From Benchtop to Bedside and Beyond: the development and application of low- and high-throughput, single-cell RNA-Seq platforms for precision medicine pipelines

by

Marc Havens Wadsworth II

B.S., Miami University (2014)

Submitted to the Department of Chemistry
In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2020

Signature of Author ……………………………………………………………………………
Department of Chemistry
February 28, 2019

Certified by……………………………………………………………………………………
Alex K. Shalek
Pfizer-Laubach Career Development Associate Professor of Chemistry
Thesis Supervisor

Accepted by……………………………………………………………………………………
Robert W. Field
Haslam and Dewey Professor of Chemistry
Chair, Departmental Committee on Graduate Students

From Benchtop to Bedside and Beyond: the development and application of low- and high-throughput, single-cell RNA-Seq platforms for precision medicine pipelines

by

Marc Havens Wadsworth II

Submitted to the Department of Chemistry
on February 28, 2020 in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy in Chemistry

Signatures

…………………………………………………………………………………………………

Alex K. Shalek
Pfizer-Laubach Career Development Associate Professor of Chemistry
Thesis Supervisor

…………………………………………………………………………………………………

Matthew D. Shoulders
Whitehead Career Development Associate Professor
Thesis Chair

…………………………………………………………………………………………………

J. Christopher Love
Raymond A. (1921) and Helen E. St. Laurent Professor of Chemical Engineering
Thesis Committee Member

From Benchtop to Bedside and Beyond: the development and application of low- and high-throughput, single-cell RNA-Seq platforms for precision medicine pipelines

by

Marc Havens Wadsworth II

Submitted to the Department of Chemistry
on February 28, 2020 in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy in Chemistry

Abstract

The development and application of single-cell technologies have revolutionized how we study health and disease. By deconstructing complex biological systems, like human tissues, into the fundamental building blocks of life, single cells, we can not only learn what makes each cell unique (intracellular circuitry) but also investigate how each interaction among them (intercellular circuits) lead to system-level functions. Single-cell approaches have the potential to be particularly crucial for precision medicine pipelines, where comprehensive cellular profiles of system-level functions could be leveraged to guide diagnosis and treatment of disease.

Here, to demonstrate the promise of these new technologies, we have developed and implemented single-cell RNA-Sequencing (scRNA-Seq) techniques to profile low-input clinical samples across a multitude of diseases, providing critical insight into how patient-specific scRNA-Seq profiles can help improve clinical treatment. More specifically, first, we applied scRNA-Seq to dissect the multicellular ecosystem of metastatic melanoma, profiling 4,645 single cells isolated from 19 patients to examine both malignant and non-malignant phenotypes and their interactions, as well as to propose potential targets for new therapies. Next, to overcome the limitations of low-throughput scRNA-Seq platforms, we developed Seq-Well, a high-throughput platform for low-input clinical samples, that is not only competitive with other scRNA-Seq technologies but also significantly cheaper and portable, enabling the democratization of scRNA-Seq technologies by empowering scientists in high- and low-resource settings. Finally, we drastically improved the gene and transcript capture of Seq-Well by introducing a step called Second Strand Synthesis (S^3) into the protocol and applied it to construct an atlas of skin inflammation across five conditions, resolving previously unappreciated adaptive and innate cellular phenotypes, as well as propose potential targets for therapeutic intervention unique to each inflammatory disease. Collectively, our work demonstrates the power of scRNA-Seq technologies and how they can be implemented in precision medicine pipelines to improve clinical outcomes.

Thesis Supervisor: Alex K. Shalek

Title: Pfizer-Laubach Career Development Associate Professor of Chemistry

Acknowledgments

1,989 Days. This time roughly translates to 47,736 hours or 2,864,160 minutes. While these numbers demarcate my time in graduate school, they cannot capture how rich this experience has been. It is safe to say that graduate school has been one of the hardest and most fulfilling experiences of my life. Throughput this process, I have met extraordinary individuals who have challenged me physically, mentally, and spiritually to be the best version of myself; these individuals will be lifelong friends as I continue to the next adventure!

I want to start by thanking my thesis advisor, Alex K. Shalek, for his support and guidance throughout the graduate school process. Alex has been an incredible mentor and challenged me to persevere when projects get tough. Importantly, he has taught me that it is never a matter of ‘if’ you can solve the problem, but rather ‘when’ you will and ‘how’ you will do it. Through his mentorship, I have learned the importance of collaboration and how to be a productive scientist. Thank you, Alex, for taking the time to invest in me! I would also like to thank my committee members, Matt Shoulders, and J. Christopher Love, for their guidance throughout this process; my labmates, specifically Travis K. Hughes for his friendship and inspiring scientific insight; Aleth Gaillard, Alex Genshaft, Sanjay Prakadan, Sam Kazer, Kellie Kolb, and Carly Ziegler (i.e., the Shalek O.Gs) for their friendship and support throughout this process; Peter Winter and Andrew Navia for setting a high bar when it comes to late-night Seq-Well…as well as being supportive friends; and, finally, the entire lab for making this an incredible experience. I am grateful to my external collaborators around the world, for their scientific insight and friendship, as we have worked together to push the frontiers of this field. Finally, I would like to thank my wife, Emily K. Ashby, for her love and support throughout this process.

This experience has been life-changing; I have thoroughly enjoyed the opportunity to collaborate on a wide variety of exciting projects and work with many incredible scientists!

Table of Contents

Lay Summary

Humans consist of trillions of cells, yet we lack the tools to deeply characterize the immense space of cellular identity and behavior that defines health and disease. A wide range of methods exist for sampling tissues in many clinical contexts (e.g., infection, cancer, autoimmunity); however, without high fidelity, comprehensive strategies for profiling them, we are limited in our capacity to identify how constituent cells and their interactions impact prognosis, and to select and develop precision therapeutics based thereupon. One potential disruptive technology, capable of thoroughly examining cellular phenotypic diversity in these precious clinical isolates, is single-cell RNA-Sequencing (scRNA-Seq). The development of scRNA-Seq methodologies enables genome-wide molecular profiling of transcriptomes of individual cells from which we can identify, absent bias, the cellular players (i.e., types and states) in nearly any sample, and their molecular drivers. In the following work, we describe scRNA-Seq and how the power and limitations of low-throughput implementations motivated the development of massively parallel, high-throughput methods. Further, we discuss how their application in precision medicine pipelines can potentially improve clinical outcomes.

To begin, we describe the application of a powerful low-throughput, plate-based method to construct single-cell profiles of the melanoma tumor microenvironment (chapter 2). Importantly, this work demonstrates how the complexity of diseases, like cancer, requires single-cell resolution to examine clinically relevant patient profiles. In melanoma, we showed that malignant cells exist on a spectrum of heterogeneity for the MITF and AXL expression programs, which are canonically thought to be separate programs based on bulk RNA-Seq profiles. This phenomenon potentially explains why targeted treatment of either tumor based on bulk profiles can lead to an outgrowth of drug-resistant tumor phenotypes. Moreover, our study highlights important limitations of low-throughput technologies, specifically the lack of scalability, high operational costs, and time-intensive protocols, thus necessitating the need for high-throughput assays.

As one solution to this, as well as the limitations of existing high-throughput assays (i.e., Drop-Seq, inDrop, 10x Genomics), we developed Seq-Well, a massive-parallel, microwell-based platform for low-input clinical samples. We benchmarked this technology against other high-throughput platforms in terms of single-cell resolution (i.e., species-mixing experiments), sensitivity (i.e., cell type resolution in peripheral blood mononuclear cells (PBMCs)), and portability. Importantly, we demonstrated the portability of Seq-Well by running the assay in a BSL3 facility, profiling macrophages exposed and unexposed to M. Tuberculosis (mTB) – the first scRNA-Seq experiment to examine the host response to mTB infection. By establishing a sensitive, portable, low-input platform, our work enabled the integration of scRNA-Seq technologies into precision medicine pipelines, reducing costs, and infrastructure requirements.

Finally, to improve the power of Seq-Well to reliably phenotype unique and rare immune subsets, limiting its application in clinical contexts, we developed Seq-Well S^3 (Second Strand Synthesis), a modified Seq-Well protocol with dramatically improved gene and transcript capture. Using this improved pipeline, we constructed an atlas of skin

inflammation, profiling immune, and parenchymal cell subsets. Importantly, with our improved sensitivity, we were able to profile previously unappreciated diversity in adaptive and innate immune subsets and identify phenotypes unique to the different inflammatory diseases. For example, we uncovered a population of dysfunctional T cells that were over-represented in patients with psoriasis. We were also able to propose biomarker targets, both unique and conserved across the inflammatory diseases, for therapeutic intervention.

Overall, our work demonstrates the utility of scRNA-Seq technologies in precision medicine pipelines, and, more specifically, how they can be leveraged to provide not only critical diagnostic insight, but also reveal patient-specific biomarkers to be targeted for therapeutic intervention.

Statement of Contributions

The work presented in the subsequent chapters is the culmination of years of highly collaborative projects in which I was honored to work closely with a multitude of scientists. In this section, I specify my contributions to each of the projects presented.

In chapter 2, I helped optimize the experimental design, modified the processing protocols to improve transcriptome coverage, worked with others to sequence and align samples, and helped with manuscript preparation.

In chapter 3, I worked with Travis Hughes and Todd Gierahn to optimize the surface chemistry of the Seq-Well assay, perform the subsequent scRNA-Seq experiments to demonstrate single-cell resolution (i.e., species-mixing) and cell-type resolution (i.e., PBMCs). In collaboration with Bryan Bryson, we demonstrated the portability of Seq-Well by having him perform the assay in a BSL3 facility. Finally, working with Travis Hughes and Todd Gierahn, we performed all the sequencing, alignment, analysis, as well as figure and manuscript preparation.

In chapter 4, I worked with Travis Hughes and Todd Gierahn to optimize the molecular biology of second-strand synthesis, as well as designed and performed the experiments for both technique benchmarking (i.e., species-mixing and PBMC experiments) and patient application (i.e., Atlas of skin inflammation). Travis Hughes and I performed all library generation, sequencing, and alignments. Finally, I worked with Travis Hughes to perform all computational analyses, as well as figure and manuscript preparation.

In chapter 5, I worked with Travis Hughes, Nil Gural, and Liliana Mancia on the host/pathogen study (section 5.1) to optimize and implement the experimental design. Travis Hughes and I performed all subsequent library preparation, sequencing, and alignments, while I built the genome and performed all downstream analyses. In section 5.2, I worked with Travis Hughes and Roisin O’Sullivan-Floyde to optimize the experimental design. Róisín Floyd-O’Sullivan and I implemented the experimental design and are currently optimizing the technology. In section 5.3, I worked with Saleem Aldajani to optimize and implement the experimental designs; we are currently optimizing the processing pipeline, as well as benchmarking our pipeline against current technologies. Finally, in section 5.4, I worked with Travis Hughes to process, sequence, align, and analyze all samples; we also worked with Patricia Darrah on figure and manuscript preparation.

In Appendix E, I worked with Alex Tsankov to optimize the single-cell experimental design. I implemented the design and performed all subsequent processing (i.e., library preparation, sequencing, alignment, and preliminary analyses). I worked with Benjamin Mead and Sam Allon to optimize and execute the small molecule FISH experiment. Sam Allon performed the subsequent smFISH analyses. Finally, I helped with figure and manuscript preparation.

Chapter 1: The Development of Single-Cell RNA-Sequencing Platforms for Clinical Application

1.1 Single-Cell Sequencing Technologies: A Brief Overview

The development and application of single-cell genomic technologies have revolutionized our understanding of complex biology systems[1-32], enabling us to comprehensively deconstruct complex biological systems, like the human body, into the fundamental building blocks of life, the cell. With this resolution, we can understand not only what makes each cell unique (its intracellular circuitry), but also examine how each interacts with other cells (intercellular circuits) to drive system-level functions. To date, a wide variety of single-cell molecular profiling approaches have been developed[1,23,27,32-46], allowing users to profile multiple different compartments (**Figure 1**). With these measurements, we can begin to better understand how certain cellular attributes link to others – e.g., genetic or epigenetic modifications to functional phenotypes (i.e., surface protein expression).[47,48] One of the most widely used single-cell sequencing approaches is single-cell RNA-Sequencing (scRNA-Seq).[49] This method captures the transcriptome of single cells by exploiting the 3'-polyadenylated tails of many ribonucleic acids (i.e., messenger RNA; mRNA). Using this information, we can view a 'snapshot' of the cell's current intentions, allowing us to surmise which molecular circuits are current activated or deactivated and how this may lead to functional phenotypes.[50-54]

By collecting this information from both healthy and diseased states, we can better identify how a disease perturbs a cell, allowing us to ultimately identify and propose biomarkers for therapeutic intervention.[55]

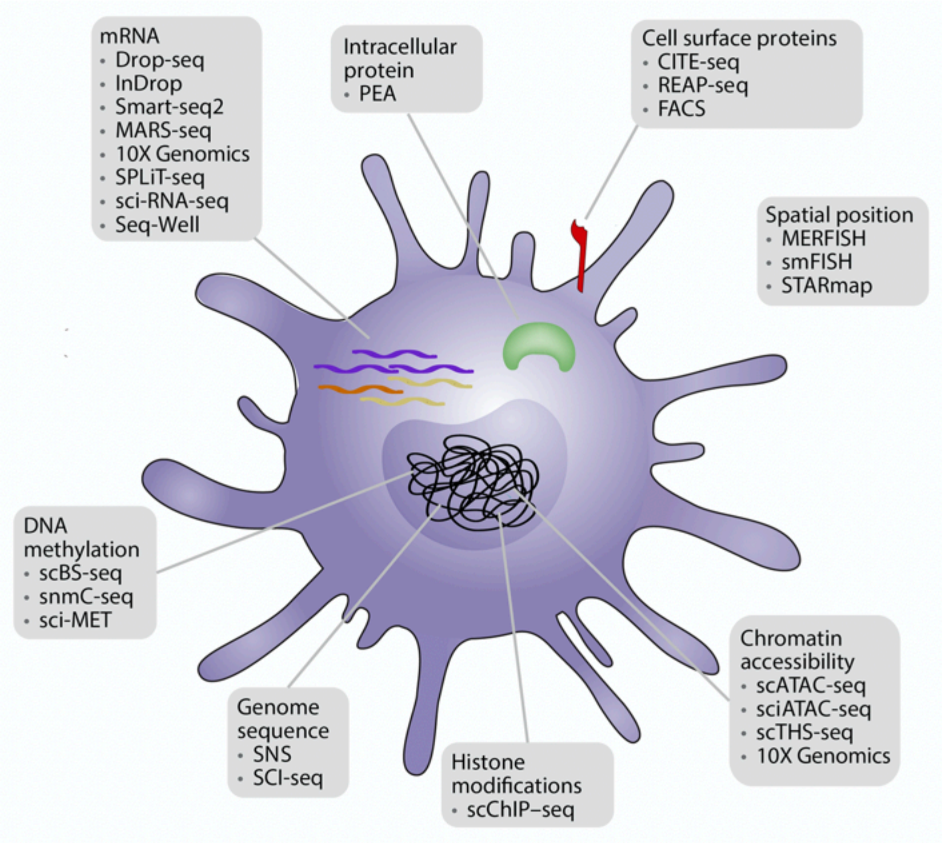


Figure 1 | Multi-omic single-cell technologies for integrated analysis. Adapted from Stuart et al.[49], to date, there are a multitude of single-cell technologies for interrogating different compartments of a single cell. Each of these sequencing technologies can provide critical insight into the overall function of a single cell and how they respond in healthy and diseased states. These methods can be divided based on the cellular parameter they measure: Transcriptome (mRNA; Drop-Seq, InDrop, Smart-Seq2, MARS-Seq, 10x Genomics, SPIT-Seq, sci-RNA-Seq, and Seq-Well); Intracellular proteins (PEA); Cell surface markers (CITE-Seq, REAP-Seq, FACS); Spatial Position (MERFISH, smFISH, and STARmap); Chromatin Accessibility (scATAC-Seq, sciATAC-Seq, scTHS-Seq, and 10x Genomics); Histone modification (scCHIP-Seq); Genome Sequence (SNS, SCI-Seq); and DNA methylation (scBS-Seq, snmC-Seq, and sci-MET).

The power of scRNA-Seq was first demonstrated in 2009 when Tang et al. leveraged their assay to profile single mouse blastomeres.[32] Cells were manually picked, using a microscope, and then lysed, after which cDNA was generated via reverse transcription using poly(T) primers for Next-Generation Sequencing (NGS) library preparation. Importantly, they showed an increase of 85% in detected genes relative to microarray techniques, as well as the sensitivity necessary to detail transcript isoform usage.[56] This seminal study was a springboard for the field, motivating experiments to capture and profile single cells more efficiently. Since, the field of scRNA-Seq has exploded, with the development of a wide range of strategies for isolating single cells and extracting their mRNA (**Figure 2**). Interestingly, a trend emerged in that, over time, with the development of new techniques, the number of single cells captured steadily increased. Nearly a decade after Tang et al.'s work, which studied a single cell, we can now capture tens of thousands of cells at once, powering the statistical analyses of these studies as well as better resolving unique and rare cellular phenotypes. Importantly, this improved statistical power and cellular resolution have enabled scientists to ask questions about health and disease previously unimaginable.

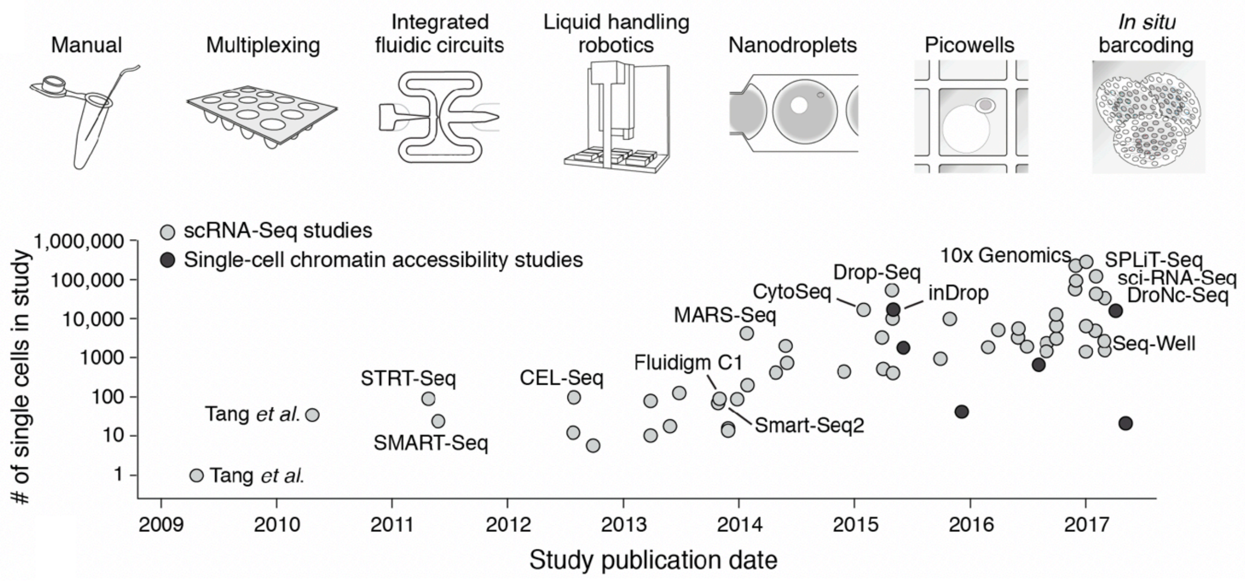


Figure 2 | Developments in experimental single-cell RNA-Seq technologies. Adopted from the HCA Consortium white paper[57], (**A**) Technologies developed for single-cell isolation. (**B**) Scatter plot showing the number of cells captured versus time for scRNA-Seq (gray circles) and scATAC-Seq (black circles) technologies (key methods indicated).

For example, this technology development has helped catalyze a global effort called the Human Cell Atlas (HCA) Consortium,[57] which seeks to realize a comprehensive reference map of all human cell types and properties to be leveraged in the understanding, monitoring, diagnosing, and treatment of human disease. Through an extensive international, collaborative effort, scientists are now actively leveraging these technologies to profile and categorize every human cell type in the body.

The development of a Human Cell Atlas will revolutionize the way we study health and disease. From an academic standpoint, this reference will be pivotal for standardizing cell type classification and contextualizing how diseases perturb cellular circuits from a healthy, steady-state (i.e., cellular reference). From a translational one, by contextualizing disease discoveries to a standardized reference, it has the potential to dramatically improve our ability to identify targets for drug validation and gene therapies (i.e., CRISPR-cas9 and lentiviral).[8,9,19] The HCA thus has the potential to bridge single-cell benchtop discoveries to translational treatments, becoming a powerful precision medicine pipeline for patient care.[57]

In the following section, we will highlight powerful implementations of scRNA-Seq and the promise this technology has to transform clinical care.

1.2 Implementation of low-throughput scRNA-Seq Technologies

Following the demonstration of the power of scRNA-Seq technology, early efforts focused on scaling up so that multiple cells could be processed in parallel. The goals were two-fold: 1. to increase throughput and 2. to decrease cost. Among these early efforts was the development of Smart-Seq[2] (Switching Mechanism at the 5'End of RNA Templates) in 2012, a plate-based method for capturing transcripts from single cells. Here, cells are flow-sorted into 96 or 384-well plates where they are lysed and immediately processed or flash-frozen and banked at -80°C. Compared with existing protocols, SMART-Seq had improved read coverage across transcripts, allowing users to generate genome-wide transcriptome profiles for cell types of interest, as well as detect alternative transcript isoforms and single-nucleotide polymorphisms (SNPs). This protocol was quickly optimized, and a second iteration was released in 2014 called Smart-Seq2[3]; the result was improved detection, coverage, and accuracy while decreasing the 3' coverage bias (**Figure 3**). Underlying these improvements were careful molecular manipulations. For example, in Smart-Seq2, Picelli et al. introduced a locked nucleic acid (LNA) at the template-switching oligo (TSO) three prime end to increase the thermal stability of the LNA:DNA base pairs, resulting in a substantial increase in cDNA yield post reverse transcription.[3]

While these early studies were powerful demonstrations of the potential of scRNA-Seq, they did not motivate widespread application because, while profiling the transcriptome of single cells was impressive, it had yet to be shown why it was important.

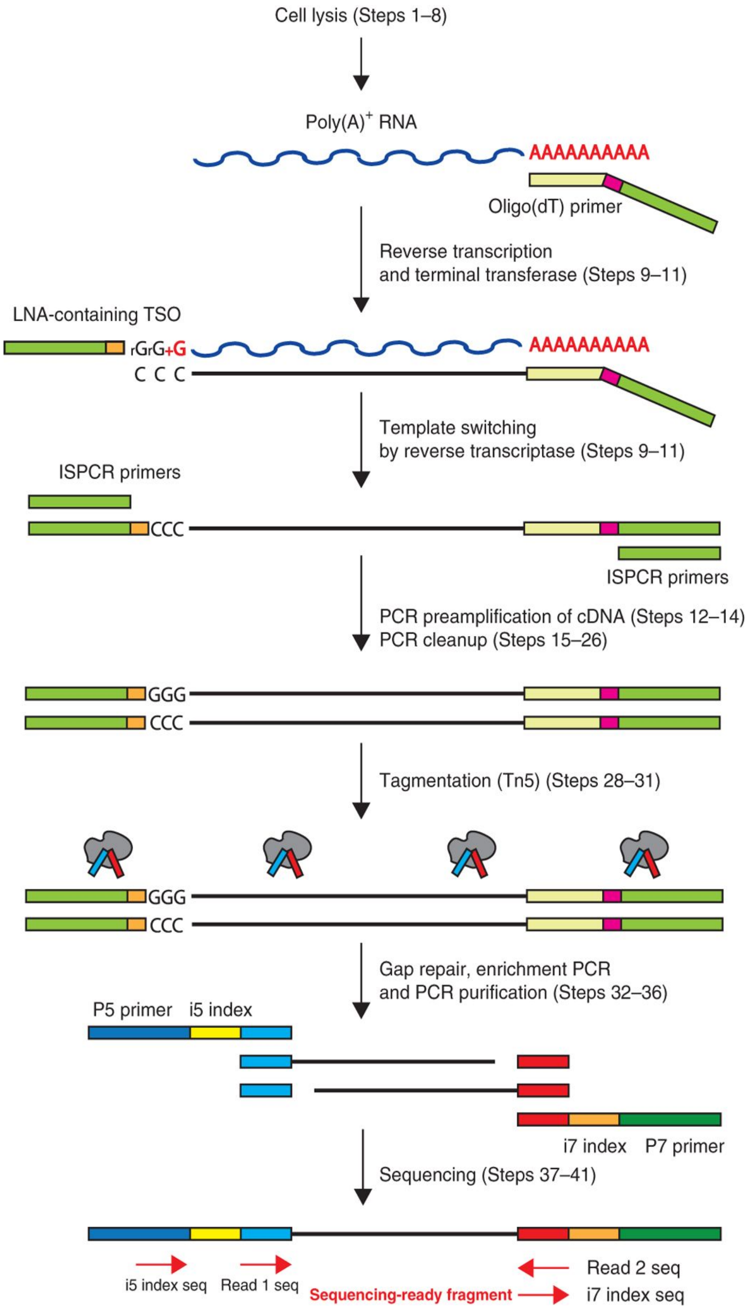


Figure 3 | Overview of Smart-Seq2 Pipeline. Adopted from Trombetta *et. al.*[1-3], cells are first lysed and mRNA transcripts captured with biotinylated poly-T SMART primers for reverse transcription (RT). Importantly, the RT enzyme deposits three ribosomal cytosines which are exploited to attach the TSO, generation full-length transcripts. Following complementary DNA (cDNA) generation, the product is amplified through PCR, tagmented by transposase (Illumina, Nextera) which deposits i7 and i5 adapters that serve as priming sights for the P7 and P5 primers. The quality of the library is validated using a BioAnalyzer (or Tape Station) and KAPA Quant. Once confirmed the library meets the necessary quality, it can be sequenced on an illumine instrument.

This was realized in 2013 when Shalek et al. leveraged scRNA-Seq, specifically Smart-Seq, to show that structure hidden within gene expression covariation across seemingly ‘identical’ dendritic cells (DCs) exposed to lipopolysaccharide (LPS) could be used to identify two distinct cell states and an interferon-driven antiviral circuit that they subsequently validated in knockout models.[21] By increasing the throughput and control of this method with microfluidic cell preparation and isolation, Shalek et al further demonstrated that a rare (1-2%), transient subset of “precocious” cells are essential to activate this antiviral response and dampen inflammation across all DCs in the population.[18] Collectively, these studies showed the promise of scRNA-Seq to uncover cell types/states, and circuity *de novo*, as well as rare cell populations that would go undetected as a result of being masked by the bulk profile.

With increased throughput and improved sample recovery, researchers began leveraging these techniques to study a wide variety of systems.[58,59] For example, there were early efforts to profile tumor biopsies, specifically to elucidate the cellular heterogeneity and the potential implications this may have for clinical outcomes. A powerful study that demonstrates this was published back in 2014, when Patel et al. leveraged scRNA-Seq to profile 430 cells from fiver primary glioblastomas, defining variability among transcriptional programs related to oncogenic signaling, proliferation, immune responses, and hypoxia.[16] Critically, by scoring cells using existing glioblastoma subtype classification signatures, they showed that each of the five glioblastomas profiled actually consisted of cells defined by multiple different subtypes. The extensive genetic and functional intra-tumoral heterogeneity they observed within freshly resected human glioblastoma tumor cells suggested potential prognostic and therapeutic implications. This was another critical steppingstone that validated the utility of scRNA-Seq technology because it was one of the first single-cell genomic studies to show how cellular heterogeneity could reveal clinically important features that, without single-cell resolution, would be lost. However, this was an n=5 study, and in order to make scRNA-Seq more applicable in clinical settings, the field needed a standardized pipeline to collect, bank, and process single cells from individual patients. In other words, the field needed a precision medicine pipeline for patients.

The development and application of this precision medicine pipeline, in which I took part (see Chapter 2), came to fruition in 2016 study where Tirosh et al. profiled tumor aggregates, and matched blood, from patients with metastatic melanoma.[60] Overall, we profiled 4,632 cells across 19 metastatic melanomas, highlight similarities and differences between tumor and immune cell populations. For example, by studying malignant cells across patients, we showed, based on bulk RNA-Seq profiles, that two programs, MITF high and AXL high, which are canonically thought to be mutually exclusive, actually define a spectrum of intra-tumor heterogeneity. This observation helped to explain why targeted treatment of either of these tumor types based on bulk RNA-Seq profiling data alone would result in the outgrowth of drug-resistant tumor phenotypes. Further, we were able to study patterns of T cell exhaustion that correlated with tumor attributes, highlighting important T cell programs and how they relate to tumor burden. Overall, this study proposed a robust precision medicine pipeline that could potentially help improve clinical treatment of metastatic melanomas, as well as served as a critical framework for a separate study where scRNA-Seq was leveraged to elucidate the developmental tumor cell hierarchy in oligodendrogliomas; this will be discussed in section 1.4.[61]

In the following section, we will address the limitations of low-throughput methods and how this motivated the development of high-throughput platforms.

1.3 Motivation and Development of High-Throughput scRNA-Seq Technologies

While early applications of low-throughput pipelines were exciting, it quickly became apparent that critical limitations needed to be addressed to permit implementation in both high- and low-resource settings. For example, while Smart-Seq2 and other early microfluidic methods (e.g., Fluidigm[62,63]) were effective at single-cell isolation and processing, they were limited in the number of cells that could be processed at a time and cost per cell. This issue was compounded by the need for dedicated equipment for either cell isolation (i.e., Fluidigm instrumentation) and/or sorting (i.e., flow-cytometer for plate-based sorting). As a result, to process the thousands of cells needed to power certain statistical analyses (i.e., profiling TCR clonality) was prohibitively expensive, thus complicating widespread adoption of scRNA-Seq methods for research application.

To address these limitations, microfluidic, droplet-based methods were developed that leveraged the power of early bead-based barcoding[35,36,38]. Here, cells are co-encapsulated in reverse-emulsion droplets where cells are lysed, and mRNA molecules are captured on uniquely barcoded oligo dT-capture beads. After each cell's transcripts are capture with unique cellular tags, the beads can be pulled into a single tube for subsequent processing (where the tags are appended during reverse transcription), thus increasing throughput while decreasing reagent costs. For example, the cost/cell for plate-based platforms (i.e., Smart-Seq2) is between 12-15 USD, while massive-parallel techniques are between 5-10 cents.[1,15,18,21,35] As a result of these improvements, tens of thousands of single cells could be captured and studied in a single experiment as opposed to hundreds in previous methods.[2,3,18,21] While impressive, these technologies were difficult to translate over to clinical specimens due to the need for several peripheral pieces of equipment and sample capture efficiency issues. For example, in Drop-Seq[36], beads and cells had to be loaded at specific concentrations to achieve double Poisson-loading to ensure single bead-cell pairings; as a result, more than 80% of the sample is lost. Other droplet-based methods were either prohibitively expensive, like 10x Genomics, to scale up or not easily portably for clinical application.

In response, I co-developed Seq-Well[44], a microwell-based platform for processing low-input clinical samples (**Figure 4**). This technology combines the power of bead-based barcoding with the simplicity and versatility of microwell arrays by co-encapsulating beads and cells in microwells. Critically, in Seq-Well, these wells are sealed with a semi-permeable membrane, facilitating buffer exchange while containing biological macromolecules in their corresponding wells.

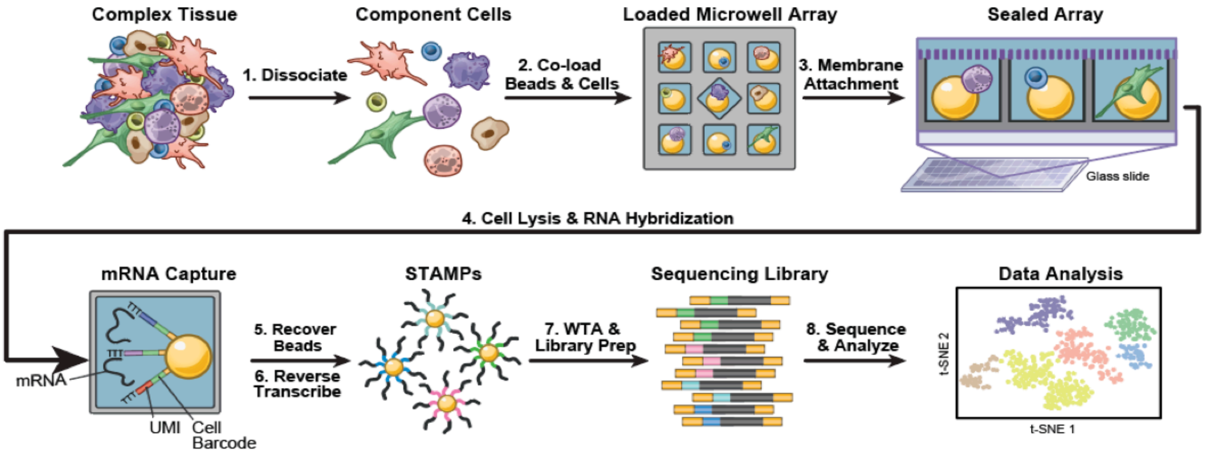


Figure 4 | Overview of Seq-Well Pipeline. Adopted from Gierahn *et. al.*[64], Following tissue dissociation and generation of a single cell suspension, uniquely-barcoded mRNA capture beads and single cells are loaded onto a functionalized Polydimethylsiloxane (PDMS) array. Critically, after beads and cells settle into wells, the array is sealed with a semi-permeable polycarbonate membrane that facilitates buffer exchange will keeping biological macromolecules e.g. RNA molecules) confined to their wells, preventing cross contamination. The mRNA molecules then hybridize to bead oligos, which are then recovered from the PDMS array and pooled for reverse transcription (RT). Following cDNA generation, the STAMPs (single-cell transcriptomes attached to microparticles) undergo subsequent processing (i.e. whole transcriptome amplification (WTA) and library prep (i.e. Nextera XT)) to generate sequencing libraries. Once libraries pass the necessary quality controls metrics, libraries are then sequenced on an illumine instrument.

This technology, along with other microwell implementations[43,44,64,65], has expanded our ability to profile, with single-cell resolution, clinical samples previously unobtainable (i.e., cerebral spinal fluid, sputum, gut-pinch biopsies). Also, because these platforms require minimal equipment and are relatively portable, it further reduces costs, democratizing scRNA-Seq technologies, and empowering researchers in low-resource settings. Thus, these technologies overcome the limitations of plate- and microfluidic-based technologies outlined in previous sections, making the implementation of a precision medicine pipeline more tangible.

In the following section, we will address the outstanding challenges of integrating scRNA-Seq technologies into precision medicine pipelines.

1.4 Applications of Single-Cell Technologies in Precision Medicine Pipelines

While there is a multitude of applications for single-cell technologies, an important implementation is in precision medicine pipelines, where these technologies can be used to build comprehensive cellular patient profiles. These profiles have the potential to pivotally guide the diagnosis and treatment of disease because they can be leveraged to find biomarkers and targets unique to that individual rather than based on a broader profile, thus overcoming current limitations both in disease treatment and drug development. For example, almost 90% of drugs developed are effective in less than 50% of patients[55,66] ; as a result, this creates an enormous financial burden in the treatment of disease. As seen in previous sections, a potential reason for this inefficiency is the intrinsic cellular heterogeneity both within and between patients.[16,29,61]

Single-cell technologies are a potential solution to this issue because they provide the necessary resolution to elucidate cellular heterogeneity in patient samples. Importantly, these tools let users profile multiple compartments within a cell (**Figure 1)**, allowing one to study how changes at the genetic or epigenetic level can manifest in the transcriptome and proteome. While profiling each of these '-omes' would be ideal, there are inherent limitations that prevent implementation. However, of the single-cell technologies at our disposal, scRNA-Seq is a major contender because of its scalability and relatively easy implementation.[4,23,35,36,38,67-70]

Early demonstrations of scRNA-Seq, to characterize diseased clinical samples, primarily focused on cancer. For example, scRNA-Seq was leveraged to study oligodendrogliomas to elucidate a developmental tumor cell hierarchy.[61] Here, scientists showed that stem-like cells could drive tumor growth and give rise to differentiated progeny, and ultimately proposed candidate gene targets for therapeutic potential. Recently, in melanoma, another scRNA-Seq study identified a resistance program expressed by malignant cells that are associated with T cell exclusion and immune evasion.[71] The authors showed that this program is expressed prior to immunotherapy and can be used as a predictive program for clinical responses to anti-PD-1 therapy. Together, these studies demonstrate the functional role scRNA-Seq has in precision medicine pipelines.

Finally, to successfully transition scRNA-Seq workflows from benchtop to bedside, there are still a series of challenges that need to be addressed. Foremost, there needs to be an efficient manner of accurately translating findings from scRNA-Seq studies to widespread clinical implementation. This is complicated by the technical (i.e., sample preparation and processing) and biological noise (i.e., stochastic biological signal) that is inherent to single-cell technologies.[72,73] For example, a challenging aspect of single-cell technologies is ‘dropout’, which is a missing value in a dataset or, in this case, missing gene detection, and is the result of failed detection events (i.e., low transcriptome coverage as a result of the technique used).[73] These technical challenges can be exacerbated by the stochastic nature of gene expression (i.e., transcriptional bursting) driven by biological, physical, and temporal properties of single cells[73,74]. While these are not major issues at the population level, it can have detrimental consequences at the level of a single cell, complicating cell type classification and subsequent phenotypic profiling.[73] However, high-throughput platforms help overcome the technical and biological noise inherent in single-cell measurements because, by increasing the number of cells sampled, high-throughput platforms allow users to study the distribution of a variable across a large population, resulting in a better profile of a system that is buffered against systematic (technical noise) and random (intrinsic noise) features.[73,74] In parallel, new computational methods are being developed to process these high-dimensional datasets, as well as standardize the analysis workflow to make it easier for non-statisticians to implement.[75] Ultimately, what will catalyze widespread application of scRNA-Seq technologies in the clinic will be the completion of the HCA reference; through this effort, both technologies and sample processing pipelines will be optimized and standardized for clinical application.[57] Also, through the completion of this reference, clinicians will have a reference to use and evaluate patient profiles, deconstructing complex, multifactorial diseases into their individual components; in doing so, it will provide a more comprehensive validation process of patient-specific features, thus transforming how disease is treated in the clinic.

1.5 Contributions of this work to the field

To date, scRNA-Seq technologies have empowered scientists around the world, providing unprecedented biological resolution of health and diseased systems and allowing scientists to ask hypotheses previously unimaginable.[76] As a result, these new discoveries have both improved our understanding of various biological systems, as well as motivate the next generation of single-cell technologies.[77] With respect to this, it is paramount to the field to understand what trends led to the technologies used today, and how that process can be optimized for future developments.[78]

In the following work, we explore the motivation, development, and application of high-throughput platforms for low-input clinical samples; from early demonstrations of Smart-Seq2 to study the melanoma tumor microenvironment[60] to the application of Seq-Well S^3 (Second Strand Synthesis) to build an atlas of skin inflammation and propose potential targets for therapeutic intervention[79]. Through these studies, we learn about the development of massively parallel scRNA-Seq technologies and how they can be implemented, in a clinical setting, to provide disease-specific profiles for drug development.

In Chapters 2, we apply scRNA-Seq, specifically Smart-Seq2[2,3], to address several biological questions by profiling the genotypic and phenotypic states of melanoma tumor microenvironments. Leveraging scRNA-Seq, we show how intratumor heterogeneity has important clinical implications, specifically in cancer treatment regimens, as previously mentioned.

In Chapters 3 and 4, we develop, optimize, and apply Seq-Well to profile low-input clinical samples. As addressed in the previous sections, while plate-based techniques[1-3] are powerful, they have inherent limitations that prevent widespread use for clinical samples. To address this, we developed Seq-Well (Chapter 3) and demonstrated both its utility for complex clinical samples, as well as it's portability. Here, we benchmarked Seq-Well using cell lines (e.g., HEK293s and NIH/3T3s) against standards in the field (i.e., Drop-Seq, inDrop, and 10x genomics), profiling human peripheral blood mononuclear cells (PBMCs) and capturing known cell populations, and finally profile macrophages exposed and unexposed to *M. tuberculosis* in a BSL3 facility to demonstrate the platforms portability. In Chapter 4, we introduce a substantially improved scRNA-Seq protocol we term Seq-Well S^3 ("Second Strand Synthesis") that enables increased efficiency in gene and transcription detection, on a per-cell basis, than other high-throughput platforms. As a result, we can better classify subtle immune phenotypes (i.e., Th1, Th17 T cells, M1/M2-like macrophages, etc.) than other best-in-class platforms. We demonstrate the power of this approach by examining five different inflammatory skin diseases, constructing an atlas of inflammation, and explore the breadth of potential immune and parenchymal cell states with improved resolution.

Finally, in Chapter 5, we explore ongoing projects and the unique application of scRNA-Seq technologies. To begin, we are leveraging Seq-Well S3 to profile the host/pathogen interactions in *P. vivax* infection to understand liver-stage infection better. With improved resolution, we can reconstruct the parasite lifecycle and work towards a comprehensive transcriptional profile of how a parasite commits to either becoming a hypnozoite (i.e., dormant-phase) or matured schizont. Critically, because we have both host and pathogen information, we can translate these parasite observations to the host environment and better understand how the host responds to hypnozoite and schizont development. Secondly, we are developing the next iteration of Seq-Well where we leverage the molecular biology of second-strand synthesis to generate sequencing libraries directly off the beads. As a result, it reduces processing time and cost because Tn5 hyperactive transposase is not required to generate the library. In parallel, to expand the utility of Seq-Well we are developing Nuc-Seq, a modified version of the Seq-Well platform that is compatible with nuclei. This modified pipeline will be compatible with previously difficult sample types (i.e., biobank frozen archived tissue) for Seq-Well, providing not only a cost-effective alternative to commercially-available platforms, but also a pipeline to overcome the cell isolation-based transcriptional artifacts that can be introduced by cellular processing pipelines.[80] Finally, we recently leveraged scRNA-Seq to elucidate the mechanisms of Bacillus Calmette-Guérin (BCG)-induced protection against pulmonary tuberculosis infection.[81] Our initial study suggests that intravenous (IV) vaccination with BCG, when compared to the standard intradermal (ID) method, dramatically alters the protective outcome of Mtb infection. Compared to ID, IV immunization induced more antigen-responsive CD4 and CD8 T cells across all lung parenchymal tissues. Specifically, we observed an IV-BCG enriched module of correlated gene expression associated with T cell survival and effector function that is enriched among T cells with a Th1/Th17 phenotype. Currently, we are profiling the alveolar space of NHPs at different IV-BCG dosages to better resolve the cellular correlates of protection and how they change with respect to dosage concentration.

As the dawn of precision medicine rapidly approaches, scRNA-Seq methods have emerged as a powerful set of tools to help facilitate this. It is through the significant contributions and developments from a wide range of biological, technology development, and computational fields that have empowered scRNA-Seq technologies and enabled the rapid development of both experimental and computational pipelines. As we begin to translate benchtop applications to bedside treatment options, we turn to the horizon and beyond as we explore the next steps of single-cell technology applications.

1.7 References

1 Trombetta, J. J. *et al.* Preparation of single-cell RNA-Seq libraries for next generation sequencing. *Current Protocols in Molecular Biology* **2014**, 4.22.21-24.22.17, doi:10.1002/0471142727.mb0422s107 (2014).

2 Picelli, S. *et al.* Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nature Methods* **10**, 1096-1100, doi:10.1038/nmeth.2639 (2013).

3 Picelli, S. *et al.* Full-length RNA-seq from single cells using Smart-seq2. *Nature Protocols* **9**, 171-181, doi:10.1038/nprot.2014.006 (2014).

4 Rodriques, S. G. *et al.* Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science* **363**, 1463-1467, doi:10.1126/science.aaw1219 (2019).

5 Manno, G. RNA velocity of single cells. *Nature* **560**, 494-498 (2018).

6 Rosenberg, A. B. Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science* **360** (2018).

7 Workman, R. E. Nanopore native RNA sequencing of a human poly(A) transcriptome. *bioRxiv*, doi:10.1101/459529 (2018).

8 Klann, T. S. CRISPR–Cas9 epigenome editing enables high-throughput screening for functional regulatory elements in the human genome. *Nat. Biotechnol.* **35**, 561-561 (2017).

9 Datlinger, P. Pooled CRISPR screening with single-cell transcriptome readout. *Nat. Methods* **14**, 297-301 (2017).

10 Müller, S. *et al.* Single-cell sequencing maps gene expression to mutational phylogenies in PDGF - and EGF -driven gliomas. *Molecular Systems Biology* **12**, 889-889, doi:10.15252/msb.20166969 (2016).

11 Setty, M. *et al.* Wishbone identifies bifurcating developmental trajectories from single-cell data. *Nature Biotechnology* **34**, 637-645, doi:10.1038/nbt.3569 (2016).

12 Kowalczyk, M. S. *et al.* Single-cell RNA-seq reveals changes in cell cycle and differentiation programs upon aging of hematopoietic stem cells. *Genome Research* **25**, 1860-1872, doi:10.1101/gr.192237.115 (2015).

13 Buenrostro, J. D. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486-490 (2015).

14 Klein, A. M. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187-1201 (2015).

15 Macosko, E. Z. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202-1214 (2015).

16 Patel, A. P. *et al.* Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344**, 1396-1401, doi:10.1126/science.1254257 (2014).

