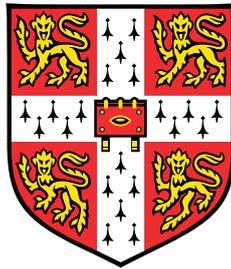


# Human cellular genetics of innate immunity



**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	xv
List of tables	xvii
Nomenclature	xix
<b>1 Introduction</b>	<b>1</b>
1.1 Human genetic variation . . . . .	1
1.1.1 The basis of genetic variation . . . . .	1
1.1.2 Approaches for studying human genetic variation . . . . .	4
1.2 Single cell RNA sequencing . . . . .	8
1.2.1 Evolution of single cell RNA sequencing technologies . . . . .	8
1.2.2 Analysis of scRNA-sequencing data . . . . .	12
1.3 The human innate immune system . . . . .	22
1.3.1 The type I interferon response . . . . .	22
1.3.2 Cell-to-cell heterogeneity in innate immunity . . . . .	25
1.3.3 Genetic variability in the innate immune response . . . . .	26
1.4 Using single-cell RNA sequencing data to study genetic variation in the innate immune response . . . . .	28

---

<b>2</b>	<b>Establishing a system for studying innate immune responses in human fibroblasts</b>	<b>29</b>
2.1	Defining optimal stimulation conditions . . . . .	30
2.1.1	Stimulation with Poly(I:C) . . . . .	32
2.1.2	Stimulation with interferons . . . . .	37
2.1.3	Innate immunity vs. apoptotic genes across conditions . . . . .	40
2.2	Large-scale stimulation experiments . . . . .	44
2.2.1	Expansion of lines . . . . .	44
2.2.2	Stimulation experiments . . . . .	44
2.3	Data processing . . . . .	47
2.4	Additional datasets . . . . .	47
2.4.1	Primary skin data . . . . .	47
2.4.2	Cross-mammalian data . . . . .	48
<b>3</b>	<b>Heterogeneity in primary human fibroblasts</b>	<b>49</b>
3.1	Introduction . . . . .	50
3.2	A comparison of <i>in vitro</i> and <i>ex vivo</i> fibroblasts . . . . .	52
3.3	Transcriptional heterogeneity in the unstimulated state . . . . .	55
3.3.1	An overview of the scRNA-seq dataset . . . . .	55
3.3.2	Clustering analysis of unstimulated fibroblasts . . . . .	57
3.4	Identifying common variants and somatic mutations in scRNA-seq data	61
3.4.1	Cardelino: a method for assigning cells to clones using scRNA-seq data . . . . .	61
3.4.2	Mutational analysis of <i>in vitro</i> fibroblasts . . . . .	63
3.4.3	Transcriptional analysis of <i>in vitro</i> fibroblasts . . . . .	67
3.5	Discussion . . . . .	73

<b>4</b>	<b>Cell-to-cell variability in the innate immune response</b>	<b>75</b>
4.1	Introduction . . . . .	76
4.2	Innate immune variability: a cross-mammalian study . . . . .	77
4.2.1	Transcriptional divergence in immune response . . . . .	77
4.2.2	Cell-to-cell variability in immune response . . . . .	78
4.2.3	Transcriptional divergence and variability of cytokines . . . . .	81
4.3	Characterising the Type I interferon response in human fibroblasts . . . . .	84
4.3.1	Single-cell RNA-sequencing data . . . . .	84
4.3.2	The temporal dynamics of the response . . . . .	84
4.3.3	Defining gene modules in the innate immune response . . . . .	90
4.4	Discussion . . . . .	96
<b>5</b>	<b>Inter-individual variability in the innate immune response</b>	<b>99</b>
5.1	Introduction . . . . .	100
5.2	Variance partitioning of gene expression . . . . .	103
5.3	eQTL analysis on bulk RNA-seq data . . . . .	104
5.4	QTL analysis on single cell phenotypes . . . . .	109
5.4.1	Mean expression . . . . .	109
5.4.2	Other response phenotypes . . . . .	112
5.5	Characterisation of QTL innate immune genes . . . . .	114
5.6	Discussion . . . . .	116
<b>6</b>	<b>Concluding remarks</b>	<b>119</b>
	<b>References</b>	<b>123</b>
	<b>Appendix A Overview of HipSci fibroblast lines</b>	<b>153</b>
	<b>Appendix B Heterogeneity in primary human fibroblasts</b>	<b>157</b>

Appendix C Manuscript: <i>Cardelino: Integrating whole exomes and single-cell transcriptomes to reveal phenotypic impact of somatic variants</i>	167
Appendix D Manuscript: <i>Gene expression variability across cells and species shapes innate immunity</i>	241
Appendix E Innate immune response modules	263

