number:11/YH/0020).

II.2.1.3 Isolation of murine leukocytes for steady-state skin dataset

To isolate leukocytes from ear tissue, ears were removed at the base, split into halves and cut into very small pieces. Tissue was digested in 3.5ml RPMI media with 0.1% BSA, 15mM Hepes, 1mg/ml collagenase D (Roche) and 450µg/ml Liberase TL (Roche) for 60 minutes at 37°C in a shaking incubator at 200rpm. Digested tissue was passed through a 18G needle to further disrupt the tissue and release cells. Cells were passed through a 70µm cell strainer, and the digestion was terminated by addition of ice-cold RPMI containing 0.1%BSA and 5mM EDTA. A three-layer (30/40/70%) Percoll

⁶ Parts of this section were adapted from [172].

density-gradient was used to enrich for the lymphocytes. Cells obtained from the digestion were layered in the 30% layer on top of the 40% and 70% layers, and centrifuged for 20 minutes at 1800rpm without brake. Cells at the 40/70% interface were collected for the subsequent analysis. Cell suspensions from spleen and bLN were prepared as described previously [112].

II.2.1.4 Isolation of murine leukocytes for steady-state colon dataset

Colons were washed twice in RPMI media with 0.1%BSA and 5mM EDTA in a shaking incubator at 200rpm at 37°C to remove epithelial cells. The tissue was then digested for an hour in the presence of RPMI/10%FCS/15mM Hepes with 100U/ml collagenase VIII. Digestion was terminated by addition of ice-cold RPMI/10% FCS/5mM EDTA. A three-layer (30/40/70%) Percoll density-gradient was used to enrich for the lymphocytes. Cells obtained from the digestion were layered in the 30% layer on top of the 40% and 70% layers, and centrifuged for 20 minutes at 1800rpm without brake. Cells at the 40/70% interface were collected for the subsequent analysis. Cell suspensions from spleen and mLN were prepared as described previously [112].

II.2.1.5 Melanoma induction and cell isolation

The melanoma induction experiments were performed in accordance with UK Home Office regulations under Project License PPL 80/2574. Protocol used was adapted from a previous publication [113]. For syngeneic tumours, 2.5×10^5 B16.F10 melanoma cells were inoculated subcutaneously into the shoulder region 6- to 14-week-old female Foxp3-IRES-GFP mice [111]. Animals were excluded only if tumours failed to form or if health concerns were reported. Control Foxp3-IRES-GFP mice were injected with 50 μ l PBS. Tumour size was monitored with calipers and the volume was calculated based on the ellipsoid formula $\pi/6 \times (\text{length} \times \text{width}^2)$. Animals were culled after 11 days. Tumour tissue, tumour-draining (brachial) lymph nodes and spleen were isolated for subsequent analysis. PBS-injected skin, draining lymph nodes (bLN) and spleen were collected from control mice, as well as steady-state skin from the lumbar region along with its draining (inguinal) lymph node (iLN). Tumour and PBS-injected skin were mechanically disrupted and digested in a 1ml mixture of 1 mg/ml collagenase A (Roche) and 0.4 mg/ml DNase I (Roche) in PBS (solution A) at 37°C for 1h with 600rpm rotation. 1ml of PBS containing 1mg/ml Collagenase D (Roche) and 0.4 mg/ml DNase I (solution B) was then added to each sample, which returned to 37 °C for 1h with 600 rpm rotation.

Lymph nodes were digested for 30min in 500 μ l of solution A, and for further 30min after the addition of 500 μ l of solution B. EDTA at the final concentration of 10mM was added to all samples. Spleens were processed as described previously [112]. Suspensions were passed through a 70 μ m mesh before immunostaining with combinations of fluorescently conjugated antibodies (Table 6). Samples from different animals were kept separated throughout processing and sorting. Cells from iLN and lumbar-region skin were later compared to bLN and PBS-injected skin cells, respectively, and shown to be identical (Figure 37E, F).

II.2.1.6 Isolation of leukocytes from Human skin

Plastic surgery skin included reticular dermis to the depth of the fat layer. The upper 200 microns of skin were harvested using a split skin graft knife. Whole skin was digested in RF10 with 1.6mg/ml type IV collagenase for 12-16 hours at 37°C and 5% CO₂. Digest was passed repeatedly through a 10ml pipette until no visible material remained. To yield a single cell suspension, digest was passed through a 100-micron filter into a polypropylene sorting tube. Wells were washed twice using cold sort buffer without calcium or magnesium to collect residual and adherent cells.

II.2.1.7 Isolation of leukocytes from Human colon

Normal regions of bowel adjacent to resected colorectal tumours were prepared as previously described, with minor modifications [114,110]. In brief, mucosa was dissected and washed in 1mM dithiothreitol (DTT) solution for 15 min at room temperature to remove mucus. Specimens were then washed three times in 0.75mM EDTA to deplete epithelial crypts and were digested for 2h in 0.1mg/ml collagenase A solution (Roche, UK). For enrichment of mononuclear cells, digests were centrifuged for 30 min at 500g in a four-layer Percoll gradient and collected at the 40%/60% interface.

II.2.1.8 Peripheral blood mononuclear cell isolation

10ml blood from skin donors were collected into EDTA. Density centrifugation with Lymphoprep was performed according to manufacturer instructions. Recovered cells were cryopreserved by pelleting and resuspending in 1ml heat-inactivated fetal calf serum containing 10% DMSO, and storing at -80°C.

Cryovials were later thawed in water bath, then rapidly being transferred to warmed medium (RPMI 1640 with 100IU/ml penicillin, 10ug/ml streptomycin, 2mM L-glutamine, 10% heat-inactivated fetal calf serum) and filtered through a 100-micron filter.

II.2.1.9 Flow cytometry and single-cell RNA sequencing

Mouse and human cell suspensions were stained with the antibodies in Table 6 and DAPI. Single-cells were sorted in 2µl of Lysis Buffer (1:20 solution of RNase Inhibitor (Clontech, cat. no. 2313A) in 0.2% v/v Triton X-100 (Sigma-Aldrich, cat. no. T9284)) in 96 well plates, spun down and immediately frozen at -80 degrees.

Table 6 - Antibody panels used for flow cytometry.

Species	Marker	Fluorochrome	Manufacturer	Cat no.
Mouse	CD8a	APC-Cy7	BioLegend	100714
Mouse	CD19	APC-Cy7	BioLegend	115530
Mouse	CD11b	APC-Cy7	BioLegend	101226
Mouse	TCRb	AF647	BioLegend	109218
Mouse	CD4	BV510	BioLegend	100553
Mouse	CD44	BV421	BD Biosciences	563970
Mouse	CD62L	AF700	BioLegend	104426
Mouse	ST2	Biotin (streptavidin-PeCy7)	mdBioproducts	101001B
Human	CD45	APCCy7	BD Biosciences	557833
Human	CD3	FITC	BD Biosciences	345763
Human	CD4	AF700	BD Biosciences	557922
Human	CD8	PERCPy5.5	BioLegend	300924
Human	CD25	PE	BioLegend	302606
Human	CD127	BV421	BioLegend	351309
Human	CCR7	PECy7	BioLegend	353226
Human	CD45RA	BV510	BD	563031

Smart-seq2 protocol [54] was largely followed to obtain mRNA libraries from single-cells. Oligo-dT primer, dNTPs (ThermoFisher, cat. no. 10319879) and ERCC RNA Spike-In Mix (1:50,000,000 final dilution, Ambion, cat. no. 4456740) were then added. Reverse Transcription and PCR were performed as previously published [54], using 50U of SMARTScribe™ Reverse Transcriptase (Clontech, cat. no. 639538). The cDNA libraries for sequencing were prepared either using Nextera XT DNA Sample Preparation Kit (section II.1.1.5), manually or with Hamilton 384 head robot (Hamilton Robotics). Libraries from single cells were pooled and purified using AMPure XP beads (Beckman Coulter).

Pooled samples were sequenced aiming at an average depth of 1 Million reads/cell, on an Illumina HiSeq 2500 (paired-end 100-bp reads) or Illumina HiSeq 2000 v4 chemistry (paired-end 75-bp reads).

II.2.2 Computational analyses

II.2.2.1 RNA expression quantification

Gene expression from scRNA-seq data was quantified in Transcripts Per Million (TPM) using Salmon v0.6.0 [115], with the parameters `-fldMax 150000000 -fldMean 350 -fldSD 250 -numBootstraps 100 -biasCorrect -allowOrphans -useVBOpt`. For mouse, the cDNA sequences used contain genes from GRCm38 and sequences from RepBase, as well as ERCC sequences and an EGFP sequence. Since the EGFP RNA is transcribed together with *Foxp3*, TPM from these two genes were added after quantification to represent *Foxp3* expression. For human data quantification, cDNA sequences from GRCh38 and ERCC were used.

II.2.2.2 scRNA-seq quality control

After expression quantification, TPM values for each cell were grouped in an expression matrix. ERCC expression levels were separated and TPM were rescaled to total 1 million per cell. Cells were then filtered based on different quality parameters calculated for each dataset (Figure 24, Table 7). Additionally, the output of TraCeR [68] was used to remove cells without a detected TCR sequence, as well as iNKT and $\gamma\delta$ -T cells (defined as cells with at least one γ and one δ chain detected and no $\alpha\beta$ pair). Two mouse steady-state skin Treg from the Mouse Melanoma dataset were also removed from posterior analysis because, even though they had reconstructed TCR and expression of expected T cell genes, they also presented transcripts expected to appear in other cell types, possibly keratinocytes or Langerhans cells [116,117] (Figure 25D).

Table 7 - Quality control criteria for filtering single cell transcriptomes in each dataset, and parameters for their visualization in tSNE (last two columns). Cells were kept if they passed all these filters.

	Mouse Colon	Mouse Skin	Mouse Melanoma	Human Skin and Colon
Maximum mitochondrial reads (%)	25	20	20	20
Maximum ERCC-derived reads (%)	30	40	40	60
Maximum unmapped reads (%)	40	35	40	60
Minimum number of detected genes (#)	1000	1000	1000	1500
Minimum number of mapped reads (#)	200000	200000	50000	150000
Contains TCR reads detected by TraCeR?	Y	Y	Y	Y
Number of PCs used as input for tSNE	20	20	20	50
tSNE perplexity	30	30	35	35

II.2.2.3 Dimensionality reduction methods

To obtain an overview of the datasets showing the relationships between cell population clusters, Principal Component Analysis (PCA, Figure 25A,B) and tSNE were used. tSNE was performed using the Rtsne R package, with the function Rtsne that uses the Barnes-Hutt implementation of the algorithm. For each dataset, a different number of Principal Components (PCs) and values for perplexity were used (Table 7). Datasets were treated separately as much as possible to avoid confounding batch effects from experiments performed separately.

II.2.2.4 Cell cycle analysis

To assess potential effects of cell cycle in the interpretation of the scRNA-seq datasets collected, Cyclone [105] (implemented in the scran R package) was used on all datasets. Results were projected

on the tSNE (Supplementary Figure 1). As the vast majority of cells was assigned to the default stage (G0/G1 in mouse, S in human), no cell cycle correction was performed.

II.2.2.5 Differential expression analysis

Two linear models were fitted using the `lm` function from the base R package to the expression of each gene: a full model containing the information for each population under study (*e.g.* NLT vs non-NLT, Treg vs Tmem, etc.), and a reduced model only including an intercept term. These were then compared using a likelihood-ratio test (`lmtest` R package), and a q-value was calculated to correct for multiple testing. For all tests between tissues, genes with a $q\text{-value} \leq 0.01$ and a $\log_2(\text{fold-change}) \geq 1$ were considered differentially expressed. For tests between cell types within a tissue, a q-value threshold of 0.05 was used.

II.2.2.6 Differential co-expression analysis

Cell identity is defined not only by expression of specific genes, but also interactions between them. To assess which interactions define Treg and Tmem identity in NLTs or Treg identity between NLTs, the DGCA package [118] was used, which relies on the Fisher z-score transformation of the Spearman correlation values to perform a differential correlation test on gene pairs. For these tests, only genes identified as NLT markers and expressed in more than 5 cells in both conditions tested were used. Only differentially correlating genes where at least one population had a correlation coefficient greater than 0.25 (meaning co-expression of the gene pair in the tested condition) were kept.

II.2.2.7 Obtaining a migration latent variable for steady-state Tregs

The large dimensionality of single-cell RNA-seq data has been used before to gain insights on time-dependent events [79,80] by applying methods for pseudotime inference. Although it is impossible to follow one cell through the complete process, these methods can order single-cell data into a continuous dimension, using the discrete samples as snapshots containing a multitude of intermediate states.

Immune cells are expected to migrate via blood or lymph. I assumed that this effect would be reflected as a gradual single-cell expression phenotype, which could be captured as a latent variable of the data. To achieve this, I used Bayesian Gaussian Process Latent Variable Modelling (BGPLVM)

[119], implemented in the python package GPy (<https://github.com/SheffieldML/GPy>) as “GPy.models.BayesianGPLVM”, which was already used before for dimensionality reduction in scRNA-seq data to model Th1-Tfh differentiation [80]. BGPLVM was used asking for 6 latent variables, and the two most significant by Automatic Relevance Determination (ARD, Figure 32) were plotted. Spleen cells were excluded from this analysis to avoid making assumptions about the migration direction between the three tissues, and focus on general LN to NLT trafficking. This analysis was performed separately in Treg and Tmem, to avoid a cell type-specific confounding effect (Figure 31A,D, Figure 34). Indeed, while skin Tmems appear to have latent effects similar to those present in Treg (Figure 34F), the same can not be said for Tmems from the mouse colon dataset, where only a division by tissue can be observed (Figure 34). Gene correlations with colon LV1 and skin LV0 are correlated (Figure 32C), and a high proportion of those genes are associated with NLT identity (Figure 32D). The effect from these LVs is also present in each tissue individually, unlike what is observed for colon LV0 or skin LV1 (Figure 33). Thus, colon LV1 and skin LV0 were identified as representing tissue adaptation in each dataset.

II.2.2.8 Identifying a common tissue migration trajectory in control and melanoma

Similarly to the steady state, migration from the LN to the skin with a melanoma challenge is also expected. A common between-tissue Treg migration trajectory in control and melanoma conditions was obtained using Manifold Relevance Determination [120] (MRD). MRD works by having an underlying BGPLVM model whose dimensions can be shared or private between sections of the data. The importance of each section in each latent variable is shown in the ARD plot (Figure 39). Having the prior knowledge that a cell cycle effect is present in the data (Supplementary Figure 1) and with the goal of obtaining a latent variable explaining tissue recruitment in both conditions, the melanoma dataset was divided into three sections for input: one with the expression in all cell cycle associated genes, one with marker genes for any tissue, and one with the remaining genes. The model was run allowing for 12 latent variables as output, and the one highly influenced by tissue-specific genes but not cell cycle or other genes was used as a migration trajectory for both conditions. The effects captured by these latent variables can be observed in BGPLVM projections for the individual conditions (Figure 40A-D), and the latent variable with the largest tissue identity is also the most correlated to the LV used in the skin steady-state (Figure 40E). However, the cell cycle associated

latent variable also includes tissue-related effects, which can be explained by the expansion of Treg just before leaving the LN and upon arriving in the skin.

II.2.2.9 Switch-like genes in the migration latent variable

Gene expression changes in a continuous trajectory can be interpreted as a series of switch-like events. These can be modeled using a sigmoid curve, described by the following equation:

$$S = \frac{2 \times \mu_0}{1 + e^{-k(t-t_0)}}$$

where μ_0 is the mean expression between the sigmoid “on” and “off” states, t_0 is the point in which the switch in expression happens, and k defines the sigmoid inclination and can be interpreted as the activation strength. Parameter k will additionally inform on the direction of the switch (activation or inhibition) from its signal.

The R package `switchde` [87] was used to model gene expression as a sigmoid in the inferred migration trajectories, using the appropriate latent variable as pseudotime. In the steady-state datasets, `switchde` was applied for genes detected as Treg markers in either dataset (Treg vs Tmem differential expression regardless of tissue), tissue markers for each tissue in each dataset, or genes with an absolute correlation greater than 0.25 with the latent variable chosen. For the melanoma dataset, genes expressed in at least 5 cells in both conditions were tested. Only genes with a q -value ≤ 0.05 and that had a t_0 within the LV range were kept for further interpretation.

II.2.2.10 Detection of expanded clonotypes

T cell receptor (TCR) sequences were reconstructed from single cell RNA-seq data and used to infer clonality using the TraCeR software tool [68]. TraCeR was used with the parameters `- loci A B D G, - max_junc_len 120` to allow reconstruction of TCR α , TCR β , TCR δ and TCR γ chains in each cell and to permit TCR γ chains with long CDR3 regions.

II.2.2.11 GO Term enrichment

To test for enriched GO Biological Processes or KEGG Pathways in gene sets, the gprofiler R package [121] was used, with the option of moderate hierarchical filtering enabled.

II.2.2.12 Clustering analysis

To search for subpopulations in the NLT obtained Treg and Tmem cells, consensus clustering algorithm implemented in the SC3 R package [122] was used. I looked for 2 to 4 clusters to be able to exclude cases where small spurious subpopulations are identified. For each tissue/cell combination, I chose to compare the largest subpopulations detected that also were the farthest apart in tSNE space. Differential expression between clusters was performed on all expressed genes in them.

CHAPTER III

Single-cell RNA-seq resolves self-antigen
expression during mTEC development

CHAPTER III - Single-cell RNA-seq resolves self-antigen expression during mTEC development⁷

III.1 Abstract

The crucial ability of T cells for discrimination between self and non-self peptides is based on negative selection of developing thymocytes by medullary thymic epithelial cells (mTECs). mTECs purge autoreactive T cells by expression of cell type specific genes referred to as tissue-restricted antigens (TRAs). Here, scRNA-seq approaches were used to look into the mTEC cell population (PI-CD45⁻EpCAM⁺Ly51⁻UEA-1⁺), dissect their development and determine trends of TRA expression. Three subpopulations of cells were found computationally within the general mTEC population: jTEC, the recently described mTEC progenitors, mTEChi, characterized by strong *Aire* expression and considered the apex of mTEC function, and mTEClo, a later stage characterized by declining *Aire* expression. Gene expression differences between these groups were determined and suggest that interactions between mTEC and T cells might change during mTEC development. A gene regulatory network based on TF-binding sites and gene-to-gene correlation values is put forward, highlighting known and novel TFs and respective targets in mTEC development (IRF and NF- κ B families, Notch-related *Hes1* and *Zbtb7a*, *Vdr*, *Plagl1*, *Zbtb7a*, *Hnf4g*). TRA expression patterns were then compared between subpopulations. No major differences in coverage of TRAs was observed, suggesting that all mTEC subpopulations seem able to express them, despite doing so with less efficiency in early stages. It is also shown that *Aire*-dependent TRA expression in genomic clusters is only switched on during jTEC-mTEC transition. Finally, it was shown that after *Aire* expression peaks, *i.e.* in mTEClo, the number of TRAs expressed per cell still increases.

Taken together, this chapter characterizes mTEC development, specifically the mTEClo stage, supporting the “terminal differentiation model” according to which individual mTECs express more TRAs all the way through their development. Simultaneously, it also contradicts the widely spread notion that *Aire* downregulation impairs TRA expression.

⁷ Parts of this section, including figures, were adapted from [170].

III.2 Results

III.2.1 Single-cell mRNA sequencing of mTEC identifies global characteristics of TRA expression

To resolve the heterogeneity of the mTEC population, and to dissect the patterns of TRA expression, transcriptomes of murine mTECs (PI-CD45-EpCAM⁺Ly51⁻UEA-1⁺) at the single-cell level were sequenced and analyzed (Figure 7A, Figure 8). By using SMARTer chemistry (Clontech) on the C1 autoprep system (Fluidigm), cDNA libraries from 216 cells were obtained and sequenced, 164 of which met quality control criteria (Figure 7B, Figure 9A) and were kept for downstream analysis.

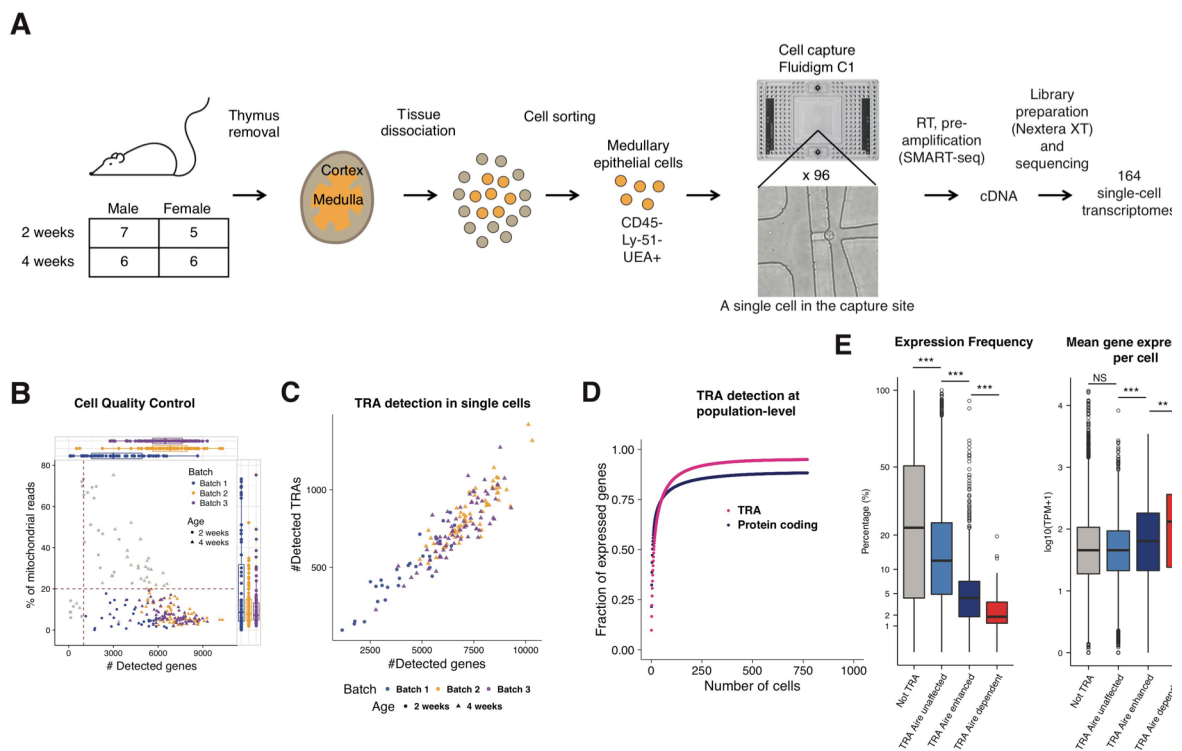


Figure 7 - Experimental workflow and expression of tissue-restricted antigens on population level.

- A)** Experimental workflow: mTECs in single-cell suspension were sorted to be run in Fluidigm C1 system.
- B)** Quality control of scRNA-seq for all 3 batches of cells processed with the Fluidigm C1 system based on number of genes detected, percentage of mitochondrial reads and number of mapped reads. Each batch corresponds to a different colour, and the age of the mice used match the shapes.
- C)** Number of expressed genes vs number of expressed TRAs in each cell.
- D)** Number of expressed genes as a function of the number of mTECs considered. Each point was calculated based on the average of 100 random orders of the 692 cells of all datasets analysed.
- E)** Comparing genes from different categories in terms of expression frequency and mean expression level across all cells.

*** p-value<0.001, ** p-value<0.01, * p-value<0.05, NS – not significant, according to Mann-Whitney-Wilcoxon test, p-value adjusted using Bonferroni correction.

Number of mapped reads, number of detected genes and percentage of reads mapped to mitochondrial genes were considered as quality control variables. All of which constitute proxies for the technical quality, *e.g.* lower concentration of cDNA in the final pool of cells, and biological quality, *e.g.* cell death during the harsh process to obtain single-cell suspensions, of the cell being analyzed. Batches of cells coming from different experiments and/or C1 machines were compared by differential expression analysis, and did not show gene expression biases (Figure 9B). TRA genes, as defined by Sansom *et al.* [107] and excluding genes coding for MHC I proteins, totalled 6611 genes expressed in both datasets. On average, TRAs accounted for approximately 10% of all genes expressed in a single cell (Figure 7C). In line with previous reports [28,107,123], the repertoires of TRAs expressed in single mTECs did not exhibit significant enrichment for any particular peripheral tissues (Figure 10).

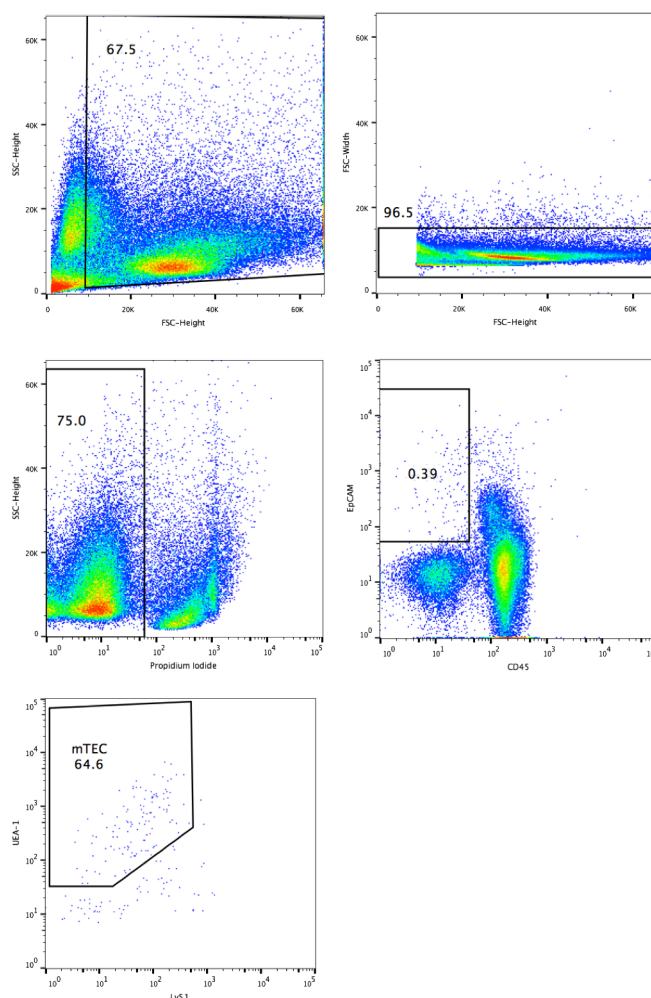


Figure 8 - Flow cytometry cell sorting strategy for isolation of mTECs.

mTECs were sorted as $PI^-CD45^-EpCAM^+Ly51^-UEA-1^+$ using a MoFlo™ XDP (Beckman Coulter, Inc.).

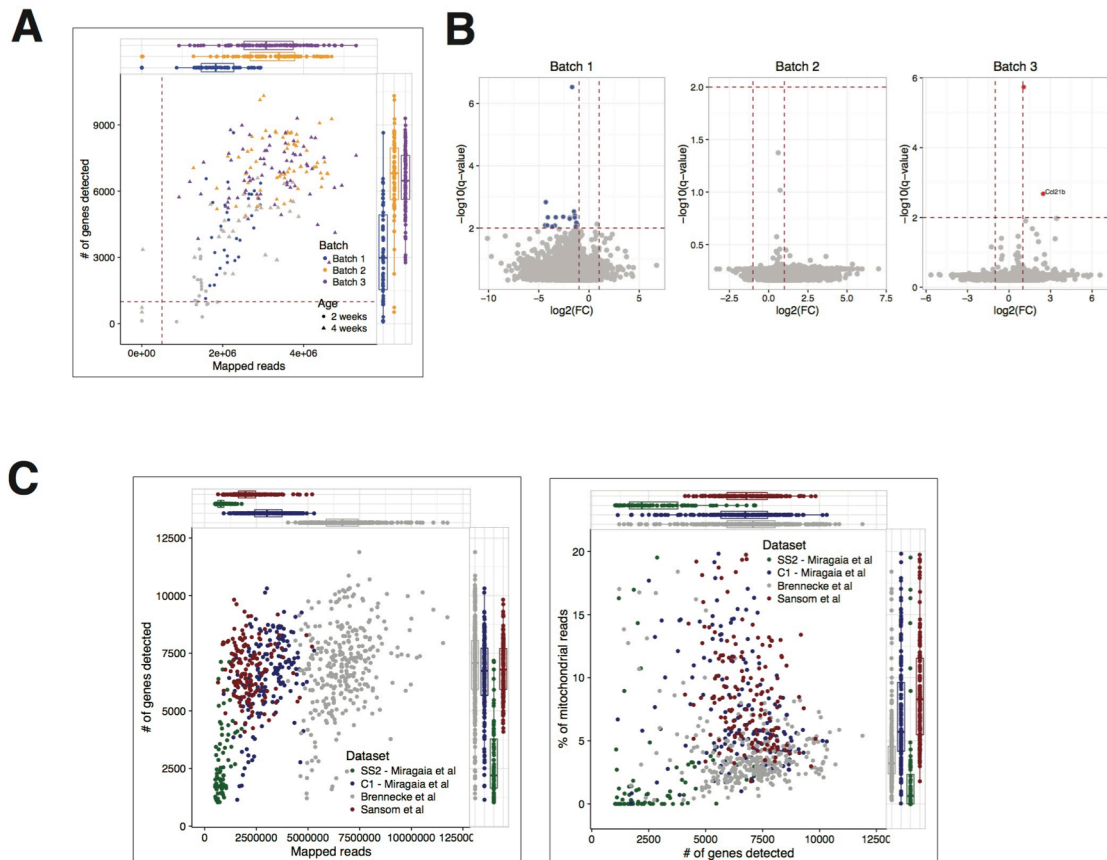


Figure 9 – mTEC scRNA-seq quality control.

A) Number of mapped reads and genes detected in the C1 dataset. Cells with less than 1000 genes, less than 500,000 mapped reads or more than 20% of mitochondrial reads were excluded (gray symbols).

B) Differential expression between different batches within the C1 dataset.

C) QC metrics for all datasets used. Same thresholds were used across all datasets.

Integrating publicly available datasets with my data, it is apparent that the majority of protein coding genes, including TRAs, are covered in a couple of hundreds of mTECs (Figure 7D), also supporting previous observations [8,28]. Further interactions of thymocytes with additional mTECs would therefore result in minimal increase in the variety of TRAs they are exposed to.

To achieve greater resolution, TRAs were then divided into subsets of genes, of which expression is either completely dependent on *Aire* (*Aire*-dependent), enhanced by it (*Aire*-enhanced), or unchanged in its absence (*Aire*-unaffected) according to previous data in *Aire*-deficient mice [107]. Analyzing the expression frequency and expression level separately for the genes of each of these subsets evidences *Aire*-dependent genes as a notably distinct group, with significantly lower expression frequency and higher mean expression level than genes in the other subsets (Figure 7E), as previously reported [107,124,123]. Despite the differences between subsets of *Aire* genes, it is worth noticing that all of them were expressed on average at equal or higher levels than all other genes in mTECs. This

indicates that *Aire* genes of all subsets, especially *Aire*-dependent, are actively expressed and not merely products of passive and/or residual level expression. The behaviour of TRAs controlled by the recently discovered *Fezf2* regulator [26] was then evaluated (Figure 11). In contrast to *Aire*-induced TRAs, *Fezf2*-induced TRAs were expressed as frequently as normal genes (“non-TRA”), although their expression level was higher, similarly to *Aire* TRAs (Figure 11). Such differences are likely to stem from the different mechanisms of gene activation by these two transcription factors.