17 Jaitin, D. A. *et al.* Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* **343**, 776-779, doi:10.1126/science.1247651 (2014).

18 Shalek, A. K. *et al.* Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature* **510**, 363-369, doi:10.1038/nature13437 (2014).

19 Gilbert, L. A. Genome-scale CRISPR-mediated control of gene repression and activation. *Cell* **159**, 647-661 (2014).

20 Smallwood, S. A. Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nat. Methods* **11**, 817-820 (2014).

21 Shalek, A. K. *et al.* Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature* **498**, 236-240, doi:10.1038/nature12172 (2013).

22 Smith, Z. D. & Meissner, A. Vol. 14 204-220 (2013).

23 Hashimshony, T., Wagner, F., Sher, N. & Yanai, I. CEL-Seq: Single-Cell RNA-Seq by Multiplexed Linear Amplification. *Cell Reports* **2**, 666-673, doi:10.1016/j.celrep.2012.08.003 (2012).

24 Yamanaka, Y. J. *et al.* Single-cell analysis of the dynamics and functional outcomes of interactions between human natural killer cells and target cells. *Integrative Biology (United Kingdom)* **4**, 1175-1184, doi:10.1039/c2ib20167d (2012).

25 Han, Q. *et al.* Polyfunctional responses by human T cells result from sequential release of cytokines. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 1607-1612, doi:10.1073/pnas.1117194109 (2012).

26 Buganim, Y. *et al.* Single-cell expression analyses during cellular reprogramming reveal an early stochastic and a late hierarchic phase. *Cell* **150**, 1209-1222, doi:10.1016/j.cell.2012.08.023 (2012).

27 Ramsköld, D. *et al.* Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nature Biotechnology* **30**, 777-782, doi:10.1038/nbt.2282 (2012).

28 Bendall, S. C. *et al.* Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. *Science* **332**, 687-696, doi:10.1126/science.1198704 (2011).

29 Wagle, N. *et al.* Dissecting therapeutic resistance to RAF inhibition in melanoma by tumor genomic profiling. *Journal of Clinical Oncology* **29**, 3085-3096, doi:10.1200/JCO.2010.33.2312 (2011).

30 Baitsch, L. *et al.* Exhaustion of tumor-specific CD8+ T cells in metastases from melanoma patients. *Journal of Clinical Investigation* **121**, 2350-2360, doi:10.1172/JCI46102 (2011).

31 Hansen, K. D. *et al.* Increased methylation variation in epigenetic domains across cancer types. *Nature Genetics* **43**, 768-775, doi:10.1038/ng.865 (2011).

32 Tang, F. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nature Methods* **6**, 377-382, doi:10.1038/nmeth.1315 (2009).

33 Sinnamon, J. R. *et al.* The accessible chromatin landscape of the murine hippocampus at single-cell resolution. *Genome Research* **29**, 857-869, doi:10.1101/gr.243725.118 (2019).

34 Fan, H. C., Fu, G. K. & Fodor, S. P. A. Combinatorial labeling of single cells for gene expression cytometry. *Science* **347**, doi:10.1126/science.1258367 (2015).

35 Klein, A. M. *et al.* Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187-1201, doi:10.1016/j.cell.2015.04.044 (2015).

36 Macosko, E. Z. *et al.* Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202-1214, doi:10.1016/j.cell.2015.05.002 (2015).

37 Ramani, V. *et al.* Massively multiplex single-cell Hi-C. *Nature Methods* **14**, 263-266, doi:10.1038/nmeth.4155 (2017).

38 Zheng, G. X. Y. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8** (2017).

39 Marcus, J. S., Anderson, W. F. & Quake, S. R. Microfluidic single-cell mRNA isolation and analysis. *Analytical Chemistry* **78**, 3084-3089, doi:10.1021/ac0519460 (2006).

40 Cusanovich, D. A. *et al.* Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* **348**, 910-914, doi:10.1126/science.aab1601 (2015).

41 Genshaft, A. S. Multiplexed, targeted profiling of single-cell proteomes and transcriptomes in a single reaction. *Genome Biol.* **17**, 1-15 (2016).

42 Dixit, A. Perturb-Seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell* **167**, 1853-1866 (2016).

43 Bose, S. *et al.* Scalable microfluidics for single-cell RNA printing and sequencing. *Genome Biology* **16**, doi:10.1186/s13059-015-0684-3 (2015).

44 Gierahn, T. M. *et al.* Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput. *Nature Methods* **14**, 395-398, doi:doi:10.1038/nmeth.4179 (2017).

45 Vitak, S. A. *et al.* Sequencing thousands of single-cell genomes with combinatorial indexing. *Nature Methods* **14**, 302-308, doi:10.1038/nmeth.4154 (2017).

46 Smallwood, S. A. *et al.* Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nature Methods* **11**, 817-820, doi:10.1038/nmeth.3035 (2014).

47 Stoeckius, M. *et al.* Simultaneous epitope and transcriptome measurement in single cells. *Nature Methods* **14**, 865-868, doi:10.1038/nmeth.4380 (2017).

48 Peterson, V. M. Multiplexed quantification of proteins and transcripts in single cells. *Nat. Biotechnol.* **161**, 1202-1202 (2017).

49 Stuart, T. & Satija, R. Vol. 20 257-272 (Nature Publishing Group, 2019).

50 Buettner, F. *et al.* Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nature Biotechnology* **33**, 155-160, doi:10.1038/nbt.3102 (2015).

51 Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology* **36**, 411-420, doi:10.1038/nbt.4096 (2018).

52 Matuła, K., Rivello, F. & Huck, W. T. S. (Wiley-VCH Verlag, 2019).

53 Vallejos, C. A., Marioni, J. C. & Richardson, S. BASiCS: Bayesian Analysis of Single-Cell Sequencing Data. *PLoS Computational Biology* **11**, doi:10.1371/journal.pcbi.1004333 (2015).

54 Wolock, S. L., Lopez, R. & Klein, A. M. Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. *Cell Systems* **8**, 281-291.e289, doi:10.1016/j.cels.2018.11.005 (2019).

55 Shalek, A. K. & Benson, M. Single-cell analyses to tailor treatments. *Science Translational Medicine* **9**, eaan4730, doi:10.1126/scitranslmed.aan4730 (2017).

56 Tang, F. mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* **6**, 377-382 (2009).

57 Regev, A. *et al.* The human cell atlas. *eLife* **6**, doi:10.7554/eLife.27041 (2017).

58 Brennecke, P. *et al.* Single-cell transcriptome analysis reveals coordinated ectopic gene-expression patterns in medullary thymic epithelial cells. *Nature Immunology* **16**, 933-941, doi:10.1038/ni.3246 (2015).

59 Tan, L., Li, Q. & Xie, X. S. Olfactory sensory neurons transiently express multiple olfactory receptors during development. *Molecular Systems Biology* **11**, 844, doi:10.15252/msb.20156639 (2015).

60 Tirosh, I. *et al.* Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352**, 189-196, doi:10.1126/science.aad0501 (2016).

61 Tirosh, I. *et al.* Single-cell RNA-seq supports a developmental hierarchy in human oligodendroglioma. *Nature* **539**, 309-313, doi:10.1038/nature20123 (2016).

62 Cimetta, E. *et al.* Microfluidic bioreactor for dynamic regulation of early mesodermal commitment in human pluripotent stem cells. *Lab Chip* **13**, 355-364, doi:10.1039/c2lc40836h (2013).

63 Treutlein, B. *et al.* Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature* **509**, 371-375, doi:10.1038/nature13173 (2014).

64 Han, X. *et al.* Mapping the Mouse Cell Atlas by Microwell-Seq. *Cell* **172**, 1091-1107.e1017, doi:https://doi.org/10.1016/j.cell.2018.02.001 (2018).

65 Yuan, J. & Sims, P. A. An Automated Microwell Platform for Large-Scale Single Cell RNA-Seq. *Scientific Reports* **6**, doi:10.1038/srep33883 (2016).

66 Vol. 30 1-1 (2012).

67 Gierahn, T. M. *et al.* Seq-Well: Portable, low-cost rna sequencing of single cells at high throughput. *Nature Methods* **14**, 395-398, doi:10.1038/nmeth.4179 (2017).

68 Hochgerner, H. STRT-seq-2i: dual-index 5' single cell and nucleus RNA-seq on an addressable microwell array. *Sci. Rep.* **7** (2017).

69 Ramani, V. Massively multiplex single-cell Hi-C. *Nat. Methods* **14**, 1-6 (2017).

70 Vitak, S. A. Sequencing thousands of single-cell genomes with combinatorial indexing. *Nat. Methods* **14**, 302-308 (2017).

71 Jerby-Arnon, L. *et al.* A Cancer Cell Program Promotes T Cell Exclusion and Resistance to Checkpoint Blockade. *Cell* **175**, 984-997.e924, doi:10.1016/j.cell.2018.09.006 (2018).

72 Raj, A., van den Bogaard, P., Rifkin, S. A., van Oudenaarden, A. & Tyagi, S. Imaging individual mRNA molecules using multiple singly labeled probes. *Nature Methods* **5**, 877-879, doi:10.1038/nmeth.1253 (2008).

73 Prakadan, S. M., Shalek, A. K. & Weitz, D. A. Scaling by shrinking: empowering single-cell 'omics' with microfluidic devices. *Nat Rev Genet* **18**, 345-361, doi:10.1038/nrg.2017.15 (2017).

74 Elowitz, M. B., Levine, A. J., Siggia, E. D. & Swain, P. S. Stochastic gene expression in a single cell. *Science* **297**, 1183-1186, doi:10.1126/science.1070919 (2002).

75 Shalek, A. K. & Benson, M. Single-cell analyses to tailor treatments. *Science Translational Medicine* **9**, doi:10.1126/scitranslmed.aan4730 (2017).

76 Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e1821, doi:10.1016/j.cell.2019.05.031 (2019).

77 Hwang, B., Lee, J. H. & Bang, D. Vol. 50 (Nature Publishing Group, 2018).

78 The Next Quarter Century. *Immunity* **50**, 769-777, doi:https://doi.org/10.1016/j.immuni.2019.03.029 (2019).

79 Hughes, T. K. *et al.* Highly Efficient, Massively-Parallel Single-Cell RNA-Seq Reveals Cellular States and Molecular Features of Human Skin Pathology. *bioRxiv*, 689273, doi:10.1101/689273 (2019).

80 Bakken, T. E. *et al.* Single-nucleus and single-cell transcriptomes compared in matched cortical cell types. *PLOS ONE* **13**, e0209648, doi:10.1371/journal.pone.0209648 (2018).

81 Darrah, P. A. *et al.* Prevention of tuberculosis in macaques after intravenous BCG immunization. *Nature* **577**, 95-102, doi:10.1038/s41586-019-1817-8 (2020).

Chapter 2: Dissecting the Multicellular Ecosystem of Metastatic Melanoma by Single-Cell RNA-Seq

This chapter is adapted in accordance with AAAS's open access policy from the following article published in Science:

Itay Tirosh*, Benjamin Izar*, Sanjay M. Prakadan, Marc H. Wadsworth II, Daniel Treacy, John J. Trombetta, Diana Lu, Asaf Rotem, Christine Lian, George Murphy, Ofir Cohen, Eli van Allen, Monica Bertagnolli, Alex Genshaft, Travis K. Hughes, Carly G. K. Ziegler, Samuel W. Kazer, Aleth Gaillard, Kellie E. Kolb, Judit Valbuena, Charles Yoon, Orit Rozenblatt-Rosen, Alex K. Shalek, Aviv Regev and Levi Garraway, "Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq," *Science*, 352, (2016).

** Denotes equal authorship*

Abstract

Tumors are multicellular assemblies that encompass many distinct genotypic and phenotypic states. Here, we applied single-cell RNA-seq to examine 3,249 single cells isolated from 16 melanomas, profiling malignant, immune, stromal and endothelial cells. Malignant cells within the same tumor displayed transcriptional heterogeneity associated with the cell cycle, spatial context, and melanoma oncogenic programs. All tumors harbored malignant cells from two distinct transcriptional cell states, such that treatment-naïve "MITF-high" tumors also contained "AXL-high" tumor cells, which have been implicated in resistance to RAF/MEK inhibition. Distinct immune and stromal cell types, meanwhile, are associated with prognosis and suggest specific interactions between components of the melanoma microenvironment. Finally, an analysis of tumor-infiltrating T-cell sub-types reveals that co- inhibitory T-cell exhaustion genes may be regulated through both T-cell activation dependent and independent mechanisms. This work begins to unravel the cellular ecosystem of tumors and shows that single cell genomics may offer new insights with implications for both targeted and immune therapies.

INTRODUCTION

2.1 Background

Tumors are complex ecosystems defined by spatiotemporal interactions between heterogeneous cell types, including malignant, immune, and stromal cells.[1] Each tumor's cellular composition, as well as the interplay between these components, may exert critical roles in cancer development.[2] However, the specific components, their salient biological functions, and the means by which they collectively define tumor behavior remain incompletely characterized.

Tumor cellular diversity poses both challenges and opportunities for cancer therapy. This is exemplified by the varied clinical efficacy achieved in malignant melanoma with targeted therapies and immunotherapies. Immune checkpoint inhibitors can produce clinical responses in many patients with metastatic melanomas[3-7]; however, the genomic and molecular determinants of response to these agents remain incompletely understood. Although tumor neoantigens and PD-L1 expression clearly correlate with this response[8-10], it is likely that other factors from subsets of malignant cells, the microenvironment, and tumor-infiltrating lymphocytes (TILs) also play essential roles.[11]

Melanomas that harbor the BRAF^{V600E} (V600E: Val^{600} → Glu^{600}) mutation are commonly treated with inhibitors of rapidly accelerated fibrosarcoma kinase (RAF) and mitogen-activated protein kinase (MEK), before or after immune checkpoint inhibition. Although this regimen improves survival, virtually all tumors eventually develop resistance to these drugs.[12,13] Unfortunately, no targeted therapy currently exists for patients whose tumors lack BRAF mutations—including NRAS mutant tumors, those with inactivating NF1 mutations, or rarer events (such as RAF fusions). Collectively, these factors highlight the need for a deeper understanding of melanoma composition and its effect on the clinical course.

The next wave of therapeutic advances in cancer will probably be accelerated by technologies that assess the malignant, microenvironmental, and immunologic states most likely to inform treatment response and resistance. Ideally, we would be able to assess salient cellular heterogeneity by quantifying variation in oncogenic signaling pathways; drug-resistant tumor cell subsets; and the spectrum of immune, stromal, and other cell states that may inform immunotherapy response. Toward this end, single-cell genomic approaches enable detailed evaluation of genetic and transcriptional features present in hundreds to thousands of individual cells per tumor.[14-16] In principle, this approach may allow us to identify all major cellular components simultaneously, determine their individual genomic and molecular states[15], and ascertain which of these features may predict or explain clinical responses to anticancer agents. To explore this question, we used single-cell RNA sequencing (RNA-seq) to examine heterogeneities in malignant and nonmalignant cell types and states and to infer their possible drivers and interrelationships in the complex tumor cellular ecosystem.

RESULTS

2. 2 Profiling individual cells from patient-derived melanoma tumors

We measured single-cell RNA-seq profiles from 4,645 malignant, immune, and stromal cells isolated from 19 freshly procured human melanoma tumors that span a range of clinical and therapeutic backgrounds (**Appendix A, Table S1**). These included 10 metastases to lymphoid tissues (9 to lymph nodes and 1 to the spleen), 8 to distant sites (5 to subcutaneous or intramuscular tissue and 3 to the gastrointestinal tract), and one primary acral melanoma. Genotypic information was available for 17 of the 19 tumors, of which 4 had activating mutations in BRAF and 5 in NRAS oncogenes; eight patients had BRAF/NRAS wild-type melanomas (**Appendix A, Table S1**).

To isolate viable single cells that are suitable for high-quality single-cell RNA-seq, we developed and implemented a rapid translational workflow (**Figure 1A**).[15] We processed tumor tissues immediately after surgical procurement and generated single-cell suspensions within ~45 min, using an experimental protocol optimized to reduce artifactual transcriptional changes introduced by dis- aggregation, temperature, or time.[17] Once in suspension, we recovered individual viable immune ($CD45^{+}$) and nonimmune ($CD45^{-}$) cells (including malignant and stromal cells) by flow cytometry (fluorescence-activated cell sorting). Next, we pre- pared cDNA from the individual cells,

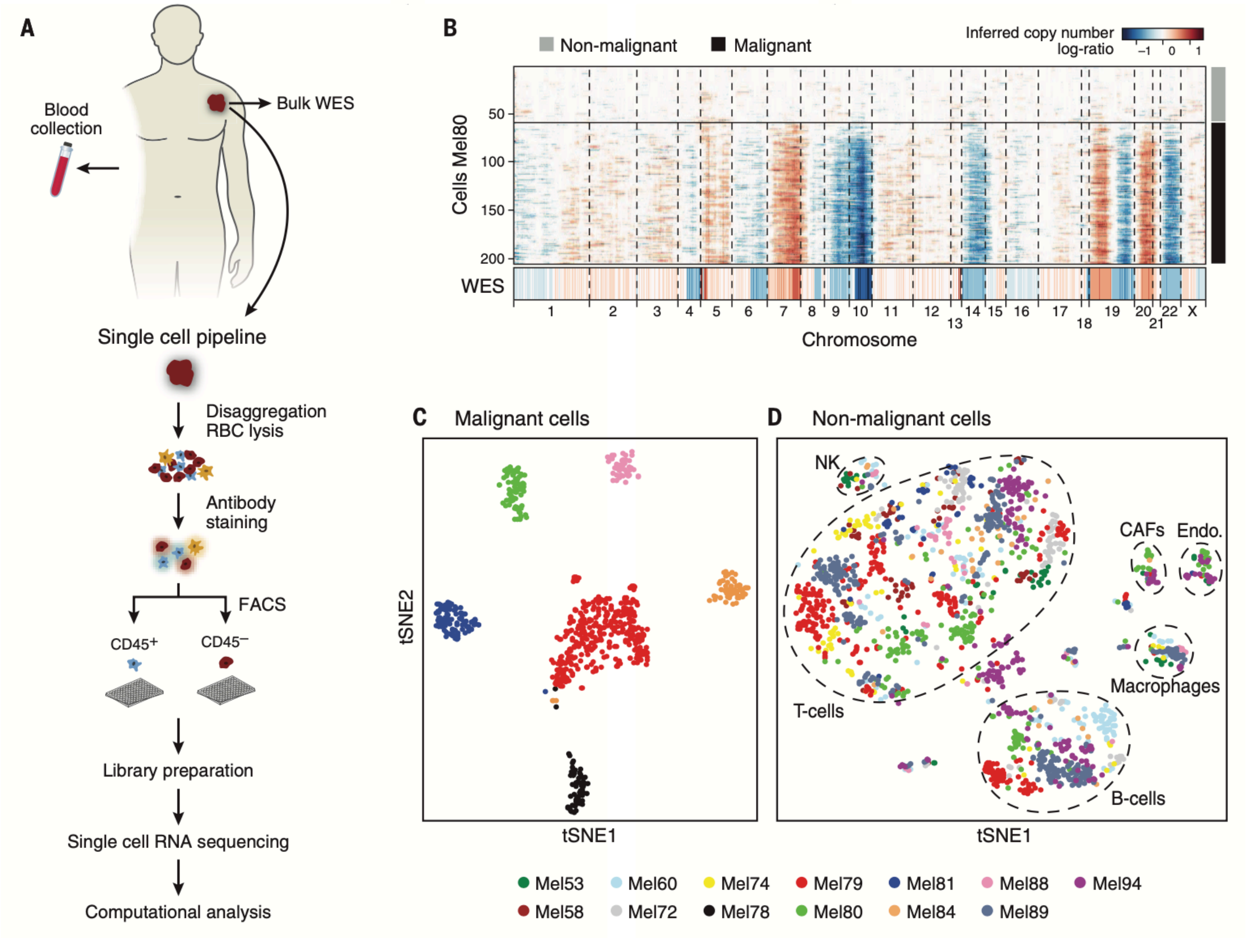


Figure 1 | Dissection of melanoma with single-cell RNA-seq. (A) Overview of workflow. WES, whole-exome sequencing; RBC, red blood cell; FACS, fluorescence-activated cell sorting. **(B)** Chromosomal landscape of inferred large-scale CNVs allows us to distinguish malignant from nonmalignant cells. The Mel80 tumor is shown with individual cells (y axis) and chromosomal regions (x axis). Amplifications (red) or deletions (blue) were inferred by averaging expression over 100-gene stretches on the respective chromosomes. Inferred CNVs are concordant with calls from WES (**bottom**). **(C and D)** Single-cell expression profiles allow us to distinguish malignant and nonmalignant cell types. Shown are t-SNE plots of malignant **(C)**, shown are the six tumors, each with >50 malignant cells] and nonmalignant **(D)** cells [as called from inferred CNVs as in **(B)**] from 11 tumors with >100 cells per tumor (see color code below the panels). Clusters of non- malignant cells [called by DBScan[17]] are marked by dashed ellipses and were annotated as T cells, B cells, macrophages, CAFs, and endothelial (Endo.) cells from preferentially expressed genes (**Appendix A, Figure S2 and tables S2 and S3**). NK, natural killer cells.

followed by library construction and massively parallel sequencing. The average number of mapped reads per cell was ~150,000[17], with a median library complexity of 4659 genes for malignant cells and 3438 genes for immune cells, comparable to previous studies of only malignant cells from fresh glioblastoma tumors.[15]

2.3 Single-cell transcriptome profiles distinguish cell states in malignant and nonmalignant cells

We used a multistep approach to distinguish the different cell types within melanoma tumors on the basis of both genetic and transcriptional states (**Figure 1, B to D**). First, we inferred large-scale copy number variations (CNVs) from expression profiles by averaging expression over stretches of 100 genes on their respective chromosomes[15] (**Figure 1B**). For each tumor, this approach revealed a common pattern of aneuploidy, which we validated in two tumors by bulk whole-exome sequencing (WES) (**Figure. 1B; Appendix B, Figure S1A**). Cells in which aneuploidy was inferred were classified as malignant cells (**Figure 1B; Appendix B, Figure S1**).

Second, we grouped the cells according to their expression profiles (**Figure 1, C and D**; **Appendix A, Figure S2**). To do this, we used nonlinear dimensionality reduction [t-distributed stochastic neighbor embedding (t-SNE)][18], followed by density clustering.[19] Generally, cells designated as malignant by CNV analysis formed a separate cluster for each tumor (**Figure 1C**), suggesting a high degree of intertumor heterogeneity. In contrast, the nonmalignant cells clustered by cell type (**Figure 1D; Appendix A, Figure S2**), independent of their tumor of origin and metastatic site (**Appendix A, Figure S3**). Clusters of nonmalignant cells were annotated as T cells, B cells, macrophages, endothelial cells, cancer-associated fibroblasts (CAFs), and natural killer cells on the basis of their preferentially or distinctively expressed marker genes (**Figure 1D; Appendix A, Figure S2 and tables S2 and S3**).

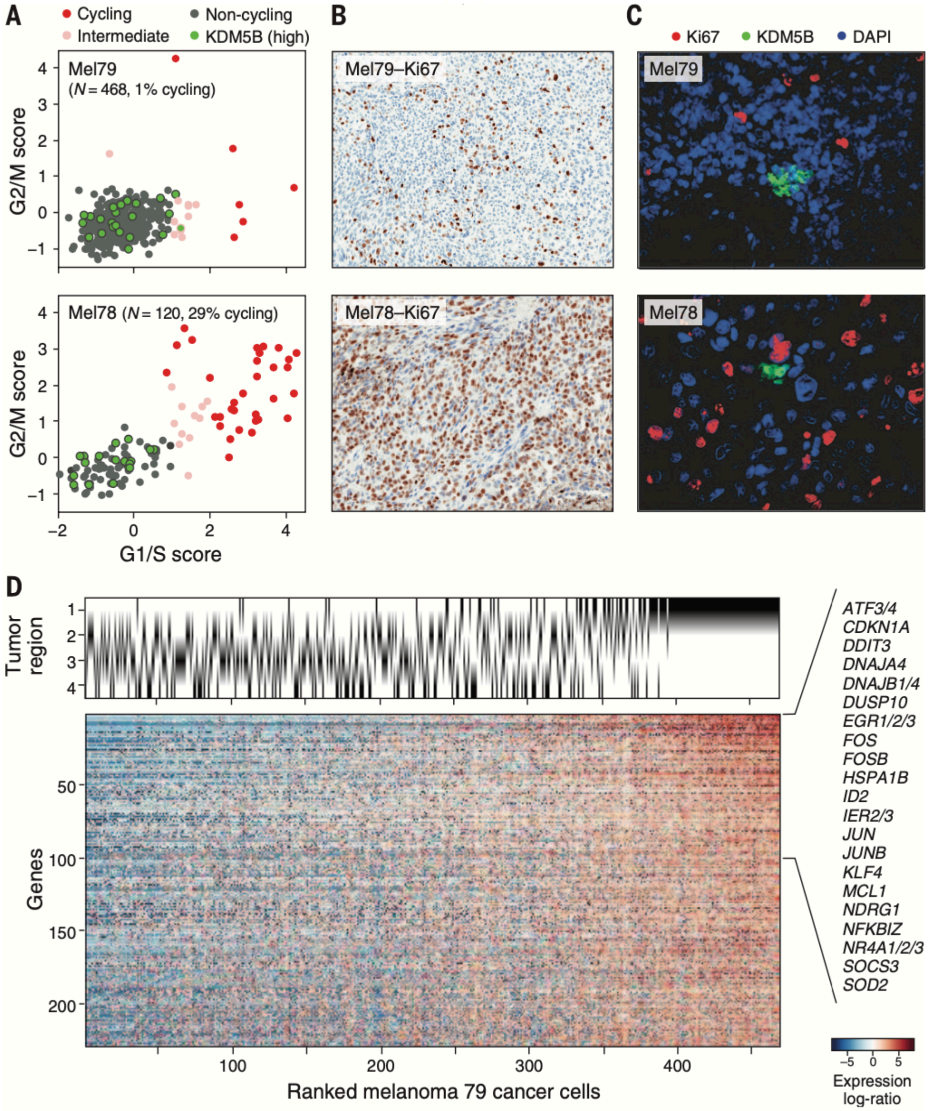


Figure 2 | Single-cell RNA-seq distinguishes cell cycle and other states among malignant cells. (A) Estimation of the cell cycle state of individual malignant cells (circles) on the basis of relative ex- pression of G1/S (x axis) and G2/M (y axis) gene sets in a low-cycling tumor (Mel79, top) and a high- cycling tumor (Mel78, bottom). Cells are colored by their inferred cell cycle states: cycling cells, red; intermediate, pink; and noncycling cells, gray. Cells with high expression of KDM5B (z score > 2) are shown in green. N denotes number of cells. **(B)** Immunohistochemistry staining (40X magnification) for Ki67+ cells shows concordance with the signature-based frequency of cycling cells for Mel79 and Mel78 (as for other tumors; fig S4C). **(C)** KDM5B and Ki67 staining (40X magnification) in corresponding tissue showing small clusters of KDM5B-high expressing cells negative for Ki67 (Appendix A, Figure S4). DAPI, 4′,6-diamidino-2-phenylindole. **(D)** An expression program specific to region one of Mel79, identified on the basis of multifocal sampling. The relative expression of genes (rows) is shown for cells (columns) ordered by the average expression of the entire gene set. The region of origin of each cell is indicated in the top panel (**Appendix A, Figure S5**).

2.4 Analysis of malignant cells reveals heterogeneity in cell cycle and spatial organization

We next used unbiased analyses of the individual malignant cells to identify biologically relevant melanoma cell states. After controlling for inter- tumor differences[17], we examined the six top components from a principal component analysis (PCA) (**Appendix A, table S4**). The first component correlated highly with the number of genes detected per cell and probably reflects technical aspects, whereas the other five significant principal components highlighted biological variability.

The second component (PC2) was associated with the expression of cell cycle genes (Gene Ontology project: “cell cycle” $P < 10^{-16}$; hyper- geometric test). To characterize cycling cells more precisely, we used gene signatures that have previously been shown to denote G1/S or G2/M phases in both synchronization[20] and single-cell[16] experiments in cell lines. Cell cycle phase–specific signatures were highly expressed in a subset of malignant cells, distinguishing cycling cells from noncycling cells (**Figure 2A; Appendix A, Figure S4A**). These signatures revealed variability in the fraction of cycling cells across tumors (13.5% on average, ±13 SD) (**Appendix A, Figure S4B**), allowing us to designate both low-cycling (1 to 3%, e.g., Mel79) and high-cycling tumors (20 to 30%, e.g., Mel78), consistent with Ki67+ staining results (**Figure 2B; Appendix A, Figure S4C**).

A core set of cell cycle genes was induced (**Appendix A, Figure S4D, red dots; and table S5**) in both low-cycling and high-cycling tumors, with one notable exception: cyclin D3 (CCND3), which was induced in cycling cells only in high-cycling tumors (**Appendix A, Figure S4D**). In contrast, KDM5B (JARID1B) showed the strongest association with noncycling cells (**Figure 2A,** green dots), mirroring Patel et al.’s findings in glioblastoma (15). KDM5B encodes a H3K4 histone demethylase associated with a subpopulation of slow-cycling and drug-resistant melanoma stemlike cells[21,22] in mouse models. Immunofluorescence (IF) staining validated the presence and mutually exclusive expression of KDM5B and Ki67. KDM5B-expressing cells were grouped in small clusters, consistent with observations in mouse and in vitro models[21] (**Figure 2C; Appendix A, Figure S4E**).

Two principal components (PC3 and PC6) primarily segregated different malignant cells from one treatment-naïve tumor (Mel79). In this tumor, we analyzed 468 malignant cells from four distinct regions after surgical resection (**Appendix A, Figure S5A**). We identified 229 genes with higher expression in the malignant cells of region one compared with those of other tumor regions [**Figure 2D**, false discovery rate (FDR) < 0.05; and **Appendix A, table S6**]. A similar expression program was found in T cells from region one (**Appendix A, Figure S6 and table S6**), suggesting a spatial effect that influences multiple cell types. The genes with the highest preferential expression in region one are also generally coexpressed across melanoma tumors profiled in bulk in The Cancer Genome Atlas (TCGA)[23] (**Appendix A, Figure S6**). Many of these genes encode immediate early-activation transcription factors linked to inflammation, stress responses, and a melanoma oncogenic pro- gram (e.g., ATF3, FOS, FOSB, JUN, JUNB). Several of these transcription factors (e.g., FOS, JUN, NR4A1/2) are regulated by cyclic adenosine monophosphate (cAMP) and cAMP response element–binding protein signaling, which has been implicated as a mitogen-activated protein kinase (MAPK)–independent resistance module in BRAF-mutant melanomas treated with RAF and MEK inhibition.[24] Other top genes differentially up-regulated in region one included those involved in survival (MCL1), stress responses (EGR1/2/3, NDRG, HSPA1B), and NF-kB signaling (NFKBIZ), which has also been associated with resistance to RAF and MEK inhibition.[25] Immunohistochemistry analysis confirmed the elevated NF-kB and JunB levels in cells of region one compared with cells in the other regions of this tumor (**Appendix A, Figure S5B**).

2.5 Heterogeneity in the abundance of a dormant, drug-resistant melanoma subpopulation

Collectively, the above observations imply that pretreatment melanoma tumors may harbor subsets of malignant cells that are less likely to respond to targeted therapy. The transcriptional programs associated with principal components PC4 and PC5 were highly correlated with expression of the MITF gene (microphthalmia-associated transcription factor), which encodes the master melanocyte transcriptional regulator and a melanoma lineage-survival oncogene.[26] Scoring genes by their correlation to MITF across single cells, we identified a "MITF-high" program consisting of MITF itself and several MITF target genes, including TYR, PMEL, and MLANA (**Appendix A, table S7**). A second transcriptional program, negatively correlated with the MITF program and with PC4 and PC5 (Pearson correlation $P < 10^{-24}$), included AXL and NGFR (p75NTR), a marker of resistance to various targeted therapies[27,28] and a putative melanoma cancer stem cell marker[29], respectively (**Appendix A, table S8**). Thus, these transcriptional programs resemble reported[25,30-32] MITF-high, as well as MITF-low and AXL-high ("AXL-high"), transcriptional profiles that can distinguish melanoma tumors, cell lines, and mouse models. Notably, the AXL-high program has been linked to intrinsic resistance to RAF and MEK inhibition.[25,30,31]

Although at the bulk tumor level each melanoma could be classified as MITF-high or AXL-high (**Figure 3A**), at the single-cell level every tumor contained malignant cells corresponding to both transcriptional states. Using single-cell RNA-seq to examine each cell's expression of the MITF and AXL gene sets, we observed that MITF-high tumors, including treatment-naïve melanomas, harbored a subpopulation of AXL-high melanoma cells that was undetectable through bulk analysis, and vice versa (**Figure 3B**). The malignant cells thus spanned the continuum between AXL- high and MITF-high states in every investigated tumor (**Figure 3B; Appendix A, Figure S7**). We performed IF staining to further validate the mutually exclusive expression of the MITF-high and AXL-high programs in cells from the same bulk tumors (**Figure 3C; Appendix A, Figure S8**).

Because malignant cells with AXL-high and MITF-high transcriptional states coexist in melanoma, we hypothesized that treatment with RAF and MEK inhibitors would increase the prevalence of AXL-high cells after the development of drug resistance. To test this, we analyzed RNA-seq data from a cohort[13] of six paired $BRAF^{V600E}$ melanoma biopsies taken before treatment and after resistance to single-agent RAF inhibition (vemurafenib; 1 patient) or combined RAF and MEK inhibition (dabrafenib and trametinib; 5 patients), respectively (**Appendix A, tables S9 and S10**). We ranked the 12 transcriptomes on the basis of the relative expression of all genes in the AXL-high program compared with those in the MITF-high program. In each pair, we observed a shift toward the AXL-high program in the drug-resistant sample [**Figure 3D**; $P < 0.05$ for same effect in six of six paired samples, binomial test; $P < 0.05$ for four of six individual paired-sample comparisons shown by black arrows[17]]. RNA-seq data from an independent cohort[33] also showed that a subset of drug-resistant samples exhibited increased expression of the AXL program (**Appendix A, Figure S9**). Other genes previously implicated in resistance to RAF and MEK inhibition were also increased in a subset of the drug-resistant samples. PDGFRB (platelet-derived growth factor receptor β)[34] was up-regulated in a similar subset as the AXL program, whereas MET[33] was up-regulated in a mutually exclusive subset (**Appendix A, Figure S9**), suggesting that AXL and MET may reflect distinct drug-resistant states.

To further assess the connection between the AXL program and resistance to RAF and MEK inhibition, we studied single-cell AXL expression in 18 melanoma cell lines from the Cancer Cell Line Encyclopedia[35] (**Appendix A, table S11**). Flow cytometry analysis revealed a wide distribution of the proportion of AXL-positive cells, from <1 to 99% per cell line, which correlated with bulk mRNA levels and was inversely associated with sensitivity to small-molecule RAF inhibition (**Appendix A, table S11**).

We treated 10 cell lines[17] with increasing doses of a combination of RAF and MEK inhibitors (dabrafenib and trametinib) and found an increase in the proportion of AXL-positive cells in 6 cell lines initially composed of a small (<3%) pretreatment AXL-positive population (**Appendix A, Figure S10A**). In contrast, cell lines with an intrinsically high proportion of AXL expression showed modest or no changes (**Appendix A, Figure S10B**). We obtained similar results by multiplexed quantitative single-cell IF, which also demonstrated that the increased fraction of AXL-positive cells after inhibition of RAF and MEK is associated with rapid decreases in extracellular signal–regulated kinase (ERK) phosphorylation (reflecting MAP kinase signaling inhibition) (**Figure 3E; Appendix A, Figures S11 and S12**). In summary, both melanoma tumors and cell lines demonstrate drug-resistant tumor cell sub- populations that precede treatment and become enriched after MAP kinase–targeted treatment.

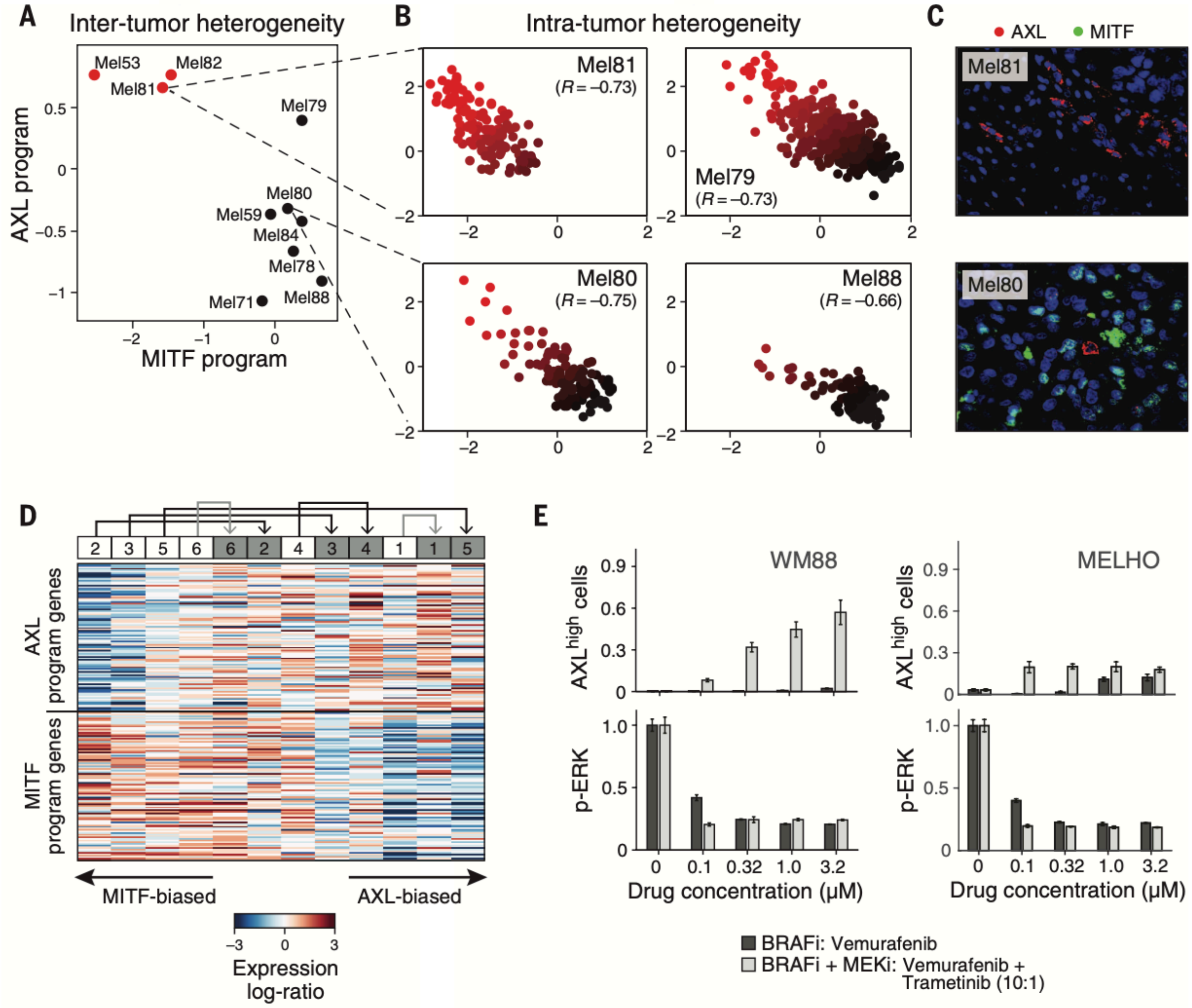


Figure 3 | MITF- and AXL-associated expression programs vary between andwithin tumors, as well as after treatment. (**A**) Average expression signatures for the AXL program (y axis) or the MITF program (x axis) stratify tumors into MITF- high (black) or AXL-high (red) categories. (**B**) Single-cell profiles show a negative correlation between the AXL program (y axis) and the MITF program (x axis) across individual malignant cells within the same tumor. Cells are colored by the relative expression of the MITF (black) and AXL (red) programs. Cells in both states are found in all examined tumors, including three tumors (Mel79, Mel80, and Mel81) without prior systemic treatment, indicating that dormant resistant (AXL-high) cells may be present in treatment-naïve patients. (**C**) Mel81 and Mel80 IF staining of MITF (green nuclei) and AXL (red), validating the mutual exclusivity among individual cells within the same tumor (**Appendix A, Figure S8**). (**D**) Relative expression (centered) of the AXL program genes (**top**) and MITF program genes (**bottom**) in six matched pretreatment (**white boxes**) and post relapse (gray boxes) samples from patients who progressed through therapeutic RAF and MEK inhibition. Numbers at the top indicate patient index. Samples are sorted by the average relative expression of the AXL versus MITF gene sets. In all cases, the relapsed samples had an increased ratio of AXL-to-MITF expression compared with their pretreatment counterparts. This consistent shift of all six patients is statistically significant (P < 0.05, binomial test), as are the individual increases in AXL and MITF for four of the six sample pairs (P < 0.05, t test; black and gray arrows denote increases that are individually significant or nonsignificant, respectively). (**E**) Quantitative, multiplex single-cell IF for AXL expression (**top y axes**) and MAP kinase pathway inhibition (p-ERK levels, **bottom y axes**) in the example cell lines WM88 and MELHO treated with increasing concentrations (x axis) of either a RAF inhibitor alone (dark gray bars) or a combination of RAF and MEK inhibitors (light gray bars). We observed an increasing fraction of AXL-high cells (**top panels**) as well as a dose-dependent decrease of p-ERK (bottom panels). (**Appendix A, Figures S11 and S12** show the results for additional cell lines).

2.6 Nonmalignant cells and their interactions within the melanoma microenvironment

Various nonmalignant cells make up the tumor microenvironment. The composition of the microenvironment has an important effect on tumorigenesis and also in the modulation of treatment responses.[1] Tumor infiltration with T cells, for example, is predictive for the response to immune checkpoint inhibitors in various cancer types.[36]

To resolve the composition of the melanoma microenvironment, we used our single-cell RNA-seq profiles to define distinct expression signatures of each of five distinct nonmalignant cell types: T cells, B cells, macrophages, endothelial cells, and CAFs. Because our signatures were derived from single-cell profiles, we could avoid confounders and ensure that each signature is determined by cell type–specific profiles.[17] Next, we used these signatures to infer the relative abundance of those cell types in a larger compendium of tumors[17] (**Figure 4A; Appendix A, Figure S13**). We found a strong correlation (correlation coefficient R ~ 0.8) between our estimated tumor purity and that predicted from DNA analysis[37] (**Figure 4A, first lane below the heat map**).

We partitioned 471 tumors from TCGA into 10 distinct microenvironment clusters on the basis of their inferred cell type composition (**Figure 4A**). Clusters were mostly independent of the site of metastasis (**Figure 4A, second lane**), with some exceptions (e.g., clusters 8 and 9). Next we examined how these different microenvironments may relate to the phenotype of the malignant cells. In particular, CAF abundance is predictive of the AXL-MITF distinction, with CAF-rich tumors expressing the AXL-high signature (**Figure 4A, bottom lane**). Interestingly, an AXL-high program was expressed by both melanoma cells and CAFs. However, we distinguished AXL-high genes that are preferentially expressed by melanoma cells (“melanoma-derived AXL program”) from those that are preferentially expressed by CAFs (“CAF- derived AXL program”). Both sets of genes were correlated with the inferred CAF abundance in tumors from TCGA (**Appendix A, Figure S14**).[38] Furthermore, the MITF-high program, which is specific to melanoma cells, was negatively correlated with inferred CAF abundance. Taken together, these results suggest that CAF abundance may be linked to preferential expression of the AXL-high over the MITF-high program by melanoma cells. Thus it is possible that specific tumor-CAF interactions may shape the melanoma cell transcriptome.

Interactions between cells play crucial roles in the tumor microenvironment.[1] To assess how cell-to-cell interactions may influence tumor composition, we searched for genes expressed by cells of one type that may influence or reflect the proportion of cells of a different type in the tumor (**Appendix A, Figure S15**). For example, we searched for genes expressed primarily by CAFs (but not T cells) in single-cell data that correlated with T cell abundance (as inferred by T cell–specific genes) in bulk tumor tissue from the TCGA data set.[23] We identified a set of CAF-expressed genes that correlated strongly with T cell infiltration (**Figure 4B,** red circles). These included known chemotactic (CXCL12 and CCL19) and immune-modulating (PD-L2) genes, which are expressed by both CAFs and macrophages (**Appendix A, Figure S16**). A separate set of genes, exclusively expressed by CAFs, that correlated with T cell infiltration (fig. S16) included multiple complement factors [C1S, C1R, C3, C4A, CFB, and C1NH (SERPING1)]. Notably, these complement genes were specifically expressed by fresh- ly isolated CAFs but not by cultured CAFs (**Appendix A, Figure S17**) or macrophages (**Appendix A, Figure S16**). These findings are intriguing, as studies have implicated complement activity in the recruitment and modulation of T cell–mediated antitumor immune responses [in addition to their role in augmenting innate immunity[39]].

We validated a high correlation (R > 0.8) be- tween complement factor 3 (C3) levels (one of the CAF-expressed complement genes) and infiltration of CD8+ T cells. We performed dual IF staining and quantitative slide analysis of two tissue microarrays with a total of 308 core biopsies, including primary tumors, metastatic lesions, normal skin with adjacent tumor, and healthy skin controls (**Figure 4C; Appendix A, Figure S18**).[17] To test the generalizability of the association between CAF-derived complement factors with T cell in- filtration, we expanded our analysis to bulk RNA- seq data sets across all TCGA cancer types (**Figure 4D**). Consistent with the results in melanoma, complement factors correlated with inferred T cell abundance in many cancer types and more highly than in normal tissues (e.g., R > 0.4 for 65% of cancer types but only for 14% of normal tissue types). Although correlation analysis can- not determine causality, this indicates a potential in vivo role for cell-to-cell interactions.

