

Tissue-specific adaptations of cell types



Tomás Pires de Carvalho Gomes

Wellcome Sanger Institute
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

King's College

January 2020

To the bigger picture!

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Contributions. This dissertation contains fewer than 60,000 words including appendices, bibliography, footnotes, tables and equations.

Tomás Pires de Carvalho Gomes
January 2020

Acknowledgements

I would need an additional very long chapter to fully and fairly acknowledge each and every person that contributed to this endeavour.

I would like to acknowledge my supervisor, Sarah Teichmann, for the opportunity to pursue this PhD. It has been a long and winding journey, and I could not have done it without her guidance. Sarah was always ready to provide valuable, original insights on the problems and questions I had, teaching me so much more than I have ever expected. I fell truly fortunate to have her as a mentor that I can rely on.

Besides being a brilliant scientist, one of Sarah's greatest achievements has been to put together and maintain a highly collaborative lab, full of outstanding people always ready to inspire and help. It is hard to find someone who has not had an impact on my progress and research, but I would like to name a few. To Ricardo Miragaia, the other half of the Portuguese dynamic duo, for the relentless understanding and support provided. Through good and bad times, we had a lot of fun together, and I would not have it any other way. I also want to acknowledge my "Jedi master" Roser Vento, for the interesting, deep discussions and crucial encouragement throughout my PhD. I am lucky to have such a brilliant and dedicated friend. To Tzachi Hagai, who is a constant source of inspiration, enthusiasm and skepticism. He has taught me to question everything, and through better knowing my limitations make the most of them. Valentine Svensson was one of the people that I learned the most from. Likely the most talented young scientist I have ever met, he was crucial in my understanding of single-cell computational methods, and always a fun company to be around. Raghda Rostom, my "PhD sister", whose funny antics and incisive comments have helped keep my head grounded. To Kylie James, for the uplifting conversations about life, they have always left me with a feeling of hope for the future. And to Rasa Elmentaite, for her inspiring dedication and contagious desire to learn and ask questions.

I am grateful to the whole ENLIGH-TEN consortium, which funded my PhD and provided me with the opportunity to meet several other early stage researchers and

their supervisors. The support from this network has been crucial for my research, and gave me a true sense of what it is like to do science across borders.

I could not have started this PhD without the assistance of my former advisors Maria do Carmo Fonseca and Ana Rita Grosso, as well as collaborators Nick Proudfoot and Taka Nojima. They were crucial in starting my career in bioinformatics and genomics, and are overall great scientists and supportive mentors.

I would also like to thank all my friends, in particular to Tiago Pires, who despite the distance always gave me an escape from the hectic PhD life. I would also like to thank Mariana, Elsa, Gonçalo, and Joana for the fun times we had in Cambridge and with the Cambridge University Portuguese Speakers' Society. And to Gianmarco Raddi, who I could always count on to keep my spirits up.

I am deeply grateful to all my family, my father José Carlos, my mother Ana Luísa, my brother Miguel, my aunt Fernanda, and my grandmother Maria do Carmo. They have been relentless in their assistance and understanding, and it is thanks to them that I got the opportunity to work on what I love. I also want to posthumously thank my grandfather Fernando for all the inspiration and all he taught me.

Finally, and most importantly, I want to thank my partner Hajrabibi Ali. I am deeply indebted to her for her encouragement, patience and self-sacrifice, that ultimately help me pursue this degree. I am hoping that I can ever repay her for her time and help in keeping me focused and reminding me of the greater picture.

Contributions

The multidisciplinary nature of the studies here presented required the valuable contributions of my collaborators. This will be further detailed at the start of each chapter, but will also be here summarised.

- In Chapter 2, the original experiments were designed by Ricardo J Miragaia, who also assisted in data interpretation.
- In Chapter 3, the original concept was conceived together with Valentine Svensson.

