Human cellular genetics of innate immunity

Raghd Rostom

Wellcome Sanger Institute

University of Cambridge

This dissertation is submitted for the degree of

Doctor of Philosophy

Christ's College August 2019

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 Acknowledgements. This dissertation contains fewer than 60,000 words excluding tables, footnotes, bibliography and appendices.

Raghd Rostom

August 2019

Human Cellular Genetics of Innate Immunity

Raghd Rostom

The type I interferon response is a key part of the innate immune system, responding to infection and inducing an antiviral intracellular state. While there is known to be variability in this signalling pathway between individuals, alongside cell-to-cell heterogeneity in a genetically identical cell population, the basis of this variation is not fully understood.

In this PhD, I established large-scale single-cell RNA sequencing experiments to study cellular variation in the innate immune response in fibroblasts of 70 healthy human individuals from the HipSci initiative. Chapter 2 describes optimisation of stimulation conditions to induce an antiviral response, and the experimental work carried out on the panel of donors.

In Chapter 3, I analyse heterogeneity in resting (unstimulated) fibroblasts. By comparing to *ex vivo* skin data containing multiple cell types, I confirm the relative homogeneity of the *in vitro* cultured fibroblasts used, mapping to one sub-population of *ex vivo* skin fibroblasts. Using matched whole exome sequencing data, somatic mutations in sub-populations of cells within each donor were detected, and clonal populations identified. A novel computational method, cardelino, was developed for inference of the clonal tree configuration and the clone of origin of individual cells that have been assayed using scRNA-seq. Applying cardelino to 32 fibroblast lines identifies hundreds of differentially expressed genes between cells from different somatic clones, with cell cycle and proliferation pathways frequently enriched.

Returning to innate immunity, Chapters 4 and 5 centre on variability in the type I interferon response. I first describe work linking variability in the innate immune response and evolutionary divergence across mammalian species. Focusing on human variability, the large dataset described above is used to characterise the innate immune response at single cell resolution, elucidating the dynamics of the response across donors in Chapter 4. Chapter 5 describes the application of quantitative trait loci approaches to innate immune phenotypes. This work characterises both inter- and intra-individual heterogeneity in innate immunity.

Acknowledgements

I am deeply grateful to Sarah Teichmann and Oliver Stegle for giving me the opportunity to undertake this research, and for the mentorship throughout. Their immense knowledge, motivation and drive inspired me to take on this PhD, and continues to inspire me going forward. I would also like to thank the members of both the Teichmann and Stegle groups for their support over the years, both scientifically and socially, without whom day-to-day research would have been an entirely different experience. In particular, I would like to express my gratitude to Tzachi Hagai and Davis McCarthy, whose advice and assistance along the way had a huge impact upon my PhD, along with Mike Stubbington and Kerstin Meyer for their invaluable research guidance.

I gratefully acknowledge the funding received towards my PhD from the Biomedical and Biological Sciences Research Council (BBSRC) and the Wellcome Sanger Institute (WSI), along with my second research home at the European Molecular Biology Laboratory - European Bioinformatics Institute (EMBL-EBI).

My deepest appreciation goes to my friends, both nearby and far away, for keeping me sane and reminding me about the world outside of research. They have truly shaped the non-scientific part of my PhD experience, both in Cambridge and beyond.

Finally, I could not have made it to this point without my loving family, who always believed in me and encouraged me to follow my dreams. Their inspiration and support throughout has motivated me to work hard and kept me going through challenging times.

Contributions

Chapter 1

The section on single cell RNA sequencing analysis was adapted from a review written with the input of Valentine Svensson, published in FEBS journal.

Chapter 2

Bulk RNA sequencing data for protocol optimisation was generated by Tzachi Hagai. During the expansion and stimulation of HipSci lines, invaluable support was provided by the Cellular Genotyping and Phenotyping facility. Data processing was conducted with the help of Davis McCarthy and the Cellular Genetics Informatics team, WSI.

Chapter 3

Primary skin data was generated by the lab of Muzlifah Haniffa.
The study of clonal structure in fibroblasts was carried out as part of a close collaboration with Davis McCarthy and Yuanhua Huang, who developed the computational method - cardelino - underpinning this analysis, and final figures for the paper. The full manuscript is included in Appendix B.

Chapter 4

The cross-mammalian dataset presented in Section 4.1 was produced by Tzachi Hagai. This work was published in Nature, 2018, and the full paper is included in Appendix C.

Chapter 5

QTL analysis was conducted using a pipeline developed by Marc Jan Bonder, and run with the support of Ni Huang.

Table of contents

List of figures

List of tables

Nomenclature

Acronyms / Abbreviations

AMD	Age-related macular degeneration
BASiCS	Bayesian analysis of single-cell sequencing
BBKNN	Batch balanged k nearest neighbours
CGI	CpG island
CyTOF	Cytometry by time of flight
DC	Diffusion component
DM	Distance to median
DPT	Diffusion pseudotime
eQTL	Expression quantitative trait loci
FACS	Fluorescence-activated cell sorting
FISH	Fluorescence in situ hybridisation
FPKM	Fragments per kilobase per million
GLM	Generalised linear model

GPLVM	Gaussian process latent variable model
GWAS	Genome wide association study
HipSci	Human Induced Pluripotent Stem Cell Initiative
IFNs	Interferons
IIG	Innate immune gene
IVT	*In vitro* transcription
LF	Lipofectamine
LMM	Linear mixed model
LPS	Lipopolysaccharide
LRT	Likelihood ratio test
MDS	Multidimensional Scaling
MNN	Mutual nearest neighbour
MST	Minimum spanning tree
NLRs	NOD-like receptors
LD	Linkage disequilibrium
PAMPs	Pathogen associated molecular patterns
PCA	Principal Component Analysis
pDCs	Plasmacytoid dendritic cells
Poly(I:C)	Polyinosinic:polycytidylic acid

PRRs	Pattern recognition receptors
RLRs	RIG-I-like receptors
ROS	Reactive oxygen species
RT	Reverse transcription
scDNA-seq	Single cell DNA sequencing
SCG	Single Cell Genotyper
scLVM	Single cell latent variable model
scMT-seq	Single-cell methylome and transcriptome sequencing
scRNA-seq	Single-cell RNA-sequencing
scRNA-seq	Single cell RNA sequencing
scRRBS	Single cell reduced representation bisulfite sequencing
SNN	Shared nearest neighbour
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variants
TLRs	Toll-like receptors
TPM	Transcripts per million
tSNE	t-Distributed Stochastic Neighbour Embedding
UMAP	Uniform manifold approximation and projection
UMIs	Unique Molecular Identifiers

WGCNA Weighted gene co-expression network analysis

ZIFA Zero-inflated factor analysis