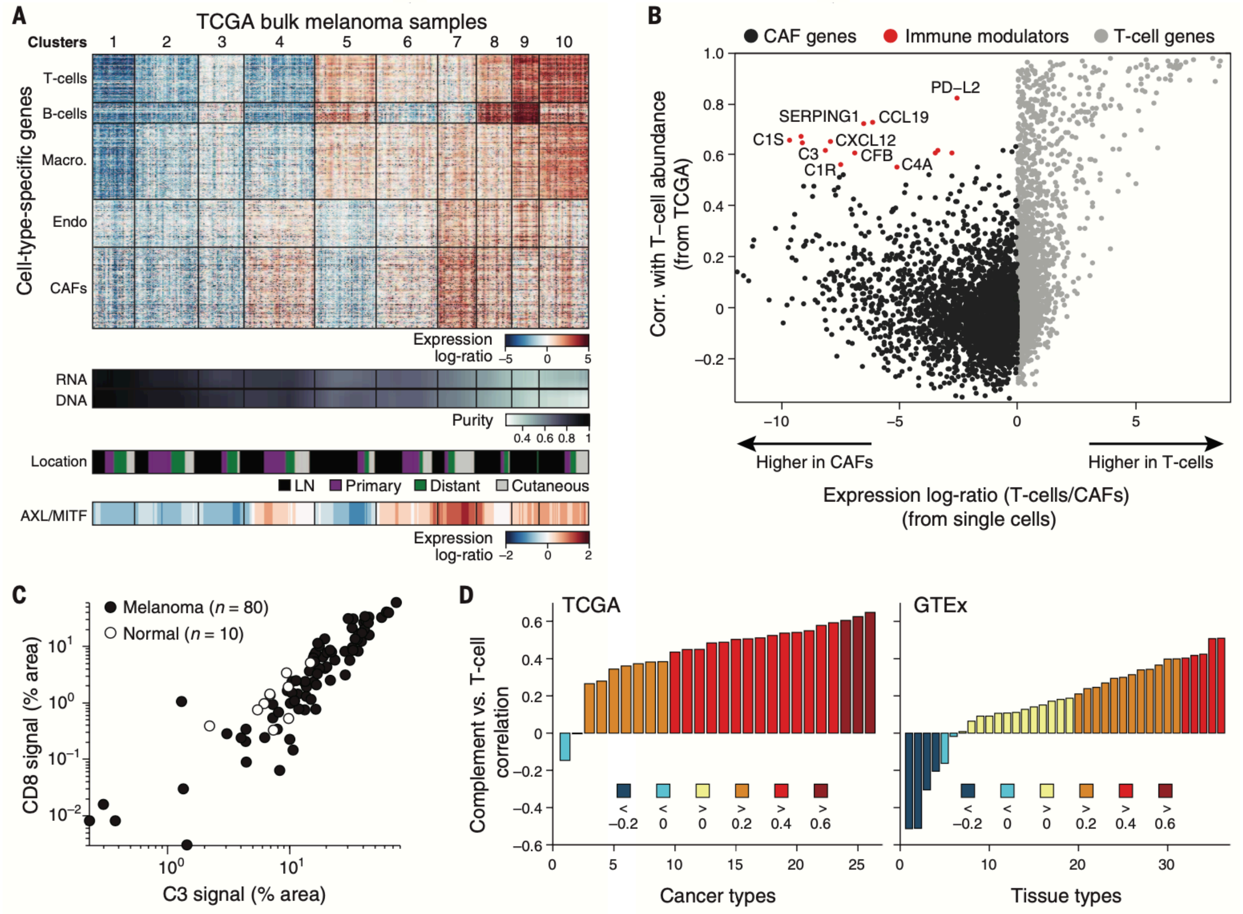


Figure 4 | Deconvolution of bulk melanoma profiles reveals cell-to-cell interactions. (A) Bulk tumors segregate to distinct clusters on the basis of their inferred cell type composition. (**Top panel**) Heat map showing the relative expression of gene sets defined from single-cell RNA-seq, as specific to each of five cell types from the tumor microenvironment (y axis) across 471 melanoma TCGA bulk-RNA signatures (x axis). Each column represents one tumor, and tumors are partitioned into 10 distinct patterns identified by k-means clustering (**vertical lines and cluster numbers at the top**). Endo, endothelial cells; Macro., macrophages. (**Lower panels, from top to bottom**) Tumor purity estimated by ABSOLUTE (DNA) and RNA-seq analysis (RNA), specimen location (from TCGA), and AXL/MITF scores. Tumors with a high abundance of CAFs are correlated with an increased ratio of AXL-to-MITF expression (**bottom**). LN, lymph node. (**B**) Inferred cell-to-cell interactions between CAFs and T cells. The scatter plot compares, for each gene (circle), the correlation of its expression with inferred T cell abundance across bulk tumors (y axis, from TCGA transcriptomes) to the specificity of its expression in CAFs (black) versus T cells (gray) (x axis, based on single-cell transcriptomes). Genes that are highly specific to CAFs in a single-cell analysis of tumors but are also associated with high T cell abundance in bulk tumors (red) are candidates for interaction between CAF cells and T cells. (**C**) Of the 90 samples, 80 tumor specimens (black dots) show a correlation (R = 0.86) between C3 and CD8 signals, as analyzed by quantitative IF. Ten normal control specimens (gray dots) are also shown (**Appendix A, Figure S18, A to F**, shows normalization and additional specimens). **(D)** Correlation coefficient (y axis) between the average expression of CAF-derived complement factors shown in (B) and that of T cell markers (CD3/D/E/G, CD8A/B) across 26 TCGA cancer types with >100 samples (**x axis, left panel**) and across 36 GTEx (Genotype-Tissue Expression Project) tissue types with >100 samples (**x axis, right panel**). Bars are colored on the basis of correlation ranges, as indicated at the bottom.

2.7 Diversity of tumor-infiltrating T lymphocytes and their functional states

The activity of TILs, particularly CD8$^+$ T cells, is a major determinant of successful immune surveillance. Under normal circumstances, effector CD8$^+$ T cells exposed to antigens and costimulatory factors may mediate lysis of malignant cells and control tumor growth. However, this function can be hampered by tumor-mediated T cell exhaustion, such that T cells fail to activate cytotoxic effector functions.[40] Exhaustion is promoted through the stimulation of coinhibitory checkpoint molecules on the T cell surface (PD-1, TIM-3, CTLA-4, TIGIT, LAG3, and others)[41]; blockade of checkpoint mechanisms has shown clinical benefit in subsets of melanoma and other malignancies.[3,10,42,43] Although checkpoint ligand expression (e.g., PD-L1) and neoantigen load clearly contribute[9,44,45], no biomarker has emerged that reliably predicts the clinical response to immune checkpoint blockade. We reasoned that single-cell analyses might yield features to elucidate response determinants and possibly identify new immunotherapy targets.

Thus, we analyzed the single-cell expression patterns of 2068 T cells from 15 melanomas. We identified T cells and their main subsets [CD4+, regulatory T cells (Tregs), and CD8+] on the basis of the expression levels of their respective defining surface markers (**Figure 5A, top; Appendix A, table S12**). Within both the CD4$^+$ and CD8+ populations, a PCA distinguished cell subsets and heterogeneity of activation states on the basis of the expression of naïve and cytotoxic T cell genes (**Figure 5, A and B; Appendix A, Figure S19**).

Next we sought to determine the exhaustion status of each cell from the expression of key coinhibitory receptors (PD1, TIGIT, TIM3, LAG3, and CTLA4). In several cases, these coinhibitory receptors were coexpressed across individual cells; we validated this phenomenon for PD1 and TIM3 by IF staining (**Figure 5C**). However, exhaustion gene expression was also highly correlated with the expression of both cytotoxicity markers and overall T cell activation states (**Figure 5B**). This observation resembles an activation-dependent exhaustion expression program, such as those reported previously.[46,47] Accordingly, expression of coinhibitory receptors (alone or in combinations) may not be

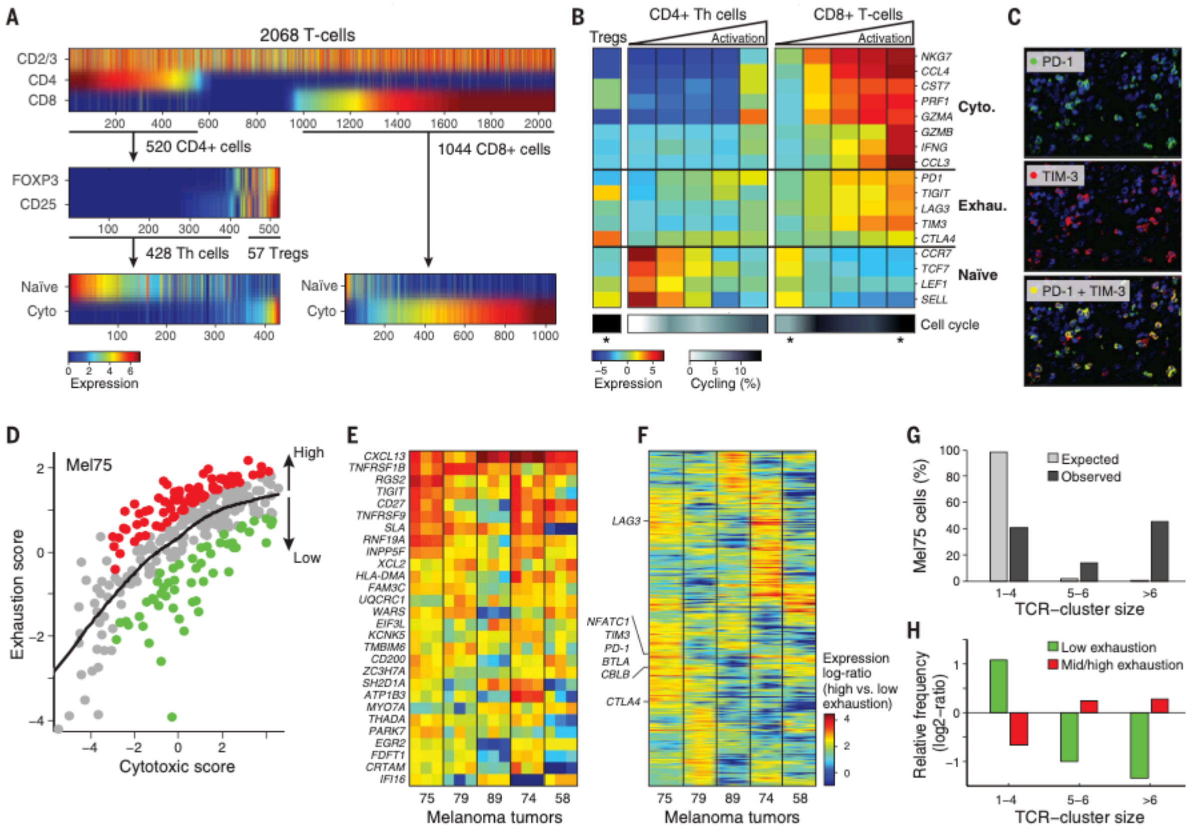


Figure 5 | Activation-dependent and -independent variation in T cell exhaustion markers. (A) Single T cell stratification into CD4$^+$ and CD8$^+$ cells (**top panel**), CD25$^+$FOXP3$^+$ and other CD4 cells (**middle panel**), and their inferred activation state [**lower panels**, from average expression of the cytotoxic and naïve gene sets in **(B)**]. Th, T helper cells; Tregs; regulatory T cells. **(B)** (**Top**) Average expression of markers of cytotoxicity (Cyto.), exhaustion (Exhau.), and naïve cell states (**rows**) in (**from left to right**) Tregs, CD4$^+$ T helper cells, and CD8$^+$ T cells. CD4+ and CD8+ T cells are each further divided into five bins by their cytotoxic score (ratio of cytotoxic to naïve marker expression levels), showing activation-dependent coexpression of exhaustion markers. (**Bottom**) Proportion of cycling cells (calculated as in **Figure 2B**). Asterisks denote significant enrichment or depletion of cycling cells in a specific subset, as compared with the corresponding set of CD4+ or CD8+ T cells ($P < 0.05$, hypergeometric test). **(C)** IF analysis of PD-1 (**top**, green), TIM-3 (**middle**, red), and their overlay (**bottom**) validates their coexpression. **(D)** Activation- independent variation in exhaustion states within highly cytotoxic T cells. The scatter plot shows the cytotoxic score (**x axis**) and exhaustion score (**y axis**, average expression of the Mel75 exhaustion program as in **Appendix A, Figure S21**) of each CD8+ T cell from Mel75. In addition to the overall correlation between cytotoxicity and exhaustion, the cytotoxic cells can be subdivided into cells with high (red) and low exhaustion (green), based on comparison to a LOWESS (locally weighted scatter plot smoothing) regression (black line). **(E and F)** Relative expression ($\log_2$ fold-change) in high- versus low-exhaustion cytotoxic CD8+ T cells from five tumors (x axis), including 28 genes that were significantly up-regulated ($P < 0.05$, permutation test) in high-exhaustion cells across most tumors **(E)** and 272 genes that were variably associated with high-exhaustion cells across tumors (F). Three independently derived exhaustion gene sets were used to define high- and low-exhaustion cells (Mel75)[17,46,48], and the corresponding results are represented as distinct columns for each tumor. **(G)** Expanded TCR clones. Cells were as- signed to clusters of TCR segment usage (dark gray bars) (**Appendix A, Figure S23**), and cluster size (**x axis**) was evaluated for significance by control analysis in which TCR segments were shuffled across cells (light gray bars). The percentage of Mel75 cells (**y axis**) is shown for clusters of small size (one to four cells) that probably represent nonexpanded cells, medium size (five or six cells) that may reflect expanded

(*continues onto the next page*)

clones (FDR = 0.12), and large size (more than six cells) that most likely reflect expanded clones (FDR = 0.005). **(H)** Expanded clones are depleted of nonexhausted cells and enriched for exhausted cells. Mel75 cells were divided according to exhaustion score into categories of low exhaustion (green, bottom 25% of cells) and medium-to-high exhaustion (red, top 75%). Shown is the relative frequency of these exhaustion subsets (y axis) in each TCR-cluster group [x axis, as defined in **(G)**], defined as the log2 ratio of the frequency in that group compared with the frequency across all Mel75 cells. All values were significant (P < 10,, binomial test).

sufficient by itself to characterize the salient functional state of tumor-associated T lymphocytes in situ or to distinguish exhaustion from activation.

To define an activation-independent exhaustion program, we leveraged single-cell data from CD8+ T cells sequenced in a single tumor (Mel75, 314 cells). These data allowed cytotoxic and exhaustion programs to be deconvolved. Specifically, PCA of Mel75 T cell transcriptomes identified a robust expression module that consisted of all five coinhibitory receptors and other exhaustion-related genes, but not cytotoxicity genes (**Appendix A, Figure S21 and table S13**).

We used the Mel75 exhaustion program, along with previously published exhaustion programs[46,48], to estimate the exhaustion state of each cell. An exhaustion state was defined as high or low expression of the exhaustion program relative to that of the cytotoxicity genes (**Figure 5D**).[17] Accordingly, we defined exhaustion states in Mel75 and in four additional tumors with the highest number of CD8+ T cells (68 to 214 cells per tumor). We identified the top preferentially expressed genes in high-exhaustion cells compared with low-exhaustion cells (both defined relative to the expression of cytotoxicity genes). This allowed us to define a core exhaustion signature across cells from various tumors.

Our core exhaustion signature yielded 28 genes that were consistently up-regulated in high-exhaustion cells of most tumors, including coinhibitory (TIGIT) and costimulatory (TNFRSF9/ 4-1BB and CD27) receptors (**Figure 5E; Appendix A, table S14**). In addition, most genes that were significantly up-regulated in high-exhaustion cells of at least one tumor had distinct associations with exhaustion across the different tumors (**Figure 5F**, 272 of 300 genes with P < 0.001 by permutation test; **Appendix A, Figure S22, A and B; and table S14**). These tumor- specific signatures included variable expression of known exhaustion markers (**Appendix A, table S14**) and could be linked to immunotherapeutic response or might reflect the effects of previous treatments. For example, CTLA-4 was highly up-regulated in exhausted cells of Mel75 and weakly up-regulated in three other tumors but was completely decoupled from exhaustion in Mel58. Interestingly, Mel58 was derived from a patient with an initial response and subsequent development of resistance to CTLA-4 blockade with ipilimumab (**Figure 5F; Appendix A, Figure S22C**). Another variable gene of interest was the transcription factor NFATC1, previously implicated in T cell exhaustion.[49] NFATC1 and its target genes were preferentially associated with the activation-independent exhaustion phenotype in Mel75 (**Appendix A, Figure S22, D and E**), suggesting a potential role of NFATC1 in the underlying variability of exhaustion programs among patients.

Finally, we explored the relationship between T cell states and clonal expansion. T cells that recognize tumor antigens may proliferate to generate discernible clonal subpopulations defined by an identical T cell receptor (TCR) sequence.[50] To identify potentially expanded T cell clones, we used RNA-seq reads that map to the TCR to classify single T cells by their isoforms of the V and J segments of the a and b TCR chains, and we searched for enriched combinations of TCR segments. Most observed combinations were found in few cells and were not enriched. However, approximately half of the CD8+ T cells in Mel75 had one of the seven enriched combinations identified (FDR = 0.005) and thus may represent expanded T cell clones (**Figure 5G; Appendix A, Figure S23**). This putative T cell expansion was also linked to exhaustion (**Figure 5H**), such that low-exhaustion T cells were depleted in expanded T cells (TCR clusters with more than six cells) and enriched in nonexpanded T cells (TCR clusters with one to four cells). In particular, the nonexhausted cytotoxic cells are almost all non- expanded cells (**Figure 5H**). Overall, this analysis suggests that single-cell RNA-seq may allow for the inference of functionally variable T cell populations that are not detectable with other pro-filing approaches (**Appendix A, Figure S24**). This knowledge may empower studies of tumor response and resistance to immune checkpoint inhibitors.

2.8 Conclusions

Our analysis has uncovered intra- and interindividual, spatial, functional, and genomic heterogeneity in melanoma cells and associated tumor components that shape the microenvironment, including immune cells, CAFs, and endothelial cells. We identified a cell state in a subpopulation of all melanomas studied that is linked to resistance to targeted therapies, and we used a variety of approaches to validate the presence of a dormant drug-resistant population in a number of melanoma cell lines.

By leveraging single-cell profiles from a few tumors to deconvolve a large collection of bulk profiles from TCGA, we discovered different micro- environments associated with distinct malignant cell profiles. We also detected a subset of genes expressed by one cell type (e.g., CAFs) that may influence the proportion of other cell types (e.g., T cells); this indicates the importance of intercellular communication for tumor phenotype. Putative interactions between stromal-derived factors and immune cell abundance in melanoma core biopsies suggest that future diagnostic and therapeutic strategies should account for tumor cell composition rather than bulk expression. Furthermore, our data suggest potential bio- markers for distinguishing exhausted and cytotoxic T cells that may aid in selecting patients for immune checkpoint blockade.

Although future work is necessary to clarify the interplay between these cell types and functional states in space and time, the ability to carry out a number of highly multiplexed single- cell observations within a tumor allows us to identify meaningful cell subpopulations and gene expression programs that may inform both the analysis of bulk transcriptional data and precision treatment strategies. Conceivably, single-cell genomic profiling may soon enable a deeper understanding of the complex interplay among cells within the tumor ecosystem and its evolution in response to treatment, thereby providing a versatile new tool for future translational applications.

2.9 References

1 Hanahan, D. & Weinberg, R. A. Vol. 144 646-674 (2011).

2 Meacham, C. E. & Morrison, S. J. Vol. 501 328-337 (2013).

3 Hodi, F. S. *et al.* Improved survival with ipilimumab in patients with metastatic melanoma. *New England Journal of Medicine* **363**, 711-723, doi:10.1056/NEJMoa1003466 (2010).

4 Brahmer, J. R. *et al.* Safety and activity of anti-PD-L1 antibody in patients with advanced cancer. *New England Journal of Medicine* **366**, 2455-2465, doi:10.1056/NEJMoa1200694 (2012).

5 Brahmer, J. R. *et al.* Phase I study of single-agent anti-programmed death-1 (MDX-1106) in refractory solid tumors: Safety, clinical activity, pharmacodynamics, and immunologic correlates. *Journal of Clinical Oncology* **28**, 3167-3175, doi:10.1200/JCO.2009.26.7609 (2010).

6 Topalian, S. L. *et al.* Safety, activity, and immune correlates of anti-PD-1 antibody in cancer. *New England Journal of Medicine* **366**, 2443-2454, doi:10.1056/NEJMoa1200690 (2012).

7 Hamid, O. *et al.* Safety and tumor responses with lambrolizumab (anti-PD-1) in melanoma. *New England Journal of Medicine* **369**, 134-144, doi:10.1056/NEJMoa1305133 (2013).

8 Weber, J. S. *et al.* Safety, efficacy, and biomarkers of nivolumab with vaccine in ipilimumab-refractory or -naive melanoma. *Journal of Clinical Oncology* **31**, 4311-4318, doi:10.1200/JCO.2013.51.4802 (2013).

9 Mahoney, K. M. & Atkins, M. B. Prognostic and predictive markers for the new immunotherapies. *Oncology (Williston Park)* **28 Suppl 3**, 39-48 (2014).

10 Larkin, J. *et al.* Combined nivolumab and ipilimumab or monotherapy in untreated Melanoma. *New England Journal of Medicine* **373**, 23-34, doi:10.1056/NEJMoa1504030 (2015).

11 Snyder, A. *et al.* Genetic basis for clinical response to CTLA-4 blockade in melanoma. *New England Journal of Medicine* **371**, 2189-2199, doi:10.1056/NEJMoa1406498 (2014).

12 Wagle, N. *et al.* Dissecting therapeutic resistance to RAF inhibition in melanoma by tumor genomic profiling. *Journal of Clinical Oncology* **29**, 3085-3096, doi:10.1200/JCO.2010.33.2312 (2011).

13 Van Allen, E. M. *et al.* The genetic landscape of clinical resistance to RAF inhibition in metastatic melanoma. *Cancer Discovery* **4**, 94-109, doi:10.1158/2159-8290.cd-13-0617 (2014).

14 Shalek, A. K. *et al.* Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature* **498**, 236-240, doi:10.1038/nature12172 (2013).

15 Patel, A. P. *et al.* Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344**, 1396-1401, doi:10.1126/science.1254257 (2014).

16 Macosko, E. Z. *et al.* Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202-1214, doi:10.1016/j.cell.2015.05.002 (2015).

17 Sos, B. C. *et al.* Characterization of chromatin accessibility with a transposome hypersensitive sites sequencing (THS-seq) assay. *Genome Biology* **17**, doi:10.1186/s13059-016-0882-7 (2016).

18 van der Maaten, L. & Hinton, G. Visualizing Data using t-SNE. *Journal of Machine Learning Research* **9**, 2579--2605 (2008).

19 Ester, M., Kriegel, H.-P., Sander, J. & Xu, X. in *Proc. of 2nd International Conference on Knowledge Discovery and* 226-231 (1996).

20 Whitfield, M. L. *et al.* Identification of Genes Periodically Expressed in the Human Cell Cycle and Their Expression in Tumors. *Molecular Biology of the Cell* **13**, 1977-2000, doi:10.1091/mbc.02-02-0030 (2002).

21 Roesch, A. *et al.* A Temporarily Distinct Subpopulation of Slow-Cycling Melanoma Cells Is Required for Continuous Tumor Growth. *Cell* **141**, 583-594, doi:10.1016/j.cell.2010.04.020 (2010).

22 Shapiro, G. *et al.* A first-in-human phase I study of the CDK4/6 inhibitor, LY2835219, for patients with advanced cancer. *Journal of Clinical Oncology* **31**, 2500-2500, doi:10.1200/jco.2013.31.15_suppl.2500 (2013).

23 Gierahn, T. M. *et al.* Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput. *Nature Methods* **14**, 395-398, doi:10.1038/nmeth.4179 (2017).

24 Johannessen, C. M. *et al.* A melanocyte lineage program confers resistance to MAP kinase pathway inhibition. *Nature* **504**, 138-142, doi:10.1038/nature12688 (2013).

25 Konieczkowski, D. J. *et al.* A melanoma cell state distinction influences sensitivity to MAPK pathway inhibitors. *Cancer Discovery* **4**, 816-827, doi:10.1158/2159-8290.CD-13-0424 (2014).

26 Garraway, L. A. *et al.* Integrative genomic analyses identify MITF as a lineage survival oncogene amplified in malignant melanoma. *Nature* **436**, 117-122, doi:10.1038/nature03664 (2005).

27 Zhang, Z. *et al.* Activation of the AXL kinase causes resistance to EGFR-targeted therapy in lung cancer. *Nature Genetics* **44**, 852-860, doi:10.1038/ng.2330 (2012).

28 Wu, X. *et al.* AXL kinase as a novel target for cancer therapy. *Oncotarget* **5**, 9546-9563, doi:10.18632/oncotarget.2542 (2014).

29 Boiko, A. D. *et al.* Human melanoma-initiating cells express neural crest nerve growth factor receptor CD271. *Nature* **466**, 133-137, doi:10.1038/nature09161 (2010).

30 Hoek, K. S. *et al.* In vivo switching of human melanoma cells between proliferative and invasive states. *Cancer Research* **68**, 650-656, doi:10.1158/0008-5472.CAN-07-2491 (2008).

31 Müller, S. *et al.* Single-cell sequencing maps gene expression to mutational phylogenies in PDGF - and EGF -driven gliomas. *Molecular Systems Biology* **12**, 889-889, doi:10.15252/msb.20166969 (2016).

32 Li, F. Z., Dhillon, A. S., Anderson, R. L., McArthur, G. & Ferrao, P. T. Vol. 5 (Frontiers Research Foundation, 2015).

33 Hugo, W. *et al.* Non-genomic and Immune Evolution of Melanoma Acquiring MAPKi Resistance. *Cell* **162**, 1271-1285, doi:10.1016/j.cell.2015.07.061 (2015).

34 Nazarian, R. *et al.* Melanomas acquire resistance to B-RAF(V600E) inhibition by RTK or N-RAS upregulation. *Nature* **468**, 973-977, doi:10.1038/nature09626 (2010).

35 Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603-607, doi:10.1038/nature11003 (2012).

36 Fridman, W. H., Pagès, F., Sautès-Fridman, C. & Galon, J. r. Vol. 12 298-306 (2012).

37 Carter, S. L. *et al.* Absolute quantification of somatic DNA alterations in human cancer. *Nature Biotechnology* **30**, 413-421, doi:10.1038/nbt.2203 (2012).

38 Roadmap Epigenomics, C. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-329, doi:10.1038/nature14248 (2015).

39 Markiewski, M. M. *et al.* Modulation of the antitumor immune response by complement. *Nature Immunology* **9**, 1225-1235, doi:10.1038/ni.1655 (2008).

40 Wherry, E. J. Vol. 12 492-499 (2011).

41 Chen, L. & Flies, D. B. Vol. 13 227-242 (2013).

42 Borghaei, H. *et al.* Nivolumab versus docetaxel in advanced nonsquamous non-small-cell lung cancer. *New England Journal of Medicine* **373**, 1627-1639, doi:10.1056/NEJMoa1507643 (2015).

43 Motzer, R. J. *et al.* Nivolumab versus everolimus in advanced renal-cell carcinoma. *New England Journal of Medicine* **373**, 1803-1813, doi:10.1056/NEJMoa1510665 (2015).

44 Rizvi, N. A. *et al.* Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. *Science* **348**, 124-128, doi:10.1126/science.aaa1348 (2015).

45 Van Allen, E. M. *et al.* Genomic correlates of response to CTLA-4 blockade in metastatic melanoma. *Science* **350**, 207-211, doi:10.1126/science.aad0095 (2015).

46 Wherry, E. J. *et al.* Molecular Signature of CD8+ T Cell Exhaustion during Chronic Viral Infection. *Immunity* **27**, 670-684, doi:10.1016/j.immuni.2007.09.006 (2007).

47 Fuertes Marraco, S. A., Neubert, N. J., Verdeil, G. & Speiser, D. E. Vol. 6 (Frontiers Media S.A., 2015).

48 Baitsch, L. *et al.* Exhaustion of tumor-specific CD8+ T cells in metastases from melanoma patients. *Journal of Clinical Investigation* **121**, 2350-2360, doi:10.1172/JCI46102 (2011).

49 Martinez, G. J. *et al.* The Transcription Factor NFAT Promotes Exhaustion of Activated CD8+ T Cells. *Immunity* **42**, 265-278, doi:10.1016/j.immuni.2015.01.006 (2015).

50 Blackburn, S. D., Shin, H., Freeman, G. J. & Wherry, E. J. Selective expansion of a subset of exhausted CD8 T cells by αPD-L1 blockade. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 15016-15021, doi:10.1073/pnas.0801497105 (2008).

Chapter 3: Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput

This chapter is adapted in accordance with Nature's open access policy from the following article published in Nature Methods:

Gierahn, T.M.*, Wadsworth II, M.H.*, Hughes, T.K., Bryson, B.D., Butler, A., Satija, R., Fortune, S., Love, J.C., and Shalek, A.K. "Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput," *Nature Methods*, 14, 395-398 (2017).

** Denotes equal authorship*

Abstract

Single-cell RNA-seq can precisely resolve cellular states but applying this method to low-input samples is challenging. Here, we present Seq-Well, a portable, low-cost platform for massively parallel single-cell RNA-seq. Barcoded mRNA capture beads and single cells are sealed in an array of subnanoliter wells using a semipermeable membrane, enabling efficient cell lysis and transcript capture. We use Seq-Well to profile thousands of primary human macrophages exposed to *Mycobacterium tuberculosis*.

INTRODUCTION

3.1 Background

The emergence of single-cell genomics has enabled new strategies for identifying the cellular and molecular drivers of biological phenomena.[1-19] Patterns in genome-wide mRNA expression measured by single-cell RNA-seq (scRNA-seq) can be leveraged to uncover distinct cell types, states and circuits within cell populations and tissues.[1–5,9–13] To inform our understanding of healthy and diseased behaviors and eventually guide precision diagnostics and therapeutics, we need broadly applicable scRNA-seq methods that are easy to use and enable high-throughput studies of cellular phenotypes, particularly for low-input (≤104 cells) samples such as clinical specimens.

Typically, scRNA-seq involves isolating and lysing individual cells, then independently reverse transcribing and amplifying their mRNAs before generating barcoded libraries that are pooled for sequencing. Although manual picking[2,5,8], FACS sorting[1,3,4] or integrated microfluidic circuits[7,9,10] can isolate single cells, these approaches are constrained in scale by cost, time and labor. Recently developed massively parallel methods assign unique bar-codes to each cell's mRNAs during reverse transcription, enabling ensemble processing while retaining single-cell resolution. These techniques typically yield single-cell libraries of lower complexity, but higher throughput reduces the impact of the technical and intrinsic noise associated with each cell in analyses.[11,12] The most commonly used approach relies on microfluidic devices to generate reverse-emulsion droplets that couple single cells with uniquely barcoded mRNA capture beads[11,12]. Droplet-based techniques, however, can have inefficiencies in encapsulation, introduce technical noise through differences in cell lysis time and require specialized equipment—limiting where, when and at what scale scRNA-seq can be performed.

One alternative to droplets is to use arrays of subnanoliter wells loaded by gravity, which reduces the need for peripheral equipment, decreases dead volumes and facilitates parallelization. As a proof of principle, cells and beads have been co-confined in nanowell arrays to perform targeted single-cell transcriptional profiling[13], yet the absence of a seal significantly impairs capture efficiency and increases cross-contamination (**Appendix B,**

Supplementary Figure 1). Nanowells have also been combined with microfluidic channels that facilitate oil-based single-cell isolation via fluid exchange.[14] Nevertheless, this design limits buffer exchange and necessitates integrated temperature and pressure controllers, impacting ease of use and portability.[15] Semiporous-membrane-covered nanowells have been used to link pairs of specific transcripts from single cells[16]; however, this approach used many beads per well, precluding the creation of unique single-cell libraries, and transcript capture and sealing efficiency were not addressed.

RESULTS

3.2 The Development of Seq-Well

To overcome these challenges, we developed Seq-Well, a simple, portable platform for massively parallel scRNA-seq (**Appendix B, Supplementary Figure 2**). Seq-Well confines single cells and bar-coded poly(dT) mRNA capture beads in a PDMS array of ~86,000 subnanoliter wells. Wells accommodate only one bead, enabling single-bead loading efficiencies of ~95% (**Figure 1a; Appendix B, Supplementary Figure 3a**). A simplified cell-loading scheme in turn permits capture efficiencies of around 80% (see **Online Methods** and **Appendix B, Supplementary Figure 3b**), with a dual occupancy rate that can be tuned by adjusting the number of cells loaded and visualized before processing (**Appendix B, Supplementary Figure 3c**).

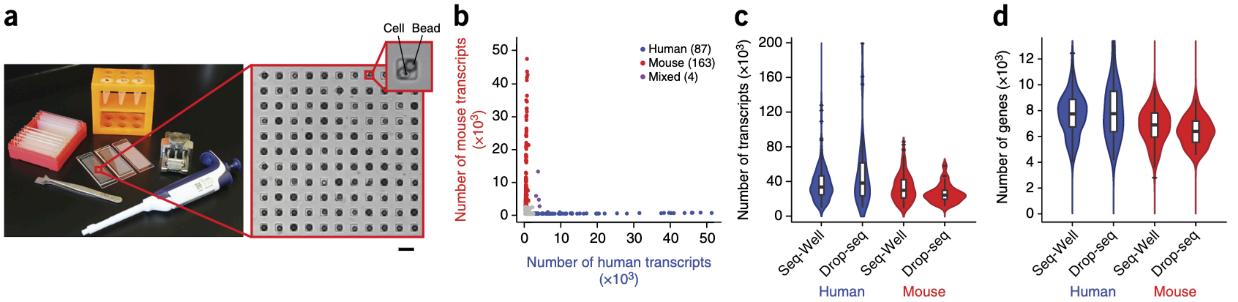


Figure 1 | Seq-Well: a portable, low-cost platform for high-throughput single-cell RNA-seq of low-input samples. (a) Equipment and array used to capture and lyse cells, respectively. Scale bar, 100 µm. **(b)** Sequencing a mix of human (HEK293) and mouse (NIH/3T3) cells reveals distinct transcript mapping and single-cell resolution. Cells with >2,000 human and <1,000 mouse transcripts are labeled as human, and cells with >2,000 mouse and <1,000 human transcripts are labeled as mouse. Of the 254 cells identified, 4 (1.6%) had a mixed phenotype. (c,d) Number of transcripts **(c)** and genes **(d)** detected in single-cell libraries generated by Seq-Well or Drop-seq (ref. 12; center-line: median; limits, first and third quartile; whiskers, ±1.5 IQR; points; values, >1.5 IQR). Using Seq-Well (Drop-seq), an average of 37,878 (48,543) transcripts or 6,927 (7,175) genes were detected among human HEK cells (n = 159 for Seq-Well; n = 48 for Drop-seq); and an average of 33,586 (26,700) transcripts or 6,113 (5,753) genes were detected among mouse 3T3 cells (n = 172 for Seq-Well; n = 27 for Drop-seq) at an average depth of 164,238 (797,915) reads per HEK cell and 152,488 (345,117) reads per 3T3 cell.

A key unique feature of Seq-Well is the use of selective chemical functionalization to facilitate reversible attachment of a semipermeable polycarbonate membrane (10-nm pore size) in physiologic buffers. This enables rapid solution exchange for efficient cell lysis but traps biological macromolecules for improved transcript capture and reduced cross-contamination (**Appendix B, Supplementary Figure 4a**). The array's three-layer surface functionalization comprises an aminosilane base[20] crosslinked to a bifunctional poly(glutamate)–chitosan top via a p-phenylene diisothiocyante intermediate (see **Online Methods** and **Appendix B, Supplementary Figure 4**). In the outer layer, poly(glutamate) at the inner nanowell surfaces prevents nonspecific binding of mRNAs, while chitosan on the array's top sur-face encourages efficient sealing to the membrane (see **Online Methods** and **Appendix B, Supplementary Figure 4b,c**). To test sealing and buffer exchange, we monitored the fluorescence of dye-labeled, cell-bound antibodies before and after adding a guanidinium-based lysis buffer. We observed rapid diffusion of the antibodies throughout the wells within 5 minutes of buffer addition and—unlike in unsealed or previously described, membrane-covered BSA-blocked arrays[16]— we observed little change in fluorescent signal over 30 min, suggesting robust retention of biological macromolecules despite the use of a strong chaotrope (see **Online Methods** and **Appendix B, Supplementary Figure 5**).

After lysis, cellular mRNAs are captured by bead-bound poly(dT) oligonucleotides that also contain a universal primer sequence, a cell barcode and a unique molecular identifier (UMI) (see **Online Methods** and **Appendix B, Supplementary Table 1**). Next, the membrane is peeled off, and the beads are removed for subsequent bulk reverse transcription, amplification, library preparation and paired-end sequencing, as previously described[12] (see **Online Methods**). Critically, beyond a disposable array and membrane, Seq-Well only requires a pipette, a manual clamp, an oven and a tube rotator to achieve stable, barcoded single-cell cDNAs (**Figure 1a**), so it can be performed almost anywhere.

3.3 Validation of Capture Efficiency and Single-cell Resolution

To assess transcript capture efficiency and single-cell resolution, we profiled a mixture of 5 × 103 human (HEK293) and 5 × 103 mouse (3T3) cells using Seq-Well. The average fraction of reads mapping to exonic regions was 77.5% (**Appendix B, Supplementary Figure 6**), demonstrating high-quality libraries. Shallow sequencing from a fraction of an array revealed highly organism-specific libraries, suggesting single-cell resolution and minimal cross-contamination (**Figure 1b** and **Appendix B, Supplementary Figure 7a–c**). In the absence of membrane sealing, by comparison, we obtained poor transcript and gene detection as well as substantial cross-contamination (**Appendix B, Supplementary Figure 1**). From deeper sequencing of a fraction of a second array, we detected an average of 37,878 mRNA transcripts from 6,927 genes in HEK cells and 33,586 mRNA transcripts from 6,113 genes in 3T3 cells, comparable to a droplet-based approach using the same mRNA capture beads (Drop-seq)[12] (**Figure 1c,d** and **Appendix B, Supplementary Figures 7 and 8**). Upon downsampling to match read depths, we also observed levels of transcript and gene detection consistent with those of other massively parallel bead-based scRNA-seq methods (see **Online Methods** and **Appendix B, Supplementary Figure 7d–g**). Moreover, bulk RNA-seq data were strongly correlated with populations constructed in silico from individual HEK cells (R = 0.751 ± 0.073 to R = 0.983 ± 0.0001 for populations of 1–1,000 single cells, respectively), suggesting representative cell and transcript sampling (see **Online Methods** and **Appendix B, Supplementary Figure 9**). To examine Seq-Well's ability to resolve populations of cells in complex primary samples, we loaded human peripheral blood mononuclear cells (PBMCs) into arrays in triplicate before beads, allowing us to perform on-array multicolor imaging cytometry (see Online Methods; **Figure 2a,b**; **Appendix B, Supplementary Tables 2 and 3**).

Sequencing one-third of the beads recovered from each array yielded 3,694 high-quality single-cell libraries (see **Online Methods**). Unsupervised graph-based clustering revealed unique subpopulations corresponding to major PBMC types (see **Online Methods**, **Figure 2b**; **Appendix B, Supplementary Figures 10–12** and **Table 4**). Each array yielded similar subpopulation frequencies (**Figure 2c**), with detection efficiencies comparable to those of other massively parallel technologies (**Appendix B, Supplementary Figure 13**). The proportion of each subpopulation determined by sequencing also matched on-array immunophenotyping results (**Figure 2a,b**).

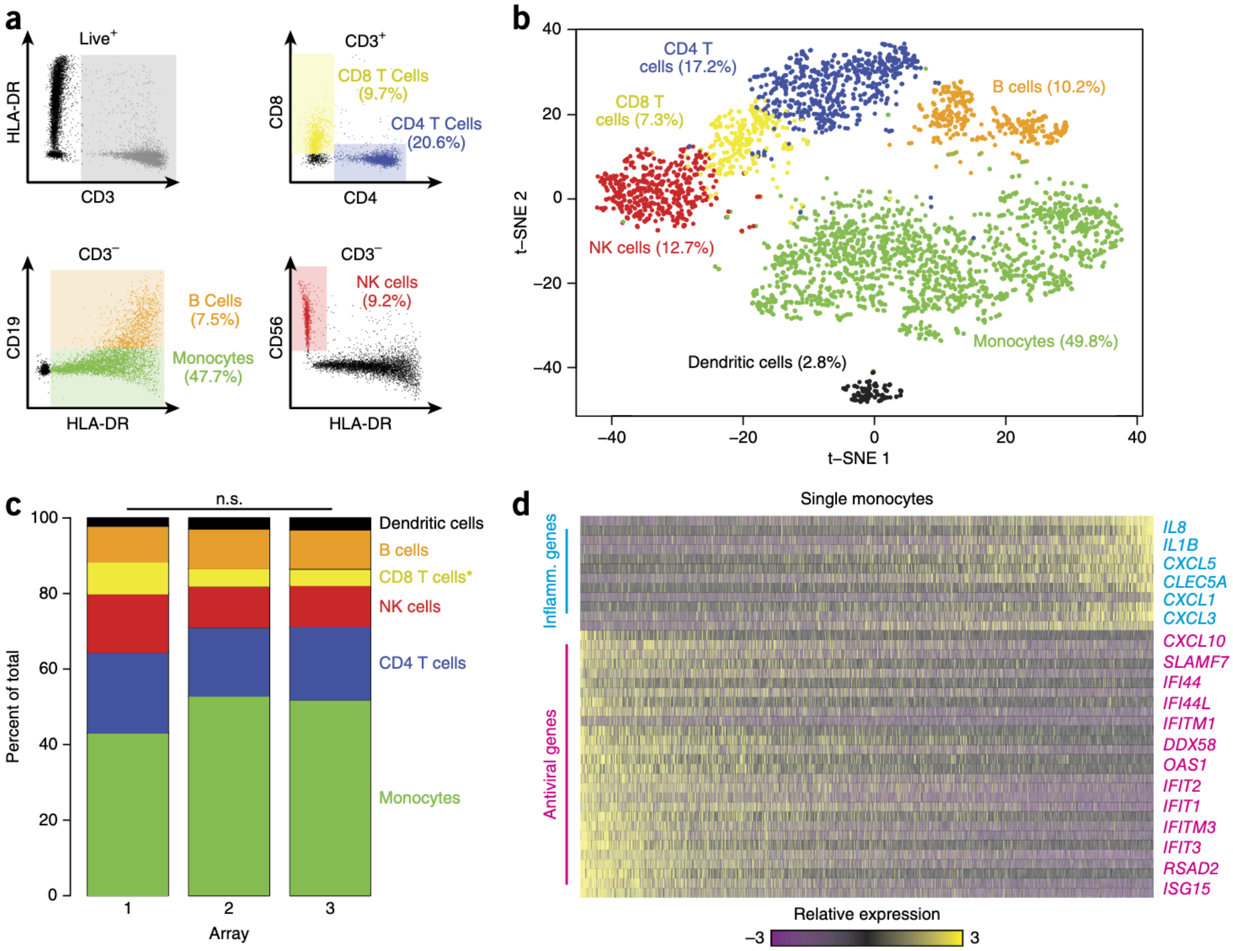


Figure 2 | Combined image cytometry and scRNA-seq of human PBMCs. (a) Hierarchical gating scheme used to analyze PBMCs labeled with a panel of fluorescent antibodies, loaded onto three replicate arrays and imaged before bead loading (see **Online Methods**). Myeloid cells (green) were identified as the population of hCD3(−) HLA-DR(+) CD19(−) cells; B cells (orange) as the subset of hCD3(−) HLA-DR(+) CD19 (+) cells; CD4 T cells (blue) as the subset of CD3(+) CD4(+) cells; CD8 T cells (yellow) as the CD3(+) CD8(+) subset of cells; and NK cells (red) as the subset of CD3(−) HLA-DR(−) CD56(+) CD16(+) cells. **(b)** t-SNE visualization of clusters identified among 3,694 human Seq-Well PBMCs single-cell transcriptomes recovered from the imaged array and the two additional arrays (see **Online Methods**). Clusters (subpopulations) are labeled based on annotated marker gene (supplementary fig. 0). **(c)** Distribution of transcriptomes captured on each of the biological replicate arrays, run on separate fractions of the same set of PBMCs. No shifts are statistically significant (n.s. = not significant; see **Online Methods**) except for a slightly elevated fraction of CD8 T cells in array 1 (*, P = 1.0 × 10−11; Chi-square test, Bonferroni corrected). **(d)** Relative expression level of a set of inflammatory and antiviral genes among cells identified as monocytes. Inflamm., inflammatory.

Critically, sequencing provided additional information; in addition to resolving dendritic cells from monocytes (**Figure 2b**), we found significant variation among the monocytes (captured in PC3) due to differential expression of inflammatory and antiviral gene programs (**Figure 2d**).[1,3] Our results show that characterizing a sample in two ways using a single platform increases the amount of information that can be extracted from a precious specimen, while allowing analysis of one measurement to be interpreted in the context of the other.