Abstract

Cells are the building blocks of life, forming the vast diversity of tissues and organisms in Nature. Across these, common cellular morphologies and functions have been identified. High-throughput, multifactorial profiling of cells has grown exponentially in recent years with the advent of single-cell RNA-sequencing (scRNA-seq), increasingly unravelling cell diversity. Nonetheless, it is not yet known how different environments affect cellular phenotypes.

The work presented on this Thesis reports on the transcriptional variation of cell types across tissues, by use of single-cell RNA-sequencing. This technology, developed in the last 10 years, has greatly impacted our ability to distinguish cellular heterogeneity by their gene expression in various tissues or conditions.

Chapter 1 outlines the impact of single-cell RNA-sequencing in cell biology, presenting the technology as the natural progression of lower throughput or low-resolution methods. The chapter then shows how cellular heterogeneity can be deconstructed by analysing this type of genomics data. It then expands on how individual datasets can be used to build models of cell type identity for automatic annotation, ultimately outlining the need to create a global cell type census of a whole organism. A cell compendium like this should be useful for automatic annotation, as well as to obtain a cross-tissue integrative overview of cell identity.

The same chapter also delves into the topic of heterogeneity in immune cells. Due to the evolutionary pressure they are subject to and ubiquitous nature across the organism, these are some of the most diverse cell types in multicellular organisms. Chapter 2 presents a deconstruction of T-regulatory cells' phenotypes in different mouse and human tissues using single-cell RNA-sequencing. The analysis in this chapter will show how these cells are structured in subpopulations, and how they adapt when migrating between lymphoid and non-lymphoid tissues. It will also assess the conservation of gene expression programmes for the same populations between mouse and human.

The creation of a global cell type reference is an endeavour that can facilitate analysis of new data, and reveal novel insights about cell and tissue biology. Several

datasets have now been produced, and a method that can efficiently integrate them and prepare them for use as a reference is necessary. Chapter 3 details the development of such method, exploring its strengths and how it can be improved, in a mouse dataset. Chapter 4 then applies this pipeline to a collection of human data, and shows how cell types relate across tissues, as well as how the human reference can be used in a practical case.

Lastly, Chapter 5 summarises all chapters, providing an overview on how single-cell sequencing has changed what we know about tissue biology, and how listing cell types and compiling them as a functional reference can help future developments in life sciences.

Table of contents

List of figures	xvii
List of tables	xix
Nomenclature	xxiii
1 Cellular identity in the genomics era	1
1.1 Cell type discovery and definition	1
1.2 Defining cell types using scRNA-seq	3
1.3 Methods for cell type classification	7
1.4 Cell identity in the immune system	11
1.5 Tissue-specific gene expression	16
1.6 Insights and scope of this thesis	19
2 Tissue adaptation of T-regulatory cells	21
2.1 Introduction	22
2.2 Results	23
2.2.1 Treg and Tmem cell identity in NLTs is driven by a common expression module	23
2.2.2 Heterogeneity within LT and NLT Treg cell populations	24
2.2.3 Treg cells adapting to skin and colon share a transcriptional trajectory	28
2.2.4 Treg cell recruitment into skin and melanoma relies on common mechanisms	30
2.2.5 Conserved NLT identity in mouse and human	33
2.2.6 Classification of Treg cell populations across species	35
2.3 Discussion	37
2.4 Methods	40
2.4.1 RNA expression quantification and normalisation	40