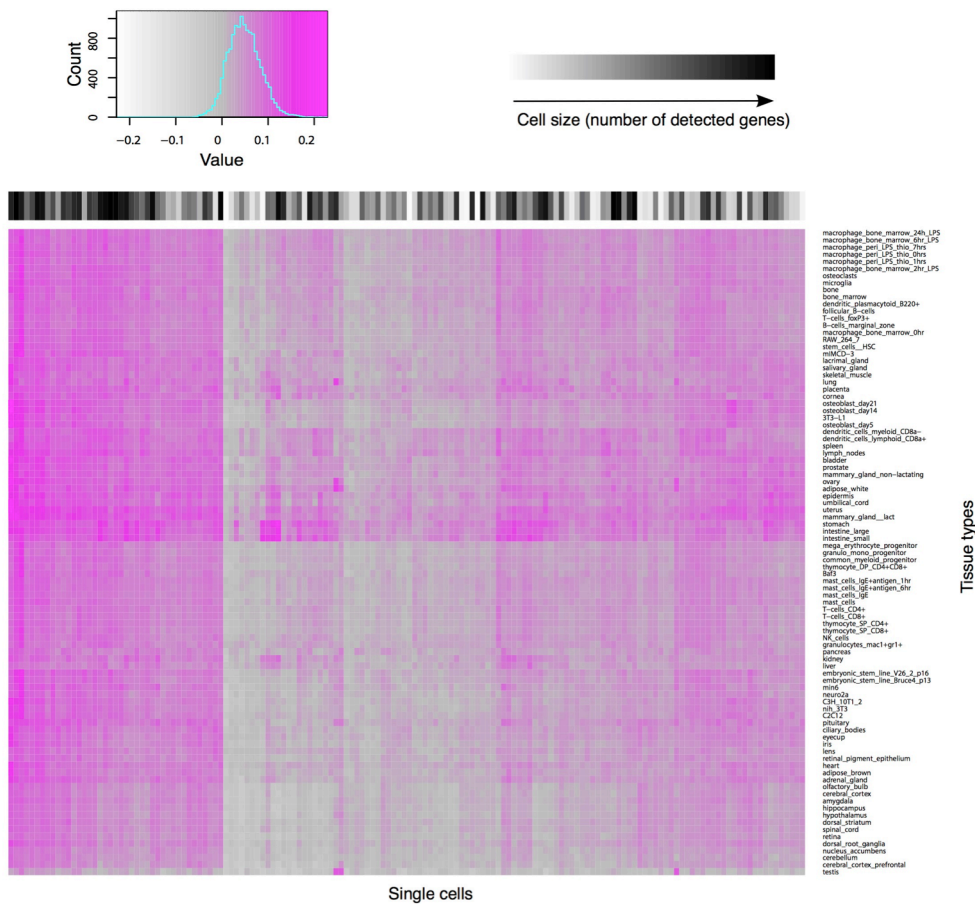


Figure 10 - Correlation between individual mTECs and tissues.

Correlations between individual mTECs and microarray data from multiple tissues⁵⁶ were measured and clustered. All expressed genes were used in the calculations. Colour bar denotes cell size, as estimated by number of detected genes.

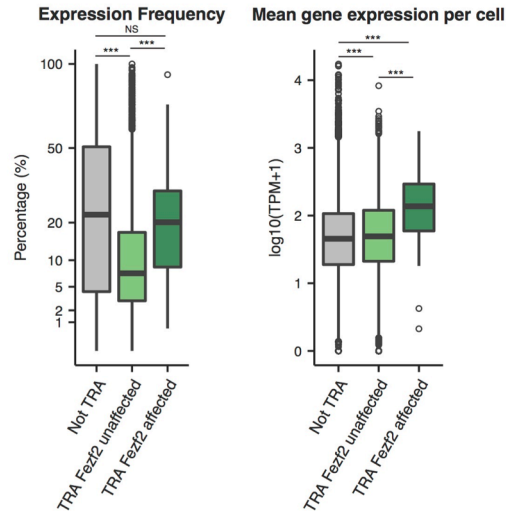


Figure 11 - Expression of Fezf2-regulated TRAs at the single-cell level.

Comparing TRAs from different Fezf2 categories in terms of expression frequency and mean expression level across all cells.

*** p-value<0.001, ** p-value<0.01, * p-value<0.05, NS – not significant, according to Mann-Whitney-Wilcoxon test, p-value adjusted using Bonferroni correction.

III.2.2 scRNA-seq resolves three major subpopulations along mTEC differentiation

Principal component analysis (PCA) was performed to explore the subpopulation structure within mTECs. It was noticed that a great source of variability came from cell size (number of detected genes), which correlated strongly with the most important PC1 (Spearman rho 0.92, Figure 12A). The analysis was thus focused on the next two PCs, markedly less affected by this variable (Figure 12A-C). Importantly, cells that were isolated from different mice and processed on different C1 integrated fluidic circuits were dispersed among each other, suggesting that batch effects did not contribute significantly to the overall heterogeneity (Figure 12B). In addition to cell size and batch, single-cell RNA-seq data can be profoundly affected with variation associated with cell cycle. To assess how much the data was biased by cell cycle, the Cyclone package was used to assign single cells into cell cycle phases [105] (Figure 12D). All but six mTECs were in G1/G0 phase, suggesting relatively modest cell cycle effects over gene expression in this population.

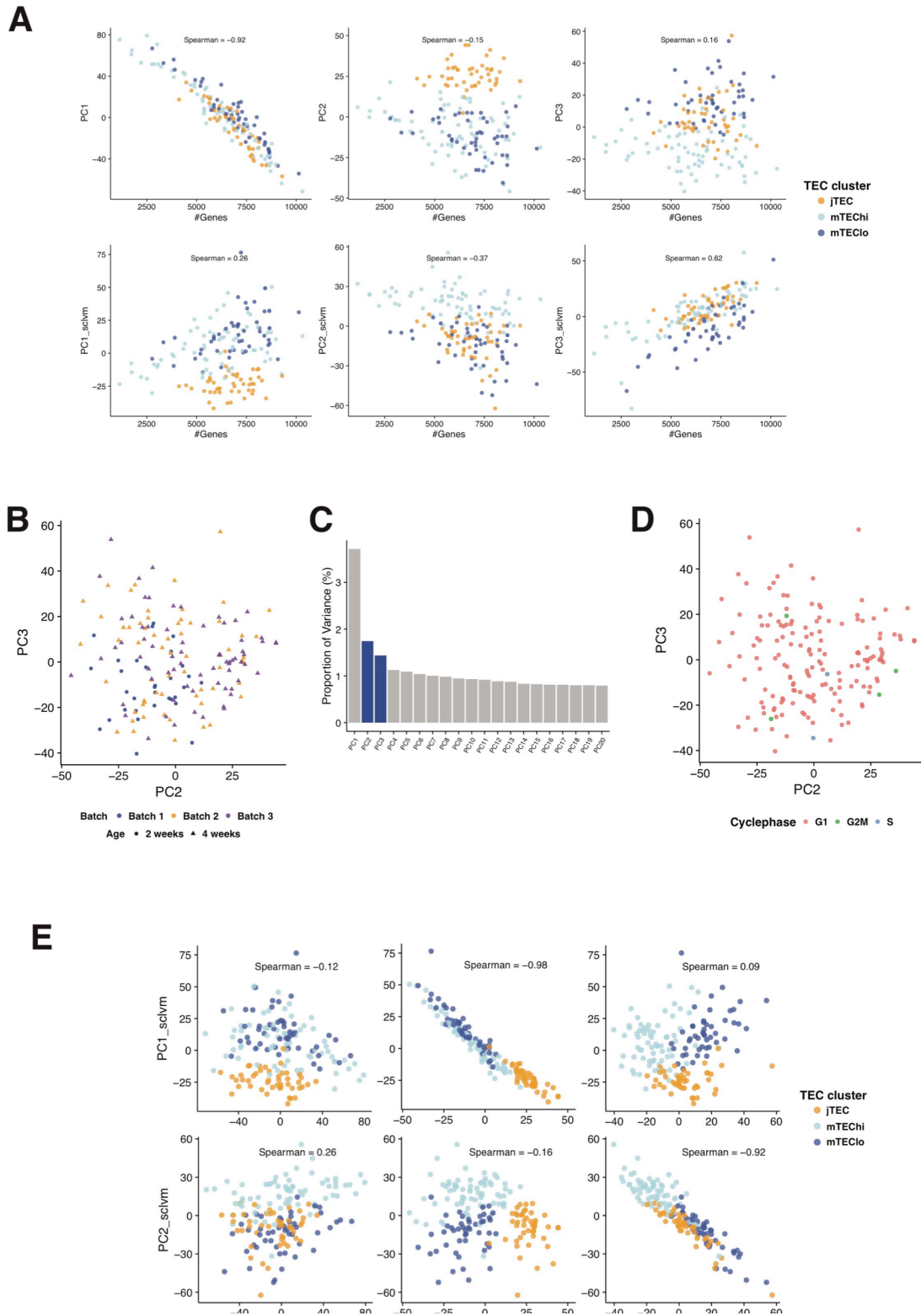


Figure 12 - Principal Component Analysis of C1 dataset.

- A) Correlation between the first three PCs and the number of detected genes per cell, either in uncorrected (top) or sLVM corrected data (bottom).
- B) Distribution of the three C1 batches along Principal Component 2 (PC2) and PC3.
- C) Proportion of variance (%) for the first 20 PCs. PC2 and PC3 seem to contain most biological variance.
- D) Cell cycle classification of single-cells (G1/G0, G2M, S) as determined by the Cyclone package.
- E) Correlation between PC1 and PC2 in sLVM corrected data and PC1, PC2, PC3 in the uncorrected data.

Nonetheless, this possibility was investigated further by running scLVM package to evaluate and regress out any cell cycle related biases. Performing PCA on the scLVM-corrected data, it was observed that PC1, PC2 and PC3 all correlated with the cell size to some extent (Figure 12A, bottom), in contrast with the uncorrected data where cell size appears limited to PC1 (Figure 12A, top). Simultaneously, scLVM-corrected PC1 was highly correlated with the original PC2 and scLVM-corrected PC2 with the original PC3 (Figure 12E). Therefore, on this particular case in which cell cycle effect is minimal, technical effects could be more easily deconvoluted from biological variability simply by focusing on higher components on the uncorrected dataset. For the publicly available datasets in which cell cycle seemed to have a stronger effect, scLVM correction was adopted.

For the main dataset, PC2 scores of single cells were correlated with the expression of several established markers of thymic development, such as *Cldn3&4*, *Pdpm* and *Cd80* [32,35,125] (Figure 13A). On the other hand, among the top PC3 loadings were *Aire*, *Cd40*, *Icosl* and other genes associated with the mature mTEC phenotype. This further suggested that the variability associated with mTEC development was primarily distributed along these two components. Using the expression data from top PC2 and PC3 genes (loadings above 0.02 and below -0.02), a consensus clustering approach was used to assign single cells to three distinct sub-populations (Figure 13B and Methods). Only approximately 20% of these top genes are TRAs, which suggests that TRA expression patterns do not explain the three sub-populations found. Cell size also does not seem to influence this classification of mTECs (Figure 14A).

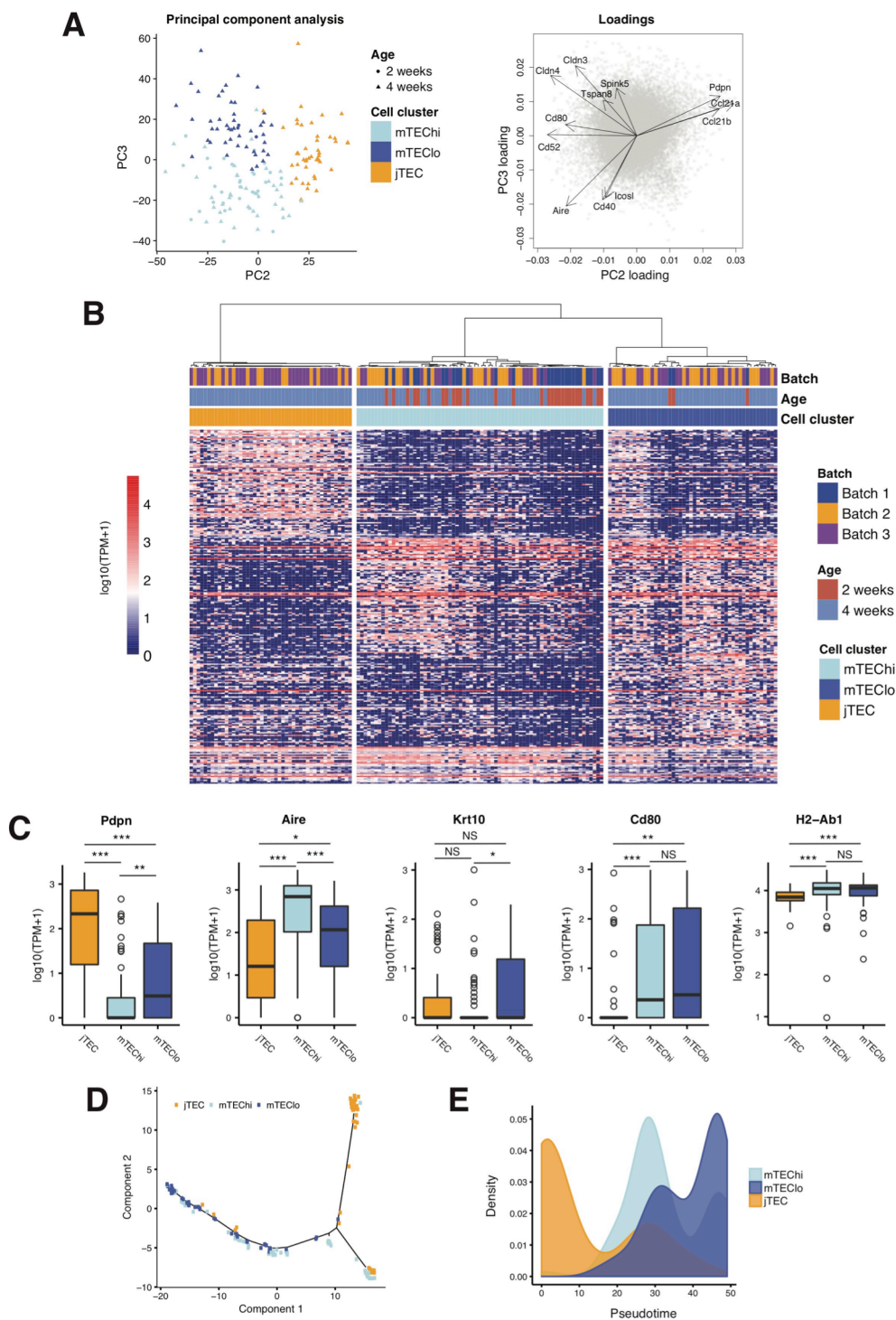


Figure 13 - Analyses of mTEC subpopulation structure.

A) Principal Component Analysis (PCA) of mTECs using all genes. Batches match colour and age of mice matches shapes (left). The PC2 and PC3 loadings of key genes of interest are highlighted (right).

B) Hierarchical clustering of 164 transcriptomes of single-cells, based on the top positively and inversely correlated genes with PC2 and PC3. Three clusters of cells were identified (jTEC, mTEChi and mTEClo). Cell cluster, mice age and batch are depicted.

C) Expression of selected marker genes in the jTEC, mTEChi, and mTEClo populations. *** p-value < 0.001, ** p-value < 0.01, * p-value < 0.05, NS – not significant, according to Mann-Whitney-Wilcoxon test, p-value adjusted using Bonferroni correction.

- D) Ordering of single cells in pseudotime by degree of maturation using the Monocle2 algorithm. Data points represent single cells, colours denoting cell clusters (B).
 E) Distribution of single cells of each cluster (B) along pseudotime.

An initial inspection of established marker genes painted a general picture of the identity of these subpopulations: the first of these clusters was characterized by the expression of *Pdpr*, a marker of a recently identified population of junctional TEC precursors (jTEC), which gives rise to fully differentiated mTEC [32]; the different levels of *Aire* expression between the second and third clusters led me to classify them as *Aire*-high mTEC (mTEChi) and *Aire*-low mTEC (mTEClo). The cells of the mTEClo population, although resembling a post-*Aire* state for their lower expression of *Aire* and higher expression of Keratin 10 (*Krt10*) [33,34], expressed similar levels of *Cd80* and HLA (class II) genes as the mTEChi cells (Figure 13C). Statistically significant differences for other proposed markers of mTEC stages (i.e., *Gp2*, *Gad1*, *Ceacam1*, *Tspan8*) [28,36] were not observed, although *Gad1* and *Tspan8* tend to characterize mTEChi and mTEClo, respectively (Figure 14B), which is in line with previous reports [36].

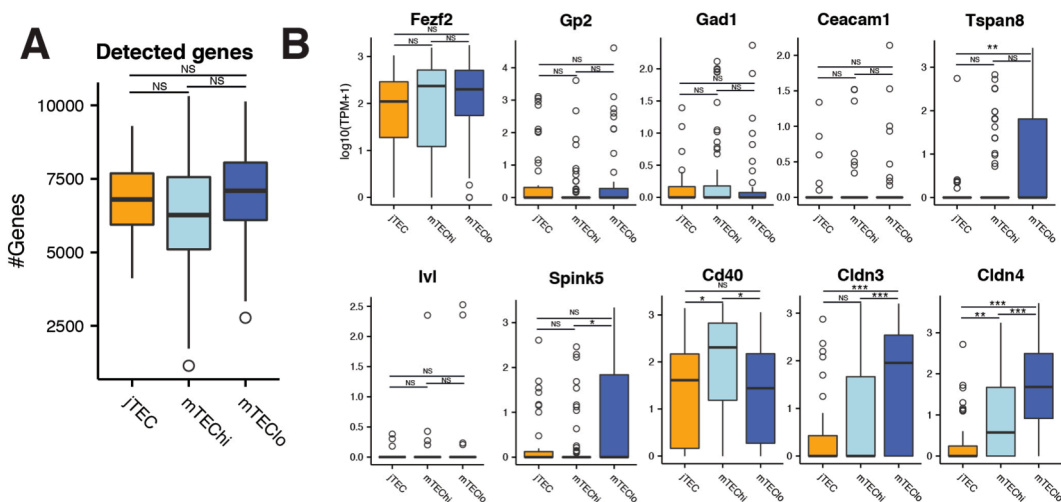


Figure 14 - Subpopulation comparisons for in-house C1 dataset.

A) Number of genes detected per subpopulation C1 dataset.

B) Expression of selected marker genes in the jTEC, mTEChi, and mTEClo populations.

*** p-value<0.001, ** p-value<0.01, * p-value<0.05, NS – not significant, according to Mann-Whitney-Wilcoxon test, p-value adjusted using Bonferroni correction.

To confirm the robustness of these findings, additional cells using the Smartseq2 protocol were sequenced [126]. In addition, two recently published single-cell datasets [28,107], were included and processed using the same pipeline and QC parameters used for the original dataset (Figure 9C).

Importantly, cell clustering and expression of marker genes remained largely consistent across these datasets (Figure 15 and Figure 16).

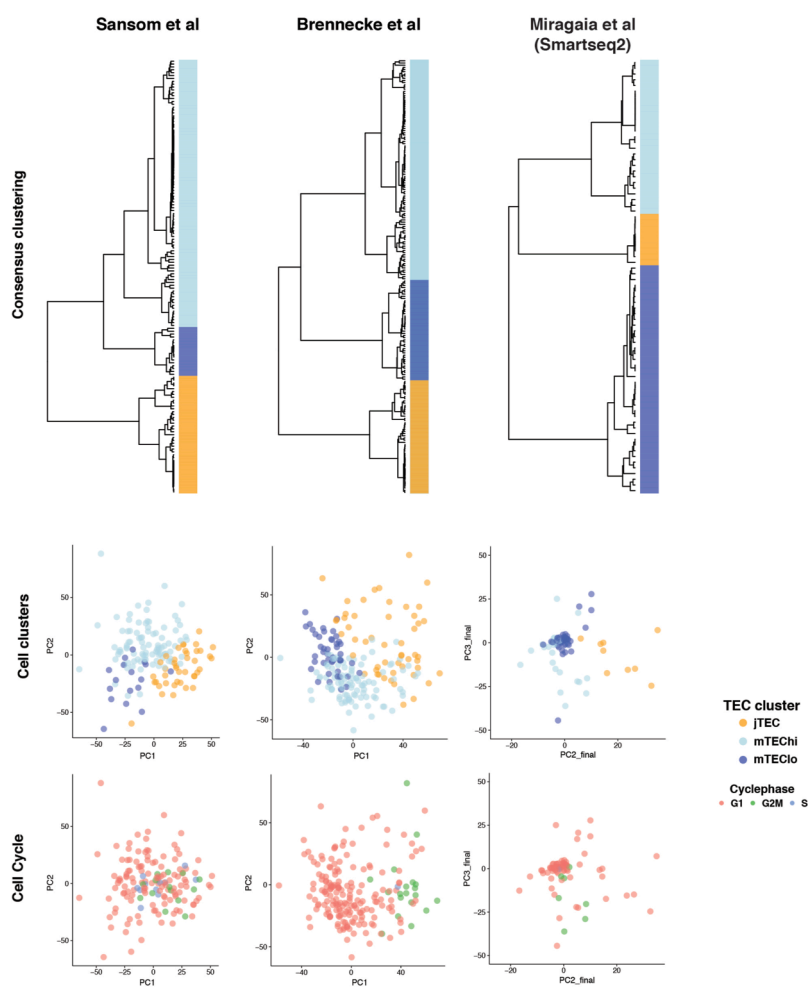


Figure 15 - PCA and consensus clustering on publicly available and in-house Smartseq2 mTEC datasets.

Consensus clustering (top) and PCA of Sansom *et al*, Brennecke *et al* and in-house Smartseq2 datasets (middle), coloured by cell clusters as determined by hierarchical clustering. Sansom *et al* and Brennecke *et al* datasets were subject to sLVM correction (see text). PCA coloured by cell cycle, as determined by the Cyclone package (bottom).

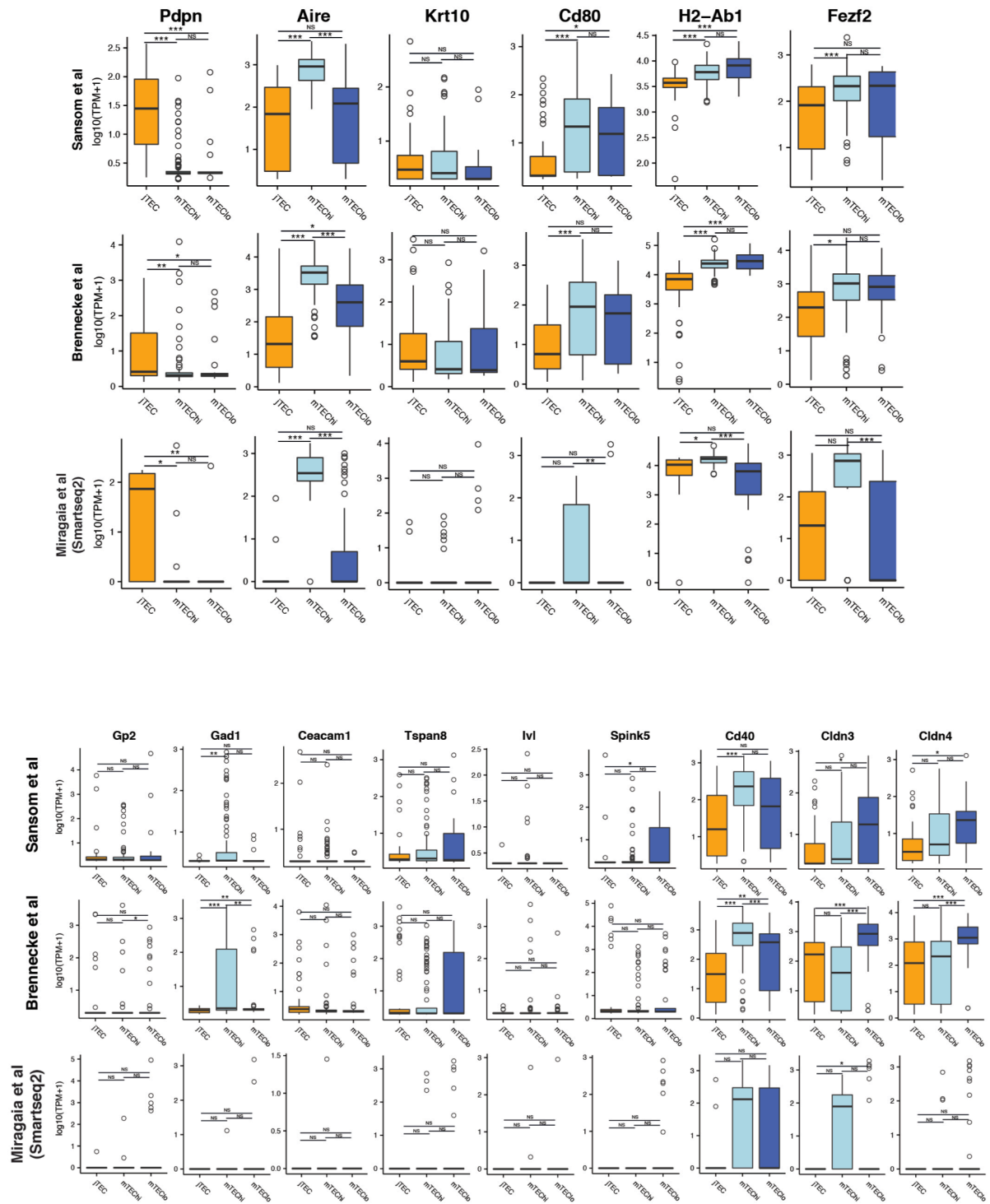


Figure 16 - Differential expression of genes across mTEC subpopulations in publicly available and in-house Smartseq2 mTEC datasets.

*** p-value<0.001, ** p-value<0.01, * p-value<0.05, NS – not significant, according to Mann-Whitney-Wilcoxon test, p-value adjusted using Bonferroni correction.

To further investigate this differentiation process, Monocle2 [79] was used to infer pseudotime, representing a measure of cells progression along the differentiation trajectory. jTEC positioned towards the beginning of the pseudotime, followed by mTEChi and ending with mTEClo cells. mTEChi and mTEClo overlapped to some extent, suggesting a close relationship between them (Figure 13D and E). In contrast, the clear gap between jTEC and mTEC indicated more profound differences between these subpopulations. Notably, no cells from the 2-week old mice fell into the jTEC population, potentially resulting from age-associated changes in frequencies of thymic subsets [127]. Then, differential expression analysis was performed to systematically identify genes that are specific to each subpopulation (Figure 17A). The jTEC state was associated with upregulation of 383 genes and downregulation of 63 genes ($q\text{-value} < 0.05$, $|\log_2(\text{FC})| > 1$). In the mTEChi cells, 50 genes were upregulated (including *Aire*) and 90 were downregulated. The mTEClo population was characterised by 109 upregulated and 81 downregulated genes. It is worth mentioning that of the genes differentially expressed between the mTEC subpopulations, only 45 encoded for TRAs. Notably, almost all of these (43) were *Aire*-unaffected TRAs. The notable absence of *Aire*-regulated TRAs among these genes is probably explained by their relatively lower expression frequency. Furthermore, some of the jTEC-specific TRAs (such as *Adm*, *Cdh3*, *Krt14* and *Krt17*) are associated with epithelial development, and are thus likely to be required for a specific functional role despite being considered TRAs. Several of the markers now identified seem to be related to particular states of maturation, interactions with the surroundings and/or specific functions of each subpopulation. For example, *Jag1* upregulation by mTEChi cells, together with *Notch2*, *Hes1* and *Hes6* upregulation in jTECs suggest that Notch signalling might be involved in the jTEC to mTEC transition. Based on the expression of genes such as *Jag1*, *Cd40* and *Icosl* in mTEChi³³⁻³⁵ [128–130], Skint-family genes (*Skint7* and *Skint9*) [131], galectins and related genes (*Lgals1*, *Lgals9*, *Lgals3bp*) in jTEC and mTEClo [132,133], the way each subpopulation instructs thymocytes is likely to be slightly different, e.g. role of ICOSL in the expansion of regulatory T cells (Treg) in humans [130]. From a practical point of view, membrane proteins in these sets of marker genes can potentially be used in the future to sort out each subpopulation for further studies (e.g. *Lypd8* to distinguish mTEChi and mTEClo).

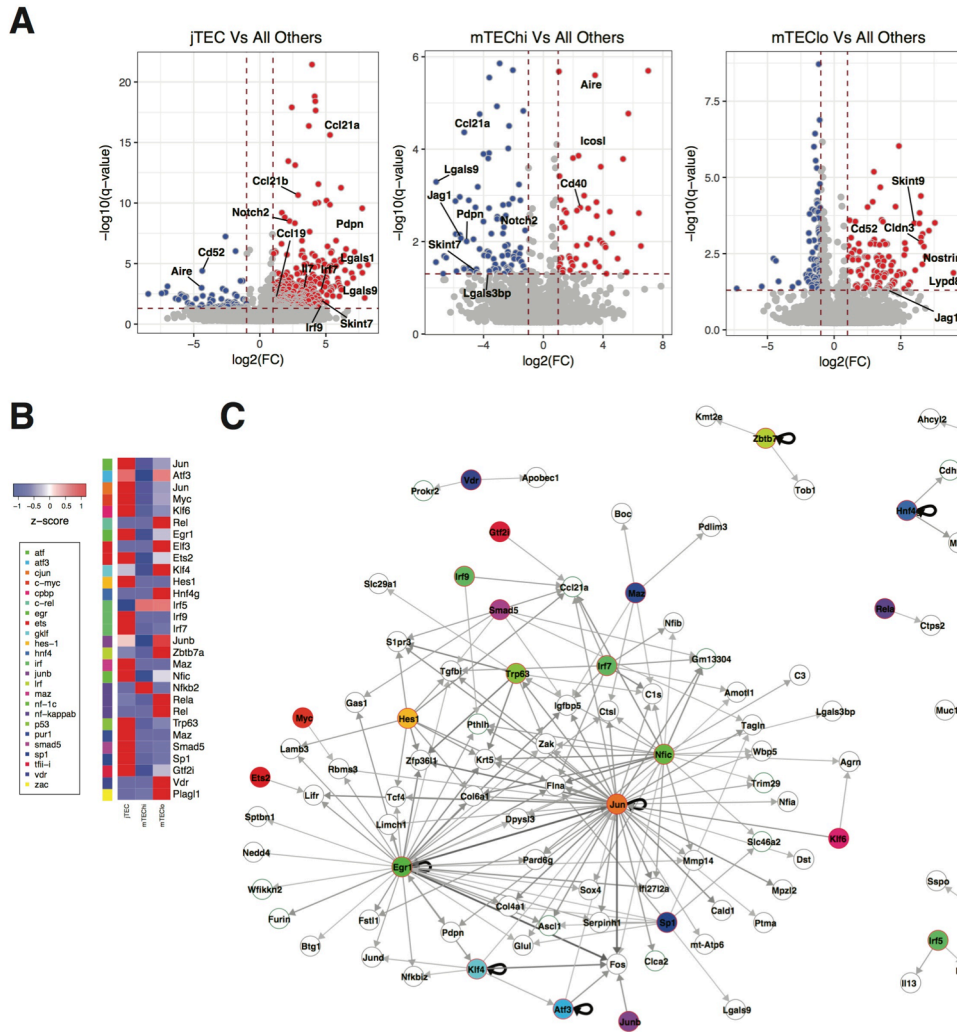


Figure 17 - Differential expression of genes across mTEC subpopulations.

A) Genes differentially expressed (DE) between each cell population and the remaining populations. The significance of this DE was calculated using a linear model. $q\text{-value} < 0.01$ and $|FC| > 1$.

B) Scaled median expression levels of transcription factors with enriched binding motifs (C) across jTEC, mTEChi and mTEClo subpopulations. The TFs were grouped according to TF families, as denoted by the color bar.

C) Network visualisation of TFs and their putative target genes. Target genes were identified by binding motif analysis using gProfiler. The results were further filtered based on co-expression, based on Spearman correlation and Jaccard index. Colours of the nodes denote TF families, as shown in (B). Thicker and darker edges represent higher Spearman correlation.