3.4 Demonstration of Portability

Finally, to test the portability of Seq-Well, we profiled primary human macrophages exposed to Mycobacterium tuberculosis (H37Rv) in a BSL3 facility (see **Online Methods**). In total, we recovered 14,218 macrophages (from a total of 40,000 loaded across experiments) with greater than 1,000 mapped transcripts from an M. tuberculosis-exposed and an unexposed array. Unsupervised analysis of 4,638 cells with greater than 5,000 transcripts per cell revealed five distinct clusters (**Figure 3a,b; Appendix B, Supplementary Figure 14a,b and Table 5**). Two clusters had lower transcript capture and high mitochondrial gene expression (suggestive of low-quality libraries)[17] and were removed; the remaining three (2,560 cells) were identified in both the exposed and unexposed samples (**Figure 3a; Appendix B, Supplementary Figures 14c,d and 15**), and they likely represent distinct subphenotypes present in the initial culture.

We next examined common and cluster-specific gene enrichments (see **Online Methods**). Although clusters 1 and 3 did not present strong stimulation-independent enrichments, cluster 2 uniquely expressed several genes associated with metabolism (**Appendix B, Supplementary Tables 6 and 7**). Intriguingly, within each cluster we observed pronounced shifts in gene expression in response to M. tuberculosis (see **Online Methods**; **Figure 3c**; **Appendix B, Supplementary Table 8**), with common enrichments for gene sets previously observed in response to intracellular infection, LPS stimulation and activation of TLR7/8 receptor (**Appendix B, Supplementary Tables 9 and 10**). Cluster 1 uniquely displayed stimulation-induced shift in several genes associated with cell growth, cluster 3 in transcripts associated with hypoxia, and cluster 2 (again) in genes linked to metabolism. Overall, these data suggest that basal cellular heterogeneity may influence ensemble M. tuberculosis responses. Equally important, they demonstrate Seq-Well's ability to acquire large numbers of single-cell transcriptomes in challenging experimental environments.

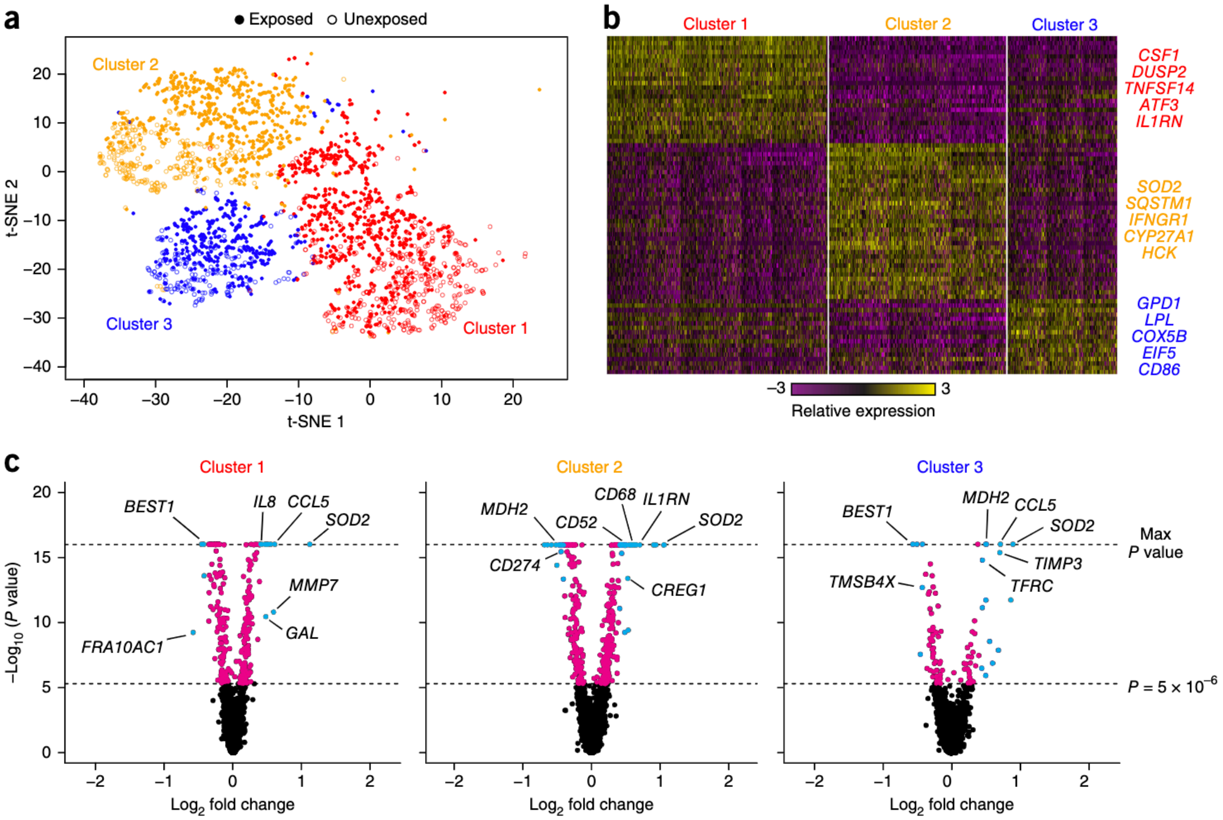


Figure 3 | Sequencing of TB-exposed macrophages in a BSL3 facility using Seq-Well. (a) t-SNE visualization of single-cell clusters identified among 2,560 macrophages (1,686 exposed, solid circles; 874 unexposed, open circles) generated using five principal components across 377 variable genes (see Online Methods). **(b)** Marker genes for the three phenotypic clusters of macrophages highlighted in a. **(c)** Differential expression between exposed and unexposed macrophages within each cluster showing genes enriched in cells exposed to M. tuberculosis. Cyan, genes with P values less than 5.0 × 10−6 (threshold for statistical significance, determined by a likelihood ratio test) and absolute log2 fold changes greater than 0.4 (threshold used for differential expression). Magenta, genes with P values less than 5.0 × 10−6 but absolute log2 fold changes less than 0.4. Black, remaining genes.

3.5 Conclusion

In conclusion, Seq-Well is a robust platform for scalable, single-cell transcriptomics applicable to almost any cellular suspension for which a reference genome or transcriptome exists. The technique is inexpensive, user friendly, portable, and efficient; it enables scRNA-seq to accelerate scientific and clinical discovery even when working with limited samples. Furthermore, the ability to measure protein secretion and cell-surface expression on the same platform[18,19] foreshadows multi-omic single-cell measurements at scale.

METHODS

3.6 Running Seq-Well

See Protocol Exchange[21], http://www.shaleklab.com, or Appendix D for a step-by-step Seq-Well protocol. In this work, we used the primers listed in Appendix B, Supplementary Table 1.

3.7 Bead synthesis

Barcoded oligo-dT beads (as described in Macosko et al.[12]) were purchased from Chemgenes (Wilmington, Massachusetts, USA; cat. no. MACOSKO-2011-10) at 10 umol scale (~100 arrays). Bead functionalization and reverse phosphoramidite synthesis was performed by Chemgenes Corporation using Toyopearl HW-65S resin (30 micron mean particle diameter) obtained from Tosoh Biosciences (cat. no. 19815). Surface hydroxyls were reacted with a PEG derivative to obtain an 18-carbon linker to serve as a support for oligo synthesis. Reverse-direction phosphoramidite synthesis was performed using an Expedite 8909 DNA/RNA synthesizer at 10 micromole scale with a coupling time of 3 min. Initially, a conserved PCR handle was synthesized followed by 12 rounds of split and pool synthesis to generate 16,777,216 unique barcode sequences. Addition of an 8-mer random sequence was performed to generate unique molecular identifiers (UMIs) on each capture oligo. Finally, a 30-mer poly-dT capture sequence was synthesized to enable capture of polyadenylated mRNA species.

4.8 Imaging differential surface functionalization

Differential labeling of the top and inner well surfaces was visualized by substituting 1 ug/mL PE–strepavidin for chitosan (**Appendix D**; step 8, Seq-Well Protocol[21]) and 1 ug/mL AlexaFluor488-Streptavidin for the poly-glutamate (**Appendix D**; step 10, Seq-Well Protocol[21]) in the standard functionalization protocol (**Appendix B, Supplementary Figure 3**). Carboxylation of the inner well surfaces was visualized by treating the functionalized array with 100 µg/mL EDC/10 µg/mL NHS MES (pH 6.0) solution for 10 min, washing twice with MES buffer, once with sodium borate buffer (pH 8.5), and incubating overnight with 1 µg/mL Alexa-Fluor 568-labeled antibody. Arrays were washed three times with phosphate-buffered saline (PBS) and imaged using Alexa Fluor 568 channel (**Appendix B, Supplementary Table 2**).

3.9 Visualizing lysate retention (imaging)

PBMCs were labeled with αCD45–AF647 (BioLegend 304020, diluted 1:20). Cells were washed and loaded onto two arrays previously blocked with 1% BSA solution for 30 min and one array functionalized with chitosan as described above. A polycarbonate membrane was attached to the chitosan-functionalized array as described above. The array was submerged in PBS and imaged for AF647 fluorescence to identify wells containing cells. The BSA-blocked arrays were imaged before membrane attachment because the membrane would detach when submerged in media. After imaging, a plasma-treated polycarbonate membrane was attached to one of the BSA-blocked arrays as described[16]. Briefly, the membrane was placed on the array with forceps, and all excess media were aspirated from the array. The open BSA-blocked array and the chitosan array were submerged in 5 mL of 5 M GCTN lysis buffer. 500 µL of lysis buffer was placed on the top of membrane attached to the BSA-blocked array as described16. 5 and 30 minutes later, 100 block positions were imaged on each array, encompassing 12,100 individual wells. Automated image analysis software was used to background subtract each image, identify cell and well locations and extract AF647 signal intensity of the cells and the well volumes (**Appendix B, Supplementary Figure 4**).

3.10 Calculating bead loading efficiency

Bead loading efficiencies were determined by loading two functionalized arrays with beads as outlined above (**Appendix B, Supplementary Figure 1**). Arrays were imaged in transmitted light and AF488 channel (**Appendix B, Supplementary Table 2**) to capture bead autofluorescence. Automated image analysis was used to identify well locations and extract the 75th percentile fluorescence intensity in each well. Histogram analysis of fluorescence intensities was used to identify empty wells and wells containing beads. Finally, manual review of 50 randomly selected image positions, each containing 121 nanowells, of a total of 690 positions was used to calculate the frequency of wells containing two beads.

3.11 Calculating cell loading efficiency

To calculate cell loading efficiencies and well occupancy distributions (**Appendix B, Supplementary Figure 3**), HEK293 and 3T3 cells were labeled with Calcein AM (Life Technologies) and Calcein Violet (Life Technologies), respectively, per the manufacturer's recommendations. 200 μL of serial dilutions of a 1:1 mix of the cells at an estimated concentration of 1,000, 10,000 and 100,000 cells/mL were loaded in functionalized arrays in triplicate using the standard protocol. To determine the distribution of cells present in 200 μL of these solutions, the same volume of each solution was added to 12 wells of a 96-well plate. 690 array positions on each array were imaged in the transmit-ted light, AF488 and AF405 spectral channels (Supplementary Table 2). Overlapping images of each well of the 96-well plate were acquired in the same channels. Automated image analysis was used to identify well and cell locations in the array images. The overlapping images of the 96-well plate were stitched together based on x–y location of each image and analyzed in a similar manner to identify cell locations. All three dilutions were used to determine the distribution of well occupancy as a function of the number of cells loaded. The 10,000 cells/mL dilution were used to calculate cell loading efficiency.

3.12 Species-mixing experiments

Murine NIH/3T3 cells (ATCC, CRL-1658) were cultured in Dulbelcco's modified Eagle's medium (DMEM) with glutamate and supplemented with 10% fetal bovine serum (FBS) at 37 °C and 5% CO2. Human 293T cells (ATCC, CRL-11268) were cultured at 37 °C and 5% CO2 in DMEM with glutamate supplemented with 10% FBS. The media were removed from the culture flasks, which were then rinsed with 5 mL of 1× PBS. Cells were detached from the surface of the culture flasks by applying 3.5 mL of Trypsin-LE (Life Technologies) and incubating at room temperature for 5 min. Once cells had adhered, 10 mL of complete media was added, and cells were pelleted by spinning at 500× g for 10 min. Cell pellets were resuspended in 1 mL of media, and a 10 μL aliquot was used to count cells. A total of 100,000 HEK and 3T3 cells were again pelleted and resuspended in 1 mL of media. For species-mixing experiments, a total of 200 μL of a single-cell suspension containing 5,000 HEK and 5,000 NIH/3T3 cells was applied to the surface of two nanowell devices loaded with beads. In the first experiment, of the 60,000 beads collected from the array, 9,600 beads were pooled for subsequent processing and sequencing, from which we identified 254 high-quality cells with greater than 2,000 transcripts. In the second experiment, of the 25,000 beads collected from the array, 15,000 beads were pooled for subsequent processing and sequencing, from which we identified 331 high-quality cells with greater than 10,000 transcripts, greater than 2,000 genes, and greater than 90% transcript purity (i.e. >90% of transcripts from the same species). Also, as in Drop-Seq, we attempted to validate capture efficiency using ERCC spike-ins; however, this required us to load ERCCs onto the nanowell array by pipetting, which proved inefficient to properly assess capture efficiency since we could not evenly distribute ERCCs to nanowells.

3.13 HEK population experiments

HEK293 cells were cultured in RPMI supplemented with 10% FBS. A total of 10,000 HEK293 cells were applied to a Seq-Well device and scRNA-seq libraries were generated from 24,000 beads and sequenced on a NextSeq 500. For the bulk RNA-seq sample, cellular lysate from 40,000 HEK293 cells in 200 µL of lysis buffer (5 M GTCN, 1% 2-mercaptoethanol, 1 mM EDTA, and 0.1% Sarkosyl in 1× PBS, pH 6.0) was combined with 40,000 mRNA capture beads in a PCR tube and rotated end over end for 1 h. Afterward, the beads were washed, and a population sequencing library was constructed in an identical manner to that of the single-cell Seq-Well libraries but with reads from the different bead barcodes combined into one population. In silico populations were created by randomly sampling 1, 10, 100 or 1,000 HEK cells from a total of 1,453 cells with greater than 3,000 transcripts obtained from a Seq-Well array. Average Pearson correlation coefficients and their s.d. were calculated between 100 randomly generated in silico populations for each number of cells and the bead population (**Appendix B, Supplementary Figure 9**).

3.14 Human PBMC experiments

Leukocytes isolated from a leukocyte reduction filter used during platelet aphoresis were purchased from Key Biologics (Memphis, Tennessee). The cells were shipped overnight at room temperature. PBMCs were isolated from the sample using a Ficoll–Hypaque (GE) gradient, washed two times with HBSS buffer, and frozen in 90% FBS/10%DMSO in aliquots of 107 cells. The day before the experiment, an aliquot was thawed and rested overnight in RPMI-1640 supplemented with 10% FBS, Pen–Strep, nonessential amino acids, sodium pyruvate, and HEPES buffer (RP10) at 106 cells/mL in a 50 mL conical tube. Cells were counted the next day, and 5 × 105 cells were pelleted, resuspended in 1 mL of CellCover solution, and processed as described above.

3.15 Array loading for imaging (PBMCs)

To quantify cell surface marker protein expression levels on array (**Figure 2a**), PBMCs were loaded first and imaged before bead addition to avoid potential detection issues associated with bead autofluorescence. Here, cells were resuspended in cold CellCover (Anacyte), an RNA stabilization reagent, and placed at 4C for 1 h. Cells were spun down and resuspended in a cocktail containing αCD45-AF647 (BioLegend; HI30), αCD3-PerCP (BioLegend; UCHT1), αCD4-PECy5.5(eBioscience; SK3), αCD56-PECy5(BD Biosciences; B159), αCD8-APCCy7 (BioLegend; RPA-T8), αHLA-DR-PECy7 (BD Biosciences; L243), and αCD19-PE (BioLegend; HIB19) with all antibodies diluted 1:20 in RP10 media and were incu-bated at 4 °C for 30 min. Cells were washed twice with PBS and resuspended in CellCover10 buffer (CellCover supplemented with 10% FBS and 100 mM sodium carbonate (pH 10) buffer). Functionalized arrays were washed with 5 mL of CellCover10 buffer. 2.0 × 104 cells were loaded onto the array and washed twice with CellCover10 buffer, and finally the array was placed in 5 mL CellCover. Arrays were imaged with a Zeiss AxioVision microscope with Lumencor light source and EMCCD camera using the settings described in Appendix B, Supplementary Table 2. Automated imaging software was used to identify cell locations within the images and extract signal intensities in each spectral channel. To generate spillover coefficients for each fluorophore, α−mouse beads (Bangs Labs) were stained individually with each antibody using the same protocol as the cells. Images of the singly stained beads were used to generate spillover coefficients for each fluorophore that were then used to calculate the amount of each fluorophore on each cell as previously described.[22] After imaging, arrays were washed with 5 mL CellCover10 media. Barcoded beads suspended in CellCover10 media were loaded into the array through gentle agitation. Arrays were washed 3x with

CellCover10 without FBS and finally washed with CellCover. Arrays were then moved on to membrane attachment.

3.16 Human monocyte isolation

Primary human monocytes were isolated from deidentified human buffy coats obtained from the Massachusetts General Hospital Blood Bank using a standard Ficoll gradient and subsequent CD14 positive selection (Stemcell Technologies). Enriched monocytes were cultured in low-adherence flasks (Corning) for 9 d with RPMI media (Invitrogen) supplemented with 10% heat-inactivated FCS (Sigma-Aldrich).

3.17 Mycobacterium tuberculosis culture

Mycobacterium tuberculosis (Mtb) H37Rv expressing the E2-Crimson fluorescent protein was grown in Difco Middlebrook 7H9 media supplemented with 10% OADC, 0.2% glycerol, 0.05% Tween-80 and Hygromycin B (50 ug/mL).

3.18 Macrophage infection and flow cytometry

The Mtb culture was pelleted by centrifugation and washed once with RPMI + 10% FCS, sonicated briefly, and filtered through a 5 μm syringe filter. Monocyte-derived macrophages (MDM) were infected at an MOI of 10 for 4 h and then washed 3× with RPMI + 10% FCS. 24 h after infection, cells were washed briefly with 1× PBS. 10× Trypsin (Life Technologies) was added, and cells were incubated briefly at 37 °C to allow for cell detachment. Detached cells were spun down and resuspended in 1× PBS supplemented with 2% FCS and 1 mM EDTA and then passed through a mesh filter to eliminate clumps. Uninfected and infected cells were sorted by flow cytometry on an Aria II flow cytometer. Mtb-infected cells were identified by the presence of an E2-Crimson signal above the background autofluorescence of uninfected cells.

3.19 Transcriptome alignment and barcode collapsing

Read alignment was performed as in Macosko et al.[12]. Briefly, for each NextSeq sequencing run, raw sequencing data was converted to FASTQ files using bcl2fastq2 that were demultiplexed by Nextera N700 indices corresponding to individual samples. Reads were first aligned to both HgRC19 and mm10, and individual reads were tagged according to the 12-bp barcode sequence and the 8-bp UMI contained in read 1 of each fragment. Following alignment, reads were binned and collapsed onto 12-bp cell barcodes that corresponded to individual beads using Drop-seq tools (http://mccarrolllab.com/dropseq). Barcodes were collapsed with a single-base error tolerance (Hamming distance = 1), with additional provisions for single insertions or deletions. An identical collapsing scheme (Hamming distance = 1) was then applied to UMIs to obtain quantitative counts of individual mRNA molecules. Quality metrics are presented in Supplementary Figures 5 and 8 (**Appendix B)**.

3.20 Data normalization

Digital gene expression matrices were obtained by collapsing filtered and mapped reads for each gene by 8-bp UMI sequences within each cell barcode. For each cell, we performed library-size normalization. UMI-collapsed gene expression values for each cell barcode were scaled by the total number of transcripts and multiplied by 10,000. Scaled expression data were then natural-log transformed before analysis using Seurat.[23]

3.21 Analyzing species-mixing experiments

In the first experiment, HEK cells were identified as those barcodes with greater than 2,000 human transcripts and less than 1,000 mouse transcripts, while barcodes with greater than 2,000 mouse transcripts and less than 1,000 human transcripts were identified as 3T3 cells. Cells with fewer than 2,000 total transcripts were considered indeterminate, while any cell with greater than 5,000 total transcripts and more than 1,000 nonmouse or nonhuman transcripts was considered a multiplet (**Figure 1d**). In the second experiment, HEK cells were identified as those barcodes with greater than 10,000 human transcripts, greater than 2,000 human genes, and greater than 90% human transcript alignment; while barcodes with greater than 10,000 mouse transcripts, greater than 2,000 mouse genes, and greater than 90% mouse transcript alignment were identified as 3T3 cells. Cells with fewer than 10,000 total transcripts were considered indeterminate, while any cells with greater than 10,000 total transcripts and more than 1,000 nonmouse or nonhuman transcripts were considered multiples (**Figure 1c** and **Appendix B, Supplementary Figure 8**).

3.22 PBMC analysis

We reduced the dimensionality of our data to 11 principle components that account for the majority of the variation (51.6% cumulative variance) among variable genes to achieve optimal discrimination of cell types identified through image cytometry. We identified seven distinct clusters of cells using the FindClusters function in Seurat with k.param = 50 (a measure of neighborhood size) and resolution = 0.75 (**Appendix B, Supplementary Figure 11**). Clusters corresponding to CD4 T cells, CD8 T cells, B cells, NK cells, monocytes, and dendritic cells were all identified on the basis of significant enrichment using an ROC test implemented in Seurat (also see **Appendix B, Supplementary Figures 10 and 11**). We removed 602 cells that comprised a distinct cluster enriched for expression of mitochondrial genes (**Appendix B, Supplementary Figure 11**) and a lower mapping rate of new transcripts and genes per sequencing read (**Appendix B, Supplementary Figure 12**), which likely rep-resented single-cell libraries of low complexity. We then applied t-distributed stochastic neighbor embedding (t-SNE) using the cell loadings for the previously chosen 11 principle components to visualize the cells in two dimensions. Following sequence alignment, we analyzed a total of 4,296 cells in which at least 10,000 reads, 1,000 transcripts and 500 genes were detected with mRNA alignment rate greater than 65% (**Figure 2b–d**), which resulted in filtering of 1,670 cells with greater than 1,000 transcripts. We analyzed a total of 6,713 genes that were detected in at least 2.5% of filtered cells across six sequencing runs from three separate arrays. We identified 687 variable genes with log-mean expression values greater than 0.5 and dispersion (variance/mean) greater than 0.5. We observed optimal discrimination of cell types identified through image cytometry using 11 principal components that account for the majority of the variation (51.6% cumulative variance) among variable genes and visualized using the t-distributed stochastic neighbor embedding (t-SNE) algorithm. We performed 1,000 iterations of the Barnes–Hut implementation of the t-SNE algorithm using a ‘perplexity’ value of 40. We identified seven distinct clusters of cells using the FindClusters function in Seurat with k.param = 50 (a measure of neighborhood size) and resolution = 0.75 (**Appendix B, Supplementary Figure 11**). Clusters corresponding to CD4+ T cells, CD8+ T cells, B cells, NK cells, monocytes and dendritic Cells were all identified on the basis of significant enrichment using an ROC test implemented in Seurat (also see **Appendix B, Supplementary Figures 10 and 11**). We removed 602 cells that comprised a distinct cluster enriched for expression of mitochondrial genes (**Appendix B, Supplementary Figures 11**) and a lower mapping rate of new transcripts and genes per sequencing read (**Appendix B, Supplementary Figures 12**), which likely represented single-cell libraries of low complexity. We examined proportions of various cell types across arrays and sequencing runs among 3,694 cells that passed the aforementioned filtering criteria. Statistical significance of differences in the proportion of clusters between separate arrays and sequencing runs was performed using a Chi-square test (**Figure 2c**). We further examined phenotypic variation within myeloid cells among identified principal components (**Figure 2d**) by ranking cells on the basis of their PC score among genes with highest loadings for each principal component.

3.23 Comparison of Seq-Well PBMCs to 10X genomics data

We performed comparisons of gene detection and transcript capture among PBMC cell types conserved between 3,590 PBMCs (excluding dendritic cells) obtained using Seq-Well and 2,700 PBMCs from the 10x Genomics platform (http://support.10xgenomics.com/single-cell/datasets/pbmc3k). To classify PBMC cell types within the 10x Genomics data, we first identified 446 variable genes with log-mean expression values greater than 0.5 and dispersion (variance/mean) greater than 0.5. We then performed graph-based clustering using 13 principal components, k.param of 50 and resolution of 0.75. Cell type identity of each cluster was established on the basis of gene enrichments. Comparisons of genes and transcripts were initially performed between B cells, CD4 T cells, CD8 T cells, monocytes and NK cells using raw data matrices. We refined these comparisons by separately downsampling genes and transcripts within each cell type in Seq-Well data to an average read depth of 69,000 reads per cell to match the reported sequencing depth using in publicly available 10x Genomics data.

3.24 Mycobacterium tuberculosis analysis

Following sequence alignment, we identified a total of 14,218 cells with greater than 1,000 mapped transcripts. Initially, we analyzed a subset of 4,638 macrophages with greater than 5,000 detected transcripts (**Appendix B, Supplementary Figure 14a**) and a total of 9,381 genes expressed in at least 5% of filtered cells. Principal components analysis was performed among a set of 377 variables genes, defined by genes with log-mean expression greater than 0.5 and dispersion (variance/mean) greater than 0.5. We performed graph-based clustering, as described below, using the first five principal components since we observed that they captured the majority of the biological variation in our data set (63% cumulative variance), and that each additional principal component contributed less than 1% to the total variance. We performed 1,000 iterations of the t-SNE algorithm (Barnes–Hut implementation) using a 'perplexity' value of 30. We identified five distinct clusters of cells in the t-SNE plot using the FindClusters function in Seurat with k.param = 40 and resolution = 0.25 (**Appendix B, Supplementary Figure 14**).We removed two clusters comprised of cells with reduced gene detection, transcript capture and enrichment for expression of mitochondrial genes. Following removal of low-quality cells, we analyzed three distinct clusters with total of 2,560 high-quality cells (**Figure 3a** and **Appendix B, Supplementary Figure 14**). Differential expression analysis was performed between clusters, and cells exposed and unexposed to TB within each t-SNE cluster using a likelihood ratio test in Seurat (**Figure 3c** and **Appendix B, Supplementary Table 9**). We performed gene set enrichment analysis to examine association of expression differences observed between control macrophages exposed and unexposed to M. tuberculosis with previously published gene sets using GSEA. For each cluster, expression patterns between exposed and unexposed cells were made to complete GSEA databases (**Appendix B, Supplementary Tables 7, 8, 10 and 11**).

3.25 Regressing out latent technical effects

Technical parameters governing sequencing data, such as the number of genes detected or the transcriptomics alignment rate, often vary significantly across single cells. We sought to conservatively remove these technical effects using a 'latent-variable' approach similar to that of Buettner et al.[24] Briefly, we fit a linear model to predict the expression value of each gene based on a set of technical metrics, as well as the total number of unique genes detected in that cell. In our analyses, we constructed models to adjust gene expression values for alignment rate of each cell. We considered the residual expression from this model as a 'corrected' gene expression value, and we used these values as input to the downstream clustering analyses.

3.26 Graph-based clustering of single-cell transcriptomes

For all single-cell clustering analyses, we used an approach similar to that of our recently proposed clustering strategy for Drop-seq data. Briefly, as in Macosko et al.[12], we first identified the set of genes that was most variable across our data set after controlling for the relationship in single-cell RNA-seq data that inherently exist between mean expression and variability by binning genes into 20 bins based on their average expression level and z-scoring dispersion (mean/variance) estimates within a bin. We excluded all genes which were detected in less than 2.5% of PBMCs (5% of monocytes for the Mtb experiments) and used a dispersion cutoff of 0.5 to select variable genes, resulting in the selection of 687 variable genes across 4,296 PBMCs and 377 variable genes across 4,638 macrophages. We next reduced the dimensionality of our data set, using principal components analysis. As previously described in Macosko et al.12, we ran PCA using the prcomp function in R. We then selected PCs for further downstream analysis (11 PCs in PBMC analysis and 5 PCs in TB analysis). As expected, markers for distinct cell types were highly represented among the genes with the largest scores along these PCs. We then applied t-distributed stochastic neighbor embedding (t-SNE) using cell loadings for the significant principal components as input to visualize the structure of our data in two dimensions. Here we used graph-based clustering methods, similar to those that have been recently proposed for both single-cell RNA-seq and mass cytometry data[25,26]. We first construct a Euclidean distance matrix on the loadings for the significant principal components as described above and use this to construct a K-nearest neighbor graph (KNN, k = 50 in PBMC analysis; k = 40 in TB analysis). Our goal was to identify 'quasi-cliques'[26], or 'communities'[25], of cells that were highly interconnected across this graph. Therefore, we first converted the KNN graph into a weighted shared nearest neighbor (SNN) graph, where the weight between any two cells was represented by the percent overlap in their respective K-nearest neighborhoods (Jaccard distance), and we pruned low-quality edges with a Jaccard distance of <0.1 (less than 10% overlap in local neighborhoods). Finally, to group the cells into clusters, we used a recently developed method for modularity optimization, which aims to optimize a function describing the density of connections within a cluster versus connections between clusters, essentially to identify highly interconnected nodes within the SNN graph. Here, we applied the smart local moving algorithm, which is similar to the widely used 'Louvain' algorithm for community detection but implements a local moving heuristic that enables communities to be split up and iteratively reorganized in an attempt to improve the overall partition modularity. This grants the SLM algorithm additional freedom in identifying an optimal clustering solution, and we empirically observed increased sensitivity and consistency applying this approach to single-cell data.

3.27 References

1 Shalek, A. K. *et al.* Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature* **498**, 236-240, doi:10.1038/nature12172 (2013).

2 Lohr, J. G. *et al.* Whole-exome sequencing of circulating tumor cells provides a window into metastatic prostate cancer. *Nature Biotechnology* **32**, 479-484, doi:10.1038/nbt.2892 (2014).

3 Shalek, A. K. *et al.* Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature* **510**, 363-369, doi:10.1038/nature13437 (2014).

4 Tirosh, I. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352**, 189-196 (2016).

5 Hashimshony, T., Wagner, F., Sher, N. & Yanai, I. CEL-Seq: Single-Cell RNA-Seq by Multiplexed Linear Amplification. *Cell Reports* **2**, 666-673, doi:10.1016/j.celrep.2012.08.003 (2012).

6 Bendall, S. C. *et al.* Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. *Science* **332**, 687-696, doi:10.1126/science.1198704 (2011).

7 Buenrostro, J. D. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486-490 (2015).

8 Smallwood, S. A. *et al.* Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nature Methods* **11**, 817-820, doi:10.1038/nmeth.3035 (2014).

9 Zeisel, A. *et al.* Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* **347**, 1138-1142, doi:10.1126/science.aaa1934 (2015).

10 Treutlein, B. *et al.* Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature* **509**, 371-375, doi:10.1038/nature13173 (2014).

11 Klein, A. M. *et al.* Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187-1201, doi:10.1016/j.cell.2015.04.044 (2015).

12 Macosko, E. Z. *et al.* Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202-1214, doi:10.1016/j.cell.2015.05.002 (2015).

13 Fan, H. C., Fu, G. K. & Fodor, S. P. A. Combinatorial labeling of single cells for gene expression cytometry. *Science* **347**, doi:10.1126/science.1258367 (2015).

14 Bose, S. *et al.* Scalable microfluidics for single-cell RNA printing and sequencing. *Genome Biology* **16**, doi:10.1186/s13059-015-0684-3 (2015).

15 Yuan, J. & Sims, P. A. An Automated Microwell Platform for Large-Scale Single Cell RNA-Seq. *Scientific Reports* **6**, doi:10.1038/srep33883 (2016).

16 Dekosky, B. J. *et al.* High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire. *Nature Biotechnology* **31**, 166-169, doi:10.1038/nbt.2492 (2013).

17 Ilicic, T. *et al.* Classification of low quality cells from single-cell RNA-seq data. *Genome Biology* **17**, doi:10.1186/s13059-016-0888-1 (2016).

18 Yamanaka, Y. J. *et al.* Single-cell analysis of the dynamics and functional outcomes of interactions between human natural killer cells and target cells.

Integrative Biology (United Kingdom) **4**, 1175-1184, doi:10.1039/c2ib20167d (2012).

19 Han, Q. *et al.* Polyfunctional responses by human T cells result from sequential release of cytokines. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 1607-1612, doi:10.1073/pnas.1117194109 (2012).

20 Steinberg, G., Stromsborg, K., Thomas, L., Barker, D. & Zhao, C. Strategies for Covalent Attachment of DNA to Beads. *Biopolymers* **73**, 597-605, doi:10.1002/bip.20006 (2004).

21 Hughes, T. K. *et al.* Highly Efficient, Massively-Parallel Single-Cell RNA-Seq Reveals Cellular States and Molecular Features of Human Skin Pathology. *bioRxiv*, 689273, doi:10.1101/689273 (2019).

22 Roederer, M. Compensation in Flow Cytometry. *Current Protocols in Cytometry* **22**, 1.14.11-11.14.20, doi:10.1002/0471142956.cy0114s22 (2002).

23 Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology* **33**, 495-502, doi:10.1038/nbt.3192 (2015).

24 Buettner, F. *et al.* Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nature Biotechnology* **33**, 155-160, doi:10.1038/nbt.3102 (2015).

25 Levine, J. H. *et al.* Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. *Cell* **162**, 184-197, doi:10.1016/j.cell.2015.05.047 (2015).

26 Xu, C. & Su, Z. Identification of cell types from single-cell transcriptomes using a novel clustering method. *Bioinformatics* **31**, 1974-1980, doi:10.1093/bioinformatics/btv088 (2015).

Chapter 4: Highly Efficient, Massively-Parallel Single-Cell RNA-Seq Reveals Cellular States and Molecular Features of Human Skin Pathology

This chapter is adapted in accordance with Cold Spring Harbor Laboratory's open access policy from the following article published in BioRxiv:

Travis K Hughes*, Marc H Wadsworth II*, Todd M Gierahn*, Tran Do , David Weiss , Priscilla R. Andrade , Feiyang Ma , Bruno J. de Andrade Silva , Shuai Shao , Lam C Tsoi , Jose Ordovas-Montanes, Johann E Gudjonsson , Robert L Modlin, and Alex K Shalek

** Denotes equal authorship*

Abstract

The development of high-throughput single-cell RNA-sequencing (scRNA-Seq) methodologies has empowered the characterization of complex biological samples by dramatically increasing the number of constituent cells that can be examined concurrently. Nevertheless, these approaches typically recover substantially less information per-cell as compared to lower-throughput microtiter plate-based strategies. To uncover critical phenotypic differences among cells and effectively link scRNA-Seq observations to legacy datasets, reliable detection of phenotype-defining transcripts – such as transcription factors, affinity receptors, and signaling molecules – by these methods is essential. Here, we describe a substantially improved massively-parallel scRNA-Seq protocol we term Seq-Well S^3 ("Second-Strand Synthesis") that increases the efficiency of transcript capture and gene detection by up to 10- and 5-fold, respectively, relative to previous iterations, surpassing best-in-class commercial analogs. We first characterized the performance of Seq-Well S^3 in cell lines and PBMCs, and then examined five different inflammatory skin diseases, characterized by distinct types of inflammation, to explore the breadth of potential immune and parenchymal cell states. Our work presents an essential methodological advance and a critical resource of the cellular and molecular features that inform human skin inflammation.

INTRODUCTION

4.1 Background

Although a nascent technology, single-cell RNA-sequencing (scRNA-Seq) has already helped define, at unprecedented resolution, the cellular composition of many healthy and diseased tissues.[1-5] The development of high-throughput methodologies has been crucial to this process, empowering the characterization of increasingly complex cellular samples. Unfortunately, current scRNA-Seq platforms typically demonstrate an inverse relationship between the number of cells that can be profiled at once and the amount of biological information that can be recovered from each cell. As a result, one must choose between quantity and quality – and thus comprehensiveness and fidelity – or alternatively employ two distinct approaches in parallel.[6] Indeed, inefficiencies in transcript capture among massively-parallel methods have limited our ability to resolve the distinct cell states that comprise broad cell types,[7] as well as their essential molecular attributes and often lowly-expressed molecular features, such as transcription factors, affinity receptors, and signaling molecules (**Figure 1A**).

Improving the fidelity of these methodologies is particularly important for resolving differences within heterogeneous populations of immune cells like lymphocytes and myeloid cells.[8] Here, subtle differences in surface receptor, transcription factor and/or cytokine expression can profoundly impact cellular function, particularly in the setting of human pathology.[9] Enhancing data quality in high-throughput scRNA-Seq would facilitate a greater appreciation of the underlying molecular features that describe such cellular variation. Similarly, it would ease integration with legacy datasets that often rely on lowly-expressed biomarkers, such as transcription factors, that are false-negative prone to discriminate subsets of cells.

Most high-throughput scRNA-Seq methods currently rely on early barcoding of cellular contents to achieve scale. Typically, these techniques recover single-cell transcriptomes for thousands of cells at once by leveraging reverse-emulsion droplets or microwells to isolate individual cells with uniquely barcoded poly-dT oligonucleotides which can then capture and tag cellular mRNAs during reverse transcription.[10] Afterward, an additional priming site is added to the 3' end of the synthesized cDNA to enable PCR-based amplification of all transcripts using a single primer (whole transcriptome amplification, WTA). A number of techniques have been described to add this second priming site.[11,12] The most common uses the terminal transferase activity of certain reverse transcription enzymes to facilitate a "template-switch" from the original mRNA to a second defined oligonucleotide handle.[13] While simple to implement, this process has the potential to be highly inefficient, leading to the loss of molecules that have been captured and converted to cDNA but not successfully tagged with a secondary PCR priming site (**Figure 1A; Appendix C, Figure S1A**).

To overcome these limitations, we have developed a new massively-parallel scRNA-Seq protocol we call Seq-Well S^3 (for "Second-Strand Synthesis"). Seq-Well S^3 increases the efficiency of the second PCR handle addition by amending it through a randomly-primed second-strand synthesis after reverse transcription (**Figure 1A**). Working with cell lines and peripheral blood mononuclear cells (PBMCs), we demonstrate that Seq-Well S^3 enables significant improvements in transcript and gene capture across sample types, facilitating studies of complex immune tissues at enhanced resolution (**Figure 1; Appendix C, Figures S1** and **S2)**.

To illustrate the utility of S^3, we apply it to generate a resource of single-cell transcriptional states spanning multiple inflammatory skin conditions. Skin represents the largest barrier tissue in the human body and is comprised of numerous specialized cell-types that help maintain both immunological and physical boundaries between our inner and outer worlds.[14] The dermis and epidermis – the two primary compartments of human skin – play complementary roles in tissue structure and function (**Figure 2A**).[14] The epidermis consists primarily of keratinized epithelial cells, which provide a physical barrier to the outside world; the dermis, meanwhile, provides structural support for the skin, with fibroblasts producing collagen and elastin fibrils along with the other components of the extracellular matrix. Crucially, within the cellular ecosystem of human skin, there are numerous tissue-resident immune and parenchymal cells essential to homeostatic barrier function. Using Seq-Well S^3, we examine the cellular composition of normal skin and altered cellular phenotypes in multiple inflammatory skin conditions, including acne, alopecia areata, granuloma annulare, leprosy and psoriasis. With conditions that span autoimmune (alopecia), autoinflammatory (psoriasis), reactive (acne), and granulomatous (granuloma annulare and leprosy) inflammation, we uncover a diverse spectrum of immune and parenchymal cellular phenotypes, as well as their molecular features, across multiple inflammatory skin conditions. Overall, our work presents an essential methodological advance as well as a critical resource for understanding how diverse inflammatory responses can impact a single tissue and the range of cellular phenotypes that are possible upon perturbation.

RESULTS

4.2 Second-Strand Synthesis (S^3) Leads to Improved Transcript Capture and Gene Detection

We hypothesized that use of “template-switching” to append a second PCR handle during reverse transcription might limit the overall recovery of unique transcripts and genes from individual cells in some massively-parallel scRNA-Seq methods such as Seq-Well and Drop-Seq.[2,15] Thus, we incorporated a randomly primed second-strand synthesis following first-strand cDNA construction (**Figures 1A; Appendix C, Figure S1A**). Briefly, after reverse transcription, barcoded mRNA capture beads are washed with 0.1 molar sodium hydroxide to remove attached RNA template strands and then a random second-strand synthesis is performed to generate double-stranded cDNA labeled on one end with the SMART sequence and its reverse complement on the other (**Figure 1A; Appendix C, Figure S1A; STAR* Methods**).[13,16]

To examine the effectiveness of Seq-Well S^3 and optimize its performance, we first tested a number of conditions using cell lines (**Appendix C, Figure S1B; STAR* Methods**). In these experiments, we observed that S^3 led to marked improvements in library complexity (Seq-Well V1: 0.22 transcripts/aligned read, Seq-Well S^3: 0.68 transcripts/ aligned read) and was able to function in the absence of a template switching oligo (TSO); Seq-Well V1, meanwhile, failed to generate appreciable product without a TSO (**Appendix C, Figure S1B-D**). In species-mixing experiments using HEK293 (human) and NIH-3T3 (mouse) cell lines, the use of the S^3 protocol resulted in significant increases in the numbers of unique transcripts captured and genes detected per cell compared to our original protocol for Seq-Well (P < 0.05, Mann-Whitney U Test; **Appendix C, Figure S1B; STAR* Methods**).

To fully understand how S^3 would perform on more challenging primary cells, we next applied it to human PBMCs (**Appendix C, Figure S1C** and **S2**; **STAR* Methods**), benchmarking against our original Seq-Well protocol as well as a commercial technology (10X genomics, V2 chemistry; hereafter 10x v2). For these comparisons, we downsampled all resulting data to an average of 42,000 reads per cell to account for differences in sequencing depth across technologies. Critically, Seq-Well S^3 resulted in significant improvements in the complexity of our sequencing libraries compared to 10x v2 as determined by the number of transcripts and genes detected at matched read depth (P < 0.05, Mann-Whitney U Test & Linear Regression**; Figure 1B-C**; **STAR* Methods**). To confirm that these overall improvements were not driven by changes in the relative frequencies of different cell types captured by each technology, we also examined each subset independently (**Appendix C, Figure S2A-B**). For each cell type detected, we observed significant increases in the numbers of transcripts captured and genes detected using S^3 for each pairwise comparison between techniques (P < 0.05, Mann-Whitney U Test; CD4$^+$ T cells, Seq-Well V1: 1,044 ± 62.3 UMIs/cell; 10x v2: 7,671 ± 103.9 UMIs/cell; Seq-Well S^3: 13,390 ± 253.4 UMIs/cell; Mean ± SEM) (**Appendix C, Figure S2**; **STAR* Methods**). Both Seq-Well S^3 and 10x v2 displayed increased sensitivity for transcripts and genes relative to Seq-Well v1, but Seq-Well S^3 showed the greatest efficiency (defined as genes recovered at matched read depth) to detect genes for each cell type (**Figure 1D-E**; **Appendix C, Figure S2**).

We sought to further understand whether these improvements resulted in enhanced detection of biologically relevant genes typically under-represented in high-throughput single-cell sequencing libraries.[6] Importantly, genes that were differentially detected (i.e., higher in S^3) within each cell type include numerous transcription factors, cytokines and cell-surface receptors (**Figure 1D-E**; **Appendix C, Table S1**).