2.4.2	scRNA-seq quality control	41
2.4.3	Dimensionality reduction methods	42
2.4.4	Subpopulation detection in 10x data	42
2.4.5	Differential expression analysis	43
2.4.6	Mapping cells to known populations using logistic regression classification	43
2.4.7	Obtaining a migration latent variable for steady-state Treg cells	44
2.4.8	Identifying a common tissue migration trajectory in control and melanoma	45
2.4.9	Switch-like genes in the migration latent variable	45
2.4.10	RNA velocity estimation	46
2.4.11	Detection of expanded clonotypes	46
2.4.12	GO Term enrichment	46
2.4.13	Cell-cycle analysis	47
2.5	Conclusions and future work	47
3	Developing a method to integrate and classify cell types across tissues	51
3.1	Introduction	51
3.2	Methodology	53
3.2.1	Per-tissue clustering to approximate cell type annotations . .	53
3.2.2	Combining cell clusters across tissues using tissue-specific clas- sifiers	56
3.2.3	Generating updatable transparent-box models for cell type classification	60
3.3	Results	62
3.3.1	Training <i>CellTypist</i> on the <i>Tabula Muris</i> dataset	62
3.3.2	Training <i>CellTypist</i> on a collection of human data	63
3.4	Discussion	68
4	Application and biological insights of the <i>CellTypist</i> model	71
4.1	Introduction	72
4.2	Results	73
4.2.1	<i>CellTypist</i> as an operational reference for annotation	73
4.2.2	Matching cell identity across tissues	78
4.2.3	Gene expression hallmarks of cell identity	82
4.3	Discussion	85
4.4	Methods	87

4.4.1	<i>CellTypist</i> parameter optimisation and training	87
4.4.2	Obtaining gene group lists	88
4.4.3	Clustering	89
4.4.4	Enrichment of gene groups	90
5	Concluding remarks	91
5.1	Cells and genes trade-offs in single-cell profiling	91
5.2	Building a transcriptomic atlas of cell types	92
5.3	Defining cellular identity	94
	References	97
	Appendix A Additional information to Chapter 2	127
A.1	Additional Experimental Methods	127
A.1.1	Mice	127
A.1.2	Human samples	127
A.1.3	Murine leukocytes isolation in steady-state skin dataset	128
A.1.4	Murine leukocytes isolation in steady-state colon dataset	128
A.1.5	Melanoma induction and cell isolation	128
A.1.6	Isolation of human CD4+ T cells	129
A.1.7	Flow cytometry and single-cell RNA sequencing	130
A.2	Supplementary Tables and Figures	133
A.3	Data and Code Accessibility	146
A.4	Full author list and contributions	146
A.4.1	Acknowledgements	146
	Appendix B Additional information to Chapter 3	149
B.1	Supplementary Figures	149
B.2	Supplementary Tables	154
	Appendix C Additional information to Chapter 4	175
C.1	Supplementary Figures	175
C.2	Supplementary Tables	189
	Appendix D Publications contributed to during the PhD degree	201

List of figures

1.1	Timeline of scRNA-seq technology development	4
1.2	Gene and protein structure of TCR	12
1.3	T-helper cell heterogeneity and key marker genes	14
2.1	Steady-state scRNA-seq datasets of CD4 ⁺ T cells from LT and NLT . .	24
2.2	Heterogeneity within LT and NLT Treg populations	26
2.3	Reconstruction of Treg cell recruitment from lymphoid to non-lymphoid tissues in steady-state	29
2.4	Recruitment and adaptation of Treg cells to the tumour environment recapitulates steady-state migration	32
2.5	Human-mouse comparison of NLT Treg cell marker genes.	34
2.6	Examining models for cross-species Treg classification	36
2.7	Enrichment of genes from the TNF pathway in NLT T cells	38
3.1	Data reprocessing per-tissue	54
3.2	Cross-tissue matching of cell types	57
3.3	Evaluation of clusters matched across tissues	59
3.4	Model training outline and evaluation	61
3.5	Evaluating model trained on cross-tissue integrated clusters	62
3.6	Cell numbers in the human dataset collection	64
3.7	Running <i>CellTypist</i> on a human scRNA-seq data collection	66
4.1	<i>CellTypist</i> predictions for lung data from (Madisson et al., 2019) . .	74
4.2	Classification accuracy for the (Madisson et al., 2019) dataset . . .	76
4.3	Cell identity relationships across tissues	80
4.4	Top gene groups for cell identification across human tissues	83
4.5	Top gene groups for cell identification across mouse tissues	84
A.1	Sorting and identification of Treg and Tmem cells	138

