UNIVERSITY OF CALIFORNIA

Los Angeles

Methods for the Analysis and Interpretation of Single Cell RNA Sequencing Data

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy
in Molecular Biology

by

Feiyang Ma

2020

ABSTRACT OF THE DISSERTATION

Methods for the Analysis and Interpretation of Single Cell RNA Sequencing Data

by

Feiyang Ma
Doctor of Philosophy in Molecular Biology
University of California, Los Angeles, 2020
Professor Matteo Pellegrini, Chair

3’ RNA sequencing provides an alternative to whole transcript analysis. However, we do not know a priori the relative advantage of each method. Thus, a comprehensive comparison between the whole transcript and the 3’ method is needed to determine their relative merits. Single cell RNA sequencing (scRNA-seq) enables the profiling of the transcriptomes of individual cells. Cell type identification is one of the major goals in scRNA-seq. Current methods for assigning cell types have several limitations, such as unwanted sources of variation that influence clustering and a lack of canonical markers for certain cell types. Thus, new methods need to be developed. We first used two commercially available library preparation kits, the KAPA Stranded mRNA-seq kit (traditional method) and the Lexogen QuantSeq 3’ mRNA-seq kit (3’ method), to determine the advantages and disadvantages of these two approaches. We found that the 3’ RNA-seq method detected more short transcripts than the whole transcript method. With regard to differential expression analysis, we found that the whole transcript method detected more differentially expressed genes, regardless of the level of sequencing depth. We then developed ACTINN (Automated Cell Type Identification using Neural Networks), which employs a neural network to predicts cell types for scRNA-seq datasets. We trained and tested ACTINN on multiple datasets, the results showed that ACTINN is fast and accurate, and should therefore be a useful tool to complement existing scRNA-seq pipelines. Lastly, we performed scRNA-seq to study gene networks associated with host defense comparing lesions from reversal reaction vs. lepromatous lesions from leprosy patients. We constructed an antimicrobial ecosystem by integrating the *IFNG* and *IL1B* antimicrobial targets with the cell-cell co-abundance in lesions, which revealed that the interaction of dendritic cells, macrophages, T cells, keratinocytes and fibroblasts contributes to the capacity of granulomas to eliminate the pathogen in leprosy.

The dissertation of Feiyang Ma is approved.

Robert L. Modlin

M. Luisa Iruela-Arispe

Xia Yang

Jingyi Li

Matteo Pellegrini, Committee Chair

University of California, Los Angeles

2020

Table of Contents

List of Figures

List of Tables

Acknowledgements

I would like to thank my advisors, Dr. Matteo Pellegrini and Dr. Robert Modlin (unofficial), for their unwavering supports for the research projects, it wouldn't have been possible without them. I would like to thank Dr. Luisa Arispe for assisting with research and being a great collaborator. I would like to thank my committee members, Dr. Xia Yang and Dr. Jingyi Li, for their advice and support in research. I want to thank the members of Pellegrini, Modlin, Arispe and other collaborating labs, they helped a lot as I grow and learn at UCLA and made my time enjoyable. I am also highly thankful of my roommates, friends, classmates and basketball mates, they brought numerous joys in daily life. I would also like to acknowledge my previous mentors who helped mold me into a scientist, including Drs. Mingchun Li, Lingyi Chen, Qilin Yu and In-Hyun Park. Lastly, I want to thank my family, including my fiancée, for their unlimited encouragement and supports.

Chapter 2 is adapted from Ma F, Fuqua BK, Hasin Y, Yukhtman C, Vulpe CD, Lusis AJ, Pellegrini M. "A comparison between whole transcript and 3' RNA sequencing methods using Kapa and Lexogen library preparation methods." *BMC Genomics*. 2019. 20(1):9. doi: 10.1186/s12864-018-5393-3. The work was supported by the U.S. Department of Energy Office of Science, Office of Biological and Environmental Research program under award number DE-FC02-02ER63421 and National Institutes of Health under award number R01GM083198. Author contributions were: MP, AJL and CDV designed the work; BKF performed the animal husbandry; BKF and CY prepared the sequencing library; FM and YH processed the transcriptome data and did bioinformatics analyses; FM and MP drafted the article; All authors read and approved the final manuscript.

Chapter 3 is adapted from Ma F, Pellegrini M. ACTINN: Automated Identification of Cell Types in Single Cell RNA Sequencing. *Bioinformatics*. 2020 Jan 15;36(2):533-538. doi: 10.1093/bioinformatics/btz592. This work was supported by the U.S. Department of Energy Office of Science, Office of Biological and Environmental Research program under Award Number DE-FC02-02ER63421 and Whitcome Fellowship from UCLA. Author contributions were: Conceptualization, Methodology, Analysis, Writing, Visualization, Validation - FM, MP; Supervision, Funding Acquisition - MP.

Chapter 4 is partly adapted from a manuscript in preparation. I would like to thank everyone who contributed to the work including Travis Hughes, Rosane MB Teles, Priscila Andrade, Bruno JA Silva, Tran Do, Marc Wadsworth, Maria Teresa Ochoa, Euzenir Sarno, Luisa Arispe, Bryan Bryson, Alex Shalek, Barry Bloom, Johann Gudjonsson, Matteo Pellegrini and Robert L. Modlin. Author contributions were: Conceptualization, Visualization, Project Administration - FM, MP, RLM; Methodology, Formal Analysis, Investigation - FM, TH, RMT, PA, BJS, TD, MW, LA, BB, AS, BB, JG, MP, RLM; Writing - FM, MP, RLM; Resource - MTO, ES; Supervision: MP, RLM; Funding Acquisition - MP RLM.

VITA

EDUCATION AND TRAINING

2012-2016	B.S.	Nankai University.
2013-2016	Research Assistant.	Nankai University. Advisor: Mingchun Li.
2014-2014	Research Assistant.	Yale University. Advisor: In-Hyun Park.
2016-	Ph.D. Student.	UCLA. Advisor: Matteo Pellegrini.

AWARDS AND HONORS

2013	National Encouragement Scholarship, Nankai University.
2014-2016	National Innovation Experiment Grant for College Students, Nankai University.
2016	Excellent Dissertation for Bachelor's Degree, Nankai University.
2019-2020	Whitcome Fellowship, UCLA.

PUBLICATIONS

1. Yu Q, Zhang B, **Ma F**, Jia C, Xiao C, Zhang B, Xing L, Li M. Novel mechanisms of surfactants against Candida albicans growth and morphogenesis. *Chemico-Biological Interactions.* 2015 Feb 5; 227:1-6. doi: 10.1016/j.cbi.2014.12.014.

2. **Ma F**, Zhang Y, Wan Y, Wan Y, Miao Y, Ma T, Yu Q, Li M. Role of Aif1 in regulation of cell death under environmental stress in Candida albicans. *Yeast*. 2016 Sep; 33(9):493-506. doi: 10.1002/yea.3167.

3. Hilfenhaus G, Nguyen DP, Freshman J, Prajapati D, **Ma F**, Song D, Ziyad S, Cuadrado M, Pellegrini M, Bustelo XR, Iruela-Arispe ML. Vav3-induced cytoskeletal dynamics contribute to heterotypic properties of endothelial barriers. *J Cell Biol*. 2018;217(8):2813–2830. doi: 10.1083/jcb.201706041.

4. McDonald AI, Shirali AS, Aragón R, **Ma F**, Hernandez G, Vaughn DA, Mack JJ, Lim TY, Sunshine H, Zhao P, Kalinichenko V, Hai T, Pelegrini M, Ardehali R, Iruela-Arispe ML. Endothelial Regeneration of Large Vessels is a Biphasic Process Driven by Local Cells with Distinct Proliferative Capacities. *Cell Stem Cell*. 23(2):210-225. doi: 10.1016/j.stem.2018.07.011.

5. **Ma F**, Fuqua BK, Hasin Y, Yukhtman C, Vulpe CD, Lusis AJ, Pellegrini M. A comparison between whole transcript and 3' RNA sequencing methods using Kapa and Lexogen library preparation methods. *BMC Genomics*. 2019;20(1):9. doi: 10.1186/s12864-018-5393-3.

6. Montoya DJ, Andrade P, Silva BJA, Teles RMB, **Ma F**, Bryson B, Sadanand S, Noel T, Lu J, Sarno E, Arnvig KB, Young D, Lahiri R, Williams DL, Fortune S, Bloom BR, Pellegrini M, Modlin RL. Dual RNA-Seq of Human Leprosy Lesions Identifies Bacterial Determinants Linked to Host Immune Response. *Cell Rep*. 2019 Mar 26;26(13):3574-3585.e3. doi: 10.1016/j.celrep.2019.02.109.

7. Bonora G, Rubbi L, Morselli M, **Ma F**, Chronis C, Plath K, Pellegrini M. DNA methylation estimation using methylation-sensitive restriction enzyme bisulfite sequencing (MREBS). *PLoS One.* 2019 Apr 4;14(4):e0214368. doi: 10.1371/journal.pone.0214368.

8. Weiss DI, **Ma F**, Merleev AA, Maverakis E, Gilliet M, Balin SJ, Bryson BD, Ochoa MT, Pellegrini M, Bloom BR, Modlin RL. IL-1β Induces the Rapid Secretion of the Antimicrobial Protein IL-26 from Th17 Cells. *J Immunol*. 2019 Jun 24. pii: ji1900318. doi: 10.4049/jimmunol.1900318.

9. R Andrade P, Mehta M, Lu J, M B Teles R, Montoya D, O Scumpia P, Nunes Sarno E, Ochoa MT, **Ma F**, Pellegrini M, Modlin RL. The cell fate regulator NUPR1 is induced by Mycobacterium leprae via type I interferon in human leprosy. *PLoS Negl Trop Dis*. 2019 Jul 25;13(7):e0007589. doi: 10.1371/journal.pntd.0007589. PMID: 31344041; PMCID: PMC6684084.

10. **Ma F**, Pellegrini M. ACTINN: Automated Identification of Cell Types in Single Cell RNA Sequencing. *Bioinformatics*. 2020 Jan 15;36(2):533-538. doi: 10.1093/bioinformatics/btz592.

11. Koseler A, **Ma F**, Kilic ID, Morselli M, Kilic O, Pellegrini M. Genome-wide DNA Methylation Profiling of Blood from Monozygotic Twins Discordant for Myocardial Infarction. *In Vivo*. 2020 Jan-Feb;34(1):361-367. doi: 10.21873/invivo.11782. PMID: 31882500; PMCID: PMC6984093.

12. Browne L, Mead A, Horn C, Chang K, A Celikkol Z, L Henriquez C, **Ma F**, Beraut E, S Meyer R, Sork VL. Experimental DNA Demethylation Associates with Changes in Growth and Gene Expression of Oak Tree Seedlings. *G3 (Bethesda)*. 2020 Mar 5;10(3):1019-1028. doi: 10.1534/g3.119.400770. PMID: 31941723; PMCID: PMC7056980.

13. Wang H, Jiang H, Teles RMB, Chen Y, Wu A, Lu J, Chen Z, **Ma F**, Pellegrini M, Modlin RL. Cellular, Molecular, and Immunological Characteristics of Langhans Multinucleated Giant Cells Programmed by IL-15. *J Invest Dermatol*. 2020 Feb 22. pii: S0022-202X(20)30158-5. doi: 10.1016/j.jid.2020.01.026.

14. Presicce P, Cappelletti M, Senthamaraikannan P, **Ma F**, Morselli M, Jackson CM, Mukherjee S, Miller LA, Pellegrini M, Jobe AH, Chougnet CA, Kallapur SG. TNF-Signaling Modulates Neutrophil-Mediated Immunity at the Feto-Maternal Interface During LPS-Induced Intrauterine Inflammation. *Front Immunol*. 2020 Apr 3;11:558. doi: 10.3389/fimmu.2020.00558. PMID: 32308656; PMCID: PMC7145904.

15. Kelly-Scumpia KM, Choi A, Shirazi R, Bersabe H, Park E, Scumpia PO, Ochoa MT, Yu J, **Ma F**, Pellegrini M, Modlin RL. ER Stress Regulates Immunosuppressive Function of Myeloid Derived Suppressor Cells in Leprosy that Can Be Overcome in the Presence of IFN-γ. *iScience*. 2020 Apr 12;23(5):101050. doi: 10.1016/j.isci.2020.101050.

16. Morgan RA, **Ma F**, Unti MJ, Brown D, Ayoub PG, Tam C, Lathrop L, Aleshe B, Kurita R, Nakamura Y, Senedheera S, Wong RL, Hollis RP, Pellegrini M, Kohn DB. Creating New β-globin-Expressing Lentiviral Vectors by High-Resolution Mapping of Locus Control Region Enhancer Sequences. *MOL THER-METH CLIN D*. 2020. doi: 10.1016/j.omtm.2020.04.006.

Chapter 1 - Introduction

RNA sequencing

High-throughput RNA sequencing (RNA-seq) is a powerful tool to characterize and quantify transcriptomes and is now widely used in biomedical research. RNA-seq is primarily used to quantify the abundance and relative changes in gene expression across sample groups [1]. It enables a relatively unbiased analysis of the transcriptome, and has single base pair resolution, a wide dynamic range of detection, and low background noise [2]. Moreover, the cost of RNA-seq is continuously dropping as the cost of sequencing decreases, enabling varied investigations of molecular biology in a more precise and comprehensive manner than is possible with competing technologies [1].

Since the initial application of RNA-seq, many library preparation methods and sequencing platforms have been established, resulting in a number of choices for users. In the classic whole transcript method, extracted mRNAs are first randomly sheared into fragments, which are then reverse transcribed into cDNAs (Figure 1-1). As cDNA fragments are sequenced, the number of reads corresponding to each transcript is proportional to the number of cDNA fragments rather than the number of transcripts. Since longer transcripts are generally sheared into more fragments, more reads will be assigned to them than shorter transcripts. Consequently, when carrying out differential expression analysis, the differentially expressed genes are more likely to be enriched for longer than shorter transcripts, as the statistical power is higher for longer transcripts due to the larger counts. Recently, new 3’ RNA-seq methods, such as Tag-seq [3] and QuantSeq [4], have been developed to minimize this bias. I n the 3’ RNA-seq method, mRNAs are not fragmented before reverse transcription. Instead, the cDNAs are only reverse transcribed from the 3′ end of the mRNAs, and only one copy of cDNA is generated for each transcript. Thus, when the cDNAs are sequenced, the number of reads directly reflects the number of transcripts of a certain gene, and the longer and shorter transcripts should have the same coverage of reads. In chapter 2, we present a comparison between whole transcript and 3' RNA sequencing methods using Kapa and Lexogen library preparation methods.

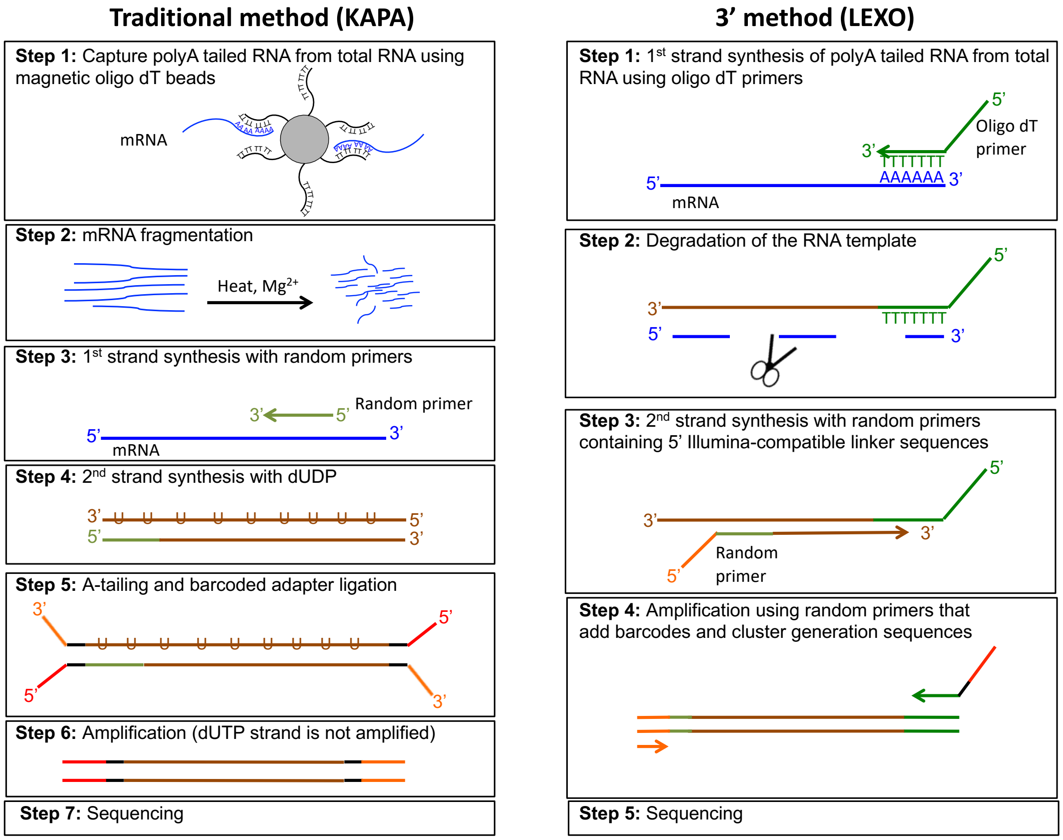

Traditional method (KAPA)
Step 1: Capture polyA tailed RNA from total RNA using magnetic oligo dT beads
mRNA
Step 2: mRNA fragmentation
Heat, Mg2+
Step 3: 1st strand synthesis with random primers
Random primer
3'
5'
5'
3'
mRNA
Step 4: 2nd strand synthesis with dUDP
U U U U U U U U U
Step 5: A-tailing and barcoded adapter ligation
Step 6: Amplification (dUTP strand is not amplified)
Step 7: Sequencing
3' method (LEXO)
Step 1: 1st strand synthesis of polyA tailed RNA from total RNA using oligo dT primers
Oligo dT primer
TTTTTTT
AAAAAA
mRNA
Step 2: Degradation of the RNA template
Step 3: 2nd strand synthesis with random primers containing 5' Illumina-compatible linker sequences
Random primer
Step 4: Amplification using random primers that add barcodes and cluster generation sequences
Step 5: Sequencing

Figure 1-1. Key library preparation steps for the Trad-KAPA (left) and 3'-LEXO (right) methods.

Single cell RNA sequencing

Single cell RNA sequencing (scRNA-seq) enables the profiling of the transcriptomes of individual cells, thus characterizing the heterogeneity of samples in manner that was not possible using traditional bulk RNA-seq [5]. However, scRNA-seq experiments typically yield high volumes of data, especially when the number of cells is large (often many thousands). Thus, fast and efficient computational methods are essential for scRNA-seq analyses.

One common goal of scRNA-seq analyses is to identify the cell type of each individual cell that has been profiled. To accomplish this, typically cells are first grouped into different clusters in an unsupervised way, and the number of clusters allows us to approximately determine how many distinct cell types are present in the sample. Each cluster should contain cells with similar expression profiles, and so the aggregated profile of a cluster increases the signal to noise of the expression estimates. To attempt to interpret the identity of each cluster, marker genes are found as those that are uniquely highly expressed in a cluster, compared to all the other clusters. These canonical markers are then used to assign the cell types for the clusters, by cross referencing the markers with lists of previously characterized cell type specific markers. While this process is able to identify cell types, there are some limitations: 1. Since the clustering method is unsupervised, all sources of variation influence the formation clusters, including effects that are not directly related to cell types such as differential expression induced by cell cycles. 2. It is often difficult to find an optimal match between the marker genes associated with each cluster and the canonical markers for specific cell types. Moreover, depending on the clustering parameters used, one cluster might contain multiple cell types, or one cell type could be split into multiple clusters. 3. Using canonical markers to assign cell types requires background knowledge of cell type specific markers, and sometimes these are not well characterized or difficult to find in the literature. Moreover, some canonical markers may be expressed by more than one cell type, and some cell types may have no known markers. 4. The same types of cells processed by two distinct scRNA-seq techniques tend to cluster separately due to technical batch effects, which complicates cell type identification in composite datasets. 5. Cell subtypes are often very similar to each other, which limits efforts to separate them accurately into different clusters. To overcome many of the limitations of existing approaches, new methods need to be developed.

Neural network

Artificial neural networks provide a popular framework for machine learning algorithms which can be used to interpret complex datasets. An artificial neural network is connected by multiple layers of neurons, like biological neural networks, each neuron contains a signal that will be transmitted to the neurons in the next layer (Figure 1-2). In artificial neural networks, the signal is a real number. In signal transmission, a linear function is applied to the signal from last layer, and the output is transmitted to the next layer. Recently, artificial neural networks have been widely used in many fields, including for the analysis of scRNA-seq data [6-9]. Since the output data from scRNA-seq is feature-enriched and well-structured, it is well suited as an input for neural networks. In chapter 3, we present ACTINN (Automated Cell Type Identification using Neural Networks) for scRNA-seq cell type identification.

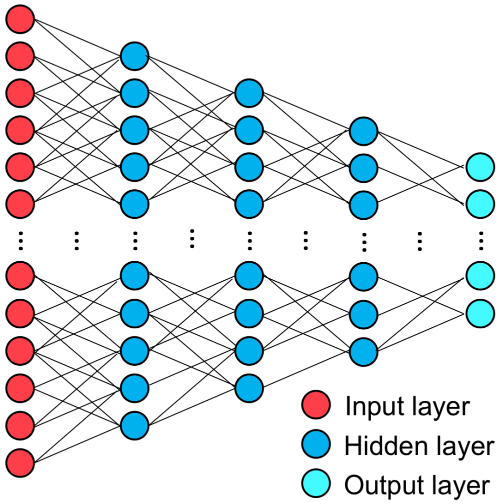


Figure 1-2. Neural network configuration.

Antimicrobial response in leprosy

The hallmark of the chronic inflammatory response to a foreign substance that has resisted destruction by an acute inflammatory response is the granuloma. In the most cited article on granulomas, Gordon defined granulomas as structures “which are formed by the immune-mediated recruitment of white blood cells, and particularly rich in macrophages" [10]. In the context of infectious diseases, the function of the granuloma is to sequester and degrade microbial pathogens that have evaded the early immune response.

Leprosy offers an attractive model to investigate the mechanisms by which the human immune system combats intracellular bacteria as the disease presents as a clinical/immunologic spectrum [11]. Because it manifests as a spectrum of disease in skin, the dynamics are accessible to study, in contrast to tuberculosis granulomas. At one end of the disease spectrum, tuberculoid leprosy typifies the host's antimicrobial response, which controls the pathogen: there are few lesions; *Mycobacterium leprae* bacilli are rare; and patients eliminate the infection. At the opposite end of the spectrum, lepromatous leprosy (L-lep) represents susceptibility to disseminated infection, with numerous skin lesions and abundant bacilli (Figure 1-3). The disease spectrum is dynamic, as patients may undergo a reversal reaction (RR), in which patients generally upgrade, either spontaneously or during chemotherapy, from the lepromatous to the tuberculoid pole. The structure of granulomas is distinct across the spectrum of leprosy. The

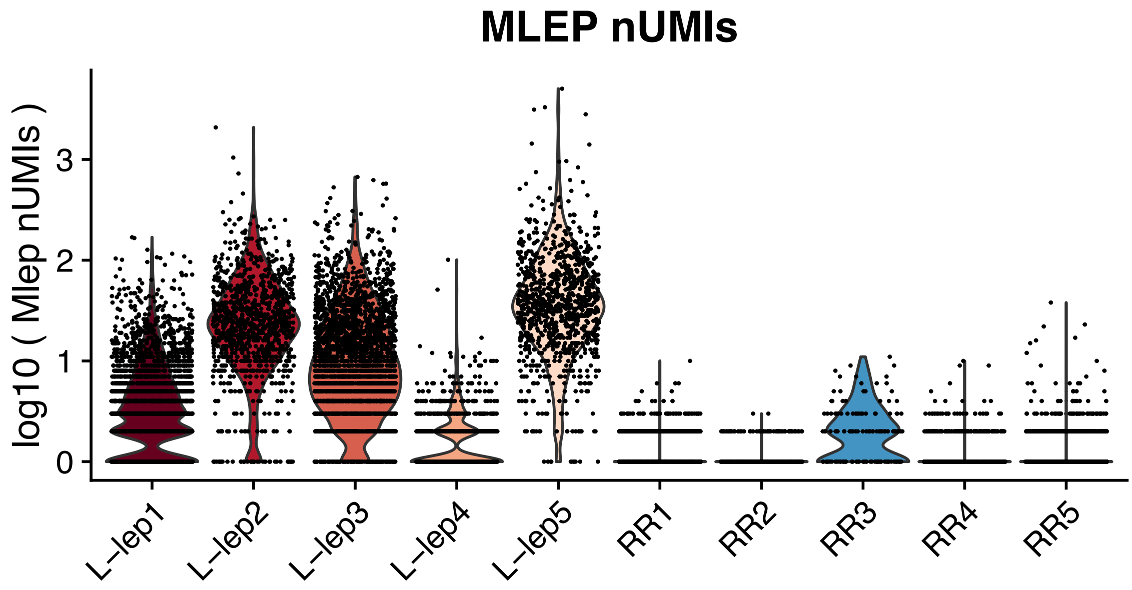


Figure 1-3. Number of M. leprae transcripts detected in individual cells for each patient.

granulomas in tuberculoid leprosy contain a core of mature macrophages with occasional multinucleated giant cells. These granulomas are organized with lymphocytes forming a mantle zone at the periphery of the granuloma. Granulomas in RR lesions are histologically similar to those in tuberculoid leprosy with the presence of intercellular edema. In lepromatous leprosy, the granulomas are disorganized, immature lipid-laden macrophages are prominent with lymphocytes scattered throughout.

The study of leprosy lesions has provided insight regarding the host immune response to intracellular bacteria and the architecture of granulomas. Through various approaches, it has been possible to define functional subpopulations of human T cells [12-15] and macrophages [16], their microanatomic distribution as well as the patterns of cytokine secretion that influence the outcome of infections caused by pathogenic mycobacteria [17-20].

Given that the resolution of the granulomatous response requires destruction of the foreign invader, the antimicrobial mechanisms that result in the death of the pathogen are central to understanding how granulomas contribute to host defense. A few pathways have been identified by the study of human cells that can lead to an antimicrobial activity against intracellular mycobacteria. Through activation via TLRs and secretion of IFN-γ, the innate and adaptive immune systems trigger the vitamin D-dependent induction of the antimicrobial proteins encoded by *CAMP* and *DEFB4A* [16, 21, 22]. T cells release antimicrobial proteins encoded by *GNLY* and *IL26*, which can enter infected macrophages and exert a direct antimicrobial activity [13, 14, 23, 24]. These human pathways are not present in mice, which utilize other mechanisms such as the release of nitric oxide to kill mycobacteria. The advent of scRNA-seq provides an opportunity to elucidate the cell-cell networks that define antimicrobial responses at the site of infection. We used this approach to study and compare the immune responses in RR vs. L-lep patient skin lesions to gain insight into mechanisms of host defense used by granulomas to eliminate an intracellular bacterium.

Chapter 2 - A Comparison Between Whole Transcript and 3’ RNA Sequencing Methods using Kapa and Lexogen Library Preparation Methods

Abstract

3’ RNA sequencing provides an alternative to whole transcript analysis. However, we do not know a priori the relative advantage of each method. Thus, a comprehensive comparison between the whole transcript and the 3’ method is needed to determine their relative merits. To this end, we used two commercially available library preparation kits, the KAPA Stranded mRNA-seq kit (traditional method) and the Lexogen QuantSeq 3’ mRNA-seq kit (3’ method), to prepare libraries from mouse liver RNA. We then sequenced and analyzed the libraries to determine the advantages and disadvantages of these two approaches. We found that the traditional whole transcript method and the 3’ RNA-seq method had similar levels of reproducibility. As expected, the whole transcript method assigned more reads to longer transcripts, while the 3’ method assigned roughly equal numbers of reads to transcripts regardless of their lengths. We found that the 3’ RNA-seq method detected more short transcripts than the whole transcript method. With regard to differential expression analysis, we found that the whole transcript method detected more differentially expressed genes, regardless of the level of sequencing depth.

Introduction

High-throughput RNA-sequencing (RNA-seq) is a powerful tool to characterize and quantify transcriptomes and is now widely used in biomedical research. RNA-seq is primarily used to quantify the abundance and relative changes in gene expression across sample groups

[1]. It enables a relatively unbiased analysis of the transcriptome, and has single base pair resolution, a wide dynamic range of detection, and low background noise [2]. Moreover, the cost of RNA-seq is continuously dropping as the cost of sequencing decreases, enabling varied investigations of molecular biology in a more precise and comprehensive manner than is possible with competing technologies [1].

Since the initial application of RNA-seq, many library preparation methods and sequencing platforms have been established, resulting in a number of choices for users. In the classic whole transcript method, extracted mRNAs are first randomly sheared into fragments, which are then reverse transcribed into cDNAs (Figure 1-1). Although RNA-seq is generally considered unbiased, it is important to note that fragmentation and library construction can introduce some biases into RNA-seq results [2]. As cDNA fragments are sequenced, the number of reads corresponding to each transcript is proportional to the number of cDNA fragments rather than the number of transcripts. Since longer transcripts are generally sheared into more fragments, more reads will be assigned to them than shorter transcripts. Consequently, when carrying out differential expression analysis, the differentially expressed genes are more likely to be enriched for longer than shorter transcripts, as the statistical power is higher for longer transcripts due to the larger counts [25]. Recently, new 3’ RNA-seq methods, such as Tag-seq [3] and QuantSeq [4], have been developed to minimize this bias. In the 3’ RNA-seq method, mRNAs are not fragmented before reverse transcription. Instead, the cDNAs are only reverse transcribed from the 3′ end of the mRNAs, and only one copy of cDNA is generated for each transcript (Figure 1-1). Thus, when the cDNAs are sequenced, the number of reads directly reflects the number of transcripts of a certain gene, and the longer and shorter transcripts should have the same coverage of reads.

Since the establishment of 3’ RNA-seq, it has been used in many studies. For example, Meyer *et al*. used Tag-Seq to profile gene expression responses of coral larvae [3], Barbash *et al*. used QuantSeq to quantify gene expression in the human brain [26], and Oberlin *et al*. used QuantSeq in a genome-wide transcriptome and translatome analysis of Arabidopsis transposons [27]. In all the above-mentioned studies, the genome of the organism that was studied (coral, human and Arabidopsis) was already characterized. However, when little genomic information is available for the species, Tandonnet *et al*. found that classic RNA-seq methods worked better than 3’ RNA-seq methods in quantifying the transcriptome [28].

To determine whether to use the classic whole transcript RNA-seq method or the 3’ method for a large mouse study where the primary goal is to identify expression quantitative trait loci, we used both methods to prepare RNA-seq libraries from the livers of mice on two diets, an iron-loaded diet and a control diet. We used the KAPA Stranded mRNA-seq Kit (Trad-KAPA) to prepare libraries using the whole transcript method, and the Lexogen Quant-Seq 3’ mRNA-seq Library Prep Kit-FWD (3’-LEXO) to prepare 3′ libraries. We then sequenced the libraries on the Illumina platform. The sequencing results for the Trad-KAPA and 3’-LEXO libraries were compared to determine their relative advantages and disadvantages. We first mapped the reads to the mouse genome, and confirmed that the Trad-KAPA reads covered the whole transcript, while 3’-LEXO reads only covered the 3′ end. Next, we determined the number of reads assigned to transcripts with different lengths and then used sub-sampling to determine how sequencing depth affects the read distributions. We also compared the reproducibility of the two methods, and carried out differential expression analysis for both methods.

Materials and Methods

Animal Husbandry

Eight female SJL/J mice (cat #686, purchased from The Jackson Laboratory, Bar Harbor, ME) housed at 4 mice per cage were placed on an AIN-93G "control" diet containing 50ppm iron (cat #515005, Dyets, Bethlehem, PA) upon arrival at 3 weeks of age. At 6-weeks of age, one cage of these mice was changed to an AIN-93G "high iron" diet containing 2% carbonyl iron (cat #515007, Dyets). At 11 weeks of age, the mice were fasted starting at 6:30am, and tissues were collected between 11:30am and 1pm. Blood was taken from the retroorbital plexus under isoflurane anesthesia using a heparin-coated capillary tube, and then mice were perfused via the heart with ice-cold phosphate buffered saline to flush remaining blood from the tissues. Tissues were collected and frozen in liquid nitrogen and stored at -80°C until analysis.