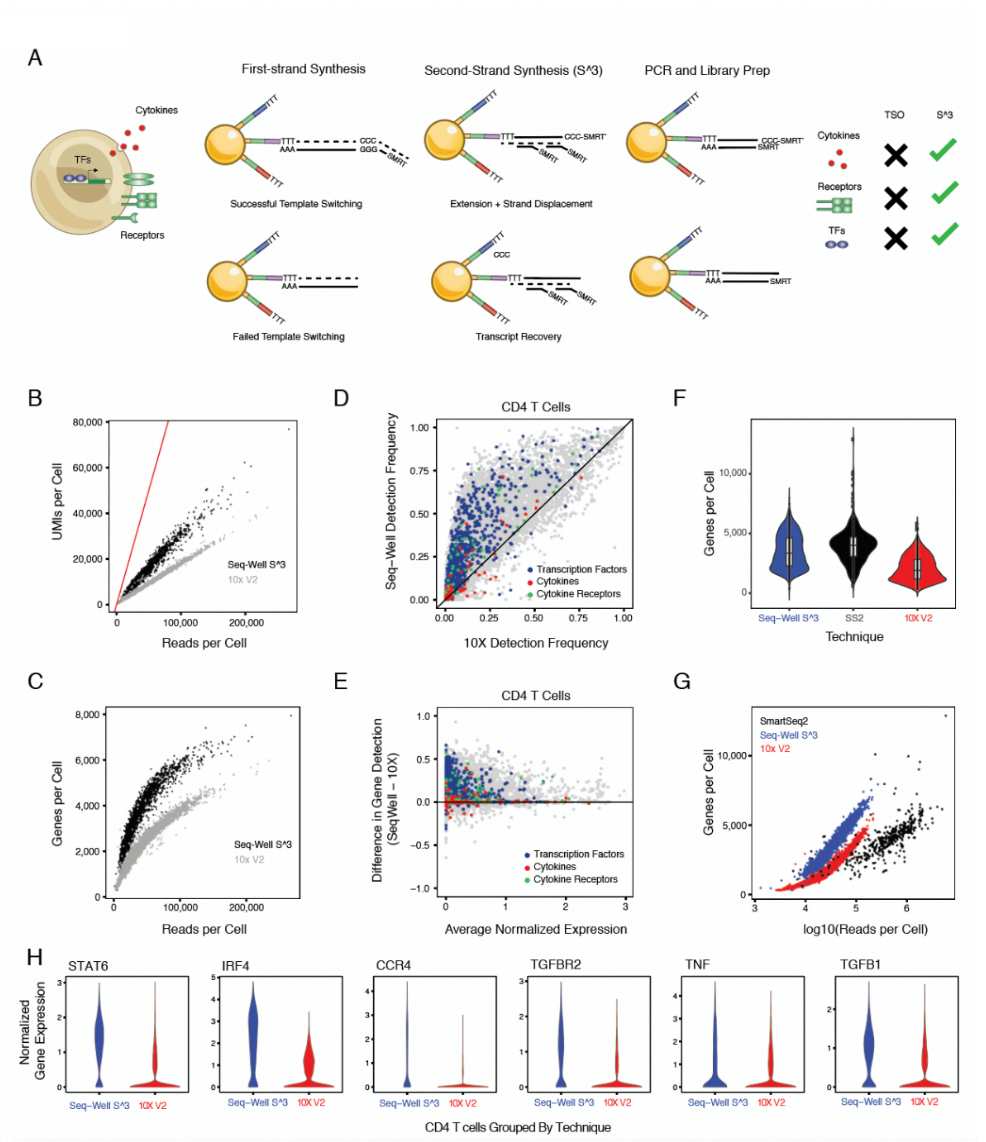


Figure 1 | Overview of Second Strand Synthesis (S^3). (A) Conceptual illustration of the molecular features that define immune phenotypes – including transcription factors, cytokines and receptors – as well as the Seq-Well second-strand synthesis method (Seq-Well S^3) and how it improves detection of key genes and transcripts. **(B)** Scatterplot showing differences in per-cell transcript capture (y-axis) as a function of aligned reads per cell (x-axis) between 10x Genomics v2 (grey) and Seq-Well S^3 (black). Red line indicates uniform line where transcripts per cell and aligned reads would be equivalent. **(C)** Scatterplot shows the differences in per-cell gene detection (y-axis) as a function of aligned reads per cell (x-axis) between 10x v2 (grey) and Seq-Well S^3 (black). **(D)** Scatterplot comparing gene detection rates in CD4$^+$ T cells between 10x v2 (x-axis) and Seq-Well S^3 (y-axis). Black line indicates point of equivalence in gene detection frequency between methods. Colors correspond to classes of genes including transcription factors (blue), cytokines (magenta), and receptors (green; **Table S1**). (**E**) Scatterplot comparing gene detection frequency (y-axis) between Seq-Well S^3 (positive values) and 10x v2 (negative values) as a function of the aggregate expression levels of an individual gene (x-axis). Black line indicates point of equivalence in gene detection frequency between methods. Colors correspond to classes of genes including transcription factors (blue), cytokines (magenta), and receptors (green; **Table S1**). **(F)** Violin plot (boxplots median +- quartiles) showing the distribution of per-cell transcript capture for Seq-well S^3 (blue; n = 1,485), 10x v2 (red; n = 2995), and Smart-Seq2 (black, n = 382). **(G)** Scatterplot showing the relationship between aligned reads and genes detected per cell between Seq-Well S^3 (blue), 10x v2 (red) and Smart-Seq2 (black) in sorted PBMC CD4$^+$ T cells. **(H)** Violin plots showing the distribution of normalized expression values for select transcription factors, cytokines and cytokine receptors between Seq-Well S^3 and 10x v2.

For example, among CD4^{+} T cells, we observe significantly increased detection of cytokines (e.g., *TGFB1* and *TNF*), surface receptors (e.g., *TGFBR* and *CCR4*) and transcription factors (e.g., *STAT6,* and *IRF4*) (P< 0.05, Chi-Square Test, **Figure 1H; Appendix C, Figure S2** and **Table S1**).

Finally, we performed an additional comparison of enriched human CD4^{+} T cells profiled using Seq-Well S^3 and 10X v2, as well as by Smart-Seq2, a commonly implemented microtiter plate-based approach (**Figure 1F-G; STAR* Methods**).[13] Integrated analysis of aggregate gene detection revealed that Seq-Well S^3 detects more genes per cell than 10x v2 and nearly as many genes per cell as Smart-Seq2 in pairwise comparison of techniques (10x v2: 2,057 ± 18.7 genes/cell , Seq-Well S^3: 3,514 ± 36.2 genes/cell , SS2: 3,975 ± 74.0 genes/cell; mean ± SEM) (P < 0.05, Mann-Whitney Test; **Figure 1F; STAR* Methods**). Further, comparing the frequency of gene detection between methods revealed crucial differences for transcription factors, cytokines and receptors/ligands (**STAR* Methods**). Surprisingly, we observe similar rates of gene detection between S^3 and Smart-Seq2 for a large number of biologically informative genes (**Appendix C, Figure S2F**). Critically, while comparable numbers of genes were detected across methods, Seq-Well S^3 detected more genes per aligned read than either 10x v2 or SS2 in pairwise comparisons (P<0.05, Mann-Whitney U Test; **Figure 1G; STAR* Methods**).

4.3 A Resource of Cellular States Across Healthy and Inflamed Skin

To demonstrate the utility of Seq-Well S^3 to comprehensively describe cellular states across human pathology at unprecedented resolution, we applied it to profile human skin samples spanning multiple, complex inflammatory skin conditions (**Figure 2**) – including acne, alopecia areata, granuloma annulare, leprosy, psoriasis – as well as normal skin (**Figure 2A-B; Appendix C, Figure S3A-C** and **Table S2; STAR* Methods**). In total, we processed nine skin biopsies by S^3 and, after data quality filtering, retained 20,903 high-quality single-cell transcriptomes (**Figure 2A-B; STAR* Methods**).

To examine similarities and differences among these cells across the high-dimensional gene expression space, we selected variable genes, performed UMAP dimensionality reduction, and identified 33 clusters through Louvain clustering in Scanpy[17] (**Figure 2; Appendix C, Figure S3A-C**; **STAR* Methods**). To collapse clusters to cell-types, we performed enrichment analyses to identify cluster-defining genes (**Appendix C, Figure S3B**) and then manually assigned cell-type identities based on the expression of known lineage markers (**Figure 2B; Appendix C, Table S3**; **STAR* Methods**). We also generated aggregate gene expression profiles and performed hierarchical clustering using a combined list of the top 50 cluster-defining genes for each cluster to further support our annotations and groupings (**Appendix C, Figure S3C**; **STAR* Methods**). Ultimately, we recovered a total of 16 primary cell-types, within which there was considerable heterogeneity. The identified cell types include: B cells (marked by expression of *MS4A1* and *CD79A*), dendritic cells (*FCER1G* and *CLEC10A*), endothelial cells (*SELE* and CD93), fibroblasts (*DCN* and *COL6A2*), hair follicles (*SOX9*), keratinocytes (*KRT5* and *KRT1*), macrophages (*CD68* and *CTSS*), mast cells (*CPA3* and *IL1RL1*), muscle (*NEAT1* and *KCNQ1OT1*), plasma cells (*IGHG1*), Schwann cells (*SCN7A*), and T cells (*CD3D* and *TRBC2*) (**Figure 2b**; **Figure S3A-E** and **Table S3**). We next sought to define nuanced cell states within these immune, stromal and parenchymal populations – including T cells, myeloid cells, endothelial cells, dermal fibroblasts, and keratinocytes – across the spectrum of skin inflammation.

4.4 Seq-Well S^3 describes T cell states across inflammatory skin conditions

To determine the range of biological diversity that can be captured using Seq-Well S^3, we first focused on further characterizing T cells across the inflammatory skin conditions examined since each is known to significantly skew T cell phenotypes (**Figure 3**).[18,19] We performed dimensionality reduction and sub-clustering across T cells alone (**Figure 3A-B; STAR* Methods**). Our analysis revealed nine sub-clusters that closely correspond to NK cells and CD8 T cells, as well as several known CD4$^+$ T-helper cell (Th) subsets. As before, we used the enhanced sensitivity of S^3 for lineage defining transcripts to help annotate the identity of each sub-cluster; for example, in T cell sub-clusters 5 and 6, respectively, we detected distinct expression of canonical regulatory T cell and Th17 T cell transcription factors (e.g., *FOXP3* and *RORC,* respectively*)* and immune receptors (e.g. *TIGIT* and *CXCR6* respectively) (**Figure 3C-E**; **Figure S4 and Table S4**).

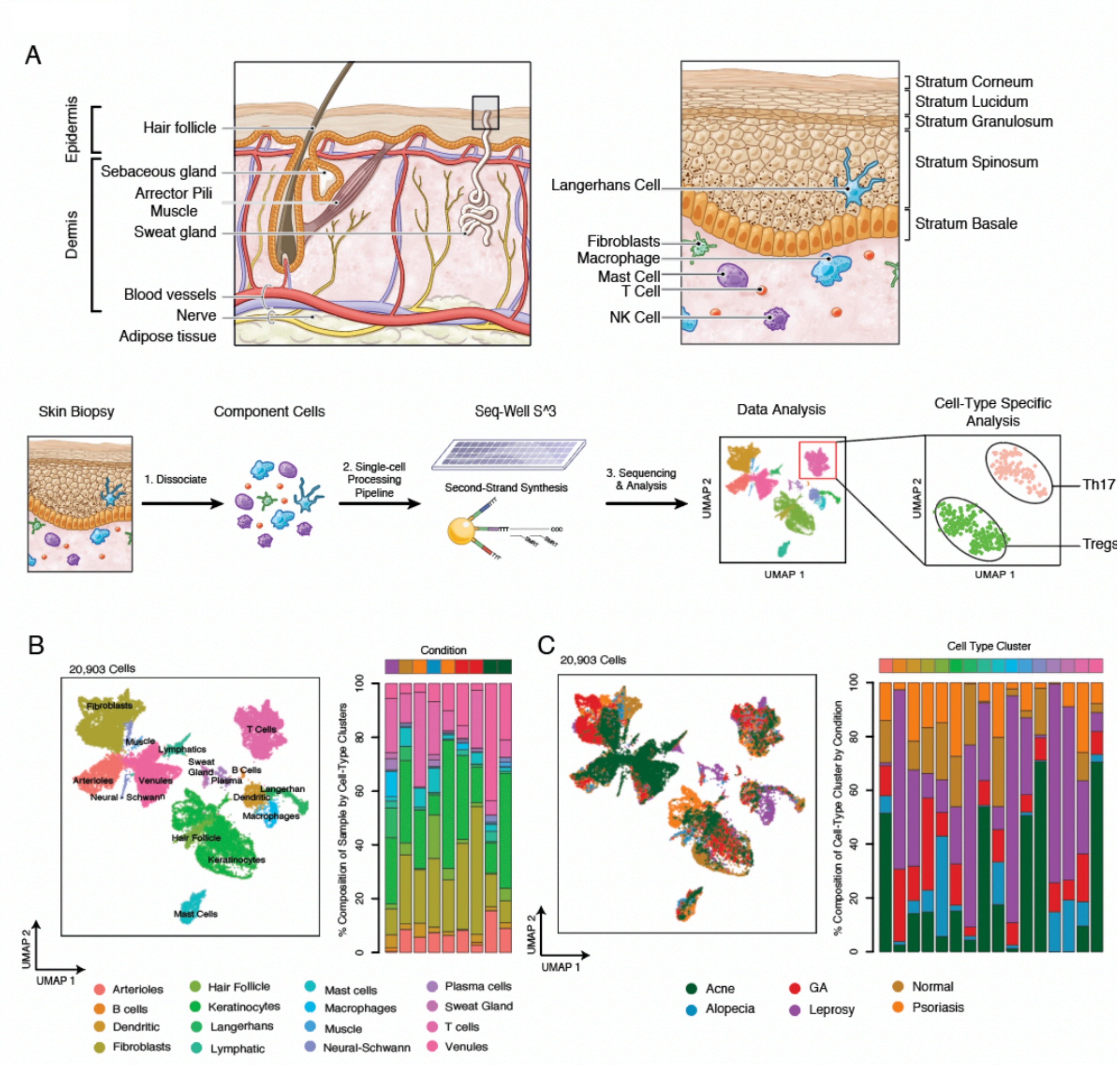


Figure 2 | Cell Types Recovered across Inflammatory Skin Conditions. (**A**) (**Top-Left**) Diagram illustrating the anatomic organization and major features of human skin. (**Top-Right**) Cell-type composition of the epidermis and dermis. (**Bottom**) Sample processing pipeline used to generate a collection cellular states across skin inflammation. (**B**) (**Left**) UMAP plot for 20,903 cells colored by cell-type cluster. (**Right**) Stacked barplot showing the cell-type composition for each of the nine skin biopsies. (**C**) (**Left**) UMAP plot for 20,308 cells colored by inflammatory skin condition. (**Right**) Stacked barplot showing the proportion of cells from each skin condition within phenotypic clusters.

Additionally, we cross-referenced each sub-cluster's marker genes against a series of curated signatures in the Savant database[20] to confirm our assignments. This analysis highlighted similarity to previously characterized T cell and NK cell populations (**Appendix C, Figure S4B; STAR* Methods**).

We next examined T cell phenotypes across inflammatory skin conditions to explore variability in T cell subset composition by skin pathology (**Figure 3B**). This analysis revealed potentially varied contributions to different classes of cutaneous inflammation. For example, sub-cluster 6 is enriched for expression of canonical Th17 genes including *RORC*, which encodes the Th-17 lineage-defining transcription factor RORγt[21] and is observed predominantly within the leprosy sample. While either Th1 or Th2 responses are typically thought to predominate in the immune response to leprosy, a role for Th17 cells in controlling disease has been previously demonstrated.[22] We further found that sub-cluster 1, which express *NR4A1*, a transcription factor that is a marker of dysfunctional T cells[23], and sub-cluster 3, enriched for genes involved in nuclear organization (*ANKRD36*, *XIST*, and *NEAT1*), were over-represented in both patients from psoriasis (**Figure 3B-D**). In alopecia areata, we detected a unique population of T cells (sub-cluster 7) that overexpress *PDE4D,* which has been shown to plays a role in TCR-dependent T cell activation (**Figure 3C**; **Appendix C, Supplemental Table S4**).[24]

We also uncovered considerable variation across cytotoxic T cells and NK cells. Directed analysis within CD8 T cells (sub-cluster 0) revealed a sub-grouping of activated CD8 T cells that express elevated levels of several inflammatory cytokines (*TNF*, *CCL4*, and *XCL1),* as well as specific affinity receptors (*FASLG* and *TNFRSF9*) and transcription factors (*KLF9* and *EGR2*); this phenotypic skewing was observed primarily in a patient with granuloma annulare (**Figure 3F; Figure S4B** and **Table S5; STAR* Methods**). Meanwhile, we found the highest degree of cytotoxic gene expression (*GNLY*, *GZMB*, and *PRF1*) among cells in sub-cluster 8, suggesting that this sub-cluster may represent a diverse set of NK cells, gamma-delta T cells, and activated cytotoxic T cells. Indeed, further analysis of sub-cluster 8 revealed 3 distinct component sub-groups of cytotoxic cells: a sub-group of CD8$^+$ T cells (T.8.1; *TNFSF8*, *SLAMF1*, *CLEC2D*, *CD5*) expressing various TCR genes; a second sub-group of CD16+ cells (T.8.2) expressing cytotoxic effector molecules (*GNLY*, *PRF1*, *GZMB*) and NK surface receptors, consistent with either NK cell or tri-cytotoxic CTL; and a third sub-group of NK cells (T.8.3) enriched for

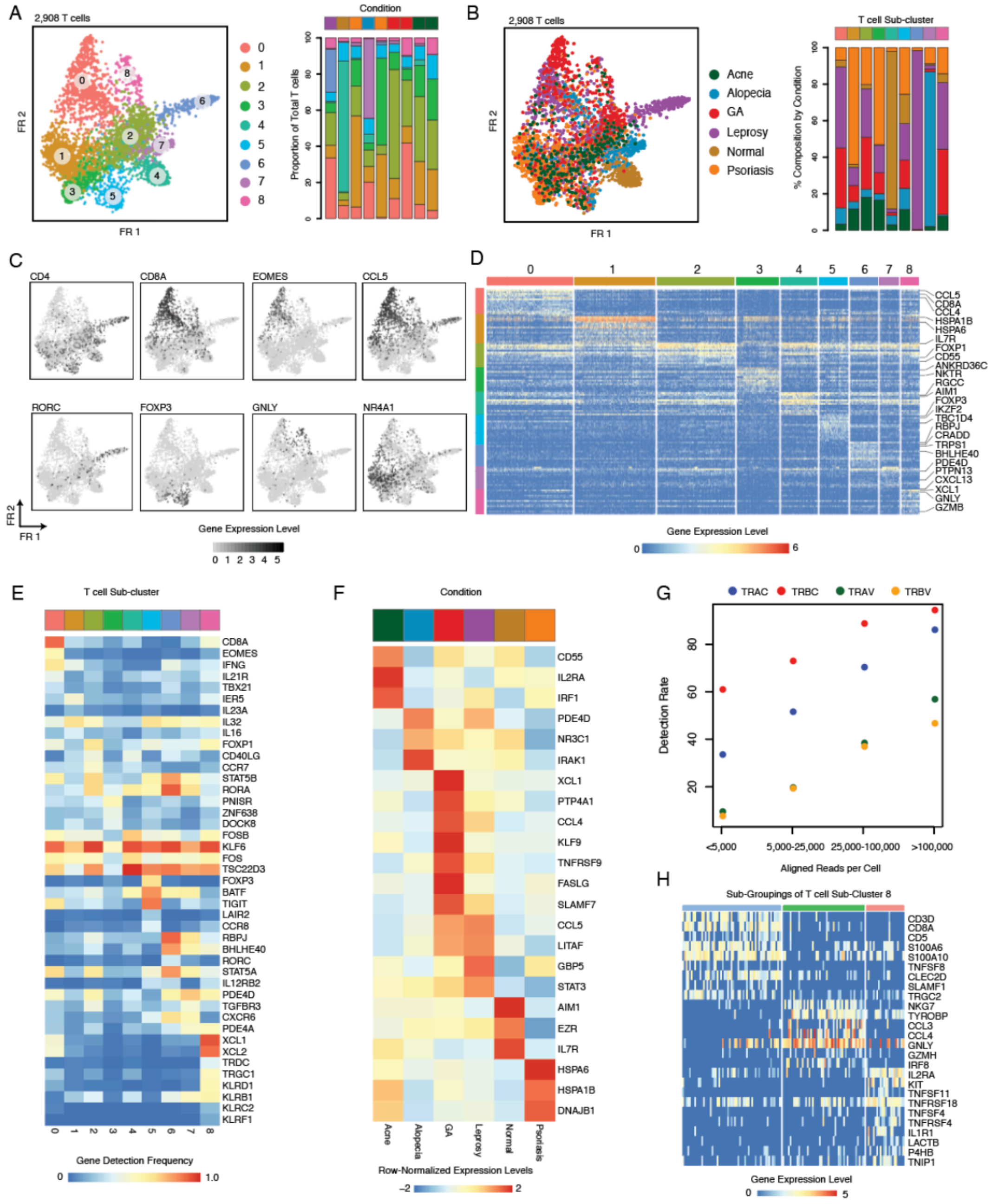


Figure 3 | Identification of Inflammatory T cell States using Seq-Well S^3. (**A**) (**Left**) Force-directed graph of 2,908 T cells colored by the nine phenotypic sub-clusters identified by Louvain clustering. (**Right**) Stacked barplots showing the distribution of these T cell sub-clusters within each skin biopsy. (**B**) (**Left**) Force-directed graph of 2,908 T cells colored by inflammatory skin condition. (**Right**) Stacked barplots showing the contribution of each inflammatory skin condition to the T cell sub-clusters. (**C**) T cell force-directed graphs displaying log-normalized expression of a curated group of sub-cluster-defining gene. Higher expression values are shown in black. (**D**) Heatmap showing log-normalized gene expression values for a curated list of sub-cluster-defining genes across nine T cell sub-clusters. (**E**) Heatmap showing the rate of detection for lineage-defining transcription factors, cytokines, and cytokine receptors across T cellphenotypic clusters. (**F**) Heatmap showing average expression of genes enriched across T cells by inflammatory skin condition (row-normalized average expression values). (**G**) Plot showing rates of detection of TCR genes from human skin T cells across a range of sequencing depths. (**H**) Heatmap showing normalized gene expression values for genes enriched in the sub-group analysis of T cell sub-cluster 8.

expression of *c-KIT*, *RANKL* (TNFSF11) and *GITR* (TNFSFR18) (**Figure 3H** and **S4B; Table S6**).[25]

Profiling of T cell receptor expression is critical to understand T cell antigen specificity.[26] Importantly, among CD4$^+$ T cells obtained from peripheral blood, we recovered most TCR-V and TCR-J genes at a higher frequency using Seq-Well S^3 as compared to 10x v2 (P< 0.05, Chi-square Test; **Appendix C, Figure S4C; STAR* Methods**). Among CD4$^+$ T cells from peripheral blood, we observed paired detection of TRAC and TRBC in 1,293 of 1,485 CD4$^+$ T cells (87.1% Paired Detection Rate; **Appendix C, Figure S4C**). In the setting of skin inflammation, we explored TCR detection rates across a range of sequencing read depths. Overall, we detected TRAC in 54.5%, TRBC in 75.5%, and paired detection in 46.4% of T cells (**Figure 3G**). Among T cells with at least 25,000 aligned reads, we recovered paired alpha and beta chains in 66.7%. Among cells from sub-cluster 8, we observe expression of gamma and delta constant genes (TRGC and TRDC), while remaining T cell clusters exclusively express alpha and beta TCR constant genes (**Appendix C, Figure S4C**). These data further suggest that sub-cluster 8 represents a diverse population of gamma delta, NK, and cytotoxic CD8 T cells that share common gene expression features and, potentially, roles in inflammation.

4.5 Spectrum of Myeloid Cell States in Skin Inflammation

In the setting of cutaneous inflammation, myeloid cells play a key role in maintaining tissue homeostasis, wound healing and response to pathogens.[27] Using Seq-Well S^3, we were able to identify numerous myeloid cell subpopulations defined by a combinations of surface markers, cytokines and lineage-defining transcription factors. Specifically, we independently analyzed 2,371 myeloid cells and identified nine sub-clusters representing 4 primary myeloid cell types based on expression of canonical lineage markers and comparison to cell-type signatures in the Savant database: dendritic cells (*CLEC10A*), Langerhans cells (*CD207* and *CD1A*), macrophages (*CD68* and *CD163*), and mast cells.

Skin functions as both a physical and immunologic barrier, and is the primary site of exposure to environmental antigens. As such, multiple types of antigen-presenting cells

(APCs) are distributed in both the dermis and epidermis. In the epidermis, there is a specialized population of antigen-presenting cells known as Langerhans cells. We initially identified Langerhans cells on the basis of expression of canonical markers (*CD207*, *CD1A*; **Figure 4C-D; Appendix C, Table S7**).[30] For biopsies obtained from normal skin and leprosy, we performed MACS enrichments from the epidermal section and loaded Langerhans cells as 5% of the total amount to increase recovery (**STAR* Methods**). When we directly compared Langerhans cells from leprosy and normal skin, we observed elevated expression of *IDO1*, *STAT1*, *HCAR3* and MHC class I molecules (*HLA-A*, *HLA-B* and *HLA-F*) in Langerhans cells in leprosy infection, which may suggest a role for Langerhans cells in priming CD8 T cell responses in this disease (**Figure 4E**; **Appendix C, Table S8**).[31,32]

Additionally, we found a large sub-group of dermal dendritic cells (**Figure 4A**). Further analysis of the *CD207*-negative dendritic cell sub-cluster revealed multiple sub-groupings of dermal dendritic cells across skin biopsies. Consistent with previous observations from peripheral blood[8], we saw a sub-group of dendritic cells that corresponds to cDC1 (*CLEC9A*, *IRF8*, and *WDFY4*) ($P<0.05$, Permutation Test; **Appendix C, Figure S4H; STAR* Methods**). We further report another sub-group that represents cDC2 cells (*IRF4*, *SOCS2*, *SLCO5A1, CD1B, CD1E*) (**Figure 4B-C**; **Figure S4F-H**; **STAR* Methods**).[33] Importantly, we detect expression of IL12B, a subunit of the IL-23 cytokine, within the sub-group of IRF4+ cDC2 cells (**Appendix C, Figure S4I-J**), which have previously been shown to promote mucosal type 17 inflammation via secretion of IL-23.[34] Further, this sub-grouping of cDC2 cells express high levels of *CCL17* and *CCL22*, chemokines involved in T cell chemotaxis (**Figure S4J**).[35]

We further identified three sub-groups of dermal dendritic cells that are broadly distinguished from conventional dendritic cell clusters by expression of *CLEC10A* (**Appendix C, Figure S4J**)*,* which has been shown to influence T cell cytokine responses in skin.[36,37] Cells from dermal DC sub-group 1 show elevated expression of *CD44*, *IL8*

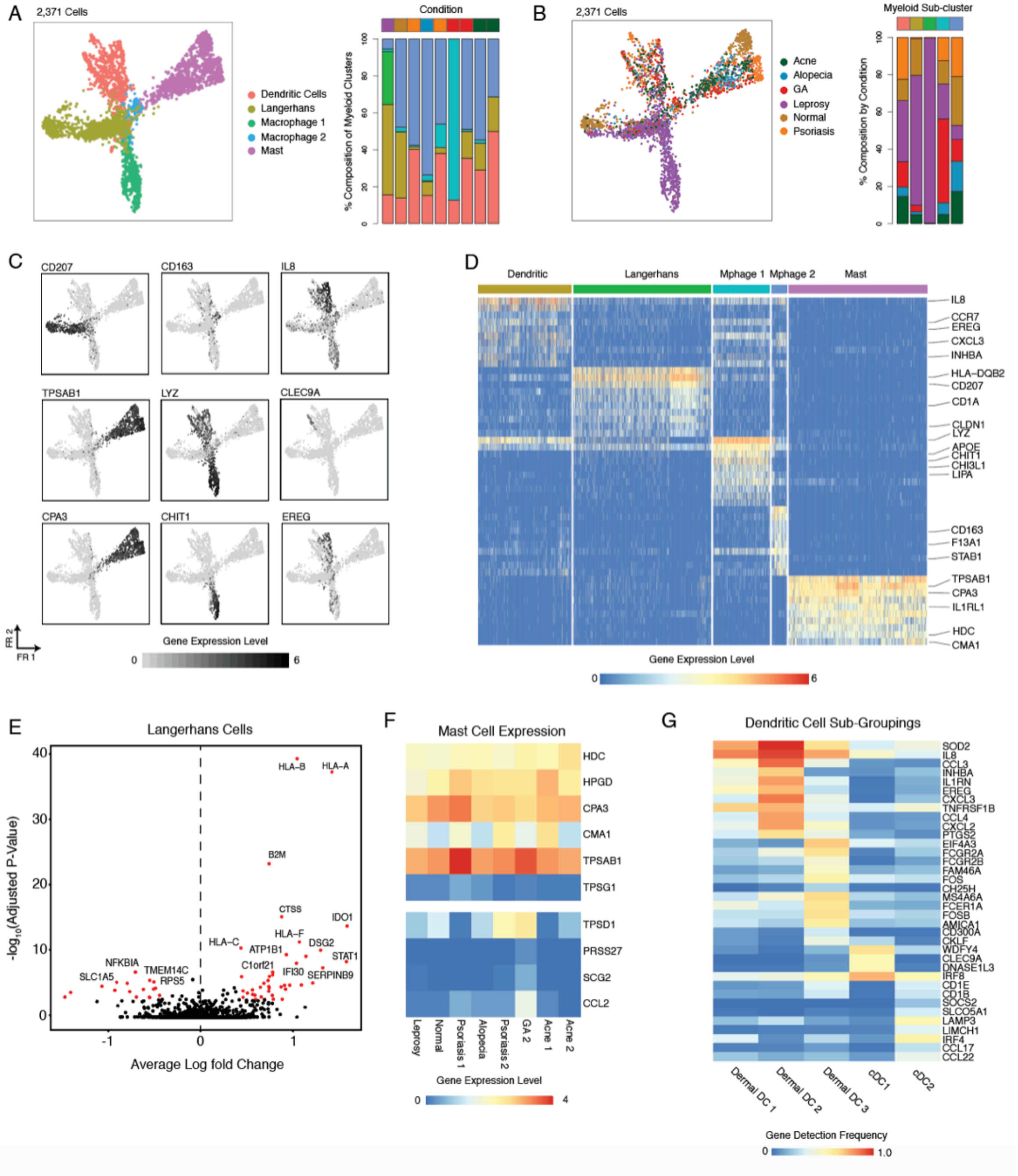


Figure 4 | Diverse Myeloid Cell States Uncovered using Seq-Well S^3. (A) (**Left**) Force-directed graph of 2,371 myeloid cells colored by five phenotypic sub-clusters. (**Right**) Stacked barplots showing the distribution of myeloid sub-clusters within each skin biopsy. **(B)** (**Left**) Force-directed graph of 2,371 myeloid cells colored by inflammatory skin condition. (**Right**) Stacked barplots showing the contribution of each inflammatory skin condition to each myeloid sub-cluster. **(C)** Force-directed graphs of 2,371 myeloid cells that highlighting expression of a curated group of sub-cluster defining genes. **(D)** Heatmap showing the expression of a curated list of myeloid cell-type cluster-defining genes. **(E)** Volcano plot showing genes differentially expressed in Langerhans cells between leprosy (n_{cells} = 56) and normal skin (n_{cells} = 120). Log10-fold change values are shown on the x-axis and -log10 adjusted p-values are shown on the y-axis. **(F)** Heatmaps showing the expression of mast-cell proteases across inflammatory skin conditions. **(G)** Heatmap showing detection frequencies for transcription factors, surface receptors, and cytokines across DC sub-populations.

and *SOD2* (**Figure 4G; Appendix C, Figure S4I and TableS9**). Cells from dermal DC sub-group 2 display elevated expression of pro-inflammatory chemokines up-regulated during DC maturation (*CXCL3*, *CCL2* and *CCL4*)[38] and soluble mediators (*EREG* and *INHBA*). Finally, a third sub-grouping of dermal DCs (Dermal DC3) was distinguished by expression of *FCER1A*, *FCGR2A*, and *FCGR2B*, which are important for interfacing with humoral immunity (**Appendix C, Figure S4I**).[39]

In the skin, mast cells are most commonly associated with allergic responses, but mast cell proteases serve additional roles in inflammation and pathogen defense.[40] Among skin mast cells, we detect core expression of *HDC* (Histidine decarboxylase), *HPGD*, and *TPSAB1* (Tryptase alpha/beta 1) (**Figure 4F**).[41] Importantly, we observe variable expression of mast cell proteases *TPSD1* (Tryptase D1) and *CMA1* (Chymase A1), which are primary mast cells effector molecules[40], which may have functional consequences. By performing analysis across inflammatory conditions and patients, we identify a distinct pattern of mast cells with elevated expression of proteases (*TPSD1*, Tryptase D1 and *PRSS27*, serine protease 27), *SCG2* (secretogranin 2), and *CCL2* in a patient with granuloma annulare (**Figure 4F).**

4.6 Detection of Endothelial Heterogeneity and Vascular Addressin Expression

Multiple types of endothelial cells exist within the dermis of the skin. As in most tissues, arterioles shuttle oxygenated blood to tissues terminating in a capillary bed that gives rise to post-capillary venules. Importantly, DARC+ post-capillary venules are the primary site of egress of immune cells from circulation into tissues.[42] Using the improved sensitivity of Seq-Well S^3, we sought to understand the spectrum of endothelial cell diversity and vascular addressin expression across multiple instances of skin inflammation.[43] We performed sub-clustering and dimensionality reduction across 4,996 endothelial cells (**Figure S5A-B** and **STAR* Methods**) and identified three primary sub-clusters of dermal endothelial cells defined by distinct expression patterns: vascular smooth muscle (*TAGLN*), endothelial cells (*CD93*) and lymphatic endothelial cells (*LYVE1*) (**Figure S5C**). Importantly, we found multiple sub-clusters of CD93+ endothelial across normal and inflamed skin biopsies (**Appendix C, Figure S5A-B**). For example, we observe two distinct populations of endothelial cells: a population of DARC-negative, CD93+ endothelial cells (Venule sub-cluster 3) that displays elevated expression of *SLC9A3R2*, which is involved in endothelial homeostasis (Bhattacharya et al., 2012), and another cluster of proliferating endothelial cells (Venule sub-cluster 4) (**Appendix C, Figure S5D**).

Further, we sought to understand the distribution of vascular addressins expressed by DARC+ endothelial sub-populations, the site primary site of lymphocyte egress into tissues (**Appendix C, Figure S5E**).[44] Notably, across sub-populations of CD93+ endothelial cells (Venule sub-clusters 1-4), we observe variation in expression of vascular addressins (**Appendix C, Figure S5E**). Among post-capillary venules, we observe broadly elevated expression of *ITGA5*, *ITGA6*, *ITGB4*, *ICAM2*, and *ITGA2*, while arterioles express higher levels of *ITGA7*, *ITGA8*, and *ITGB5*. Further, we observe the highest expression of *ITGA4*, *ITGA9*, *ITGB2* and *ITGB8* among lymphatic endothelial cells (**Figure 5E**).

4.7 Altered Dermal Fibroblast Identities in Skin Inflammation

Dermal fibroblasts provide structural support and are the primary source of extracellular matrix components within the skin. Previous studies have demonstrated significant variation among dermal fibroblasts based on their relationship to anatomic features of the skin.[45,46] To deeply catalogue diverse fibroblast cell states across inflamed skin, we performed dimensionality reduction and sub-clustering within the 4,189 fibroblasts identified across all samples and conditions (**Appendix C, Figure S5F-G; STAR* Methods**). In comparison to inflamed biopsies, fibroblasts from normal skin display enrichments in *LTBP4*, *IGFBP5*, and *TCF4*. Consistent with previous single-cell studies of dermal fibroblasts, we observe a sub-population of fibroblasts (Cluster 6) that express *COL11A1*, *DPEP1* and *RBP4,* where these cells were suggested to have a role in connective tissue differentiation (**Appendix C, Figure S5H**).[47]

Fibroblasts from GA patient 1 (sub-cluster 2) express elevated levels of *SPOCK1*, *CRLF1,* and *COMP*, a cartilage protein that is upregulated in matrix-producing fibroblasts following myocardial infarction[48] (**Appendix C, Figure S5H-I**). Further, fibroblasts from GA patient

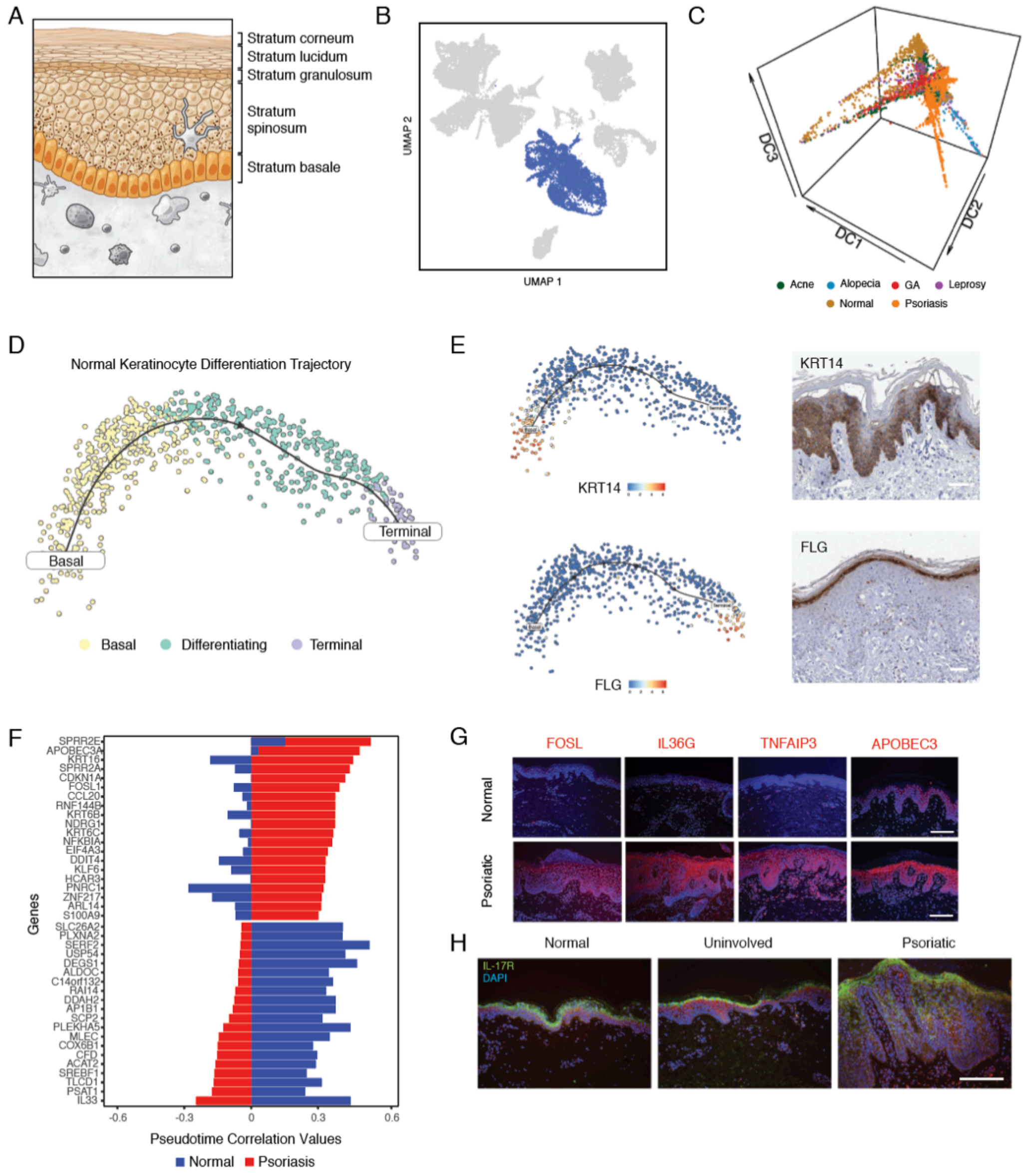


Figure 5 | Keratinocyte Differentiation Trajectories. (**A**) Diagram showing the layers of the epidermis and morphologic changes associated with keratinocyte differentiation. (**B**) UMAP embedding of 20,903 cells with all keratinocyte and hair follicle populations highlighted in blue. (**C**) Diffusion map of 5,141 keratinocytes colored by inflammatory skin condition. (**D**) Plot showing differentiation trajectory of keratinocytes from normal skin from basal cells (yellow) through differentiating cells (aqua) and terminal keratinocytes (purple). (**E**) (**Top-left**) tSNE plot of normal keratinocytes with normalized *KRT14* expression values overlayed. (**Top-right**) Immunohistochemistry staining showing the expression of *KRT14* from the human protein atlas.[56] (**Bottom-left**) tSNE plot of normal keratinocytes with normalized *FLG* expression values overlayed. (Bottom-**right**) Immunohistochemistry staining of *FLG* from the human protein atlas.[56] Scale bars = 50 microns. (**F**) Stacked barplot showing genes with the highest differential pseudo-time correlation between normal keratinocytes (blue) and psoriatic keratinocytes (red) sorted by correlation values in psoriatic keratinocytes. Correlation values shown on the x-axis represent Pearson correlation coefficients between normalized gene expression and diffusion pseudotime. (**G**) (**Top**) Immunofluorescence staining in normal (above) and psoriatic (**below**) for FOSL, IL36G, TNFAIP3, and APOBEC3. All images stained for nuclei (DAPI) and gene of interest (Red Fluorescence). Scale bar = 100 microns. (**H**) Immunofluorescence staining for IL-17R expression (green) in normal (**left**), uninvolved (**middle**), and psoriatic skin (**right**). Scale bar = 100 microns.

2 (sub-cluster 0) display elevated expression of protease inhibitor 16 (*PI16*), which inhibits the function of MMP2[49], and *ITIH5,* a protease inhibitor important for maintenance of dermal hyaluronic acid that is overexpressed in skin inflammation (**Appendix C, Figure S5H-I**).[50] Finally, among fibroblasts from acne patients, we observed elevated expression of multiple metallothioneins (**Appendix C, Figure S5H-I**). Specifically, the expression levels of *MT1E* and *MT2A* are highest in fibroblasts and endothelial cells in acne (**Appendix C, Figure S5H**). As seen among endothelial cells, fibroblast expression patterns in acne are consistent with a wound healing response.[51]

4.8 Keratinocyte Differentiation Trajectories

Within the epidermis, keratinocytes undergo a stereotyped differentiation process in which cells acquire altered morphology and phenotype as they mature (**Figure 5A-C**).[52] Under physiologic conditions, basal keratinocytes are characterized by expression of *KRT14* and *TP63*, and continuously divide to give rise to the remaining cells of the epidermis.[53] Using keratinocytes from normal skin, we performed pseudo-temporal analysis to reconstruct the differentiation process of normal epidermal keratinocytes (**Figure 5D**; **STAR* Methods**). More specifically, in normal skin, we first identified a population of keratinocytes enriched for expression of *TP63* and *KRT14*, markers of basal keratinocytes (**Appendix C, Figure S6B**).[54] We then used known patterns of cytokeratin expression to infer localization of keratinocytes along a supervised differentiation trajectory (**Figure 5E**; **Appendix C, Figure S6A**).[4] Our trajectory analysis revealed patterns of transcription factor and cytokeratin expression that closely correspond to previously established signatures of keratinocyte maturation.[55] Consistent with immunohistochemical staining from the Human Protein Atlas (**Figure 5E**)[56], we observed enrichment of filaggrin (*FLG*), a protein in the outer layers of the epidermis[57], *mRNA* among keratinocytes that lie at the terminal points in the pseudo-temporal ordering (**Figure 6E; Appendix C, Figure S6B**).

Having established a trajectory for normal keratinocyte differentiation, we next examined patterns of keratinocyte differentiation across pathologic conditions. To identify conserved and unique patterns across conditions, we constructed a combined diffusion map using the 5,141 keratinocytes recovered across all samples (**Figure 5C**; **STAR* Methods**). While keratinocytes from most conditions closely align with normal differentiation, we observe marked deviation in the differentiation trajectory of psoriatic keratinocytes (**Figure 5C**). Consistent with previous observations, differential expression analysis reveals significant up-regulation of antimicrobial peptides (*S100A7*, *S100A8*, *S100A9*) and pro-inflammatory cytokines (*IL36G*, *IL36RN*) in psoriatic keratinocytes (**Appendix C, Table S12**).[58]

Based on increased sensitivity of Seq-Well S^3 to detect transcription factors observed in peripheral lymphocytes, we hypothesized that our data might enable identification of novel transcriptional regulators of psoriatic keratinocytes. To identify potential drivers of the psoriatic disease process within the epidermis, we performed differential pseudo-time correlation analysis between psoriatic and normal keratinocytes (**STAR* Methods**). Specifically, we separately constructed pseudo-time trajectories for normal and psoriatic keratinocytes, calculated correlation values between diffusion pseudo-time and gene expression levels, and examined the difference in correlation values between psoriatic and normal keratinocytes (**Figure 5F**; **Appendix C, Figure S6A-B and Table S13**). Notably, we observed positive correlation of *FOSL1*, an AP-1 transcription factor, with diffusion pseudo-time in psoriatic keratinocytes, implying that *FOSL1* is preferentially expressed along the differentiation trajectory of psoriatic keratinocytes. To validate this observation, we performed immunofluorescence staining for FOSL1 protein, and measured increased levels of FOSL1 in psoriatic skin (**Figure 5G,** STAR* Methods). We further validated the distribution of additional genes overexpressed or differentially correlated with diffusion pseudo-time in psoriatic keratinocytes (including *TNFAIP3, IL36G*, and *APOBEC3*) at the protein level (**Figure 5G**; **Appendix C, Figure S6A** and **Tables S12-13**; **STAR* Methods**).

To further define differences in gene expression patterns between normal and psoriatic keratinocytes, we scored the expression levels of known cytokine response signatures using a series of reference signatures gene lists derived from population RNA-Seq of cultured keratinocytes exposed to IL-17 (**Appendix C, Figure S6C** and **Table S14**,

STAR* Methods). While IL-17 has been previously implicated in the pathogenesis of psoriasis, here we infer the identity of cells that dominate the IL-17 response, localizing the expression of IL-17 responsive genes to spinous keratinocytes.[4] To validate this observation, we performed immunofluorescent staining for IL-17R protein and measured the highest staining within spinous keratinocytes exclusively within psoriatic skin (**Figure 5H; STAR* Methods**). Collectively, these data provide novel insights into the localization IL-17 response in psoriatic keratinocytes.

