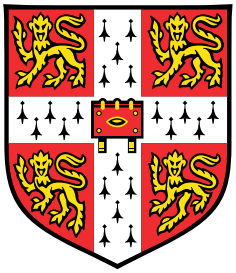
Tissue-specific adaptations of cell types



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. Raghd 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 greatful 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 assitance 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

List of tables

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