Liver RNA purification

Total RNA was extracted from a 20mg piece of the large lobe of six livers (3 per diet group) using the Qiagen miRNeasy Mini kit (cat# 217004, Qiagen) per the manufacturer's instructions. In brief, samples were homogenized in QIAzol lysis reagent using a rotor stator homogenizer. Chloroform was added and the extract was vigorously shaken and then centrifuged at 12,000 g to phase separate the organic and aqueous phases. Total RNA was purified from the aqueous phase using the kit spin column. DNA was digested on-column per the manufacturer's instructions using the RNase-Free DNase Set (cat# 79254, Qiagen). RNA concentration was measured using the Qubit RNA BR Assay (cat# Q10211, Molecular Probes) and RNA integrity was measured with an Agilent 2200 Tapestation instrument using the Agilent RNA ScreenTape and Sample Buffer (cat#5067-5576 and cat#5067-5577, Agilent, Santa Clara, CA). All samples had RINe values greater than 8.

Library generation

Libraries were prepared from the extracted RNA using two different kits, the QuantSeq 3'mRNA-seq Library Prep Kit-FWD (cat #15, Lexogen, Vienna, Austria), denoted here as "3'-LEXO", and the KAPA Stranded mRNA-seq Kit (cat #KK8421, KAPA Biosystems, Wilmington, MA), denoted here as "Trad-KAPA", per the manufacturers' instructions using 1 µg of RNA per library.

For the Trad-KAPA libraries, RNA was heated in a thermocycler for 6 minutes at 94°C for the fragmentation step, and KAPA Pure Beads (cat #KK8002, KAPA Biosystems) were used for cDNA capture. For the Trad-KAPA adapter ligation reactions, aliquots of 700 nM stock adapters (prepared from 30 µM original stock, cat #KK8700, KAPA Biosystems) were added to give final adapter concentrations of 50 nM. Ten cycles of library amplification were performed, and the libraries were eluted in 23.5 uL 10 mM Tris-HCl (pH 8). The double stranded DNA concentration was quantified using two methods: the Qubit dsDNA BR Assay Kit (cat #Q32853, Molecular Probes), which gave concentrations ranging from 42.1 to 46.7 ng/ µL, and by the KAPA Library Quantification Kit (cat #KK4824, KAPA Biosystems), which gave values approximately 2.5 higher. The molar concentration of cDNA molecules in the individual Trad-KAPA libraries was calculated from the double stranded DNA concentration (as determined by the KAPA Library Quantification Kit) and the region average size (determined by analyzing each sample on an Agilent 2200 Tapestation instrument using the Agilent D1000 ScreenTape and Sample Buffer (cat#5067-5582 and cat#5067-5583, Agilent, Santa Clara, CA). Aliquots from each library were diluted to 10 nM cDNA molecules in 10 mM Tris-HCl (pH 8) + 0.01% Tween-20 (cat #P1379-25ML, Sigma, St. Louis, MO), and equal volumes were pooled to make the final pooled library for sequencing.

For the 3'-LEXO libraries, indices from the first two columns of the i7 Index Plate for QuantSeq/SENSE for Illumina adapters 7001-7096 (cat #044, Lexogen) were used, and 11 cycles of library amplification were performed. Libraries were eluted in 22 µL of the kit's Elution Buffer. The double stranded DNA concentration was quantified using the Qubit dsDNA HS Assay Kit (cat #Q32854, Molecular Probes), and by the KAPA Library Quantification Kit, both which gave similar concentrations for each sample that ranged from 1.7 to 4.3 ng/ µL. The molar concentration of cDNA molecules in the individual 3'-LEXO libraries was calculated from the double stranded DNA concentration and the region average size (determined by analyzing each sample on an Agilent 2200 Tapestation instrument using the Agilent High Sensitivity D1000 ScreenTape and Sample Buffer (cat#5067-5584 and cat#5067-5585, Agilent, Santa Clara, CA). Aliquots containing an equal number of nmoles of cDNA molecules from each library were pooled to give a pooled library with a concentration of 10 nM cDNA molecules. Per the manufacturer's advice, the final pool was purified once more (to remove any free primers to prevent index-hopping) by adding 0.9x volumes of PB and proceeding from Step 30 onwards in the QuantSeq User Guide protocol. The library was eluted in 22 µL of the kit's Elution Buffer.

Sequencing

The pooled libraries were sequenced in an Illumina HiSeq4000 instrument (Illumina, San Diego, CA).

Transcript Coverage

The reads were mapped with STAR 2.5.3a to the mouse genome (mm10 / GRCm38). After mapping, all 12 BAM files were used as input for RSeQC v2.6.4 to calculate transcript coverage. For visualization of the Unc50 gene coverage, control sample 1 BAM files from Trad-KAPA and 3'-LEXO were visualized in Integrative Genomics Viewer.

Reads subsampling

We randomly sampled 1, 2.5, 5, and 10 million reads that are uniquely mapped to a gene's exonic regions from each sample. We considered genes to be detected if they had at least 1 read. The transcript length was calculated by adding the lengths of all the exons from the gene.

Correlation between Trad-KAPA and 3'-LEXO samples

For comparison between samples sequenced by the same method, raw read counts were modified by the addition of 0.01 before log10 transformation, then Pearson correlation coefficients were calculated between each comparison. For comparisons between Trad-KAPA and 3'-LEXO samples, Trad-KAPA raw read counts were divided by transcript length and multiplied by 1000, then the samples were treated as comparison within one method.

Differential expression analysis

We used DESeq2 to find differentially expressed transcripts in control diet and iron-loaded diet samples for each sequencing depth. The FDR was adjusted to 0.05, and the other parameters were set to default. The number of overlapping differentially expressed transcripts in Trad-KAPA and 3'-LEXO was calculated. For 1, 2.5 and 5 million reads, the overlap between differentially expressed transcripts in subsampled pools and the initial 10 million read sample was computed. The log fold changes from DESeq2 were used to calculate the correlations between the two methods.

Real-time quantitative PCR

All primers are listed in Table 2-2. cDNA for real-time quantitative polymerase chain reaction (RT-qPCR) reactions was prepared with High Capacity cDNA Reverse Transcription Kit (cat# 4368814, Life Technologies) using the same liver RNA stock used for the Trad-KAPA and 3'-LEXO library synthesis. KAPA SYBR FAST qPCR reaction mix (cat# KK4611, Roche) was added with primers and run in triplicate on a LightCycler 480 Instrument (Roche). PCR products gave a strong single peak by melt curve analysis. For each mouse and transcript, housekeeping-normalized expression values were calculated as 2-(Cp GOI – Cp housekeeper), where GOI is the gene of interest and Cp is the cycle number where fluorescence reached a set threshold. Three housekeeping genes (TBP, Beta-actin, and HPRT) were selected to control for variation in cDNA amounts. Students' t-test was performed for each gene and housekeeper to compare expression levels between the three control and three iron loaded mice, and the average t-test p-value across all three housekeepers was calculated. For each gene, housekeeper, and animal, housekeeping-normalized expression values for each gene were then normalized to the average level in animals on the control diet by dividing each housekeeping-normalized expression value by the average control group housekeeping-normalized expression value. These fold change values versus control were then averaged for all three housekeepers used, to give a final average fold change value versus control for each gene.

Results

Library preparation and RNA-sequencing

We extracted RNA from the large lobe of the liver from 3 mice on an iron-loaded diet and 3 mice on an iron sufficient control diet and then prepared RNA-seq libraries using both the Trad-KAPA and 3'-LEXO methods for all six samples. An overview of the key library preparation steps for the two methods are described in Figure 1-1. After library preparation, we pooled and sequenced the libraries using single-end sequencing with 50 bp reads on an Illumina HiSeq4000 instrument (Illumina, San Diego, CA). We obtained an average of 22.9 million and 18.4 million reads for Trad-KAPA and 3'-LEXO libraries, respectively. The reads were mapped with STAR 2.5.3a [29] to the mouse genome (mm10 / GRCm38). 80% of the Trad-KAPA reads and 82% of the 3'-LEXO reads were uniquely mapped. As the percentages of mapped reads from the two methods were similar, we randomly sampled 10 million uniquely mapped reads in each sample for further analysis, to make sure that each library had the same sequencing depth.

3'-LEXO reads mapped to the 3' region

After sequencing and read mapping, we used RSeQC [30] to determine the distribution of the reads along transcripts. As expected, Trad-KAPA reads covered transcripts uniformly, with only a slight decrease in coverage at the 5' end (Figure 2-1A). By contrast, 3'-LEXO reads preferentially mapped to the 3' end. This suggests that most of the 3'-LEXO reads originated from the 3' region of the gene. The individual Trad-KAPA libraries (red lines) had very similar transcript coverage profiles, while the individual 3'-LEXO samples (blue lines) exhibited some variation near the middle of the transcript.

We show an example of the coverage differences between Trad-KAPA and 3'-LEXO in Figure 2-1B. The mouse Unc50 gene has 6 exons and encodes an inner nuclear membrane RNA binding protein. We used the integrative genomics viewer [31] to visualize Trad-KAPA and 3'-LEXO read coverage. Trad-KAPA re ads covered all the exons uniformly, with only a slight decrease in the 5' exon. There were also some Trad-KAPA reads that mapped to the introns of Unc50, suggesting that some of the introns are not fully spliced. By contrast, most of the 3'-LEXO reads mapped only to the last exon of the gene.

Trad-KAPA assigned more reads to longer transcripts

Since Trad-KAPA reads originated from the entire transcript while 3'-LEXO reads originated primarily from the 3' end, we expected that the Trad-KAPA libraries would generate more reads for longer transcripts while the 3'-LEXO libraries would produce equal numbers of reads for transcripts independently of their lengths. To determine whether this is the case, we selected transcripts that have a length range from 500 bp to 8500 bp and have at least 100 read counts, and measured the distribution of coverage levels. For Trad-KAPA libraries, median read counts increased with transcript length (Figure 2-2A), indicating that as expected these libraries generate more reads for longer transcripts. By contrast, the median read counts from 3'-LEXO libraries did not change significantly with length (Figure 2-2B). This is expected, since the strong 3' bias found in 3'-LEXO libraries is not significantly affected by transcript length. Thus, for datasets of the same sequencing depth, Trad-KAPA samples contain more reads from longer transcripts, while 3'-LEXO samples appear to be insensitive to transcript length.

3'-LEXO recovers more short transcripts as sequencing depth drops

To determine whether 3'-LEXO detects more short transcripts than Trad-KAPA as sequencing depth drops, we subsampled 1, 2.5 and 5 million uniquely mapped reads for all the samples, and determined how many transcripts with lengths ranging from 0 bp to 10000 bp were detected (Figure 2-3A). As sequencing depth dropped, shorter transcripts were detected less frequently than longer ones in both the Trad-KAPA and 3'-LEXO libraries. When the sequencing depth dropped to 5 million, we found that we detected about 300 more transcripts that are shorter than 1000 bp from the 3'-LEXO libraries than from the Trad-KAPA libraries. With only 2.5 million reads, the difference became even more significant, approaching about 400 transcripts. However, when the sequencing depth dropped to 1 million, the difference became smaller. For transcripts longer than 1000 bp and shorter than 2000 bp, as sequencing depth drops, the detection difference between Trad-KAPA and 3'-LEXO inverted, with 3'-LEXO libraries leading to the detection of slightly more transcripts. For transcripts longer than 2500 bp, while Trad-KAPA always detected slightly more transcripts than 3'-LEXO at all the sequencing depths, the differences were very small.

We also compared the 1, 2.5 and 5 million read depths to 10 million read depth to see how many transcripts were detected by each method as sequencing depth drops. As shown in Figure 2-3B, 3'-LEXO detected 10% more transcripts than Trad-KAPA for transcripts shorter than 1000 bp. For transcripts longer than 1000 bp and shorter than 3000 bp, 3'-LEXO only recovered slightly more than Trad-KAPA. For transcripts longer than 3000 bp, the two methods detected about the same percentage of transcripts.

Trad-KAPA and 3'-LEXO have similar levels of reproducibility

To compare the reproducibility of the two library preparation methods, we calculated the correlation within and between Trad-KAPA and 3'-LEXO samples. Biological replicates of samples made with each of the two protocols were correlated at comparable levels (Figure 2-4AC), with correlation coefficients around 0.95. The control and diet samples were also highly correlated in both cases (Figure 2-4BD), although slightly lower than that found for the biological replicates. Finally, we also compared libraries generated from the same RNA stock but with the two different library preparation methods (Figure 2-4EF), and found that the correlation coefficient was around 0.85. We found that Trad-KAPA detects some genes that are missed by 3'-LEXO (shown in the red rectangle area in Figure 2-4EF), but generally the agreement between the two libraries was quite high.

Trad-KAPA detects more differentially expressed genes

One major application of RNA sequencing is the identification of differentially expressed genes (DEGs). We used DESeq2 [32] to carry out differential expression analysis on the control and iron loaded diet samples with subsampling. We adjusted the FDR to 0.05 and detected 1982 and 1157 differentially expressed transcripts for Trad-KAPA and 3'-LEXO, respectively (Table 2-1). Among those transcripts, 882 were detected by both methods. As sequencing depth drops, the number of differentially expressed transcripts detected by Trad-KAPA and 3'-LEXO decreased, and this trend can also be seen in the MA plots in Supplementary Figure 2-1. However, samples sequenced by Trad-KAPA always resulted in more differentially expressed transcripts when comparing the two libraries at the same sequencing depth. Not surprisingly, more than 95% of the differentially expressed transcripts detected in the subsampled datasets were also detected in the analysis of the initial 10 million read dataset. These results indicate that Trad-KAPA libraries lead to a higher detection of differentially expressed transcripts compared to 3'-LEXO libraries, at all sequencing depths.

We also looked at the lengths of the differentially expressed transcripts detected by the two methods. As shown in Supplementary Figure 2-2, some short transcripts were only detected as differentially expressed in 3'-LEXO samples (blue bins). As the transcript length increases, the number of differentially expressed transcripts detected only by 3'-LEXO drops. By contrast, most of the longer transcripts were only detected as differentially expressed by Trad-KAPA. This may be due to the fact that Trad-KAPA assigned more reads to the longer transcripts, which gained enough statistical power to be detected as differentially expressed.

Validation of the differential expression analysis

To understand why some genes were only detected as significantly differentially expressed in one method, we selected DEGs (1100 from Trad-KAPA and 275 from 3'-LEXO) and compared their expression and log fold changes across both methods (Supplementary Figure 2-3). We found that most genes had higher expression and larger log fold changes in the method that detected them as significantly differentially expressed compared to the other method. However, we also found that the correlation coefficients for the log fold changes and expression levels are 0.87 and 0.83, indicating that the Trad-KAPA and 3'-LEXO methods overall yield consistent results. We compared the expression level of the DEGs detected in only one method to the expression level of the DEGs that were identified in common by both methods and found that these had on average 36% higher expression than the DEGs detected in only one method. Thus, we think the reason for genes being detected as DEGs in only one method was due to lower expression in the other method. This can be explained by the differences that the two methods use in assigning reads to the genes.

We also used RT-qPCR to examine the expression of a subset of the genes that were found to be detected only by either the Trad-KAPA or 3'-LEXO method (mean expression across all six samples [control and iron loaded] >10 by one method and <1 by the other). We tested 11 genes that were only detected by the 3'-LEXO method, and 7 genes that were only detected by the Trad-KAPA method (Table 2-2). Of note, for some of these genes with several reported splice variants, we used multiple primer sets but obtained similar results. For most of these genes, differential expression analysis gave different results for the two RNA-seq methods. The crossing point-PCR-cycle (Cp) values for 3 of the 3'-LEXO only genes were greater than 30. Of the 8 tested 3'-LEXO only genes that had Cp values less than 30, 5 genes' RT-qPCR fold change results comparing iron loaded to control diet agreed better with the 3'-LEXO results, 2 agreed better with the Trad-KAPA results, and 1 gave an intermediate result. All of the RT-qPCR results from the 7 tested Trad-KAPA only genes agreed better with the Trad-KAPA results. Thus, as expected, genes that were more highly detected by one method tended to give differential expression results that better agreed with RT-qPCR results.

Differential expression in iron metabolism

To validate if the differentially expressed genes detected by each method overlap in terms of biological function, we carried out functional enrichment analyses using the DEGs from both the Trad-KAPA and 3'-LEXO methods using KEGG pathways. We found that the enriched pathways determined from the data from the Trad-KAPA and 3'-LEXO largely overlapped, although there were some pathways specific to each method (Supplementary Figure 2-4AB). The overlapping pathways were related to amino acid and lipid metabolism. Lipid metabolism in particular has been previously reported to be affected by iron status [33]. We also performed differential expression analysis on previously published microarray data from iron loaded and control C57BL/6J mice livers [34] and obtained 792 DEGs. We then performed functional enrichment analysis on these DEGs in the same way as for the RNA-seq results (Supplementary Figure 2-4C). Again, pathways related to amino acid and lipid metabolism were shared between all 3 analyses.

To further determine if the results from both methods were consistent, we examined 13 genes known to be involved in iron metabolism by RT-qPCR, and compared the results with those from both the Trad-KAPA and 3'-LEXO. All 13 genes tested were well represented in both RNA-seq data sets and had Cp values less than 30 by qPCR. 8 genes were found to have significantly increased expression in the iron loaded livers compared to controls by at least one of the methods (Table 2-2). Bmp6 and Hamp1 increased 5-6 fold. Atoh8, Smad7, and Id1 increased 3-4 fold. Lcn2 and Cp increased 2-3 fold in all studies. The results for Ftl1 differed between the methods, with Trad-KAPA giving no difference, 3'-LEXO giving a 3 fold increase, and RT-qPCR results about 2 fold increase. The expression of these genes has been reported previously to increase with iron loading [34, 35]. Two tested genes exhibited significantly decreased expression by at least one method. Bdh2 decreased 2-4 fold, and Hamp2 decreased 3-4 fold. The decreased expression of Bdh2 is in agreement with a previous study, but the Hamp2 results (found by all methods) were different than those previously reported for other mouse strains [36]. Finally, 3 genes (Hfe2, Slc11a2, and Tfrc) known to be involved in iron metabolism had little to no difference in expression reported at the mRNA level in the liver with iron loading and also had slight to no differences in expression by the three methods tested here [37, 38].

Thus, the results for both RNA-seq methods agreed well with both the RT-qPCR results and with previously reported studies.

Discussion

With the development and advancement of RNA-sequencing technology, many library preparation methods and sequencig platforms have become available. Here, we used a classic whole transcript RNA-seq method (Trad-KAPA) and a 3’ RNA-seq method (3’-LEXO) to prepare sequencing libraries from livers of iron-loaded diet and control diet mice, and sequenced the libraries on the Illumina platform. We then compared the sequencing results to determine the advantages and disadvantages of the two approaches.

We identified the gene body coverage of the Trad-KAPA and 3’-LEXO libraries by mapping the reads back to the genome. As expected, Trad-KAPA reads covered transcripts uniformly, with a slight decrease at the 5’ end. One reason for the decrease might be that the secondary structure of the mRNA can cause early termination of reverse transcription [39], making it difficult to reach the cap site (5’ end). It is also possible that many of the transcripts are partially degraded, so that the polyadenylation capture biases the coverage towards the 3’ end. By contrast, 3’-LEXO reads mapped mostly to the 3’ end. 3’-LEXO reads that mapped to the middle of the transcript showed significant coverage variation from library to library. The variation might be caused by the randomness in the reverse transcription start site on the cDNA. In the classic whole transcript method, mRNAs are first sheared into fragments, then the fragments are reverse transcribed to generate cDNAs. Hence, it is expected that the longer a transcript is, the more fragments it should have. The 3’ RNA-seq method however generates only one read for each transcript, so the number of reads directly reflects the level of gene expression.

We counted the reads mapped to transcripts that have lengths ranging from 500 bp to 8500 bp and found that Trad-KAPA libraries had more reads assigned to longer transcripts. By contrast, 3’-LEXO read counts remained uniform as transcript length increased.

As Trad-KAPA assigned more reads to longer transcripts and 3’-LEXO assigned a similar number of reads to transcripts with different lengths, we expected to see fewer short transcripts and more long transcripts detected by Trad-KAPA as sequencing depth drops. For transcripts shorter than 1000 bp, 3’-LEXO detected about 10% more than Trad-KAPA when sequencing depth dropped. However, for transcripts longer than 1000 bp, there was only a small difference between the number detected by Trad-KAPA and 3’-LEXO. Since a 3’ RNA-seq method only captures reads from the 3’ end of the mRNA, it is difficult for this method to detect differences in isoforms close to the 5’ end of longer genes. In our study, 15% of uniquely mapped Trad-KAPA reads contain splices, while only 6% of uniquely mapped 3’-LEXO reads contain splices. As a result, the 3’ RNA-seq method is not recommended for novel transcript or splice variant discovery. We also compared Trad-KAPA and 3’-LEXO reproducibility, and found that both methods showed very high reproducibility between biological replicates. When comparing the sequencing results generated with the same mouse using the Trad-KAPA versus 3’-LEXO methods, we found the two methods generally agreed with each other. Although there were a few transcripts detected only by Trad-KAPA, they turned out to be non-coding RNAs.

One major application of RNA-sequencing is to detect differentially expressed transcripts. We subsampled the reads generated by both the methods and carried out differential expression analysis using DESeq2. We found that Trad-KAPA detected more differentially expressed transcripts at all four sequencing depths tested. Interestingly, Xiong *et al* also detected more DEGs using the traditional method compared the 3’ method [40], while Tandonnet *et al* detected more DEGs using the 3’ method. We think the differences were caused by removing duplicated reads. Xiong *et al* did not remove duplicates in their traditional method but rather used unique molecular identifier to remove the PCR duplicates in their 3’ method. Tandonnet *et al* removed all the duplicates in both methods. In our study, we did not remove duplicates, as we believe that instead of PCR over-amplification, the major cause of duplicated reads is very high expression of a small number of genes [41].

Among all the DEGs we found, some of the very short transcripts (shorter than 500 bp) were only detected to be differentially expressed by 3’-LEXO, while many of the long transcripts, especially those longer than 7500 bp, were only detected as differentially expressed by Trad-KAPA. As Trad-KAPA assigns more reads to longer transcripts, the statistical power to detect differences increases. Thus, the probability that those transcripts are detected differentially expressed is higher. It is also clear that as sequencing depth drops, both methods will detect fewer differentially expressed transcripts. Thus, if users want to use RNA-seq to detect differentially expressed transcripts, Trad-KAPA will likely generate larger lists than 3’-LEXO, biased towards longer transcripts.

Author contributions

MP, AJL and CDV designed the work. BKF performed the animal husbandry. BKF and CY prepared the sequencing library. FM and YH processed the transcriptome data and did bioinformatics analyses. FM and MP drafted the article. All authors read and approved the final manuscript.

Figures

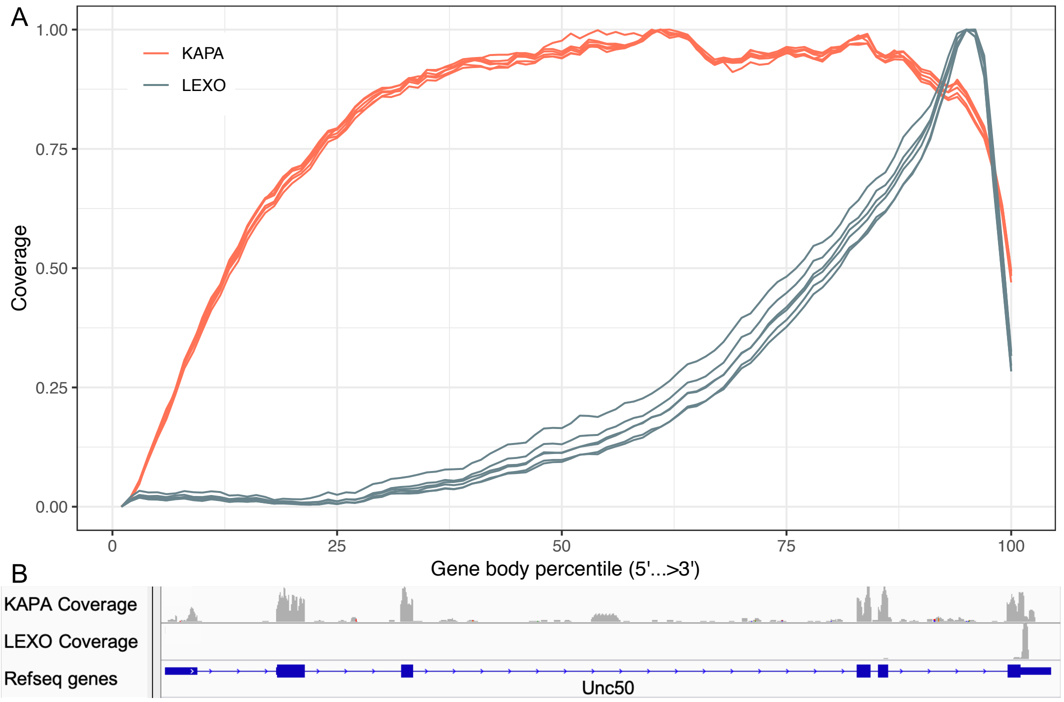


Figure 2-1. Gene body coverage.

(A) Gene body coverage from the Trad-KAPA and 3'-LEXO libraries.

(B) Unc50 gene body coverage from the Trad-KAPA and 3'-LEXO libraries.

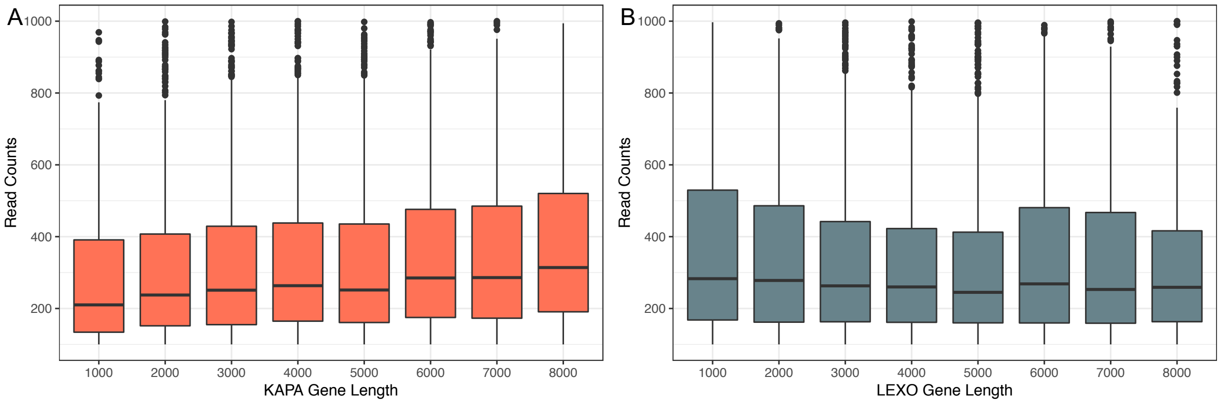


Figure 2-2. Read counts for transcripts of different length.

(A) Trad-KAPA read counts for transcripts with different length.

(B) 3'-LEXO read counts for transcripts with different length.

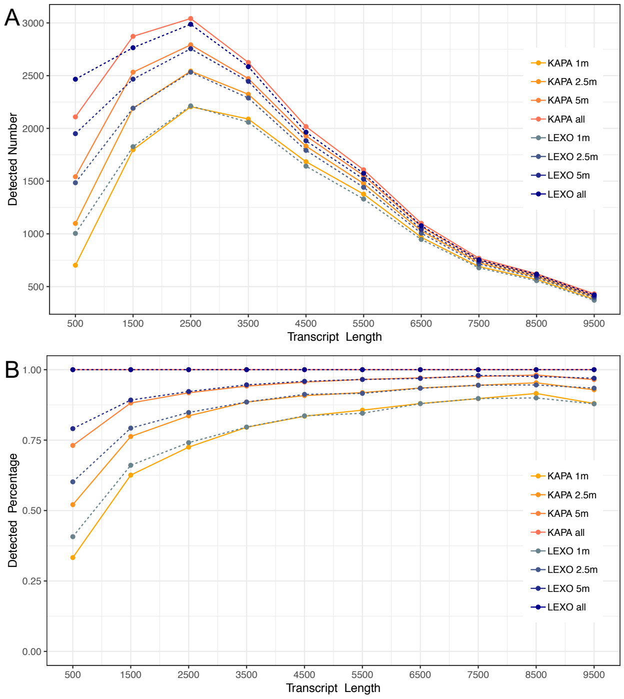


Figure 2-3. Transcripts of different length detected after subsampling.

(A) The number of transcripts of different length detected after subsampling.

(B) Percent of transcripts of different length detected after subsampling, compared to sampling at 10 million reads.

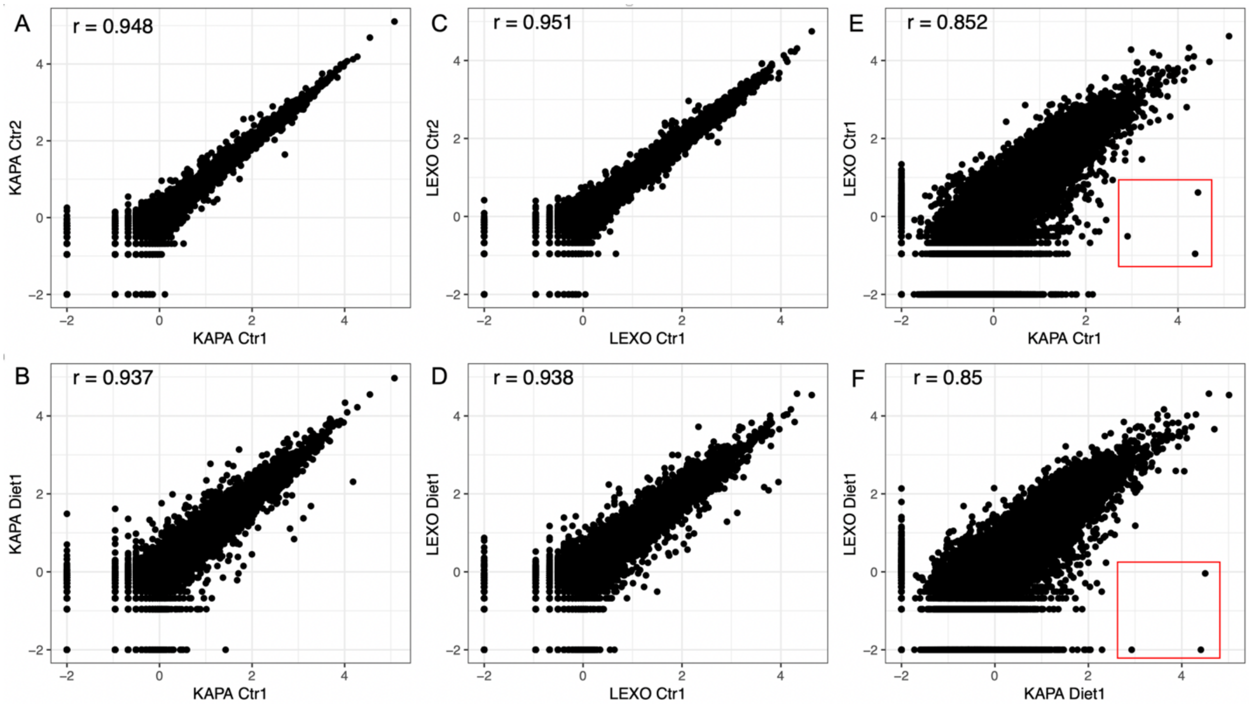


Figure 2-4. Correlation between Trad-KAPA and 3'-LEXO samples.

(A) Correlation between the Trad-KAPA control samples 1 and 2.

(B) Correlation between the Trad-KAPA control sample 1 and the iron loaded diet sample 1.

(C) Correlation between the 3'-LEXO control samples 1 and 2.

(D) Correlation between the 3'-LEXO control sample 1 and the iron loaded diet sample 1.

(E) Correlation between the Trad-KAPA and 3'-LEXO control sample 1.

(F) Correlation between the Trad-KAPA and 3'-LEXO iron loaded diet sample 1.

Sequencing Depth	Trad-KAPA	Intersection (with 10m)	3'-LEXO	Intersection (with 10m)	Intersection (Trad-KAPA and 3'-LEXO)
1 million	343	339 (98.8%)	257	249 (96.9%)	177
2.5 million	758	742 (97.9%)	474	460 (97.0%)	329
5 million	1234	1194 (96.8%)	777	740 (95.2%)	562
10 million	1982	1982	1157	1157	882

Table 2-1. The number of differentially expressed transcripts detected by the Trad-KAPA and 3'-LEXO, before and after subsampling from 10 million reads.