III.2.3 Binding motif and coexpression analysis highlights potential drivers of mTEC maturation

Hierarchical clustering and pseudotime reconstruction revealed three distinct stages during mTECs lifetime. Considering the importance of mTEC development for central tolerance, it would be of great relevance to pinpoint transcription factors (TFs) involved in this process, as well as their respective

target genes. To address this question, the top PC2 and PC3 genes, which were also used for cell clustering, were considered. Scanning the genomic regions upstream of these genes (1 kb), I found an enrichment of binding-sites for several TFs, most of which were preferentially expressed in one of the three mTEC subpopulations (Figure 17B). Top PC2 and PC3 genes possessing a binding-site for a given TF were considered as being potentially regulated by it. The gene-to-gene correlation between TF and respective target was then calculated for all TF-target pairs. The most significant TF-target relationships were filtered using stringent correlation and Jaccard index threshold values. These relationships were thus considered to potentially indicate direct regulation of the target by the TF, and were visualised as a co-expression network (Figure 17C). In this network, the most prominent TF hubs were *Egr1* and *Jun*. Although both are characteristic of jTEC, the program they set in motion is likely to also span mTEChi and mTEClo (e.g. *Klf4*). *Ccl21*, a key marker of the jTEC subpopulation, was identified as a putative target of seven different TFs. Notably, *Nfkb* and *Irf* transcription factor families were well represented across all three subpopulations: jTEC expressing *Irf7* and *Irf9*, mTEChi expressing *Irf5* and *Nfkb2*, and mTEClo expressing *Irf5*, *Rel* and *Rela*. Both classical and non-classical NF- κ B signaling (through TRAF6 and NIK, respectively) have been proven necessary for the development of Aire-positive mTECs [134]. More recently, the *Irf* family has also been implicated in the development of mTECs [10] and shown to contribute to TRA expression along with AIRE [22]. *Hes1* expression by jTEC and *Zbtb7a* [135] by mTEClo suggest the involvement of the Notch pathway in this progression. Finally, *Vdr*, *Plagl1*, *Zbtb7a*, *Hnf4g*, most of which have previously been detected in mTEC [36,136,137], assume particular relevance as presumptive drivers of the late stages of mTEC differentiation.

III.2.4 TRA expression during mTEC development

Having identified three subpopulations of mTECs spanning different stages of maturation, investigating TRA expression patterns across them was naturally of great interest. Several models have been put forward in efforts to explain how TRA expression is regulated during mTECs lifetime to guarantee a comprehensive negative selection of self-reactive thymocytes. Do mTECs progressively express a higher number of TRAs as they differentiate (“terminal differentiation model”)? Do they begin with the capacity to express significant numbers of TRAs, and then progressively (and independently of other mTECs) limit the range of TRAs expressed (“progressive restriction”) [138]? Are certain sets of TRAs co-expressed, or in a predefined sequence [28]?

With these questions in mind, the extent to which mTEC populations differ in their TRA expression coverage was addressed (Figure 18A). Very few genes (in any of the TRA-subsets) were uniquely expressed by jTECs, in line with the notion that they are the most immature population. Surprisingly, jTECs covered as many as 84% of *Aire*-unaffected TRAs and the majority (66%) of *Aire*-enhanced TRAs. Nonetheless, the percentage of *Aire*-unaffected TRAs and *Aire*-enhanced TRAs shared exclusively between mTEChi and mTEClo (11% and 23%, respectively) was larger than between either population and jTECs. *Aire*-dependent TRAs were expressed by jTEC to a lesser extent (46%) than the other TRAs, and once again, mTEChi and mTEClo shared 37% exclusively between them and expressed a higher percentage of unique genes (14% for mTEChi and 8% for mTEClo) (Figure 18A). Together, these observations indicate that the TRA repertoire is largely equivalent between the three maturation stages, except for some compartmentalization of *Aire*-dependent TRAs.

Next, the performance of individual cells within these groups was assessed. To measure their competence in driving the negative selection process, the level of expression and the number of TRAs expressed by individual cells was measured in jTEC, mTEChi and mTEClo. As expected, mTEChi expressed significantly elevated levels of *Aire*-enhanced and *Aire*-dependent TRAs (Figure 19A). Then, the number of TRAs expressed on a cell-by-cell basis was assessed, normalized by the number of detected genes, thus accounting for differences in sequencing efficiency of single cells (Figure 18B). jTEC stood out as the least competent subpopulation for each of the TRA subgroups. Surprisingly, mTEClo were at the other end of the spectrum, expressing the highest number of TRAs per cell. Specifically, mTEClo expressed the highest number of *Aire*-unaffected TRAs, and a similar number of *Aire*-enhanced and even *Aire*-dependent TRAs compared to mTEChi (Figure 18B) an equivalent number of *Aire*-enhanced TRAs compared to mTEChi. mTEChi seem to perform slightly better with respect to *Aire*-dependent TRAs, although the advantage over mTEClo is not significant.

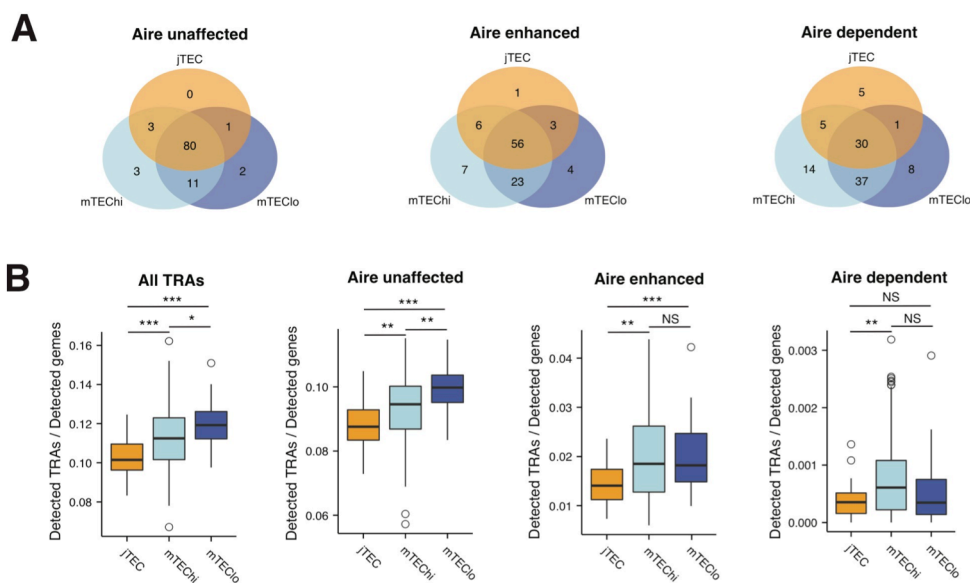


Figure 18 - Expression of TRAs across mTEC subpopulations.

A) The detection of TRA genes across jTEC, mTEChi, and mTEClo populations. The numbers indicate the fraction of TRAs detected in the respective subset of cells.

B) Number of TRA genes expressed in the jTEC, mTEChi, and mTEClo populations. To account for differences in library sizes, the number of detected TRA genes were normalised to the number of detected genes per cell.

*** p-value < 0.001, ** p-value < 0.01, * p-value < 0.05, NS – not significant, according to Mann-Whitney-Wilcoxon test, p-value adjusted using Bonferroni correction.

Genes controlled by *Fezf2*, a master regulator of TRAs besides *Aire* [26], were shown to increase consistently in number and level of TRA expression along differentiation (Figure 19B).

In summary, all three subpopulations of mTECs collectively expressed most of the TRA genes. However, they did differ in their expression efficiency in terms of level and number of TRAs expressed, with mTEChi holding an advantage for *Aire*-regulated TRAs, as expected. Nevertheless, mTEClo, a state that has been classically regarded as a passive step towards mTEC death, showed extreme competence at expressing *Aire*-enhanced and *Aire*-unaffected TRAs. These observations are largely confirmed in other datasets (Figure 20). They are also in line with previous reports [33,34] and indicate that mTEClo are active players in negative selection, with TRA expression increasing during the entire mTEC lifetime and remaining high after *Aire* expression declines.

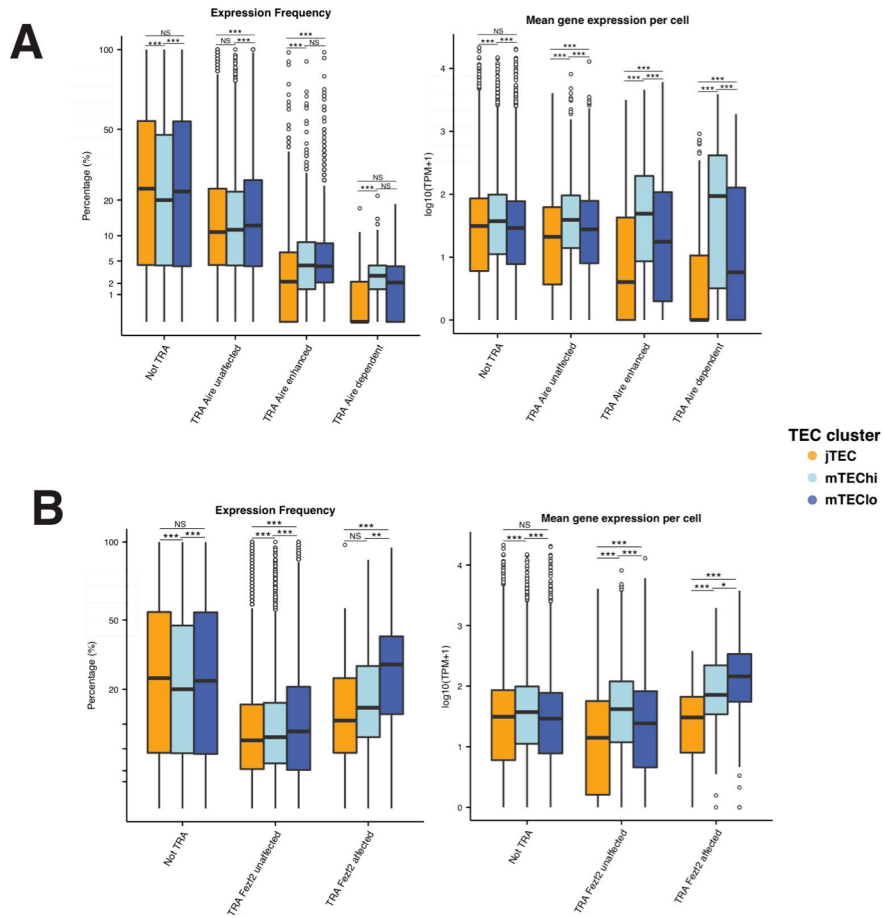


Figure 19 - TRA expression sectioned by *Aire* and *Fezf2* dependency and mTEC subpopulations in C1 dataset.

A) Comparing TRAs from different Aire categories in terms of expression frequency and mean expression level across all cells.

B) Comparing TRAs from different Fezf2 categories in terms of expression frequency and mean expression level across all cells.

*** p-value<0.001, ** p-value<0.01, * p-value<0.05, NS – not significant, according to Mann-Whitney-Wilcoxon test, p-value adjusted using Bonferroni correction.

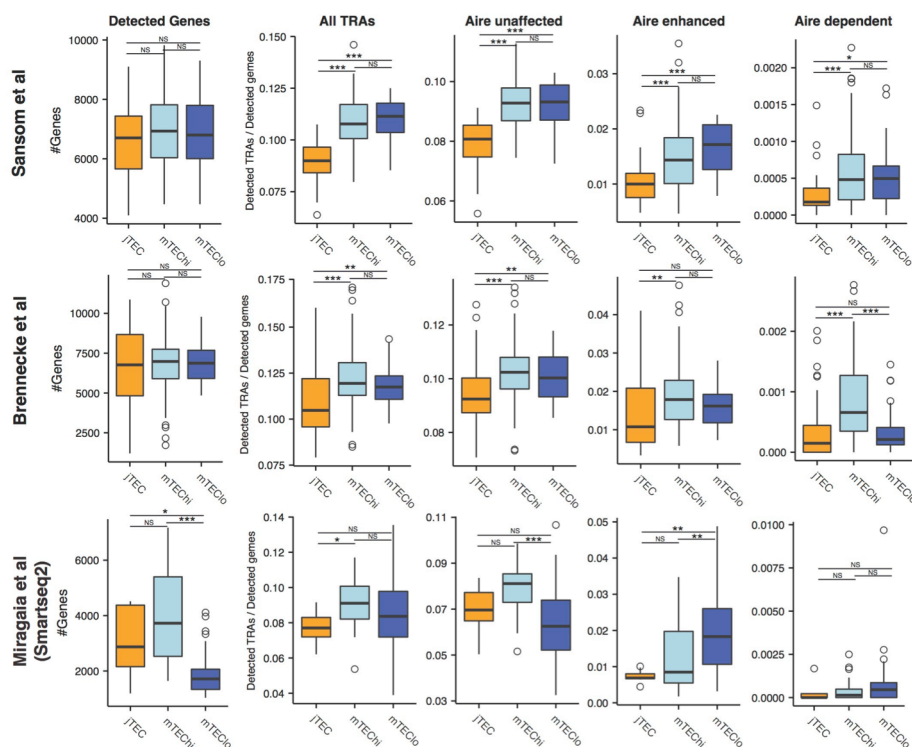


Figure 20 - Characterization of publicly available and in-house Smartseq2 mTEC datasets.

Number of genes detected per mTEC subpopulation in each dataset and number of TRA genes expressed in the jTEC, mTEChi, and mTECdo populations, in each dataset. To account for differences in library sizes, the number of detected TRA genes were normalised to the number of detected genes per cell.

*** p-value<0.001, ** p-value<0.01, * p-value<0.05, NS – not significant, according to Mann-Whitney-Wilcoxon test, p-value adjusted using Bonferroni correction.

III.2.5 Expression of Aire-dependent TRAs is associated with strong genomic enrichment at single-cell level and is induced during jTEC-mTEC transition

About a decade ago, *Aire*-regulated genes were shown to be in close linear chromosomal proximity to each other, forming genomic microclusters [14,139]. Since then, it became accepted that AIRE's ability to recruit transcription factors to regions of closed chromatin would induce remodelling of such segments, thus facilitating the co-expression of neighbouring genes [140]. The discovery that AIRE binds super-enhancers [24] supports this idea, providing a model that explains both intra- and interchromosomal coexpression patterns of *Aire*-regulated genes [123]. Although pivotal to understand the mechanisms behind TRA expression, genomic clustering of expressed TRAs was never investigated at the single-cell level. Single-cell qPCR of mTEC has been reported [24], but it covered a very restricted number of TRAs and genomic clustering was not addressed.

In this section, several aspects of TRA genomic clustering are addressed at the single-cell level, namely how does clustering evolve along mTEC development, how differently are *Aire*-enhanced and

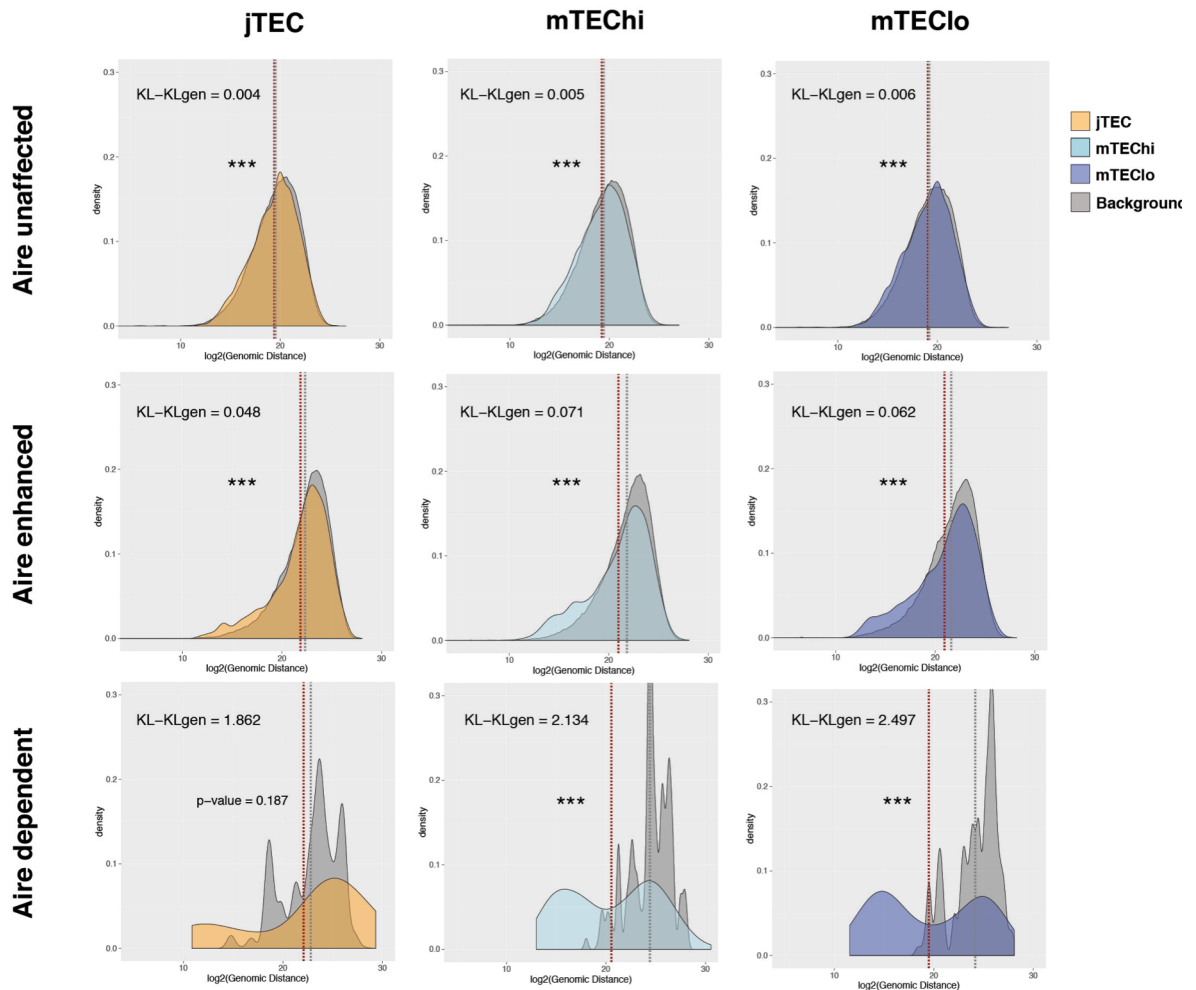


Figure 21 – Genomic distribution of expressed TRAs across mTEC subpopulations.

Histograms represent distribution of distance to the nearest neighbour gene for TRA genes. The background histograms represent the distribution of distances to the nearest neighbour of randomly sampled expressed genes from a control distribution (Methods) in the corresponding cell.

*** p-value<0.001, ** p-value<0.01, * p-value<0.05, according to Mann-Whitney-Wilcoxon test

Aire-dependent TRAs affected, and whether these principles can be generalized to all TRAs.

First, the tendency of expressed TRAs to sit closer together along the genome was tested at the single-cell level. Sub-groups of TRAs based on *Aire*-regulation were considered separately, as well as the three developmental stages identified, jTEC, mTEChi and mTEClo. For each expressed TRA, the base-pair distance to the nearest expressed TRA was calculated on a cell-by-cell basis. The resulting distribution of distances was then compared to a control distribution built to account for the genomic location of TRA genes, as well as the generic clustering effect reported for any set of expressed genes [14] (see Material and Methods).

The mean values for the actual and the background distances (p -value from Mann-Whitney-Wilcoxon test) were compared, along with the magnitude of the divergence between these distributions (Kullback-Leibler divergence (KL)) (Figure 18C). Although statistically significant, it is clear that genomic clustering of *Aire*-unaffected TRAs is extremely weak for all three subpopulations (KLs \leq 0.006). For *Aire*-enhanced TRAs, the clustering effect does increase (KLs \leq 0.071), remaining quite modest nonetheless particularly for jTECs (KL = 0.048). *Aire*-dependent TRAs, on the other hand, exhibit a distribution of distances indicating strong genomic clustering profile in both mTEChi and mTEClo (KL= 2.134 and 2.297, respectively). In turn, for jTECs there is no statistically significant difference between the *Aire*-dependent TRA distance mean and the background, and the divergence between these two distributions (KL = 1.862) is lower than mTEChi and mTEClo counterparts.

For *Aire*-enhanced TRAs, the majority of observed distances overlap the control, suggesting that for most genes the induction of expression mediated by *Aire* does not rely on the activation of stretches of chromatin. Nonetheless, there is a minor proportion of shorter distances hinting that such a mechanism might be important for a reduced number of *Aire*-enhanced TRA genes. Finally, for *Aire*-dependent TRA genes, the distribution of real distances is markedly distinct from the control background, exhibiting a clear peak of short distances.

Overall, these results evidence that the genomic clustering tendency affects only a minority of TRA genes. They also highlight that such mechanism preferentially affects *Aire*-dependent genes in comparison to *Aire*-enhanced genes, which might actually explain why and how AIRE affects them differently. Finally, this clustering effect seems to be established only during the progression from jTEC to mTEC, as both the number of TRAs per cell and genomic clustering of *Aire*-dependent TRAs were very limited in the jTEC population.

III.3 Discussion

III.3.1 Unbiased analysis resolves mTEC development in adult

Here, the analysis that I performed on scRNA-seq data of mTEC cells constitutes, to the best of my knowledge, the first attempt to use single-cell transcriptomics to dissect the development of this pivotal cell type in central immune tolerance. Although similar scRNA-seq data has been produced previously, these studies focused solely on TRA co-expression patterns across mTECs, overlooking their development.

I identify and characterize three distinct stages of mTEC maturation: early mTEC in the cortex-medulla junction (jTEC), *Aire*-expressing mTECs (mTEChi), and mTECs entering the post-*Aire* stage (mTEClo) [32–34]. This dissection is particularly interesting for two main reasons: first, is the first time that these populations are analyzed as a whole, *i.e.* with no bias introduced by sorting strategies, allowing the capture of a more representative range of cell states; second, it identifies a post-*Aire* state that is often masked by the sorting strategies used.

For each of the three stages I provide markers that can potentially reduce the need for transgenic mice or intra-cellular staining of AIRE in future studies. While pseudotime analysis indicates that in general, mTEChi stage preceded mTEClo, there was some overlap between the two populations. This suggested that the transition between these states might not be a tightly programmed event and individual cells might undergo it at asynchronous rates. Notably, hierarchical clustering did not clearly segregate cells expressing high or low levels of the widely-used maturation markers *Cd80* and MHC Class II. This might relate to the fact that CD80^{low} mTECs appear to represent a mixed population, containing both immature precursor cells and terminally differentiated post-*Aire* mTECs [23,141].

I also identify potential new regulators of mTEC development. Besides the expected *Nfkb* and *Irf* families of genes, the regulatory network inferred for mTEC development (Figure 3C) puts forward a few candidate TFs as drivers of this process: *Egr1* mainly on jTEC, and *Vdr*, *Plagl1*, *Zbtb7a*, *Hnf4g* on mTEClo.

Surprisingly, having mTECs from 2 and 4 week-old mice gave some interesting insight into the changes happening in the mTEC compartment during the first weeks of development. The fact that most mTECs from 2 week-old mice have been classified as mTEChi suggests that jTEC and mTEClo are lowly represented so early in the post-natal life. One can hypothesize that fewer mTECs have undergone the full development process at this stage, contributing to a low number of mTEClo cells being present. jTEC absence can be explained by the fact that the thymus is still undergoing

developmental changes at week 2, having a less defined cortex-medulla interface, where jTEC are later located. Nonetheless, one has to consider the low number of sorted cells potentially affecting the representation of all the three subpopulations. Future experiments using some sort of droplet-based microfluidics would be ideal to address these questions.

III.3.2 Relevance of post-*Aire* states in central tolerance

Historically, mTEChi, as key expressers of most TRAs and being particularly competent in antigen presentation, have been considered the main player in negative selection in the thymus, while pre- and post-*Aire* stages would be of limited relevance. However, my observations indicate that mTEClo, and to a more limited extent, jTEC cells, might also contribute to this process. The competence of mTEClo in terms of number of TRAs expressed per cell was in fact equal or greater than that of mTEChi (Figure 4B), suggesting that mTECs progressively express more TRAs as they mature, even after *Aire* expression declines. Moreover, the expression levels of *Fezf2*-affected TRAs were progressively higher from jTEC, to mTEChi and then to mTEClo (Figure 19B). In parallel, differentially expressed genes like *Jag1*, *Cd40*, *Icosl*, Skint-family genes and galectins suggested distinct functions/interactions of mTEC subsets during thymocyte development. Overall, these observations were in line with previous studies [33,34] inspection of MHCII and individual TRA expression in post-*Aire* mTECs, consolidating mTEClo as a key stage in the maintenance of central tolerance. In terms of TRA repertoire, even jTEC could cover most of the *Aire*-unaffected and the *Aire*-enhanced genes, and most TRAs in general (Figure 4A). However, the cell-to-cell ability to express them was significantly impaired (Figure 4B), which likely explains the low levels of TRAs detected in previous reports, namely for *Aire*-dependent TRAs [32]. Taken together, these results indicated that while AIRE is important for turning on TRA expression during jTEC-mTEC transition, the TRA expression was maintained in mTEClo even in the absence of AIRE. In summary, AIRE seems to be critical for inducing TRA expression, but not for maintaining it.

III.3.3 Lack of order optimizes the negative selection process

Theoretically, a couple of hundred mTECs are enough to collectively express the entire repertoire of TRA genes (Figure 1D and [28]), and further contacts of thymocytes with additional mTECs would therefore be unnecessary for increasing their exposure to new self-antigens. This number of mTECs

fits perfectly with the observation that thymocytes visit only a small number of confinement areas (each containing 100-200 mTECs) [142], and appears to be a highly energy/time-effective strategy for covering as much of the TRA repertoire possible. In line with this scenario, it was observed little divergence between the TRA repertoires of each maturation stage (Figure 4A), meaning that the TRAs encountered by thymocytes anywhere across the medulla should not depend largely on the maturation stage of the surrounding mTECs. This is particularly relevant as mTECs tend to re-locate to different regions of the medulla during their maturation (jTEC in the cortex-medulla junction, mTEChi towards the periphery of the medulla and mTEClo towards the centre of the medulla) [33]. Nonetheless, the possibility of subtle differences along the differentiation process having been overlooked (*e.g.* *Tspan8* and *Gad1* trends (Figure 14B and B), or that subtle TRA biases are present within each subpopulation of mTECs cannot be excluded. It remains possible that such patterns can be elucidated by future studies employing technologies with higher cell throughput and higher transcript detection sensitivity.

CHAPTER IV

Single-cell transcriptomics of regulatory T cells reveals
trajectories of tissue adaptation

CHAPTER IV - Single cell transcriptomics of regulatory T cells reveals trajectories of tissue adaptation⁸

IV.1 Abstract

Non-lymphoid tissues (NLTs) harbour a pool of adaptive immune cells distinct from their counterparts in lymphoid tissues, and their development and phenotype remains largely unexplored. Here, I used scRNA-seq to survey CD4⁺ T regulatory (Treg) and memory T (Tmem) cells in spleen, lymph nodes, skin and colon in an unbiased way. I establish comprehensive gene signatures for cell types and tissues, and model a continuous lymphoid-to-NLT trajectory for steady-state murine Tregs to dissect T cell recruitment. Reshaping of Tregs' transcriptomes begins with NLT priming in the lymph nodes and continues with adaptation in the NLT, and a core signature of this process is shared between skin and gut. Predicted kinetics are validated using a melanoma-induction model, emphasizing the relevance of key regulators and receptors such as *Batf*, *Rora*, *Ccr8*, *Samsn1*. Finally, I profile human blood and NLT cells and identify conserved tissue signatures. In sum, I provide a new conceptual framework for modelling scRNAseq data to recapitulate and uncover new genes involved in Treg NLT adaptation.

⁸ Parts of this section, including figures, were adapted from [172].

IV.2 Results

IV.2.1 Transcriptomes of CD4⁺ Treg and Tmem cells cluster by cell type and tissue

To compare T cell transcriptomes in different tissues, CD4⁺Foxp3⁺ (Treg) and CD4⁺Foxp3⁺CD44^{high} memory (Tmem) T cells were isolated from a Foxp3-GFP-KI mouse reporter line [110] from two barrier NLT sites - the colonic lamina propria, designated as colon for simplicity, and the skin. These were profiled together with their lymphoid counterparts: the respective draining lymph nodes (LN), mesenteric and brachial lymph nodes (mLN and bLN), and the spleen (Figure 22A, Figure 23). Each NLT (and accompanying lymphoid tissues) was obtained separately, and henceforth will be referred to as “colon” or “skin” datasets. Treg and Tmem populations together will be referred as CD4⁺ T cells.

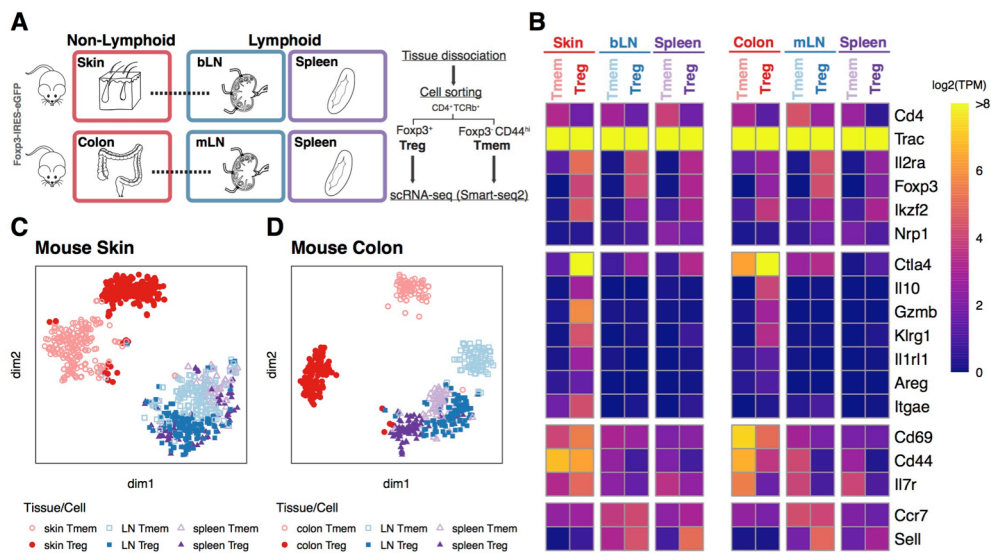


Figure 22 - Steady-state scRNA-seq datasets of CD4⁺ T cells from LT and NLT.

(A) Two independent scRNA-seq datasets composed of Treg (TCRb+CD4+Foxp3+) and Tmem (TCRb+CD4+Foxp3-CD44+) cells from an NLT (colon or skin), draining lymph node (LN, mesenteric LN for the colon and brachial LN for the skin, respectively) or spleen, sorted from Foxp3⁺ reporter mice were generated.

(B) Mean expression levels of characteristic markers across CD4⁺ T cell populations and tissues are recapitulated by scRNA-seq.

(C and D) t-SNE dimensionality reduction of colon (C) and skin (D) datasets, showing Treg and Tmem from NLTs (colon and skin), LNs (mLN and bLN) and spleen. Treg and Tmem are represented by filled and open symbols, respectively. Colours and symbols match tissue: NLTs in red circles, LNs in blue squares, spleen in purple triangles.

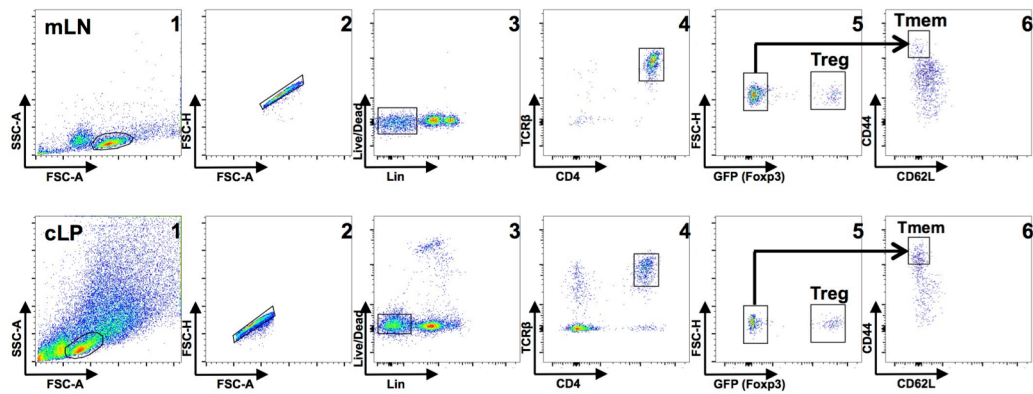


Figure 23 - Sorting scheme for murine Treg and Tmem.