4.9 Discussion

Here, we present an enhanced technique for high-throughput scRNA-Seq – Seq-Well S^3 – that affords improved sensitivity for transcript capture and gene detection. Through use of a templated second-strand synthesis, S^3 recovers information typically lost in bead-based high-throughput scRNA-Seq protocol such as Seq-Well or Drop-Seq. Specifically, S^3 reclaims mRNA molecules that are successfully captured and reverse transcribed but not labeled with a second primer sequence through template switching (**Figure 1**; **Appendix C, Figure S1**). Using Seq-Well S^3, we obtain a 5-10 fold increase in the number of unique molecules captured from cells at similar sequencing depth relative Seq-Well v1 (**Figure 1**; **Appendix C, Figures S1** and **S2**).[15] Beyond aggregate increases in the number of transcripts recovered per-cell, the improvements in sensitivity made possible by Seq-Well S^3 enable enhanced detection, and thus deeper examination, of lineage-defining factors in immune and parenchymal cells – such as transcription factors, cytokines, and cytokine receptors among lymphocytes (**Figure 1**; **Appendix C, Figures S2**) – which are often transiently or lowly expressed.[59] Among $CD4^+$ T cells isolated from PBMCs, for example, we observed rates of gene detection similar to those observed in Smart-Seq2, a best-in-class microtiter plate-based method (**Figure 1F-G**; **Appendix C, Figure S2F**).

Similarly, using Seq-Well S^3, we report improved paired detection of alpha and beta TCR sequences from T cells in peripheral blood and tissue biopsies (**Figures 3G**; **Appendix C, Figure S4C**). Among $CD4^+$ T cells from PBMCs, we recover paired TCR alpha and beta constant genes in 87.1% of cells. Together with targeted enrichment, amplification and sequencing, we anticipate that Seq-Well S^3 will enable improvements in TCR reconstruction and deep characterizations of clonotype-phenotype relationships at scale.[26] Collectively, our validation experiments show that Seq-Well S^3 significantly augments the amount of information that can be recovered in massively-parallel scRNA-seq experiments, enabling high-resolution profiling of low-input biopsy samples at scale.

With this enhanced method, here, we move towards a draft atlas of human skin inflammation by creating a compendium of cell-types and states for the broader research community.[60] Through use of Seq-Well S^3, we survey, at unprecedented resolution, the diversity of cell-types and states – e.g., among tissue resident T cells and myeloid cells – present across multiple types of skin inflammation. For example, GA and leprosy are two granulomatous diseases characterized by aggregates of lymphocytes and macrophages within the dermis, which are both thought to arise from a delayed-type hypersensitivity response to *M. leprae* infection (leprosy) and an unknown agent (GA).[61,62] Here, we find that both are characterized by the presence of T cell sub-cluster 0 (Immature CD8-CTL) and T cell sub-cluster 8 (mature CTL effectors containing CD8+ T-CTL, $\gamma\delta$ and NK cells; **Figure 3**). Although both conditions contain CD163+ dermal macrophages and various DC subpopulations, M1-like macrophages were present only in leprosy, which host the intracellular pathogen *M. leprae*.[63,64] Moreover, GA uniquely contained specific populations of fibroblasts expressing *SPOCK1*, *CRLF1,* and *COMP* (**Appendix C, Figure S5**), which likely reflect remodeling of the dermis with mucin deposition and alternation of elastin fibers.[65,66]

Acne, meanwhile, is an inflammatory disease thought to arise in response to infection with *P. acnes*, resulting in the formation of lesions that resemble a wound following eruption of the hair follicle into the dermis.[67] Here, we observe 2 clusters of endothelial cells marked by expression of *SLC9A3R2*, a marker of endothelial homeostasis, and a signature of proliferation (**Venule clusters 3 and 4**; **Appendix C, Figure S5**). This increased angiogenesis and endothelial proliferation is most consistent with the proliferative phase of wound healing in acne.[68]

Alopecia areata and psoriasis both arise from autoimmune and autoinflammatory processes, yet there were distinct differences in their underlying cell states. For example, alopecia areata is thought to be driven by a population of CD8 T cells that target hair follicles.[69] Notably, in alopecia, we report a sub-cluster of T cells characterized by expression of *PDE4D* (**Figure 3**). PDE4 inhibitors have recently shown demonstrated efficacy in the treatment of alopecia[20,70], and it is intriguing to speculate that these inhibitors might work by targeting this subset of T cells.

In psoriasis, T cells are thought to be a primary driver of inflammation, with dendritic cells playing a central role in the recruitment and polarization of T cells that contribute to the hyperproliferation of keratinocytes in psoriasis.[19] In both patients with psoriasis, we report a sub-cluster of DCs (IRF4+ cDC2) that display elevated expression of *CCL17*, *CCL22* and *IL12B* (**Figure 4G**; **Appendix C, Figure S4I**). Importantly, a similar population of dermal cDC2 cells has recently been shown to drive psoriatic inflammation in mice and humans through the recruitment of inflammatory T cells.[71,72] Although we detected a diversity in T cell subtypes in psoriatic lesions, we note few Th17-like cells.[73]

Leveraging the increased sensitivity of Seq-Well S^3, we performed pseudo-time correlation analysis to uncover an altered differentiation trajectory of keratinocytes compared to normal skin (**Figure 5**; **Appendix C, Figure S6**). From our pseudo-time correlation analysis, we detected FOSL1 as a putative transcription factor involved in psoriatic differentiation, a finding which we validated through immunofluorescent staining of healthy and psoriatic skin (**Figure 5G**). Further, previous studies using *in vitro* keratinocyte based systems have suggested that more differentiated keratinocytes were the main responders to IL-17A, given larger effect sizes in differentiated compared to monolayer keratinocyte.[74] Using data generated with Seq-Well S^3 cross-analyzed against an IL-17 response signature in keratinocytes, we show that IL-17 responses are observed in keratinocytes from all layers of the epidermis, but that these responses are stronger in keratinocytes derived from more differentiated layers of the psoriatic epidermis (**Appendix C, Figure S6C**). This observation is corroborated by co-localization of the IL-

17 receptor subunits (IL-17RA/IL-17RC) in the upper layers of psoriatic epidermis (**Figure 5H**).

4.10 Conclusion

In conclusion, we describe a powerful massively-parallel scRNA-Seq protocol that enables improved transcript capture and gene detection from low-input clinical samples. Here, Seq-Well S^3 provides novel insights into putative mechanisms and the cellular localization of previously appreciated and unknown responses to specific inflammatory mediators in immunologic skin conditions in a fashion not previously achievable. Increases in the sensitivity of gene and transcript detection are increasingly important as single-cell atlasing efforts shift from detection of large differences between cell types within normal tissue to identification of subtle differences in cell state across cell types within diseased tissues. The increased sensitivity of gene detection and transcript capture afforded by S^3 enhances the strength of inferences that can be drawn from these types of single-cell data, as evidenced by the range of immune, stromal and parenchymal cell states uncovered in human skin inflammation. The S^3 protocol is easy to integrate into current bead-based RNA-Seq platforms, such as Drop-Seq, making it broadly useful for the single-cell community, particularly in the setting of human disease. Importantly, S^3's increases in library complexity and sequencing efficiency reduce costs relative to plate-based protocols and providing researchers with a powerful and cost-effective alternative to commercial solutions in a format that can be deployed almost anywhere.

METHOD DETAILS

4.11 Single-Cell Processing Pipeline

We utilized Seq-Well, a massively-parallel, low-input scRNA-Seq platform for clinical samples, to capture the transcriptome of single cells. A complete, updated protocol for Seq-Well S^3 is included as a Supplementary Protocol and is hosted on the Shalek Lab website (www.shaleklab.com). Briefly, 10-15,000 cells were loaded onto a functionalized-polydimethylsiloxane (PDMS) array preloaded with uniquely-barcoded mRNA capture beads (Chemgenes; MACOSKO-2011-10). After cells had settled into wells, the array was then sealed with a hydroxylated polycarbonate membrane with pore sizes of 10 nm, facilitating buffer exchange while confining biological molecules within each well. Following membrane-sealing, subsequent buffer exchange permits cell lysis, mRNA transcript hybridization to beads, and bead removal before proceeding with reverse transcription. The obtained bead-bound cDNA product then underwent Exonuclease I treatment (New England Biolabs; M0293M) to remove excess primer before proceeding with second strand synthesis.

4.12 Templated Second Strand Synthesis

Following Exonuclease I treatment, beads were washed once with 500uL of a TE-SDS (0.5% SDS) solution, and twice in 500uL of a TE-Tween (0.01% Tween) solution. After the second TE-TW wash the beads were solvated with 500uL of 0.1M NaOH and mixed for 5 minutes at room temperature using an end-over-end rotator with intermittent agitation to denature the mRNA-cDNA hybrid product on the bead. Following denaturing, the NaOH was removed and beads were washed once with 1M TE, and then combined with a mastermix consisting of 40uL 5x maxima RT buffer, 80uL 30% PEG8000 solution, 20uL 10mM dNTPs, 2uL 1mM dn-SMART oligo, 5uL Klenow Exo-, and 53ul of DI ultrapure water. Second strand synthesis was carried out by incubating the beads for 1 hour at 37°C with end-over-end rotation and intermittent agitation. Following incubation, beads were sequentially washed twice with 0.5 mL of TE buffer with 0.01% Tween 20, and once with 0.5 mL of TE. Immediately prior to PCR amplification, beads were washed once with 0.5 mL of water and resuspended in 0.5 mL of water.

4.13 PCR Amplification

After second strand synthesis, PCR amplification was performed using KAPA HiFi PCR Mix (Kapa Biosystems KK2602). Specifically, a 40uL PCR Mastermix consisting of 25 uL of KAPA 5X Mastermix, 0.4 uL of 100 uM ISPCR oligo, and 14.6 uL of nuclease-free water was combined with 2,000 beads per reaction. For each sample, the total number of PCR reactions performed varied based on the number of beads recovered following second strand synthesis. PCR amplification was performed using the following cycling conditions: an initial denaturation at 95°C for 3 minutes, then 4 cycles of 98°C for 20 seconds, 65°C for 45 seconds, and 72°C for 3 minutes, followed by 9-12 cycles of 98°C or 20 seconds, 67°C or 20 second, and 72°C for 3 minutes, and then a final extension of 72°C for 5 minutes. Following PCR amplification, WTA products were isolated through two rounds of SPRI purification using Ampure Spri beads (Beckman Coulter, Inc.) at both 0.6X and 0.8x volumetric ratio and quantified using a Qubit.

4.14 Optimization of Second Strand Synthesis

We performed a series of experiments to validate the performance of the second-strand synthesis protocol relative other techniques. For the comparison of the Seq-Well protocol with and without second-strand synthesis, we performed species-mixing experiments and PBMC comparisons. For species-mixing experiments, we applied a mixture of 5,000 HEK293 and 5,000 NIH-3T3 cells to a loaded Seq-Well device, while for PBMC comparisons, we loaded a total of 10,000 PBMCs. In optimization experiments, PBMCs were thawed and immediately loaded directly onto Seq-Well devices without stimulation. Following bead removal, beads were split into separate reverse transcription reactions with and without the template-switching oligo. After reverse transcription and ExoI treatment, beads for each comparison were processed separately with and without the second-strand synthesis protocol.

Specifically, we performed a series of optimization experiments to validate the effectiveness of Seq-Well S^3. Specifically, we performed a series of control experiments using beads from a single Seq-Well array loaded with 10,000 PBMCs. For each array we split the beads into six equal fractions and performed the following controls: (1) we performed PCR amplification without the use of second-strand synthesis. (2) we performed random second-strand synthesis followed by PCR amplification. (3) we omitted the template switching oligo without the use of second-strand synthesis. (4) we omitted the template switching oligo but subsequently performed random second-strand synthesis. (5) we examined the effect of heat inactivation of the reverse transcription reagent without the use of second strand synthesis. (6) we examined the effect of heat inactivation of the reverse transcription reagent followed by random second strand synthesis (**Appendix C, Figure S1B-C**). Following PCR amplification, products were obtained from all conditions with the exception of Condition 3 (Seq-Well V1/ No TSO), which did not yield appreciable WTA product.

4.15 Comparison of 10X Genomics (V2 Chemistry) and Seq-Well S^3

Human PBMC were thawed and rested overnight. Cells were stimulated for 18 hours by adding aCD3 (UCHT1) and aCD28 (CD28.2) antibodies were added to the bulk PBMC culture at a concentration of 1mg/mL and 5 mg/mL, respectively, and CD4$^+$ T cells were enriched following stimulation using magnetic negative selection (Stemcell Technologies). Following isolation, T cells were stained with calcein violet live stain (Thermo), Sytox dead stain (Thermo), and aCD45-AF647 (HI30) antibody at 4C for 30 minutes. After two washes, aliquots of the cells were placed on ice and delivered to facilities for flow sorting directly into RLT buffer for Smart-Seq2 processing and another unstained sample for 10X Chromium analysis. Once the cells were delivered, a third aliquot was loaded onto a Seq-Well array. Single-cell libraries were generated using the Smart-Seq 2, 10X V2, and Seq-Well S^3 protocols.

4.16 Sequencing Library Preparation

A total of 1ng of WTA product at a concentration of 0.2 ng/uL was combined with 10 uL of Buffer TD and 5 ul of Buffer ATM and incubated at 55°C for 5 minutes. Following tagmentation, 5 uL of Buffer NT was added and incubated at room temperature for 5 minutes to neutralize the reaction. A total of 8 uL of nuclease-free water, 15uL of buffer NPM, 1 uL of Custom P5 hybrid Oligo, and 1 uL of N700 Index oligo were combined and PCR amplification was performed using the following cycling conditions: an initial denaturation of 95°C for 30 seconds, then 12 cycles of 95°C for 10 seconds, 55°C for 30 seconds, and 72°C for 30 seconds, followed by a final extension of 72°C for 5 minutes. PCR products were isolated through two rounds of SPRI purification (0.6x and 0.8x volumetric ratios) and quantified using a Qubit. Library size distributions were determined using an Agilent Bioanalyzer D1000 High Sensitivity Screen tape.

DATA ANALYSIS OF COMPARISON EXPERIMENTS

4.17 DNA Sequencing and Alignment of PBMC Optimization samples and Downsampling

PBMC optimization experiments were all sequenced on NextSeq500 75 cycle kits. Sequencing read alignment was performed using version 1 of the Drop-Seq pipeline (Macosko et al., 2015). NextSeq runs were loaded at a final concentration of 2.2pM using NextSeq 550 v2 sequencing kits at the Ragon Institute. Briefly, for each sequencing run, raw sequencing reads were converted from bcl files to FASTQs using bcl2fastq using Nextera N700 indices that corresponded to individual samples. Demultiplexed FASTQs were then aligned using an implementation of DropSeqTools v1.0 maintained by the Broad Institute for data analysis, and aligned to the Hg19 genome using standard parameters. Individual reads were tagged with a 12-bp barcode and 8-bp unique molecular identifier (UMI) contained in Read 1 of each sequencing fragment. Following alignment, aligned read 2 sequences were grouped by the 12-bp cell barcodes and subsequently collapsed by the 8-bp UMI for digital gene expression (DGE) matrix extraction and generation.

4.18 PBMC Comparison Experiments

We generated data matrices for PBMC data from 10x genomics and Seq-Well S^3. Initially, we performed downsampling to an average sequencing depth of 42,000 reads per cell. Specifically, downsampling was performed on Seq-Well S^3 to match the sequencing depth of 10x Genomics v2. For each data set, we performed variable gene identification and selected variable genes for downstream analysis (Seq-Well S^3, 856 variable genes and 10x Genomics v2, 516 genes). We performed principal components analysis and selected the first 20 principal components to perform a t-SNE dimensionality reduction. We then performed cluster identification and discovered clusters representing CD4$^+$ T cells, CD8/NK cells, and B cells for each of the technology platforms (**Appendix C, Figure S2A**). We examined the proportion of cell types recovered between Seq-well S^3 and 10x Genomics v2 and performed a Chi-Square test to examine differences in the proportion of recovered cell types (P = 0.971).

Within each cell type identified between Seq-Well S^3 and 10x Genomics V2, we examined differences in aggregate gene detection and transcript capture (**Appendix C, Figure S2C**). We initially performed a Lilliefors test to assess normality of the distribution of genes and UMIs for each technique. Based on these results, we determined to use a Mann-Whitney U Test to determine difference in aggregate gene and transcript detection between techniques (**Figure 1; Appendix C, S1, and S2**). As a measure of library complexity, we examined the linear relationship between the number of UMIs captured and aligned sequencing reads. Specifically, across cell types for each technique, we plotted the number of UMIs against the number of aligned reads and calculated the slope of the regression line for each condition (**Figure 1B-C**). For comparisons of library complexity, we constructed a multivariable linear regression model in which the number of transcripts per cell was modeled follows: nUMI ~ Intercept + B1*nReads + B2*Technique + B3*nReads*Technique. From these models, we determined statistical significance of the difference in slope (i.e. library complexity) based on p-values for the interaction term B3*nReads*Technique, the magnitude and significance of which correspond to a difference in slope (i.e. library complexity or the number of UMIs per aligned read) (**Figure 1B**; **Appendix C, S1B-C**). For example, in a library of low-complexity application of additional sequencing reads might result in detection of a new transcript in every 20th aligned read (i.e. slope = 0.05). Conversely, a library of high complexity, might result in detection of a new transcript with every 4 aligned reads (i.e. slope = 0.25). Critically, these comparisons should be performed on libraries that have been sequenced or down-sampled to similar depths as over-sequencing can augment the relative perception of differences in library complexity. Specifically, libraries that have been "over-sequenced" will appear to have lower complexity because unique molecular identifiers will eventually accumulate additional reads upon saturation.

4.19 Comparison of Gene Detection Rates

For each cell-type cluster, we calculated the rate of detection for each gene as the proportion of cells with a non-zero expression value. Gene detection rates were separately calculated across CD4$^+$ T cells, B cells, CD8/NK cells, and monocytes for both Seq-Well S^3 and 10x Genomics v2. We further examined differences in gene detection rates among transcription factors, cytokines and surface receptors (**Appendix C, Table S1**). For comparisons of relationship between gene-detection rates and overall expression levels, we calculated the expression level of individual genes as the average normalized expression value within each cell type for all cells identified in both Seq-Well S^3 and 10x v2 data (**Figure 1**; **Appendix C, Figure S2**). To test the statistical significance of differences in gene detection frequencies, we performed a chi-square test using the number of cells in which a given gene had a non-zero expression values for each technique.

PROFILING CELL STATES IN HUMAN SKIN INFLAMMATION

4.20 IRB Statement

Informed written consent was obtained from human subjects under a protocol approved by the institutional review boards of the University of Michigan and University of California Los Angeles (UCLA). This study was conducted according to the Declaration of Helsinki Principles.

4.21 Processing of Human Skin

Skin biopsies were obtained from a total of 9 patients at the University of California, Los Angeles and University of Southern California Hansen's Clinic. 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.

4.22 Sequencing and Alignment of Skin Samples

Sequencing read alignment was performed using version 2 of the Drop-seq pipeline previously described in Macosko et al. Briefly, for each Nova-Seq sequencing run, raw sequencing reads were converted from bcl files to FASTQs using bcl2fastq based on Nextera N700 indices that corresponded to individual samples. Demultiplexed FASTQs were then aligned to the Hg19 genome using STAR and the DropSeq Pipeline on a cloud-computing platform maintained by the Broad Institute. Individual reads were tagged with a 12-bp barcode and 8-bp unique molecular identifier (UMI) contained in Read 1 of each sequencing fragment. Following alignment, reads were grouped by the 12-bp cell barcodes and subsequently collapsed by the 8-bp UMI for digital gene expression (DGE) matrix extraction and generation.

4.23 Tissue Immunofluorescence Staining

Formalin fixed, paraffin-embedded tissue slides obtained from psoriasis patients and normal controls were heated for 30 min at 60°C, rehydrated, and epitope retrieved with Tris-EDTA, pH 6. Slides were blocked, incubated with primary antibody APOBEC3 (LS-C98892-400; Lifespan bioscience), FOSL (A03927; Boster), IL-36G (sc-80056; Santa Cruz Biotechnology), TNFAIP3 (ab74037, Abcam), IL-17RC (LS-C400522, Lifespan bioscience), and IL-17RA (LS-C359381, Lifespan bioscience) overnight at 4 °C. Slides were then washed and incubated with Donkey anti-Rabbit IgG 594, Donkey anti-Mouse IgG 488, or Donkey anti-Rat IgG 594 (all from Invitrogen) for 1 h at room temperature. Slides were washed and prepared in mounting medium with 4',6-diamidino-2-phenylindole (DAPI) (VECTASHIELD Antifade Mounting Medium with DAPI, H-1200, VECTOR). Images were acquired using Zeiss Axioskop 2 microscope and analyzed by SPOT software 5.1. Images presented are representative of at least three experiments using biological replicates.

DATA ANALYSIS OF SKIN SAMPLES

4.24 Cell Quality Filtering

Cells were initially filtered on the basis of gene detection (> 500 genes per cell) and transcript detection (> 700 umis per cell) for inclusion in downstream analysis. Further, cells with fractional representation of mitochondrial genes greater than 40% were excluded. To account for potential transcript spreading, we removed any duplicated or hamming=1 barcodes among samples sequenced on the same Nova-Seq runs. For each sample, we performed variable gene identification and calculated 30 principal components. Within each sample, we performed jackstraw simulations to identify significant principal components that were then used to perform t-SNE dimensionality reduction and clustering for each sample using only significant principal components. Within each sample, clusters defined exclusively by mitochondrial gene expression, indicative of low-quality cells, were removed from downstream analysis.

4.25 Removal of Ambient RNA Contamination

Within each sample, we removed ambient RNA contamination using SoupX (Young and Behjati, 2018). Initially, we determined appropriate UMI thresholds to estimate background contamination using EmptyDrops (Lun et al., 2019). Specifically, we examined the distribution of P-values UMI thresholds between 30 and 100 and selected the UMI threshold in 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 S^3 array.

4.26 Doublet Removal

We performed doublet removal for each sample individually using DoubletFinder (McGinnis et al., 2018). For each sample, we calculated the expected doublet rate based on the cell loading density. For each sample, a total of 20,000 cells were loaded to a loaded Seq-Well device containing 85,000 wells (lambda = 20,000). For each array, we calculated an expected doublet rate of 2.37%. For each array, we generated pseudo-doublets using the following parameter values in DoubletFinder: proportion.artificial = 0.25 and proportion.NN = 0.01. Cells were identified as doublets based on their rank order in the distribution of the proportion of artificial nearest neighbors (pANN). Specifically, we identified the pANN value for the cell at the expected doublet percentile and used the corresponding pANN value as a threshold to remove additional cells with pANN greater than or equal to this value.

4.27 Analysis of Combined Skin Dataset

After background and doublet correction, we performed integrated analysis on a combined dataset of 25,468 cells. We performed variable gene identification and dimensionality reduction to identify 34 cell type clusters using Louvain clustering (Resolution = 2.0). We identified genes enriched across clusters to identify generic cell types. We performed an initial round of dimensionality reduction and cluster identification among cell types used in subsequent analysis (i.e. T cells, myeloid cells, B and plasma cells, endothelial cells, fibroblasts, and keratinocytes). Based on the initial sub-clustering results for each cell type, we removed sub-clusters defined by residual contamination not corrected for by SoupX background correction and doublet filtering. In total, we filtered 5,160 cells from sub-clusters defined by residual contamination: 968 from the T cell sub-analysis, 473 from the myeloid sub-analysis, 787 from B and Plasma Cells, 1,049 from the endothelial sub-analysis, 1,051 from the fibroblast sub-analysis, and 832 from the keratinocyte sub-analysis.

After this stringent quality control filtering step, a total of 20,308 cells were included in downstream analysis of the atlas of skin inflammation. We first performed variable gene identification and identified 8,927 genes as variably expressed. We performed UMAP dimensionality reduction to generate a 2-dimensional representation of gene expression data, and we identified a total of 33 cell type clusters using Louvain clustering (Resolution = 2.0) in Scanpy (Wolf et al., 2018). To understand similarity of identified clusters, we performed hierarchical clustering of identified cell type clusters (**Appendix C, Figure S2C**). Specifically, across clusters we generated a list composed of the top 25 cluster-defining genes from each cluster. Average gene expression values within each across the 511 unique cluster-defining genes was used to perform hierarchical clustering. A dendrogram was generated to display the similarity of clusters, and the observed relationships were used to inform rational combination of related cell type clusters for combined analysis (**Appendix C, Figure S2C**). Cell type assignments were assigned through a combination of literature-based assessment of expression signatures and manual curation. In total, we identified clusters representing 16 major cell types (**Appendix C, Table S3**) including arterioles, B cells, dendritic cells, fibroblasts, hair follicle, keratinocytes, Langerhans cells, lymphatics, mast cells, macrophages, muscle, neurons, plasma cells, sweat gland, T cells, and venules.

4.28 Identification of T cell Sub-Clusters

We performed sub-analysis for numerous cell-types to examine additional variation within major cell types. Among the 2,908 T cells identified in the total dataset, we identified 5,574 variable genes that were used to construct a force-directed graph. We further used this set of variable genes to perform Louvain clustering (Resolution = 0.8) and identified a total of nine T cell sub-clusters. Cell-type identities were established by examining the expression of known marker genes corresponding to $CD4^+$ T helper and CD8 T cell subsets. We performed comparison of identified T cell signatures to previously identified signatures in Savant (Lopez et al., 2017) (**Figure S4A**). We identified genes enriched by phenotypic condition across the set of 2,908 by performing a bimodal expression test in Seurat (**Figure 3F, Table S4**). To further define variation among T cell sub-clusters 0 and 8, we performed additional sub-grouping analyses. In both cases, we performed sub-analyses in which t-SNE was performed using a total of 5 principal components calculated across variable genes using Seurat. For T sub-cluster 0 (CD8 T cells), we identified sub-groupings in Seurat using a resolution of 0.5, while a resolution of 1.10 was used for sub-grouping analysis for T cell cub-cluster 8 (Cytotoxic cells) (**Figure 3F**; **Appendix C, S4B**).

4.29 T Cell Receptor Detection

Initially, we examined the detection rates for TCR alpha and beta (Constant, V and J genes) among $CD4^+$ T cells from experiments performed on PBMCs using the Seq-Well

V1, Seq-Well S^3 protocol and 10x v2 (**Appendix C, Figure S4C**). Specifically, detection of constant genes was determined by non-zero values for either *TRAC* or *TRBC2* genes for alpha and beta constant genes, respectively. Similarly, detection of TRAV and TRBV sequences was determined on the basis of a non-zero value for any TRAV or TRBV gene, We further examined the rate of TCR detection across 2,908 T cells obtained from human skin biopsies. Specifically, we examine detection rates across multiple sequencing depths: <5,000, 5,000-25,000, 25-000-100,000, and > 100,000 aligned reads per cell. (**Figure 3G**).

4.30 Identification of Myeloid Heterogeneity

We performed sub-analysis of myeloid populations observed in skin, which include Dendritic cells, Macrophages, Mast cells, and Langerhans cells identified in global analysis of 20,308 total cells. Using a combined dataset of 2,371 myeloid cells, we performed variable gene identification and dimensionality reduction in Scanpy. Specifically, we constructed a force-directed graph across 6,599 variable genes and performed Louvain clustering (resolution = 1.0), and we obtained 9 sub-clusters of myeloid cells (**Appendix C, Figure S4D**).

To understand differences in Langerhans cells in normal skin, we performed differential expression analysis within each cluster of Langerhans cells (**Appendix C, Figure S4D**). Results presented in **Figure 4E** represent differential expression results between Langerhans cells from Cluster 1 between normal and leprosy skin biopsies. To identify mast cell genes associated with inflammatory skin conditions, we performed enrichment analysis using a bimodal expression test in Seurat (**Figure 4F**).

We performed additional sub-grouping among 502 dendritic cells and performed UMAP dimensionality reduction across 4,333 variable genes. We once again used Louvain clustering and identified 5 sub-groupings of dendritic cells, and identified genes enriched within each cluster by performing a bimodal expression test in Seurat. To understand how dendritic cells related to previous findings, we performed comparisons to published signatures of dendritic cell phenotypes (Villani et al., 2017). Specifically, we generated expression scores using the top 10 genes using the AddModuleScore function in Seurat and examined the distribution of signature scores (**Appendix C, Figure S4H**). We determined significance of cluster enrichment by performing 1,000 permutations in which cell and signature score identifiers were randomly re-assigned. Upon permutation testing, we observed significant enrichment among DC sub-group 4 (cDC1) for a signature corresponding to CLEC9A+ cDC1 cells (P<0.05, Permutation Test; **Appendix C, Figure S4H**).

4.31 Identification of Endothelial Heterogeneity

We performed variable gene identification across 4,996 endothelial cells identified in the global analysis across 9 skin biopsies. We generated a force-directed graph and performed Louvain clustering among 6957 variable genes, which revealed 7 sub-clusters of endothelial cells used in downstream analysis. To identify genes enriched in each endothelial sub-cluster, we performed a bimodal expression test in Seurat (**Table S10**). We specifically examined the distribution of addressin expression across endothelial subsets identified using Louvain clustering as well as the distribution across the spectrum of skin inflammation. Directed analysis of addressins was performed using a curated list of addressins using only the set of genes identified in endothelial cells (**Figure S5E**).

5.32 Identification of Fibroblast Heterogeneity

We performed variable gene identification across 4,189 fibroblasts identified in global analysis. We performed dimensionality reduction across 6,931 variable genes and performed Louvain clustering (Resolution = 0.8), revealing 9 sub-clusters of fibroblasts. For each fibroblast sub-cluster, we examined the relative contribution of cells from each sample and condition. Further, we examined the distribution of fibroblast sub-clusters within each sample. For each sub-cluster, we performed enrichment analysis to identify cluster-defining genes (**Appendix C, Table S11**).

4.33 Pseudo-temporal Reconstruction of Epidermal Keratinocytes

Initially, we performed sub-clustering analysis on a combined dataset of 5,141 keratinocytes observed across normal skin and pathologic conditions to assess heterogeneity. Diffusion analysis across all keratinocytes was performed using the Diffmap function in Scanpy, which implements a method for diffusion (Haghverdi et al., 2015). Within normal and psoriatic kerainocytes, a population of basal keratinocytes was identified on the basis of expression of TP63 and KRT14 (**Figure 5**). Initially, we performed pseudo-temporal analysis within normal keratinocyte separately, using the basal keratinocyte population as the origin of the pseudotemporal ordering. After observing distinct developmental trajectories among psoriatic keratinocytes (**Figure 5C**), we performed differential expression analysis between normal and psoriatic keratinocytes in Seurat using a bimodal expression test (**Appendix C, Table S12**)

Differentiation trajectories for individual samples were constructed using dyno (https://rdrr.io/github/dynverse/dyno/man/dyno.html). Specifically, we used SlingShot to generate diffusion pseudotime reconstructions of normal and psoriatic keratinocytes. We examined gene expression patterns correlated with pseudo-temporal order across normal keratinocyte populations. For normal and psoriatic keratinocytes from psoriasis patient 1, we performed differential pseudo-time correlation analysis. Here, for each pseudo-temporal reconstruction, we performed linear regression between pseudotime values and gene expression values for either normal or psoriatic keratinocytes (**Appendix C, Table S13**). We then calculated the difference in pseudotime correlation between psoriatic and normal keratinocytes to identify genes that are uniquely involved in the development of psoriatic keratinocytes (**Figure 5F**).

4.34 Keratinocyte Cytokine-Response Profiles

Among both normal and psoriatic keratinocytes, we generated cytokine response scores using a series of reference datasets. Specifically, bulk RNA-sequencing references were previously generated from in vitro experiments in which cultured keratinocytes were stimulated with cytokines, individually or in combination. We used expression signatures generated by exposing keratinocytes to IL-17A, IL17-A + TNF-alpha, TNF-alpha, IFN-alpha, IL-4, IL-13, and IFN-gamma. Expression signature were generated relative an unstimulated control population of keratinocytes. For each cytokine condition, we used the top 100 differentially expressed genes to generate a cytokine response score across both psoriatic and normal keratinocytes (**Appendix C, Table S14**). We then examined the extent of cytokine response across basal, differentiating and terminal keratinocytes between normal and psoriatic keratinocytes.

4.10 References

1. Klein, A.M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D.A., and Kirschner, M.W. (2015). Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells. Cell 161, 1187–1201.
2. Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A.R., Kamitaki, N., Martersteck, E.M., et al. (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell 161, 1202–1214.
3. Montoro, D.T., Haber, A.L., Biton, M., Vinarsky, V., Lin, B., Birket, S.E., Yuan, F., Chen, S., Leung, H.M., Villoria, J., et al. (2018). A revised airway epithelial hierarchy includes CFTR-expressing ionocytes. Nature 560, 319.
4. Ordovas-Montanes, J., Dwyer, D.F., Nyquist, S.K., Buchheit, K.M., Vukovic, M., Deb, C., Wadsworth, M.H., Hughes, T.K., Kazer, S.W., Yoshimoto, E., et al. (2018). Allergic inflammatory memory in human respiratory epithelial progenitor cells. Nature 560, 649.
5. Vento-Tormo, R., Efremova, M., Botting, R.A., Turco, M.Y., Vento-Tormo, M., Meyer, K.B., Park, J.-E., Stephenson, E., Polański, K., Goncalves, A., et al. (2018). Single-cell reconstruction of the early maternal–fetal interface in humans. Nature 563, 347.
6. Tabula Muris Consortium, Overall coordination, Logistical coordination, Organ collection and processing, Library preparation and sequencing, Computational data analysis, Cell type annotation, Writing group, Supplemental text writing group, and Principal investigators (2018). Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. Nature 562, 367–372.
7. Braga, F.A.V., Kar, G., Berg, M., Carpaij, O.A., Polanski, K., Simon, L.M., Brouwer, S., Gomes, T., Hesse, L., Jiang, J., et al. (2019). A cellular census of human lungs identifies novel cell states in health and in asthma. Nat. Med. 1.
8. Villani, A.-C., Satija, R., Reynolds, G., Sarkizova, S., Shekhar, K., Fletcher, J., Griesbeck, M., Butler, A., Zheng, S., Lazo, S., et al. (2017). Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. Science 356, eaah4573.
9. Puel, A., Ziegler, S.F., Buckley, R.H., and Leonard, W.J. (1998). Defective IL7R expression in T - B + NK + severe combined immunodeficiency. Nat. Genet. 20, 394.
10. Prakadan, S.M., Shalek, A.K., and Weitz, D.A. (2017). Scaling by shrinking: empowering single-cell “omics” with microfluidic devices. Nat. Rev. Genet. 18, 345–361.
11. Sasagawa, Y., Nikaido, I., Hayashi, T., Danno, H., Uno, K.D., Imai, T., and Ueda, H.R. (2013). Quartz-Seq: a highly reproducible and sensitive single-cell RNA

sequencing method, reveals non-genetic gene-expression heterogeneity. Genome Biol. 14, 3097.

12. Shishkin, A.A., Giannoukos, G., Kucukural, A., Ciulla, D., Busby, M., Surka, C., Chen, J., Bhattacharyya, R.P., Rudy, R.F., Patel, M.M., et al. (2015). Simultaneous generation of many RNA-seq libraries in a single reaction. Nat. Methods 12, 323–325.
13. Picelli, S., Björklund, Å.K., Faridani, O.R., Sagasser, S., Winberg, G., and Sandberg, R. (2013). Smart-seq2 for sensitive full-length transcriptome profiling in single cells. Nat. Methods 10, 1096–1098.
14. Kabashima, K., Honda, T., Ginhoux, F., and Egawa, G. (2019). The immunological anatomy of the skin. Nat. Rev. Immunol. 19, 19–30.
15. Gierahn, T.M., Wadsworth Ii, M.H., Hughes, T.K., Bryson, B.D., Butler, A., Satija, R., Fortune, S., Love, J.C., and Shalek, A.K. (2017). Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput. Nat. Methods 14, 395–398.
16. Picelli, S., Faridani, O.R., Björklund, Å.K., Winberg, G., Sagasser, S., and Sandberg, R. (2014). Full-length RNA-seq from single cells using Smart-seq2. Nat. Protoc. 9, 171–181.
17. Wolf, F., Angerer, P. & Theis, F. (2018). SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol* 19**,** 15.
18. Diani, M., Altomare, G., and Reali, E. (2015). T cell responses in psoriasis and psoriatic arthritis. Autoimmun. Rev. 14, 286–292.
19. Lowes, M.A., Suárez-Fariñas, M., and Krueger, J.G. (2014). Immunology of psoriasis. Annu. Rev. Immunol. 32, 227–255.
20. Lopez, D., Montoya, D., Ambrose, M., Lam, L., Briscoe, L., Adams, C., Modlin, R.L., and Pellegrini, M. (2017). SaVanT: a web-based tool for the sample-level visualization of molecular signatures in gene expression profiles. BMC Genomics 18.
21. Ivanov, I.I., McKenzie, B.S., Zhou, L., Tadokoro, C.E., Lepelley, A., Lafaille, J.J., Cua, D.J., and Littman, D.R. (2006). The Orphan Nuclear Receptor RORγt Directs the Differentiation Program of Proinflammatory IL-17+ T Helper Cells. Cell 126, 1121–1133.
22. Saini N, Roberts SA, Klimczak LJ, Chan K, Grimm SA, Dai S, et al. (2016) The Impact of Environmental and Endogenous Damage on Somatic Mutation Load in Human Skin Fibroblasts. PLoS Genet 12 (10).
23. Liu, X., Wang, Y., Lu, H., Li, J., Yan, X., Xiao, M., Hao, J., Alekseev, A., Khong, H., Chen, T., et al. (2019). Genome-wide analysis identifies NR4A1 as a key mediator of T cell dysfunction. Nature 567, 525.
24. Peter, D., Catherine Jin, S.L., Conti, M., Hatzelmann, A., Zitt, C., (2007) Differential Expression and Function of Phosphodiesterase 4 (PDE4) subtypes in Human Primary CD4+ T Cells: Predominant Role of PDE4D. J Immunol 178 (8), 4820-4831.
25. Söderström, K., Stein, E., Colmenero, P., Purath, U., Müller-Ladner, U., Matos, C.T. de, Tarner, I.H., Robinson, W.H., and Engleman, E.G. (2010). Natural killer cells trigger osteoclastogenesis and bone destruction in arthritis. Proc. Natl. Acad. Sci. 107, 13028–13033.

26. Zhang, L., Yu, X., Zheng, L., Zhang, Y., Li, Y., Fang, Q., Gao, R., Kang, B., Zhang, Q., Huang, J.Y., et al. (2018). Lineage tracking reveals dynamic relationships of T cells in colorectal cancer. Nature 564, 268.
27. Malissen, B., Tamoutounour, S., and Henri, S. (2014). The origins and functions of dendritic cells and macrophages in the skin. Nat. Rev. Immunol. 14, 417–428.
28. Fuentes-Duculan, J., Suárez-Fariñas, M., Zaba, L.C., Nograles, K.E., Pierson, K.C., Mitsui, H., Pensabene, C.A., Kzhyshkowska, J., Krueger, J.G., and Lowes, M.A. (2010). A subpopulation of CD163-positive macrophages is classically activated in psoriasis. J. Invest. Dermatol. 130, 2412–2422.
29. Di Rosa, M., Malaguarnera, G., De Gregorio, C., Drago, F., and Malaguarnera, L. (2013). Evaluation of CHI3L-1 and CHIT-1 Expression in Differentiated and Polarized Macrophages. Inflammation 36, 482–492.
30. Romani, N., Holzmann, S., Tripp, C.H., Koch, F., and Stoitzner, P. (2003). Langerhans cells – dendritic cells of the epidermis. APMIS 111, 725–740.
31. Hunger, R.E., Sieling, P.A., Ochoa, M.T., Sugaya, M., Burdick, A.E., Rea, T.H., Brennan, P.J., Belisle, J.T., Blauvelt, A., Porcelli, S.A., et al. (2004). Langerhans cells utilize CD1a and langerin to efficiently present nonpeptide antigens to T cells. J. Clin. Invest. 113, 701–708.
32. Pinheiro, R.O., Schmitz, V., Silva, B.J. de A., Dias, A.A., de Souza, B.J., de Mattos Barbosa, M.G., de Almeida Esquenazi, D., Pessolani, M.C.V., and Sarno, E.N. (2018). Innate Immune Responses in Leprosy. Front. Immunol. 9.
33. Guilliams, M., Dutertre, C.-A., Scott, C.L., McGovern, N., Sichien, D., Chakarov, S., Van Gassen, S., Chen, J., Poidinger, M., De Prijck, S., et al. (2016). Unsupervised High-Dimensional Analysis Aligns Dendritic Cells across Tissues and Species. Immunity 45, 669–684.
34. Schlitzer A, McGovern N, Teo P, et al. IRF4 transcription factor-dependent CD11b+ dendritic cells in human and mouse control mucosal IL-17 cytokine responses. *Immunity*. 2013;38(5):970–983.
35. Stutte S, Quast T, Gerbitzki N, et al. Requirement of CCL17 for CCR7- and CXCR4-dependent migration of cutaneous dendritic cells. *Proc Natl Acad Sci U S A*. 2010;107(19):8736–8741. doi:10.1073/pnas.0906126107
36. Kashem, S.W., Riedl, M.S., Yao, C., Honda, C.N., Vulchanova, L., and Kaplan, D.H. (2015). Nociceptive Sensory Fibers Drive Interleukin-23 Production from CD301b+ Dermal Dendritic Cells and Drive Protective Cutaneous Immunity. Immunity 43, 515–526.
37. Kumamoto, Y., Linehan, M., Weinstein, J.S., Laidlaw, B.J., Craft, J.E., and Iwasaki, A. (2013). CD301b+ Dermal Dendritic Cells Drive T Helper 2 Cell-Mediated Immunity. Immunity 39, 733–743.
38. Jin P, Han TH, Ren J, et al. (2010). Molecular signatures of maturing dendritic cells: implications for testing the quality of dendritic cell therapies. *J Transl Med*. 2010;8:4
39. Guilliams, M., Bruhns, P., Saeys, Y., Hammad, H., and Lambrecht, B.N. (2014). The function of Fcγ receptors in dendritic cells and macrophages. Nat. Rev. Immunol. 14, 94–108.

40. Pejler, G., Rönnberg, E., Waern, I., and Wernersson, S. (2010). Mast cell proteases: multifaceted regulators of inflammatory disease. Blood 115, 4981–4990.
41. Dwyer, D.F., Barrett, N.A., Austen, K.F., The Immunological Genome Project Consortium, Dwyer, D.F., Barrett, N.A., Austen, K.F., Kim, E.Y., Brenner, M.B., Shaw, L., et al. (2016). Expression profiling of constitutive mast cells reveals a unique identity within the immune system. Nat. Immunol. 17, 878–887.
42. Schön, M.P., Zollner, T.M., and Boehncke, W-H. (2003) The Molecular Basis of Lymphocyte Recruitment to the Skin: Clues for Pathogeneis and Selective Therapies of Inflammatory Disorders, J Invest. Dermatol. 121 (5), 951-962.
43. Andrian, U.H. von, and Mempel, T.R. (2003). Homing and cellular traffic in lymph nodes. Nat. Rev. Immunol. 3, 867.
44. Thiriot, A., Perdomo, C., Cheng, G., Novitzky-Basso, I., McArdle, S., Kishimoto, J.K., Barreiro, O., Mazo, I., Triboulet, R., Ley, K., et al. (2017). Differential DARC/ACKR1 expression distinguishes venular from non-venular endothelial cells in murine tissues. BMC Biol. 15, 45.
45. Driskell, R.R., and Watt, F.M. (2015). Understanding fibroblast heterogeneity in the skin. Trends Cell Biol. 25, 92–99.
46. Driskell, R.R., Lichtenberger, B.M., Hoste, E., Kretzschmar, K., Simons, B.D., Charalambous, M., Ferron, S.R., Herault, Y., Pavlovic, G., Ferguson-Smith, A.C., et al. (2013). Distinct fibroblast lineages determine dermal architecture in skin development and repair. Nature 504, 277–281.
47. Tabib, T., Morse, C., Wang, T., Chen, W., and Lafyatis, R. (2018). SFRP2/DPP4 and FMO1/LSP1 Define Major Fibroblast Populations in Human Skin. J. Invest. Dermatol. 138, 802–810.
48. Fu X, Khalil H, Kanisicak O, et al. Specialized fibroblast differentiated states underlie scar formation in the infarcted mouse heart. *J Clin Invest*. 2018;128(5):2127–2143.
49. Hazell, G.G.J., Peachey, A.M.G., Teasdale, J.E., Sala-Newby, G.B., Angelini, G.D., Newby, A.C., and White, S.J. (2016). PI16 is a shear stress and inflammation-regulated inhibitor of MMP2. Sci. Rep. 6, 39553.
50. Huth, S., Heise, R., Vetter-Kauczok, C.S., Skazik, C., Marquardt, Y., Czaja, K., Knüchel, R., Merk, H.F., Dahl, E., and Baron, J.M. (2015). Inter-α-trypsin inhibitor heavy chain 5 (ITIH5) is overexpressed in inflammatory skin diseases and affects epidermal morphology in constitutive knockout mice and murine 3D skin models. Exp. Dermatol. 24, 663–668.
51. Iwata N, Inazu N, and Satoh T (1990) The purification and properties of aldosereductase from rat ovary. Arch Biochem Biophys, 282:70 –77.
52. Fuchs, E. (1990). Epidermal differentiation: the bare essentials. J. Cell Biol. 111, 2807–2814.
53. Fuchs, E., and Raghavan, S. (2002). Getting under the skin of epidermal morphogenesis. Nat. Rev. Genet. 3, 199.
54. Pellegrini, G., Dellambra, E., Golisano, O., Martinelli, E., Fantozzi, I., Bondanza, S., Ponzin, D., McKeon, F., and Luca, M.D. (2001). p63 identifies keratinocyte stem cells. Proc. Natl. Acad. Sci. 98, 3156–3161.