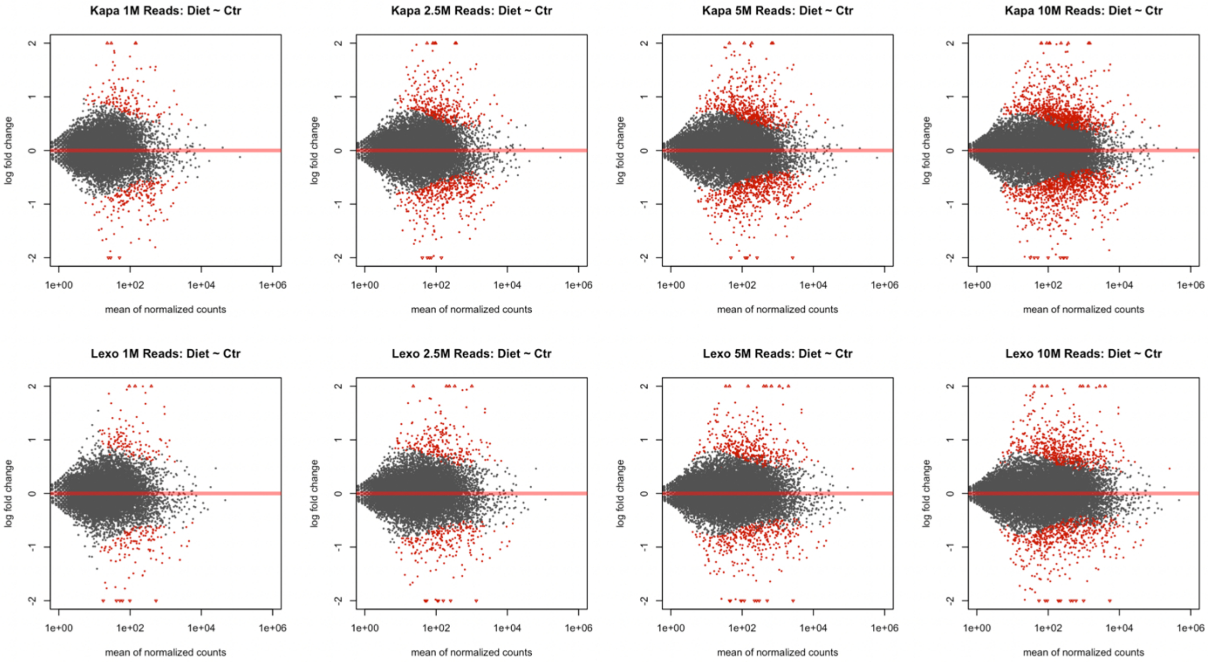
The first column denotes the sequencing depth (i.e. the total number of mapped reads from the library examined). The second column denotes the number of differentially expressed transcripts detected by Trad-KAPA. The third column denotes the number of differentially expressed transcripts detected after subsampling that overlap with those from the 10 million sequencing depth. The fourth and fifth columns denote the results for the 3'-LEXO method. The sixth column denotes the number of differentially expressed transcripts detected by both the Trad-KAPA and the 3'-LEXO methods at listed sequencing depth.

Gene name	Primer set used	RT-qPCR fold change	Trad-KAPA fold change	3'-LEXO fold change	RT-qPCR result match which RNA-Seq method	Group
Adnp	mAdnp-ex2-3	0.83	1.15	5.02	Trad-KAPA	Trad-KAPA only
Cd7a	mCd7a-ex3-4	0.69	0.79	5.11	Trad-KAPA	Trad-KAPA only
Fv1	mFv1-F169	0.55	0.54	10.48	Trad-KAPA	Trad-KAPA only
Mid1	mMid1ex4-5	0.77	0.53	5.12	Trad-KAPA	Trad-KAPA only
Mid1	mMid1ex8-9	0.83	0.53	5.12	Trad-KAPA	Trad-KAPA only
Mmp28	mMmp28ex2-3	3.24	4.52	8.55	Trad-KAPA	Trad-KAPA only
Unkl	mUnkl-ex5-6	0.75	1.11	5.12	Trad-KAPA	Trad-KAPA only
Unkl	mUnkl-ex2-3	0.90	1.11	5.12	Trad-KAPA	Trad-KAPA only
Zfp647	mZfp647-204ex4-5	0.55	0.42	8.46	Trad-KAPA	Trad-KAPA only
Zfp647	mZfp647-201ex3-4	0.59	0.42	8.46	Trad-KAPA	Trad-KAPA only
Bcl2a1b	mBcl2a1bEx1-2	2.91	1.44	5.76	In between	3'-LEXO only
Hist4h4	mHist4h4	1.83	0.26	0.27	Neither	3'-LEXO only
Mir5136	mMir5136	1.47	5.07	0.88	3'-LEXO	3'-LEXO only
Mt-Tq	mMt-Tq	1.09	0.95	0.30	Trad-KAPA	3'-LEXO only
Rps27rt	mRps27rt	1.27	0.27	1.40	3'-LEXO	3'-LEXO only
S100a4	mS100a4ex1-2	1.93	1.42	2.31	3'-LEXO	3'-LEXO only
S100a4	mS100a4ex2-3	2.06	1.42	2.31	3'-LEXO	3'-LEXO only
Schip1	mSchip1ex7-8	0.85	0.51	0.90	3'-LEXO	3'-LEXO only
Snord118	mSnord118	0.46	0.26	0.60	3'-LEXO	3'-LEXO only
Snord13	mSnord13	0.92	0.98	0.48	Trad-KAPA	3'-LEXO only
Spink1	mSpink1ex3-4	9.49	2.66	8.28	3'-LEXO	3'-LEXO only
Tceal5	mTceal5ex3-4	5.82	10.48	30.09	Trad-KAPA	3'-LEXO only
Tceal5	mTceal5ex1-2	6.07	10.48	30.09	Trad-KAPA	3'-LEXO only
Atoh8	mAtoh8	3.97	3.10	3.19	Both	Iron metabolism
Bdh2	mBdh2	0.28	0.35	0.39	Both	Iron metabolism
Bmp6	mBmp6	4.83	6.01	6.20	Both	Iron metabolism
Cp	mCp	1.77	1.88	1.94	Both	Iron metabolism
Ftl1	mFtl1	1.75	0.98	3.26	In between	Iron metabolism
Hamp1	mHamp1	5.75	5.19	5.73	Both	Iron metabolism
Hamp2	mHamp2	0.26	0.28	0.33	Both	Iron metabolism
Hfe2	mHfe2	0.61	0.66	0.67	Both	Iron metabolism
Id1	mId1F205&200	4.05	3.43	3.19	Both	Iron metabolism
Lcn2	mLcn2	2.91	2.92	2.25	Both	Iron metabolism
Slc11a2	mSlc11a2	0.66	0.80	0.73	Both	Iron metabolism
Smad7	mSmad7	3.33	3.85	2.91	In between	Iron metabolism
Tfrc	mTfrc	1.16	1.26	1.32	Both	Iron metabolism

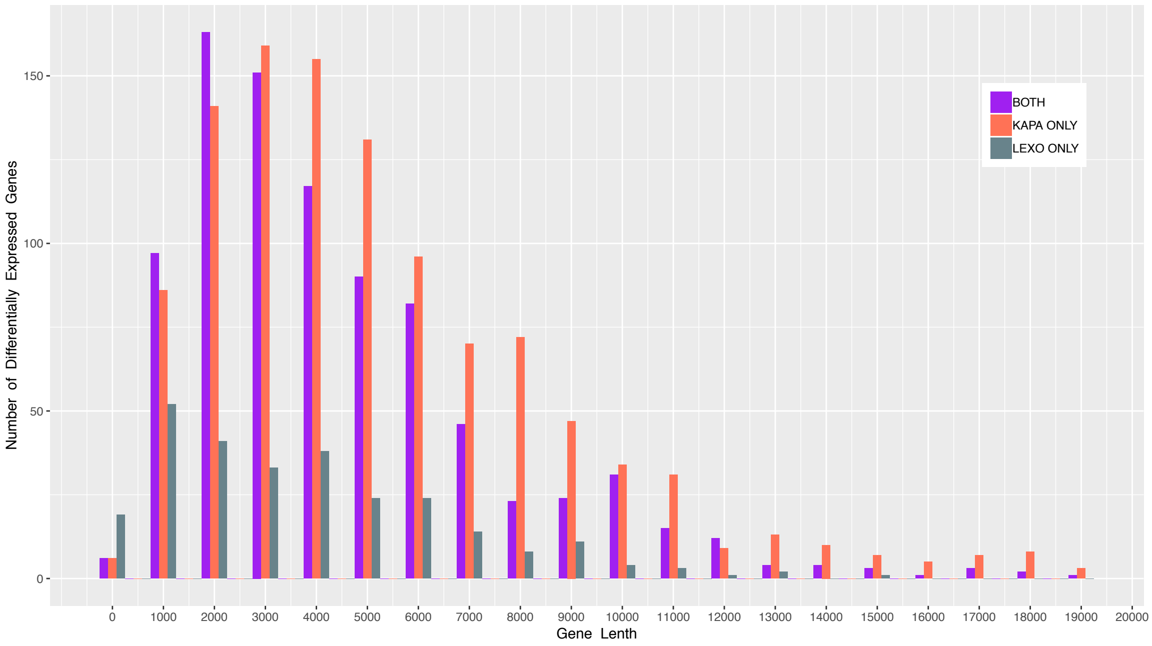
Table 2-2. RT-qPCR results.

Column 3-5 give the log2 fold difference in expression between the iron loaded and control samples by RT-qPCR, Trad-KAPA, and 3'-LEXO. Column 6 indicates if the RT-qPCR results matched better to one RNA-seq method. Column 7 denotes the group of the genes: detected only in Trad-KAPA (Trad-KAPA only), detected only in 3'-LEXO (3'-LEXO only) or iron metabolism related (Iron metabolism).

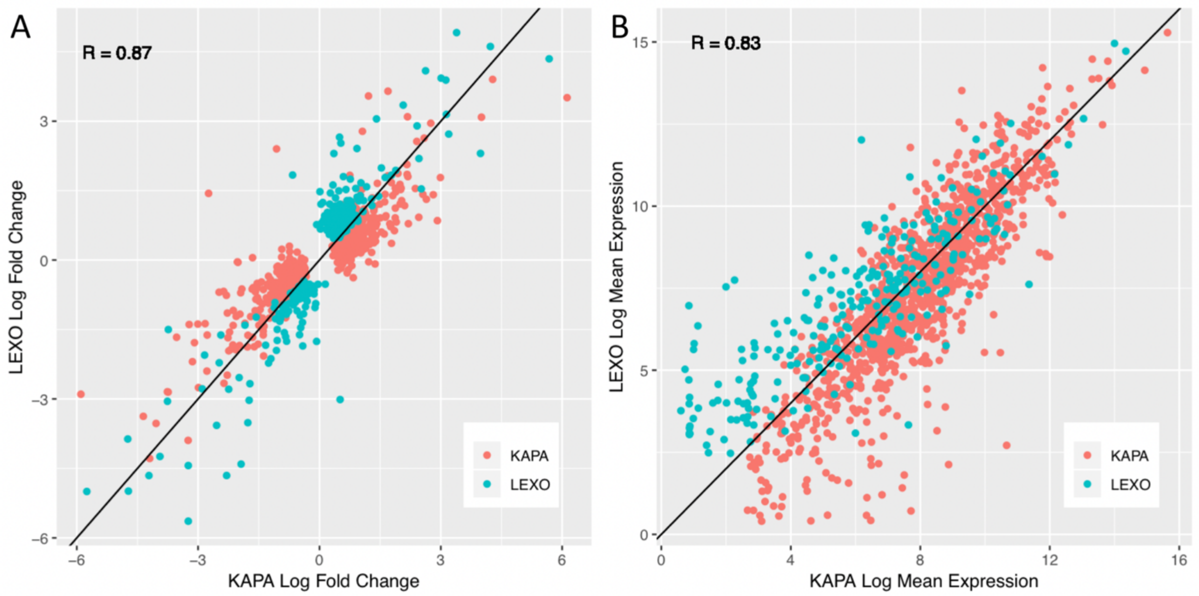
Supplemental information



Supplementary Figure 2-1. MA plots showing the differentially expressed transcripts detected by Trad-KAPA and 3’-LEXO with subsampling.

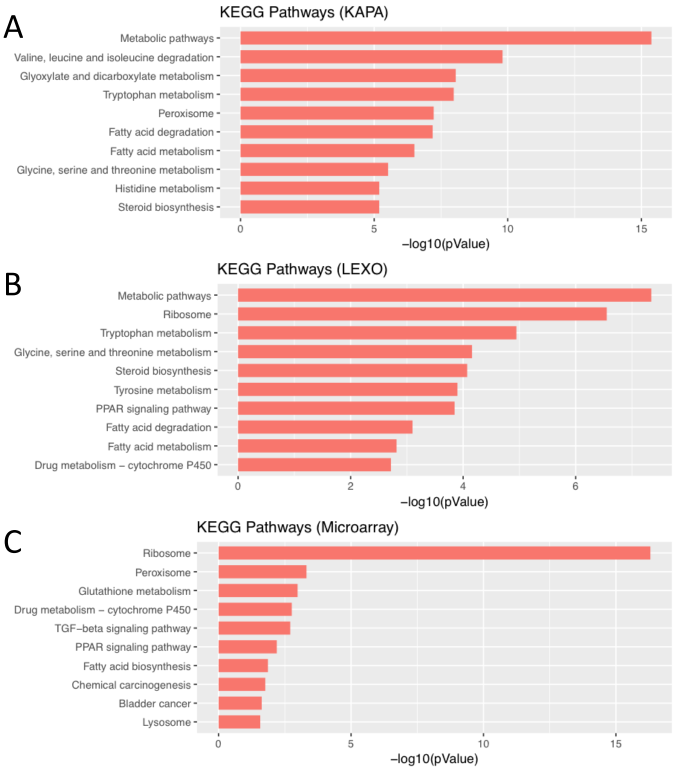


Supplementary Figure 2-2. The number of differentially expressed transcripts, grouped by transcript length, detected only by Trad-KAPA (red), only by 3'-LEXO (blue) and by both methods (purple).



Supplementary Figure 2-3. Comparing DEGs detected in only one method.

Genes here are DEGs detected in only KAPA (red) or in only LEXO (blue), log2 fold changes (A) and log2 mean expression (B) are compared between the two methods.



Supplementary Figure 2-4. KEGG Pathways enriched by Trad-KAPA (A), 3’-LEXO (B) and Microarray (C) DEGs.

Chapter 3 - ACTINN: Automated Identification of Cell Types in Single Cell RNA Sequencing

Abstract

Cell type identification is one of the major goals in single cell RNA sequencing (scRNA-seq). Current methods for assigning cell types typically involve the use of unsupervised clustering, the identification of signature genes in each cluster, followed by a manual lookup of these genes in the literature and databases to assign cell types. However, there are several limitations associated with these approaches, such as unwanted sources of variation that influence clustering and a lack of canonical markers for certain cell types. Here, we present ACTINN (Automated Cell Type Identification using Neural Networks), which employs a neural network with 3 hidden layers, trains on datasets with predefined cell types, and predicts cell types for other datasets based on the trained parameters. We trained the neural network on a mouse cell type atlas (Tabula Muris Atlas) and a hu-man immune cell dataset, and used it to predict cell types for mouse leukocytes, human PBMCs and human T cell sub types. The results showed that our neural network is fast and accurate, and should therefore be a useful tool to complement existing scRNA-seq pipelines.

Introduction

Single cell RNA sequencing (scRNA-seq) enables the profiling of the transcriptomes of individual cells, thus characterizing the heterogeneity of samples in manner that was not possible using traditional bulk RNA-seq [5]. However, scRNA-seq experiments typically yield high volumes of data, especially when the number of cells is large (often many thousands). Thus, fast and efficient computational methods are essential for scRNA-seq analyses.

One common goal of scRNA-seq analyses is to identify the cell type of each individual cell that has been profiled. To accomplish this, typically cells are first grouped into different clusters in an unsupervised way, and the number of clusters allows us to approximately determine how many distinct cell types are present in the sample. Each cluster should contain cells with similar expression profiles, and so the aggregated profile of a cluster increases the signal to noise of the expression estimates. To attempt to interpret the identity of each cluster, marker genes are found as those that are uniquely highly expressed in a cluster, compared to all the other clusters. These canonical markers are then used to assign the cell types for the clusters, by cross referencing the markers with lists of previously characterized cell type specific markers. While this process is able to identify cell types, there are some limitations: 1. Since the clustering method is unsupervised, all sources of variation influence the formation clusters, including effects that are not directly related to cell types such as differential expression induced by cell cycles. 2. It is often difficult to find an optimal match between the marker genes associated with each cluster and the canonical markers for specific cell types. Moreover, depending on the clustering parameters used, one cluster might contain multiple cell types, or one cell type could be split into multiple clusters. 3. Using canonical markers to assign cell types requires background knowledge of cell type specific markers, and sometimes these are not well characterized or difficult to find in the literature. Moreover, some canonical markers may be expressed by more than one cell type, and some cell types may have no known markers. 4. The same types of cells processed by two distinct scRNA-seq techniques tend to cluster separately due to technical batch effects, which complicates cell type identification in composite datasets. 5.

Cell subtypes are often very similar to each other, which limits efforts to separate them accurately into different clusters. To overcome many of the limitations of existing approaches, new methods need to be developed.

Neural networks provide a popular framework for machine learning algorithms which can be used to interpret complex datasets. As a result, neural networks have been widely used in many fields, including for the analysis of scRNA-seq data [6-9]. Since the output data from scRNA-seq is feature-enriched and well-structured, it is well suited as an input for neural networks. Here, we present ACTINN (Automated Cell Type Identification using Neural Networks) for scRNA-seq cell type identification. To overcome many of the limitations of traditional cell type identification approaches described above, we used a neural network with 3 hidden layers, trained it on scRNA-seq datasets with predefined cell types, and predicted cell types in other datasets based on the trained parameters. We tested our neural network with several published datasets and show that it is fast, efficient and accurate.

Materials and Methods

Data normalization

We used several publicly available datasets in our analyses. The mouse cell atlas datasets were collected from https://tabula-muris.ds.czbiohub.org. The CD45 sorted leukocyte datasets were published by Winkels *et al* [42]. The T cell subtypes and PBMC datasets were collected from https://support.10xgenomics.com/single-cell-gene-expression/datasets. To filter and normalize the data, we first identified genes that were detected in both training set and test set. The training set and the test set were then merged into one matrix based on the common genes. Next, each cell's expression value was normalized to its total expression value and multiplied by a scale factor of 10,000. The counts were increased by 1, and the log2 value was calculated. To filter out outlier genes, the genes with the highest 1% and lowest 1% expression were removed. The gene with the highest 1% and the lowest 1% standard deviation were also removed. Finally, the matrix was split into the training set and the test set.

Neural network configuration

We used a neural network that contains an input layer, 3 hidden layers, and an output layer. The input layer had a number of nodes equal to the number of genes in the training set. The 3 hidden layers had 100, 50 and 25 nodes, respectively. The output layer had a number of nodes equal to the number of cell types in the training set. Forward propagation was implemented as:

$$x^{[i]} = g\left(W^{[i]}x^{[i-1]} + b^{[i-1]}\right)$$

Where $x^{[i]}$ represents the output of the ith layer ($x^{[0]}$ represents the input layer), $b^{[i]}$ represents the intercept of the ith layer, $W^{[i]}$ represents the weight matrix of the ith layer, and g represents the activation function used in the neural network. Specifically, for the activation function, the rectified linear unit (ReLU) function was used for the input and hidden layers, which is defined as:

$$ReLU(x) = max(0, x)$$

For the output layer, the softmax function was used, which is defined as:

$$softmax(x_{[j]}) = \frac{\exp\,(x_{[j]})}{\sum_{j=1}^{k} \exp\,\left(x_{[j]}\right)}$$

Where $x_{[j]}$ represents the jth element of the input vector for the output layer, which has k elements, representing a total of k cell types in the training set. For the loss function, we used the cross-entropy function, which is defined as:

$$H(y',y) = \sum_{j=1}^{k} \left(y_{[j]} log\left(y'_{[j]}\right) + \left(1 - y_{[j]}\right) log\left(1 - y'_{[j]}\right) \right)$$

Where vector y represents the true label for the cell, $y_{[j]}$ is defined to be 1 if the cell is the jth cell type, and the other elements in y are defined to be 0. y' represents the output of the output layer, and $y'_{[j]}$ represents the posterior probability that the cell is the jth cell type. L2 regularization was added to the loss function.

Parameters used in the neural network

The neural network model was implemented using TensorFlow (https://www.tensorflow.org), and the code was written in python. The parameters were initialized with Xavier initializer [43]. The starting learning rate was set to 0.0001 with staircase exponential decay, the decay rate was set to 0.95, and the decay step was set to 1000. This means that after every 1000 global steps, the learning rate would be the original learning rate multiplied by 0.95. 50 epochs were used to train the neural network with a mini batch size of 128, which is the number of samples used in training at every global step. The L2 regularization rate was set to 0.005.

Unsupervised single cell analysis

To identify different cell types and find signature genes for each cell type, Seurat [44] was used to analyze the digital expression matrix generated by scRNA-seq. Specifically, in Seurat, cells with less than 1000 unique molecular identifiers (UMIs) and genes detected in less than 10 cells were first filtered out. Second, highly variable genes were detected and used for further analysis. Third, the data was scaled for sequencing depth of each cell. Fourth, principle component analysis (PCA) and t-distributed stochastic neighbor embedding (tSNE) were used to reduce the dimension and plot the data on a two-dimensional graph. Lastly, a graph-based clustering approach was used to cluster the cells, then signature genes were found and used to define cell type for each cluster.

Results

Overview of the neural network

We used a neural network with 3 hidden layers, each containing 100, 50 and 25 nodes, respectively (Figure 1-3). For the activation functions, we used the softmax function for the ouput layer and the rectified linear unit (ReLU) function for the other layers. We used the cross-entropy function as the loss function. The neural network model was implemented using TensorFlow, and the code was written in python. We trained the neural network on 6 Intel(R) Xeon(R) CPU E5-2660 v3 nodes, and the training process took 0.5 minute to complete with 1000 cells, 11 minutes with 32,000 cells and 21 minutes with 56,000 cells. The maximum memory used in training with 56,000 cells was 18 GB. The code and datasets used in this study are available at https://github.com/mafeiyang/ACTINN.

ACTINN model for murine cell types

We used 2 datasets from the Tabula Muris Consortium (The Tabula Muris Consortium. 2018) to train and test our neural network. The datasets contain 100,605 cells from 20 mouse organs, and were sequenced by two distinct techniques, 10X Genomics (10X) and Smart-seq2 (SS2). To ensure we are using cells with high quality, we filtered out cells with less than 300 detected genes, clustered the cells, and identified marker genes for each cluster using Seurat. The details of the Seurat analysis can be found in the methods section. We manually assigned cell types for each cluster based on canonical markers (Figure 3-1A). To make the analysis easier to interpret, we merged similar cell types into one single cell type. For example, we merged B cells, naïve B cells, immature B cells, pro-B cells and late pro-B cells from the TMA datasets into B cells. We focused on 12 cell types and selected cells that have the same labels between our analyses and the Tabula Muris Consortium's. This process resulted in 56,112 cells (Figure 3-1B). Cells processed by 10X have a median of 4,787 unique molecular identifiers (UMIs) and 1,558 genes detected, and cells processed by SS2 have a median of 623,799 UMIs and 2,448 genes detected.

To test the robustness of our neural network's performance, we first trained and tested it on cells processed by each scRNA-seq platform separately. To this end, we randomly sampled 3000 cells for testing, and used the remainder of cells for training. We repeated this process 10 times, and the average training accuracies for the 10X dataset and the SS2 dataset were 99.997% and 99.963%, respectively, and the average testing accuracies were 99.883% and 99.660%, respectively (Figure 3-1D). These results show that our neural network can achieve very high accuracy when training and testing on datasets generated by the same technique.

ACTINN overcomes batch effects introduced by different techniques

Different scRNA-seq techniques can introduce significant batch effects [45] with the same cell types clustering separately due to technical artifacts (Figure 3-1C). To test our neural network's performance accounting for the batch effects introduced by different techniques, we trained it on cells processed by one platform and tested it on cells processed by the other. We first trained the neural network on all the 10X cells and tested in on all the SS2 cells. The training accuracy was 99.997% and the testing accuracy was 98.625%. Among the 288 incorrectly predicted cells, 118 monocytes were predicted as B cells, 64 monocytes were predicted as epithelial cells, 47 NK cells were predicted T cells (Supplementary Table 3-1). We then trained the neural network on the SS2 dataset and tested it on the 10X dataset. The training accuracy was 100% and the testing accuracy was 99.195%. Among the 283 incorrectly predicted cells, 150 endothelial cells were predicted as epidermis, 46 T cells were predicted as NK cells, and there were several other mispredictions (Supplementary Table 3-2).

Early stopping prevents overfitting of the training set

To prevent overfitting the parameters on the training set, we randomly sampled 5,000 cells from the 10X dataset and 5,000 cells from the SS2 dataset. We trained the neural network on the 10X cells and tested it on the SS2 cells. During the training process, we recorded the accuracy and the cost after each epoch. The accuracy was defined as the percentage of cells whose cell type was correctly predicted, and the cost was the output of the cost function after each epoch. We found that the training accuracy saturated early (5 epochs), and the testing accuracy saturated at around 50 epochs (Figure 3-1E), and the cost decreased very slowly after 50 epochs (Figure 3-1F). These results indicate that early stopping can be used to reduce training time and prevent overfitting.

Cell type prediction using the mouse cell atlas

Since the cell types from the two mouse cell atlas datasets can be accurately predicted, we combined the two datasets and used the combined dataset as the reference to predict cell types for other datasets. We first tried to predict cell types for a dataset that contains flow cytometry sorted leukocytes from mouse aorta [42]. All cells were predicted as leukocytes except for 1 erythrocyte, which we think is a doublet of an erythrocyte and B cell as high expression of hemoglobin genes was detected (Figure 3-2A). We also carried out unsupervised analysis on the dataset and clustered the cells using Seurat. Then we used the canonical markers to assign the cell types for each cluster (Figure 3-2B). Most cells had the same cell type assignment by the two methods. However, our neural network detected some natural killer (NK) cells, which were in the same cluster with the T cells, and were assigned as T cells in the unsupervised clustering. We checked the expression of CD3D, CD8A and GZMA (Figure 3-2C), and found no expression of CD3D and CD8A, but high expression of GZMA in the NK cells, which suggests that these are likely NK cells. To test if ACTINN produces consistent results from run to run, we trained the neural network on the combined TMA dataset, tested it on the mouse leukocytes dataset, and repeated this process 10 times. We found that most of the cells were assigned the same label across all 10 runs (Supplementary Figure 3-1A), and the frequency for each cell type was also consistent between different runs (Supplementary Figure 3-1B).

It is generally thought that human and mouse share similar cell types, and the same cell type from human and mouse share similar expression profiles. To test this, we trained our neural network on the mouse cell atlas datasets and used the parameters to predict the cell types for a human peripheral blood mononuclear cell (PBMC) dataset. We found 4 main populations in the PBMC dataset, namely, B cells, monocytes, NK cells and T cells (Figure 3-2D). We plotted the canonical markers for these 4 populations (Figure 3-2E) and found that the predicted cell types matched the expected marker expression. These results suggest that the mouse cell atlas datasets can be used as a reference to identify cell types for both human and mouse cells.

ACTINN accurately identifies cell types not in the reference

An scRNA-seq experiment may be performed on tissues where not all the cell types in the data of interest are included in the reference dataset. If a cell cannot be classified as a known cell type in the training data, we would label it “uncertain”. To test if ACTINN can identify cell types that are not in the reference, we trained the neural network on the TMA datasets and tested it on the mouse leukocytes plus 109 mouse nerve cells (the nerve cells are not in the training data). We output the probabilities for each cell being one of the cell types in the training data, and labelled the cell “uncertain” if its highest probability is smaller than 0.95. We found that most of the B cells, T cells, NK cells, monocytes and granulocytes were assigned correctly (Supplementary Figure 3-2A). By contrast, 105 out of 107 nerve cells were assigned “uncertain” (Supplementary Figure 3-2B). These results show that ACTINN is able to identify cell types that are not in the training dataset.

ACTINN accurately predicts cell subtypes

Although it is relatively easy to distinguish different cell types in scRNA-seq using the unsupervised clustering methods, it is more difficult to further divide one cell type into cell subtypes. Here, we collected 5 publicly available datasets [46], each containing one flow cytometry sorted T cell subtype. We merged these datasets and selected the cells that have the same labels between our analyses and the flow cytometry sorting, and then used these cells as a reference for the neural network. We then clustered the selected cells and identified markers (Figure 3-3A and 3-3B) for each sub cell type using Seurat. For the test set, we used the T cells from the human PBMC datasets mentioned above.

To test our neural network's ability to predict cell subtypes, we trained it on the T cell subtype reference, and predicted the subtypes for the T cells from the PBMC dataset (Figure 3-3D). We then identified marker genes for each predicted subtype. As expected, the marker genes matched the ones from the reference (Figure 3-3E). These results show that our neural network can be used to accurately identify cell subtypes. We found that the subtypes predicted by the neural network did not perfectly match the cell types associated with the Seurat clusters (Figure 3-3C). Some clusters contained different subtypes and some subtypes were composed of several clusters. We think the difference was influenced by two factors: 1. Unsupervised clustering considers all variance in the data, while the neural network is trained to find the difference between the subtypes; 2. It is difficult to set the parameters optimally for the unsupervised analysis, which can result in multiple cell types in one cluster or multiple clusters for one cell type.

Comparison to other cell type identification tools

As the field of scRNA-seq is evolving rapidly, new ideas and methods are being published frequently. Several supervised scRNA-seq cell type identification methods were proposed recently. SuperCT [47] uses a neural network, CaSTLe [48] uses XGboost, and SingleCellNet [49] uses a random forest to annotate cell types in scRNA-seq experiment. We found that these 3 methods convert the expression values to binary signals (SuperCT and XGboost) or 4 categories (CaSTLe) before training the data. This conversion may significantly decrease the complexity of the expression data, which makes it difficult to distinguish between small changes in expression. We compared the performance of the 3 methods to ACTINN in sub cell type identification. We trained CaSTLe and SingleCellNet using the T cell subtype reference, and trained SuperCT on its human cell reference as it does not allow user defined reference. Then we predicted the subtypes for the T cells from the PBMC dataset. CaSTLe and SingleCellNet failed to define most of the naïve T cells and regulatory T cells, and SuperCT failed to distinguish T cell subtypes (Supplementary Figure 3-3ABCD). Based on the predictions and marker gene expression, we manually set the labels for the T cell subtypes (Supplementary Figure 3-3E). Then we calculated the prediction accuracy for ACTINN (73%), CaSTLe (59%) and SingleCellNet (57%) (Supplementary Figure 3-3F). These results show that ACTINN outperforms the 3 tools in finding small changes between subtypes.

Discussion

scRNA-seq provides high resolution profiling of the transcriptomes of single cells. Typically, the first step in scRNA-seq analysis is to assign each cell a cell type based on our prior knowledge of marker genes. Current methods for cell type assignment first cluster the cells in an unsupervised manner and rely on the canonical markers to identify the cell types for each cluster. However, this approach has several limitations, including the fact that the clusters may not optimally segregate single cell types, and certain cell types may not have previously characterized markers. Moreover, these methods are computationally intensive, especially when the number of cells becomes large. To render cell type identification in scRNA-seq more efficient, we employed a neural network, trained it on cells with predefined cell types, and used it to predict cell types for new datasets.

We first obtained and cleaned two datasets from the Tabula Muris Consortium, then trained and tested our neural network on these datasets with or without batch effect introduced by different scRNA-seq platforms. The training accuracy always approached 100%, and the testing accuracy was around 99.8% within a platform and 99.0% when testing and training are performed across different platforms. As the cell types in the two Tabula muris atlas datasets can be mutually predicted using our neural network, we merged them and used the combined datasets as the reference to predict cell types for other datasets. The predicted cell types were well matched with the cell types assigned using the canonical markers for both the mouse and human datasets. We also trained and tested the neural network on 5 T cell subtypes and found that the predicted subtypes showed the same markers as the reference subtypes, which suggests that our neural network can be used to predict sub cell types as well.