FACS-sorting strategy for sorting Treg and Tmem cells from (top) lymphoid (mLN) and (bottom) non-lymphoid (colonic lamina propria, cLP, as an example) organs.

From single cell suspensions, individual Treg and Tmem cells were index-sorted into 96-well plates. Poly-adenylated mRNA was then reverse transcribed and amplified using the Smart-seq2 protocol [54], and subsequently sequenced after library preparation. After mapping and quality control, 485 cells were retained from the colon dataset, and 796 cells from the skin dataset, expressing on average 3118 genes (TPM>0) (Figure 24A; see Material and Methods). The thresholds used for quality control were adapted to the distributions found in each dataset. TCR sequences reconstructed using TraCeR [68] were used to exclude non-T cell contaminations, as well as few iNKT cells sorted as Tmem in the spleen and $\gamma\delta$ -T cells in the skin (Table 7).

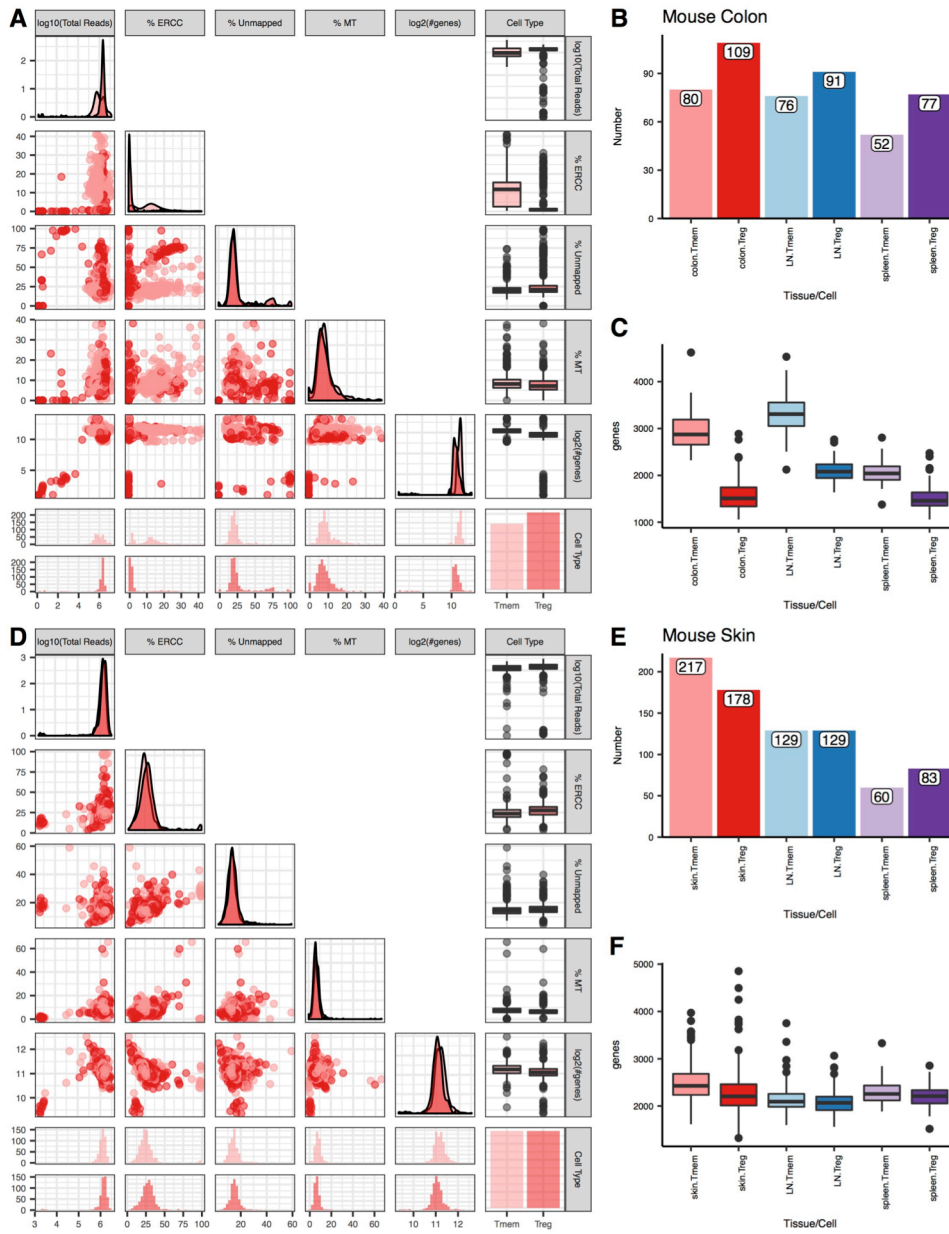


Figure 24 - Quality-control of steady-state colon and skin scRNA-seq datasets.

(A and D) Multiple QC metrics and their relationships in the (A) colon and (D) skin dataset. Measurements for individual cells are depicted in scatterplots. The respective distributions are represented as density plots (diagonal), and histograms (bottom). Distribution per cell type is represented in boxplots (right). Treg and Tmem are marked as dark and light red, respectively.

(B and E) Number of single-cells from colon (B) and skin (D) datasets retained after QC, per tissue and cell type.

(C and F) Distribution of the number of genes for the colon (C) and skin (F) datasets per tissue and cell type.

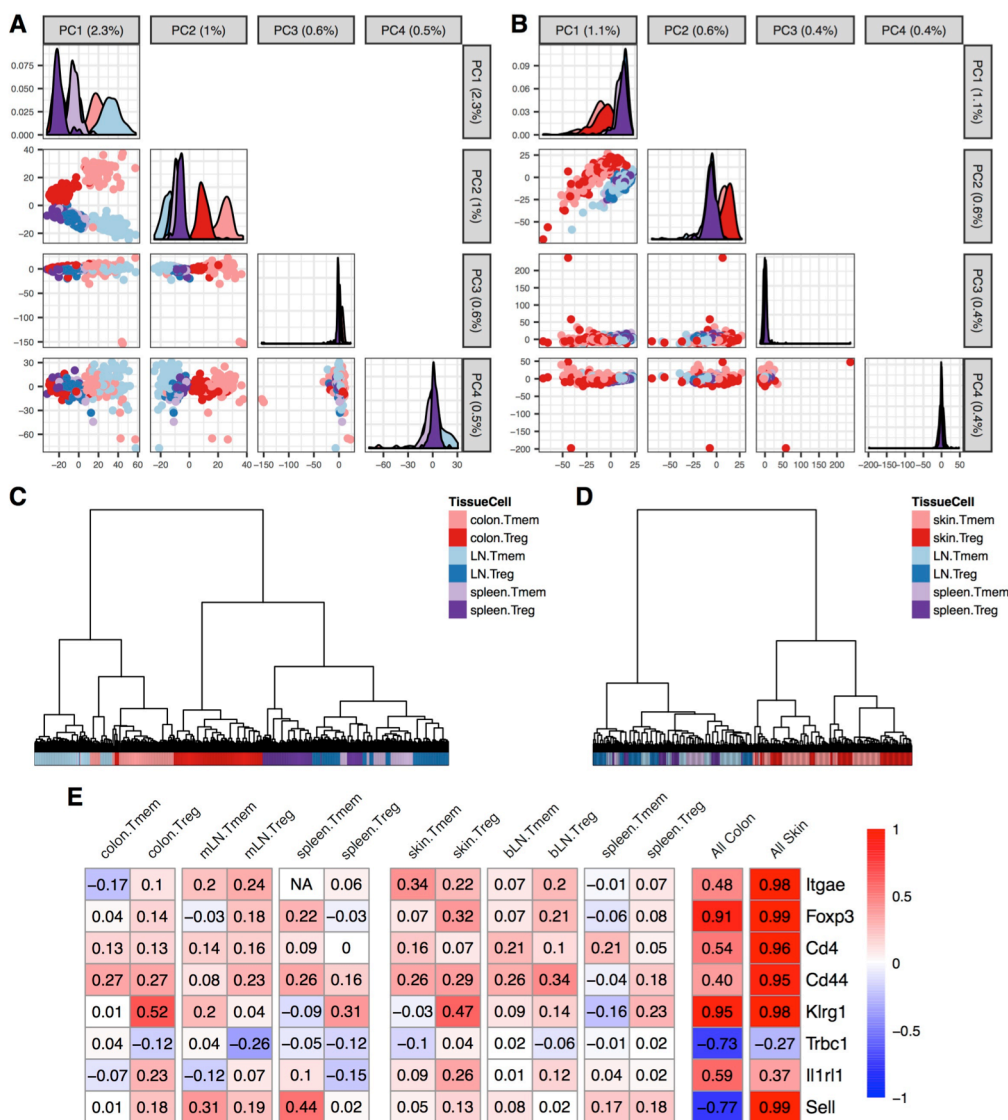


Figure 25 - Additional characterization of sorted CD4⁺ T cell populations from steady-state mouse tissues.

(A and B) Principal Component Analysis (PCA) projections of the mouse colon (A) and skin (B) datasets. The first four PCs were plotted against each other, and the percentage of variance explained by each of them is presented in the title boxes. (C and D) Hierarchical clustering of cell-to-cell Spearman correlations across colon (C) and skin (D) datasets. (E) Spearman correlation values between mRNA and protein expression the single-cell level by tissue and cell type. The two rightmost columns refer to correlations between the average levels of mRNA and the average levels of protein for all subpopulations in colon and skin datasets, respectively. Protein expression values were obtained from FACS index sorting data.

In these datasets, the expression levels of classical Treg and Tmem marker genes reproduce previously known patterns (Figure 22D). *Foxp3* expression is reliably detected in Tregs while being absent from Tmem cells, and effector Treg molecules such as *Ctla4*, *Il10*, *Gzmb*, *Klrg1* and *Il1r1l*

(ST2), were highly expressed only in Tregs isolated from the NLTs. *Sell* (CD62L) characterises Treg and Tmem from LT²² [143], while *Ccr7* is specific to lymph nodes, as previously described [144].

To get a global picture of the relationships between the cell populations isolated, a dimensionality reduction algorithm (tSNE) was applied to skin and colon datasets individually (Figure 22B,C). Both datasets were treated separately as much as possible to avoid introduction of batch effects. Overall, populations in both datasets are structured in a similar fashion, and they contain only small numbers of cycling cells (Supplementary Figure 1). First and foremost, Treg and Tmem populations from NLTs are clearly separated from the respective lymphoid counterparts. This can be seen in the tSNE and along PC2 of the Principal Component Analysis (PCA) (Figure 22C,D, Figure 25A-D), evidencing that these cell types exhibit distinct gene signatures in NLTs, in line with previous studies with bulk RNA-seq in human have shown [145]. Secondly, Treg and Tmem from spleen and LNs form a tight cluster, albeit with distinct regions for each population, underscoring the similarity between lymphoid tissue CD4⁺ T cells.

From a technical perspective, scRNA-seq plate-based approaches allow for the comparison between mRNA levels and protein levels of the markers included for cytometry. As observed by others [146], mRNA-protein correlations are pushed to low values, likely by a mixture of post-translational regulation and high technical drop-out rates (Figure 25E). Nonetheless, the agreement between the average values of mRNA and protein across the six populations considered is high for most markers, *i.e.* if a population has higher average level of protein, scRNA-seq will detect higher average level of mRNA.

tSNE projections do not provide evidence of clear sub-clusters within each cell type in the skin or colon, which is at odds with reports of phenotypically and functionally diverse subpopulations of CD4⁺ T cells within NLTs [47,147,148]. Therefore, to search for subpopulations within the NLT Treg and Tmem cells, the consensus clustering algorithm implemented in the SC3 R package [122].

The search ranged from 2 to 4 clusters to be able to exclude cases where small spurious subpopulations are identified. For each tissue/cell combination, a comparison was made between the largest subpopulations detected, which incidentally were also the farthest apart in tSNE space. Differential expression between clusters was performed on all expressed genes in them. This confirmed that colonic Tregs remained a homogenous population (Figure 26A-C), with no differentially expressed genes detected between them.

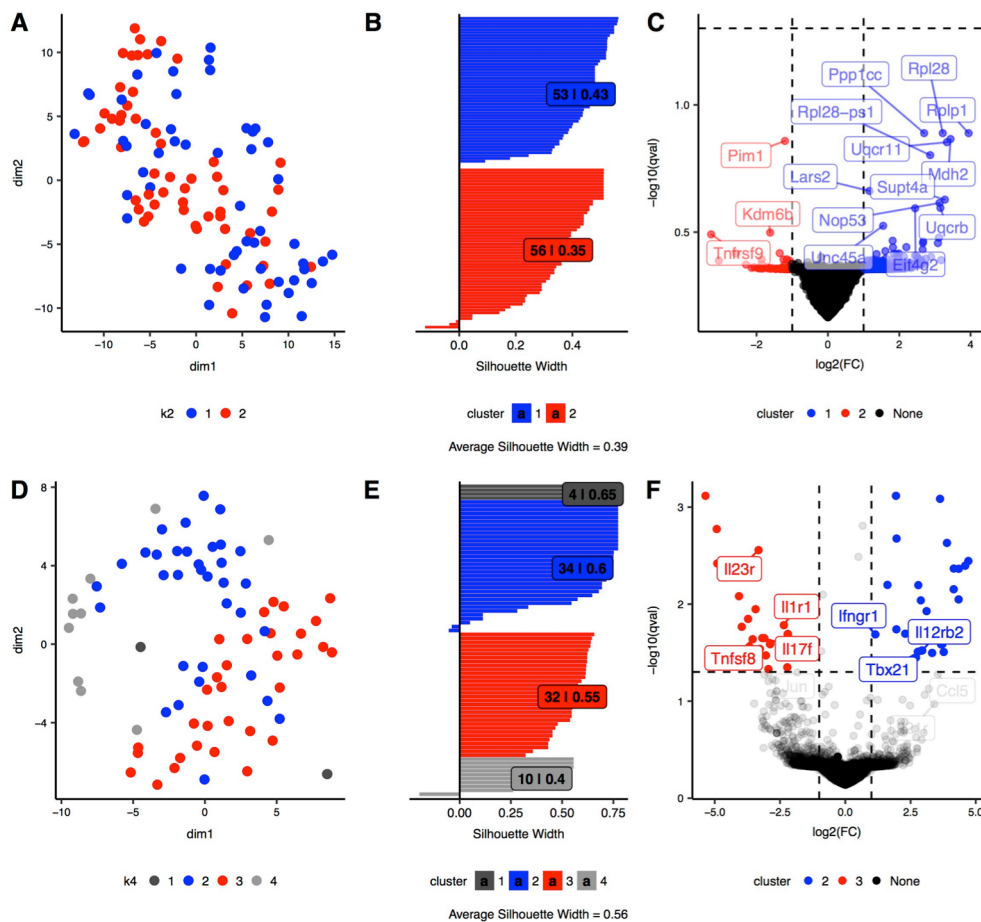


Figure 26 - Assessment of intra-population heterogeneity by consensus clustering in colonic cell populations.

(A) t-SNE dimensionality reduction of colon Treg. Red and blue identify the two clusters found employing a consensus clustering approach.

(B) Silhouette plot showing cluster stability within and between colon Treg clusters.

(C) Differential expression between subpopulations identified within colon Treg ($q\text{-value} \leq 0.05$ and $\log_2(\text{FC}) \geq 1$). Genes are coloured based on the cluster they are upregulated in (based on fold-change). Transparent genes are not significantly DE. In this case, no genes were found to be differentially expressed.

(D) t-SNE dimensionality reduction of colon Tmem. Colors identify the four clusters found employing a consensus clustering approach. Red and blue identify the two clusters considered for comparison.

(E) Silhouette plot showing cluster stability within and between colon Tmem clusters.

(F) Differential expression between two subpopulations identified within colon Tmem ($q\text{-value} \leq 0.05$ and $\log_2(\text{FC}) \geq 1$). Genes are coloured based on the cluster they are upregulated in (based on fold-change). Transparent genes are not significantly DE.

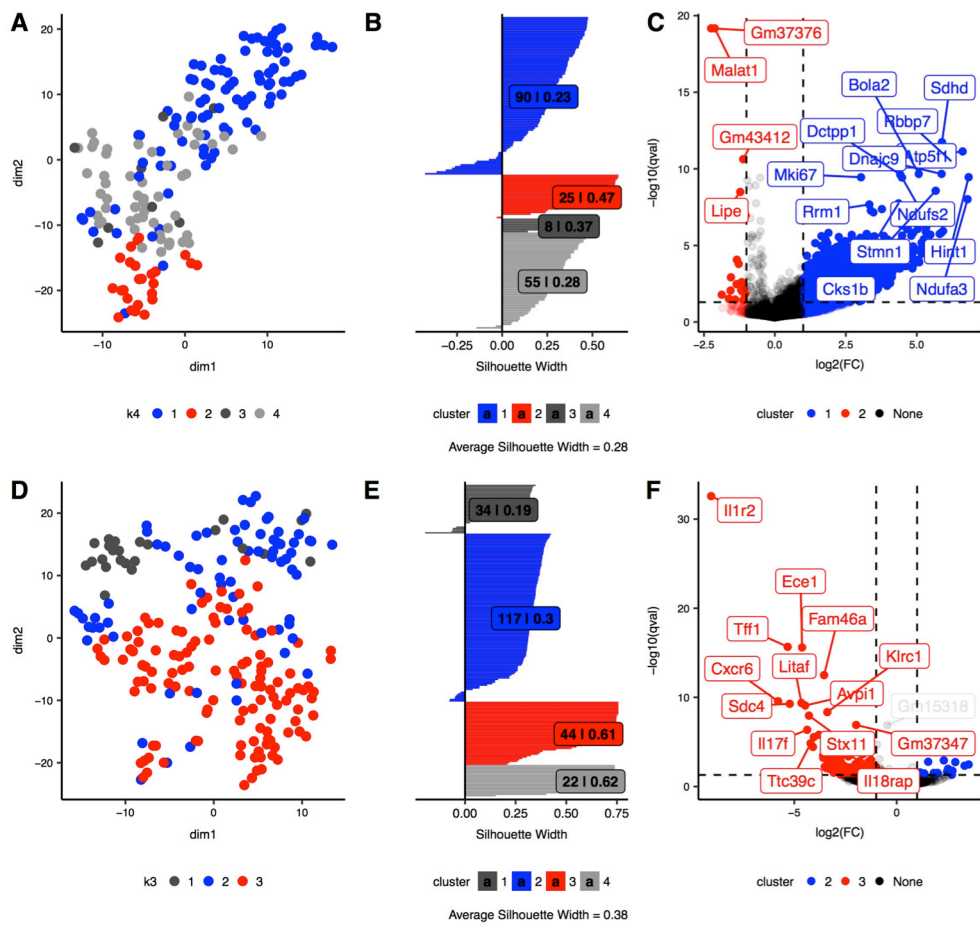


Figure 27 - Assessment of intra-population heterogeneity by consensus clustering in skin cell populations.

(A) t-SNE dimensionality reduction of skin Treg. Colors identify the four clusters found employing a consensus clustering approach. Red and blue identify the two clusters considered for comparison.
 (B) Silhouette plot showing cluster stability within and between skin Treg clusters.
 (C) Differential expression between two subpopulations identified within skin Treg ($q\text{-value} \leq 0.05$ and $\log_2(\text{FC}) \geq 1$). Genes are coloured based on the cluster they are upregulated in (based on fold-change). Transparent genes are not significantly DE.
 (D) t-SNE dimensionality reduction of skin Tmem. Colors identify the four clusters found employing a consensus clustering approach. Red and blue identify the two clusters considered for comparison.
 (E) Silhouette plot showing cluster stability within and between skin Tmem clusters.
 (F) Differential expression between two subpopulations identified within skin Tmem ($q\text{-value} \leq 0.05$ and $\log_2(\text{FC}) \geq 1$). Genes are coloured based on the cluster they are upregulated in (based on fold-change). Transparent genes are not significantly DE.

In contrast with Treg, colonic Tmem cells present more diversity, segregating into a cluster characterized by higher expression of Th1-associated genes, such as *Tbx21* (*T-bet*), *Il12rb* and *Ifngr1*, and a second cluster characterized by higher expression of Th17-associated genes, including *Il23r*, *Il17f* and *Il1rl1* (Figure 26D-F). Skin Treg and Tmem populations appear to exhibit small gradients of activation and cytokine expression within each population (Figure 27). It is worth noting that absence of evidence for the existence of subpopulations is not evidence of their absence. Indeed, it has

previously been shown that the number of cells in a scRNA-seq experiment is more important for subpopulation detection than the number of detected genes [149]. Therefore, I underline that the presented datasets are mostly useful for a deep characterization of the sorted populations.

Overall, these single-cell transcriptomic data validate the expression of known CD4⁺ T cell markers, and provide evidence of dramatic phenotypic changes between Treg and Tmem populations in NLT adaptation.

IV.2.2 Gene expression signatures of NLT regulatory and memory CD4⁺ T cells

Next, the transcriptomic differences between NLT and LT CD4⁺ T cells were explored. Separately for skin and colon datasets, a differential expression test was performed to determine which genes characterize CD4⁺ T cells in the NLTs. Then, to investigate cell type-specific NLT adaptations, NLT markers were classified as Treg, Tmem or shared, based on differential expression between the two populations (Figure 28A,B). In colon, 31 genes were unique to Treg, 289 unique to Tmem, and 83 were shared between the two. In the skin, 85 genes were unique to Treg, 146 unique to Tmem and 527 were shared.

The common signature expressed by CD4⁺ T cell found in the colon and skin (“Both NLT” in Figure 28C) included transcription factors with known functions in T cells, such as *Id2*, *Crem* and *Maf*, kinases *Sik1* and *Pim1*, and histone demethylases *Jmjd1c* and *Kdm6b*. In addition, NLT Tregs (as compared to Tmems) were enriched in the expression of TNF receptors (*Tnfrsf4*, *Tnfrsf9*, *Tnfrsf18*), homing receptors, including *Itgav* and *Ccr2*, and immunosuppressive molecules *Fgl2*, *Tigit* and *Il10* (“Both NLT” in Figure 28C). In line with previous reports, the chemoattractant receptor *Gpr15* was highly expressed in both colonic Tmem and Treg cells [150] (“Colon” in Figure 28C). *Skil*, *Gata3* and *Ccr8* were preferentially expressed in colonic Tregs *versus* Tmems. In contrast, colonic Tmem cells expressed high levels of genes important for T cell survival and function *Runx1*, *Irf4* and *Ifng*, which are absent from Tregs.

The skin NLT signature (“Skin” in Figure 28C) includes the chemokine receptor *Ccr6*, along with vimentin, *Vim*, two galectin-encoding genes, *Lgals1* and *Lgals3*, and transcription factors, such as *Mxi1* and the TGFb family member *Smad7*. On top of this, the skin Treg signature presents high expression of *Batf*, the NF-κB modulator *Bcl3*, effector Treg markers *Il1r1* and *Itgae*, and genes with hitherto undefined function in the context of Treg biology, including the chemokine-like factor *Cmtm7*,

triglyceride metabolism enzyme *Dgat2* and putative signal transduction modulator *Gem*. In sum, scRNA-seq characterisation reveals a rich landscape of global and cell type-specific expression patterns in non-lymphoid tissues.

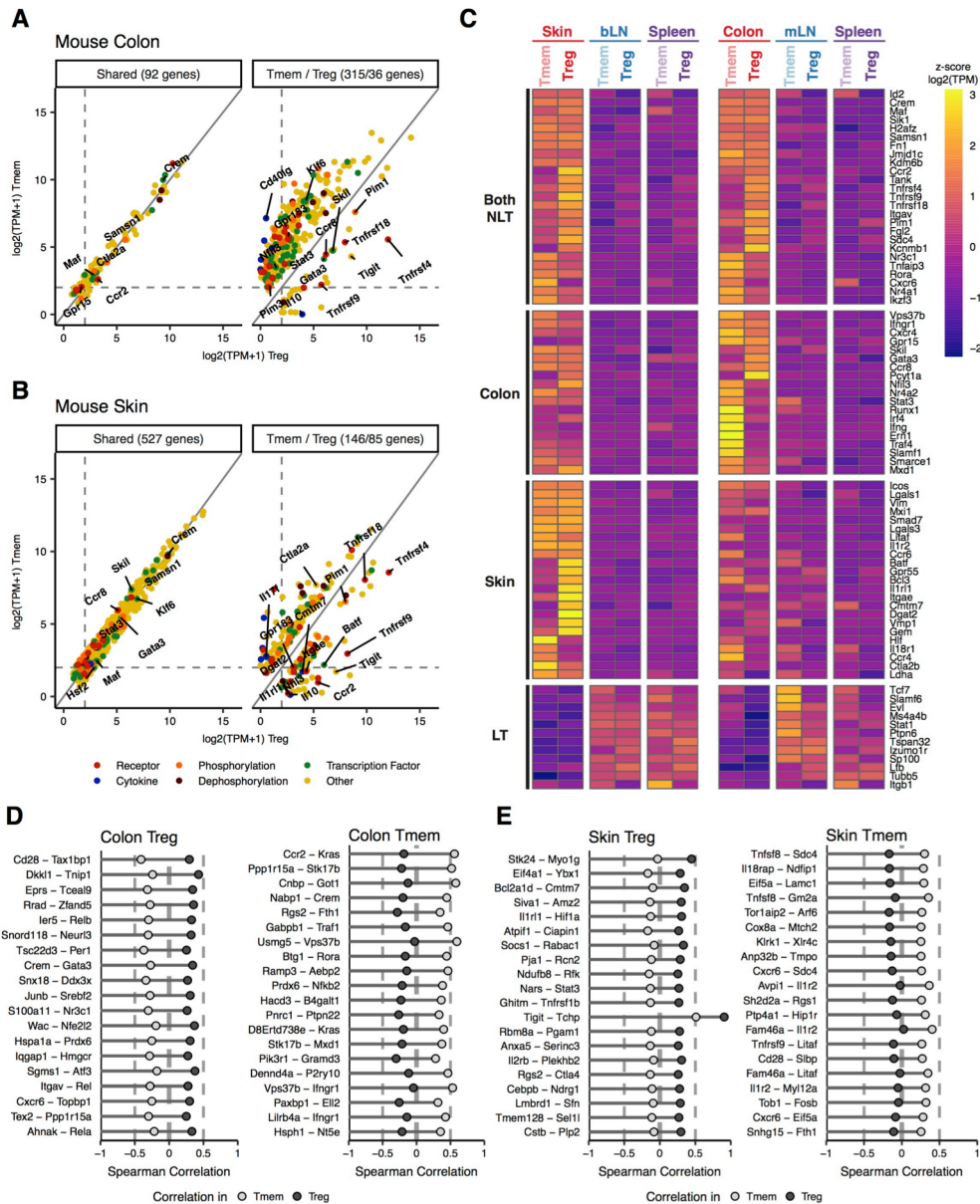


Figure 28 - Characterization of Treg and Tmem NLT adaptations.

(A and B) Differential expression of NLT markers between NLT Treg and Tmem, in colon (A) and skin (B). Per dataset, genes were first identified as NLT- or LT-associated ($q\text{-value} \leq 0.01$, $\log_2 \text{fold-change (FC)} \geq 1$). This subset of genes was then characterized as Treg, Tmem or shared between cell types based on $q\text{-value} \leq 0.05$ and $\log_2 \text{FC} \geq 1$ (see Methods).

(C) Z-score of mean expression levels of newly identified markers across cell types and tissues, grouped by the populations where they are differentially expressed (left labels).

(D) Top 20 differentially correlated gene pairs in colonic Treg (left) and Tmem (right), ordered by difference in correlation between both populations.

(E) Top 20 differentially correlated gene pairs in skin Treg (left) and Tmem (right), ordered by difference in correlation between both populations.

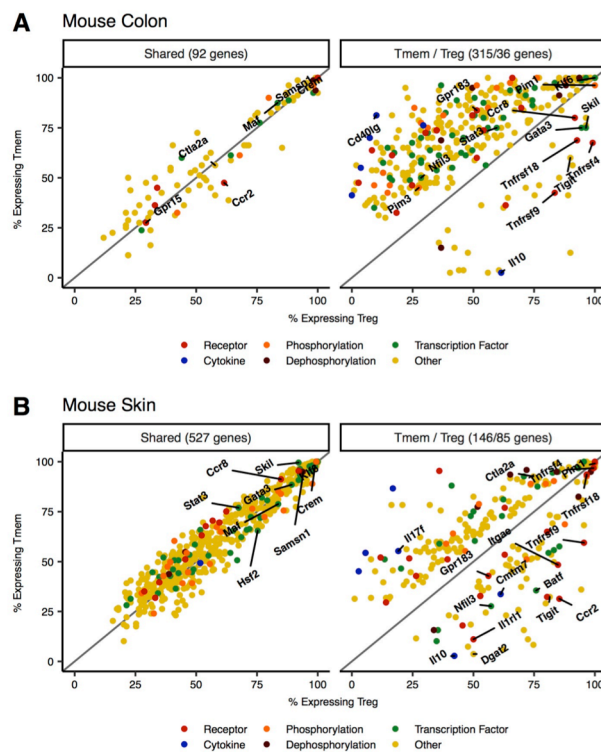


Figure 29 - Percentage of cells expressing NLT markers from mouse colon (A) and skin (B).

Although preferentially expressed in either NLT Treg or Tmem populations, most marker genes are present in both cell types to a certain extent (Figure 28A-C, Figure 29). This observation suggests that the interactions established by genes within each cell type can be more relevant than the expression values per se. Therefore, a differential correlation test between cell types in each NLT was used to identify differential gene-gene interactions between Treg and Tmem using (refer to Methods for detailed description).

With this approach, the coexpressed pairs of genes exhibiting cell type-specific differences were ranked by the magnitude of this difference (correlations of top pairs are the most different between cell types). In colon (Figure 28D), *Crem* and *Gata3* are strongly correlated in Treg, mirroring their induction in Th2 cells [151]. Within colon Tmem, there is a strong relationship between *Nfkb2* and *Prdx6*, which has been shown in the context of ROS homeostasis [152] in a different system. In skin (Figure 28E), *Stat3* is potentially inducing *Nars* expression [153] in Treg, while Tmem exhibit signs of exhaustion by co-expressing *Sh2d2a* and *Rgs1* [154]. Skin Tmem also show a high correlation between the pair *Fosb* and *Tob1*, upregulated by platelet-derived growth factor [155] in another system.

NLT adaptations to skin and colon were extensively identified and characterized, specifically emphasizing the similarities and differences between Treg and Tmem. Differential expression analyses were then complemented with the examination of differential gene-gene interactions.

IV.2.3 Tregs in skin and colon present contrasting biases

Treg adaptation across NLTs was then directly addressed. To do so, both populations were compared by differential expression (Figure 30A) to understand which transcriptomic changes underlie Treg adaptation to either the colon or the skin environment. As expected from the previous analyses, there is a core NLT Treg signature shared across tissue Tregs, e.g. *Kdm6b*, *Itgav*, *Tnfrsf4*, *Fgl2*. In addition to this shared signature, transcription factors *Batf* and *Nfil3*, and the surface receptors *Tnfrsf9* and *Il1rl1*, and the Ca²⁺-binding protein *S100a6*, are all upregulated in the skin Treg cells. In turn, Treg in the colon are characterized by the activation markers *Cd83* and *Ccr8*, the receptor *Gpr15*, and the immunosuppressive cytokine *Il10*. This seems to suggest that in the skin environment there is a stronger induction of typical NLT Treg markers [49], while the colonic environment promotes a rather activated Treg state, potentially due to the very close contact between the mucosa and foreign antigens.

Similarly to the comparison between Treg and Tmem, information about gene-gene interactions established in each NLT was then extracted (Figure 30B). The differential correlation test highlighted several pairs of genes that shift in relevance from colon to skin. For example, in colon Treg, previously described MARCH-mediated ubiquitination effects on *Cd44* stability [156] might underlie *March7* correlation with *Cd44*. The interaction between *Il1rl1*, a known player in visceral adipose tissue (VAT) Treg development, and *Hif1a*, involved in T cell metabolism reprogramming [157], can be directly associated with skin Treg differentiation. Also in skin, there is a high correlation between *Rel* and *Il1r2*, whose promoter contains a potential binding site for *Rel* [158], and a high correlation between the MAPK-pathway genes *Rgcc* and *Klf4*, which have been shown to be correlated in T cell activation settings [159].