A.2	Heterogeneity in SS2 and Tmem cell populations	140
A.3	Additional information on BGPLVM for the 10x dataset	142
A.4	Additional information on BGPLVM for the Smart-seq2 datasets . . .	143
A.5	Additional details on the MRD-BGPLVM projection	144
A.6	Additional information about the Human Smart-seq2 dataset	145
B.1	Cell numbers in the <i>Tabula Muris</i> dataset	149
B.2	Expression of <i>PTPRC</i> and <i>EPCAM</i> in human data collection	150
B.3	<i>CellTypist</i> parameters grids with other statistics	151
B.4	Grouping of annotated cell types and datasets in human pancreas data	152
B.5	Training statistics for other <i>CellTypist</i> models	153
C.1	Number of tissue-specific genes determined per tissue for mouse and human	176
C.2	Relating number of per-tissue clusters and number of cells	177
C.3	Enrichment of tissue gene modules in other <i>CellTypist</i> models	178
C.4	Clusters merged across tissues in the different models	179
C.5	Enrichment of tissue gene modules in merged clusters of different <i>CellTypist</i> models	180
C.6	Correlation between gene expression and importance in the human <i>CellTypist</i> model	181
C.7	Correlation between gene expression and importance in the <i>Tabula Muris CellTypist</i> model	182
C.8	Gene upset plots of different <i>CellTypist</i> models	183
C.9	<i>CellTypist</i> predictions for oesophagus data from (Madisson et al., 2019)	184
C.10	<i>CellTypist</i> predictions for spleen data from (Madisson et al., 2019) .	185
C.11	Matching <i>CellTypist</i> predictions in lung with annotations in the data collection	186
C.12	Clusters matching lung annotated cell types in other <i>CellTypist</i> models	187
C.13	Lung annotated cell types matching clusters in other <i>CellTypist</i> models	188

List of tables

1.1	Current methods for single-cell RNA-seq	5
1.2	Methods for automated cell state matching	9
A.1	Batch details for the Mouse steady-state Smart-seq2 data	133
A.2	Batch details for the Mouse melanoma Smart-seq2 data	134
A.3	Batch details for the Human steady-state Smart-seq2 data	135
A.4	Batch details for the Mouse steady-state Chromium 10x data	135
A.5	Quality control criteria for filtering scRNA-seq	136
A.6	Clinical information on human donors included in this study	137
B.1	F1 scores and class sizes for <i>CellTypist</i> trained on the <i>Tabula Muris</i> with cell type labels	155
B.2	F1 scores and class sizes for <i>CellTypist</i> trained on the <i>Tabula Muris</i> with cell type labels (continued)	156
B.3	F1 scores and class sizes for <i>CellTypist</i> trained on the <i>Tabula Muris</i> with integrated cluster labels	157
B.4	F1 scores and class sizes for <i>CellTypist</i> trained on the <i>Tabula Muris</i> with integrated cluster labels (continued 1)	158
B.5	F1 scores and class sizes for <i>CellTypist</i> trained on the <i>Tabula Muris</i> with integrated cluster labels (continued 2)	159
B.6	F1 scores and class sizes for <i>CellTypist</i> trained on the <i>Tabula Muris</i> with integrated cluster labels (continued 3)	160
B.7	Human scRNA-seq datasets collected and corresponding cell numbers	161
B.8	F1 scores and class sizes for <i>CellTypist</i> trained on the human collection with integrated cluster labels	162
B.9	F1 scores and class sizes for <i>CellTypist</i> trained on the human collection with integrated cluster labels (continued 1)	163