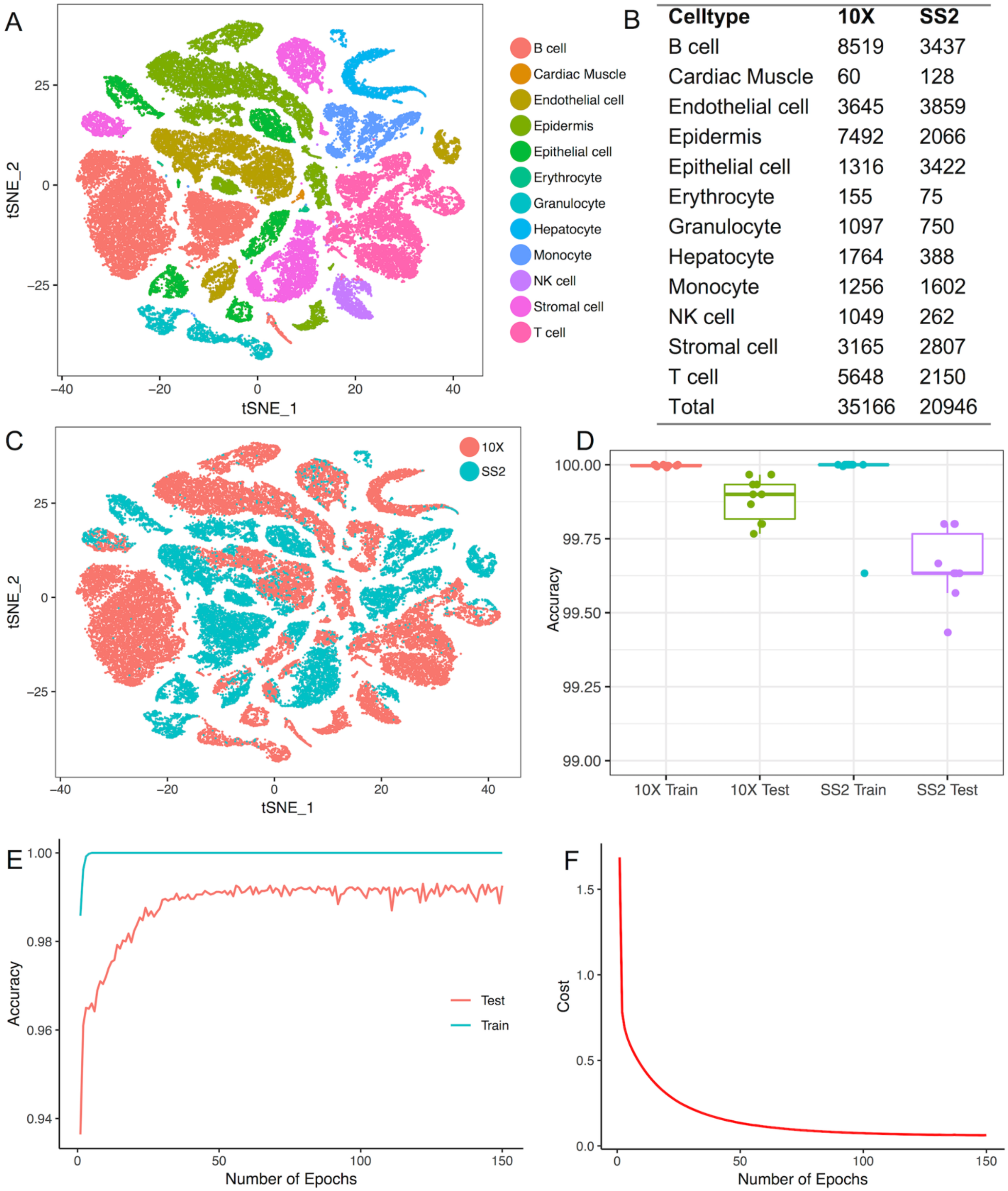
Compared to the traditional unsupervised methods used for cell type identification, our neural network has the following advantages: 1. It uses all the genes to capture the features for each cell type instead of relying on a limited number of canonical markers. 2. It focuses the analysis on the signal associated with the variance between cell types, while unsupervised clustering tends to be affected by other sources of cell type independent variation (i.e. platform or cell cycle). 3. It requires no background knowledge of cell type markers, while the unsupervised method requires users to have prior knowledge of canonical markers for each cell type in their data. 4. It is much more computationally efficient than the traditional approach. Moreover, users can subsample the reference cells to make the computation of the neural network less compute intensive and more memory efficient. We also compared ACTINN to 3 other cell type prediction tools, and the results showed that ACTINN performs better in finding small changes between subtypes.

There are some aspects of our approach that could be improved in the future. As the neural network is supervised, the quantity and quality of the reference data are critical. We anticipate that with time more cell types from larger atlases should be used to train a more comprehensive neural network. Also, better pairing of reference and test sets will undoubtedly improve performance. For example, the soon to be developed human cell atlas should be used to predict human cell types instead of the mouse cell atlas. Nonetheless, we showed that even with the current reference data our neural network is computationally efficient and accurate, and should improve cell type identification pipelines.

Author contributions

Conceptualization, Methodology, Analysis, Writing, Visualization, Validation – FM, MP; Supervision, Funding Acquisition – MP.

Figures



Celltype	10X	SS2
B cell	8519	3437
Cardiac Muscle	60	128
Endothelial cell	3645	3859
Epidermis	7492	2066
Epithelial cell	1316	3422
Erythrocyte	155	75
Granulocyte	1097	750
Hepatocyte	1764	388
Monocyte	1256	1602
NK cell	1049	262
Stromal cell	3165	2807
T cell	5648	2150
Total	35166	20946

Figure 3-1. Training and testing of the neural network on the Tabula Muris Atlas.

(A) Cell types obtained from the TMA.

(B) Number of cells obtained for each cell type from each technique.

(C) The same cell type tends to cluster separately by techniques.

(D) Training and testing accuracy of the neural network when trained and tested using cells processed by the same technique.

(E) Training and testing accuracy after each epoch when trained with 5,000 10X cells and tested with 5,000 SS2 cells.

(F) Cost after each epoch when trained with 5,000 10X cells and tested with 5,000 SS2 cells.

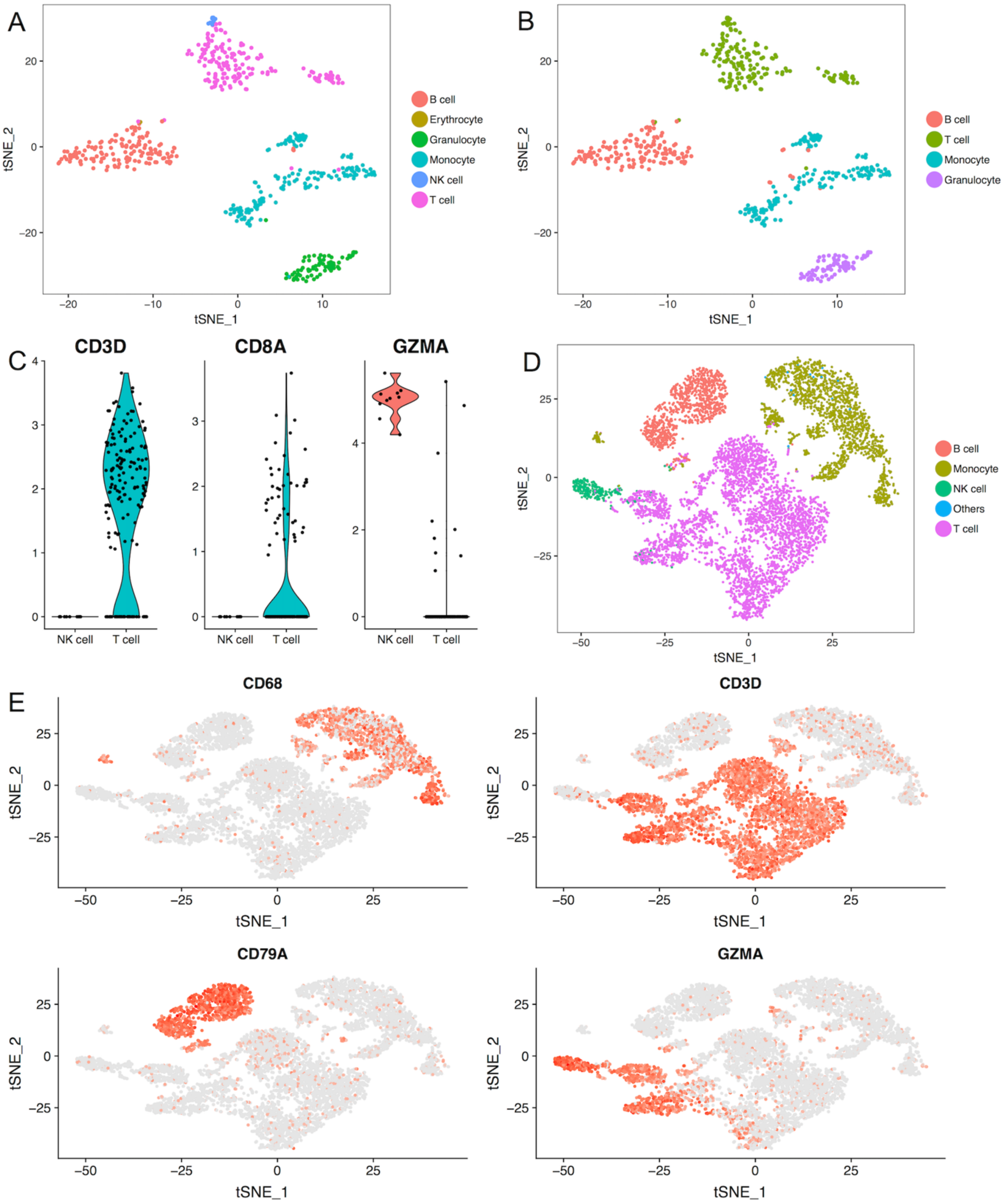

A
tSNE_2
tSNE_1
B cell
Erythrocyte
Granulocyte
Monocyte
NK cell
T cell
B
tSNE_2
tSNE_1
B cell
T cell
Monocyte
Granulocyte
C
CD3D
CD8A
GZMA
NK cell
T cell
D
tSNE_2
tSNE_1
B cell
Monocyte
NK cell
Others
T cell
E
CD68
CD3D
CD79A
GZMA
tSNE_2
tSNE_1

Figure 3-2. Neural network predicts cell types for human and mouse datasets.

(A) Cell types predicted by the neural network for the mouse leukocyte dataset.

(B) Cell types identified by unsupervised clustering and canonical markers for the mouse leukocyte dataset.

(C) Violin plots showing 3 genes' expression level in the NK and T cells from the mouse leukocytes.

(D) Cell types predicted by the neural network for the human PBMC dataset.

(E) TSNE plots showing 4 marker genes' expression for the human PBMC dataset.

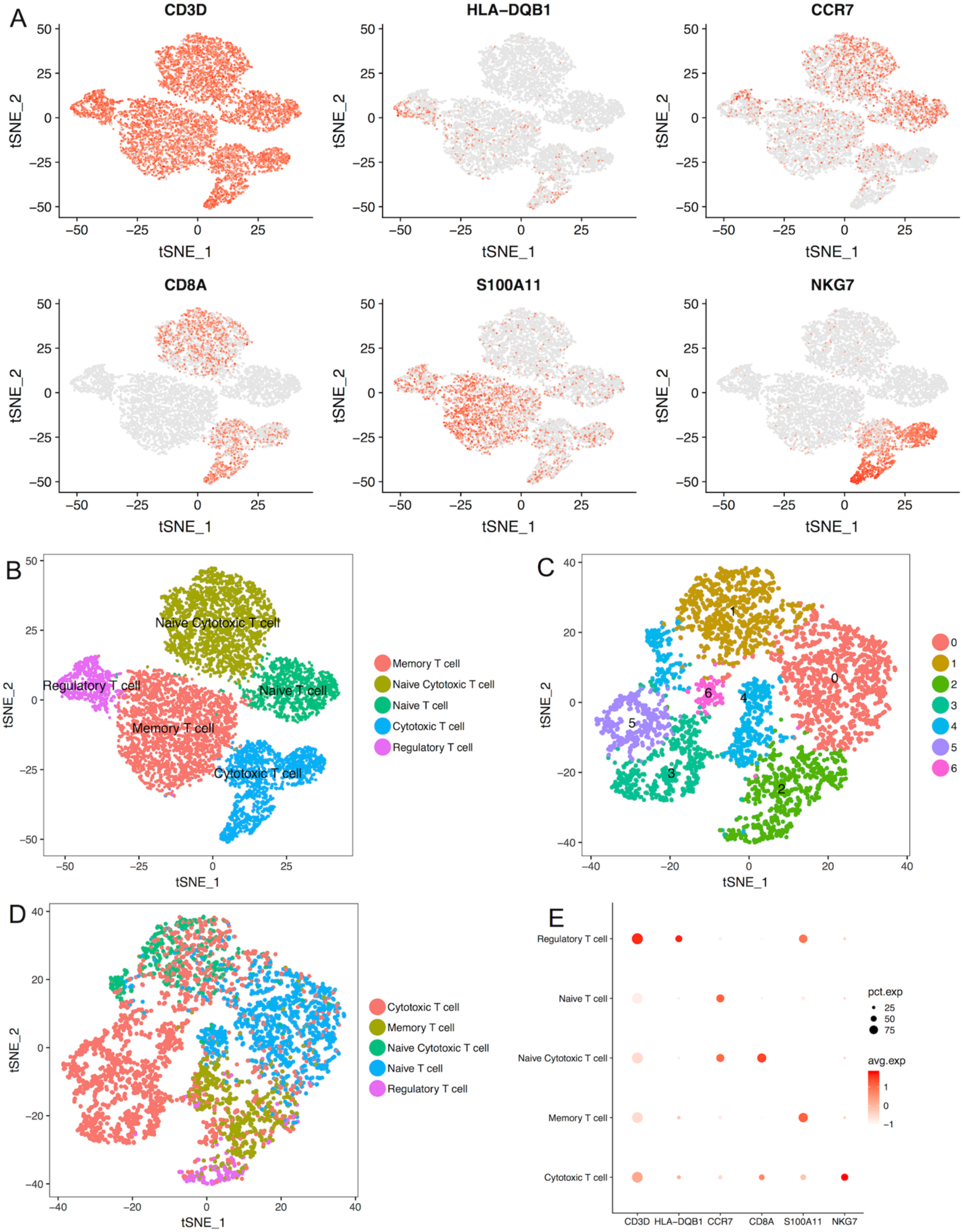

A
CD3D
HLA−DQB1
CCR7
CD8A
S100A11
NKG7
tSNE_1
tSNE_2
50
25
0
−25
−50
B
Naive Cytotoxic T cell
Regulatory T cell
Naive T cell
Memory T cell
Cytotoxic T cell
Memory T cell
Naive Cytotoxic T cell
Naive T cell
Cytotoxic T cell
Regulatory T cell
C
0
1
2
3
4
5
6
40
20
0
−20
−40
D
Cytotoxic T cell
Memory T cell
Naive Cytotoxic T cell
Naive T cell
Regulatory T cell
E
Regulatory T cell
Naive T cell
Naive Cytotoxic T cell
Memory T cell
Cytotoxic T cell
CD3D
HLA−DQB1
CCR7
CD8A
S100A11
NKG7
pct.exp
25
50
75
avg.exp
1
0
−1

Figure 3-3. Neural network predicts sub cell types.

(A) TSNE plots showing 6 maker genes' expression for the reference T cell subtypes.

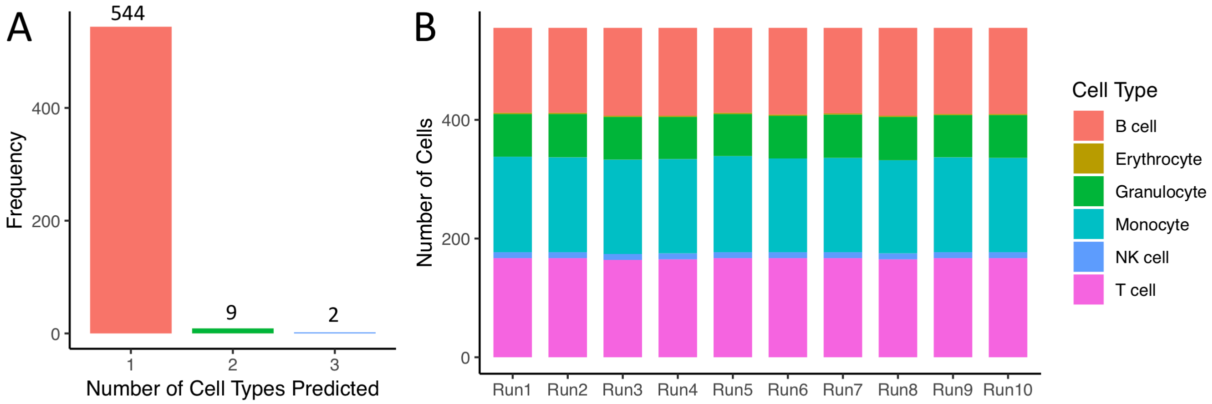
(B) T cell subtypes obtained to train the neural network.

(C) T cells from the human PBMC were grouped into 7 clusters by the unsupervised method.

(D) Subtypes predicted for the T cells from the human PBMC.

(E) Dot plot showing the expression of 6 genes for the predicted subtypes, dot size represents the percentage of cells expressing the gene, color scale represents the expression level of the gene.

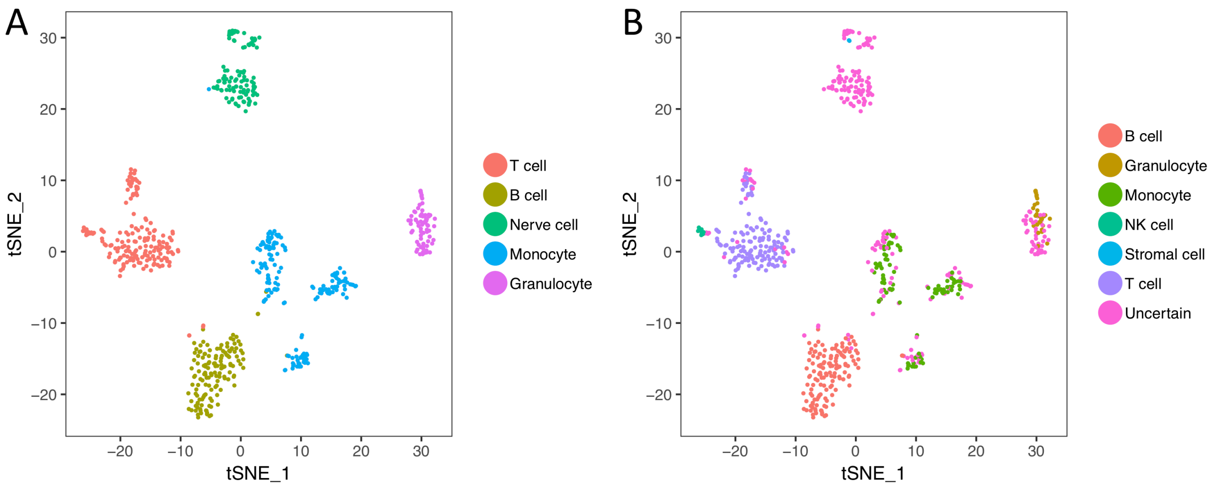
Supplemental information



Supplementary Figure 3-1. ACTINN produce consistent results from run to run.

(A) Number of cell types predicted for each cell across 10 runs.

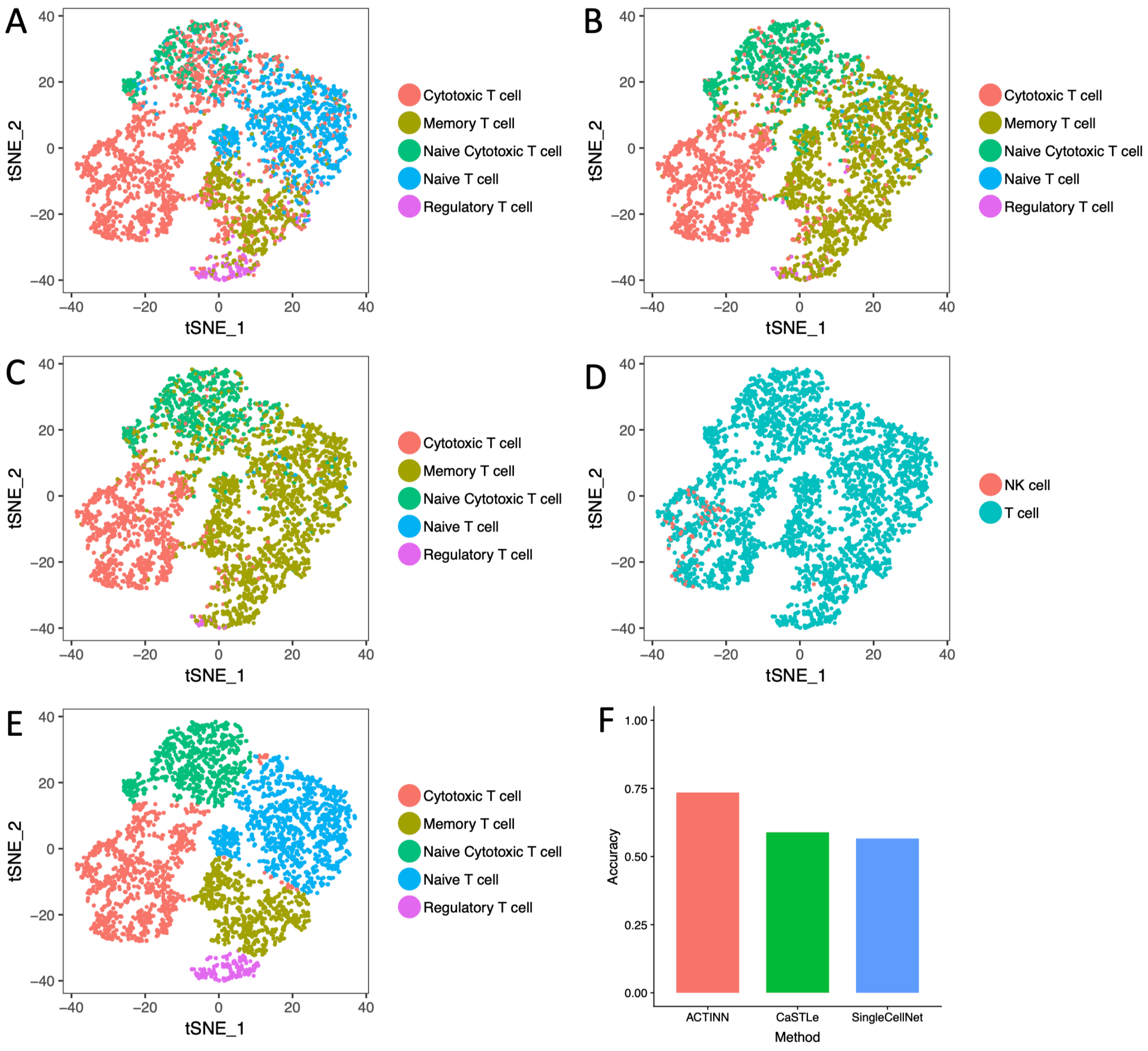
(B) Frequency for each cell type in each run.



Supplementary Figure 3-2. ACTINN accurately identifies cell types not in the reference.

(A) Cell types identified by unsupervised analysis.

(B) Cell types identified by ACTINN. A cell was labelled "uncertain" if its highest probability is smaller than 0.95.



Supplementary Figure 3-3. T cell subtypes predicted by 4 tools.

(A) ACTINN. (B) CaSTLe. (C) SingleCellNet. (D) SuperCT.

(E) True label inferred by all the predictions and gene expression.

(F) Accuracy for the predictions.

Cell Type	Predicted Type	Number
B cell	B cell	3437
Cardiac Muscle	Cardiac Muscle	123
Endothelial cell	Endothelial cell	3858
Epidermis	Epidermis	2065
Epithelial cell	Epithelial cell	3421
Erythrocyte	Erythrocyte	45
Granulocyte	Granulocyte	744
Hepatocyte	Hepatocyte	388
Monocyte	Monocyte	1416
NK cell	NK cell	215
Stromal cell	Stromal cell	2803
T cell	T cell	2143
Cardiac Muscle	Epidermis	2
Cardiac Muscle	Erythrocyte	2
Cardiac Muscle	Stromal cell	1
Endothelial cell	Erythrocyte	1
Epidermis	Epithelial cell	1
Epithelial cell	B cell	1
Erythrocyte	B cell	2
Erythrocyte	Endothelial cell	10
Erythrocyte	Epidermis	2
Erythrocyte	Epithelial cell	8
Erythrocyte	Monocyte	3
Erythrocyte	NK cell	1
Erythrocyte	Stromal cell	3
Erythrocyte	T cell	1
Granulocyte	B cell	1
Granulocyte	Monocyte	5
Monocyte	B cell	118
Monocyte	Epithelial cell	64
Monocyte	Granulocyte	1
Monocyte	NK cell	3
NK cell	T cell	47
Stromal cell	Endothelial cell	1
Stromal cell	Epithelial cell	2
Stromal cell	Erythrocyte	1
T cell	B cell	4
T cell	NK cell	3

Supplementary Table 3-1. Predicted cell type compared to original cell type when training on the 10X dataset and predicting on the SS2 dataset

Cell Type	Predicted Type	Number
B cell	B cell	8516
Cardiac Muscle	Cardiac Muscle	60
Endothelial cell	Endothelial cell	3641
Epidermis	Epidermis	7492
Epithelial cell	Epithelial cell	1159
Erythrocyte	Erythrocyte	130
Granulocyte	Granulocyte	1096
Hepatocyte	Hepatocyte	1764
Monocyte	Monocyte	1244
NK cell	NK cell	1016
Stromal cell	Stromal cell	3164
T cell	T cell	5601
B cell	Monocyte	3
Endothelial cell	B cell	1
Endothelial cell	Epithelial cell	1
Endothelial cell	Stromal cell	2
Epithelial cell	B cell	3
Epithelial cell	Epidermis	150
Epithelial cell	Erythrocyte	1
Epithelial cell	Monocyte	3
Erythrocyte	B cell	5
Erythrocyte	Granulocyte	15
Erythrocyte	Monocyte	2
Erythrocyte	T cell	3
Granulocyte	T cell	1
Monocyte	Granulocyte	12
NK cell	B cell	13
NK cell	Stromal cell	20
Stromal cell	Endothelial cell	1
T cell	Monocyte	1
T cell	NK cell	46

Supplementary Table 3-2. Predicted cell type compared to original cell type when training on the SS2 dataset and predicting on the 10X dataset.

Chapter 4 – Single Cell Transcriptomics Identifies a Cellular Ecosystem in Leprosy Lesions Encoding an Antimicrobial Response Network

Abstract

Granulomas are thought to be complex cellular responses comprised predominantly of macrophages and lymphocytes for containing and killing invading pathogens. Here, we investigated the antimicrobial response of single cells in human leprosy granulomas by studying reversal reactions (RR), a dynamic process in which some patients with the disseminated and immunologically unresponsive form of the disease, lepromatous leprosy, transition to a reactional state that is characteristic of tuberculoid leprosy, the self-limiting form able to generate an effective antimicrobial response. Using single cell RNA-seq to compare the granulomatous response in RR vs. lepromatous leprosy lesions, we were able to cluster cells into subtypes of T cells, myeloid cells, keratinocytes (KC), endothelial cells (EC) and fibroblasts (FB). We identified 1,124 genes encoding proteins known to be involved in the antimicrobial response (AMGs) enriched in RR cells across cell types, most strongly regulated by IFN-γ. In addition, we identified pseudotime trajectories for macrophages and KC that mapped the progression of cells from lepromatous to RR, revealing IL-1β as a key upstream regulator of both. We constructed an granuloma ecosystem by integrating the IFN-γ and IL-1β antimicrobial targets with the cell-cell co-abundance in lesions, which revealed that antimicrobial pathways in granulomas extend beyond the interaction simply of T cells and macrophages to include activated keratinocytes and inflammatory fibroblasts.

Introduction

The hallmark of the chronic inflammatory response to a foreign substance that has resisted destruction by an acute inflammatory response is the granuloma. In the most cited article on granulomas, Gordon defined granulomas as structures “which are formed by the immune-mediated recruitment of white blood cells, and particularly rich in macrophages” [10]. In the context of infectious diseases, the function of the granuloma is to sequester and degrade microbial pathogens that have evaded the early immune response.

Leprosy offers an attractive model to investigate the mechanisms by which the human immune system combats intracellular bacteria as the disease presents as a clinical/immunologic spectrum [11]. Because it manifests as a spectrum of disease in skin, the dynamics are accessible to study, in contrast to tuberculosis granulomas. At one end of the disease spectrum, tuberculoid leprosy typifies the host’s antimicrobial response, which controls the pathogen: there are few lesions; *Mycobacterium leprae* bacilli are rare; and patients eliminate the infection. At the opposite end of the spectrum, lepromatous leprosy (L-lep) represents susceptibility to disseminated infection, with numerous skin lesions and abundant bacilli. The disease spectrum is dynamic, as patients may undergo a reversal reaction (RR), in which patients generally upgrade, either spontaneously or during chemotherapy, from the lepromatous to the tuberculoid pole. The structure of granulomas is distinct across the spectrum of leprosy. The granulomas in tuberculoid leprosy contain a core of mature macrophages with occasional multinucleated giant cells. These granulomas are organized with lymphocytes forming a mantle zone at the periphery of the granuloma. Granulomas in RR lesions are histologically similar to those in tuberculoid leprosy with the presence of intercellular edema. In lepromatous leprosy, the granulomas are disorganized, immature lipid-laden macrophages are prominent with lymphocytes scattered throughout.

The study of leprosy lesions has provided insight regarding the host immune response to intracellular bacteria and the architecture of granulomas. Through various approaches, it has been possible to define functional subpopulations of human T cells [12-15] and macrophages [16], their microanatomic distribution as well as the patterns of cytokine secretion that influence the outcome of infections caused by pathogenic mycobacteria [17-20].

Given that the resolution of the granulomatous response requires destruction of the foreign invader, the antimicrobial mechanisms that result in the death of the pathogen are central to understanding how granulomas contribute to host defense. A few pathways have been identified by the study of human cells that can lead to an antimicrobial activity against intracellular mycobacteria. Through activation via TLRs and secretion of IFN-γ, the innate and adaptive immune systems trigger the vitamin D-dependent induction of the antimicrobial proteins encoded by *CAMP* and *DEFB4A* [16, 21, 22]. T cells release antimicrobial proteins encoded by *GNLY* and *IL26*, which can enter infected macrophages and exert a direct antimicrobial activity [13, 14, 23, 24]. These human pathways are not present in mice, which utilize other mechanisms such as the release of nitric oxide to kill mycobacteria. The advent of scRNA-seq provides an opportunity to elucidate the cell-cell networks that define antimicrobial responses at the site of infection. We used this approach to study and compare the immune responses in RR vs. L-lep patient skin lesions to gain insight into mechanisms of host defense used by granulomas to eliminate an intracellular bacterium.

Materials and Methods

Processing of Human Skin

Skin biopsy specimens were obtained from patients with leprosy at University of Southern California and Brazil. Patients were classified according to standard clinical and histologic criteria. Five patients with reversal reaction are designated here as RR1, RR2, RR3, RR4 and RR5. The other five are designated here as L-lep1, L-lep2, L-lep3, L-lep4 and L-lep5, of which four were classified as LL and one as BL.

For each sample, a 4-mm punch biopsy was obtained following local anesthesia and was placed immediately into 10 mL of RPMI on ice. Initially, skin biopsies were incubated in 5mL of a 0.4% Dispase II solution (Roche Inc.) at 37°C for 1 hour with vigorous shaking. The dermis and epidermis were then carefully separated using forceps and transferred to separate tubes for additional processing. Epidermal samples were placed in 3mL of 0.25% Trypsin and 10U/mL DNAse for 30 minutes at 37°C. Trypsin was neutralized with 3mL of fetal calf serum (FCS), and the tissue was passed through a 70-micron nylon cell strainer which was washed with 5mL of RPMI. Epidermal cells were then pelleted at 300xg for 10 minutes and counted. Dermal samples were minced with a scalpel and incubated in a solution of 0.4% collagenase 2 and 10 U/mL DNAse for 2 hours at 37°C with agitation. The cell suspension was passed through a 70-micron cell strainer and washed with 5mL of RPMI. Cells were pelleted at 300xg for 10 minutes, resuspended in 1mL of RPMI and counted. MACS enrichment for CD1a$^+$ cells was performed for epidermis from three RR patients.