In sum, Tregs show a general NLT phenotype which, in colon, is biased towards activation, and in skin is biased towards a more differentiated state. Some of the gene-gene interactions determined are likely to be involved in these tissue differences.

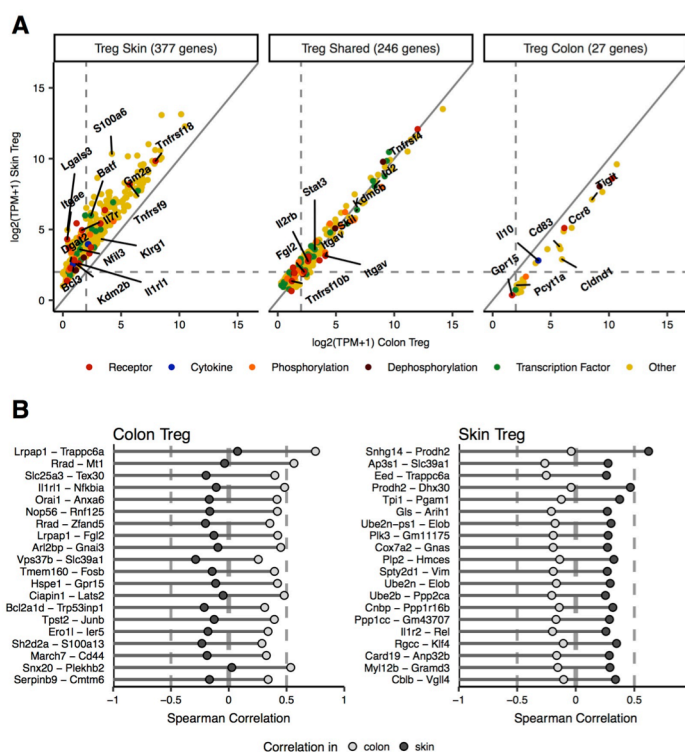


Figure 30 - Specific Treg adaptations to skin and colon environments.

(A) Differential expression of NLT markers between skin and colon Treg. Genes are characterized as colon-specific, skin-specific or shared based on based on $q\text{-value} \leq 0.01$ and absolute $\log_2\text{FC} \geq 1$ (see Methods).

(E) Top 20 differentially correlated gene pairs in colon Treg (left) and skin Treg (right), ordered by difference in correlation between both populations.

IV.2.4 Pseudospace alignment of Tregs from LNs to NLTs reveals stages of tissue adaptation

The analyses performed so far demonstrate how NLT Tregs differ from their LT counterparts in terms of expression of a set of NLT-specific genes. It is known that NLT populations are progressively established throughout an individual's lifetime, as Tregs accumulate in the NLTs of mice under steady-state conditions [160,161]. However, the actual transition process from LT to NLT phenotypes is far from being completely understood.

To investigate whether there is evidence of cell trafficking between lymphoid and non-lymphoid tissues, TCR sequences were reconstructed using TraCeR [68] (Supplementary Figure 2A,B) and T cell clonal relationships were investigated. TCR clonotype analysis reveals that clonal relationships occur exclusively within the Tmem and Treg compartments, rather than across these cells types. Tmem and Treg clones are observed both within and across lymphoid and non-lymphoid

environments, suggesting trafficking of members of the same clone across locations. It is worth mentioning that for these steady-state experiments, the cells captured come from pools of mice and not a single animal, which is likely to cause an underestimation of clonotype sharing frequency.

Given the absence of subpopulations within Treg populations, phenotypic changes associated with migration and adaptation to NLTs are likely to constitute a rather continuous process, from activation of Tregs in draining-LNs, proliferation, migration from LN towards the respective NLT, and potential priming for NLT adaptation along the way. On the NLT side, it is also likely that Tregs in the NLT can present various degrees of adaptation depending on the time spent in the NLT environment. Therefore, to identify such trends in the Treg data, this “pseudospace” relationship between cells reflecting recruitment and adaptation of Tregs to NLTs, was computationally reconstructed. Bayesian Gaussian Process Latent Variable Modelling (BGPLVM) [119] (see Methods) were employed to Treg transcriptomes from NLTs and their respective draining-LN (Figure 31). The non-linear latent variables (LVs) representing combinations of gene expression patterns can then be interpreted biologically. Per dataset, two main LVs that explain most of the transcriptomic differences present across cells and tissues (LV0 and LV1) (Figure 31A,D; Figure 32A,B) were identified. Importantly, the general organization along LVs (Figure 31A,D), as well as the genes associated with each of them are significantly similar in both mLN-to-colon and bLN-to-skin transitions (Figure 32C; see Methods). This pair of latent variables contains a larger proportion of NLT-associated genes (Figure 32D), and the gradient is retained even when analysing NLT and LN populations individually (Figure 33). LV1 in colon and LV0 in skin could thus be used to represent the continuum of Treg activation and trafficking between LNs and NLTs. Although trajectories for Tmem populations were also calculated (Figure 34), no downstream analyses were conducted, and only Treg were further investigated.

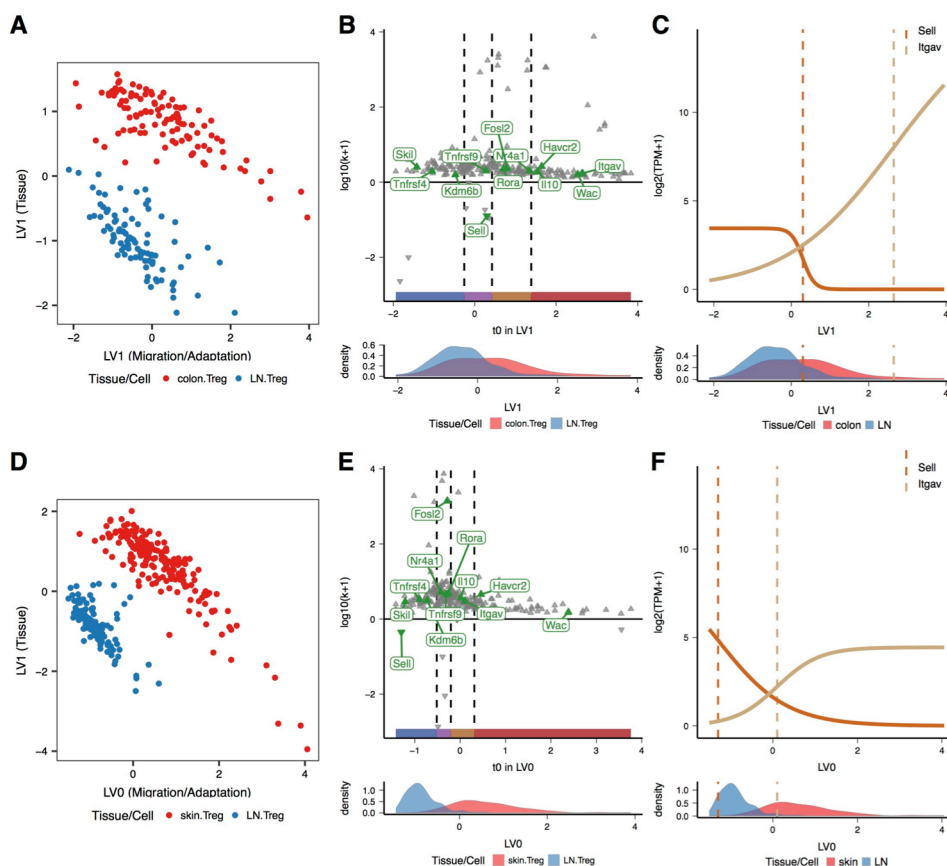


Figure 31 - Reconstruction of Treg recruitment from lymphoid to non-lymphoid tissues in steady-state.

(A and D) Top two latent variables found with BGPLVM of Treg in (A) skin and (C) colon datasets, associated with either tissue of origin or activation/recruitment to non-lymphoid tissue. Colours match tissue of origin: non-lymphoid tissue in red, lymph nodes in blue.

(B and E) Activation moment (t_0) and intensity (k) of genes significant in migration to colon or skin. (Top) Time of activation/deactivation (t_0) of genes along the activation/recruitment trajectory as determined by the switchDE package, in colon (B) and skin (E) datasets. (Bottom) Distribution of Treg cells from lymph nodes and the non-lymphoid tissues along trajectories.

(C and F) Sigmoid fits of *Itgav* and *Sell* expression along (C) colon and (F) skin recruitment trajectory as determined by switchDE. Dashed lines represent t_0 .

Having established a trajectory for recruitment and adaptation of Tregs from LNs to NLTs, the kinetics of up- and down-regulation of genes along it were then examined. This question was addressed by fitting a sigmoid curve to the expression values of each gene along the pseudospace trajectory (see Methods) (Figure 31E,F). This allowed the determination of each gene's point of activation/deactivation in the trajectory (t_0) (Figure 31G,H). From the 318 and 1163 genes that follow a sigmoidal trend in the bLN-to-skin and mLN-to-colon paths, 243 genes in common were identified, which are still associated with various immune functions and cell migration (Figure 35A).

Several NLT transcription factors (*Maf*, *Id2*, *Gata3*, *Irf4*, *Nfil3*, *Stat3*, *Fosl2*, *Nr3c1*, *Ikzf3*, *Nr4a1*, *Skil*, *Rora*) and chromatin regulators (*Kdm6b* and *Jmjd1c*) appear to drive this progression in both NLTs. Effector membrane proteins (*Ctla4*, *Cd44*, *Cd82*, *Tigit*, *Klrg1*, *Icos*, *Gzmb*, *Havcr2*, *Cmtm7*), as well as secreted factors (*Lgals1*, *Lgals3*, *Il10*) are upregulated, progressively increasing NLT Treg ability to interact with surrounding cells and shape the environment in the NLTs. Membrane receptor molecules (*Ccr2*, *Ccr8*, *Cxcr6*, *Gpr183*, *Itgav*) can also be found in this core NLT signature. Expression of TNF receptors (*Tnfrsf4*, *Tnfrsf9*, *Tnfrsf18*) and a TNF transducer (*Traf1*) underline the role that Nf- κ b signalling is likely to exert over Treg function and survival in the NLTs as a whole [162]. Downregulation of *Sell* (Cd62I) and *Evl*, respectively relevant for T cell trafficking to secondary lymphoid organs in steady-state and inflammation, and *Bcl2*, occurs along the trajectory ordered towards the NLT.

Interestingly, genes previously detected as more highly expressed in NLT Tmems (*Rora*) or differentially present between colon and skin Tregs (*Nfil3*, *Stat3*, *Skil*) can still play a role in the inferred migration along both trajectories, and have similar expression kinetics. As determined before (Figure 30B), different interactions can be established in each context, potentially leading to distinct overall transcriptomic signatures. Alternatively or in addition to this, the suggested differences of states of activation and differentiation between colon and skin Treg (Figure 30A) might explain the differences in expression at the population level. Interestingly, there are factors specific to the skin trajectory such as *Ikzf4* and *Il1rl1*, known to contribute to Treg stability and enhanced suppression respectively [163,164].

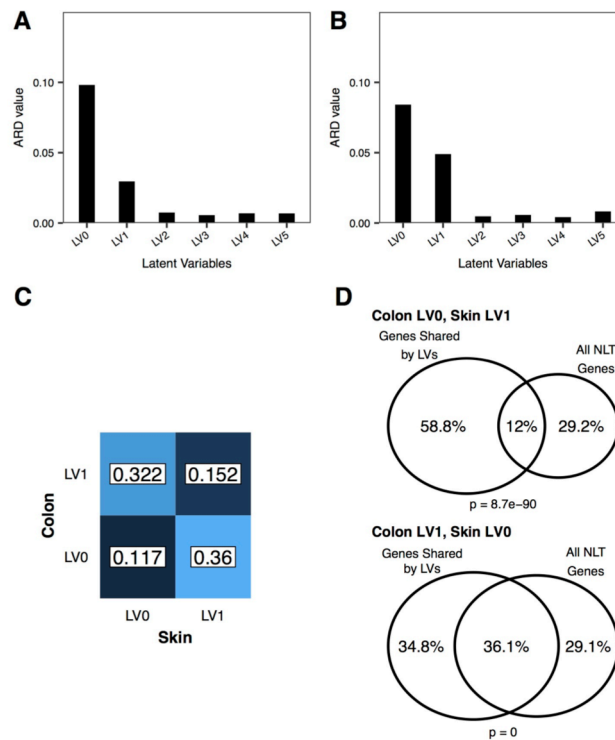


Figure 32 - BGPLVM Latent Variable relevance for LN-to-NLT transition in the colon and skin datasets.

(A-B) Automatic Relevance Determination (ARD) plots for BGPLVM of Treg in colon (A) and skin (B) datasets
 (C) Correlation between latent variables using the correlations of all genes with each latent variable. This shows that Colon dataset LV1 is more similar to Skin dataset LV0, and Colon dataset LV0 is more similar to Skin dataset LV1.
 (D) Venn diagrams intersecting genes in common between pairs of latent variables and NLT markers. (top) Genes in Colon dataset LV0 and Skin dataset LV1; (bottom) Genes in Colon dataset LV1 and Skin dataset LV0 (both used as a “pseudospacial” variable).

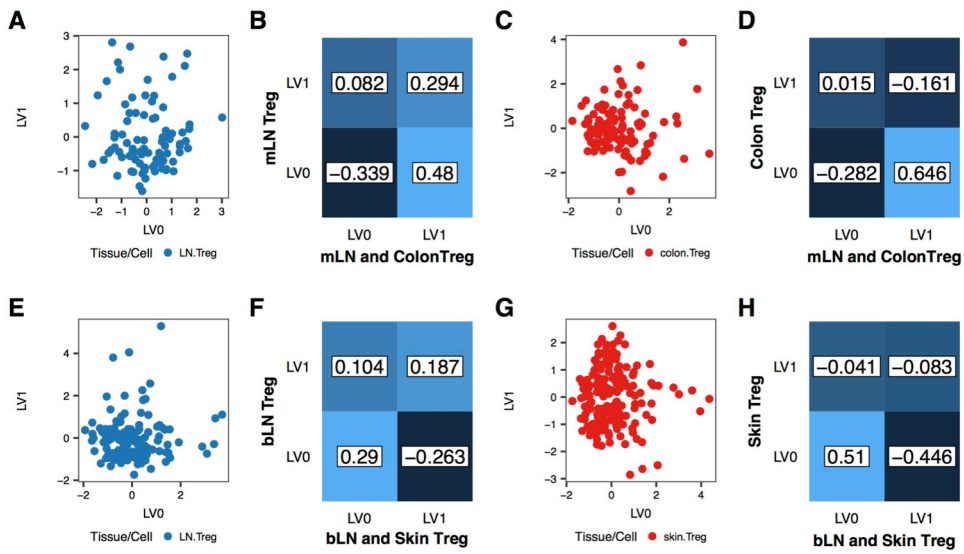


Figure 33 - The effect of the “pseudospace” latent variable can be identified in individual tissues.

- (A) BGPLVM projection of mLN Tregs, using the top two latent variables.
- (B) Correlation between latent variables using the correlations of all genes with each latent variable. X axis - Colon dataset Treg latent variables; Y axis - mLN Treg latent variables.
- (C) BGPLVM projection of colon Tregs, using the top two latent variables.
- (D) Correlation between latent variables using the correlations of all genes with each latent variable. X axis - Colon dataset Treg latent variables; Y axis - colon Treg latent variables.
- (E) BGPLVM projection of bLN Tregs, using the top two latent variables.
- (F) Correlation between latent variables using the correlations of all genes with each latent variable. X axis - Skin dataset Treg latent variables; Y axis - bLN Treg latent variables.
- (G) BGPLVM projection of skin Tregs, using the top two latent variables.
- (H) Correlation between latent variables using the correlations of all genes with each latent variable. X axis - Skin dataset Treg latent variables; Y axis - skin Treg latent variables.

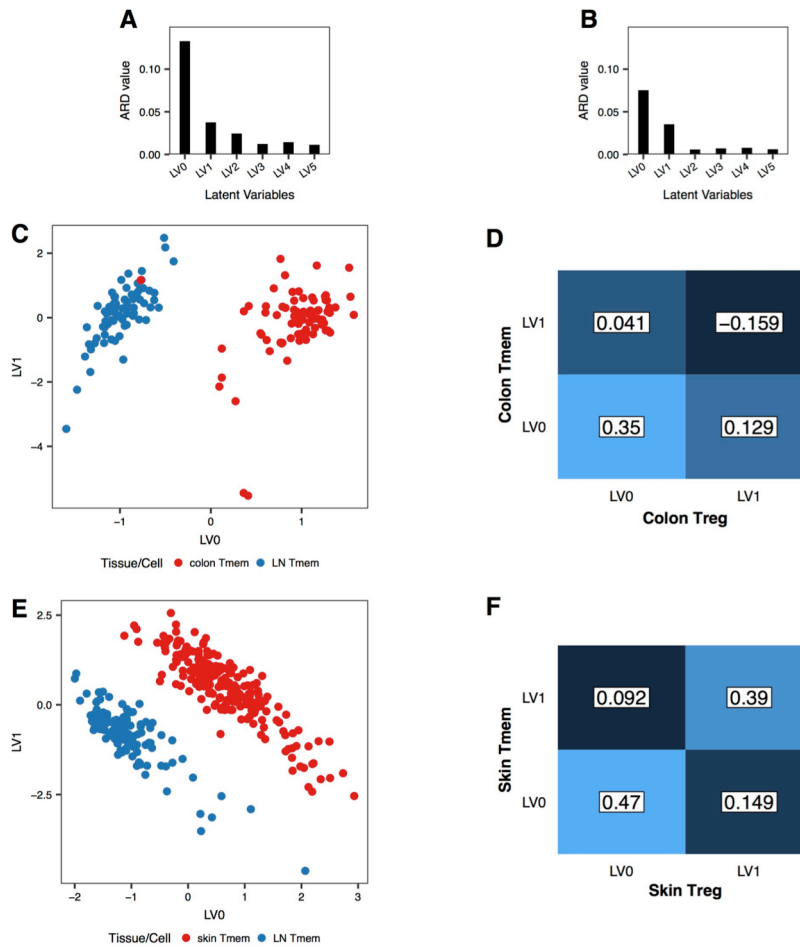


Figure 34 - Comparison of BGPLVM in Tmem and Treg from the colon and skin datasets.

(A-B) Automatic Relevance Determination (ARD) plots for BGPLVM of Tmem in colon (A) and skin (B) datasets

(C) BGPLVM projection of mLN and colon Tmems, using the top two latent variables according to ARD.

(D) Correlation between latent variables using the correlations of all genes with each latent variable. X axis - Colon dataset Treg latent variables; Y axis - Colon dataset Tmem latent variables.

(E) BGPLVM projection of bLN and skin Tmems, using the top two latent variables according to ARD.

(F) Correlation between latent variables using the correlations of all genes with each latent variable. X axis - Skin dataset Treg latent variables; Y axis - Skin dataset Tmem latent variables.

Going one step further, the conservation of the sequence of gene activation/deactivation between the two trajectories was investigated. Early NLT adaptation, *i.e.* the 1st and 2nd quartiles of activated/deactivated genes, is consistent between NLTs (Figure 36A; see Methods). The first wave of NLT adaptation is characterized by upregulation of genes such as *Ctla4*, *Kdm6b*, *Tnfrsf4* and *Tnfrsf18*, while the second wave includes the upregulation of *Icos*, *Lmna* and *Tnfrsf9*, and downregulation of *Bcl2* and *Evl*. Respectively, they are enriched in “leukocyte transendothelial migration” and “cytokine-cytokine receptor interaction” (Figure 36B,C). In contrast, late NLT adaptation (3rd and 4th quartiles) is less consistent between colon and skin, which can potentially be

explained by different environmental cues in each NLT. Nevertheless, the transcription factors *Stat3* and *Irf4*, known to be important in Treg adaptation and differentiation [165,166], are part of the third wave of adaptation, which is enriched in “Th17 cell differentiation” (Figure 36D,E).

In summary, a range of Treg stages were captured and these cells were ordered across lymphoid and non-lymphoid tissues, establishing a continuous trajectory of recruitment and adaptation that is consistent between NLTs, and characterized by a core set of genes. Furthermore, from their conserved sequential order of activation, it can be concluded that gene activation leading to NLT adaptation follows a similar regulatory sequence in both bLN-to-skin and mLN-to-colon trajectories.

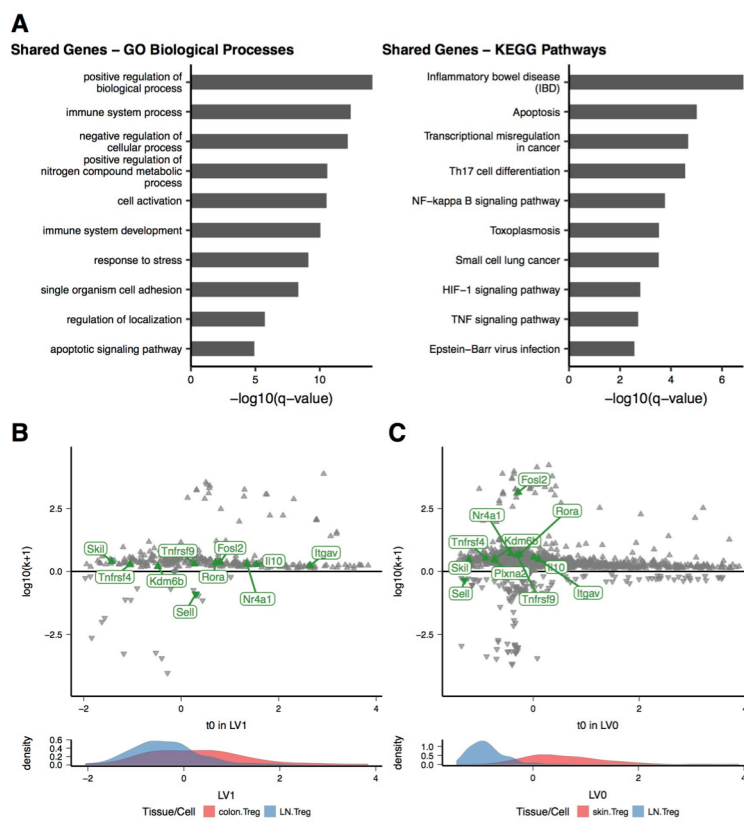


Figure 35 - Characterizing core genes in steady-state adaptation to colon and skin.

(A) Enriched GO Biological Processes (left) and KEGG Pathways (right) in common between the colon and skin migration/adaptation gradient.

(B and C) Time of activation (t_0) and intensity (k) of all genes identified as DE in the Colon (B) or Skin (C) trajectories (related to Figure 4B and 4E).

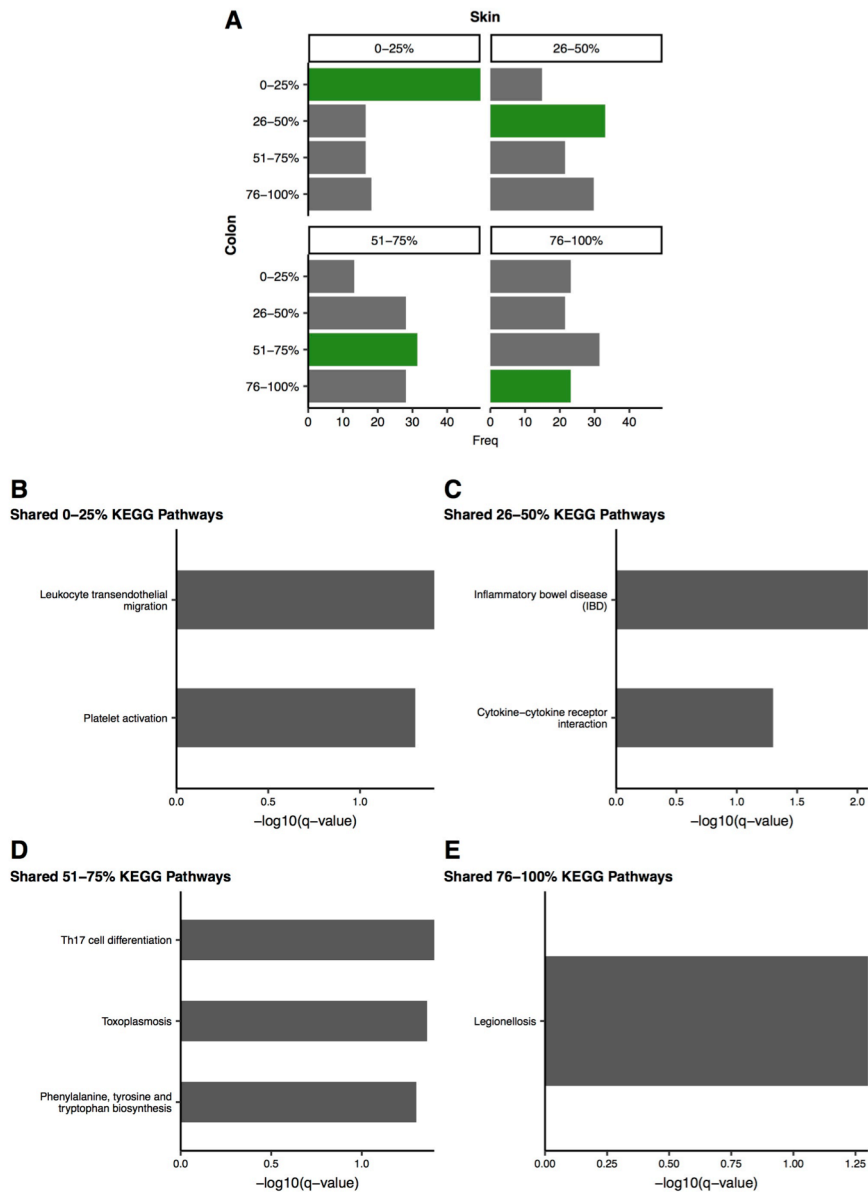


Figure 36 - Examining the migration programme shared between the colon and skin datasets.

(A) Comparing order of activation between Colon and Skin trajectories. Genes DE in both trajectories were grouped by the quartile of their t_0 value in each trajectory, and each of the four groups was intersected between datasets. (B-E) KEGG Pathways enriched in shared gene groups in each quartile

IV.2.5 Recruitment of Tregs to steady-state and challenged non-lymphoid tissues rely on shared mechanisms

The analyses above predicted the existence of a continuous sequence of stages of gene regulation in Tregs trafficking and adapting from lymph nodes to skin and colon in immune homeostasis. To validate these phenotypes, I sought to characterise the transcriptomes of Tregs in a process where dynamic recruitment of cells is known to occur between lymph nodes and skin. Namely, the system chosen was tumour Treg recruitment. Tumours are known to harbour several populations of lymphocytes, including Tregs. Based on the human TCR repertoire [167,168], previous studies have shown that tumour-Treg are likely to be recruited *de novo* from lymphoid tissues and not from the adjacent NLT, despite exhibiting a phenotype similar to that of NLT Treg [167].

CD4⁺ T cells were obtained from Foxp3-IRES-eGFP reporter mouse [111] 11 days after injection with B16.F10 melanoma cell line or PBS (Figure 38A; see Methods). Skin and tumour Treg cluster separately (Figure 38B), providing transcriptomic evidence of differences between the two conditions. In contrast to the steady-state, there is a cluster of cycling cells in the cells from tumour-injected mice (Supplementary Figure 1C). In addition, tumour-injected mice show clonotype sharing between tumour Treg and bLN Treg (Supplementary Figure 2C). Cycling Tregs on tumour and respective draining-LN, in parallel with clonotype sharing between both locations, suggest *de novo* Treg recruitment from LN (as seen in human [167]) and concomitant expansion in both tumour and draining-LN.

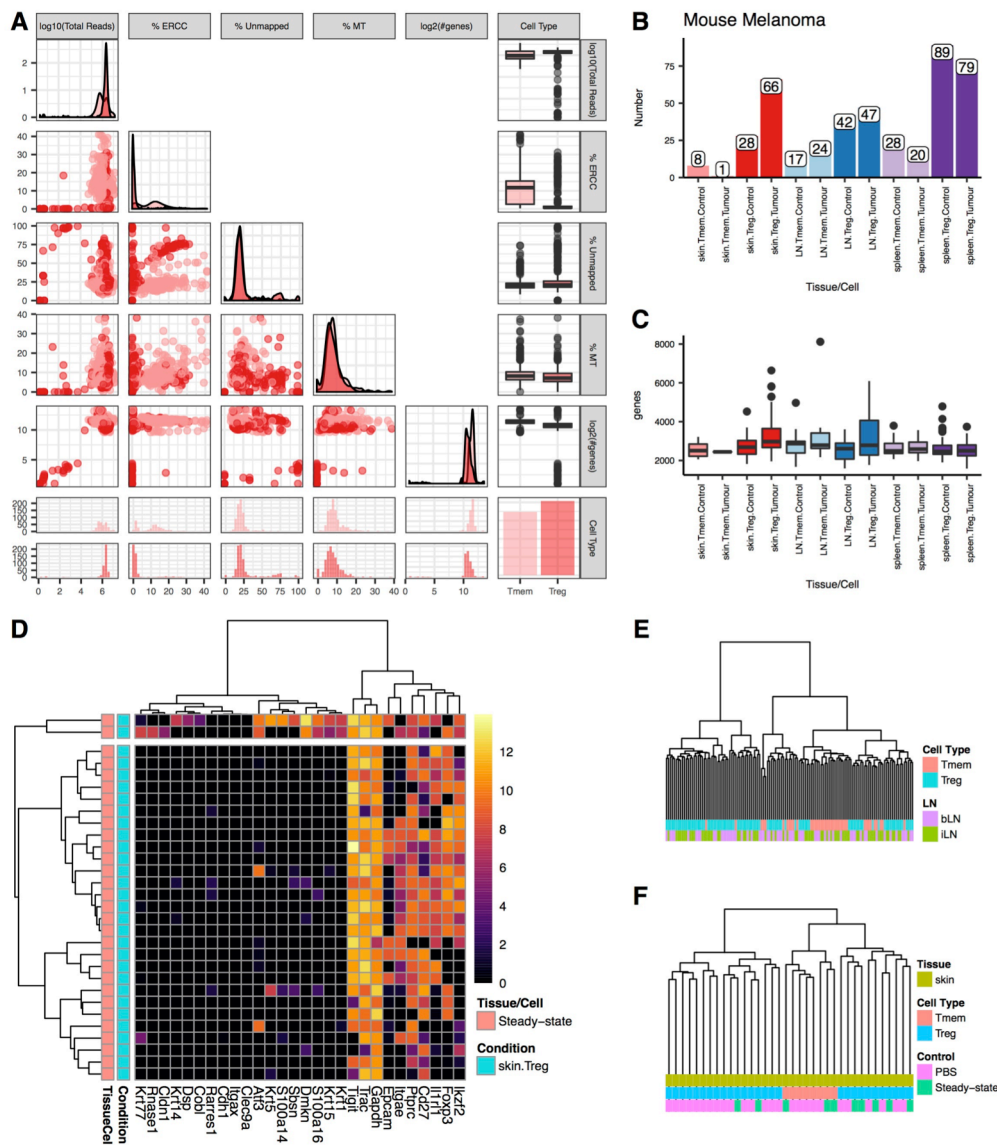


Figure 37 - Quality-control of mouse melanoma scRNA-seq dataset.

- (A) Multiple QC metrics and their relationships within the melanoma dataset. Measurements for individual cells are depicted in scatterplots. The respective distributions are represented as density plots (diagonal), and histograms (bottom). Distribution per cell type is represented in boxplots (right). Treg and Tmem are marked as dark and light red, respectively.
- (B) Number of single-cells from the melanoma dataset retained after QC.
- (C) Number of detected genes per cell type, tissue and condition.
- (D) Hierarchical clustering of steady-state skin Treg based on general Treg and keratinocyte markers. The horizontal division separates the two cells that were removed from the dataset for containing keratinocyte markers, suspected of being doublets.
- (E) Hierarchical clustering of cell-to-cell Spearman correlations across brachial and inguinal lymph nodes.
- (F) Hierarchical clustering of cell-to-cell Spearman correlations across steady-state and PBS injected skin.