55. Cheng, J.B., Sedgewick, A.J., Finnegan, A.I., Harirchian, P., Lee, J., Kwon, S., Fassett, M.S., Golovato, J., Gray, M., Ghadially, R., et al. (2018). Transcriptional Programming of Normal and Inflamed Human Epidermis at Single-Cell Resolution. Cell Rep. 25, 871–883.

56. Uhlén, M., Fagerberg, L., Hallström, B.M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, Å., Kampf, C., Sjöstedt, E., Asplund, A., et al. (2015). Tissue-based map of the human proteome. Science 347, 1260419.

57. Sandilands, A., Sutherland, C., Irvine, A.D., and McLean, W.H.I. (2009). Filaggrin in the frontline: role in skin barrier function and disease. J. Cell Sci. 122, 1285–1294.

58. Li, B., Tsoi, L.C., Swindell, W.R., Gudjonsson, J.E., Tejasvi, T., Johnston, A., Ding, J., Stuart, P.E., Xing, X., Kochkodan, J.J., et al. (2014). Transcriptome Analysis of Psoriasis in a Large Case–Control Sample: RNA-Seq Provides Insights into Disease Mechanisms. J. Invest. Dermatol. 134, 1828–1838.

59. Zhu, J., Yamane, H., and Paul, W.E. (2010). Differentiation of Effector CD4 T Cell Populations. Annu. Rev. Immunol. 28, 445–489.

60. Regev, A., Teichmann, S., Rozenblatt-Rosen, O., Stubbington, M., Ardlie, K., Amit, I., Arlotta, P., Bader, G., Benoist, C., Biton, M., et al. (2018). The Human Cell Atlas White Paper.

61. Modlin, R.L., Horwitz, D.A., Jordan, R.R., Gebhard, J.F., Taylor, C.R., and Rea, T.H. (1984). Immunopathologic Demonstration of T Lymphocyte Subpopulations and Interleukin 2 in Graneloma Annulare. Pediatr. Dermatol. 2, 26–32.

62. Terziroli Beretta-Piccoli, B., Mainetti, C., Peeters, M.-A., and Laffitte, E. (2018). Cutaneous Granulomatosis: a Comprehensive Review. Clin. Rev. Allergy Immunol. 54, 131–146.

63. Fulco, T. de O., Andrade, P.R., Barbosa, M.G. de M., Pinto, T.G.T., Ferreira, P.F., Ferreira, H., Nery, J.A. da C., Real, S.C., Borges, V.M., Moraes, M.O., et al. (2014). Effect of Apoptotic Cell Recognition on Macrophage Polarization and Mycobacterial Persistence. Infect. Immun. 82, 3968–3978.

64. Verreck, F.A.W., Boer, T. de, Langenberg, D.M.L., Hoeve, M.A., Kramer, M., Vaisberg, E., Kastelein, R., Kolk, A., Waal-Malefyt, R. de, and Ottenhoff, T.H.M. (2004). Human IL-23-producing type 1 macrophages promote but IL-10-producing type 2 macrophages subvert immunity to (myco)bacteria. Proc. Natl. Acad. Sci. 101, 4560–4565.

65. Piette, E.W., and Rosenbach, M. (2016). Granuloma annulare: Clinical and histologic variants, epidemiology, and genetics. J. Am. Acad. Dermatol. 75, 457–465.

66. Yun, J.H., Lee, J.Y., Kim, M.K., Seo, Y.J., Kim, M.H., Cho, K.H., Kim, M.B., Lee, W.S., Lee, K.H., Kim, Y.C., et al. (2009). Clinical and Pathological Features of Generalized Granuloma Annulare with Their Correlation: A Retrospective Multicenter Study in Korea. Ann. Dermatol. 21, 113–119.

67. Beylot, C., Auffret, N., Poli, F., Claudel, J.-P., Leccia, M.-T., Giudice, P.D., and Dreno, B. (2014). Propionibacterium acnes: an update on its role in the pathogenesis of acne. J. Eur. Acad. Dermatol. Venereol. 28, 271–278.

68. Xing, L., Dai, Z., Jabbari, A., Cerise, J.E., Higgins, C.A., Gong, W., de Jong, A., Harel, S., DeStefano, G.M., Rothman, L., et al. (2014). Alopecia areata is driven

by cytotoxic T lymphocytes and is reversed by JAK inhibition. Nat. Med. 20, 1043–1049.

69. Keren, A., Shemer, A., Ullmann, Y., Paus, R., et al. (2015). The PDE4 inhibitor apremilast, suppresses experimentally induced alopecia areata in human skin *in vivo*. J Dermatol Sci.

70. Kim, T.-G., Kim, S.H., Park, J., Choi, W., Sohn, M., Na, H.Y., Lee, M., Lee, J.W., Kim, S.M., Kim, D.-Y., et al. (2018). Skin-Specific CD301b+ Dermal Dendritic Cells Drive IL-17−M ediated Psoriasis-Like Immune Response in Mice. J. Invest. Dermatol. 138, 844–853.

71. Zaba, L.C., Fuentes-Duculan, J., Eungdamrong, N.J., Johnson-Huang, L.M., Nograles, K.E., White, T.R., Pierson, K.C., Lentini, T., Suárez-Fariñas, M., Lowes, M.A., et al. (2010). Identification of TNF-related apoptosis-inducing ligand and other molecules that distinguish inflammatory from resident dendritic cells in patients with psoriasis. J. Allergy Clin. Immunol. 125, 1261-1268.e9.

72. Hawkes, J.E., Yan, B.Y., Chan, T.C., and Krueger, J.G. (2018). Discovery of the IL-23/IL-17 Signaling Pathway and the Treatment of Psoriasis. J. Immunol. 201, 1605–1613.

73. Chiricozzi, A., Nograles, K.E., Johnson-Huang, L.M., Fuentes-Duculan, J., Cardinale, I., Bonifacio, K.M., Gulati, N., Mitsui, H., Guttman-Yassky, E., Suárez-Fariñas, M., et al. (2014). IL-17 Induces an Expanded Range of Downstream Genes in Reconstituted Human Epidermis Model. PLOS ONE 9, e90284.

Chapter 5: Conclusions

As demonstrated in previous sections, scRNA-Seq is a powerful tool that has transformed the way we study health and disease. However, as our capability to ask more complex biological questions continues to grow, it necessitates both the optimization and innovation of single-cell technologies in order to improve our understanding of the biological systems of interest. When I entered the field in 2014, the power of single-cell technologies had been demonstrated; however, it's application in precision medicine pipelines had yet to be realized. As a result, I was motivated to contribute to ongoing collaborations that not only demonstration why single-cell technologies are critical to precision medicine applications (see Chapter 2), but also develop new, high-throughput technologies (See Chapters 3 and 4) that provide viable options for the implementation of scRNA-Seq methods in precision medicine pipelines.

In this final chapter, we explore ongoing work, both technology development, and biological applications, in the field of scRNA-Seq.

5.1 Contributions of this work

In this work, we provide a chronology of scRNA-Seq technology development and how, through both the power and limitations of low-throughput platforms, motivated the development of massively parallel, high-throughput methods. In Chapter 2, we describe the application of a powerful low-throughput, plate-based method[1-3] to construct single-cell profiles of the melanoma tumor microenvironment. Critically, this work demonstrates how the complexity of diseases, like cancer, requires single-cell resolution to construct clinically relevant patient profiles. For example, leveraging scRNA-Seq, this study showed that malignant cells exist on a spectrum of heterogeneity for MITF and AXL programs, which are canonically thought to be separate programs based on bulk RNA-Seq profiles.[4] This phenomenon potentially explains why targeted treatment of either tumor, based on bulk profiles, can lead to an outgrowth of drug-resistant tumor phenotypes

In Chapter 3, we present Seq-Well[5], our solution to the limitations of existing low-[1-3] and high-[6-9] throughput scRNA-Seq assays for clinical implementations. This work demarcates an important time in single-cell technology development because it was a catalyst for democratizing scRNA-Seq. At that time, scRNA-Seq technologies were either prohibitively expensive[1-3] or required an extensive setup[6-8], preventing implementation in low-resource settings. Our technology was, and still is, a viable solution to these limitations because it is relatively inexpensive to run (less than $200/assay) and requires minimal equipment. As a result, we have helped democratize scRNA-Seq, setting up technologies in over 130 laboratories across 6 continents and 20 countries.

Although Seq-Well catalyzed multiple international collaborations, it could not achieve the necessary transcriptome coverage to reliably phenotype unique and rare immune subsets, limiting its application in certain contexts. In chapter 4, to address this, we present Seq-Well S^3 (Second Strand Synthesis), a modified Seq-Well protocol with dramatically improved gene and transcript capture.[10] Using this improved pipeline, we constructed an atlas of skin inflammation, profiling immune, and parenchymal cell subsets. Importantly, with our improved sensitivity, we were able to profile previously unappreciated diversity in adaptive and innate immune subsets, identifying phenotypes unique to the different inflammatory diseases. For example, we identified a population of dysfunctional T cells that were over-represented in patients with psoriasis.[10] With this improved sensitivity, we were also able to propose biomarker targets, both unique and conserved across the inflammatory diseases for therapeutic intervention.

In the subsequent sections, we will present four ongoing projects that are extensions of the presented work.

5.2 Elucidating the Host-Pathogen Interaction of liver-stage *P. vivax* infection

Leveraging the improved sensitivity of Seq-Well S^3, we are studying the host-pathogen interactions of liver-stage *P. vivax* infection in hepatocytes. To date, a major barrier to the eradication of malaria is the unique relapsing nature of *Plasmodium vivax* infection.[11] Using scRNA-Seq, we are profiling primary human hepatocytes, using co-cultures (micropatterned, primary human hepatocyte co-cultures; MPCCs)[11] to elucidate the host-pathogen dynamics over the course of liver-stage infection. Currently, we are working towards identifying how sporozoites, upon infecting a hepatocyte, commit to either maturation (i.e., replication by schizogony to produce merozoites and release into the bloodstream) or become hypnozoites, a dormant form of the parasite that can lead to relapse weeks to years after the initial infection.[11] To date, we have profiled the transcriptome of 72,536 host (hepatocyte) cells and 126 parasites, allowing us to understand how infection alters the host response (**Figure 1A-B**). Preliminary results show that, in a comparison between infected and uninfected hepatocytes, we observe downregulation of interferon gamma (INFγ) programs in the infected cells while upregulation of INFγ in neighboring hepatocytes (**Figure 1C**). Currently, we are validating the potential gene targets, derived from the scRNA-Seq analysis, using a small interfering RNA (siRNA) perturbation assay. This will not only improve our understanding of the hepatocyte response to infection but also validate our hypothesis that reactivation of INFγ programs will help eliminate the parasites.

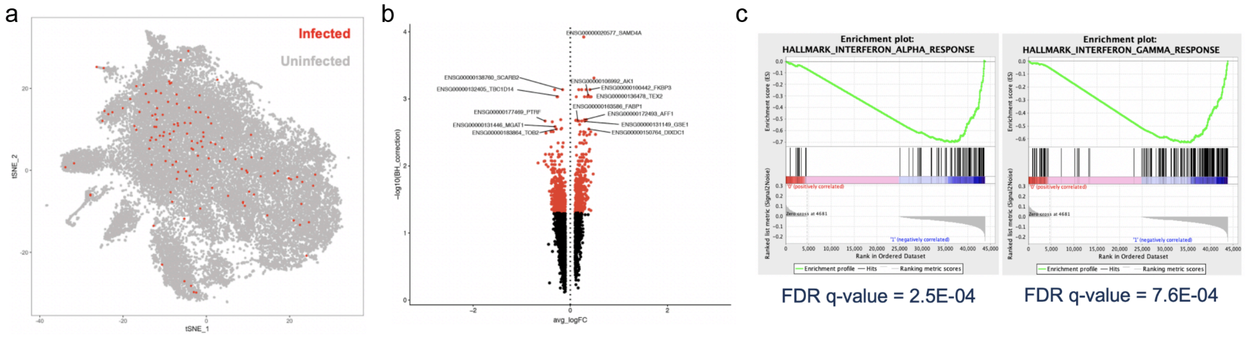


Figure 1 | Seq-Well of 72,536 hepatocytes from host/pathogen study. (**A**) tSNE plot of hepatocyte cells color by infection status. (**B**) Volcano plot of differentially expressed genes between infected (right) and uninfected (left) hepatocytes. (**C**) Gene Set Enrichment Analysis (GSEA) plots of IFNA and IFNG programs (0 = infected, 1 = uninfected).

5.3 Seq-Well V3: Bead-to-Seq

As presented in Chapter 4, the dramatic improvement of transcriptome coverage afforded by Seq-Well S^3 has enabled robust phenotyping of previously elusive immune subsets.[10] However, while Seq-Well S^3 is one of the most affordable scRNA-Seq methods available (around $200/array before sequencing), experimental costs can still accumulate when trying to scale this technology for larger studies (e.g., cellular atlasing). To address this, we are currently developing Seq-Well V4, a modified pipeline of Seq-Well S^3 that leverages second strand synthesis to generate sequencing libraries directly on the bead. In doing so, this decreases reagent costs by ~40% by eliminating the need for Illumina nextera XT kits, making Seq-Well an even more affordable alternative to other scRNA-Seq technologies. Preliminary data is promising; however, the technique still needs to be benchmarked against the standards in the field (e.g., Seq-Well S^3, Drop-Seq, inDrop, 10x Genomics, etc.) to ensure it is competitive and a usable option for researchers.

5.4 Nuc-Seq: Development and Modification of Seq-Well for nuclei

Although Seq-Well V1[5] and S^3[10] platforms have transformed the field and helped democratize single-cell technologies by empowering scientists in low- and high-resource settings, they still have critical limitations that prevent even more widespread use. For instance, the current iteration of Seq-Well is not compatible with nuclei, preventing our ability to profile specific sample types like fixed cells (e.g., formalin fixed paraffin embedded tissues), frozen tissues (e.g., bio-banked archived samples) and tissues with cell-type-specific processing compilations (e.g., neurons). To address this, in collaboration with the Regev Lab (MIT/Broad), we are developing Nuc-Seq, an optimized massively-parallel pipeline for processing single nuclei. To date, we have profiled 3,654 hepatocyte nuclei from mouse livers, recapitulating known biology (**Figure 2A and B**). Preliminary results are promising, showing that we are not only recovering transcripts known to localize to in the nucleus (e.g., Hnf4a) versus cytoplasmic transcripts (e.g., Gata2; **Figure C**) but that, when comparing snRNA-Seq with scRNA-Seq, the methods recover cells with similar transcriptional profiles, although sometimes at varying proportions (**Figure 2D** and **E)**. However, the technique still needs to be benchmarked against the standards in the field (e.g., DroNc-Seq, 10x Genomics, etc.) to ensure it is competitive and a usable option for researchers.

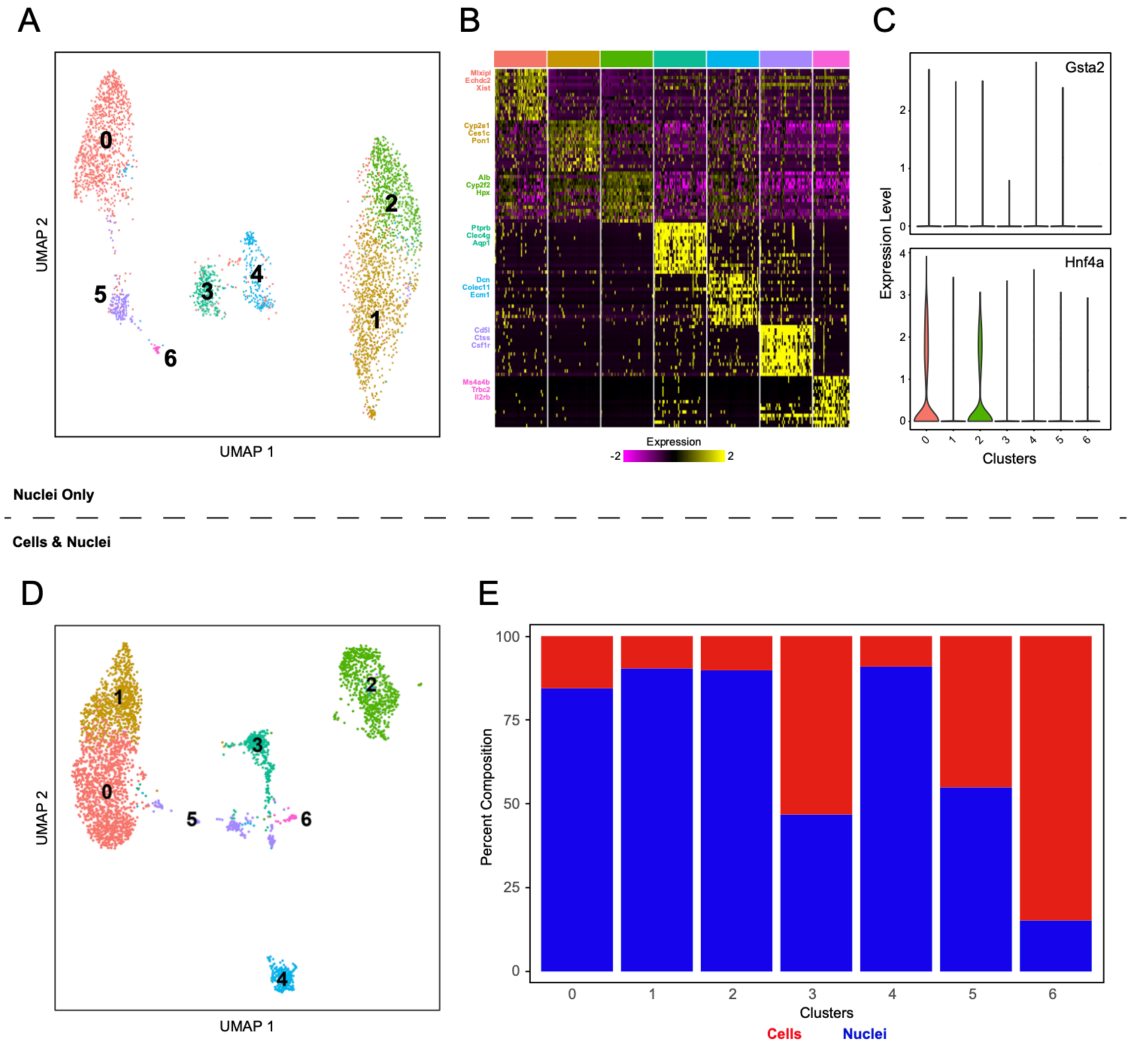


Figure 2 | Nuc-Seq of ~3,654 hepatocyte nuclei from preliminary pilot. (**A**) UMAP of nuclei recovered from mouse liver. (**B**) Heatmap of cluster-defining genes from UMAP in **A.** (**C**) Violin plots of Gata2 and Hnf4a gene expression. (**D**) UMAP of integrated hepatocyte cells and nuclei. (**E**) Proportion of sample types recovered from mouse liver broken down by UMAP clusters in **D**.

5.5 Elucidating the mechanism of BCG vaccination using single-cell RNA-Seq

To date, we have yet to realize a highly protective and durable vaccine against *Mycobacterium tuberculosis* (Mtb) infection and TB disease, which results in 10 million TB cases and 1.6 million deaths each year.[12,13] While the currently licensed vaccine, Bacillus Calmette-Guérin (BCG), confers protection against disseminated disease in infants, clinical trials have shown variable protection (0-70%) against pulmonary tuberculosis (TB) disease.[12] In order to develop a more effective vaccine, we must identify the mechanisms that confer protection and optimize our prophylactic interventions to induce them.

Recent studies have shown that protection against Mtb is, in part, dependent on T cell responses.[14,15] These results suggest that protection against pulmonary TB likely requires a vaccine that can induce high levels of tissue-resident T cells in the lung that can recognize a wide range of antigens from among the 4,000 MtB genes, and hence have the potential to immediately control and eliminate Mtb in the lung post-exposure. However, as lung resident T cells may have limited durability, it could be critical for a vaccine to generate circulating, long-lived TB-specific "central memory" T cells that can function as a reservoir of cells to repopulate the lung for long-term protection.[15] Several studies have demonstrated that the route of vaccine administration has the potential to impact the relative frequency and attributes of these responses.[16-19]

The Seder Lab (NIH/VRC) recently developed a non-human primate (NHP) model to compare immunity and protection following immunization with BCG by intradermal (ID), aerosol (AE), and intravenous (IV) routes.[20] Intriguingly, their results showed that IV administration of BCG prevented infection following Mtb challenge (6 months later) in 80% of NHPs. The other vaccine routes, meanwhile, proved less effective and displayed higher mycobacterial burden post-challenge. Comparing tissue samples from animals across this spectrum of vaccine efficacy enables the identification of cellular phenotypes that may be associated with control. For example, by flow cytometry, the Seder and Roederer Labs (NIH/VRC) identified that IV BCG leads to a significant expansion of T cells in bronchoalveolar lavages (BALs).

To characterize cellular features that might inform differences in protection, we performed high-throughput single-cell RNA-sequencing (scRNA-Seq) with Seq-Well on a non-human primate (NHP) multi-route BCG vaccination study in collaboration with the Seder and Roederer Labs. At two time points (weeks 13 (peak response) and 25 (memory phase)) post-vaccination, we profiled BALs from 15 monkeys representing five routes of BCG administration – an unvaccinated control, low-dose ID, high-dose ID, AE, and IV (3 monkeys/condition). We also performed *ex vivo* overnight PPD stimulation on a fraction of each BAL to polarize its macrophages and T cells and similarly subjected it to scRNA-

Seq. Across 60+ runs (160,000+ single cells), we identified trends in cell-type composition and T cell phenotypes that are associated with the degree and duration of protection against mucosal Mtb challenge (**Figure 3A** and **B**). Importantly, we identify an IV-BCG enriched module of correlated gene expression associated with T cell survival and effector function that is enriched among T cells with a Th1/Th17 phenotype. This is particularly intriguing since a recent study of mucosal BCG vaccination in rhesus macaques showed by flow cytometry that such Th1/Th17 CD4 T cell responses were associated with high-level protection against repeated, low dose challenge with Mtb.[21] More specifically, at 13 weeks post-vaccination, we observed a stimulation-inducible module of gene expression enriched for effector memory T cell functionality that was almost exclusively expressed in the T cells of animals that received IV BCG vaccination (**Figure 4A-D)**. Further, we observed that the degree of stimulation-based induction of this module corresponded to the level of protection conferred by different vaccine routes (**Figure 4E**). Notably, when this stimulation-inducible gene-expression module, which was identified in week 13 T cells, was projected onto week 25 post-vaccination T cells, we observed a similar trend in the distribution of cellular module-scores by vaccine route. Looking directly at the 25 weeks post-vaccination data, we uncovered multiple stimulation-specific T cell gene expression modules. Notably, two of these showed a high

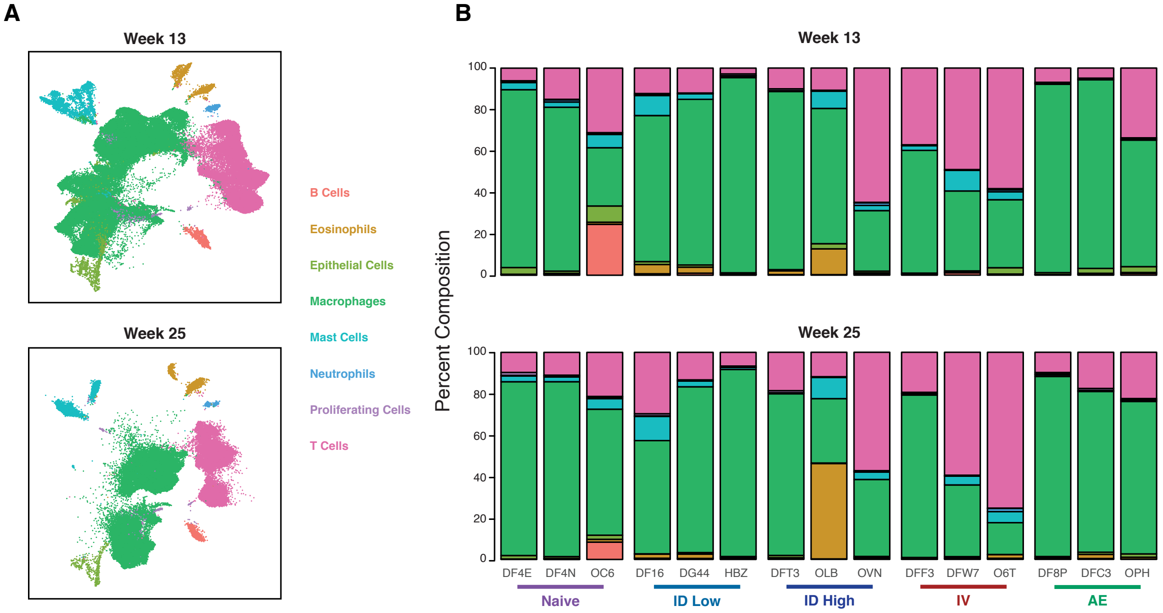


Figure 3 | Seq-Well of ~160,000 cells from a multi-route NHP BCG vaccination study. (**A**) UMAPs of cells recovered from Week 13 (top) and Week 25 (bottom)**.** (**B**) Proportion of cell types (colored as in **A**) recovered from BALs by route and by animal at Week 13 (top) and Week 25 (bottom).

degree of similarity to the module that correlated with vaccine efficacy at week 13. Further, while the IV vaccine response-signature at week 13 was largely driven by a single animal (i.e., a precocious responder; though up in each), we observed that both of these modules (as well as the projected week 13 module) were upregulated across all IV vaccinated animals at week 25. Based on these data, we believe that the onset of some vaccine-associated gene expression signatures may vary across animals, but once induced, may persist – at least over the timescales studied.

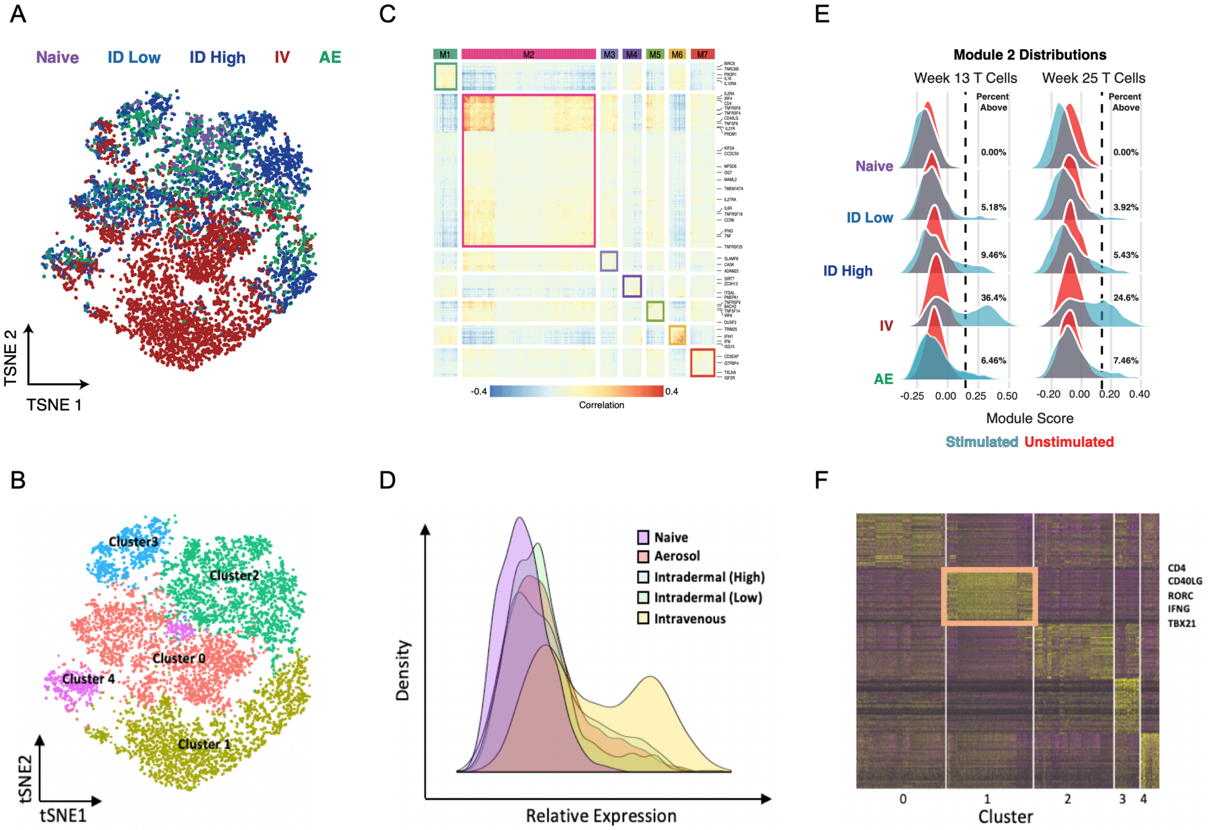


Figure 4 | A non-canonical Th1/Th17 hybrid T cell is enriched in IV-BCG NHPs. (**A**) t-SNE plot of week 13 stimulated T cells colored by vaccine route. (**B**) t-SNE plot of week 13 stimulated T cells colored by phenotypic cluster. (**C**) Gene-gene correlation matrix demonstrating correlated gene modules (effector memory module boxed). (**D**) Distribution of module score by vaccine route. (**E**) Distribution of signature score by vaccine route and timepoint. (**F**) Heatmap showing identifying features of cells expressing gene module associated with IV vaccine response (boxed).

To further examine the molecular features of the cells expressing this effector memory module, we performed dimensionality reduction and clustering across T cells (**Figure 4A and B**). We then examined the distribution of gene expression module scores across identified cell-type clusters at weeks 13 and 25 to uncover those enriched for each module. We observed that the cells that most highly expressed the effector memory module were CD4+ T cells further enriched for expression of TBX21 and RORC (**Figure 4F**). Importantly, these findings are consistent with a recent report that demonstrated that the frequency of poly-functional Th1/Th17 cells is a predictor of mucosal protection following aerosol BCG vaccination in rhesus macaques.[21]

Having demonstrated that IV-BCG immunization in NHPs as a strategy to generate lung-resident T cells, as well as systemic immunity to serve as a T cell reservoir, we are currently investigating how the dosage of IV-BCG alters the previously identified correlates of protection. Again, we are leveraging Seq-Well to transcriptionally profile bronchoalveolar lavages (BALs) performed on immunized NHPs and define gene signatures of a protective immune response to Mtb infection post IV BCG vaccination. In doing so, we hope to understand how the correlates of protection change with BCG dosage and how this impacts protection against pulmonary TB infection.

5.6 Conclusion

The development and application of scRNA-Seq technologies has transformed our understanding of biological systems and how diseases perturb them. As we enter the era of precision medicine, it is imperative that we approach benchtop science with a bedside mindset. In doing so, it will influence how scRNA-Seq technology is developed and implemented to study biological systems, allowing scientists to better contextualize and translate findings for clinical care. If achieved, the field of scRNA-Seq has the potential to transform clinical care, enabling the rapid generation of patient profiles that can be cross-referenced with a comprehensive database, providing customized, tailored medical treatments. However, in order for successful implementation of scRNA-Seq in clinical settings to be achieved, experimental and computational pipelines need to be standardized to ensure robust and reproducible results. It is through efforts like the Human Cell Atlas that single-cell technologies will become critical tools for precision medicine pipelines. Ultimately, it will be the field’s ability to contextualize benchtop findings in bedside applications that will enable it to move beyond its current limitations and transform how we treat disease.

5.7 References

1 Picelli, S. *et al.* Full-length RNA-seq from single cells using Smart-seq2. *Nature Protocols* **9**, 171-181, doi:10.1038/nprot.2014.006 (2014).

2 Picelli, S. *et al.* Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nature Methods* **10**, 1096-1100, doi:10.1038/nmeth.2639 (2013).

3 Trombetta, J. J. *et al.* Preparation of single-cell RNA-Seq libraries for next generation sequencing. *Current Protocols in Molecular Biology* **2014**, 4.22.21-24.22.17, doi:10.1002/0471142727.mb0422s107 (2014).

4 Tirosh, I. *et al.* Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352**, 189-196, doi:10.1126/science.aad0501 (2016).

5 Gierahn, T. M. *et al.* Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput. *Nature Methods* **14**, 395-398, doi:10.1038/nmeth.4179 (2017).

6 Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nature Communications* **8**, doi:10.1038/ncomms14049 (2017).

7 Macosko, E. Z. *et al.* Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202-1214, doi:10.1016/j.cell.2015.05.002 (2015).

8 Klein, A. M. *et al.* Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187-1201, doi:10.1016/j.cell.2015.04.044 (2015).

9 Fan, H. C., Fu, G. K. & Fodor, S. P. A. Combinatorial labeling of single cells for gene expression cytometry. *Science* **347**, doi:10.1126/science.1258367 (2015).

10 Hughes, T. K. *et al.* Highly Efficient, Massively-Parallel Single-Cell RNA-Seq Reveals Cellular States and Molecular Features of Human Skin Pathology. *bioRxiv*, 689273, doi:10.1101/689273 (2019).

11 Gural, N. *et al.* In Vitro Culture, Drug Sensitivity, and Transcriptome of Plasmodium Vivax Hypnozoites. *Cell Host & Microbe* **23**, 395-406.e394, doi:https://doi.org/10.1016/j.chom.2018.01.002 (2018).

12 Mangtani, P. Protection by BCG vaccine against tuberculosis: a systematic review of randomized controlled trials. *Nephrol. Dial. Transplant.* **58**, 470-480 (2014).

13 Rc Harris, T. S. G. M. K. R. G. W. Systematic review of mathematical models exploring the epidemiological impact of future TB vaccines. *Hum. Vaccin. Immunother.* **12**, 2813-2832 (2016).

14 Cooper, A. M. Cell-mediated immune responses in tuberculosis. *Annu. Rev. Immunol.* **27**, 393-422 (2009).

15 Ogongo, P., Porterfield, J. Z. & Leslie, A. Lung Tissue Resident Memory T-Cells in the Immune Response to Mycobacterium tuberculosis. *Front Immunol* **10**, 992-992, doi:10.3389/fimmu.2019.00992 (2019).

16 Barclay, W. R. Protection of monkeys against airborne tuberculosis by aerosol vaccination with bacillus Calmette-Guerin. *Am. Rev. Respir. Dis.* **107**, 351-358 (1973).

17 Wr Barclay, R. L. A. W. B. W. L. E. R. Aerosol-induced tuberculosis in subhuman primates and the course of the disease after intravenous BCG vaccination. *Infect. Immun.* **2**, 574-582 (1970).

18 Ribi, E. Efficacy of mycobacterial cell walls as a vaccine against airborne tuberculosis in the Rheusus monkey. *J. Infect. Dis.* **123**, 527-538 (1971).

19 Anacker, R. L. Superiority of intravenously administered BCG and BCG cell walls in protecting rhesus monkeys (Macaca mulatta) against airborne tuberculosis. *Z. Immunitatsforsch. Exp. Klin. Immunol.* **143**, 363-376 (1972).

20 Darrah, P. A. *et al.* Prevention of tuberculosis in macaques after intravenous BCG immunization. *Nature* **577**, 95-102, doi:10.1038/s41586-019-1817-8 (2020).

21 Dijkman, K. *et al.* Prevention of tuberculosis infection and disease by local BCG in repeatedly exposed rhesus macaques. *Nature Medicine* **25**, 255-262, doi:10.1038/s41591-018-0319-9 (2019).

Appendix A - Supplemental Information for Chapter 2: Dissecting the Multicellular Ecosystem of Metastatic Melanoma by Single-Cell RNA-Seq

This chapter is adapted in accordance with AAAS's open access policy from the following article published in Science:

Itay Tirosh*, Benjamin Izar*, Sanjay M. Prakadan, Marc H. Wadsworth II, Daniel Treacy, John J. Trombetta, Diana Lu, Asaf Rotem, Christine Lian, George Murphy, Ofir Cohen, Eli van Allen, Monica Bertagnolli, Alex Genshaft, Travis K. Hughes, Carly G. K. Ziegler, Samuel W. Kazer, Aleth Gaillard, Kellie E. Kolb, Judit Valbuena, Charles Yoon, Orit Rozenblatt-Rosen, Alex K. Shalek, Aviv Regev and Levi Garraway, "Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq," *Science*, 352, (2016).

** Denotes equal authorship*

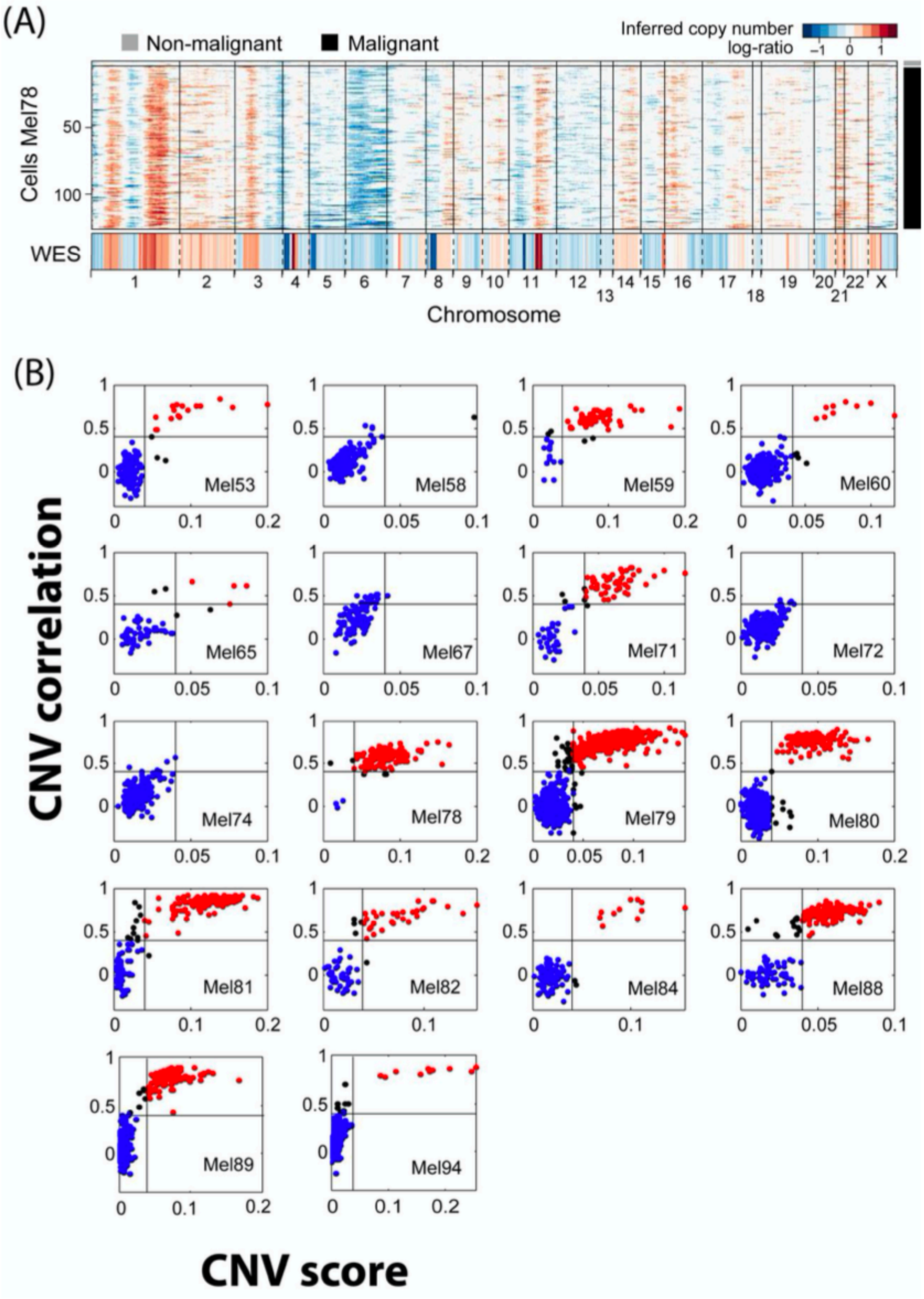

(A)
Non-malignant
Malignant
Inferred copy number
log-ratio
−1
0
1
Cells Mel78
50
100
WES
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
Chromosome
(B)
CNV correlation
Mel53
Mel58
Mel59
Mel60
Mel65
Mel67
Mel71
Mel72
Mel74
Mel78
Mel79
Mel80
Mel81
Mel82
Mel84
Mel88
Mel89
Mel94
0
0.5
1
0.05
0.1
0.2
CNV score

Figure S1 | Classification of cells to malignant and non-malignant based on inferred CNV patterns.

(A) Same as shown in Figure 1B for another melanoma tumor (Mel78). **(B)** Each plot compares two CNV parameters for all cells in a given tumor: (1) CNV score (X-axis) reflects the overall CNV signal, defined as the mean square of the CNV estimates across all genomic locations; (2) CNV correlation (Y-axis) is the Pearson correlation coefficient between each cell's CNV pattern and the average CNV pattern of the top 5% of cells from the same tumor with respect to CNV signal (i.e., the most confidently-assigned malignant cells). These two values were used to classify cells as malignant (red; CNV score > 0.04; correlation score > 0.4; grey lines mark thresholds on plot), non-malignant (blue; CNV score < 0.04; correlation score < 0.4), or unresolved intermediates (black, all remaining cells). In four tumors (Mel58, 67, 72 and 74), we sequenced primarily the immune infiltrates (CD45+ cells) and there were only zero or one malignant cells by this definition; in those cases, CNV correlation is not indicative of malignant cells (since the top 5% cells by CNV signal are primarily non-malignant) and therefore all cells except for one in Mel58 were defined as non- malignant. Note that while these thresholds are somewhat arbitrary, this classification was highly consistent with the clustering patterns of these cells (as shown in Chapter 2, Figure 1C) into clusters of malignant and non-malignant cells.

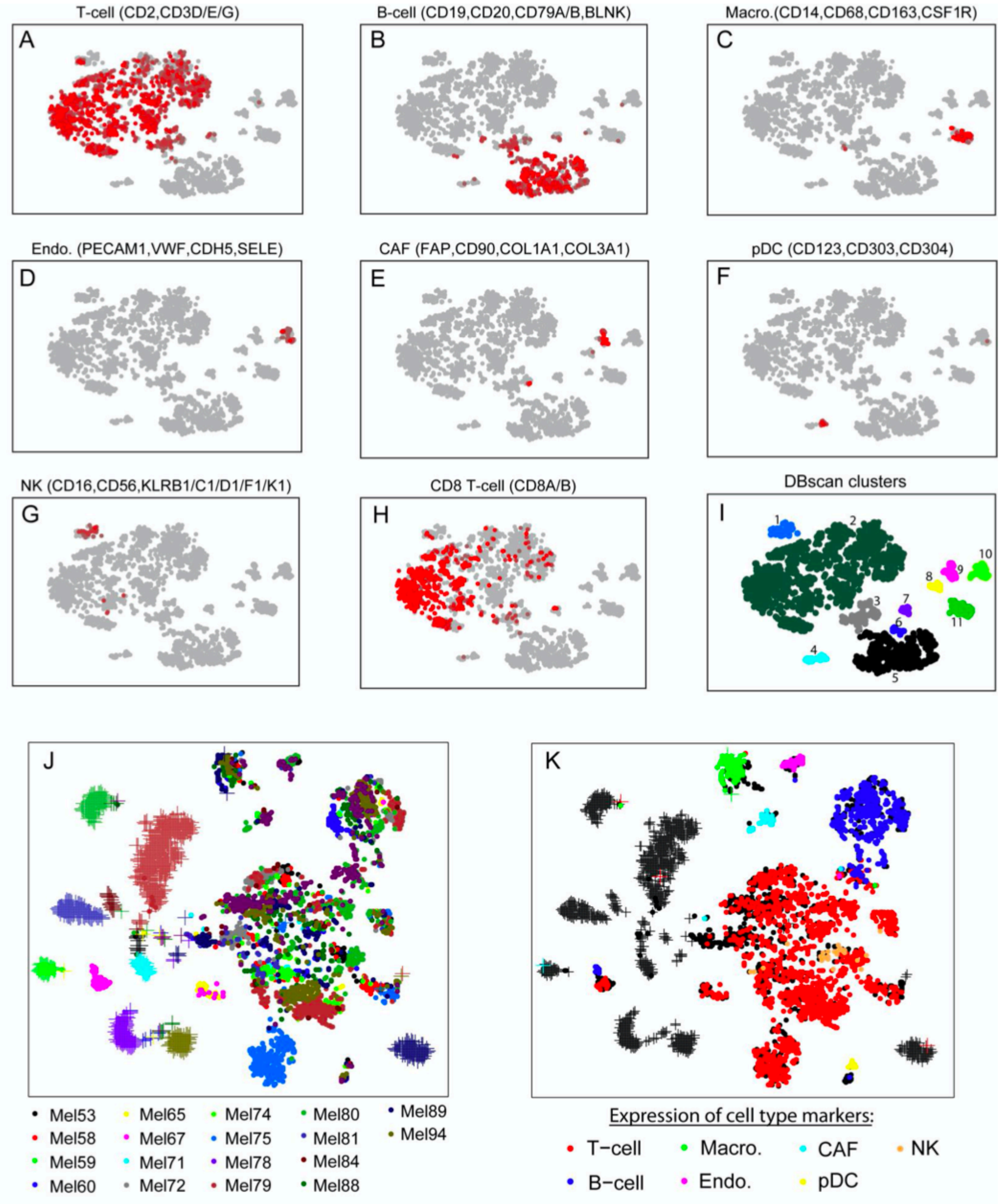

T-cell (CD2,CD3D/E/G)
A
B-cell (CD19,CD20,CD79A/B,BLNK)
B
Macro.(CD14,CD68,CD163,CSF1R)
C
Endo. (PECAM1,VWF,CDH5,SELE)
D
CAF (FAP,CD90,COL1A1,COL3A1)
E
pDC (CD123,CD303,CD304)
F
NK (CD16,CD56,KLRB1/C1/D1/F1/K1)
G
CD8 T-cell (CD8A/B)
H
DBscan clusters
I
1
2
3
4
5
6
7
8
9
10
11
J
K
● Mel53
● Mel58
● Mel59
● Mel60
● Mel65
● Mel67
● Mel71
● Mel72
● Mel74
● Mel75
● Mel78
● Mel79
● Mel80
● Mel81
● Mel84
● Mel88
● Mel89
● Mel94
Expression of cell type markers:
● T−cell
● B−cell
● Macro.
● Endo.
● CAF
● pDC
● NK

Figure S2 | Identification of non-malignant cell types by tSNE clusters that preferentially express cell type markers. (A–H) Each plot shows the average expression of a set of known marker genes for a particular cell type (as indicated at the top) overlaid on the tSNE plot of non-malignant cells, as shown in Chapter 2, Figure 1C. Gray indicates cells with no or minimal expression of the marker genes (E, average log2(TPM+1), below 4), dark red indicates intermediate expression (4 < E < 6), and light red indicates cells with high expression (E > 6). **(I)** DBscan clusters derived from tSNE coordinates, with parameters eps = 6 and min-points = 10. Eleven clusters are indicated by numbers and colors. **(J-K)** Combined tSNE plot of all cells profiled in this work. Colors indicate the tumor-of-origin in **(J)** and the expression of cell type-specific marker genes (E > 5) in **(K)**.