# List of figures

1.1	Mechanisms of DNA damage and repair. . . . .	3
1.2	Genome-wide association study design. . . . .	6
1.3	The expression quantitative trait loci (eQTL) approach . . . . .	7
1.4	Overview of analysis methods for the interpretation of scRNA-seq data. . . . .	11
1.5	Induction of the Type I Interferon response . . . . .	23
1.6	Interferons and inter-cellular communication in the immune response . . . . .	24
2.1	Effects of the poly(I:C) transfection procedure. . . . .	34
2.2	Response to poly(I:C) stimulation over time in two individuals. . . . .	36
2.3	Response to different interferon stimulations. . . . .	39
2.4	Response to IFN- $\beta$ stimulation over time in two individuals. . . . .	41
2.5	Comparison of control, poly(I:C) and IFN- $\beta$ treated cells with time. . . . .	43
2.6	An overview of stimulation experiments on HipSci fibroblast lines. . . . .	46
3.1	An overview of a pilot droplet scRNA-seq dataset. . . . .	53
3.2	Comparison of <i>in vitro</i> and <i>ex vivo</i> fibroblasts . . . . .	54
3.3	Quality control of scRNA-seq data. . . . .	56
3.4	Integration of scRNA-seq batches with Seurat - unstimulated cells. . . . .	58
3.5	Clustering analysis of unstimulated fibroblasts. . . . .	60
3.6	Overview and validation of the cardelino model. . . . .	64

3.7	Characterisation of mutational and clonal structure in 32 fibroblast lines	68
3.8	Clone-specific transcriptome profiles reveal gene expression differences for joxm, one example line. . . . .	70
3.9	Signatures of transcriptomic clone-to-clone variation across 31 lines. . .	72
4.1	Response divergence across species in innate immune response. . . . .	79
4.2	Cell-to-cell variability versus response divergence across species and conditions. . . . .	80
4.3	Promoter architecture versus transcriptional divergence and variability.	81
4.4	Transcriptional divergence in genes of different functional categories. . .	82
4.5	Cell-to-cell variability levels in cytokines, transcription factors and kinases.	83
4.6	Integration of scRNA-seq batches with Seurat - all stimulation conditions.	85
4.7	A pseudotime of poly(I:C) and interferon response pathways. . . . .	87
4.8	Functional classification of innate immune genes. . . . .	88
4.9	Expression of innate immune genes across response pseudotime. . . . .	89
4.10	Modules of co-expressed genes in the response to IFN- $\beta$ . . . . .	94
4.11	Modules of co-expressed genes in the response to poly(I:C). . . . .	95
5.1	Variance partitioning of gene expression across the scRNA-seq dataset.	103
5.2	Overlap of eQTLs across stimulation conditions in bulk RNA-seq data.	106
5.3	Expression of DDX1 varies with genotype and conditions. . . . .	107
5.4	Overlap of eQTLs across stimulation conditions in scRNA-seq derived 'pseudobulk' data. . . . .	110
5.5	Detection of a TRIM25 eQTL in scRNA-seq data. . . . .	111
5.6	Detection of a ZC3HAV1 cell proportion QTL in scRNA-seq data. . . .	113
5.7	Investigation of significant genes across QTL approaches. . . . .	115
E.1	Modules of co-expressed innate immune response genes using WGCNA.	279

# List of tables

2.1	Enrichment of apoptotic v.s. IFN response genes in response to poly(I:C).	37
2.2	Enrichment of apoptotic v.s. IFN response genes in response to IFN- $\beta$ .	40
4.1	Enrichment of IIGs in modules of co-expressed genes in the IFN- $\beta$ response.	93
4.2	Enrichment of IIGs in modules of co-expressed genes in the poly(I:C) response. . . . .	93
5.1	Phenotypes derived from scRNA-seq data. . . . .	102
5.2	Significant eQTL hits from bulk RNA-seq - known IIGs. . . . .	108
5.3	Significant eQTL hits from scRNA-seq 'pseudobulk' values - known IIGs.	109
A.1	Overview of HipSci lines used in stimulation experiments. . . . .	153
B.1	Marker genes of <i>ex vivo</i> skin clusters . . . . .	158
B.2	GO term enrichment in unstimulated fibroblast clusters . . . . .	162
E.1	GO term enrichment in IFN- $\beta$ response gene modules . . . . .	263
E.2	GO term enrichment poly(I:C) response gene modules . . . . .	269



# Nomenclature

## Acronyms / Abbreviations

AMD Age-related macular degeneration

BASiCS Bayesian analysis of single-cell sequencing

BBKNN Batch balanced 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