Differential expression analysis between non-cycling tumour Tregs and control skin Tregs reveals a relatively small number of genes are significantly different between the two populations of Treg (112 upregulated in tumour Treg and 37 in steady-state skin Treg (Figure 38C), in line with recently

published human data [167]. Unsurprisingly, tumour Tregs upregulate genes related to proteasome activity, along with other pathways indicative of adhesion and migration (Figure 39D). Specifically, they upregulate the exhaustion marker *Lag3* [169], as well as *Cxcr3*, *Nkg7*, *Lgals1*, *Ccl5*, and some LN markers, such as *Ltb*. Control skin Tregs, on the other hand, show upregulation of skin Treg markers such as *Il1rl1*, *Rora*, *Pim1*, *Sdc4*, *Kdm6b*, *Jmjd1c* and *Erdr1*. Nonetheless, tumour and skin Treg are overall very similar in terms of recruitment and NLT adaptation. Skin Treg signature genes such as *Batf*, *Tnfrsf4*, *Tnfrsf9*, *Samsn1*, *Tigit*, *Tchp*, *Ccr8*, *Ccr2* and *Itgav* are expressed at equivalent levels in tumour Treg.

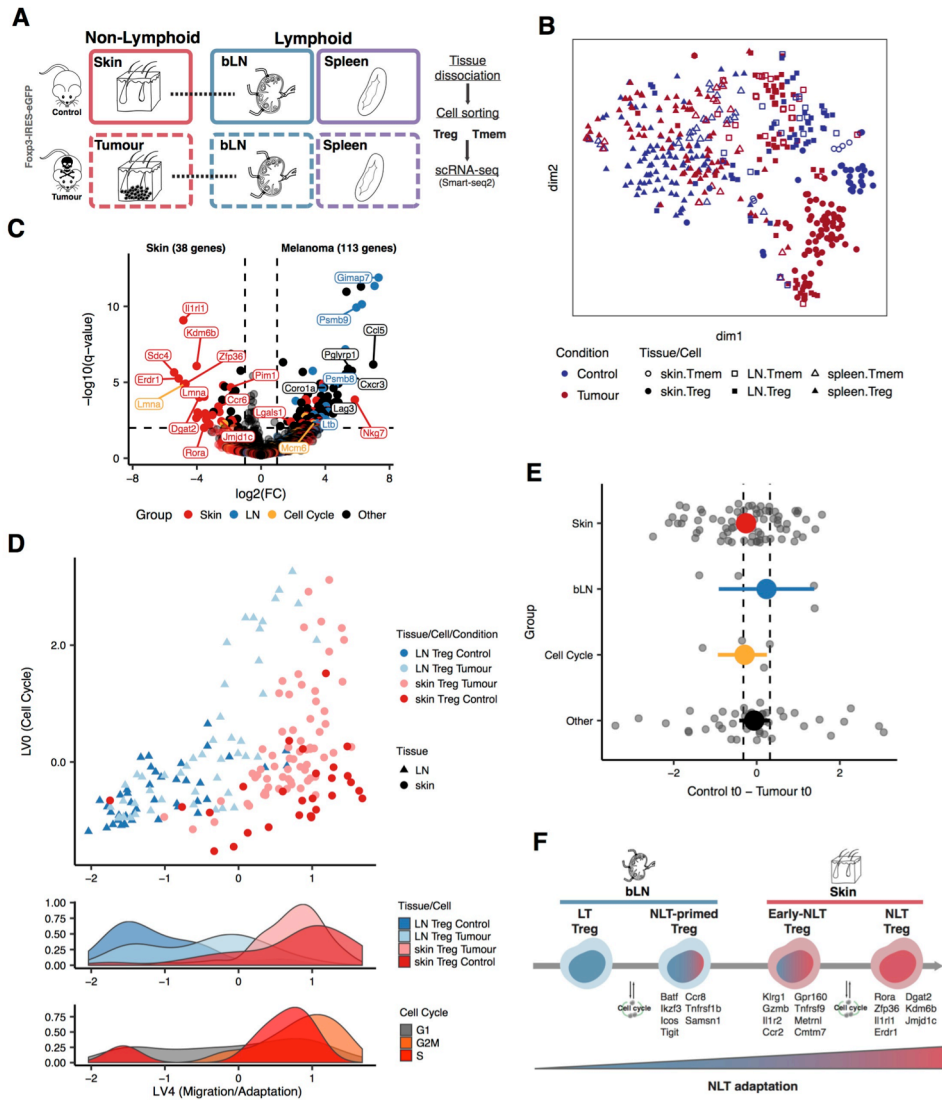


Figure 38 - Recruitment and adaptation of Treg to the tumour environment recapitulates steady-state migration.

(A) Melanoma induction strategy and tissues sampled for different cell types (similar to Figure 1A, see Methods).
 (B) t-SNE dimensionality reduction of single-cell data from control and tumour conditions, depicting Treg and Tmem steady-state skin and tumour, draining brachial lymph nodes and spleen. Treg and Tmem are represented by filled and open symbols, respectively. Colours match condition: control in dark blue, melanoma in dark red. Symbols match tissue: non-lymphoid tissues in circles, lymph nodes in squares, spleen in triangles.
 (C) Differential expression between skin and tumour (non-cycling) Treg ($q\text{-value} < 0.01$ and $\log_2(\text{FC}) > 1$).
 (D) (top) Latent variables found with BGPLVM representing cell cycle and non-lymphoid tissue recruitment/adaptation of Treg (see Methods). Tissue of origin matches colour: skin in red, lymph node in blue. Condition matches shade: control cells darker, tumour cells lighter. (bottom) Distribution of cells in different categories (Tissue and Condition, Cell Cycle phase) along the recruitment trajectory.
 (E) Comparison between time of activation of genes (t_0) in control and tumour, measured as the difference of t_0 between conditions. Genes are classified as being markers of skin, lymph node, cell cycle or other. Coloured points show mean \pm mean standard error. Vertical dashed lines represent \pm difference between mean t_0 in each condition.
 (F) Proposed model of gene expression changes in NLT adaptation of Tregs. At each stage examples of genes activated are listed below.

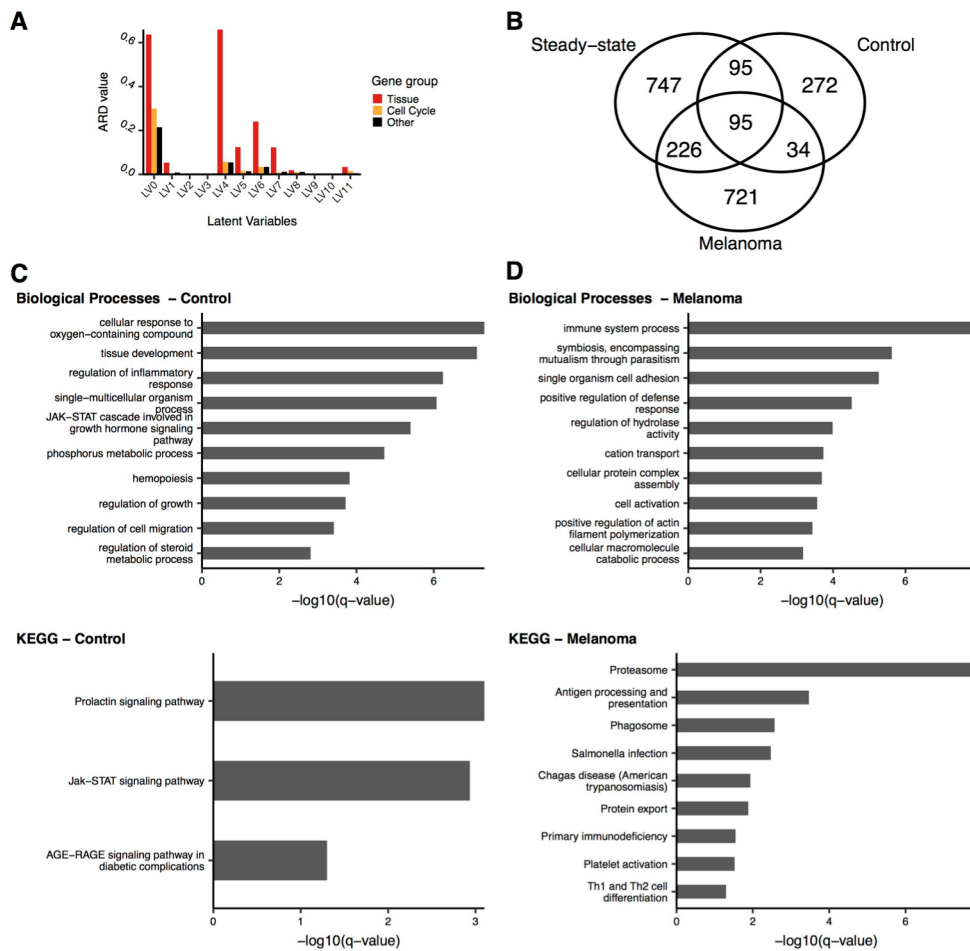


Figure 39 - Control and Melanoma response in the Tumour migration trajectory.

(A) Automatic Relevance Determination (ARD) plots for MRD-BGPLVM of Treg in control and melanoma conditions. Colours show effect of gene groups in each obtained latent variable.

(B) Venn diagram comparing obtained DE genes in the control and melanoma common trajectory, as well as the Skin steady-state trajectory previously obtained.

(C and D) GO Biological Processes (top) and KEGG Pathways (bottom) enriched in genes DE in Control (C) or Melanoma (D) cells along the common trajectory.

Then, a shared migration trajectory between control and melanoma-exposed cells was determined. To this end, MRD-BGPLVM [120] (see Methods) was used to explore gene expression trends across Treg from the control skin, tumour and respective draining-LNs together. Two main latent variables were identified, one concentrating most of cell cycle-associated variability (LV0), and one mainly associated with the NLT signature (LV4) (Figure 39A). Notably, NLT adaptation trajectory (LV4) correlates with trajectories found in control and melanoma conditions when MRD-BGPLVM is applied to each one individually (respectively, 0.63 and 0.45 Spearman correlation coefficients; Figure 40A-D, see Methods). Moreover, it also correlates with the NLT adaptation pseudotime previously determined for steady-state skin (0.35 Spearman correlation coefficient, Figure 40E). These observations support that

changes in Treg gene expression leading to NLT adaptation along bLN-to-skin and bLN-to-tumour trajectories are equivalent processes.

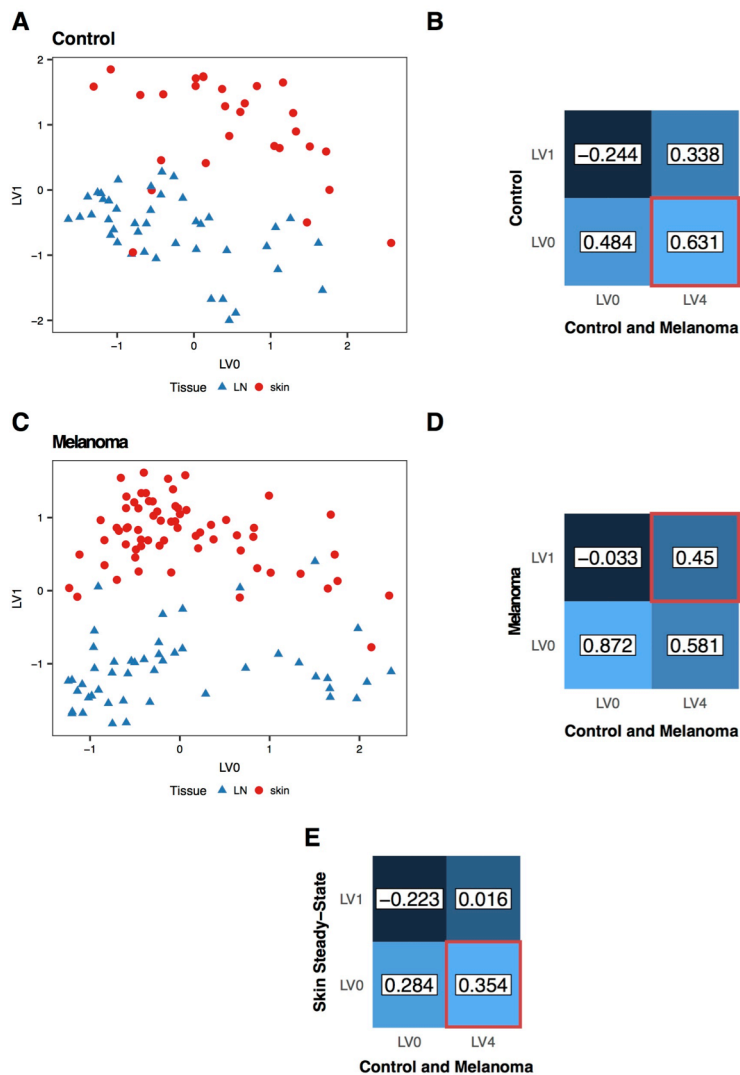


Figure 40 - The migration effect detected using MRD is present in both conditions individually.

(A and C) BGPLVM projection of bLN and skin in control (A) and melanoma (D) conditions, using the top two latent variables.

(B) Correlation between latent variables using the correlations of all genes with each latent variable. X axis - Control and Melanoma Treg latent variables; Y axis - Control Treg latent variables.

(D) Correlation between latent variables using the correlations of all genes with each latent variable. X axis - Control and Melanoma Treg latent variables; Y axis - Melanoma Treg latent variables.

(E) Correlation between latent variables using the correlations of all genes with each latent variable. X axis - Control and Melanoma Treg latent variables; Y axis - Steady-state Skin dataset Treg latent variables.

Gene kinetics along NLT adaptation (LV4) was investigated separately in control and melanoma conditions (Figure 41). 129 genes are shared between both conditions, 73% of which were also present in the steady-state skin trajectory determined previously (Figure 39B). As expected, values of t_0 remain largely unchanged between control and melanoma (Figure 38E), further suggesting that

NLT recruitment and adaptation follow the same program in homeostatic and perturbed conditions. The shared adaptation genes between control and melanoma include transcription factors (*Rora*, *Ikzf2*, *Ikzf3*, *Id2*, *Nr3c1*, *Batf*, *Nfil3*, *Mxd1*), migration and adhesion-related proteins (*Ccr2*, *Ccr8*, *Itgav*, *Plxna2*, *Iqgap1*), ligands and receptors (*Il1r2*, *Klrg1*, *Icos*, *Tigit*, *Ctla4*, *Tnfrsf9*, *Tnfrsf4*, *Tnfrsf18*, *Cmtm7*), secreted factors (*Lgals1*, *Metrn*), and the NF- κ B modulator *Bcl3*.

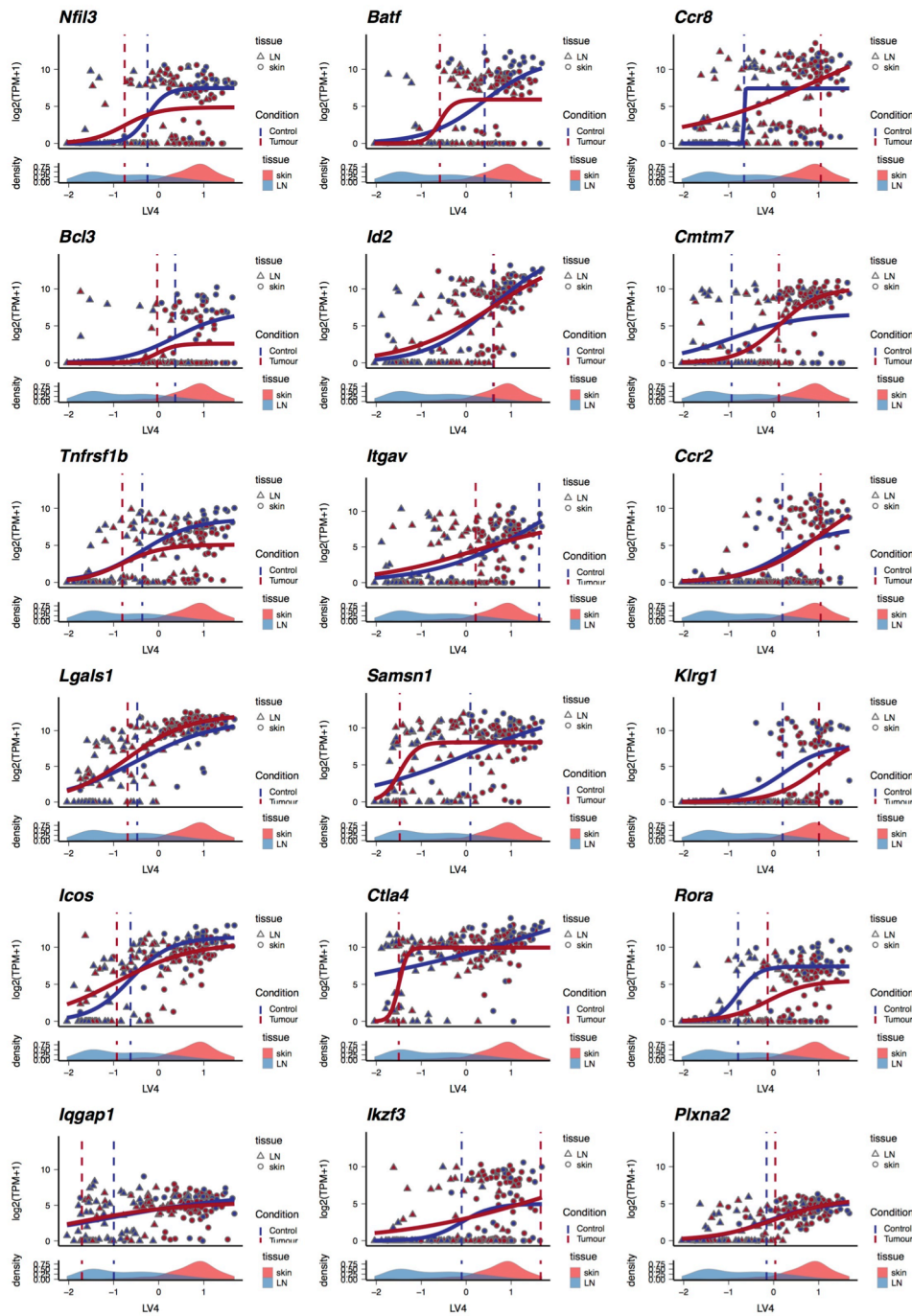


Figure 41 - Examples of individual genes varying in the control and melanoma migration trajectory.

Vertical dashed lines indicate time of activation (t_0) values determined for each condition.

Despite the similarities between melanoma and control, the distributions of cells from both conditions are not completely overlapping along NLT adaptation trajectory, and can be ordered by increasing NLT adaptation between populations (from least to most adapted: control LN Treg, melanoma LN Treg tumour Treg, and control skin Treg) (Figure 38D). This implies that in response to an ongoing challenge in the peripheral tissue, a higher fraction of Tregs in the LNs is acquiring NLT adaptations. In fact, for several NLT markers there are more cells expressing them in the tumour-draining LN compared to the control, *e.g.* *Id2* (53% vs 24%), *Batf* (49% vs 17%), *Nfil3* (43% vs 21%), *Lgals1* (85% vs 57%), *Ccr8* (72% vs 45%). This further supports the notion of priming of Treg to NLTs happening while still in the LN. Contrary to tumour-Treg, LN Tregs do not show signs of exhaustion (*Lag3* expression in Figure 38C, and data not shown), which strongly suggests that the “primed” LN Tregs do not represent instances of migration from tumour to the LN.

Finally, and as suggested by DE analysis (Figure 38C), tumour Tregs have the most cells with NLT adaptation, which brings them very close to the control skin Tregs.

Overall, Tregs from challenged mice recapitulate and fill in the gaps along steady-state NLT adaptation. This trajectory is summarized in (Figure 38F), where the stages of adaptation and key markers for each stage are shown.

IV.2.6 Evolutionary conservation between mouse and human NLT Treg cells

Insights into human NLT lymphocytes are of critical importance in the development of future therapies. To address this need, the extensive characterization of murine NLT CD4⁺ Treg and Tmem was complemented with a comparison to their human counterparts. Treg, Tcm and Tem cells were collected from blood and skin of individuals undergoing mammary reduction surgical interventions, and also from tumour-adjacent colon sections from patients undergoing colonic resection interventions (Figure 42A, Figure 43). Similarly to the mouse analysis, gene markers for human CD4⁺ T cell populations were determined (Figure 44; see Methods). Tcm and Tem cells were considered together as Tmem.

In terms of one-to-one orthologs, 86 out of 350 human skin Treg markers and 37 out of 214 human colon Treg markers overlap with the respective mouse signature. Interestingly, the general NLT signature common to both organs includes human/mouse conserved TNF-pathway receptors (*Tnfrsf1b*, *Tnfrsf4*, *Tnfrsf18*) and a regulator (*Tank*), other membrane proteins (*Ctla4*, *Icos*, and *Cd44*), and transcriptions factors (*Maf* and *Fosl2*). In tissue-specific terms, *Rora*, *Prdm1*, *Batf*, *Zfp36*, *Ccr6* and *Lgals1* in skin Treg, and *Skil*, *Ccr8*, *Cxcr6* and *Gpr183* in colon Treg are markers conserved across the two organisms (Figure 42B-E).

Close inspection of NLT markers obtained for both species revealed several instances where the expression pattern of one gene is taken by its paralog in the other organism (Figure 42F). For example, while the kinase *Pim1* is a marker of mouse NLT Tregs, and is not expressed in human, *Pim3* is not expressed in mouse, and marks human NLT Treg. A similar situation was observed for *Ccr2/Ccr4*, *Traf1/Traf3* and others. This suggests that some paralogous proteins have evolved to substitute for each other during the evolution of NLT Tregs in mammals.

Overall, this comparison suggests that despite mouse-human differences, the NLT program defined in mouse can at least in part be transposed to a human context.

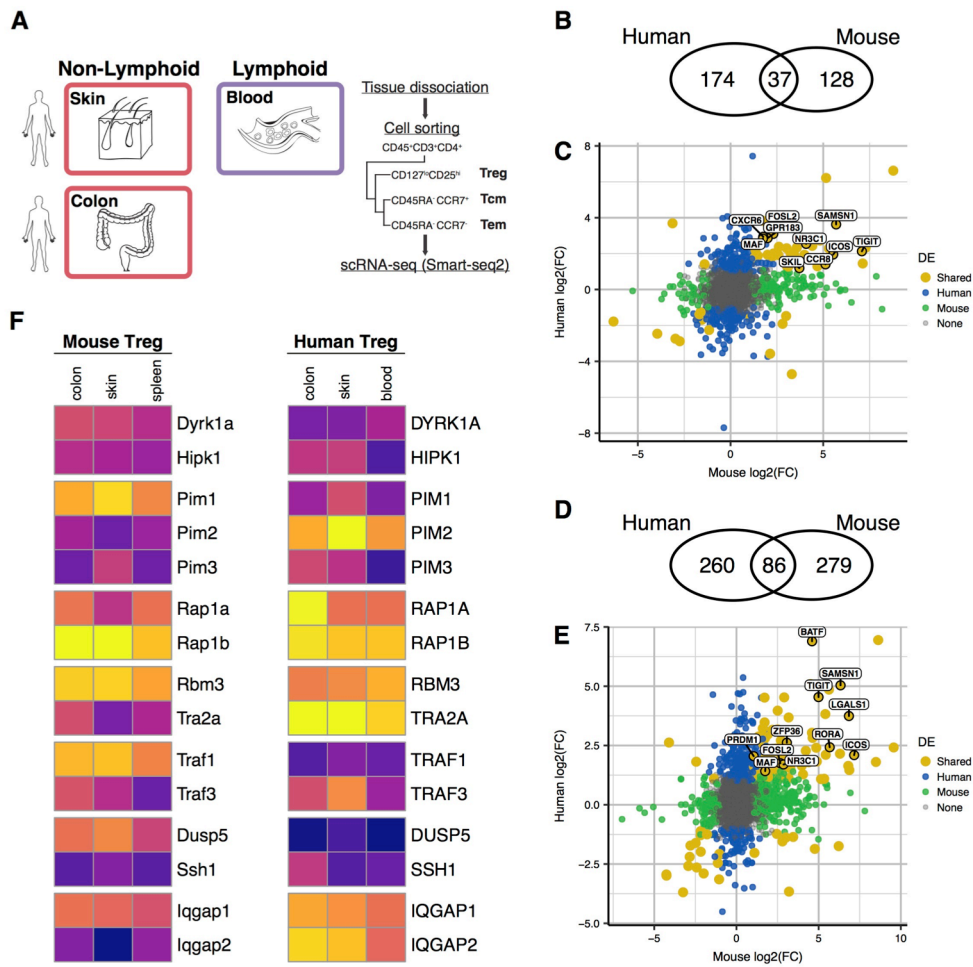


Figure 42 - Human-mouse comparison of non-lymphoid tissue Treg marker genes.

(A) Tissues sampled for different cell-types from human individuals (similar to Figure 1A, see Methods).

(B and D) Overlap between non-lymphoid tissue Treg markers detected in human and mouse, in either (B) colon or (D) skin datasets.

(C and E) Fold-change between gene expression in non-lymphoid and lymphoid tissues in mouse and human. Blood and spleen were used as lymphoid tissues in human and mouse respectively.

(F) Non-lymphoid tissue paralogs exhibiting opposing expression patterns between human and mouse.

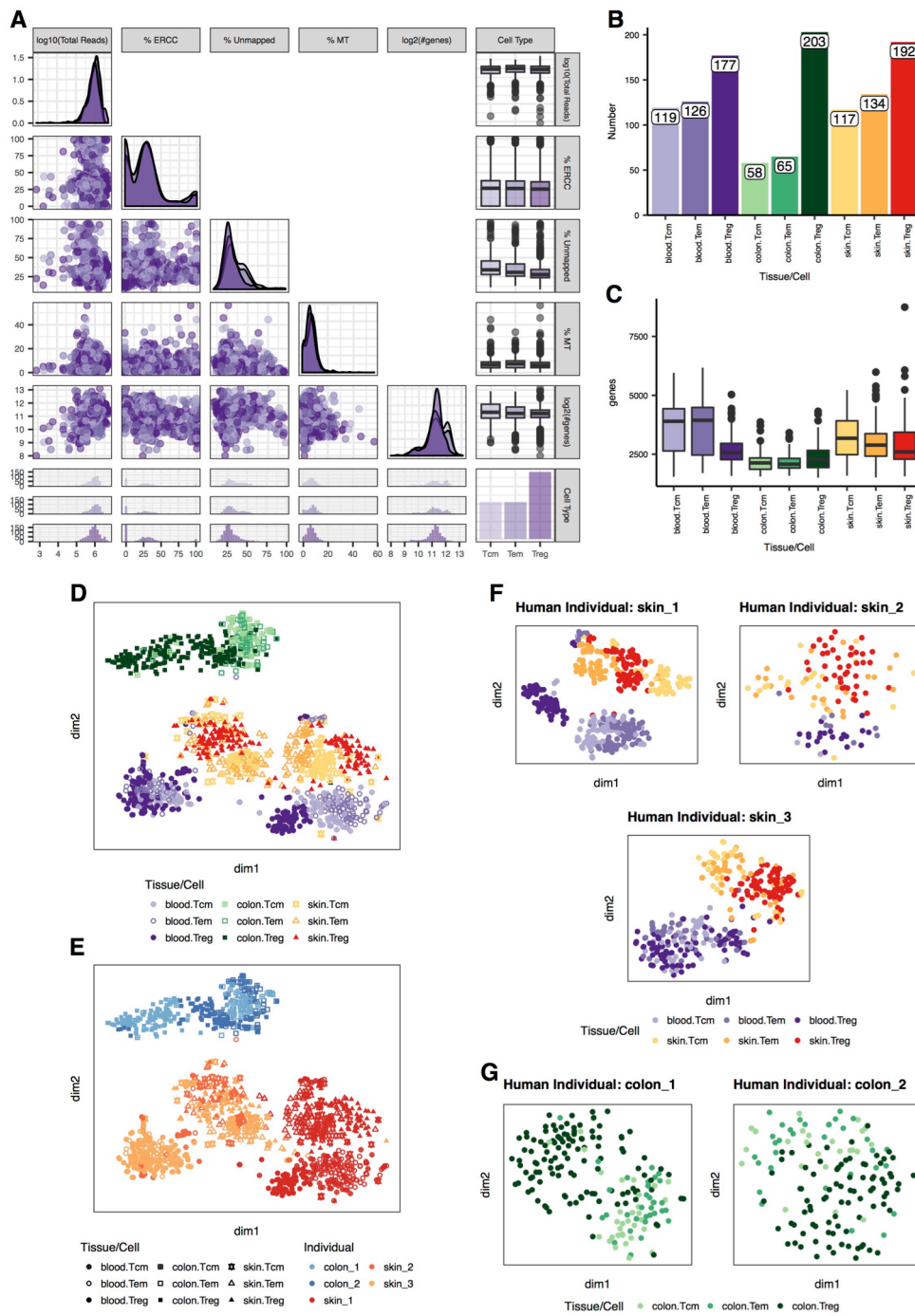


Figure 43 - Quality-control of steady-state human colon and skin scRNA-seq.

(A) Multiple QC metrics and their relationships in all human samples. Measurements for individual cells are depicted in scatterplots. The respective distributions are represented as density plots (diagonal), and histograms (bottom). Average per cell type is represented in boxplots (right). Treg, Tem and Tcm are marked in shades of purple.

(B) Number of single-cells from retained after QC.

(C) Number of genes per cell type and tissue.

(D and E) t-SNE dimensionality reduction. Shapes match cell type and tissue according to legend. Colours match either cell type and tissue (D) or sampled individual (E).

(F and G) t-SNE dimensionality reduction per individual, with cells coloured by tissue and cell type.

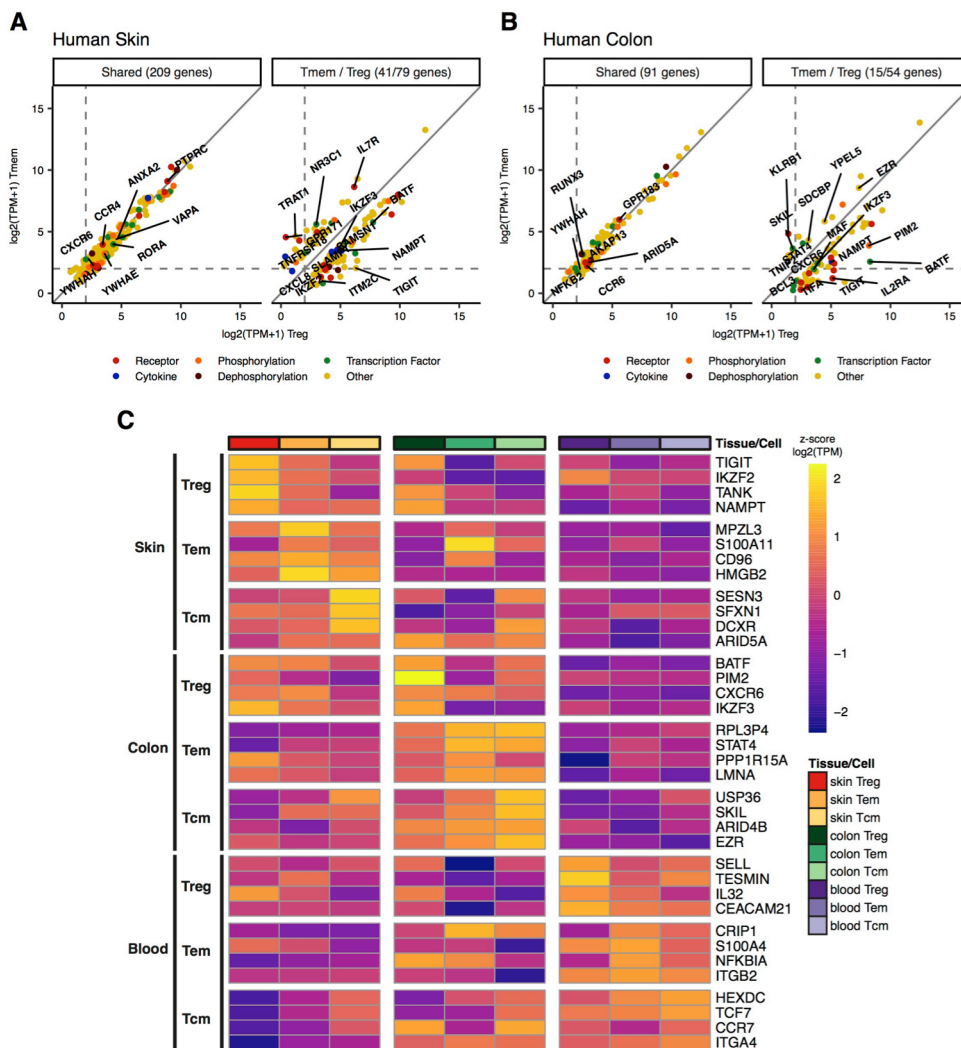


Figure 44 - NLT CD4⁺ T cell markers in human colon and skin scRNA-seq datasets.