Sequencing and alignment

Libraries were sequenced on an Illumina Nova-Seq (Illumina, San Diego, CA) as 50bp paired end reads and were converted from bcl files to fastq files using bcl2fastq. We use Nextera N700 indices to identify individual samples. The alignment was performed using Drop-seq pipelines (version 1.12) previously described [50]. Briefly, the raw reads were aligned to the concatenated human (hg38) and M. leprae genome using STAR [29]. Each read was tagged with a 12bp barcode and 8bp unique molecular identifier (UMI). After alignment, the reads were grouped by the barcodes and deduplicated using the UMIs. The number of UMIs was then counted for each gene in each cell to generate the digital expression matrix (DEM).

Removal of ambient RNA contamination

Ambient RNA contamination was removed using SoupX [51]. Specifically, we examined the distribution of UMIs for each gene and selected the genes for which the distribution most closely approximated a uniform distribution. For each sample, we calculated an array-specific "soup" profile among barcodes below the UMI threshold. To calculate estimated per-cell contamination fractions, we manually selected genes observed to be bimodally expressed across cells, which suggest that these genes are predominantly expressed in a single cell type but are observed at low levels in other cell types for which endogenous expression would not be expected. For each array, we removed individual transcripts most likely to be contamination from each single cell based on the estimated contamination fraction. Specifically, individual transcripts were sequentially removed from each single cell transcriptome until the probability of subsequent transcripts being soup-derived was less than 0.5 to generate a background-corrected UMI matrix for each Seq-Well array.

Cell clustering and cell type annotation

Digital expression matrices for human genes from all 10 samples were merged, and the R package Seurat [44] was used to cluster the cells in the merged matrix. Cells with less than 300 genes detected or more than 50% mitochondrial gene expression were first filtered out as low-quality cells. Genes detected in less than five cells were removed as low-abundance genes. The gene counts for each cell were divided by the total gene counts for the cell and multiplied by the scale factor 10,000, then natural-log transformation was applied to the counts. The FindVariableFeatures function was used to select 2,000 variable genes with default parameters. The ScaleData function was used to scale and center the counts in the dataset. Principal component analysis (PCA) was performed on the variable genes, and 13 PCs (based on the elbow point of variance explained by each PC) were used for cell clustering (resolution = 0.5) and Uniform Manifold Approximation and Projection (UMAP) dimensional reduction. The cluster markers were found using the FindAllMarkers function, and cell types were manually annotated based on the cluster markers. To generate the heatmap showing the cell type markers, the top 100 cells with the highest number of UMI detected were plotted for each cell type. The total number of M. leprae UMIs were calculated for each cell and plotted for each sample.

Cell type sub-clustering

We performed sub-clustering on endothelial cells, fibroblasts, keratinocytes, myeloid cells and T cells. The same functions described above were used to obtain the sub-clusters. To choose the number of PCs, the rank of PCs based on the percentage of variance explained was plotted, and the elbow point was chosen as the number of PCs to use in cell clustering (resolution = 0.6) and UMAP dimension reduction. Clusters that were defined exclusively by mitochondrial gene expression, indicating low quality, were removed from further analysis. To generate the heatmap with marker genes for each sub-cluster, the top 100 sub-cluster marker genes with the highest average log fold change were plotted, and five representative genes were labelled.

Interferon signature enrichment analysis

Supervised analyses were performed to identify Type I and Type II IFN regulated genes as described previously [20, 52-54]. Differentially expressed genes in TC1 (RR CTL) and TC2 (L-lep CTL) were identified using a Wilcoxon rank sum test with adjusted p value cutoff at 0.05. A list of genes specifically induced by only IFN-α/β or IFN-γ was derived from the gene expression profile data of IFN-treated human PBMC [55]. 148 IFN-α/β specific genes and 33 IFN-γ specific genes were identified, which were overlapped with TC1 and TC2 specific genes to determine the differential expression of IFN-regulated genes. Hypergeometric test was used to determine the enrichment level, a p value small than 0.05 was considered to be significantly enriched.

Pseudo-time analysis

Pseudo-time trajectories for macrophage and keratinocyte sub-clusters were constructed using the R package Monocle [56]. The raw counts for cells in the intended sub-clusters were extracted and normalized by the estimateSizeFactors and estimateDispersions functions with the default parameters. Genes with average expression larger than 0.5 and detected in more than 10 cells were retained for further analysis. Variable genes were determined by the differentialGeneTest function with a model against the sub-cluster identities. The top 500 variable genes with the lowest adjusted p value were used to order the cells. The orders were determined by the orderCells function, and the trajectory was constructed by the reduceDimension function with default parameters. Differentially expression analysis was carried out using the differentialGeneTest function with a model against the pseudotime, and genes with an adjusted p value smaller than 0.05 were clustered into 6 patterns and plotted in the heatmap.

Antimicrobial gene analysis

A list of 1,404 genes were curated by searching for genes with "antimicrobial" as a keyword in GeneCards (https://www.genecards.org/). To study the difference of antimicrobial response in L-lep and RR, the cell types were split into L-lep and RR groups. To measure the relative abundance of anti-microbial genes (AMGs), the total expression of each AMG was calculated for each L-lep and RR cell type. The AMG expression for the L-lep cell types was normalized by the total number of L-lep cells, and the AMG expression for the RR cell types was normalized by the total number of RR cells. The z scores were calculated across all L-lep and RR cell types for each AMG. A cutoff of z score > 3 was applied to obtain the specific AMGs for each cell type. A list of 1,124 AMGs was obtained as specific to at least one RR cell type. Ingenuity Pathway Analysis was applied to the 1,124 AMGs, and the upstream regulators were ranked by p value. To generate the circos plots, a list of direct antimicrobial genes was obtained from The Antimicrobial Peptide Database [57], and those regulated by *IL1B* or *IFNG* were included.

Cell type composition analysis

To calculate the sample composition based on cell type, the number of cells for each cell type from each sample were counted. The counts were then divided by the total number of cells for each sample and scaled to 100 percent. The same procedures were applied to calculate the sample composition for each subtype in endothelial cells, fibroblasts, keratinocytes, myeloid cells and T cells. The cell type (including the subtype) with more than 70% L-lep (or RR) composition was named L-lep (or RR) specific. The cell type compositions were combined, and the correlation matrix was generated by calculating the correlation for each pair of cell types. Hierarchical correlation was performed on the correlation matrix and plotted in the heatmap, with L-lep specific cell types were labelled in red and RR specific in blue. To construct the antimicrobial networks, the correlation coefficient was filtered to be at least 0.5 between the linked cell types, and the number of AMG links was filtered to be at least 10. The connections were directed from the cell types that express the upstream regulator to the cell types that express the AMGs.

Results

Major cell types in leprosy lesions

To study the transcriptional changes between RR and L-lep, we performed single cell RNA sequencing by Seq-Well on skin biopsy specimens from five RR and five L-lep patients. After quality filtering, we retained 21,318 cells, with an average 741 genes and 3,556 transcripts per cell. To study the heterogeneity of these cells, we selected variable genes, performed UMAP dimension reduction and cell clustering using the R package Seurat. We then ran differential expression analysis to find the cluster markers and overlapped the cluster markers to canonical cell type defining signature genes. Ultimately, we recovered 12 primary cell types across all 10 samples (Figure 4-1A). These annotated cell types include: T cells (TC; *CD3D* and *TRBC2*), B cells (BC; *MS4A1* and *CD79A*), plasma cells (PLC; *IGHG1* and *IGHG3*), myeloid cells (ML; *C1QA* and *LYZ*), Langerhans cells (LC; *CD207* and *CD1A*), mast cells (Mast; *CPA3* and *CTSG*), keratinocytes (KC; *KRT1* and *KRT10*), fibroblasts (FB; *COL1A1* and *DCN*), smooth muscle cells (SMC; *ACTA2* and *TAGLN*), endothelial cells (EC; *PECAM1* and *VWF*), eccrine gland cells (ECG; *DCD* and *MUCL1*) and melanocytes (MLNC; *DCT* and *PMEL*) (Figure 4-1C).

The major cell types, including T cells, myeloid cells, keratinocytes, endothelial cells and fibroblasts were found in both the RR and the L-lep lesions (Figure 4-1B, 1D). Although B cells were found in both RR and L-lep lesions, plasma cells were derived predominantly from L-lep lesions. Given that LC are more frequent in RR than L-lep lesions [12], we immunoselected CD1a^{+} cells from the epidermis from three RR patients, adding these back to the dermal cells, accounting for the high frequency of LC from these RR lesions. *M. leprae* reads were most prevalent in the multibacillary L-lep lesions, but also detected at a lower level in one RR lesion (Figure 4-1E, Supplementary Figure 4-1).

Major cell sub-clusters in leprosy lesions

We detected seven T cell sub-clusters, two predominantly derived from RR lesions and one from L-lep lesions (Figure 4-2A). T cell sub-cluster 0 (TC0) express the classic Th17 cell markers *RORC*, *RORA*, *RBPJ* and *IL23R* (Figure 4-2B, 2C), although the expression levels of the major Th17 cytokine genes were low. TC1 and TC2 are designated as cytolytic T lymphocytes (CTL) as they both contain *CD8A*, *GZMB* and *PRF1*; TC1 was derived mainly from RR lesions (RR CTL) and TC2 was mainly derived from L-lep lesions (L-lep CTL) (Figure 4-2D). We noted several type I IFN downstream genes in L-lep CTL including *IFI44L*, *MX1*, *IRF1* and *OAS3*. We ran differential expression analysis between L-lep CTL and RR CTL, and performed enrichment analysis on the differentially expressed genes using IFN signatures derived from activated human PBMC [20]. Genes up-regulated in L-lep CTL were significantly enriched in type I IFN downstream signatures (Supplementary Figure 4-2). The remaining sub-clusters contained a mixture of cells from RR and L-lep including TC3 (TCM, T-central memory, *IL7R* and *CCR7*). TC4 (naïve, *LEF1*, *JUNB*), TC5 (Treg, *FOXP3*, *CTLA4*) and TC6 (antimicrobial CTL (amCTL), *GZMB*, *PRF1* and *GNLY*) containing a mixture of tricytotoxic T cells (T-CTL) and γδ T cells (Figure 4-2C).

We previous described a functional subset of CTL, amCTL, expressing *GZMB*, *PRF1* and *GNLY*, that exert antimicrobial activity against intracellular *M. leprae* and correlate with protective immunity to tuberculosis and leprosy [58, 59]. In TC6, the expression of *GZMB*, *PRF1* and *GNLY*, were greater in cells from RR vs. L-lep lesions, with the aggregation score for these three genes, the T-CTL score that characterizes amCTL, significantly greater in RR lesions (Figure 4-2E). *IFNG* was most strongly expressed by Th17 cells (TC0) and RR CTL (TC1) but was also present in L-lep CTL (TC2) and amCTL (TC6) (Figure 4-2C). Within cells from RR lesions, the number of *IFNG*-expressing cells was similar to the number of either *GZMB*$^+$ or *PRF1*$^+$ cells in RR CTL and amCTL, as well as the total CTL from both sub-types (Figure 4-2F). Strikingly, the number of *IFNG*-expressing cells was far greater than the *GNLY*$^+$ amCTL for both the total CTL and RR CTL, but the number of *IFNG*$^+$ and *GNLY*$^+$ cells were equal in the amCTL (TC6). These data indicate that *IFNG* is a marker for all CTL, but is not a useful marker for estimating for amCTL, a smaller subset of CTL that express *GNLY* in addition to *GZMB* and *PRF1* shown to kill infected cells and the intracellular bacteria within them.

We identified five myeloid sub-clusters, three predominantly derived from RR lesions (ML0, ML3, ML4) and two from L-lep lesions (ML1, ML2) (Figure 4-3A, 3B). ML0 is comprised of a mixture of dendritic cells (DC), with distinct subpopulations expressing *CD1C* and *LAMP3* (Supplemental Figure 4-3). ML1 from L-lep lesions express type I IFN downstream genes including *IFI44L*, *MX2* and *IFIT3*. ML2 from L-lep lesions are TREM2 MΦ based on expression of *TREM2* and *APOE*. ML4 from RR lesions are M1-like MΦ, with *LYZ*, *MMP9* and *IL23A*. ML3 was derived from three RR lesions and two L-lep lesion and appears to be a transitional population between ML2 and ML4 (Figure 4-3D), expressing genes from both sub-clusters (Supplemental Figure 4-3). *TREM2* and *APOE* expression, as well as a TREM2 score comprised of nine conserved genes from seven datasets [60-66], were highest in ML2, declining in ML3 and ML4 (Figure 4-3E, Supplemental Figure 4-3).

Of seven keratinocyte sub-clusters, two were enriched in RR patients, KC3 (*FLG*$^+$ granular layer KC) and KC4 (*KRT14/15*$^+$ basal layer KC), and two enriched in L-lep patients, KC1 and KC5, both derived from spinous layer KC (Figure 4-4). KC0 (spinous-1 KC), KC2 (supraspinous KC) and KC6 (hair follicle KC) were derived from both RR and L-lep samples.

For fibroblasts, *SFRP2*$^+$ FB (FB0) and *CXCL2*$^+$ FB (FB2) were enriched in RR lesions, with two additional sub-clusters mainly derived from L-lep lesions (Supplementary Figure 4-4). The *SFRP2*$^+$ FB sub-cluster, which expresses *COL3A1*, *COL18A1* and *COMP*; have been shown to be involved in the deposition of extracellular matrix [67]. In addition, the *CXCL2*$^+$ FB sub-cluster expresses a number of inflammatory genes as specific marker genes including *IL6*, *CCL2*, *CXCL3*, *CXCL8* and *IL32*, which displays a similar expression profile to the inflammatory fibroblasts detected in atopic dermatitis skin lesions [68]. Two FB sub-clusters were predominantly derived from L-lep lesions. *MGP*$^{+}$ FB (FB1) are a population of FB found in the reticular dermis. *COL11A1*$^{+}$ FB (FB3) have also been reported in skin [67].

Of six endothelial cell sub-clusters, *LYVE1*$^{+}$ lymphatic EC (EC4) and *HEY2*$^{+}$ EC (EC5) were mainly derived from RR lesions. *MEOX2*$^{+}$ EC (EC3) were primarily found in L-lep lesions (Supplementary Figure 4-5); *MEOX2* is an inhibitor of NF-□B activation in EC [69].

Antimicrobial genes in RR lesions

Given that leprosy RRs are associated with a reduction in viable *M. leprae* bacilli in lesions, we sought to determine the array of antimicrobial genes that were present in defined cell populations. We integrated a curated list of 1,404 genes known to encode proteins that contribute to antimicrobial responses (AMGs, antimicrobial genes) with the scRNA-seq data. To do so, we divided each cell type (including subtype) by RR vs. L-lep cells, and calculated the z score for expression of each gene across all cells of each type (Methods). This metric captures the total amount of each transcript by cell type in our lesions, which we believe is relevant for measuring the extent of the antimicrobial effect of a gene. We compared the sum of z scores for L-lep and RR cell types and found that RR cell types have a higher expression pattern for the AMGs (Figure 4-5A). We identified 1124 AMGs with a z score $\geq$3 in at least one RR cell type. A high z score indicates that these cells, in aggregate, produce relatively more of the specific transcript than other cell types. We identified the upstream regulators of the 1,124 AMGs using Ingenuity Pathways Analysis, with *IFNG*, *TNF* and *IL1B* having the highest enrichment scores (Figure 4-5B). We then calculated the sum of the scores for the top 20 URs in L-lep and RR cell types and found that URs are significantly more highly expressed in RR (Figure 4-5C).

Pseudotime analysis

The curved linear shape of sub-clusters in both the myeloid and KC subpopulations suggested the linear transition of cells indicative of differentiation. By using Monocle to perform pseudotime analysis, we ordered the cells in TREM2 MΦ (ML2), transitional MΦ (ML3) and M1-like MΦ (ML4) into a linear progression, starting from L-lep enriched cells and ending with RR, which mirrors the clinical progression seen in patients that start at the lepromatous pole and subsequently develop RR (Figure 4-5D). Using a similar analysis on spinous-2 KC (KC1), supraspinous KC (KC2) and granular KC (KC3), we identified a pseudotime continuum from L-lep to RR derived cells (Figure 4-5E).

We hypothesized that the upstream regulators which trigger the antimicrobial response also induce cellular differentiation in lesions. To test this, we split the variable genes into six expression patterns for both macrophage and keratinocyte pseudotimes (Supplementary Figure 4-6) and identified the upstream regulators for each expression pattern. We calculated a module score for each upstream regulator using the targets found in all six patterns, and calculated the correlation coefficient between the module score and the pseudotime. Of the top URs for the AMGs, only the target scores for *IL1B* were highly correlated with both the macrophage and keratinocyte pseudotime (R=0.63 and 0.83, respectively) (Figure 4-6F, 6G). The target scores for *IFNG* correlated with keratinocyte (R=0.84) but not macrophage pseudotime (R= 0.04). To this end, we selected for further study the *IFNG* and *IL1B* target genes, as *IFNG* had the highest enrichment score for the AMGs, and *IL1B* was not only a top upstream regulator of the AMGs but the expression of the *IL1B* target genes was highly correlated with both macrophage and keratinocyte pseudotimes.

Antimicrobial gene network and ecosystem

To construct a gene network depicting the antimicrobial response in RR, we first determined the source of the two key upstream regulators. *IFNG* was detected (z score ≥3) in RR cells from Th17 cells (TC0) and RR CTL (TC1) and *IL1B* in RR cells from LC and DC (ML0) (Figure 4-6A). As such, our analysis reveals that source of *IL1B* and *IFNG,* which represent key upstream regulators of the antimicrobial response genes that mediate the innate and the adaptive immune responses respectively in restricting the infection.

Next, we constructed circos plots to depict the interactions of the *IL1B* and *IFNG* expressing cells with the target AMG expressing cells (Figure 4-6B). For clarity, we limited the number of interactions to AMGs that encode proteins with direct antimicrobial activity and having a z score ≥3 in at least one RR cell type (Supplemental Figure 4-7 and 8). In view of the variable expression of the genes encoding the receptors for IL-1β and IFN-γ in the scRNA-seq dataset, we inferred connections between the upstream regulators and these AMGs as identified using Ingenuity Pathway Analysis. *IL1B* was linked to 30 unique direct AMG targets with 42 connections to RR cells, and *IFNG* was linked to 28 unique direct AMG targets with 44 connections to RR cells. *IL1B* and *IFNG* shared 22 AMG targets, with 14 AMGs exclusive to only one of the upstream regulators. For both *IL1B* and *IFNG*, the majority of connections were to cell types that were predominantly associated with cells from RR lesions, and strikingly there were no connections to a cell type that was predominantly derived from L-lep lesions.

We further explored the nature of cellular communication in leprosy lesions by determining the cell-cell interactions according to cell type co-abundance correlation. We reasoned that cells that interact are more likely to be present together with correlated abundance across lesions. Two major clusters were identified, with the RR and L-lep cell types forming distinct patterns (Figure 4-6C). A major group in the RR branch contained LC, DC, Th17, RR CTL and M1-like MΦ. The L-lep branch included both B cell and Treg, although these cell types were composed of cells from both RR and L-lep lesions.

Next, a cellular ecosystem for RR lesions was derived by combining the results from the upstream regulators to AMG connections with the cell-cell correlation map, including only cell-cell co-abundance correlations ≥0.5 to limit the number of interactions (Figure 4-6D). DC expressing *IL1B* was abundantly linked by connections to AMG targets in LC, basal *KRT14/15*$^{+}$ KC, *CXCL2*$^{+}$ FB and Th17 cells, and most highly correlated with Th17 cells. LC was abundantly linked to AMG targets in *CXCL2*$^{+}$ FB and Th17 cells, but also highly correlated with RR CTL, granular *FLG*$^{+}$ KC and DC. The connections from the major *IFNG*-expressing cells, Th17 and RR CTL, were greatest to AMG targets in LC. Th17 cells were also robustly connected to AMGs in basal *KRT14/15*$^{+}$ KC, with weaker connections to the other cell types. RR CTL was highly connected with AMG targets in *CXCL2*$^{+}$ FB, but correlated by cell-cell abundance to LC, granular *FLG*$^{+}$ KC and M1-like MΦ.

Finally, we depict the connections associated with antibacterial responses between *IL1B* and *IFNG* with target AMGs in cells of co-abundance correlation ≥0.5 (Figure 4-7). We found that *IL1B* drives the differentiation of both MΦ and KC in RR. The clinical RR syndrome that develops in lepromatous patients is characterized by a change from immature to mature macrophages with a reduction in the number of intracellular bacilli. *IL1B* together with *IFNG* trigger an antimicrobial response including the induction of genes in MΦs encoding enzymes and antimicrobial peptides that can kill bacteria. In addition, the innate system, including LC, DC, *SFRP2*$^{+}$ FB, *CXCL2*$^{+}$ FB and various KC subtypes, express genes encoding antimicrobial proteins as well as chemokines known to have a direct antimicrobial response. The adaptive T cell response contributes to the antimicrobial response against intracellular bacteria in macrophage via amCTL that express *GZMB*, *PRF1* and *GNLY*, as well as Th17 cells expressing *IL26*. Thus, the cellular ecosystem is a multifaceted highly interconnected system that acts to contain infection by an intracellular bacterium in leprosy through the engagement of innate and adaptive cells, both within and outside the granuloma, to form an integrated antimicrobial network.

Discussion

The organized granulomatous response allows the immune system to wall off and eliminate intracellular bacteria that have initially evaded destruction. Investigation of the immune interactions in such granulomas has previously, and almost exclusively, focused on the role of specific myeloid and lymphocytic populations. The dynamics of the leprosy spectrum provide a unique opportunity to study pathways of host defense against intracellular bacteria. Patients classified towards the lepromatous pole have disseminated infection with many bacilli in macrophages in diffuse aggregates of macrophages and lymphocytes. They can develop RR, characterized by organized granulomas, inflamed lesions with reduced numbers of bacilli. Our premise has been that the study of the changes in the cellular response of immunologically unresponsive patients with L-lep to the immunologically reactive RR granulomas will enable understanding of mechanisms likely to contribute to the antimicrobial response.

Here, we performed single cell transcriptomics on cell types comprising the granulomatous response in leprosy skin lesions. Of 43,363 genes in 21,282 cells studied, we detected 1,124 AMGs that were differentially expressed in RR lesion-derived cells across all cell types. Analysis by scRNA-seq revealed that the immune response diverges across the spectrum of leprosy not only for distinct populations of immune cells, including subpopulations of myeloid cells and T cells, but also for subpopulations of fibroblasts, endothelial cells and keratinocytes. The expression of these antimicrobial genes as well as the upstream regulators *IL1B* and *IFNG* for which these AMGs serve as targets was significantly higher in RR compared to L-lep lesions. From this data, we formulate a cellular ecosystem by integrating cell-cell co-abundance in lesions with the links between cells expressing the upstream regulators *IL1B* and *IFNG* to RR cell types expressing the downstream AMG targets. Key antimicrobial subpopulations associated with immunity in RR included cells of the myeloid and lymphocyte lineages including LC, DC, M1-like MΦs, Th17 cells, CD8$^+$ CTL and amCTL. Strikingly, the antimicrobial responses included two distinct subpopulations of fibroblasts, *SFRP2*$^+$ FB and *CXCL2*$^+$ FB as well as various KC subpopulations.

As expected, the activation of macrophages, DC and T cells contributed to the antimicrobial response network in granulomas. We found that DC and LC express *IL1B*, and Th17 cells and RR CTL express *IFNG*, two major upstream regulators of the AMGs. In macrophages, the AMGs include genes that encode proteins with direct antimicrobial activity such as *CCL18* [70], and as reported, the vitamin D-downstream, IL-1β dependent and IFN-γ responsive genes *CAMP* and *DEB4A* [16, 21, 22, 71]. Macrophages also expressed *CYBB*, the gene encoding the gp91(phox) subunit of the phagocyte NADPH oxidase, that if deleted results in enhanced susceptibility to mycobacterial infection [72]. In addition, macrophages expressed *CCL3* and *MMP9*, reported to be involved in an antibacterial response [73]. Th17 cells expressed *IL26*, encoding an antimicrobial protein that is taken up by *M. leprae* infected macrophages, colocalizes with the intracellular bacteria and also, by binding DNA activates STING, resulting in phagolysosomal fusion and an antimicrobial activity [24, 74]. IL-26 can be induced in Th17 cells either by activation via the T cell receptor, or by an innate pathway via IL-1β, the latter results in IL-26 release without production of other Th17 cytokines [75]. Although TC6, amCTL, was present in both RR and LL, there was higher expression of *GZMB*, *PRF1* and *GNLY* in the RR cells; these genes encode proteins that act synergistically when released from amCTL to kill intracellular mycobacteria [13-15, 58, 59]. Among the two CTL sub-clusters, RR-CTL and amCTL, the RR cells in both contained *IFNG* expressing cells, but only amCTL expressed *GNLY*. RR CTL were more abundant than amCTL, such that detection of IFN-γ is an equivalent measure of all CTL, but not amCTL, which require expression of *GZMB*, *PRF1* and *GNLY* for their antimicrobial activity.

Surprisingly, the cell-cell analysis indicated that AMGs were expressed in keratinocytes and fibroblasts, cells not typically considered to contribute to the antimicrobial response, indeed expressed AMGs in RR granulomas. There were two distinct fibroblast subpopulations in RR, *CXCL2*$^{+}$ FB and *SFRP2*$^{+}$ FB. *CXCL2*$^{+}$ FB express a number of inflammatory genes that are directly antimicrobial: *ADM*, *CCL11*, *CXCL12*, *CXCL2*, *CXCL3*, *CXCL9*, *CCL26*, *CXCL10* and *CCL2* [57, 70, 76]. *CXCL2*$^{+}$ FB also express *MMP2* [77] and *MMP8* [78], known to contribute to antimicrobial responses, as well as *IL32*, which was shown to be involved in the vitamin D-dependent antimicrobial pathway and a marker of protective immunity in tuberculosis [79]. *CXCL2*$^{+}$ FB were also observed in atopic dermatitis and considered to be inflammatory [68], where these cells may contribute to host defense in granulomas as antimicrobial fibroblasts. *SFRP2*$^{+}$ FB have been shown to be involved in the deposition of extracellular matrix proteins [67]. In RR, *SFRP2*$^{+}$ FB express genes encoding vimentin (*VIM*), fibrillin (*FBN1*) and multiple collagens. Fibroblasts expressing vimentin microanatomically located at the periphery of the granuloma in tuberculosis [80], are thought to be responsible for laying down “fibrin” to build a wall to prevent bacteria from exiting the granuloma and disseminating. Our data suggest that "fibrin", an old pathologic term for the appearance of this material on hematoxylin and eosin stained microscopic sections, may be composed of multiple extracellular matrix proteins. *SFRP2*$^+$ FB also express genes encoding chemokines with direct antimicrobial activity, some more strongly than expressed in *CXCL2*$^+$ FB such as *CXCL12*, *CXCL6*, *CXCL9* and *CXCL17* [57, 70, 81, 82]. These cells express *CXCL8* as well, albeit at lower levels than LC, DC and M1-like MΦs. Of note, fibroblast secretion of CXCL8 has been demonstrated to limit the survival of *M. tuberculosis* in infected macrophages [83]. Thus, *SFRP2*$^+$ FB may represent a second subpopulation of antimicrobial fibroblasts. The epidermis is activated in RR, with hyperplasia of keratinocytes expressing MHC class II and IFN-γ inducible protein 10 (IP-10) [84, 85] indicating activation by IFN-γ. Keratinocytes in RR expressed genes encoding antimicrobial peptides, all except *PI3* [57] have antimycobacterial activity: *DEFB1 [86]*, *RNASE7 [87]*, *S100A8*/*S100A9* [88-90] and *TAC1* [91]. Lee *et al.* demonstrated that production of antimicrobial peptides from KCs would result in increased antimicrobial activity in the dermis, presumably by diffusion of the peptide across the dermal epidermal junction [92]. Endothelial cells expressed a number of AMGs including *APP*, *CCL24*, *CXCL11*, *LEAP2*, *SNCA*, *TSLP*, *VIP*, *EC1*, *EC3*, *EC4*, *CCL21* and *NTS* [57]. However, we could not link the endothelial sub-clusters expressing these genes to cells expressing the upstream regulators *IL1B* and *IFNG* in the cell-cell ecosystem. Thus, the fate of granulomas is not only dictated by the response of macrophages and lymphocytes, the classic cells per the longstanding definition of a granuloma, but by a multiplicity of cell types including fibroblasts and keratinocytes.

We focused on two major upstream regulators, *IL1B* and *IFNG*, that have a substantial effect on the immune response in RR lesions, recognizing that additional upstream regulators contribute to the antimicrobial response. *IFNG* was most highly correlated with the AMGs associated with RR, consistent with previous findings demonstrating an upregulation of IFN-γ in RR concomitant with a change from a Th2 to a Th1 response in paired samples in the same individuals from before and during RR [17, 19, 20, 93]. Of the top upstream regulators of the AMGs, only *IL1B* regulated the pseudotime trajectory of both macrophages and keratinocytes as patients transitioned from L-lep to RR. The macrophage psuedotime trajectory maps from the TREM2 macrophages in L-lep lesions, to the transitional macrophages in two L-lep patients and three RR patients, to the M1-like macrophages in RR lesions. Previously, we found that the TLR2 ligand, lipopeptide, in combination with IFN-γ, triggered macrophage plasticity in a similar trajectory, with macrophages reversing from M2-like to M1-like in vitro as observed in RR lesions [16, 94]. Since TLR2 and IL-1β both signal via MyD88, and IFN-γ is upregulated during RR, the key signals are present to facilitate the plasticity of macrophage differentiation to the M1 state known to have high antimicrobial activity. Similarly, the pseudotime mapped keratinocyte maturation, ending with a gene pattern indicates activation by both IL-1β and IFN-γ, with expression of *IL18*, which encodes for a protein that further upregulates IFN-γ in mycobacterial infection [95].

Several features of the cell sub-types that were overrepresented in L-lep lesions may contribute to pathogenesis in leprosy. Among myeloid cells, L-lep lesions contained few DC but two distinct macrophage populations. TREM2 macrophages have been identified in several diseases characterized by altered lipid metabolism including atherosclerosis, Alzheimer's disease, non-alcoholic steatohepatitis and obesity [60-66]. The gene program in TREM2 macrophages suggests that these cells are programmed to transport and process lipids, most likely they are the foam cells or foamy macrophages that characterize both atherosclerosis and L-lep. One myeloid cell subtype, Type I IFN MΦ, and one T cell subtype, L-lep CTL, both characterized by a type I IFN gene program. Previously, we found that as opposed to IFN-γ and its downstream target genes that were preferentially expressed in RR lesions, IFN-β and its downstream targets were preferentially expressed in L-lep lesions [20]. We have previously reported that IFN-β can inhibit the antimicrobial effects of IFN-γ on macrophages in vitro [20]. Of note, the RR CTL sub-cluster was remarkably similar to the L-lep CTL sub-cluster, except for the expression of type I IFN genes. Although type I IFNs initially participate to activate CD8$^+$ CTL, it is unclear how low term exposure to type I IFNs affect CD8$^+$ CTL function. Noteworthy was the presence of plasma cells almost exclusively in the L-lep lesions, consistent with previous studies [96, 97]. The high levels of antibody in L-lep patients suggests that they do not protect against infection in leprosy, although it is possible that distinct antibodies produced in RR have an antimycobacterial role [98]. Although there were few LC from the L-lep lesions, we did not immuno-select LC from these biopsy specimens, nevertheless, LC are less frequent in L-lep than in RR [12, 99]. Although Tregs were present in both RR and L-lep, we were not able to identify CD8$^+$ T suppressor cells in L-lep lesions, perhaps because of the low detection of *IL4*.

One of the practical findings of this work was that in the RR lesions there is an abundance of CTL capable of producing IFN-□, expressing GZMB and PRF1, which we presume can killing infected macrophages. But that activity alone would likely have the effect of releasing viable bacilli and disseminating the infection. It is only the amCTL subset, expressing GNLY that were previously shown to be capable of killing mycobacteria within infected macrophages that would limit the infection. In human vaccine trials, there is a critical need for correlates and biomarkers of protection. The data here suggests that measurement of IFN-□

production by CD8 T cells will simply not serve as a useful measure of amCTL, which can be estimated by specific surface markers.