B.10 F1 scores and class sizes for <i>CellTypist</i> trained on the human collection with integrated cluster labels (continued 2)	164
B.11 F1 scores and class sizes for <i>CellTypist</i> trained on the human collection with integrated cluster labels (continued 3)	165
B.12 F1 scores and class sizes for <i>CellTypist</i> trained on the human collection with integrated cluster labels (continued 4)	166
B.13 F1 scores and class sizes for <i>CellTypist</i> trained on the human collection with integrated cluster labels (continued 5)	167
B.14 F1 scores and class sizes for <i>CellTypist</i> trained on the human collection with integrated cluster labels (continued 6)	168
B.15 F1 scores and class sizes for <i>CellTypist</i> trained on the human collection with integrated cluster labels (continued 7)	169
B.16 F1 scores and class sizes for <i>CellTypist</i> trained on the human collection with integrated cluster labels (continued 8)	170
B.17 F1 scores and class sizes for <i>CellTypist</i> trained on the human collection with integrated cluster labels (continued 9)	171
B.18 F1 scores and class sizes for <i>CellTypist</i> trained on the human collection with integrated cluster labels (continued 10)	172
B.19 F1 scores and class sizes for <i>CellTypist</i> trained on the human collection with integrated cluster labels (continued 11)	173
B.20 Top genes in the largest merged clusters of each <i>CellTypist</i> model	174
C.1 Cell types from (Madisson et al., 2019) with expression programmes enriched in <i>CellTypist</i> clusters	190
C.2 Cell types from (Madisson et al., 2019) with expression programmes enriched in <i>CellTypist</i> clusters (continued 1)	191
C.3 Cell types from (Madisson et al., 2019) with expression programmes enriched in <i>CellTypist</i> clusters (continued 2)	192
C.4 Cell types from (Madisson et al., 2019) with expression programmes enriched in <i>CellTypist</i> clusters (continued 3)	193
C.5 Cell types from (Madisson et al., 2019) with expression programmes enriched in <i>CellTypist</i> clusters (continued 4)	194
C.6 Cell types from (Madisson et al., 2019) with expression programmes enriched in <i>CellTypist</i> clusters (continued 5)	195
C.7 Cell types from (Madisson et al., 2019) with expression programmes enriched in <i>CellTypist</i> clusters (continued 6)	196

C.8	Cell types from (Madisson et al., 2019) with expression programmes enriched in <i>CellTypist</i> clusters (continued 7)	197
C.9	Cell types from (Madisson et al., 2019) with expression programmes enriched in <i>CellTypist</i> clusters (continued 8)	198
C.10	Cell types from (Madisson et al., 2019) with expression programmes enriched in <i>CellTypist</i> clusters (continued 9)	199

Nomenclature

Acronyms / Abbreviations

APC Antigen-Presenting Cell

ARD Automatic Relevance Determination

BGPLVM Bayesian Gaussian Process Latent Variable Modelling

bLN brachial Lymph Nodes

CITE-seq Cellular Indexing of Transcriptomes and Epitopes by sequencing

cTreg central Treg (cells)

DE Differentially Expressed (genes)

EGFP Enhanced Green Fluorescent Protein

ERCC External RNA Controls Consortium

eTreg effector Treg (cells)

FACS Fluorescence-Activated Cell Sorting

GRCh Genome Reference Consortium human

GRCm Genome Reference Consortium mouse

HCA Human Cell Atlas

iNKT invariant Natural Killer T (cells)

LN Lymph Nodes

LT Lymphoid Tissues

LV	Latent Variable
MHC	Major Histocompatibility Complex
mLN	mesenteric Lymph Nodes
MRD-BGPLVM	Manifold Relevnce Determination-BGPLVM
NLT	Non Lymphoid Tissue
NMF	Non-negative Matrix Factorization
oNMF	orthogonal Non-negative Matrix Factorization
PBS	Phosphate Buffer Saline
PCA	Principal Component Analysis
QC	Quality Control
RNA	Ribonucleic acid
scATAC-seq	single-cell Assay for Transposase-Accessible Chromatin sequencing
scRNA-seq	Single-cell RNA sequencing
SGD	Stochastic Gradient Descent
SJ	split-join (distance)
SS2	Smart-seq2
SVM	Support Vector Machine
TCR	T Cell Receptor
Tfh	T follicular helper (cells)
Th	T-helper (cells)
Tmem	T-memory (cells)
Treg	T-regulatory (cells)
tSNE	t-Distributed Stochastic Neighbor Embedding
VAE	Variational Autoencoder
VAT	Visceral Adipose Tissue