(A and B) Differential expression of NLT markers between NLT Treg and Tmem (grouping together Tcm and Tem), in skin (A) and colon (B). Per dataset, genes were first identified as NLT- or blood-associated ($q\text{-value} \leq 0.01$, $\log_2 \text{fold-change (FC)} \geq 1$). This subset of genes was then characterized as Treg, Tmem or shared between cell types based on $q\text{-value} \leq 0.05$ and $\log_2 \text{FC} \geq 1$ (see Methods).

(C) Z-score of mean expression levels of identified markers across all sampled cell types and tissues in human.

IV.3 Discussion

IV.3.1 Phenotype of NLT CD4⁺ T cells as a set of tissue and cell type-specific modules

Until recently, the protective role of T cells was believed to be supported mostly by populations of cells recirculating between secondary lymphoid organs. Currently, not only is there evidence that T cells can reside in non-lymphoid tissues, they have been shown to play fundamental roles in the local immune response. Moreover, non-immune functions have been attributed to some of these populations, namely tissue resident Tregs (section 1.3.2). Although phenotypic differences between non-lymphoid and lymphoid tissue populations have been reported, they were not systematically or comprehensively studied before.

To address this knowledge gap, I used scRNA-seq data from steady-state CD4⁺ T cells, *i.e.* Treg and Tmem, from lymphoid and non-lymphoid tissues. My first objective was to determine the sets of specific genes that characterize each T cell compartment in comparison to the remaining ones, with particular focus on the non-lymphoid tissues skin and colon. The strategy adopted involved a “two-step differential expression”, first comparing tissues and then cell types within each tissue, allowing me to classify genes as Treg, Tmem or shared across both cell types in each given tissue.

I could then “order” gene sets from the most general to the most specific in terms of NLT adaptation. First, I could identify genes that constitute fundamental traits shared across Treg and Tmem, in the skin and colon compartments, *e.g.* *Maf*, *Id2*, *Kdm6b*, *Samsn1*. The next level of NLT adaptation to consider is cell type-specific but still shared across skin and colon, *e.g.* *Tnfrsf4* or *Gzmb*, that are functionally important for the general role of Treg in both NLTs. Finally, the most specific markers are unique to one of the cell types in either skin or colon, like *Batf* in skin Treg. General factors are thus likely to work as a basic adaptation of CD4⁺ T cells to the non-lymphoid environment, on top of which additional functional, migratory and survival features are added to attain the complete skin and colon-specific phenotypes. Cell surface proteins are a good example of this. Receptors such as *Itgav*, *Ccr2*, *Gpr183* are part of a general NLT Treg signature, and appear to set the stage for migration and retention in NLTs. A second level of receptors, *e.g.* *Gpr15* [150] in colon, *Ccr6* in skin, may determine tissue-specific Treg retention.

In future attempts to differentiate and stabilize NLT CD4⁺ T cells, general NLT genes should be upregulated. This also means that for therapies that aim at targeting specific NLT compartments,

these might not be good choices, as one would expect a general effect across all NLT CD4⁺ T cell populations

Although useful for characterization, statistic tests used to assign genes to different categories always rely on arbitrary thresholds, namely in terms of fold-change, which can oversimplify the biological processes under study. This seems to be the case with several of the marker genes found (Figure 28A,B, Figure 30A), as many of them are expressed in both tissues or both cell types, although to different extents. These observations suggest that such genes are likely to play roles in more than one condition, which prompted me to address gene relationships rather than gene expression alone to complement this characterization (Figure 28D, Figure 30B).

Lastly, I was not able to find clear subpopulations of Treg cells, which were expected particularly in the colon. While for effector and central Tregs one could argue that they are confounded with tissue of origin, *i.e.* central Treg in lymphoid tissues, effector Treg in the NLTs, the same is not true for the microbiota-dependent, peripherally-derived Treg Ror- γ ⁺ (Rorc⁺) and thymic-derived Treg Gata3⁺ colonic Treg. The fact that the majority of colonic Tregs captured do express high levels of *Gata3* and very little *Rorc* (data not shown) suggests that not many Ror- γ ⁺ cells have been sorted, which is not surprising as only 20% of colonic Treg should express it in the first place [47]. Furthermore, the animal facility where these mice were kept is known to be extremely clean (*e.g.* registering difficulties in establishing models of spontaneous inflammatory bowel disease), which might further contribute to low numbers of Ror- γ ⁺ Treg.

The systematic characterization of both Tmem and Treg cells across lymphoid and non-lymphoid tissues presented in this Chapter has major biological interest in understanding NLT adaptations, and could ultimately help harness NLT CD4⁺ T cells for future therapeutic applications.

IV.3.2 Consistency in the recruitment and adaptation of Tregs to NLTs

I used steady-state NLTs in mice to build high-resolution transcriptomic trajectories of recruitment and adaptation of Tregs from lymph nodes to NLTs. This modelling approach revealed the continuous nature of recruitment and adaptation to NLTs, with no clearly defined intermediate stages.

The core genes shared by trajectories, as well as their order of transcriptional activation and repression are similar between skin and colon, particularly for the first half of the trajectory (Figure 36).

Presumably, the first half of the process is related to triggering of recruitment and general NLT adaptations, while the differences in the later stage might reflect the bias that seems to exist towards activation or full-differentiation between colon and skin Treg, respectively.

The trajectories of LT-to-NLT trafficking are supported by clonotype sharing between cells, which were inferred by reconstruction the TCR α and β chains from the full-length mRNA transcriptome data in each cell. Although the number of Treg sharing both chains across tissues is limited (only one case between bLN and skin), it evidences that one can capture “sister-cells” in different stages of NLT adaptation across tissues, which is further supported by multiple events observed within Tmems. On the other hand, the expanded clonotypes are almost exclusively within a cell type (e.g. within Tregs or within Tmems) rather than shared across Tregs and Tmems (Supplementary Figure 2), suggesting that these are distinct, committed cell fates *in vivo* rather than plastic and interchangeable states.

IV.3.3 Tumour Treg adaptation follows the path of steady-state NLT Tregs

Perturbation often exacerbates and evidences the features of the system under study. Therefore, to further address the adaptation of Tregs to non-lymphoid environments, I chose to promote the development of a tumour, which has been suggested to induce *de novo* recruitment of Treg from draining-lymph nodes [167]. Indeed, the existence of cycling populations in both tumour and the draining-LN, in conjunction with clonotype-sharing between tumour and LN Tregs, strongly suggests *de novo* recruitment of Tregs to NLTs is taking place and has been captured in this dataset. Therefore, I have then used this melanoma-challenge to validate the steady-state observations and, on a second stage, to increase the resolution and understanding of this process.

I observed that not only the trajectory of adaptation of Treg to the tumour environment is well correlated with the steady-state bLN-to-skin trajectory (Figure 40), there is also a considerable proportion of this adaptation programme that is consistent with both the bLN-to-skin and mLN-to-colon trajectories (Figure 32C). These observations unify all three NLT adaptation programmes around a solid core set of genes.

Furthermore, analysing Treg cells from control and challenged mice together provided a more detailed perspective of this process. On one hand, I concluded that cues trigger transcriptomic changes in Tregs located in the draining-LNs, inducing transcriptomic changes that place them closer to the NLT-adapted phenotype, as indicated by higher percentage of cells expressing *Batf*, *Ccr8*, *Samsn1* and

other NLT markers in the melanoma condition. On the other hand, the NLT trajectory suggests that tumour Tregs are slightly less mature versions of NLT Tregs in immune homeostasis. This difference is likely related to the few genes (37) that are upregulated in skin Treg when compared to tumour Treg (Figure 38C), including NLT markers such as *Il1r1*, *Rora*, *Kdm6b*, and *Ccr6*.

It remains to be explained whether these genes are not expressed equally between conditions for lack of the correct environmental cues or because they only stabilize at a later stage of adaptation, and whether any (or several) of these differentially expressed genes is responsible for the final NLT differentiation. Knock-out mice and/or genome editing techniques coupled with adoptive transfers would be of great importance in answering these questions.

IV.3.4 NLT adaptation from mouse to human

The NLT compartment of CD4⁺ T cells has remained more obscure in human than in mouse due to limitations in access to tissues. Therefore, the scRNA-seq data I collected from human skin, colon and blood constitutes, in and of itself, an important resource to understand different aspects of NLT CD4⁺ T cells.

Overall, these human cell populations show conservation in structure, *i.e.*, as in mouse, there is a clear separation between cells from lymphoid and non-lymphoid tissues. Not only this, but several of the genes driving this separation are common between human and mouse (Figure 42B-E). Interestingly, I observed changes in the expression patterns of specific paralog genes between species, in particular in the PIM and the TRAF family (Figure 42F). The usage of different paralogs for the same function suggests an important role of expanded gene families in rewiring signaling pathways throughout evolution.

Reconstruction of recruitment trajectories in the human data was not attempted, mainly because no lymph node data was available. Blood-to-NLT trajectories would most likely be extremely confounded by T cells activated and primed in lymph nodes all across the organism.

Although clonotype sharing analysis show that memory and regulatory cell types can share TCRs, in contrast with mouse, these events are likely artefacts of the sorting strategy, that allows for Treg to be sorted as Tmem and vice-versa.

In sum, my observations suggest that several of the key targets identified in mouse are likely to play important roles in humans, and can therefore be considered as potential targets in future therapeutic approaches. The analyses performed on the human datasets focused mainly on validating the

presence of markers found in mouse NLT Treg, overlooking other interesting questions, such as an in-depth evolutionary comparison.

In this study, I have used genomic data to model the process of recruitment and adaptation of regulatory T cells from secondary lymphoid organs to non-lymphoid tissues. The transcriptomic changes associated with this transition form a continuous trajectory that starts in the draining-lymph nodes and continues within the NLTs, both in steady-state and immune challenge conditions.

Overall, these results reveal a dynamic adaptation of T cells as they traffic from one tissue to another, and provides a broad resource for investigating *in vivo* CD4⁺ T cell phenotypes in mouse and human.

CHAPTER V

Conclusions and Future Work

CHAPTER V - Conclusions and Future Work

In this thesis, I set myself to capitalize on state-of-the-art scRNA-seq methods to characterize elusive cell types and processes in immune tolerance. The results presented herein contribute to a better understanding of these populations of cells and their dynamics of development and function, which had been concealed until now. Single-cell resolution was instrumental in addressing these two systems, interestingly, in considerable different ways. While it allowed for the direct identification of discrete subpopulations during mTEC development, it could also be leveraged to interpret the continuous gradients underlying recruitment and adaptation of Tregs to non-lymphoid tissues.

In the work described in Chapter III, I identified three subpopulations of medullary thymic epithelial cells: jTEC, mTEChi and mTEClo. To the best of my knowledge, this is the first time that these populations have been identified in an unbiased way. Importantly, single-cell resolution allowed me to isolate mTEClo, a subpopulation generally dismissed and/or heavily biased towards late post-*Aire* stage, and clarify its role as an important and active piece in negative selection.

Chapter IV introduces a model for adaptation of Treg to non-lymphoid tissues, one of the major open questions in the field. The results shown suggest that this adaptation starts still in the lymph nodes, progressing as a continuous gradient within and across tissues. The comparisons made between tissues and cell types provide important information on how specific NLT markers are, important information for future studies and therapies.

Any scRNA-seq experiment is affected by technical drop-out events, *i.e.* genes which are expressed but not being detected. In the context of this thesis, such limitation can strongly impair the study of *Aire*-dependent and *Aire*-enhanced TRAs across mTECs, which are expressed in a low percentage of cells in the population. It is plausible that myself and others [28,123] have thus underestimated expression frequency of TRAs and overlooked small clusters of co-expressed TRAs.

This drop-out effect, coupled with post-transcriptional regulation and differences in the protein stability, leads to relatively low correlations between mRNA and protein levels. This discordance is likely to affect how intra-population heterogeneity is perceived, and might explain why reported heterogeneity within Treg populations using protein markers is not apparent at the transcriptome level.

The numbers of cells that can be processed and sequenced using scRNA-seq plate-based approaches, as done in this thesis, are situated in the hundreds of cells. This might not have been sufficient to

capture all the heterogeneity present within the populations I studied in this thesis, e.g. Ror- γ t⁺ colonic Treg. Droplet-based approaches to analyze tens of thousands of cells could be used to uncover lowly represented populations in future experiments.

Future work should also include additional validation experiments to further consolidate some of the reported observations. In the mTEC context, the most immediate path to follow includes FACS validation of marker genes identified at the mRNA level for jTEC, mTEChi, and most importantly, mTEClo. CHIP-seq and/or CRISPR would also be of interest to test to what extent the TFs identified as new regulators of mTEC development (e.g. *Vdr*, *Plagl1*, *Zbtb7a*, *Hnf4g*) are affecting the process. In the Treg context, adoptive transfer experiments of cells deficient for TFs and chromatin modifiers involved in NLT adaptation would greatly contribute to fully understand how this phenotype is established. Genes upregulated in skin and not in tumour (e.g. *Rora*, *Kdm6b*) would be an interesting set of genes to start from.

In conclusion, the work described in this thesis represents a significant step forward in discerning the highly complex and specific set of mechanisms that sustain immune tolerance, whilst also providing a resource for further characterization of mTEC and Treg populations.

References

1. Abbas AK, Lichtman AHH, Pillai S: *Cellular and Molecular Immunology*. Elsevier Health Sciences; 2007.
2. Coomes SM, Pelly VS, Wilson MS: **Plasticity within the $\alpha\beta$ CD4⁺ T-cell lineage: when, how and what for?**. *Open Biol.* 2013, **3**:120157.
3. Pepper M, Jenkins MK: **Origins of CD4⁺ effector and central memory T cells**. *Nat. Immunol.* 2011, **131**:467–471.
4. Schenkel JM, Masopust D: **Tissue-resident memory T cells**. *Immunity* 2014, **41**:886–897.
5. Mueller SN, Gebhardt T, Carbone FR, Heath WR: **Memory T Cell Subsets, Migration Patterns, and Tissue Residence**. *Annu. Rev. Immunol.* 2013, **31**:137–161.
6. Thome JJC, Bickham KL, Ohmura Y, Kubota M, Matsuoka N, Gordon C, Granot T, Griesemer A, Lerner H, Kato T, et al.: **Early-life compartmentalization of human T cell differentiation and regulatory function in mucosal and lymphoid tissues**. *Nat. Med.* 2015, **22**:72–77.
7. DuPage M, Bluestone JA: **Harnessing the plasticity of CD4(+) T cells to treat immune-mediated disease**. *Nat Rev Immunol* 2016, **16**:149–163.
8. Klein L, Kyewski B, Allen PM, Hogquist K a: **Positive and negative selection of the T cell repertoire: what thymocytes see (and don't see)**.. *Nat. Rev. Immunol.* 2014, **14**:377–391.
9. Moran AE, Hogquist KA: **T-cell receptor affinity in thymic development**. *Immunology* 2012, **135**:261–267.
10. Otero DC, Baker DP, David M: **IRF7-dependent IFN- β production in response to RANKL promotes medullary thymic epithelial cell development**.. *J. Immunol.* 2013, **190**:3289–98.
11. Hadeiba H, Lahl K, Edalati A, Oderup C, Habtezion A, Pachynski R, Nguyen L, Ghodsi A, Adler S, Butcher EC: **Plasmacytoid dendritic cells transport peripheral antigens to the thymus to promote central tolerance**.. *Immunity* 2012, **36**:438–50.
12. Aschenbrenner K, D'Cruz LM, Vollmann EH, Hinterberger M, Emmerich J, Swee LK, Rolink A, Klein L: **Selection of Foxp3⁺ regulatory T cells specific for self antigen expressed and presented by Aire⁺ medullary thymic epithelial cells**.. *Nat. Immunol.* 2007, **8**:351–8.

13. Yang S, Fujikado N, Kolodin D, Benoist C, Mathis D: **Immune tolerance. Regulatory T cells generated early in life play a distinct role in maintaining self-tolerance.** *Science (80-).* 2015, **348**:589–594.
14. Derbinski J, Pinto S, Rösch S, Hexel K, Kyewski B: **Promiscuous gene expression patterns in single medullary thymic epithelial cells argue for a stochastic mechanism..** *Proc. Natl. Acad. Sci. U. S. A.* 2008, **105**:657–62.
15. Org T, Rebane A, Kisand K, Laan M, Haljasorg U, Andreson R, Peterson P: **AIRE activated tissue specific genes have histone modifications associated with inactive chromatin..** *Hum. Mol. Genet.* 2009, **18**:4699–710.
16. Cloosen S, Arnold J, Thio M, Bos GMJ, Kyewski B, Germeraad WT V: **Expression of tumor-associated differentiation antigens, MUC1 glycoforms and CEA, in human thymic epithelial cells: implications for self-tolerance and tumor therapy.** *Cancer Res.* 2007, **67**:3919–3926.
17. Peterson P, Org T, Rebane A: **Transcriptional regulation by AIRE: molecular mechanisms of central tolerance..** *Nat. Rev. Immunol.* 2008, **8**:948–57.
18. Arstila TP, Jarva H: **Human APECED; a Sick Thymus Syndrome?.** *Front. Immunol.* 2013, **4**:313.
19. LaFlam TN, Seumois G, Miller CN, Lwin W, Fasano KJ, Waterfield M, Proekt I, Vijayanand P, Anderson MS: **Identification of a novel cis-regulatory element essential for immune tolerance.** *J. Exp. Med.* 2015, **212**:1993–2002.
20. Haljasorg U, Bichele R, Saare M, Guha M, Maslovskaja J, Kõnd K, Remm A, Pihlap M, Tomson L, Kisand K, et al.: **A highly conserved NF- κ B-responsive enhancer is critical for thymic expression of Aire in mice.** *Eur. J. Immunol.* 2015, **45**:3246–3256.
21. Guerau-de-Arellano M, Mathis D, Benoist C: **Transcriptional impact of Aire varies with cell type.** *Proc. Natl. Acad. Sci.* 2008, **105**:14011–14016.
22. Giraud M, Taubert R, Vandiedonck C, Ke X, Lévi-Strauss M, Pagani F, Baralle FE, Eymard B, Tranchant C, Gajdos P, et al.: **An IRF8-binding promoter variant and AIRE control CHRNA1 promiscuous expression in thymus.** *Nature* 2007, **448**:934–937.
23. Takahama Y, Ohigashi I, Baik S, Anderson G: **Generation of diversity in thymic epithelial cells.** *Nat. Rev. Immunol.* 2017, **17**:295–305.
24. Bansal K, Yoshida H, Benoist C, Mathis D: **The transcriptional regulator Aire binds to and activates super-enhancers.** *Nat. Immunol.* 2017, **18**:263–273.

25. Anderson MS, Venanzi ES, Klein L, Chen Z, Berzins SP, Turley SJ, von Boehmer H, Bronson R, Dierich A, Benoist C, et al.: **Projection of an immunological self shadow within the thymus by the aire protein..** *Science* 2002, **298**:1395–401.
26. Takaba H, Morishita Y, Tomofuji Y, Danks L, Nitta T, Komatsu N, Kodama T, Takayanagi H: **Fezf2 Orchestrates a Thymic Program of Self-Antigen Expression for Immune Tolerance.** *Cell* 2015, **163**:975–987.
27. Pinto S, Michel C, Schmidt-Glenewinkel H, Harder N, Rohr K, Wild S, Brors B, Kyewski B: **Overlapping gene coexpression patterns in human medullary thymic epithelial cells generate self-antigen diversity..** *Proc. Natl. Acad. Sci. U. S. A.* 2013, **110**:E3497-505.
28. Brennecke P, Reyes A, Pinto S, Rattay K, Nguyen M, Kuchler R, Huber W, Kyewski B, Steinmetz LM: **Single-cell transcriptome analysis reveals coordinated ectopic gene-expression patterns in medullary thymic epithelial cells.** *Nat. Immunol.* 2015, **16**:933–941.
29. Bleul CC, Corbeaux T, Reuter A, Fisch P, Mönning JS, Boehm T: **Formation of a functional thymus initiated by a postnatal epithelial progenitor cell.** *Nature* 2006, **441**:992–996.
30. Rossi SW, Jenkinson WE, Anderson G, Jenkinson EJ: **Clonal analysis reveals a common progenitor for thymic cortical and medullary epithelium.** *Nature* 2006, **441**:988–991.
31. Ohigashi I, Zuklys S, Sakata M, Mayer CE, Hamazaki Y, Minato N, Hollander GA, Takahama Y: **Adult Thymic Medullary Epithelium Is Maintained and Regenerated by Lineage-Restricted Cells Rather Than Bipotent Progenitors.** *Cell Rep.* 2015, **13**:1432–1443.
32. Onder L, Nindl V, Scandella E, Chai Q, Cheng H-W, Caviezel-Firner S, Novkovic M, Bomze D, Maier R, Mair F, et al.: **Alternative NF- κ B signaling regulates mTEC differentiation from podoplanin-expressing precursors in the cortico-medullary junction.** *Eur. J. Immunol.* 2015, **45**:2218–2231.
33. Metzger TC, Khan IS, Gardner JM, Mouchess ML, Johannes KP, Krawisz AK, Skrzypczynska KM, Anderson MS: **Lineage tracing and cell ablation identify a post-Aire-expressing thymic epithelial cell population.** *Cell Rep.* 2013, **5**:166–179.
34. Nishikawa Y, Hirota F, Yano M, Kitajima H, Miyazaki J-I, Kawamoto H, Mouri Y, Matsumoto M:

- Biphasic Aire expression in early embryos and in medullary thymic epithelial cells before end-stage terminal differentiation.** *J. Exp. Med.* 2010, **207**:963–971.
35. Hamazaki Y, Sekai M, Minato N: **Medullary thymic epithelial stem cells: role in thymic epithelial cell maintenance and thymic involution.** *Immunol. Rev.* 2016, **271**:38–55.
36. Rattay K, Meyer HV, Herrmann C, Brors B, Kyewski B: **Evolutionary conserved gene co-expression drives generation of self-antigen diversity in medullary thymic epithelial cells.** *J. Autoimmun.* 2016, **67**:65–75.
37. Anderson MS, Su MA: **AIRE expands: new roles in immune tolerance and beyond.** *Nat. Rev. Immunol.* 2016, **16**:247–258.
38. White AJ, Nakamura K, Jenkinson WE, Saini M, Sinclair C, Seddon B, Narendran P, Pfeffer K, Nitta T, Takahama Y, et al.: **Lymphotoxin signals from positively selected thymocytes regulate the terminal differentiation of medullary thymic epithelial cells.** *J Immunol* 2010, **185**:4769–4776.
39. Liston A, Piccirillo C a: *Developmental plasticity of murine and human Foxp3(+) regulatory T cells.* Elsevier Inc.; 2013.
40. Abbas AK, Benoist C, Bluestone J a, Campbell DJ, Ghosh S, Hori S, Jiang S, Kuchroo VK, Mathis D, Roncarolo MG, et al.: **Regulatory T cells: recommendations to simplify the nomenclature.** *Nat. Immunol.* 2013, **14**:307–8.
41. Coquet JM, Ribot JC, Bąbała N, Middendorp S, van der Horst G, Xiao Y, Neves JF, Fonseca-Pereira D, Jacobs H, Pennington DJ, et al.: **Epithelial and dendritic cells in the thymic medulla promote CD4+Foxp3+ regulatory T cell development via the CD27-CD70 pathway.** *J. Exp. Med.* 2013, **210**:715–28.
42. Ohkura N, Kitagawa Y, Sakaguchi S: **Development and maintenance of regulatory T cells.** *Immunity* 2013, **38**:414–23.
43. Povoleri GAM, Scottà C, Nova-Lamperti EA, John S, Lombardi G, Afzali B: **Thymic versus induced regulatory T cells - who regulates the regulators?.** *Front. Immunol.* 2013, **4**:169.
44. Liston A, Gray DHD: **Homeostatic control of regulatory T cell diversity.** *Nat. Rev. Immunol.* 2014, **14**:154–165.
45. Cipolletta D, Feuerer M, Li A, Kamei N, Lee J, Shoelson SE, Benoist C, Mathis D: **PPAR- γ is a major driver of the accumulation and phenotype of adipose tissue Treg**

- cells. *Nature* 2012, **486**:549–553.
46. Burzyn D, Kuswanto W, Kolodin D, Shadrach JL, Cerletti M, Jang Y, Sefik E, Tan TG, Wagers AJ, Benoist C, et al.: **A special population of regulatory T cells potentiates muscle repair.** *Cell* 2013, **155**:1282–1295.
 47. Schiering C, Krausgruber T, Chomka A, Fröhlich A, Adelman K, Wohlfert EA, Pott J, Griseri T, Bollrath J, Hegazy AN, et al.: **The alarmin IL-33 promotes regulatory T-cell function in the intestine.** *Nature* 2014, **513**:564–568.
 48. Panduro M, Benoist C, Mathis D: **Tissue Tregs.** *Annu. Rev. Immunol.* 2016, **34**:609–633.
 49. Hu Z-Q, Zhao W-H: **The IL-33/ST2 axis is specifically required for development of adipose tissue-resident regulatory T cells.** *Cell. Mol. Immunol.* 2015, **12**:521–524.
 50. Chandramouli K, Qian P-Y: **Proteomics: Challenges, Techniques and Possibilities to Overcome Biological Sample Complexity.** *Hum. Genomics Proteomics* 2009, **2009**:1–22.
 51. Schwanhäusser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, Chen W, Selbach M: **Global quantification of mammalian gene expression control.** *Nature* 2011, **473**:337–342.
 52. Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, Wang X, Bodeau J, Tuch BB, Siddiqui A, et al.: **mRNA-Seq whole-transcriptome analysis of a single cell.** *Nat. Methods* 2009, **6**:377–382.
 53. Kolodziejczyk AA, Kim JK, Svensson V, Marioni JC, Teichmann SA: **The Technology and Biology of Single-Cell RNA Sequencing.** *Mol. Cell* 2015, **58**:610–620.
 54. Picelli S, Faridani OR, Björklund AK, Winberg G, Sagasser S, Sandberg R: **Full-length RNA-seq from single cells using Smart-seq2.** *Nat. Protoc.* 2014, **9**:171–181.
 55. Macosko EZ, Basu A, Satija R, Nemes J, Shekhar K, Goldman M, Tirosh I, Bialas AR, Kamitaki N, Martersteck EM, et al.: **Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets.** *Cell* 2015, **161**:1202–1214.
 56. Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, Peshkin L, Weitz DA, Kirschner MW: **Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells.** *Cell* 2015, **161**:1187–1201.
 57. Macaulay IC, Voet T: **Single Cell Genomics: Advances and Future Perspectives.** *PLoS Genet.* 2014, **10**:e1004126.
 58. Macaulay IC, Haerty W, Kumar P, Li YI, Hu TX, Teng MJ, Goolam M, Saurat N, Coupland P,

- Shirley LM, et al.: **G&T-seq: parallel sequencing of single-cell genomes and transcriptomes.** *Nat Methods* 2015, **12**:519–522.
59. Dey SS, Kester L, Spanjaard B, Bienko M, van Oudenaarden A: **Integrated genome and transcriptome sequencing of the same cell.** *Nat Biotechnol* 2015, **33**:285–289.
60. Angermueller C, Clark SJ, Lee HJ, Macaulay IC, Teng MJ, Hu TX, Krueger F, Smallwood SA, Ponting CP, Voet T, et al.: **Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity..** *Nat. Methods* 2016, **13**:229–32.
61. Dixit A, Parnas O, Li B, Chen J, Fulco CP, Jerby-Arnon L, Marjanovic ND, Dionne D, Burks T, Raychowdhury R, et al.: **Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens.** *Cell* 2016, **167**:1853–1866.e17.
62. Jaitin DA, Weiner A, Yofe I, Lara-Astiaso D, Keren-Shaul H, David E, Salame TM, Tanay A, van Oudenaarden A, Amit I: **Dissecting Immune Circuits by Linking CRISPR-Pooled Screens with Single-Cell RNA-Seq.** *Cell* 2016, **167**:1883–1896.e15.
63. Adamson B, Norman TM, Jost M, Cho MY, Nuñez JK, Chen Y, Villalta JE, Gilbert LA, Horlbeck MA, Hein MY, et al.: **A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the Unfolded Protein Response.** *Cell* 2016, **167**:1867–1882.e21.
64. Svensson V, Natarajan KN, Ly L-H, Miragaia RJ, Labalette C, Macaulay IC, Cvejic A, Teichmann SA: **Power analysis of single-cell RNA-sequencing experiments.** *Nat. Methods* 2017, doi:10.1038/nmeth.4220.
65. Christoph Ziegenhain, Beate Vieth, Swati Parekh, Björn Reinius, Amy Guillaumet-Adkins, Martha Smets, Heinrich Leonhardt, Holger Heyn, Ines Hellmann, Wolfgang Enard: **Comparative Analysis of Single-Cell RNA Sequencing Methods.** *Mol. Cell* 2017, doi:10.1016/j.molcel.2017.01.023.
66. Kumar P, Tan Y, Cahan P: **Understanding development and stem cells using single cell-based analyses of gene expression.** *Development* 2017, **144**:17–32.
67. Welch JD, Hu Y, Prins JF: **Robust detection of alternative splicing in a population of single cells.** *Nucleic Acids Res.* 2016, **44**:e73.
68. Stubbington MJT, Lönnberg T, Proserpio V, Clare S, Speak AO, Dougan G, Teichmann SA: **T cell fate and clonality inference from single-cell transcriptomes.** *Nat. Methods* 2016, **13**:329–332.