It has long been thought that the nature of the immune responses in infection, cancer and autoimmune diseases is dictated by the principal cells of the immune system, lymphocytes and myeloid cells. Certainly, in our study the expression of *IL1B* by DC and LC, and *IFNG* by T cell subpopulations suggest these immune cells are the first responders and the key drivers of the immune response. However, a compelling aspect of our data on leprosy is that these immune cells activate lymphocytes and myeloid cells, but also other cell types such as fibroblasts and keratinocytes, cells that are beyond the traditional immune cells in the granulomatous immune response, with capability of producing antimicrobial molecules. There is mounting evidence that the connective tissue and epithelium are key components of the overall immune response. As such, the granuloma is not limited to an organized core of macrophages with lymphocytes, but extends beyond its microanatomic limits to recruit an array of cell types to combat the foreign invader. One could summarize our key findings simply by saying that it takes a village to create effective antimicrobial granulomas.

Author contributions

Conceptualization, Visualization, Project Administration - FM, MP, RLM; Methodology, Formal Analysis, Investigation - FM, TH, RMT, PA, BJS, TD, MW, LA, BB, AS, BB, JG, MP, RLM; Writing - FM, MP, RLM; Resource - MTO, ES; Supervision: MP, RLM; Funding Acquisition - MP RLM.

Figures

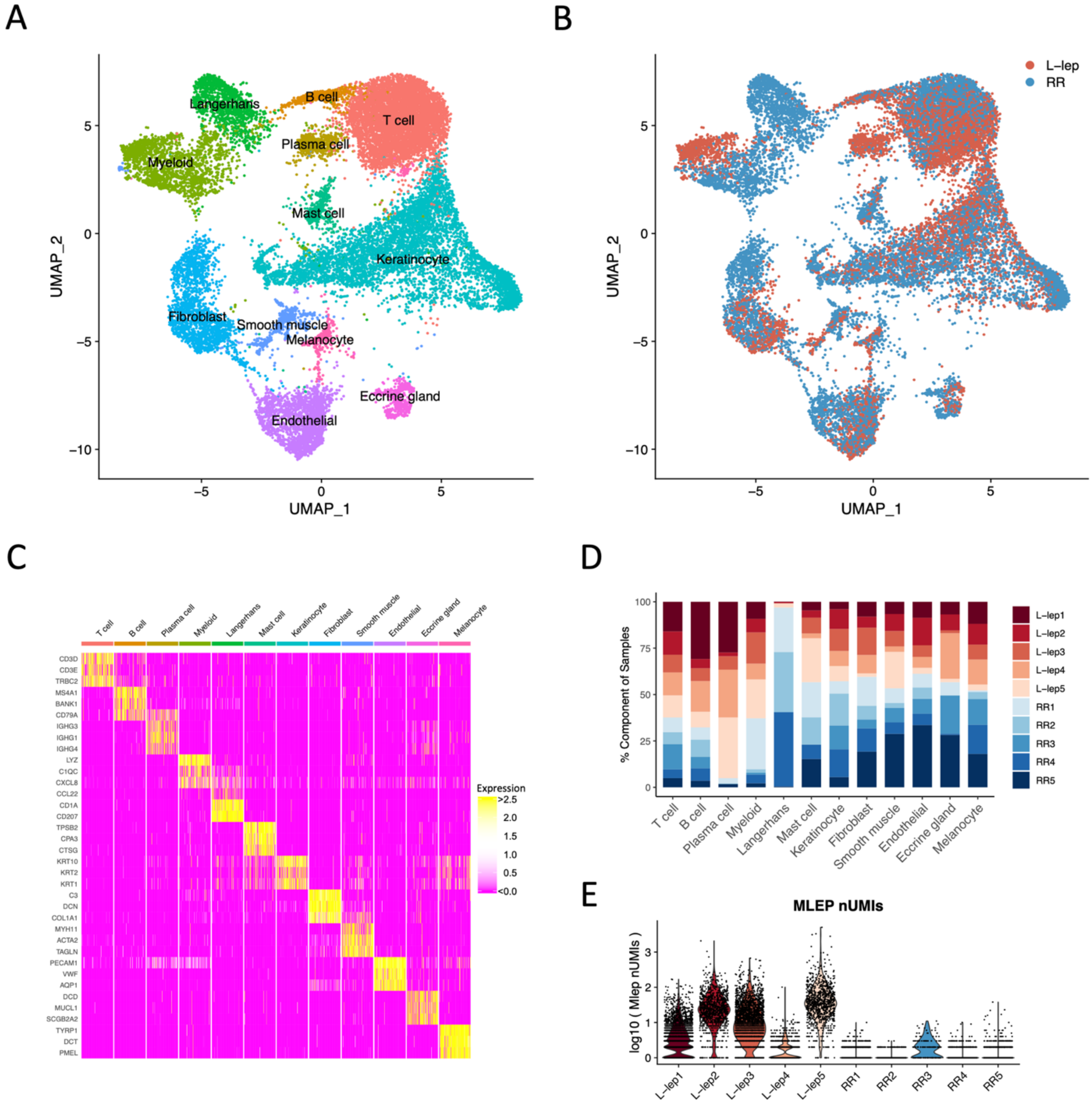

A
B
C
D
E
Langerhans
B cell
T cell
Plasma cell
Myeloid
Mast cell
Keratinocyte
Fibroblast
Smooth muscle
Melanocyte
Eccrine gland
Endothelial
UMAP_1
UMAP_2
5
0
−5
−10
L−lep
RR
T cell
B cell
Plasma cell
Myeloid
Langerhans
Mast cell
Keratinocyte
Fibroblast
Smooth muscle
Endothelial
Eccrine gland
Melanocyte
CD3D
CD3E
TRBC2
MS4A1
BANK1
CD79A
IGHG3
IGHG1
IGHG4
LYZ
C1QC
CXCL8
CCL22
CD1A
CD207
TPSB2
CPA3
CTSG
KRT10
KRT2
KRT1
C3
DCN
COL1A1
MYH11
ACTA2
TAGLN
PECAM1
VWF
AQP1
DCD
MUCL1
SCGB2A2
TYRP1
DCT
PMEL
Expression
>2.5
2.0
1.5
1.0
0.5
<0.0
% Component of Samples
100
75
50
25
0
L−lep1
L−lep2
L−lep3
L−lep4
L−lep5
RR1
RR2
RR3
RR4
RR5
MLEP nUMIs
log10 (Mlep nUMIs)
3
2
1
0
L−lep1
L−lep2
L−lep3
L−lep4
L−lep5
RR1
RR2
RR3
RR4
RR5

Figure 4-1. Cell types observed in leprosy lesions.

A. UMAP plot for 21,318 cells colored by cell types.

B. UMAP plot colored by clinical forms.

C. Heatmap showing three representative marker genes for each cell type.

D. Abundance composition across 10 patients for each cell type.

E. Violin plot showing the number of M. leprae transcripts detected in individual cells from each patient.

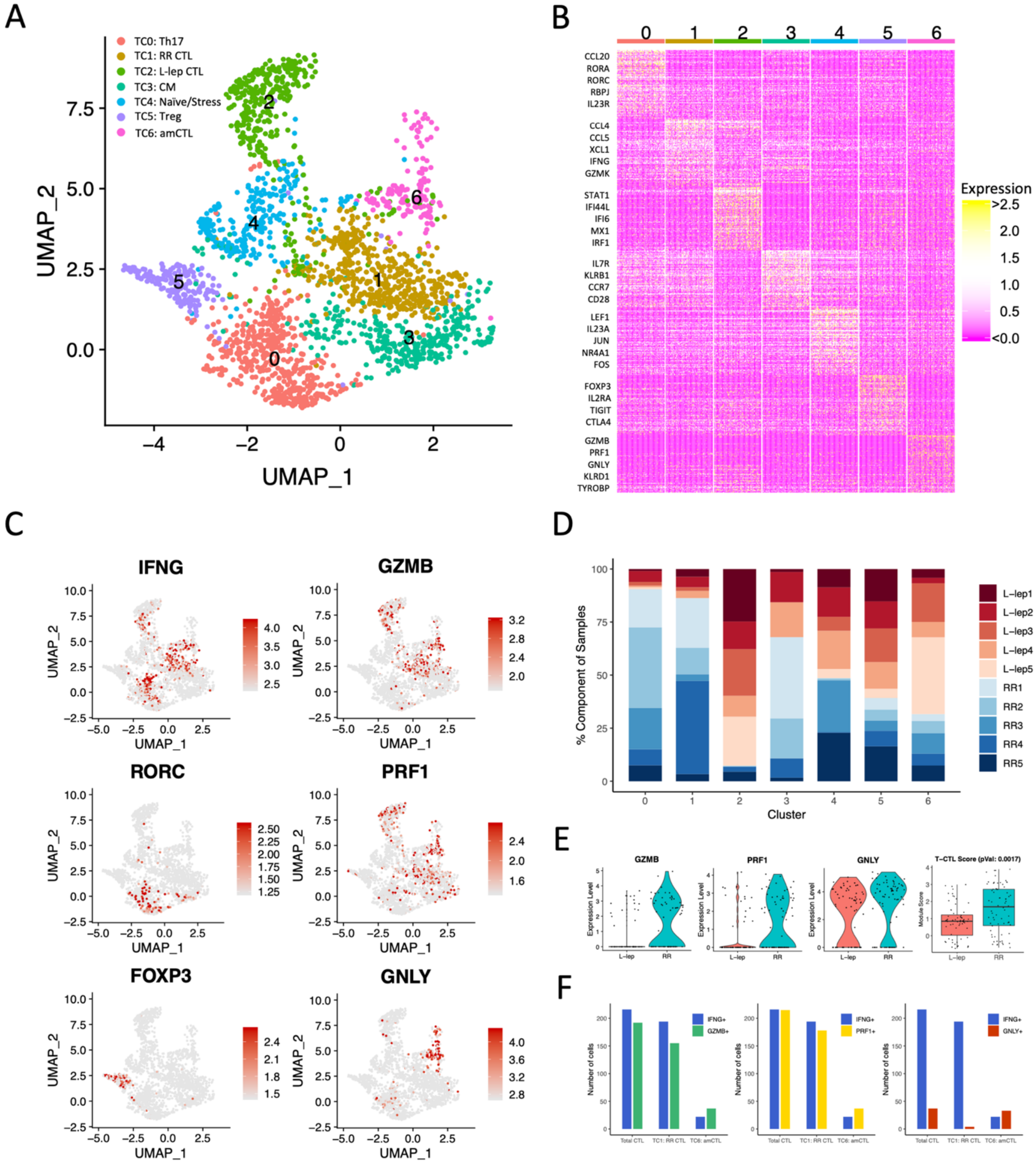

A
TC0: Th17
TC1: RR CTL
TC2: L-lep CTL
TC3: CM
TC4: Naïve/Stress
TC5: Treg
TC6: amCTL
UMAP_1
UMAP_2
B
0 1 2 3 4 5 6
CCL20
RORA
RORC
RBPJ
IL23R
CCL4
CCL5
XCL1
IFNG
GZMK
STAT1
IFI44L
IFI6
MX1
IRF1
IL7R
KLRB1
CCR7
CD28
LEF1
IL23A
JUN
NR4A1
FOS
FOXP3
IL2RA
TIGIT
CTLA4
GZMB
PRF1
GNLY
KLRD1
TYROBP
Expression
>2.5
2.0
1.5
1.0
0.5
<0.0
C
IFNG
GZMB
RORC
PRF1
FOXP3
GNLY
D
% Component of Samples
Cluster
L−lep1
L−lep2
L−lep3
L−lep4
L−lep5
RR1
RR2
RR3
RR4
RR5
E
GZMB
PRF1
GNLY
T−CTL Score (pVal: 0.0017)
Expression Level
Module Score
L−lep
RR
F
Number of cells
IFNG+
GZMB+
PRF1+
GNLY+
Total CTL
TC1: RR CTL
TC6: amCTL

Figure 4-2. Identification of T cell subtypes.

A. UMAP plot colored by T cell subtypes.

B. Heatmap showing 100 marker genes for each subtype. The representative genes are labelled.

C. UMAP plots showing six marker genes. The color scale represents normalized expression level of the gene.

D. Abundance composition across 10 patients for each T cell subtype.

E. (Left) Violin plots showing the expression for *GZMB*, *PRF1* and *GNLY* in T cell sub-cluster 6 grouped by L-lep and RR. (Right) Boxplot showing the T-CTL score in T cell sub-cluster 6 grouped by L-lep and RR, the p value was calculated from a Wilcoxon rank sum test.

F. Number of RR cells expressing *IFNG*, *GMZB*, *PRF1* and *GNLY* in RR CTL and amCTL.

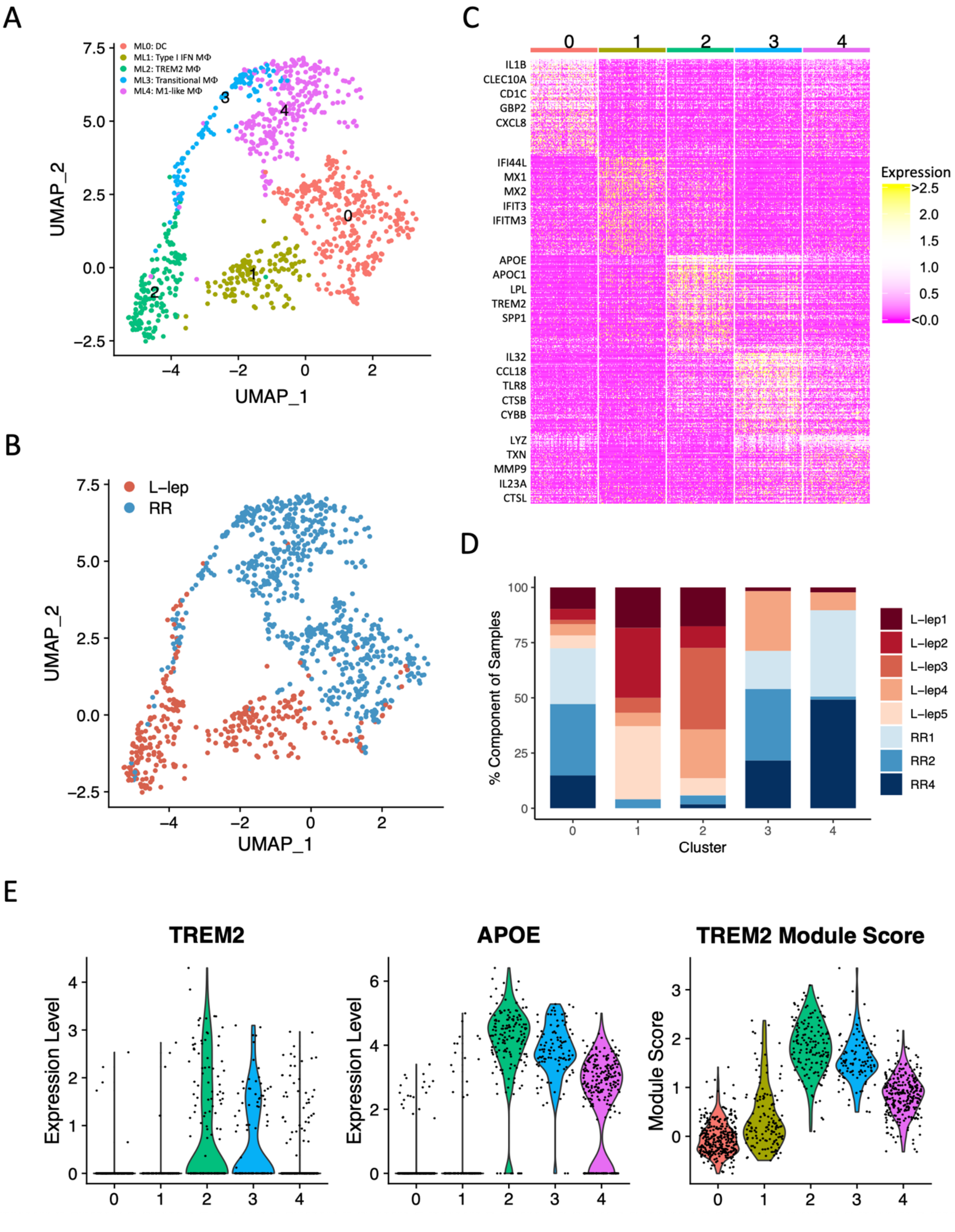

A
ML0: DC
ML1: Type I IFN MΦ
ML2: TREM2 MΦ
ML3: Transitional MΦ
ML4: M1-like MΦ
UMAP_2
UMAP_1
7.5
5.0
2.5
0.0
−2.5
−4
−2
0
2
0
1
2
3
4
B
L−lep
RR
UMAP_2
UMAP_1
C
0
1
2
3
4
IL1B
CLEC10A
CD1C
GBP2
CXCL8
IFI44L
MX1
MX2
IFIT3
IFITM3
APOE
APOC1
LPL
TREM2
SPP1
IL32
CCL18
TLR8
CTSB
CYBB
LYZ
TXN
MMP9
IL23A
CTSL
Expression
>2.5
2.0
1.5
1.0
0.5
<0.0
D
% Component of Samples
100
75
50
25
0
Cluster
L−lep1
L−lep2
L−lep3
L−lep4
L−lep5
RR1
RR2
RR4
E
TREM2
Expression Level
APOE
Expression Level
TREM2 Module Score
Module Score

Figure 4-3. Identification of myeloid cell subtypes.

A. UMAP plot colored by myeloid cell subtypes.

B. UMAP plot colored by clinical forms.

C. Heatmap showing 100 marker genes for each subtype. The representative genes are labelled.

D. Abundance composition across 10 patients for each myeloid cell subtype.

E. (Left) Violin plots showing the expression for *TREM2* and *APOE* in myeloid subtypes.

(Right) Violin plot showing the TREM2 Module score in myeloid subtypes.

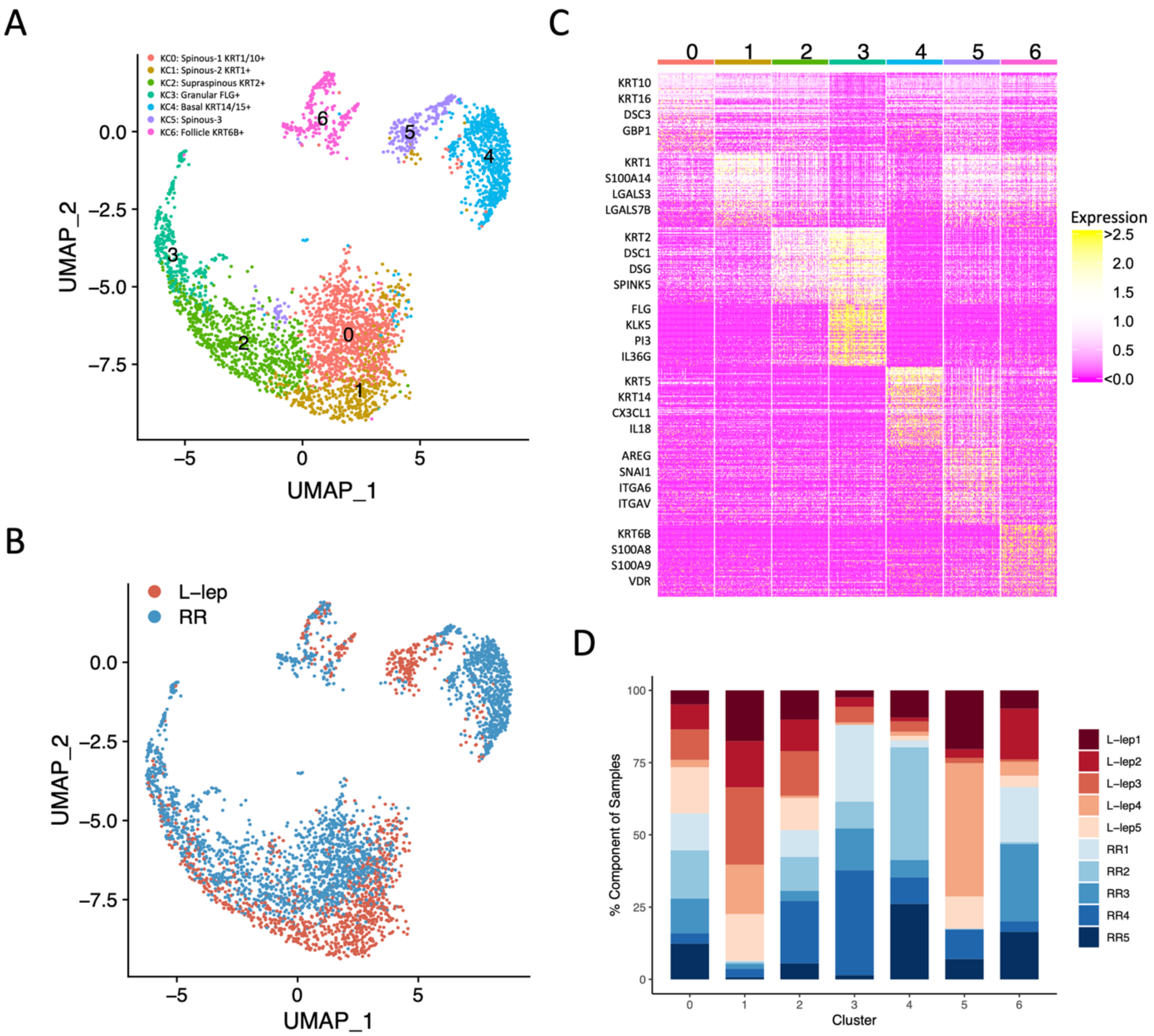


Figure 4-4. Identification of keratinocyte subtypes.

A. UMAP plot colored by keratinocyte subtypes.

B. UMAP plot colored by clinical forms.

C. Heatmap showing 100 marker genes for each subtype. The representative genes are labelled.

D. Abundance composition across 10 patients for each keratinocyte subtype.

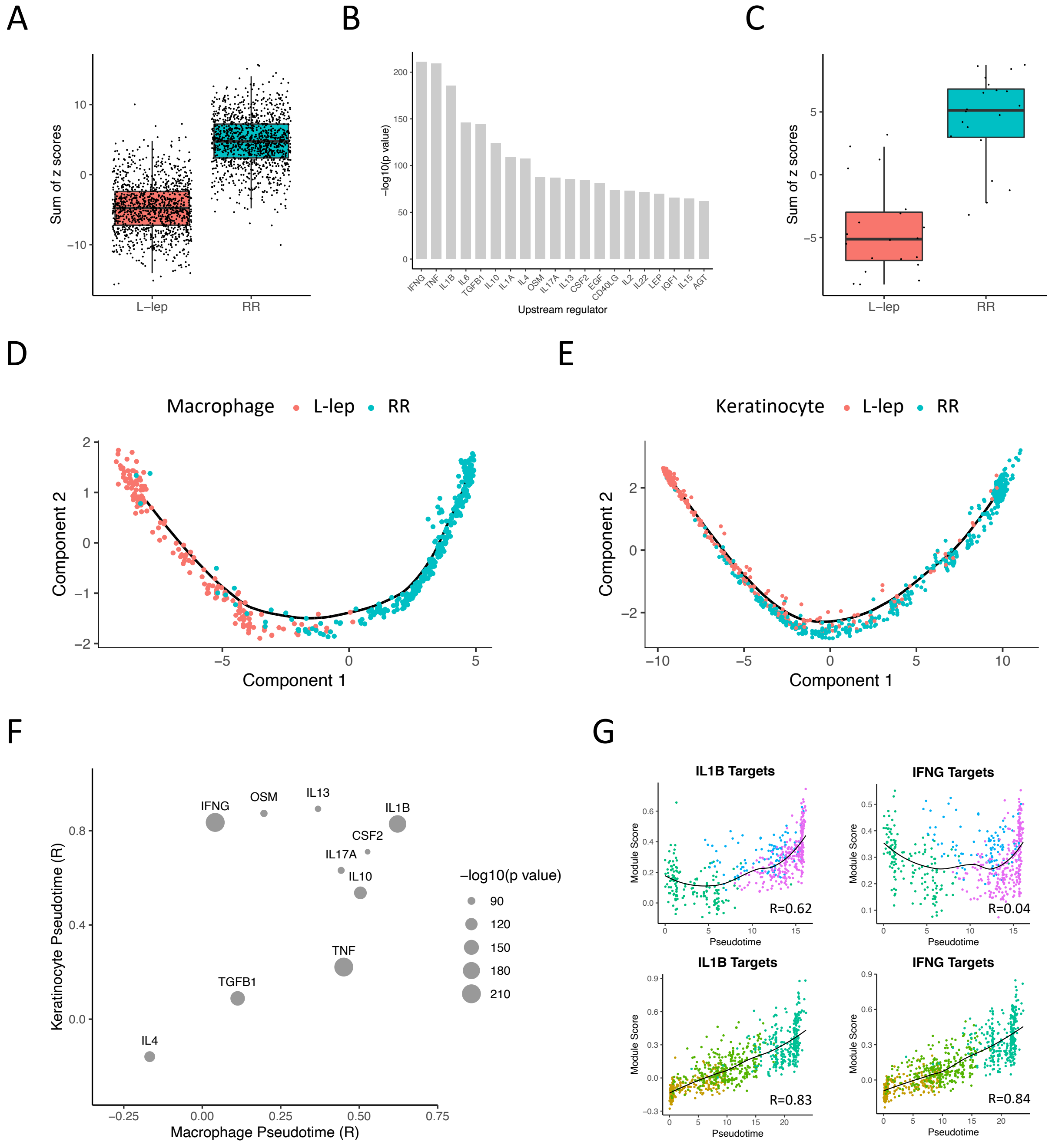

A
B
C
Sum of z scores
L–lep
RR
−log10(p value)
Upstream regulator
IFNG
TNF
IL1B
IL6
TGFB1
IL10
IL1A
IL4
OSM
IL17A
IL13
CSF2
EGF
CD40LG
IL2
IL22
LEP
IGF1
IL15
AGT
D
E
Macrophage ● L-lep ● RR
Keratinocyte ● L-lep ● RR
Component 1
Component 2
F
G
Keratinocyte Pseudotime (R)
Macrophage Pseudotime (R)
−log10(p value)
90
120
150
180
210
IFNG
OSM
IL13
IL1B
CSF2
IL17A
IL10
TNF
TGFB1
IL4
IL1B Targets
IFNG Targets
Module Score
Pseudotime
R=0.62
R=0.04
R=0.83
R=0.84

Figure 4-5. Antimicrobial gene analysis and pseudotime construction.

A. Boxplot showing the sum of 1,124 AMG z scores in L-lep and RR cell types.

B. Bar graph showing the top 20 upstream regulators ranked by p value from the enrichment analysis using the 1,124 AMGs.

C. Boxplot showing the sum of the top 20 z scores in L-lep and RR cell types.

D. Pseudotime trajectory colored by clinical form in myeloid sub-cluster 2, 3 and 4.

E. Pseudotime trajectory colored by clinical form in keratinocyte sub-cluster 1, 2 and 3.

F. Dot plot showing the correlation between the top 10 URs' module scores and macrophage/keratinocyte pseudotimes. The size of the dots represents the -log10(p value) from the enrichment analysis.

G. Scatter plot showing the correlation between macrophage (top) or keratinocyte (bottom) pseudotimes and module scores calculated using *IL1B* target genes or *IFNG* target genes from the six expression patterns. Color of the dots represents the sub-cluster identity of the cells.

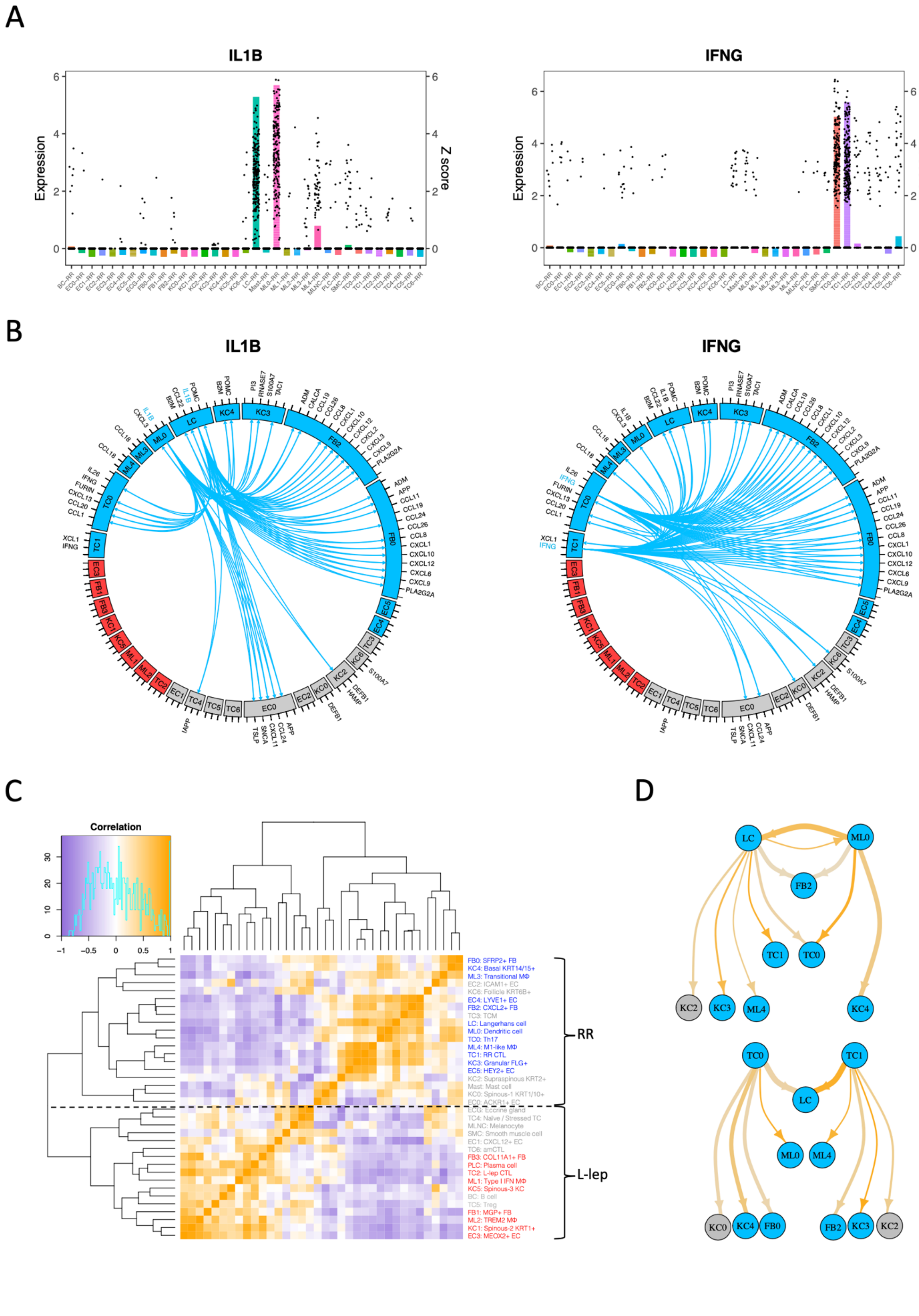

A
IL1B
IFNG
Expression
Z score
0
2
4
6
BC-RR
EC0-RR
EC1-RR
EC2-RR
EC3-RR
EC4-RR
EC5-RR
ECG-RR
FB0-RR
FB1-RR
FB2-RR
KC0-RR
KC1-RR
KC2-RR
KC3-RR
KC4-RR
KC5-RR
KC6-RR
LC-RR
Mast-RR
ML0-RR
ML1-RR
ML2-RR
ML3-RR
ML4-RR
MLNC-RR
PLC-RR
SMC-RR
TC0-RR
TC1-RR
TC2-RR
TC3-RR
TC4-RR
TC5-RR
TC6-RR
B
IL1B
IFNG
TC1
TC0
ML4
ML3
ML0
LC
KC4
KC3
FB2
FB0
EC5
EC4
TC3
KC6
KC2
KC0
EC2
EC0
TC6
TC5
TC4
EC1
TC2
ML2
ML1
KC5
KC1
FB3
FB1
EC3
XCL1
IFNG
CCL1
CCL20
CXCL13
FURIN
IFNG
IL26
CCL18
CCL18
CXCL3
IL1B
B2M
CCL22
IL1B
POMC
B2M
POMC
PI3
RNASE7
S100A7
TAC1
ADM
CALCA
CCL19
CCL26
CCL8
CXCL1
CXCL10
CXCL12
CXCL2
CXCL3
CXCL9
PLA2G2A
ADM
APP
CCL11
CCL19
CCL24
CCL26
CCL8
CXCL1
CXCL10
CXCL12
CXCL6
CXCL9
PLA2G2A
S100A7
DEFB1
HAMP
DEFB1
APP
CCL24
CXCL11
SNCA
TSLP
IAPP
C
Correlation
0
10
20
30
-1
-0.5
0
0.5
1
FB0: SFRP2+ FB
KC4: Basal KRT14/15+
ML3: Transitional MΦ
EC2: ICAM1+ EC
KC6: Follicle KRT6B+
EC4: LYVE1+ EC
FB2: CXCL2+ FB
TC3: TCM
LC: Langerhans cell
ML0: Dendritic cell
TC0: Th17
ML4: M1-like MΦ
TC1: RR CTL
KC3: Granular FLG+
EC5: HEY2+ EC
KC2: Supraspinous KRT2+
Mast: Mast cell
KC0: Spinous-1 KRT1/10+
EC0: ACKR1+ EC
ECG: Eccrine gland
TC4: Naive / Stressed TC
MLNC: Melanocyte
SMC: Smooth muscle cell
EC1: CXCL12+ EC
TC6: amCTL
FB3: COL11A1+ FB
PLC: Plasma cell
TC2: L-lep CTL
ML1: Type I IFN MΦ
KC5: Spinous-3 KC
BC: B cell
TC5: Treg
FB1: MGP+ FB
ML2: TREM2 MΦ
KC1: Spinous-2 KRT1+
EC3: MEOX2+ EC
RR
L-lep
D
LC
ML0
FB2
TC1
TC0
KC2
KC3
ML4
KC4
TC0
TC1
LC
ML0
ML4
KC0
KC4
FB0
FB2
KC3
KC2

Figure 4-6. Antimicrobial network and cell-cell co-abundance.

A. Bar plot showing the z scores of *IL1B* (left) or *IFNG* (right) expression levels in each cell type from RR lesions. The dots represent *IL1B* or *IFNG* expression in individual cells.

B. Circos plot showing the connection between *IL1B* (left) or *IFNG* (right) and the direct antimicrobial gene targets in the cell types with z score > 3.

C. Heatmap showing the cell type correlations calculated based on the co-abundance composition across 10 patients. Cell types in red have > 70% of the cells from L-lep lesions, cell types in blue have > 70% from RR lesions, the other cell type names are colored in grey.

D. Network depicting the antimicrobial connections induced by *IL1B* (Top) and *IFNG* (Bottom). Color scale of the links represent the co-abundance correlation between the cell types. Width of the links represent number of antimicrobial links.

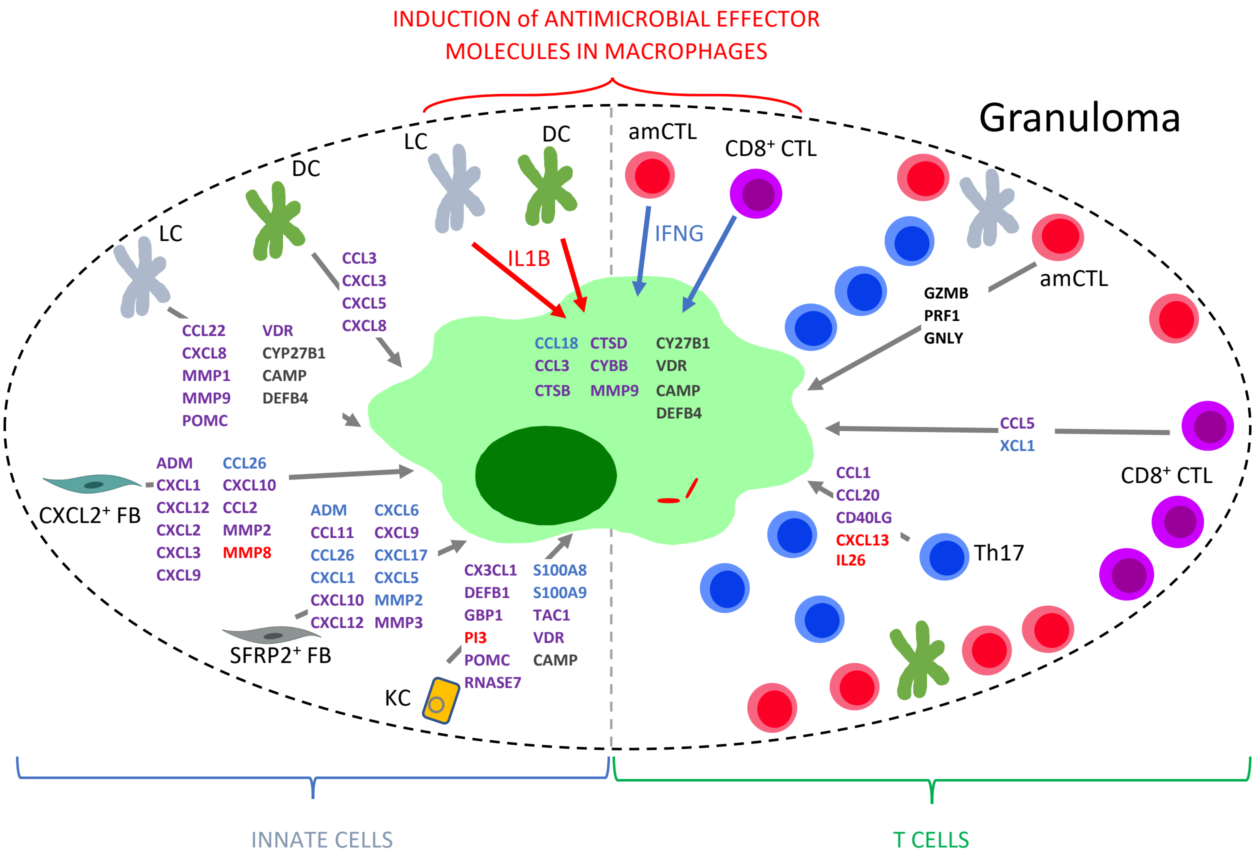
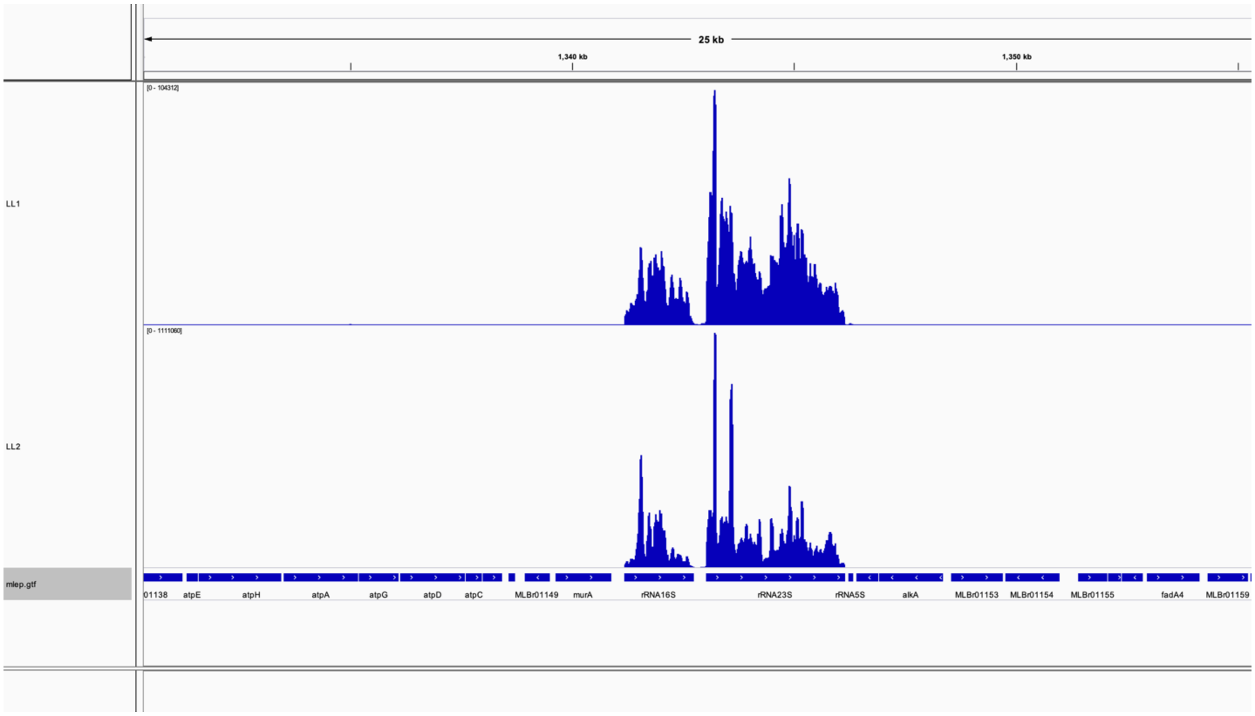


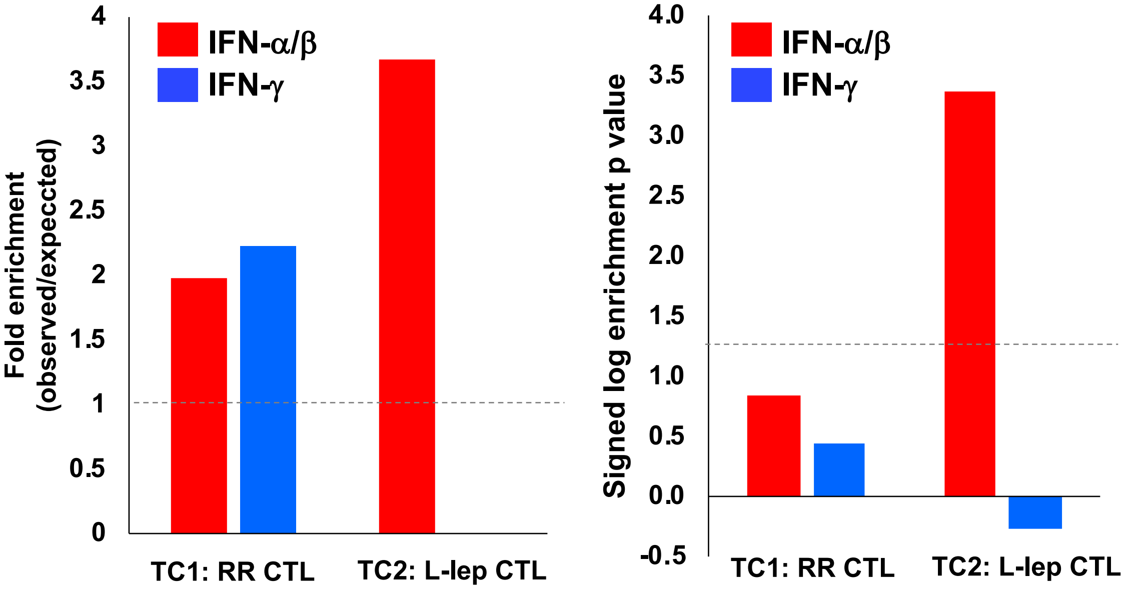
Figure 4-7. Antimicrobial ecosystem in leprosy granuloma.

Antimicrobial diagram in granuloma from leprosy lesions. Gene names in red represent targets of *IL1B*. Gene names in blue represent targets of *IFNG*. Gene names in purple represents targets of both *IL1B* and *IFNG*.

Supplemental information

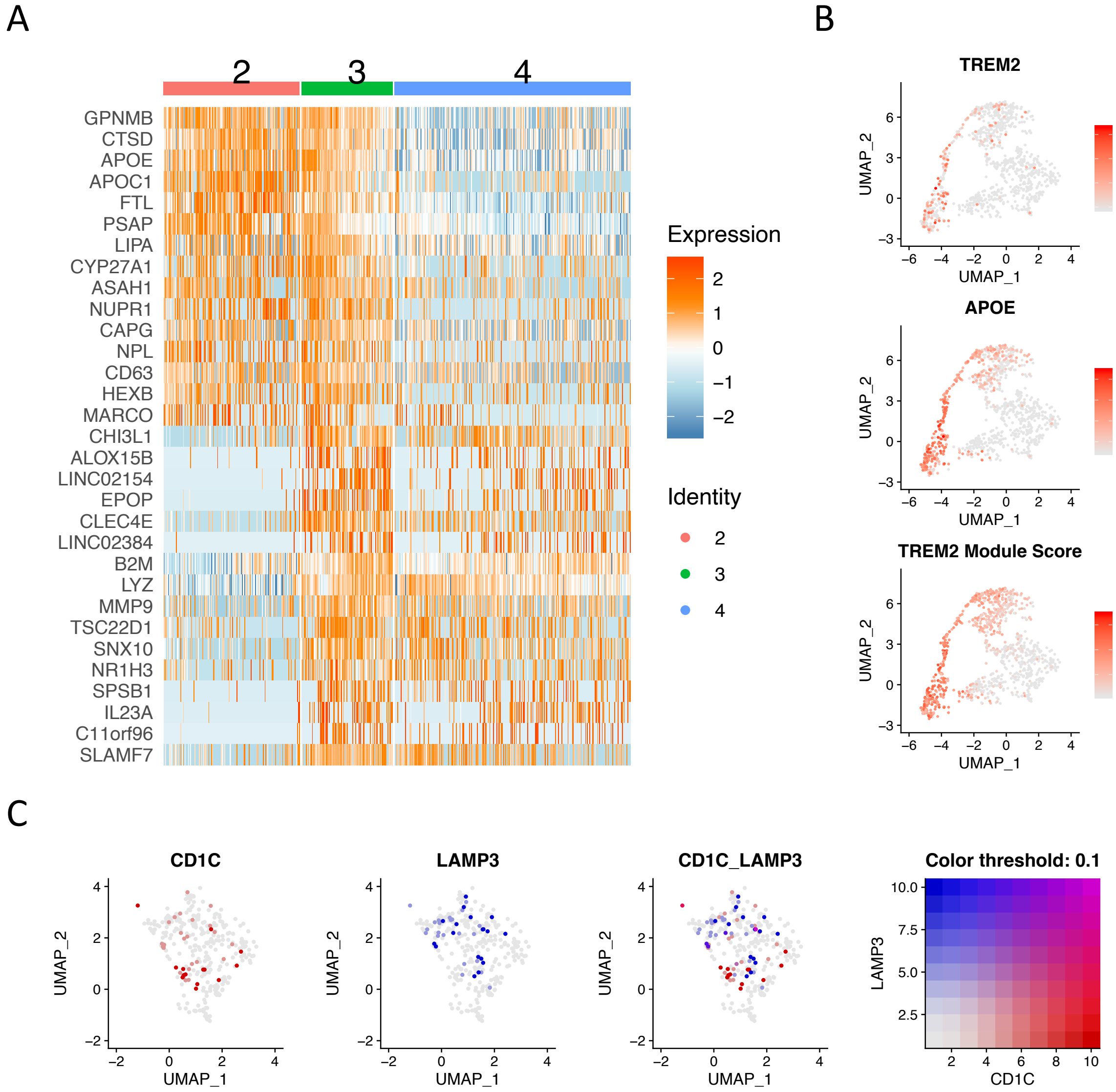


Supplemental Figure 4-1. ***M. leprae*** **reads were mainly detected in the rRNA region.**



Supplemental Figure 4-2: IFN-α/β and IFN-γ signature on CTL subtypes.

Enrichment analysis on differentially expressed genes (adjusted p value < 0.05) between TC1 (RR CTL) and TC2 (L-lep CTL) using IFN-α/β and IFN-γ specific genes identified in human PBMC. Dotted lines indicate (left) no enrichment or (right) the hypergeometric test p value of 0.05 (log p value = 1.3).

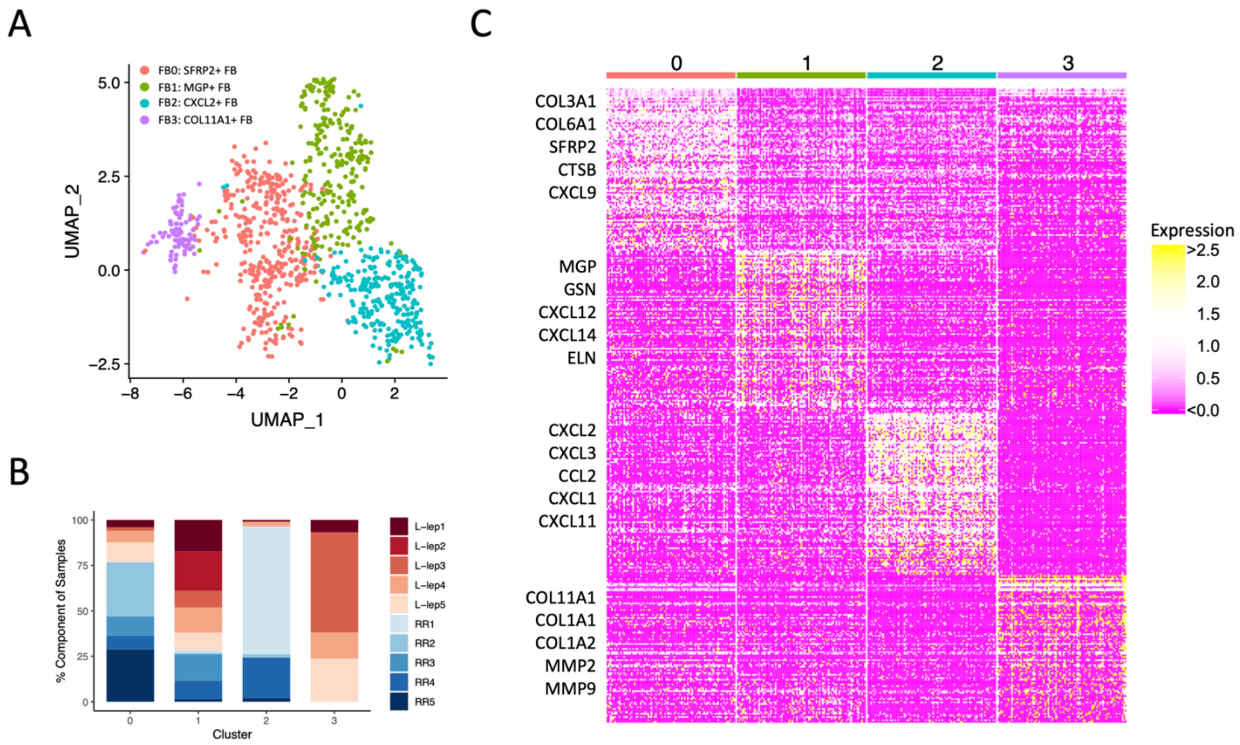


Supplemental Figure 4-3. Macrophage transition.

A. Heatmap showing top differentially expressed genes between ML2 and ML4. ML3 expressed both ML2 and ML4 specific genes.

B. UMAP plots showing *TREM2* expression, *APOE* expression and TREM2 module score. The color scale represents the expression level of the genes.

C. UMAP plots showing *CD1C* and *LAMP3* expression in ML0. Only few co-expression events were observed.

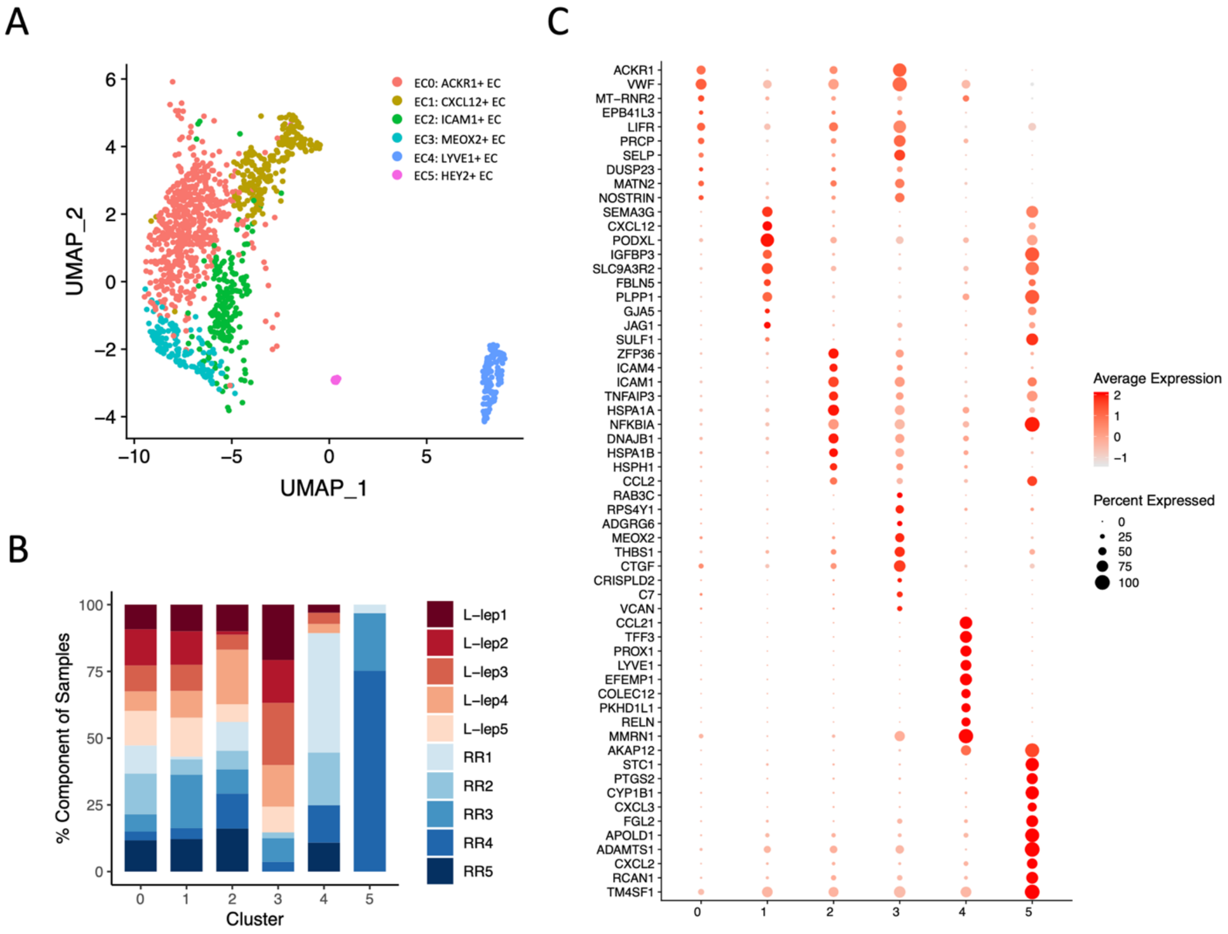


Supplemental Figure 4-4. Identification of fibroblast subtypes.

A. UMAP plot colored by fibroblast subtypes.

B. Abundance composition across 10 patients for each fibroblast subtype.

C. Heatmap showing 100 marker genes for each subtype. The representative genes are labelled.

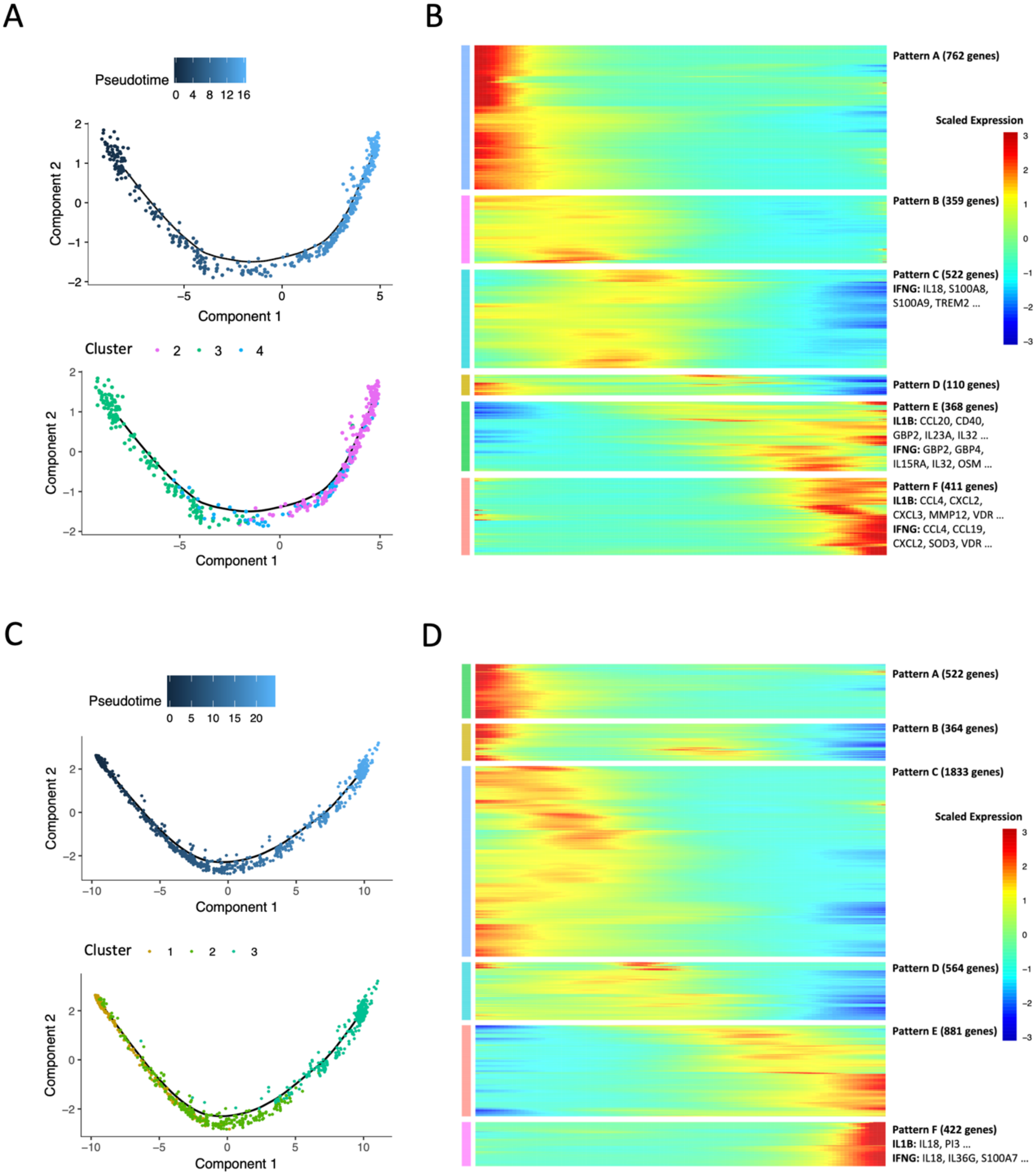


Supplemental Figure 4-5. Identification of endothelial cell subtypes.

A. UMAP plot colored by endothelial cell subtypes.

B. Abundance composition across 10 patients for each endothelial cell subtype.

C. Dot plot showing 10 marker genes for each subtype. The color scale represents the scaled expression of the gene. The size of the dot represents percentage of cells expressing the gene.


A
Pseudotime
0 4 8 12 16
Component 2
Component 1
Cluster 2 3 4
B
Pattern A (762 genes)
Pattern B (359 genes)
Pattern C (522 genes)
IFNG: IL18, S100A8, S100A9, TREM2 ...
Pattern D (110 genes)
Pattern E (368 genes)
IL1B: CCL20, CD40, GBP2, IL23A, IL32 ...
IFNG: GBP2, GBP4, IL15RA, IL32, OSM ...
Pattern F (411 genes)
IL1B: CCL4, CXCL2, CXCL3, MMP12, VDR ...
IFNG: CCL4, CCL19, CXCL2, SOD3, VDR ...
Scaled Expression
3 2 1 0 −1 −2 −3
C
Pseudotime
0 5 10 15 20
Component 2
Component 1
Cluster 1 2 3
D
Pattern A (522 genes)
Pattern B (364 genes)
Pattern C (1833 genes)
Pattern D (564 genes)
Pattern E (881 genes)
Pattern F (422 genes)
IL1B: IL18, PI3 ...
IFNG: IL18, IL36G, S100A7 ...
Scaled Expression
3 2 1 0 −1 −2 −3

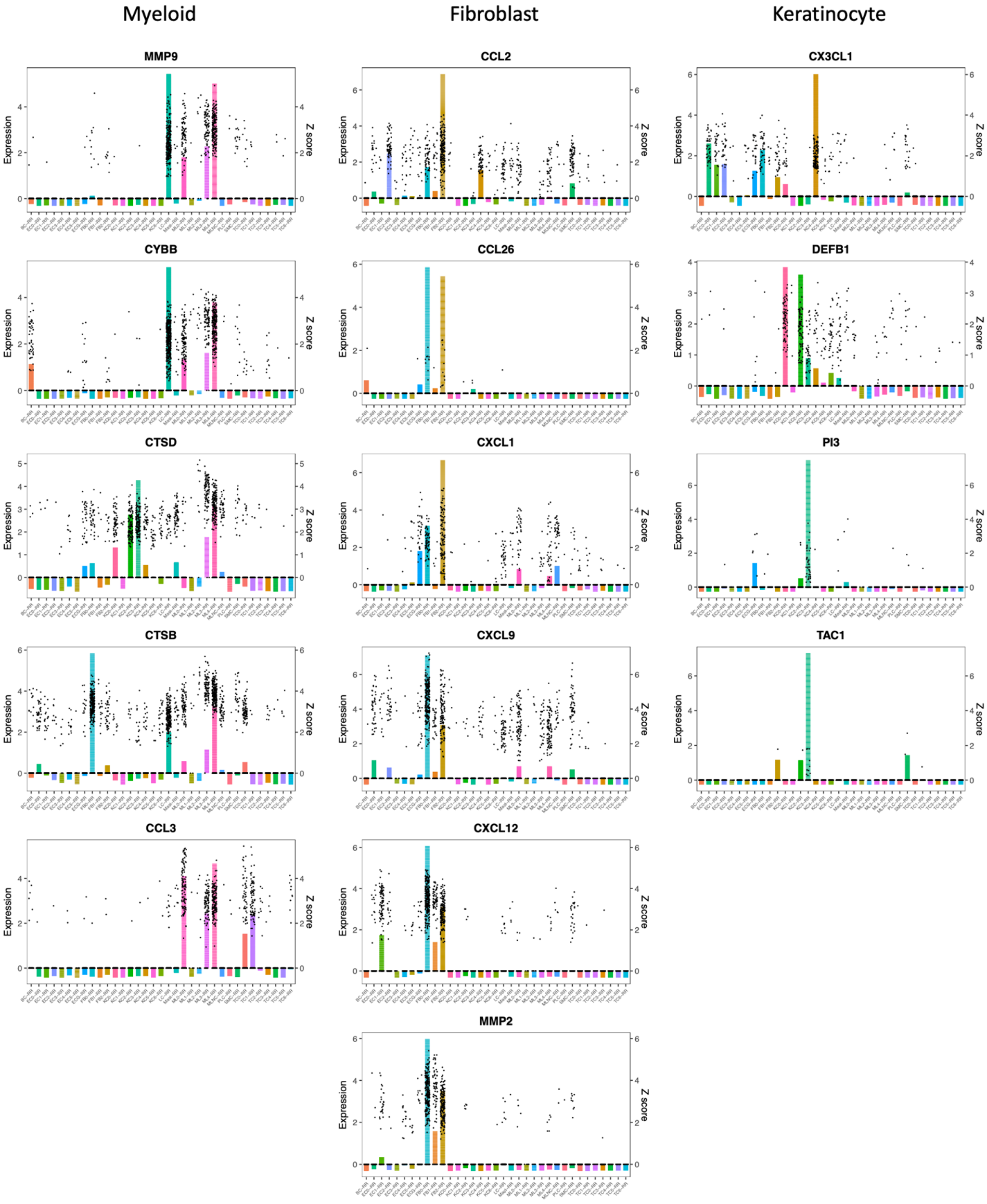
Supplemental Figure 4-6. Pseudotime construction in macrophages and keratinocytes.

A. Pseudo-temporal trajectory colored by pseudotime (top) and by sub-cluster identity (bottom) for macrophage sub-cluster 2, 3 and 4.