69. Macaulay IC, Svensson V, Labalette C, Ferreira L, Hamey F, Voet T, Teichmann SA, Cvejic A: **Single-Cell RNA-Sequencing Reveals a Continuous Spectrum of Differentiation in Hematopoietic Cells.** *Cell Rep.* 2016, **14**:966–977.
70. Stegle O, Teichmann S a., Marioni JC: **Computational and analytical challenges in single-cell transcriptomics.** *Nat. Rev. Genet.* 2015, doi:10.1038/nrg3833.
71. Ilicic T, Kim JK, Kolodziejczyk AA, Bagger FO, Mccarthy DJ, Marioni JC, Teichmann SA: **Classification of low quality cells from single-cell RNA-seq data.** *Genome Biol.* 2016, doi:10.1186/s13059-016-0888-1.
72. Rostom R, Svensson V, Teichmann SA, Kar G: **Computational approaches for interpreting scRNA-seq data.** *FEBS Lett.* 2017, doi:10.1111/ijlh.12426.
73. van Dijk D, Nainys J, Sharma R, Kathail P, Carr AJ, Moon KR, Mazutis L, Wolf G, Krishnaswamy S, Pe'er D: **MAGIC: A diffusion-based imputation method reveals gene-gene interactions in single-cell RNA-sequencing data.** *bioRxiv* 2017, doi:10.1101/111591.
74. Bacher R, Kendzioriski C: **Design and computational analysis of single-cell RNA-sequencing experiments.** *Genome Biol.* 2016, **17**:63.
75. Chen J, Schlitzer A, Chakarov S, Ginhoux F, Poidinger M: **Mpath maps multi-branching single-cell trajectories revealing progenitor cell progression during development..** *Nat. Commun.* 2016, **7**:11988.
76. Shin J, Berg DA, Zhu Y, Shin JY, Song J, Bonaguidi MA, Enikolopov G, Nauen DW, Christian KM, Ming G, et al.: **Single-Cell RNA-Seq with Waterfall Reveals Molecular Cascades underlying Adult Neurogenesis.** *Cell Stem Cell* 2015, **17**:360–372.
77. Haghverdi L, Büttner M, Wolf FA, Buettner F, Theis FJ: **Diffusion pseudotime robustly reconstructs lineage branching.** *Nat. Methods* 2016, **13**:845–848.
78. Marco E, Karp RL, Guo G, Robson P, Hart AH, Trippa L, Yuan G-C: **Bifurcation analysis of single-cell gene expression data reveals epigenetic landscape.** *Proc. Natl. Acad. Sci.* 2014, **111**:E5643–E5650.
79. Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, Lennon NJ, Livak KJ, Mikkelsen TS, Rinn JL: **The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells.** *Nat. Biotechnol.* 2014, **32**:381–386.
80. Lönnberg T, Svensson V, James KR, Fernandez-Ruiz D, Sebina I, Montandon R, Soon MSF,

- Fogg LG, Nair AS, Liligeto UN, et al.: **Single-cell RNA-seq and computational analysis using temporal mixture modeling resolves T_H1/T_{FH} fate bifurcation in malaria.** *Sci. Immunol.* 2017, **2**:1–12.
81. Satija R, Farrell JA, Gennert D, Schier AF, Regev A: **Spatial reconstruction of single-cell gene expression data.** *Nat. Biotechnol.* 2015, **33**:495–502.
 82. Achim K, Pettit J-B, Saraiva LR, Gavriouchkina D, Larsson T, Arendt D, Marioni JC: **High-throughput spatial mapping of single-cell RNA-seq data to tissue of origin.** *Nat. Biotechnol.* 2015, **33**:503–509.
 83. Kharchenko P V, Silberstein L, Scadden DT: **Bayesian approach to single-cell differential expression analysis..** *Nat. Methods* 2014, **11**:740–2.
 84. Zhang B, Horvath S: **A General Framework for Weighted Gene Co-Expression Network Analysis.** *Stat. Appl. Genet. Mol. Biol.* 2005, **4**.
 85. Setty M, Tadmor MD, Reich-Zeliger S, Angel O, Salame TM, Kathail P, Choi K, Bendall S, Friedman N, Pe'er D: **Wishbone identifies bifurcating developmental trajectories from single-cell data.** *Nat. Biotechnol.* 2016, **34**:1–14.
 86. Bendall SC, Davis KL, Amir EAD, Tadmor MD, Simonds EF, Chen TJ, Shenfeld DK, Nolan GP, Pe'Er D: **Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development.** *Cell* 2014, **157**:714–725.
 87. Campbell KR, Yau C: **switchde: inference of switch-like differential expression along single-cell trajectories.** *Bioinformatics* 2016, **33**:1241–1242.
 88. Sander J, Schultze JL, Yosef N: **ImpulseDE: detection of differentially expressed genes in time series data using impulse models.** *Bioinformatics* 2016, **33**.
 89. Gaublotte JT, Yosef N, Lee Y, Gertner RS, Yang LV, Wu C, Pandolfi PP, Mak T, Satija R, Shalek AK, et al.: **Single-Cell Genomics Unveils Critical Regulators of Th17 Cell Pathogenicity.** *Cell* 2015, doi:10.1016/j.cell.2015.11.009.
 90. Kakaradov B, Arsenio J, Ewidjaja C, He Z, Aigner S, Metz PJ, Yu B, Wehrens EJ, Lopez J, Kim SH, et al.: **Early transcriptional and epigenetic regulation of CD8+ T cell differentiation revealed by single-cell RNA-seq.** *Nat. Immunol.* 2017, doi:10.1038/ni.3688.
 91. Canzar S, Neu KE, Tang Q, Wilson PC, Khan AA: **BASIC: BCR assembly from single cells.** *Bioinformatics* 2016, **33**:btw631.
 92. Afik S, Yates KB, Bi K, Darko S, Godec J, Gerdemann U, Swadling L, Douek DC, Klenerman P,

- Barnes EJ, et al.: **Targeted reconstruction of T cell receptor sequence from single cell RNA-sequencing links CDR3 length to T cell differentiation state.** *bioRxiv* 2016, doi:10.1101/072744.
93. Redmond D, Poran A, Elemento O: **Single-cell TCRseq: paired recovery of entire T-cell alpha and beta chain transcripts in T-cell receptors from single-cell RNAseq.** *Genome Med.* 2016, **8**:80.
94. Eltahla AA, Rizzetto S, Pirozyan MR, Betz-Stablein BD, Venturi V, Kedzierska K, Lloyd AR, Bull RA, Luciani F: **Linking the T cell receptor to the single cell transcriptome in antigen-specific human T cells.** *Immunol. Cell Biol.* 2016, **94**:604–611.
95. Mooijman D, Dey SS, Boisset J-C, Crosetto N, van Oudenaarden A: **Single-cell 5hmC sequencing reveals chromosome-wide cell-to-cell variability and enables lineage reconstruction..** *Nat. Biotechnol.* 2016, **34**:852–6.
96. Junker JP, Spanjaard B, Peterson-Maduro J, Alemany A, Hu B, Florescu M, van Oudenaarden A: **Massively parallel whole-organism lineage tracing using CRISPR/Cas9 induced genetic scars.** *bioRxiv* 2016, doi:10.1101/056499.
97. Kalhor R, Mali P, Church GM: **Rapidly evolving homing CRISPR barcodes.** *Nat. Methods* 2017, **14**:55863.
98. McKenna A, Findlay GM, Gagnon JA, Horwitz MS, Schier AF, Shendure J: **Whole-organism lineage tracing by combinatorial and cumulative genome editing.** *Science (80-).* 2016, **353**:aaf7907-aaf7907.
99. Perli SSD, Cui CH, Lu TK: **Continuous Genetic Recording with Self-Targeting CRISPR-Cas in Human Cells.** *bioRxiv* 2016, **511**:53058.
100. Schmidt ST, Zimmerman SM, Wang J, Kim SK, Quake SR: **Quantitative Analysis of Synthetic Cell Lineage Tracing Using Nuclease Barcoding.** *ACS Synth. Biol.* 2017, doi:10.1021/acssynbio.6b00309.
101. Woodworth MB, And KMG, Walsh CA: **Building a lineage from single cells: genetic techniques for cell lineage tracking.** *Nat. Rev. Genet.* 2017, **18**:230–244.
102. Gray DHD, Fletcher AL, Hammett M, Seach N, Ueno T, Young LF, Barbuto J, Boyd RL, Chidgey AP: **Unbiased analysis, enrichment and purification of thymic stromal cells.** *J. Immunol. Methods* 2008, **329**:56–66.
103. Wu TD, Nacu S: **Fast and SNP-tolerant detection of complex variants and splicing in short reads..** *Bioinformatics* 2010, **26**:873–81.

104. Anders S, Pyl PT, Huber W: **HTSeq--a Python framework to work with high-throughput sequencing data.** *Bioinformatics* 2015, **31**:166–169.
105. Scialdone A, Natarajan KN, Saraiva LR, Proserpio V, Teichmann SA, Stegle O, Marioni JC, Buettner F: **Computational assignment of cell-cycle stage from single-cell transcriptome data.** *Methods* 2015, **85**:54–61.
106. Buettner F, Natarajan KN, Casale FP, Proserpio V, Scialdone A, Theis FJ, Teichmann SA, Marioni JC, Stegle O: **Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells.** *Nat. Biotechnol.* 2015, **33**:155–160.
107. Sansom SN, Shikama N, Zhanybekova S, Nusspaumer G, Macaulay IC, Deadman ME, Heger A, Ponting CP, Holländer G a: **Population and single cell genomics reveal the Aire-dependency, relief from Polycomb silencing and distribution of self-antigen expression in thymic epithelia..** *Genome Res.* 2014, doi:10.1101/gr.171645.113.
108. Wilkerson MD, Hayes DN: **ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking.** *Bioinformatics* 2010, **26**:1572–1573.
109. Reimand J, Kull M, Peterson H, Hansen J, Vilo J: **g:Profiler--a web-based toolset for functional profiling of gene lists from large-scale experiments.** *Nucleic Acids Res.* 2007, **35**:W193–200.
110. Bettelli E, Carrier Y, Gao W, Korn T, Strom TB, Oukka M, Weiner HL, Kuchroo VK: **Reciprocal developmental pathways for the generation of pathogenic effector TH17 and regulatory T cells.** *Nature* 2006, **441**:235–238.
111. Haribhai D, Lin W, Relland LM, Truong N, Williams CB, Chatila TA: **Regulatory T cells dynamically control the primary immune response to foreign antigen.** *J. Immunol.* 2007, **178**:2961–2972.
112. Uhlig HH, Coombes J, Mottet C, Izcue A, Thompson C, Fanger A, Tannapfel A, Fontenot JD, Ramsdell F, Powrie F: **Characterization of Foxp3+CD4+CD25+ and IL-10-secreting CD4+CD25+ T cells during cure of colitis.** *J. Immunol.* 2006, **177**:5852–5860.
113. Riedel A, Shorthouse D, Haas L, Hall BA, Shields J: **Tumor-induced stromal reprogramming drives lymph node transformation.** *Nat. Immunol.* 2016, **17**:1118–1127.
114. Geremia A, Arancibia-Cárcamo C V, Fleming MPP, Rust N, Singh B, Mortensen NJ, Travis SPL, Powrie F: **IL-23-responsive innate lymphoid cells are increased in inflammatory**

- bowel disease. *J. Exp. Med.* 2011, **208**:1127–1133.
115. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C: **Salmon provides fast and bias-aware quantification of transcript expression.** *Nat. Methods* 2017, **14**:417–419.
 116. Katayama S, Skoog T, Jouhilahti E-M, Siitonen HA, Nuutila K, Tervaniemi MH, Vuola J, Johnsson A, Lönnerberg P, Linnarsson S, et al.: **Gene expression analysis of skin grafts and cultured keratinocytes using synthetic RNA normalization reveals insights into differentiation and growth control.** *BMC Genomics* 2015, **16**:476.
 117. Heng TSP, Painter MW, Immunological Genome Project Consortium: **The Immunological Genome Project: networks of gene expression in immune cells.** *Nat. Immunol.* 2008, **9**:1091–1094.
 118. McKenzie AT, Katsyv I, Song W-M, Wang M, Zhang B: **DGCA: A comprehensive R package for Differential Gene Correlation Analysis.** *BMC Syst. Biol.* 2016, **10**:106.
 119. Michalis K Titsias ND: **Bayesian Gaussian Process Latent Variable Model.** *Artif. Intell.* 2010, [no volume].
 120. Andreas Damianou Michalis Titsias, Neil Lawrence CE: **Manifold Relevance Determination.** *arXiv* 2012, doi:arXiv:1206.4610.
 121. Reimand J, Arak T, Adler P, Kolberg L, Reisberg S, Peterson H, Vilo J: **g:Profiler-a web server for functional interpretation of gene lists (2016 update).** *Nucleic Acids Res.* 2016, **44**:W83–9.
 122. Kiselev VY, Kirschner K, Schaub MT, Andrews T, Yiu A, Chandra T, Natarajan KN, Reik W, Barahona M, Green AR, et al.: **SC3: consensus clustering of single-cell RNA-seq data.** *Nat. Methods* 2017, **14**:483–486.
 123. Meredith M, Zemmour D, Mathis D, Benoist C: **Aire controls gene expression in the thymic epithelium with ordered stochasticity..** *Nat. Immunol.* 2015, **16**:942–949.
 124. Brennecke P, Reyes A, Pinto S, Rattay K, Nguyen M, Kuchler R, Huber W, Kyewski B, Steinmetz LM: **Single-cell transcriptome analysis reveals coordinated ectopic gene-expression patterns in medullary thymic epithelial cells..** *Nat. Immunol.* 2015, doi:10.1038/ni.3246.
 125. Irla M, Hugues S, Gill J, Nitta T, Hikosaka Y, Williams IR, Hubert F-X, Scott HS, Takahama Y, Holländer GA, et al.: **Autoantigen-specific interactions with CD4+ thymocytes control mature medullary thymic epithelial cell cellularity.** *Immunity* 2008, **29**:451–463.

126. Picelli S, Björklund ÅK, Faridani OR, Sagasser S, Winberg G, Sandberg R: **Smart-seq2 for sensitive full-length transcriptome profiling in single cells.** *Nat. Methods* 2013, **10**:1096–1098.
127. Gray DHD, Seach N, Ueno T, Milton MK, Liston A, Lew AM, Goodnow CC, Boyd RL: **Developmental kinetics, turnover, and stimulatory capacity of thymic epithelial cells.** *Blood* 2006, **108**:3777–3785.
128. Van de Walle I, De Smet G, Gärtner M, De Smedt M, Waegemans E, Vandekerckhove B, Leclercq G, Plum J, Aster JC, Bernstein ID, et al.: **Jagged2 acts as a Delta-like Notch ligand during early hematopoietic cell fate decisions.** *Blood* 2011, **117**:4449–4459.
129. Spence PJ, Green EA: **Foxp3+ regulatory T cells promiscuously accept thymic signals critical for their development.** *Proc. Natl. Acad. Sci. U. S. A.* 2008, **105**:973–978.
130. Nazzal D, Gradolatto A, Truffault F, Bismuth J, Berrih-Aknin S: **Human thymus medullary epithelial cells promote regulatory T-cell generation by stimulating interleukin-2 production via ICOS ligand.** *Cell Death Dis.* 2014, **5**:e1420.
131. Barbee SD, Woodward MJ, Turchinovich G, Mention J-J, Lewis JM, Boyden LM, Lifton RP, Tigelaar R, Hayday AC: **Skint-1 is a highly specific, unique selecting component for epidermal T cells.** *Proc. Natl. Acad. Sci.* 2011, **108**:3330–3335.
132. Stillman BN, Hsu DK, Pang M, Brewer CF, Johnson P, Liu F-T, Baum LG: **Galectin-3 and Galectin-1 Bind Distinct Cell Surface Glycoprotein Receptors to Induce T Cell Death.** *J. Immunol.* 2006, **176**:778–789.
133. Bi S, Earl LA, Jacobs L, Baum LG: **Structural features of galectin-9 and galectin-1 that determine distinct T cell death pathways.** *J. Biol. Chem.* 2008, **283**:12248–12258.
134. Akiyama T, Shimo Y, Yanai H, Qin J, Ohshima D, Maruyama Y, Asaumi Y, Kitazawa J, Takayanagi H, Penninger JM, et al.: **The tumor necrosis factor family receptors RANK and CD40 cooperatively establish the thymic medullary microenvironment and self-tolerance.** *Immunity* 2008, **29**:423–437.
135. Lee JB, Werbowetski-Ogilvie TE, Lee J-H, McIntyre BAS, Schnerch A, Hong S-H, Park I-H, Daley GQ, Bernstein ID, Bhatia M: **Notch-HES1 signaling axis controls hemato-endothelial fate decisions of human embryonic and induced pluripotent stem cells.** *Blood*

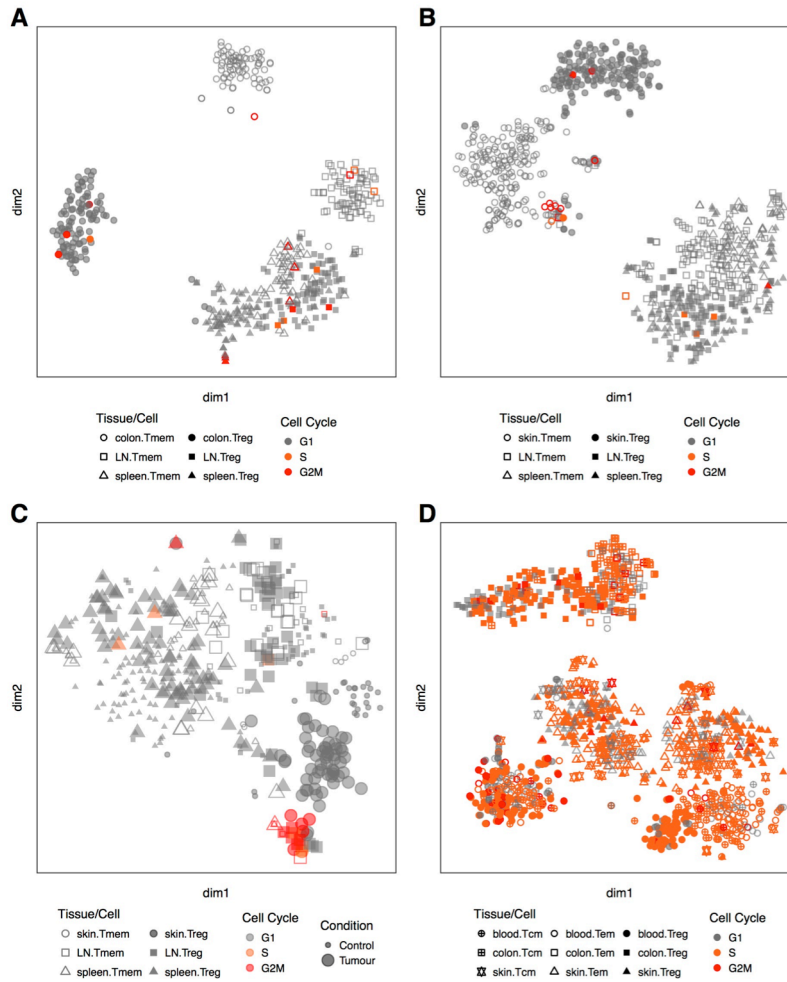
- 2013, **122**:1162–1173.
136. St-Jean JR, Ounissi-Benkalha H, Polychronakos C: **Yeast one-hybrid screen of a thymus epithelial library identifies ZBTB7A as a regulator of thymic insulin expression.** *Mol. Immunol.* 2013, **56**:637–642.
 137. Derbinski J, Gäbler J, Brors B, Tierling S, Jonnakuty S, Hergenahn M, Peltonen L, Walter J, Kyewski B: **Promiscuous gene expression in thymic epithelial cells is regulated at multiple levels.** *J. Exp. Med.* 2005, **202**:33–45.
 138. Gillard GO, Farr AG: **Contrasting models of promiscuous gene expression by thymic epithelium.** *J. Exp. Med.* 2005, **202**:15–19.
 139. Johnnidis JB, Venanzi ES, Taxman DJ, Ting JP-Y, Benoist CO, Mathis DJ: **Chromosomal clustering of genes controlled by the aire transcription factor.** *Proc. Natl. Acad. Sci. U. S. A.* 2005, **102**:7233–8.
 140. Mathis D, Benoist C: **Aire.** *Annu. Rev. Immunol.* 2009, **27**:287–312.
 141. Lkhagvasuren E, Sakata M, Ohigashi I, Takahama Y: **Lymphotoxin β receptor regulates the development of CCL21-expressing subset of postnatal medullary thymic epithelial cells.** *J. Immunol.* 2013, **190**:5110–5117.
 142. Klein L, Hinterberger M, Wirnsberger G, Kyewski B: **Antigen presentation in the thymus for positive selection and central tolerance induction.** *Nat. Rev. Immunol.* 2009, **9**:833–844.
 143. Steeber DA, Green NE, Sato S, Tedder TF: **Lymphocyte migration in L-selectin-deficient mice. Altered subset migration and aging of the immune system.** *J. Immunol.* 1996, **157**:1096–1106.
 144. Schneider MA, Meingassner JG, Lipp M, Moore HD, Rot A: **CCR7 is required for the in vivo function of CD4+ CD25+ regulatory T cells.** *J. Exp. Med.* 2007, **204**:735–745.
 145. De Simone M, Arrigoni A, Rossetti G, Gruarin P, Ranzani V, Politano C, Bonnal RJP, Provasi E, Sarnicola ML, Panzeri I, et al.: **Transcriptional Landscape of Human Tissue Lymphocytes Unveils Uniqueness of Tumor-Infiltrating T Regulatory Cells.** *Immunity* 2016, **45**:1135–1147.
 146. Björklund ÅK, Forkel M, Picelli S, Konya V, Theorell J, Friberg D, Sandberg R, Mjösberg J: **The heterogeneity of human CD127(+) innate lymphoid cells revealed by single-cell RNA sequencing.** *Nat. Immunol.* 2016, **17**:451–60.
 147. Sefik E, Geva-Zatorsky N, Oh S, Konnikova L, Zemmour D, McGuire AM, Burzyn D, Ortiz-Lopez

- A, Lobera M, Yang J, et al.: **MUCOSAL IMMUNOLOGY. Individual intestinal symbionts induce a distinct population of ROR γ ⁺ regulatory T cells.** *Science (80-)*. 2015, **349**:993–997.
148. Ohnmacht C, Park J-H, Cording S, Wing JB, Atarashi K, Obata Y, Gaboriau-Routhiau V, Marques R, Dulauroy S, Fedoseeva M, et al.: **MUCOSAL IMMUNOLOGY. The microbiota regulates type 2 immunity through ROR γ ⁺ T cells.** *Science (80-)*. 2015, **349**:989–993.
149. Shekhar K, Lapan SW, Whitney IE, Tran NM, Macosko EZ, Kowalczyk M, Adiconis X, Levin JZ, Nemesh J, Goldman M, et al.: **Comprehensive Classification of Retinal Bipolar Neurons by Single-Cell Transcriptomics.** *Cell* 2016, **166**:1308–1323.e30.
150. Kim S V, Xiang W V, Kwak C, Yang Y, Lin XW, Ota M, Sarpel U, Rifkin DB, Xu R, Littman DR: **GPR15-mediated homing controls immune homeostasis in the large intestine mucosa.** *Science (80-)*. 2013, **340**:1456–1459.
151. Sasaki T, Onodera A, Hosokawa H, Watanabe Y, Horiuchi S, Yamashita J, Tanaka H, Ogawa Y, Suzuki Y, Nakayama T: **Genome-Wide Gene Expression Profiling Revealed a Critical Role for GATA3 in the Maintenance of the Th2 Cell Identity.** *PLoS One* 2013, **8**:e66468.
152. Fatma N, Kubo E, Takamura Y, Ishihara K, Garcia C, Beebe DC, Singh DP: **Loss of NF-kappaB control and repression of Prdx6 gene transcription by reactive oxygen species-driven SMAD3-mediated transforming growth factor beta signaling.** *J. Biol. Chem.* 2009, **284**:22758–22772.
153. Hardee J, Ouyang Z, Zhang Y, Kundaje A, Lacroute P, Snyder M: **STAT3 targets suggest mechanisms of aggressive tumorigenesis in diffuse large B-cell lymphoma.** *G3* 2013, **3**:2173–2185.
154. Giordano M, Henin C, Maurizio J, Imbratta C, Bourdely P, Buferne M, Baitsch L, Vanhille L, Sieweke MH, Speiser DE, et al.: **Molecular profiling of CD8 T cells in autochthonous melanoma identifies Maf as driver of exhaustion.** *EMBO J.* 2015, **34**:2042–2058.
155. Tullai JW, Schaffer ME, Mullenbrock S, Kasif S, Cooper GM: **Identification of transcription factor binding sites upstream of human genes regulated by the phosphatidylinositol 3-kinase and MEK/ERK signaling pathways.** *J. Biol. Chem.* 2004, **279**:20167–20177.

156. Nakamura N: **The Role of the Transmembrane RING Finger Proteins in Cellular and Organelle Function.** *Membranes (Basel)*. 2011, **1**:354–393.
157. Dimeloe S, Burgener A-V, Grählert J, Hess C: **T-cell metabolism governing activation, proliferation and differentiation; a modular view.** *Immunology* 2017, **150**:35–44.
158. Gagné-Ouellet V, Guay S-P, Boucher-Lafleur A-M, Bouchard L, Laprise C: **DNA methylation signature of interleukin 1 receptor type II in asthma.** *Clin. Epigenetics* 2015, **7**:80.
159. Michel M, Demel C, Zacher B, Schwalb B, Krebs S, Blum H, Gagneur J, Cramer P: **TT-seq captures enhancer landscapes immediately after T-cell stimulation.** *Mol. Syst. Biol.* 2017, **13**:920.
160. Bapat SP, Myoung Suh J, Fang S, Liu S, Zhang Y, Cheng A, Zhou C, Liang Y, LeBlanc M, Liddle C, et al.: **Depletion of fat-resident Treg cells prevents age-associated insulin resistance.** *Nature* 2015, **528**:137–141.
161. Cipolletta D, Cohen P, Spiegelman BM, Benoist C, Mathis D: **Appearance and disappearance of the mRNA signature characteristic of Treg cells in visceral adipose tissue: age, diet, and PPAR γ effects.** *Proc. Natl. Acad. Sci. U. S. A.* 2015, **112**:482–487.
162. Griseri T, Asquith M, Thompson C, Powrie F: **OX40 is required for regulatory T cell-mediated control of colitis.** *J. Exp. Med.* 2010, **207**:699–709.
163. Molofsky AB, Van Gool F, Liang H-E, Van Dyken SJ, Nussbaum JC, Lee J, Bluestone JA, Locksley RM: **Interleukin-33 and Interferon- γ Counter-Regulate Group 2 Innate Lymphoid Cell Activation during Immune Perturbation.** *Immunity* 2015, **43**:161–174.
164. Ohkura N, Hamaguchi M, Morikawa H, Sugimura K, Tanaka A, Ito Y, Osaki M, Tanaka Y, Yamashita R, Nakano N, et al.: **T cell receptor stimulation-induced epigenetic changes and Foxp3 expression are independent and complementary events required for Treg cell development..** *Immunity* 2012, **37**:785–99.
165. Chaudhry A, Rudra D, Treuting P, Samstein RM, Liang Y, Kas A, Rudensky AY: **CD4 Regulatory T Cells Control TH17 Responses in a Stat3-Dependent Manner.** *Science (80-)*. 2009, **326**:986–991.
166. Cretney E, Xin A, Shi W, Minnich M, Masson F, Miasari M, Belz GT, Smyth GK, Busslinger M, Nutt SL, et al.: **The transcription factors Blimp-1 and IRF4 jointly control the differentiation and function of effector regulatory T cells.** *Nat. Immunol.* 2011,

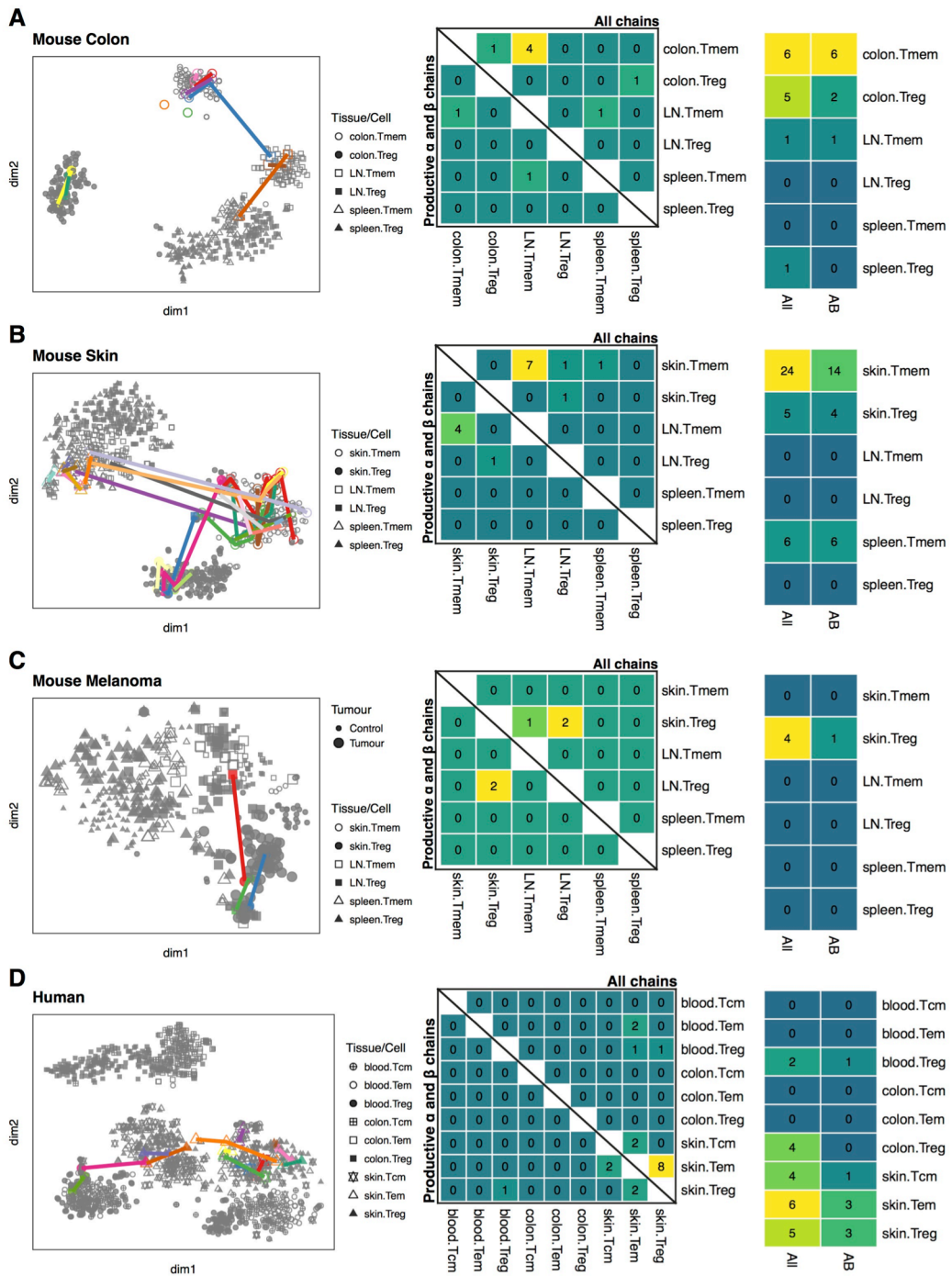
- 12:304–311.
167. Plitas G, Konopacki C, Wu K, Bos PD, Morrow M, Putintseva E V, Chudakov DM, Rudensky AY: **Regulatory T Cells Exhibit Distinct Features in Human Breast Cancer.** *Immunity* 2016, **45**:1122–1134.
 168. Sherwood AM, Emerson RO, Scherer D, Habermann N, Buck K, Staffa J, Desmarais C, Halama N, Jaeger D, Schirmacher P, et al.: **Tumor-infiltrating lymphocytes in colorectal tumors display a diversity of T cell receptor sequences that differ from the T cells in adjacent mucosal tissue.** *Cancer Immunol. Immunother.* 2013, **62**:1453–1461.
 169. Malik BT, Byrne KT, Vella JL, Zhang P, Shabaneh TB, Steinberg SM, Molodtsov AK, Bowers JS, Angeles C V, Paulos CM, et al.: **Resident memory T cells in the skin mediate durable immunity to melanoma.** *Sci Immunol* 2017, **2**.
 170. Miragaia RJ, Zhang X, Gomes T, Svensson V, Ilicic T, Henriksson J, Kar G, Lönnberg T: **Single-cell RNA-sequencing resolves self-antigen expression during mTEC development (manuscript under revision).**
 171. Miragaia RJ, Teichmann SA, Hagai T: **Single-cell insights into transcriptomic diversity in immunity.** *Curr. Opin. Syst. Biol.* 2017, **5**:63–71.
 172. Miragaia RJ, Gomes T, Chomka A, Jardine L, Riedel A, Hegazy A, Lindeman I, Emerton G, Krausgruber T, Shields J, et al.: **Single-cell transcriptomics of regulatory T cells reveals trajectories of tissue adaptation (manuscript in preparation).**

Annexes



Supplementary Figure 1 - Cell cycle phase detection by Cyclone.

(A-D) t-SNE dimensionality reduction showing cell cycle phase in (A) mouse colon, (B) mouse skin, (C) mouse melanoma and (D) human datasets.



Supplementary Figure 2 - CD4⁺ T cells TCR clonotypes in different tissues.

Clonotypes detected using TraCeR software in (A) mouse colon, (B) mouse skin, (C) mouse melanoma and (D) human datasets.

(left) t-SNE dimensionality reduction highlighting cells sharing productive α and β TCR chains. Shapes match cell type and tissue origin according to the legend.

(middle) Number of clonotypes detected spanning different tissues and cell type combinations. Top half registers all events of TCR chain sharing, bottom half only considers the sharing of productive α and β TCR chain.

(bottom) Number of clonotypes detected within each cell type and tissue, considering the sharing of any chain or productive α and β .