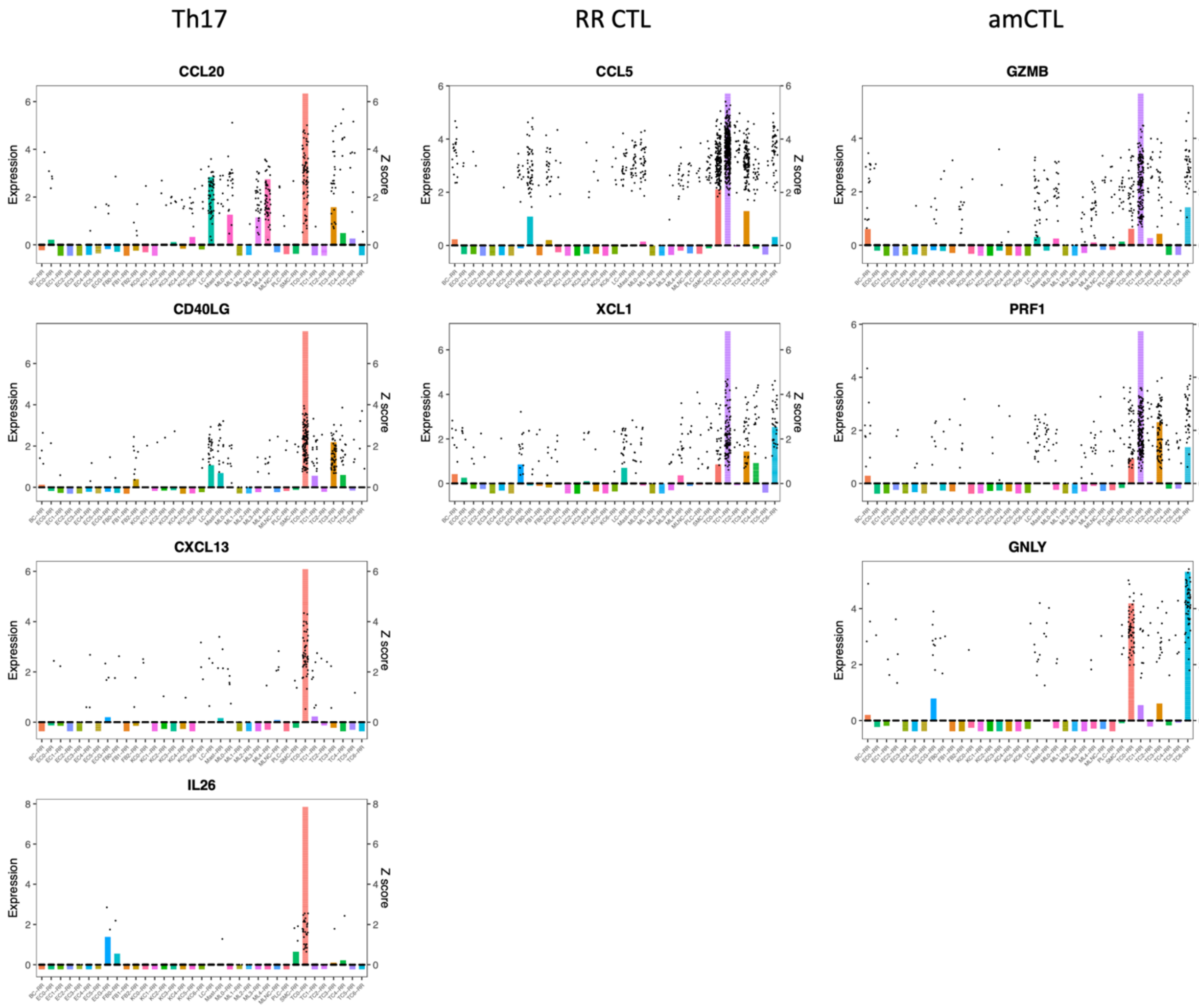
B. Heatmap showing six expression patterns along the macrophage pseudotime. Representative genes regulated by *IL1B* and *IFNG* are labelled.

C. Pseudo-temporal trajectory colored by pseudotime (top) and by sub-cluster identity (bottom) for keratinocyte sub-cluster 1, 2 and 3.

D. Heatmap showing six expression patterns along the keratinocyte pseudotime. Representative genes regulated by *IL1B* and *IFNG* are labelled.



Supplemental Figure 4-7. Representative antimicrobial genes expressed by myeloid cells, fibroblasts and keratinocytes.



Supplemental Figure 4-8. Representative antimicrobial genes expressed by Th17 cells, RR CTL and amCTL.

Chapter 5 – Conclusions

With the development and advancement of RNA-sequencing technology, many library preparation methods and sequencing platforms have become available. In chapter 2, we used a classic whole transcript RNA-Seq method (Trad-KAPA) and a 3' RNA-Seq method (3'-LEXO) to prepare sequencing libraries from livers of iron-loaded diet and control diet mice, and sequenced the libraries on the Illumina platform [100]. We then compared the sequencing results to determine the advantages and disadvantages of the two approaches.

We identified the gene body coverage of the Trad-KAPA and 3'-LEXO libraries by mapping the reads back to the genome. As expected, Trad-KAPA reads covered transcripts uniformly, with a slight decrease at the 5' end. One reason for the decrease might be that the secondary structure of the mRNA can cause early termination of reverse transcription [39], making it difficult to reach the cap site (5' end). It is also possible that many of the transcripts are partially degraded, so that the polyadenylation capture biases the coverage towards the 3' end. By contrast, 3'-LEXO reads mapped mostly to the 3' end. 3'-LEXO reads that mapped to the middle of the transcript showed significant coverage variation from library to library. The variation might be caused by the randomness in the reverse transcription start site on the cDNA. In the classic whole transcript method, mRNAs are first sheared into fragments, then the fragments are reverse transcribed to generate cDNAs. Hence, it is expected that the longer a transcript is, the more fragments it should have. The 3' RNA-Seq method however generates only one read for each transcript, so the number of reads directly reflects the level of gene expression. We counted the reads mapped to transcripts that have lengths ranging from 500 bp to 8500 bp and found that Trad-KAPA libraries had more reads assigned to longer transcripts. By contrast, 3'-LEXO read counts remained uniform as transcript length increased.

As Trad-KAPA assigned more reads to longer transcripts and 3'-LEXO assigned a similar number of reads to transcripts with different lengths, we expected to see fewer short transcripts and more long transcripts detected by Trad-KAPA as sequencing depth drops. For transcripts shorter than 1000 bp, 3'-LEXO detected about 10% more than Trad-KAPA when sequencing depth dropped. However, for transcripts longer than 1000 bp, there was only a small difference between the number detected by Trad-KAPA and 3'-LEXO. Since a 3' RNA-Seq method only captures reads from the 3' end of the mRNA, it is difficult for this method to detect differences in isoforms close to the 5' end of longer genes. In our study, 15% of uniquely mapped Trad-KAPA reads contain splices, while only 6% of uniquely mapped 3'-LEXO reads contain splices. As a result, the 3' RNA-Seq method is not recommended for novel transcript or splice variant discovery. We also compared Trad-KAPA and 3'-LEXO reproducibility, and found that both methods showed very high reproducibility between biological replicates. When comparing the sequencing results generated with the same mouse using the Trad-KAPA versus 3'-LEXO methods, we found the two methods generally agreed with each other. Although there were a few transcripts detected only by Trad-KAPA, they turned out to be non-coding RNAs.

Among all the DEGs we found, some of the very short transcripts (shorter than 500 bp) were only detected to be differentially expressed by 3'-LEXO, while many of the long transcripts, especially those longer than 7500 bp, were only detected as differentially expressed by Trad-KAPA. As Trad-KAPA assigns more reads to longer transcripts, the statistical power to detect differences increases. Thus, the probability that those transcripts are detected differentially expressed is higher. It is also clear that as sequencing depth drops, both methods will detect fewer differentially expressed transcripts. Thus, if users want to use RNA-Seq to detect differentially expressed transcripts, Trad-KAPA will likely generate larger lists than 3’-LEXO, biased towards longer transcripts.

scRNA-seq provides high resolution profiling of the transcriptomes of single cells. Typically, the first step in scRNA-seq analysis is to assign each cell a cell type based on our prior knowledge of marker genes. Current methods for cell type assignment first cluster the cells in an unsupervised manner and rely on the canonical markers to identify the cell types for each cluster. However, this approach has several limitations, including the fact that the clusters may not optimally segregate single cell types, and certain cell types may not have previously characterized markers. Moreover, these methods are computationally intensive, especially when the number of cells becomes large. To render cell type identification in scRNA-seq more efficient, we employed a neural network, trained it on cells with predefined cell types, and used it to predict cell types for new datasets.

In chapter 3, we first obtained and cleaned two datasets from the Tabula Muris Consortium, then trained and tested our neural network on these datasets with or without batch effect introduced by different scRNA-seq platforms [101]. The training accuracy always approached 100%, and the testing accuracy was around 99.8% within a platform and 99.0% when testing and training are performed across different platforms. As the cell types in the two Tabula muris atlas datasets can be mutually predicted using our neural network, we merged them and used the combined datasets as the reference to predict cell types for other datasets. The predicted cell types were well matched with the cell types assigned using the canonical markers for both the mouse and human datasets. We also trained and tested the neural network on five T cell subtypes and found that the predicted subtypes showed the same markers as the reference subtypes, which suggests that our neural network can be used to predict sub cell types as well.

Compared to the traditional unsupervised methods used for cell type identification, our neural network has the following advantages: 1. It uses all the genes to capture the features for each cell type instead of relying on a limited number of canonical markers. 2. It focuses the analysis on the signal associated with the variance between cell types, while unsupervised clustering tends to be affected by other sources of cell type independent variation (i.e. platform or cell cycle). 3. It requires no background knowledge of cell type markers, while the unsupervised method requires users to have prior knowledge of canonical markers for each cell type in their data. 4. It is much more computationally efficient than the traditional approach. Moreover, users can subsample the reference cells to make the computation of the neural network less compute intensive and more memory efficient. We also compared ACTINN to three other cell type prediction tools, and the results showed that ACTINN performs better in finding small changes between subtypes.

There are some aspects of our approach that could be improved in the future. As the neural network is supervised, the quantity and quality of the reference data are critical. We anticipate that with time more cell types from larger atlases should be used to train a more comprehensive neural network. Also, better pairing of reference and test sets will undoubtedly improve performance. For example, the soon to be developed human cell atlas should be used to predict human cell types instead of the mouse cell atlas. Nonetheless, we showed that even with the current reference data our neural network is computationally efficient and accurate, and should improve cell type identification pipelines.

In chapter 4, we performed single cell transcriptomics on cell types comprising the granulomatous response in leprosy skin lesions. Of 43,363 genes in 21,282 cells studied, we detected 1,124 AMGs that were differentially expressed in RR lesion-derived cells across all cell types. Analysis by scRNA-seq revealed that the immune response diverges across the spectrum of leprosy not only for distinct populations of immune cells, including subpopulations of myeloid cells and T cells, but also for subpopulations of fibroblasts, endothelial cells and keratinocytes. The expression of these AMGs as well as the upstream regulators *IL1B* and *IFNG* for which these AMGs serve as targets was significantly higher in RR compared to L-lep lesions. From this data, we formulate a cellular ecosystem by integrating cell-cell co-abundance in lesions with the links between cells expressing the upstream regulators *IL1B* and *IFNG* to RR cell types expressing the downstream AMG targets. Key antimicrobial subpopulations associated with immunity in RR included cells of the myeloid and lymphocyte lineages including LC, DC, M1-like MΦs, Th17 cells, $CD8^+$ CTL and amCTL. Strikingly, the antimicrobial responses included two distinct subpopulations of fibroblasts, *SFRP2*$^+$ FB and *CXCL2*$^+$ FB as well as various KC subpopulations.

We focused on two major upstream regulators, *IL1B* and *IFNG*, that have a substantial effect on the immune response in RR lesions, recognizing that additional upstream regulators contribute to the antimicrobial response. *IFNG* was most highly correlated with the AMGs associated with RR, consistent with previous findings demonstrating an upregulation of IFN-γ in RR concomitant with a change from a Th2 to a Th1 response in paired samples in the same individuals from before and during RR [17, 19, 20, 93]. Of the top upstream regulators of the AMGs, only *IL1B* regulated the pseudotime trajectory of both macrophages and keratinocytes as patients transitioned from L-lep to RR. The macrophage psuedotime trajectory maps from the TREM2 macrophages in L-lep lesions, to the transitional macrophages in two L-lep patients and three RR patients, to the M1-like macrophages in RR lesions. Previously, we found that the TLR2 ligand, lipopeptide, in combination with IFN-γ, triggered macrophage plasticity in a similar trajectory, with macrophages reversing from M2-like to M1-like in vitro as observed in RR lesions [16, 94]. Since TLR2 and IL-1β both signal via MyD88, and IFN-γ is upregulated during RR, the key signals are present to facilitate the plasticity of macrophage differentiation to the M1 state known to have high antimicrobial activity. Similarly, the pseudotime mapped keratinocyte maturation, ending with a gene pattern indicates activation by both IL-1β and IFN-γ, with expression of *IL18*, which encodes for a protein that further upregulates IFN-γ in mycobacterial infection [95].

It has long been thought that the nature of the immune responses in infection, cancer and autoimmune diseases is dictated by the principal cells of the immune system, lymphocytes and myeloid cells. Certainly, in our study the expression of *IL1B* by DC and LC, and *IFNG* by T cell subpopulations suggest these immune cells are the first responders and the key drivers of the immune response. However, a compelling aspect of our data on leprosy is that these immune cells activate lymphocytes and myeloid cells, but also other cell types such as fibroblasts and keratinocytes, cells that are beyond the traditional immune cells in the granulomatous immune response, with capability of producing antimicrobial molecules. There is mounting evidence that the connective tissue and epithelium are key components of the overall immune response. As such, the granuloma is not limited to an organized core of macrophages with lymphocytes, but extends beyond its microanatomic limits to recruit an array of cell types to combat the foreign invader. One could summarize our key findings simply by saying that it takes a village to create effective antimicrobial granulomas.

References

1. Han, Y., et al., *Advanced Applications of RNA Sequencing and Challenges.* Bioinform Biol Insights, 2015. **9**(Suppl 1): p. 29-46.

2. Wang, Z., M. Gerstein, and M. Snyder, *RNA-Seq: a revolutionary tool for transcriptomics.* Nat Rev Genet, 2009. **10**(1): p. 57-63.

3. Meyer, E., G.V. Aglyamova, and M.V. Matz, *Profiling gene expression responses of coral larvae (Acropora millepora) to elevated temperature and settlement inducers using a novel RNA-Seq procedure.* Mol Ecol, 2011. **20**(17): p. 3599-616.

4. Moll, P., et al., *QuantSeq 3' mRNA sequencing for RNA quantification.* Nature Methods, 2014. **11**(12): p. i-iii.

5. Hwang, B., J.H. Lee, and D. Bang, *Single-cell RNA sequencing technologies and bioinformatics pipelines.* Experimental & Molecular Medicine, 2018. **50**(8): p. 96.

6. Lin, C., et al., *Using neural networks for reducing the dimensions of single-cell RNA-Seq data.* Nucleic Acids Res, 2017. **45**(17): p. e156.

7. Shaham, U., et al., *Removal of batch effects using distribution-matching residual networks.* Bioinformatics, 2017. **33**(16): p. 2539-2546.

8. Lopez, R., et al., *Deep generative modeling for single-cell transcriptomics.* Nat Methods, 2018. **15**(12): p. 1053-1058.

9. Cho, H., B. Berger, and J. Peng, *Generalizable and Scalable Visualization of Single-Cell Data Using Neural Networks.* Cell Syst, 2018. **7**(2): p. 185-191.e4.

10. Gordon, S., *Alternative activation of macrophages.* Nat Rev Immunol 2003. **3**(1): p. 23-35.

11. Ridley, D.S. and W.H. Jopling, *Classification of leprosy according to immunity. A five-group system.* Int J Lepr 1966. **34**: p. 255-273.

12. Modlin, R.L., et al., *T lymphocyte subsets in the skin lesions of patients with leprosy.* J Am Acad Dermatol, 1983. **8**(2): p. 182-9.

13. Stenger, S., et al., *Differential effects of cytolytic T cell subsets on intracellular infection.* Science, 1997. **276**(5319): p. 1684-7.

14. Stenger, S., et al., *An antimicrobial activity of cytolytic T cells mediated by granulysin.* Science, 1998. **282**(5386): p. 121-5.

15. Ochoa, M.T., et al., *T-cell release of granulysin contributes to host defense in leprosy.* Nat Med, 2001. **7**(2): p. 174-9.

16. Montoya, D., et al., *Divergence of macrophage phagocytic and antimicrobial programs in leprosy.* Cell Host Microbe, 2009. **6**(4): p. 343-53.

17. Cooper, C.L., et al., *Analysis of naturally occurring delayed-type hypersensitivity reactions in leprosy by in situ hybridization.* J Exp Med, 1989. **169**(5): p. 1565-81.

18. Yamamura, M., et al., *Defining protective responses to pathogens: cytokine profiles in leprosy lesions.* Science, 1991. **254**(5029): p. 277-9.

19. Yamamura, M., et al., *Cytokine patterns of immunologically mediated tissue damage.* J Immunol, 1992. **149**(4): p. 1470-5.

20. Teles, R.M., et al., *Type I interferon suppresses type II interferon-triggered human anti-mycobacterial responses.* Science, 2013. **339**(6126): p. 1448-53.

21. Liu, P.T., et al., *Toll-like receptor triggering of a vitamin D-mediated human antimicrobial response.* Science, 2006. **311**(5768): p. 1770-3.

22. Fabri, M., et al., *Vitamin D is required for IFN-gamma-mediated antimicrobial activity of human macrophages.* Sci Transl Med, 2011. **3**(104): p. 104ra102.

23. Ochoa, M.T., et al., *Role of granulysin in immunity to leprosy.* Lepr Rev, 2000. **71 Suppl**: p. S115; discussion S115-6.

24. Dang, A.T., et al., *IL-26 contributes to host defense against intracellular bacteria.* Journal of Clinical Investigation, 2019. **129**(5): p. 1926-1939.

25. Oshlack, A. and M.J. Wakefield, *Transcript length bias in RNA-seq data confounds systems biology.* Biol Direct, 2009. **4**: p. 14.

26. Barbash, S., et al., *Neuronal-expressed microRNA-targeted pseudogenes compete with coding genes in the human brain.* Transl Psychiatry, 2017. **7**(8): p. e1199.

27. Oberlin, S., et al., *A genome-wide transcriptome and translatome analysis of Arabidopsis transposons identifies a unique and conserved genome expression strategy for Ty1/Copia retroelements.* Genome Res, 2017. **27**(9): p. 1549-1562.

28. Tandonnet, S. and T.T. Torres, *Traditional versus 3' RNA-seq in a non-model species.* Genom Data, 2017. **11**: p. 9-16.

29. Dobin, A., et al., *STAR: ultrafast universal RNA-seq aligner.* Bioinformatics, 2013. **29**(1): p. 15-21.

30. Wang, L., S. Wang, and W. Li, *RSeQC: quality control of RNA-seq experiments.* Bioinformatics, 2012. **28**(16): p. 2184-5.

31. Robinson, J.T., et al., *Integrative genomics viewer.* Nat Biotechnol, 2011. **29**(1): p. 24-6.

32. Love, M.I., W. Huber, and S. Anders, *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.* Genome Biol, 2014. **15**(12): p. 550.

33. Ahmed, U., P.S. Latham, and P.S. Oates, *Interactions between hepatic iron and lipid metabolism with possible relevance to steatohepatitis.* World J Gastroenterol, 2012. **18**(34): p. 4651-8.

34. Kautz, L., et al., *Iron regulates phosphorylation of Smad1/5/8 and gene expression of Bmp6, Smad7, Id1, and Atoh8 in the mouse liver.* Blood, 2008. **112**(4): p. 1503-9.

35. Xiao, X., et al., *Lipocalin 2 alleviates iron toxicity by facilitating hypoferremia of inflammation and limiting catalytic iron generation.* Biometals, 2016. **29**(3): p. 451-65.

36. Liu, Z., et al., *Siderophore-mediated iron trafficking in humans is regulated by iron.* J Mol Med (Berl), 2012. **90**(10): p. 1209-21.

37. Krijt, J., et al., *Effect of iron overload and iron deficiency on liver hemojuvelin protein.* PLoS One, 2012. **7**(5): p. e37391.

38. Nam, H., et al., *ZIP14 and DMT1 in the liver, pancreas, and heart are differentially regulated by iron deficiency and overload: implications for tissue iron uptake in iron-related disorders.* Haematologica, 2013. **98**(7): p. 1049-57.

39. Zhang, Y.J., H.Y. Pan, and S.J. Gao, *Reverse transcription slippage over the mRNA secondary structure of the LIP1 gene.* Biotechniques, 2001. **31**(6): p. 1286, 1288, 1290, passim.

40. Xiong, Y., et al., *A Comparison of mRNA Sequencing with Random Primed and 3'-Directed Libraries.* Sci Rep, 2017. **7**(1): p. 14626.

41. Fu, Y., et al., *Elimination of PCR duplicates in RNA-seq and small RNA-seq using unique molecular identifiers.* BMC Genomics, 2018. **19**(1): p. 531.

42. Winkels, H., et al., *Atlas of the Immune Cell Repertoire in Mouse Atherosclerosis Defined by Single-Cell RNA-Sequencing and Mass Cytometry.* Circ Res, 2018. **122**(12): p. 1675-1688.

43. Glorot, X. and Y. Bengio, *Understanding the difficulty of training deep feedforward neural networks*, in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, T. Yee Whye and T. Mike, Editors. 2010, PMLR: Proceedings of Machine Learning Research. p. 249--256.

44. Butler, A., et al., *Integrating single-cell transcriptomic data across different conditions, technologies, and species.* 2018. **36**(5): p. 411-420.

45. Haghverdi, L., et al., *Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors.* Nat Biotechnol, 2018. **36**(5): p. 421-427.

46. Zheng, G.X., et al., *Massively parallel digital transcriptional profiling of single cells.* Nat Commun, 2017. **8**: p. 14049.

47. Xie, P., et al., *SuperCT: a supervised-learning framework for enhanced characterization of single-cell transcriptomic profiles.* Nucleic Acids Res, 2019. **47**(8): p. e48.

48. Lieberman, Y., L. Rokach, and T. Shay, *CaSTLe - Classification of single cells by transfer learning: Harnessing the power of publicly available single cell RNA sequencing experiments to annotate new experiments.* PLoS One, 2018. **13**(10): p. e0205499.

49. Tan, Y. and P. Cahan, *SingleCellNet: A Computational Tool to Classify Single Cell RNA-Seq Data Across Platforms and Across Species.* Cell Syst, 2019. **9**(2): p. 207-213.e2.

50. Macosko, E.Z., et al., *Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets.* Cell, 2015. **161**(5): p. 1202-1214.

51. Young, M.D. and S. Behjati, *SoupX removes ambient RNA contamination from droplet based single-cell RNA sequencing data.* bioRxiv, 2020: p. 303727.

52. Teles, R.M.B., et al., *Identification of a systemic interferon-gamma inducible antimicrobial gene signature in leprosy patients undergoing reversal reaction.* PLoS Negl Trop Dis, 2019. **13**(10): p. e0007764.

53. R. Andrade, P., et al., *The cell fate regulator NUPR1 is induced by Mycobacterium leprae via type I interferon in human leprosy.* PLoS Negl Trop Dis, 2019. **13**(7): p. e0007589.

54. Wang, H., et al., *Cellular, Molecular, and Immunological Characteristics of Langhans Multinucleated Giant Cells Programmed by IL-15.* J Invest Dermatol, 2020.

55. Waddell, S.J., et al., *Dissecting interferon-induced transcriptional programs in human peripheral blood cells.* PLoS ONE 2010. **5**(3): p. e9753.

56. Trapnell, C., et al., *The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells.* Nat Biotechnol, 2014. **32**(4): p. 381-386.

57. Wang, G., X. Li, and Z. Wang, *APD3: the antimicrobial peptide database as a tool for research and education.* Nucleic Acids Res, 2016. **44**(D1): p. D1087-93.

58. Busch, M., et al., *Lipoarabinomannan-Responsive Polycytotoxic T Cells Are Associated with Protection in Human Tuberculosis.* Am J Respir Crit Care Med, 2016. **194**(3): p. 345-55.

59. Balin, S.J., et al., *Human antimicrobial cytotoxic T lymphocytes, defined by NK receptors and antimicrobial proteins, kill intracellular bacteria.* Sci Immunol, 2018. **3**(26).

60. Cochain, C., et al., *Single-Cell RNA-Seq Reveals the Transcriptional Landscape and Heterogeneity of Aortic Macrophages in Murine Atherosclerosis.* Circ Res, 2018. **122**(12): p. 1661-1674.

61. Jaitin, D.A., et al., *Lipid-Associated Macrophages Control Metabolic Homeostasis in a Trem2-Dependent Manner.* Cell, 2019. **178**(3): p. 686-698 e14.

62. Keren-Shaul, H., et al., *A Unique Microglia Type Associated with Restricting Development of Alzheimer's Disease.* Cell, 2017. **169**(7): p. 1276-1290 e17.

63. Lavin, Y., et al., *Innate Immune Landscape in Early Lung Adenocarcinoma by Paired Single-Cell Analyses.* Cell, 2017. **169**(4): p. 750-765 e17.

64. Xue, D., et al., *Transcriptome landscape of myeloid cells in human skin reveals diversity, rare populations and putative DC progenitors.* J Dermatol Sci, 2020. **97**(1): p. 41-49.

65. Wang, E.C.E., et al., *A Subset of TREM2(+) Dermal Macrophages Secretes Oncostatin M to Maintain Hair Follicle Stem Cell Quiescence and Inhibit Hair Growth.* Cell Stem Cell, 2019. **24**(4): p. 654-669 e6.

66. Xiong, X., et al., *Landscape of Intercellular Crosstalk in Healthy and NASH Liver Revealed by Single-Cell Secretome Gene Analysis.* Mol Cell, 2019. **75**(3): p. 644-660 e5.

67. Tabib, T., et al., *SFRP2/DPP4 and FMO1/LSP1 Define Major Fibroblast Populations in Human Skin.* J Invest Dermatol, 2018. **138**(4): p. 802-810.

68. Sieling, P.A., et al., *CD1-restricted T cell recognition of microbial lipoglycans.* Science, 1995. **269**(5221): p. 227-230.

69. Chen, Y., A.B. Rabson, and D.H. Gorski, *MEOX2 regulates nuclear factor-kappaB activity in vascular endothelial cells through interactions with p65 and IkappaBbeta.* Cardiovasc Res, 2010. **87**(4): p. 723-31.

70. Yang, D., et al., *Many chemokines including CCL20/MIP-3alpha display antimicrobial activity.* J Leukoc Biol 2003. **74**(3): p. 448-455.

71. Liu, P.T., et al., *Convergence of IL-1beta and VDR activation pathways in human TLR2/1-induced antimicrobial responses.* PLoS One, 2009. **4**(6): p. e5810.

72. Bustamante, J., et al., *Germline CYBB mutations that selectively affect macrophages in kindreds with X-linked predisposition to tuberculous mycobacterial disease.* Nat Immunol 2011. **12**(3): p. 213-221.

73. Narni-Mancinelli, E., et al., *Memory CD8+ T cells mediate antibacterial immunity via CCL3 activation of TNF/ROI+ phagocytes.* J Exp Med, 2007. **204**(9): p. 2075-87.

74. Meller, S., et al., *T(H)17 cells promote microbial killing and innate immune sensing of DNA via interleukin 26.* Nat Immunol, 2015. **16**(9): p. 970-9.

75. Weiss, D.I., et al., *IL-1beta Induces the Rapid Secretion of the Antimicrobial Protein IL-26 from Th17 Cells.* J Immunol, 2019.

76. Hoover, D.M., et al., *Antimicrobial characterization of human beta-defensin 3 derivatives.* Antimicrob Agents Chemother, 2003. **47**(9): p. 2804-9.

77. Hong, J.S., et al., *Dual protective mechanisms of matrix metalloproteinases 2 and 9 in immune defense against Streptococcus pneumoniae.* J Immunol, 2011. **186**(11): p. 6427-36.

78. Kuula, H., et al., *Local and systemic responses in matrix metalloproteinase 8-deficient mice during Porphyromonas gingivalis-induced periodontitis.* Infect Immun, 2009. **77**(2): p. 850-9.

79. Montoya, D., et al., *IL-32 is a molecular marker of a host defense network in human tuberculosis.* Sci Transl Med, 2014. **6**(250): p. 250ra114.

80. Wong, E.A., et al., *IL-10 Impairs Local Immune Response in Lung Granulomas and Lymph Nodes during Early Mycobacterium tuberculosis Infection.* J Immunol, 2020. **204**(3): p. 644-659.

81. Cole, A.M., et al., *Cutting edge: IFN-inducible ELR- CXC chemokines display defensin-like antimicrobial activity.* J Immunol, 2001. **167**(2): p. 623-7.

82. Burkhardt, A.M., et al., *CXCL17 is a mucosal chemokine elevated in idiopathic pulmonary fibrosis that exhibits broad antimicrobial activity.* J Immunol, 2012. **188**(12): p. 6399-406.

83. O'Kane, C.M., et al., *Monocyte-dependent fibroblast CXCL8 secretion occurs in tuberculosis and limits survival of mycobacteria within macrophages.* J Immunol, 2007. **178**(6): p. 3767-76.

84. Rea, T.H., J.Y. Shen, and R.L. Modlin, *Epidermal keratinocyte Ia expression, Langerhans cell hyperplasia and lymphocytic infiltration in skin lesions of leprosy.* Clin Exp Immunol, 1986. **65**(2): p. 253-9.

85. Kaplan, G., et al., *The expression of a gamma interferon-induced protein (IP-10) in delayed immune responses in human skin.* J Exp Med 1987. **166**: p. 1098-1108.

86. Fattorini, L., et al., *In vitro activity of protegrin-1 and beta-defensin-1, alone and in combination with isoniazid, against Mycobacterium tuberculosis.* Peptides, 2004. **25**(7): p. 1075-7.

87. Pulido, D., et al., *Two human host defense ribonucleases against mycobacteria, the eosinophil cationic protein (RNase 3) and RNase 7.* Antimicrob Agents Chemother, 2013. **57**(8): p. 3797-805.

88. Wang, J., et al., *MRP8/14 induces autophagy to eliminate intracellular Mycobacterium bovis BCG.* J Infect, 2015. **70**(4): p. 415-26.

89. Dhiman, R., et al., *Interleukin 22 inhibits intracellular growth of Mycobacterium tuberculosis by enhancing calgranulin A expression.* J Infect Dis, 2014. **209**(4): p. 578-87.

90. Steinbakk, M., et al., *Antimicrobial actions of calcium binding leucocyte L1 protein, calprotectin.* Lancet, 1990. **336**(8718): p. 763-5.

91. Yavvari, P.S., et al., *Clathrin-Independent Killing of Intracellular Mycobacteria and Biofilm Disruptions Using Synthetic Antimicrobial Polymers.* Biomacromolecules, 2017. **18**(7): p. 2024-2033.

92. Lee, P.H., et al., *Expression of an additional cathelicidin antimicrobial peptide protects against bacterial skin infection.* Proc Natl Acad Sci U S A, 2005. **102**(10): p. 3750-5.

93. Modlin, R.L., et al., *In situ characterization of T lymphocyte subsets in the reactional states of leprosy.* Clin Exp Immunol, 1983. **53**(1): p. 17-24.

94. Montoya, D., et al., *Plasticity of antimicrobial and phagocytic programs in human macrophages.* Immunology, 2019. **156**(2): p. 164-173.

95. Keegan, C., et al., *Mycobacterium tuberculosis Transfer RNA Induces IL-12p70 via Synergistic Activation of Pattern Recognition Receptors within a Cell Network.* J Immunol, 2018. **200**(9): p. 3244-3258.

96. Ochoa, M.T., et al., *A role for interleukin-5 in promoting increased immunoglobulin M at the site of disease in leprosy.* Immunology, 2010. **131**(3): p. 405-414.

97. Montoya, D.J., et al., *Dual RNA-Seq of Human Leprosy Lesions Identifies Bacterial Determinants Linked to Host Immune Response.* Cell Rep, 2019. **26**(13): p. 3574-3585 e3.

98. Lu, L.L., et al., *A Functional Role for Antibodies in Tuberculosis.* Cell, 2016. **167**(2): p. 433-443.e14.

99. Modlin, R.L., et al., *Comparison of S-100 and OKT6 antisera in human skin.* J Invest Dermatol, 1984. **83**(3): p. 206-9.

100. Ma, F., et al., *A comparison between whole transcript and 3' RNA sequencing methods using Kapa and Lexogen library preparation methods.* BMC Genomics, 2019. **20**(1): p. 9.

101. Ma, F. and M. Pellegrini, *ACTINN: automated identification of cell types in single cell RNA sequencing.* Bioinformatics, 2020. **36**(2): p. 533-538.