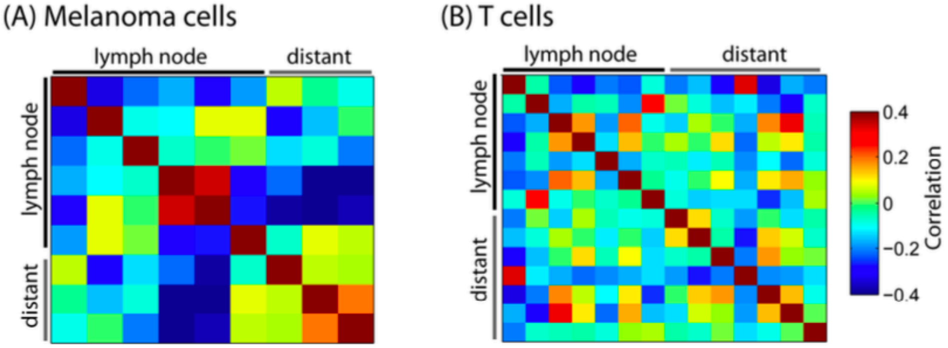


Figure S3 | Limited influence of tumor site on RNA-seq patterns. (A–B) Heat maps show correlations of global expression profiles between tumors, which were ordered by metastatic site. Expression levels were first averaged over melanoma **(A)** or T cells **(B)** in each tumor and then centered across the different tumors before calculating Pearson correlation coefficients. Differential expression analysis conducted between the two groups of tumors found zero differentially expressed genes with FDR of 0.05 based on a shuffling test for both T cells and melanoma cells.

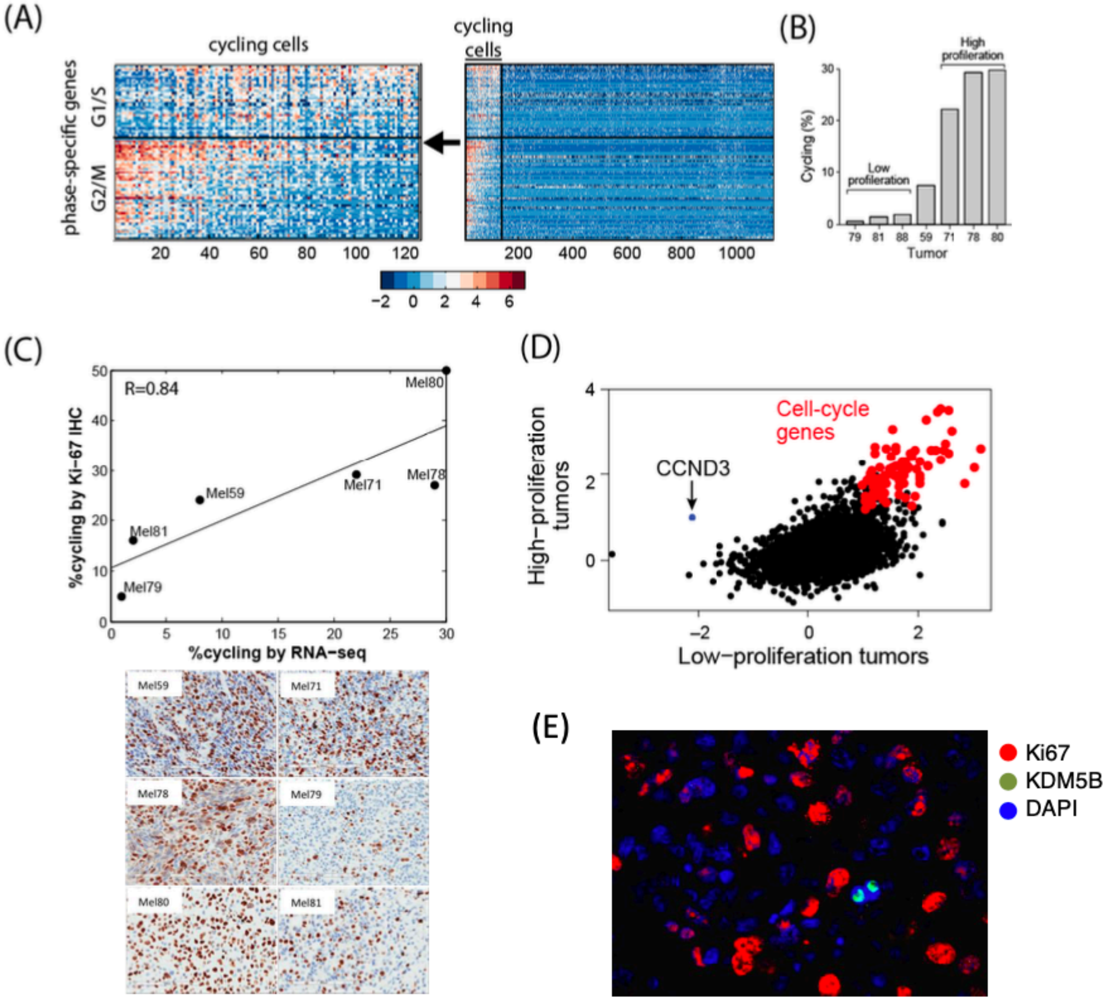

(A)
cycling cells
cycling cells
phase-specific genes
G1/S
G2/M
20 40 60 80 100 120
200 400 600 800 1000
−2 0 2 4 6
(B)
High proliferation
Low proliferation
Cycling (%)
0 10 20 30
79 81 88 59 71 78 80
Tumor
(C)
R=0.84
%cycling by Ki−67 IHC
%cycling by RNA−seq
Mel80
Mel59
Mel71
Mel78
Mel81
Mel79
0 10 20 30 40 50
0 5 10 15 20 25 30
Mel59
Mel71
Mel78
Mel79
Mel80
Mel81
(D)
Cell-cycle genes
CCND3
High−proliferation tumors
Low−proliferation tumors
4 2 0
−2 0 2
(E)
● Ki67
● KDM5B
● DAPI

Figure S4 | Identification and characterization of cycling malignant cells. (A) Heat map showing relative expression of G1/S (**top**) and G2/M (**bottom**) genes (**rows**, as defined from integration of multiple datasets; **Methods**) across cycling cells (**left panel**, columns, ordered by the ratio of expression of G1/S genes to G2/M genes) and across all cells (**right panel**, columns, cycling cells ordered as in left panel followed by non-cycling cells at random order). Cycling cells were defined as those with significantly high expression of G1/S and/or G2/M genes (FDR<0.05 by t-test, and fold-change > 4 compared to all malignant cells). **(B)** The frequency of inferred cycling cells (Y axis) in seven tumors (X axis) with > 50 malignant cells/tumors, denoting low (≤ 3%) or high (>20%) proliferation tumors. **(C, upper panel)** Significant correlation (P < 0.038) between inferred proportion of cycling cells by single-cell transcriptome analysis (horizontal axis) and Ki67+ immunohistochemistry (IHC) (**lower panel**) of corresponding tumor slides (vertical axis). **(D)** Comparison of cycling cell expression programs between low- and high-proliferation tumors. Scatter plots compared the expression log-ratio between cycling and non-cycling cells in high-proliferation (y-axis) and low-proliferation (x-axis) tumors. Genes significantly upregulated (P < 0.01, fold-change > 2) in cycling cells in both types of tumor are marked in red. CCND3 (arrow) is significantly upregulated in cycling cells in high-proliferation tumors and downregulated in cycling cells in low-proliferation tumors. **(E)** Dual KDM5B (JARID1B)/Ki67 immunofluorescence staining of tissue slide of Mel80 (40x magnification). Consistent with findings presented for Mel78 and Mel79 in Figure 2C, KDM5B-expressing cells (green nuclear staining) occurred in small clusters of two or more cells and do not express Ki67 (red nuclear staining), indicating that these cells are not undergoing cell division.

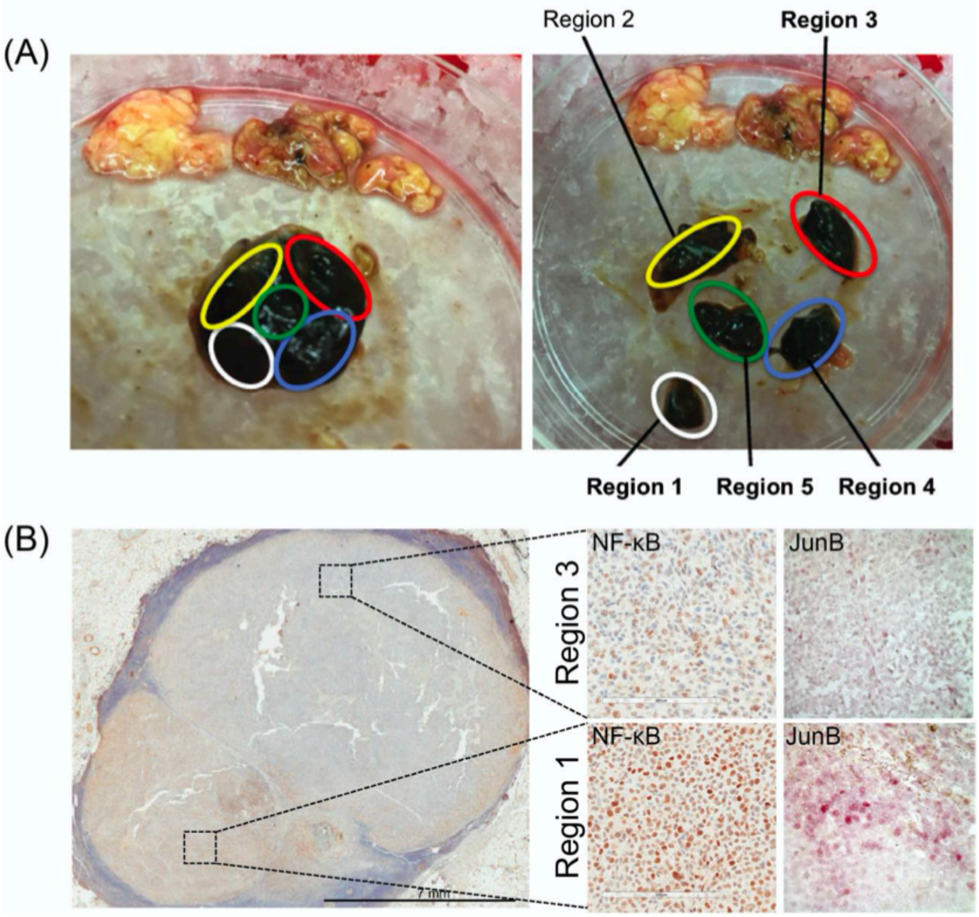


Figure S5 | Immunohistochemistry of melanoma 79 shows gross differences between tumor parts and increased NF-κB levels in Region 1. (A) Tumor dissection into five regions. Left: melanoma tumor prior to dissection. Macroscopically distinct regions are highlighted by colored ovals. Right: The tumor was dissected into five pieces, which were further processed as individual samples. Regions 1, 3, 4 and 5 were included in the single-cell RNA-seq analysis, Cells from Region 2 were lost during library construction. **(B)** Corresponding histopathological cross-section of the tumor demonstrates distinct features of Region 1 compared to the other regions. Consistent with enrichment of cells in Region 1 expressing multiple markers that are highlighted in Fig. 2D, immunohistochemistry staining revealed increased staining of NF-κB and JunB in Region 1 (right lower panel, 40x magnification), compared to region Region 3 (**right upper panel**, 40x magnification).

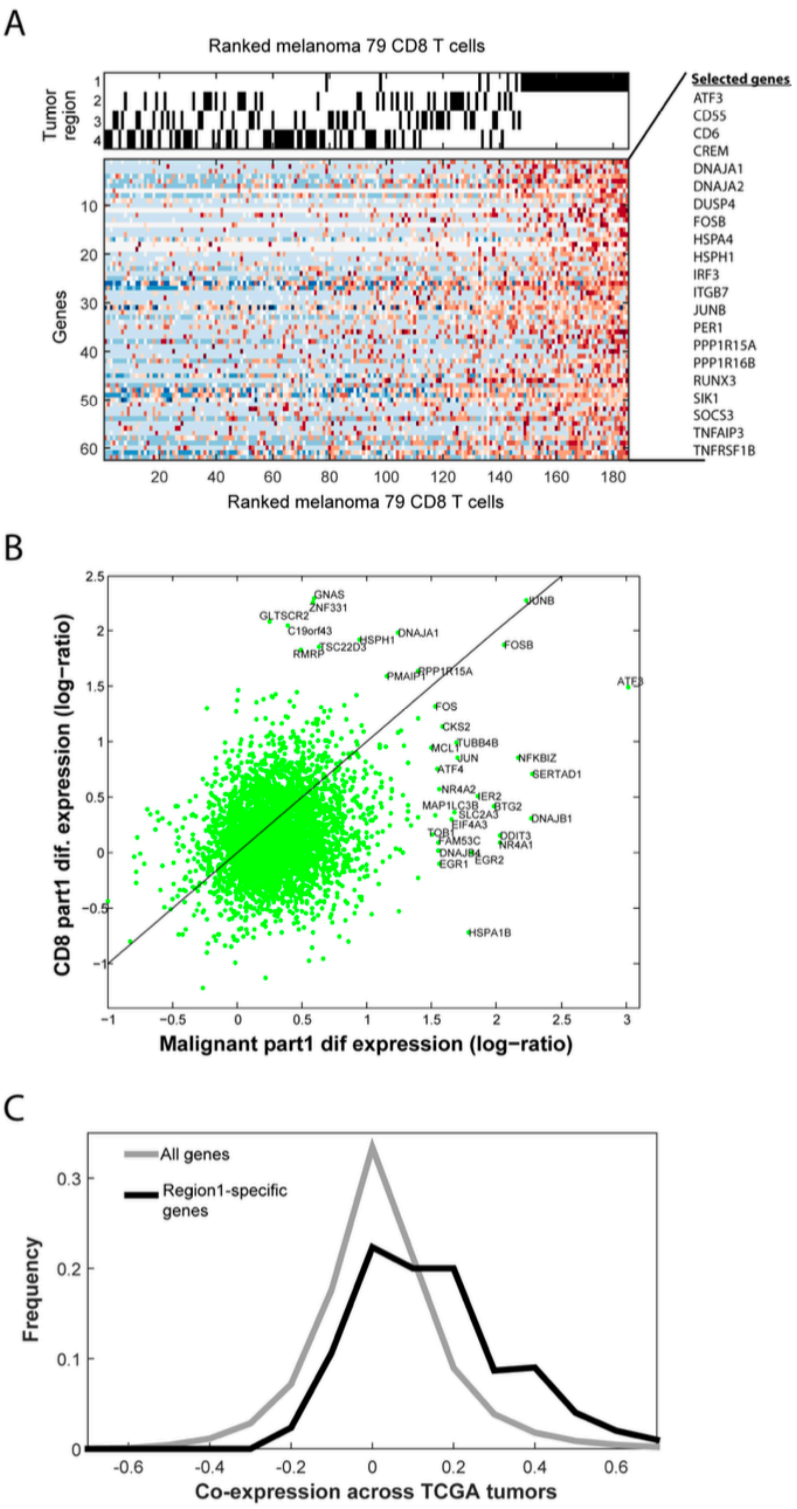

A
Ranked melanoma 79 CD8 T cells
Tumor region
1
2
3
4
Selected genes
ATF3
CD55
CD6
CREM
DNAJA1
DNAJA2
DUSP4
FOSB
HSPA4
HSPH1
IRF3
ITGB7
JUNB
PER1
PPP1R15A
PPP1R16B
RUNX3
SIK1
SOCS3
TNFAIP3
TNFRSF1B
Genes
10
20
30
40
50
60
20
40
60
80
100
120
140
160
180
Ranked melanoma 79 CD8 T cells
B
CD8 part1 dif. expression (log−ratio)
2.5
2
1.5
1
0.5
0
−0.5
−1
GNAS
ZNF331
GLTSCR2
C19orf43
HSPH1
DNAJA1
RMRP
TSC22D3
JUNB
FOSB
PMAIP1
PPP1R15A
ATF3
FOS
CKS2
MCL1
TUBB4B
JUN
NFKBIZ
ATF4
SERTAD1
NR4A2
IER2
MAP1LC3B
BTG2
SLC2A3
DNAJB1
EIF4A3
TOB1
DDIT3
FAM53C
NR4A1
DNAJB4
EGR2
EGR1
HSPA1B
−1
−0.5
0
0.5
1
1.5
2
2.5
3
Malignant part1 dif expression (log−ratio)
C
All genes
Region1-specific genes
Frequency
0.3
0.2
0.1
0
-0.6
-0.4
-0.2
0
0.2
0.4
0.6
Co-expression across TCGA tumors

Figure S6 | Spatial heterogeneity in Mel79. (A-B) As shown in Chapter 2, Figure 2D for malignant cells, we examined the expression differences between regions of Mel79 for other cell types. The only cell type for which we had >10 cells in each of the regions was CD8+ T cells. We thus focused on the differences among CD8+ T cells and found 62 genes that were preferentially expressed in region 1 (fold-change>2, FDR<0.05) and that partially overlapped the region 1-specific genes among the malignant cells (see Table S6). **(A)** Region 1-specific expression program of CD8+ T-cells (as shown in Figure 2D for malignant cells). Bottom: heat map shows the relative expression of the 62 genes preferentially expressed in region 1, in all CD8+ T-cells from Mel79, ranked by their average expression of these genes. A subset of genes of interest are noted at the right. Top: assignment of cells to the four regions of Mel79. **(B)** Comparison of region 1 preferential expression between malignant cells (X-axis) and CD8+ T-cells (Y-axis). For each cell type, the scatterplot shows the log2-ratio between the average expression of all cells in region 1 and those in all other regions. **(C)** The top region1-specific genes from analysis of malignant cells (25 genes with 3-fold upregulation in region 1 compared to all other regions of Mel79) are co-expressed across melanoma TCGA bulk tumors. Shown are the distribution of Pearson correlation coefficients among the expression profiles of the top region1-specific genes (black), which is significantly higher (P<0.001 by permutation test) than the distribution of correlations among all genes (gray).

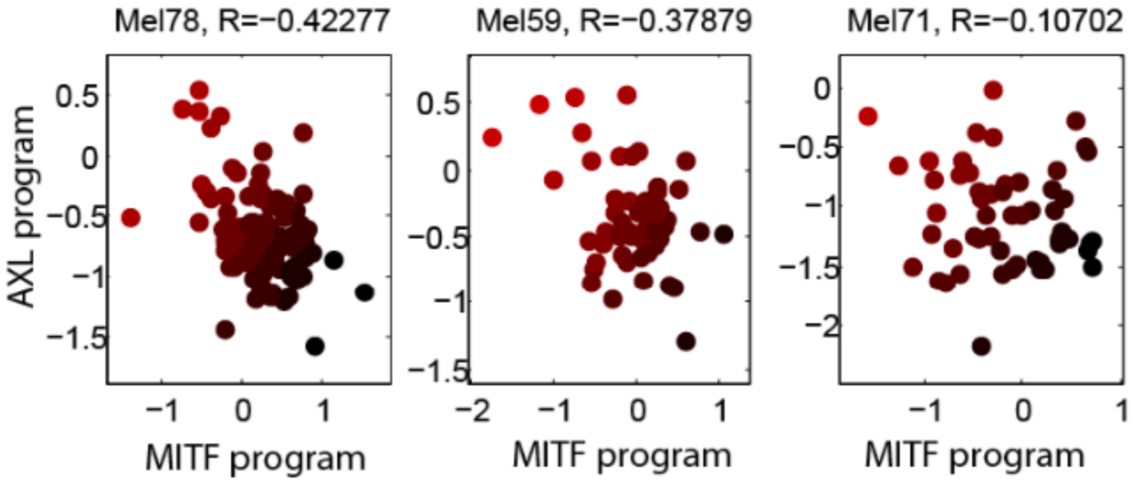


Figure S7 | Intra-tumor heterogeneity in AXL and MITF programs. AXL-program (Y-axis) and MITF-program (X-axis) scores for malignant cells in each of the three tumors with a sufficient number of malignant cells (n>50) that were not included in Figure 3B. Cells are colored from black to red by the relative AXL and MITF scores. The Pearson correlation coefficient is denoted on top.

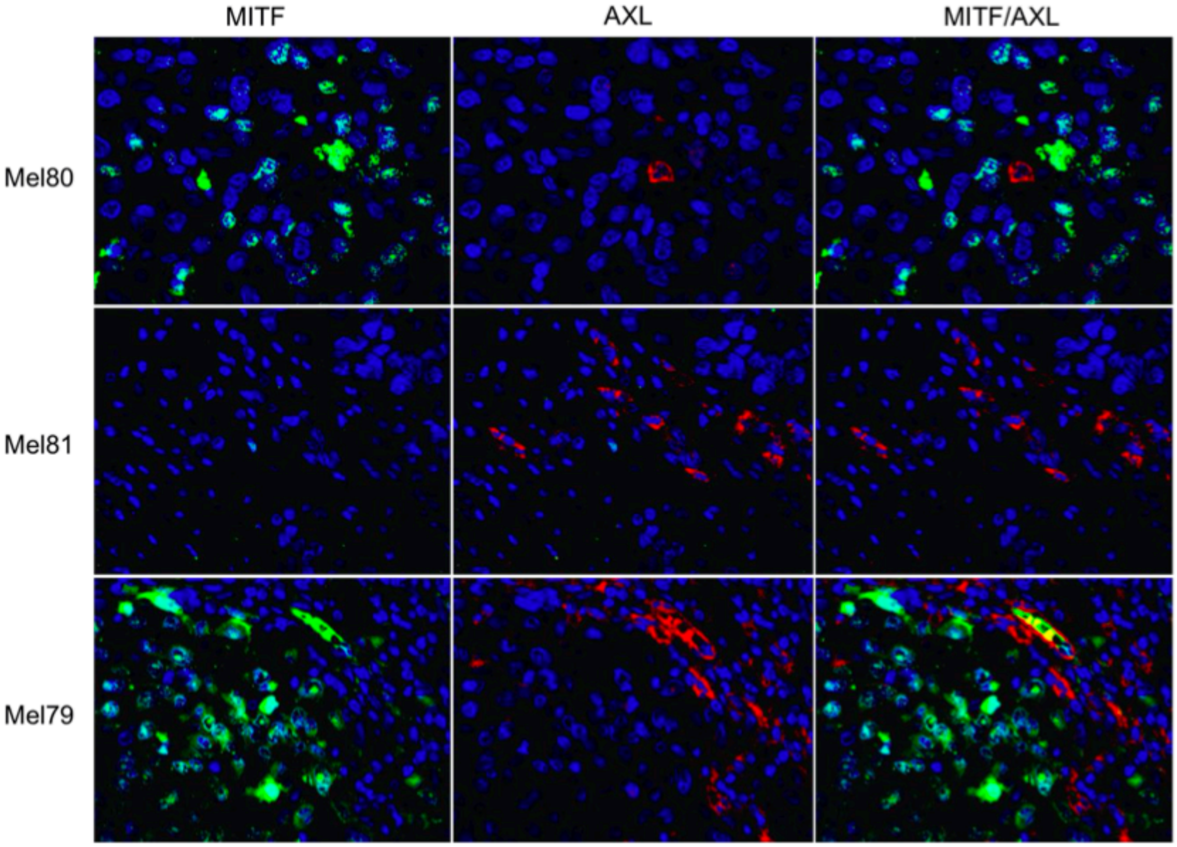


Figure S8 | AXL/MITF immunofluorescence staining of tissue slides of Mel80, Mel81 and Mel79 (40x magnification) revealed presence of AXL-expressing and MITF-expressing cells in each sample. Consistent with single-cell RNA-seq inferred frequencies of each population, Mel80 contained rare AXL-expressing cells (red, cell membrane staining) and mostly malignant MITF-positive cells (green, nuclear staining), while malignant cells of Mel81 almost exclusively consisted of AXL-expressing cells. Mel79 had a mixed population with rare cells positive for both markers, all in agreement with the inferred single-cell transcriptome data.

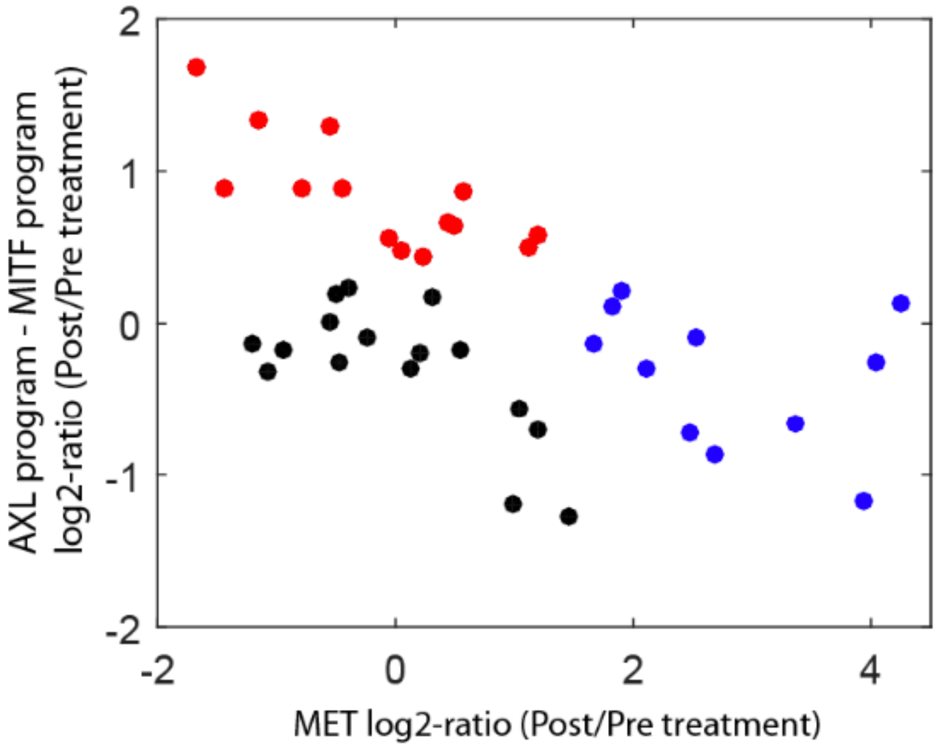


Figure S9 | AXL upregulation in a second cohort of post-treatment melanoma samples and mutual exclusivity with MET upregulation. Each point reflects a comparison between a matched pair of pre-treatment and post-relapse samples from Hugo et al., where the X-axis shows expression changes in MET, and the Y-axis shows expression changes in the AXL program minus those of the MITF program. Note that some patients are represented more than once based on multiple post-relapse samples. Fourteen out of 41 samples (34%) shown in red had significant upregulation of the AXL vs. MITF program, as determined by a modified t-test as described in Methods; these correspond to at least one sample from half (9/18) of the patients included in the analysis. Eleven out of 41 samples (27%) shown in blue had at least 3-fold upregulation of MET; these correspond to at least one sample from a third (6/18) of the patients included in the analysis. Notably, the AXL and MET upregulated samples are mutually exclusive, consistent with the possibility that these are alternative resistance mechanism.

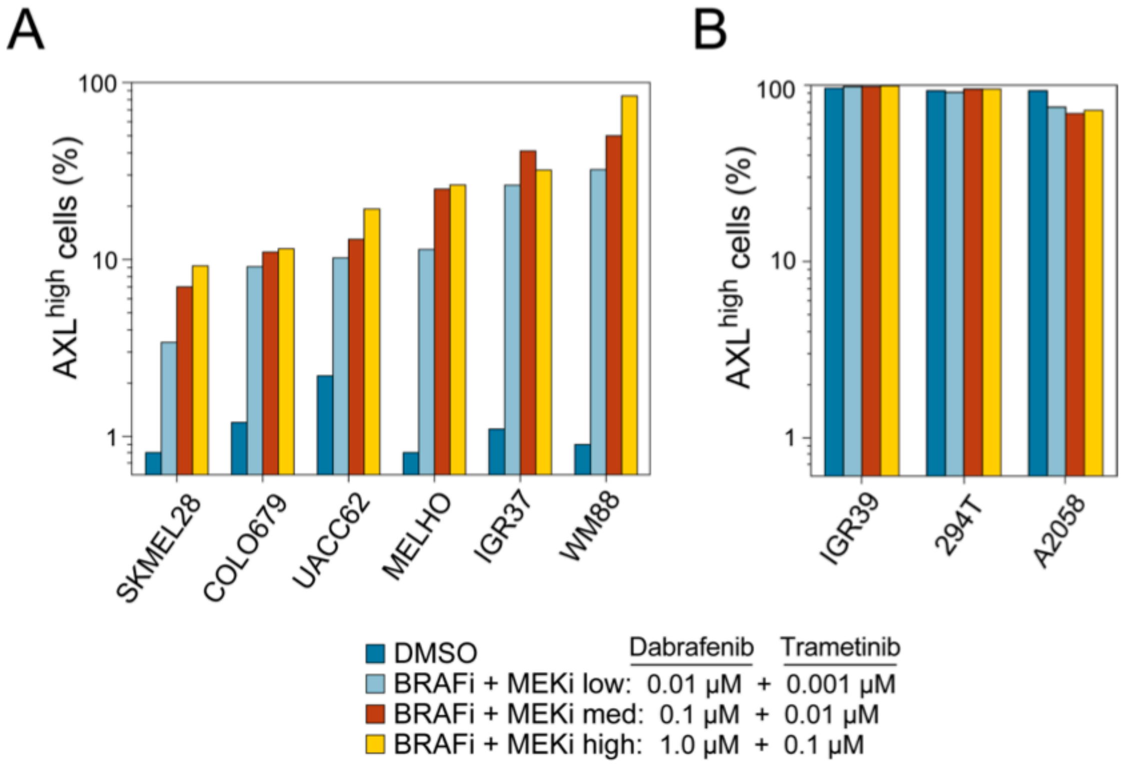
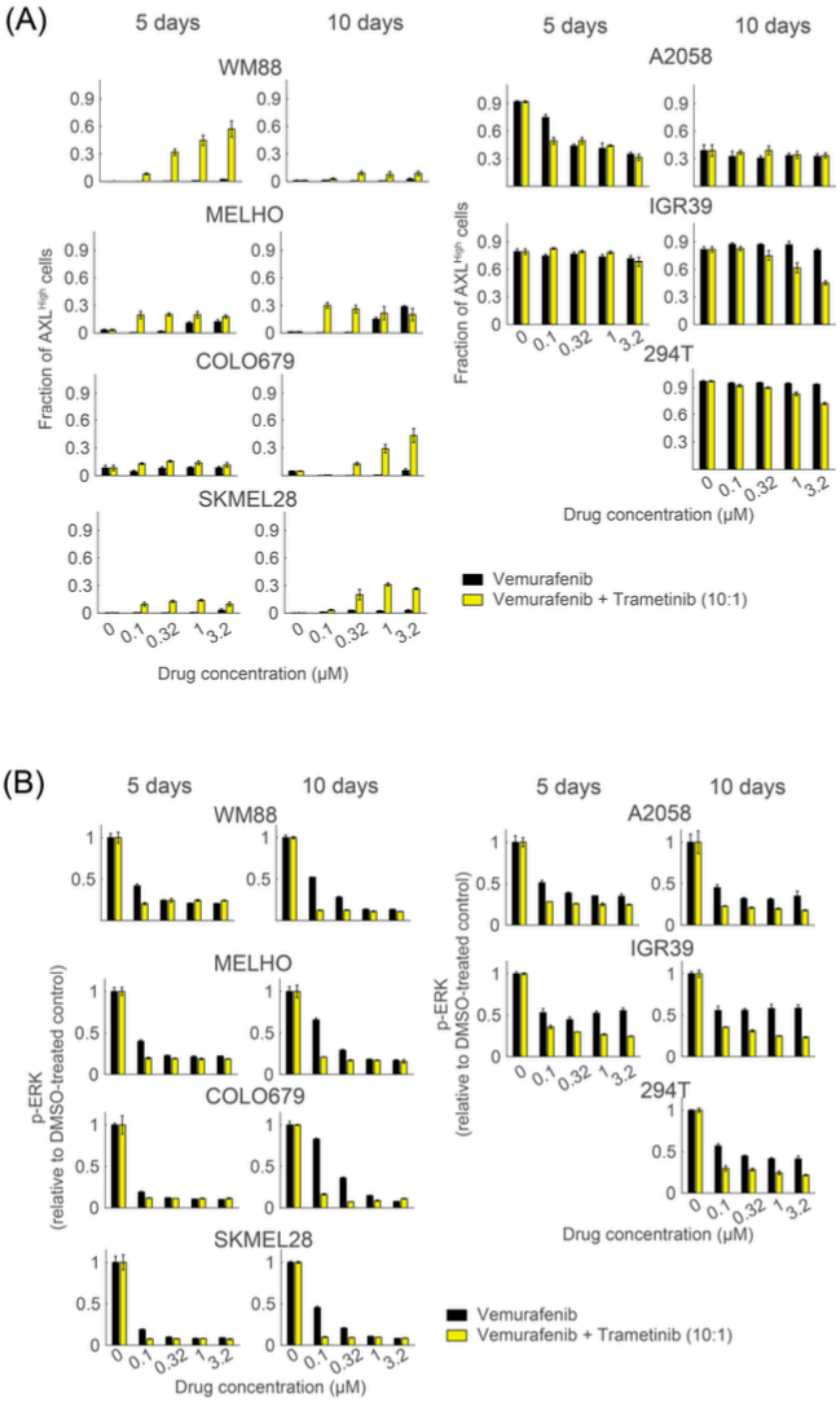


Figure S10 | Flow-cytometry of melanoma cell lines before and after treatment with RAF/MEK- inhibition. (A) Sensitive cell lines show an increased proportion of AXL-positive cells while resistant cell lines **(B)** show modest or no changes following treatment with RAF/MEK-inhibitors.


(A)
5 days
10 days
WM88
MELHO
COLO679
SKMEL28
A2058
IGR39
294T
Fraction of AXL^High cells
0.9
0.6
0.3
0
0
0.1
0.32
1
3.2
Drug concentration (μM)
■ Vemurafenib
■ Vemurafenib + Trametinib (10:1)
(B)
5 days
10 days
WM88
MELHO
COLO679
SKMEL28
A2058
IGR39
294T
p-ERK
(relative to DMSO-treated control)
1
0.5
0
0
0.1
0.32
1
3.2
Drug concentration (μM)
■ Vemurafenib
■ Vemurafenib + Trametinib (10:1)

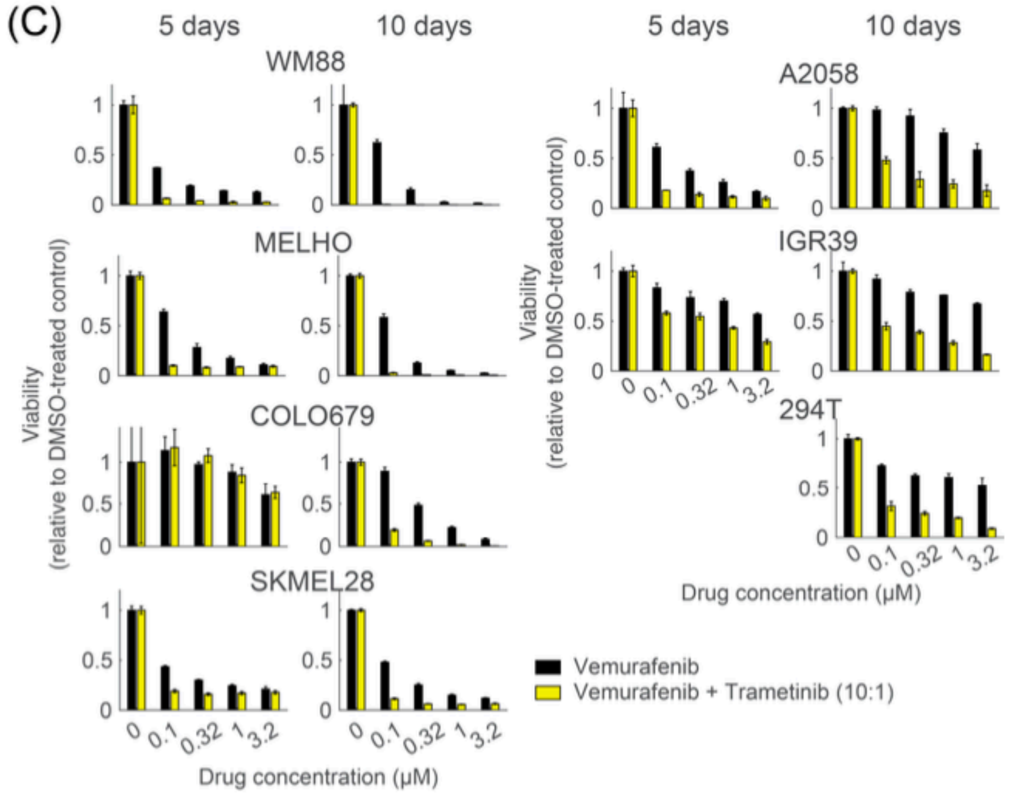

(C)
5 days
10 days
5 days
10 days
WM88
A2058
MELHO
IGR39
COLO679
294T
SKMEL28
Viability
(relative to DMSO-treated control)
Viability
(relative to DMSO-treated control)
1
0.5
0
0
0.1
0.32
1
3.2
Drug concentration (μM)
Drug concentration (μM)
Vemurafenib
Vemurafenib + Trametinib (10:1)

Figure S11 | Summary of multiplexed single-cell immunofluorescence in seven CCLE cell lines before and after treatment with BRAF/MEK-inhibition. (A) Relative fraction (compared to DMSO- treatment) of AXL-high cells (y-axis) treated for 5 or 10 days with increasing doses (as indicated on x-axis) of BRAF-inhibition alone (with vemurafenib) or in combination with a MEK-inhibitor (trametinib) with a 10:1 ratio (vemurafenib:trametinib). In all cell lines with a baseline low-fraction of AXL- expressing cells (WM88, MELHO, COLO679 and SKMEL28), there was a significant dose-dependent increase in the AXL-high cell fraction with BRAF-inhibition alone (black bars), and more pronounced with combined BRAF/MEK-inhibition (yellow bars). Cell lines with a baseline high AXL-expressing cell fraction (A2058, IGR39 and 294T) showed either minimal changes in the AXL-high cell fraction, however, A2058 demonstrated a significant decreased in the AXL-positive fraction. Although an outlier in this experiment, this indicates that alternative mechanisms of resistance with low AXL expression (Hugo et al.; Figure S9). **(B)** The increase in AXL-high cell fractions in the sensitive cell lines was correlated with a significant decrease of p-ERK indicating strong MAP-kinase pathway inhibition, and **(C)** a decrease in cell viability. Overall, these results indicate, that the increase in the AXL-high cell fraction was at least in part due to a selection process. Both effects were more pronounced when cells were treated with combined BRAF/MEK-inhibition compared BRAF-inhibition alone.

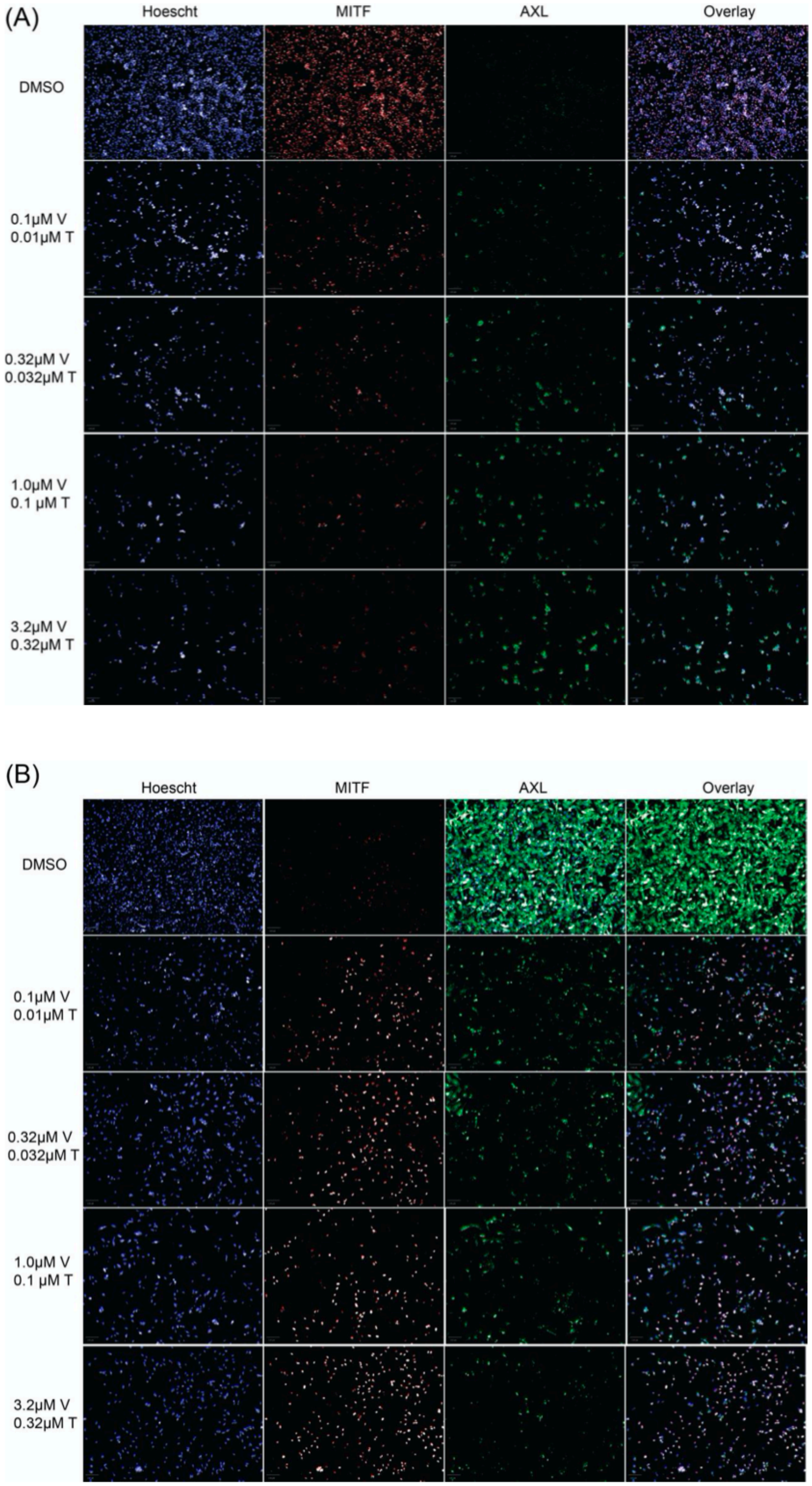

(A)
Hoescht
MITF
AXL
Overlay
DMSO
0.1µM V
0.01µM T
0.32µM V
0.032µM T
1.0µM V
0.1 µM T
3.2µM V
0.32µM T
(B)
Hoescht
MITF
AXL
Overlay
DMSO
0.1µM V
0.01µM T
0.32µM V
0.032µM T
1.0µM V
0.1 µM T
3.2µM V
0.32µM T

Figure S12 | Exemplary images of multiplexed single-cell immunofluorescence quantitative analysis for (A) an AXL-low (WM88) and (B) AXL-high cell line (A2058). Treatment with a combination of vemurafenib (V) and trametinib (T) at indicated doses on the left resulted in a dose-dependent change in the AXL-high population. In WM88, increasing drug concentrations led to killing of MITF-expressing, resulting in the emergence of a pre-existing AXL-high subpopulation. This indicates that the shift towards a higher AXL-expressing population (and possibly the AXL-high signature) is at least in part due to a selection process. While cell lines with a high baseline fraction of AXL-expressing cells showed modest to no changes in the AXL-fraction (Figure S10B), A2058 was an exception. This cell lines has a major AXL-expressing population at baseline, which decreases with treatment, while the MITF-expressing population emerges. This indicates the presence of alternative mechanisms of resistance to RAF/MEK-inhibition, consistent with a recent report by Hugo et al. and our analysis shown in Figure S9.

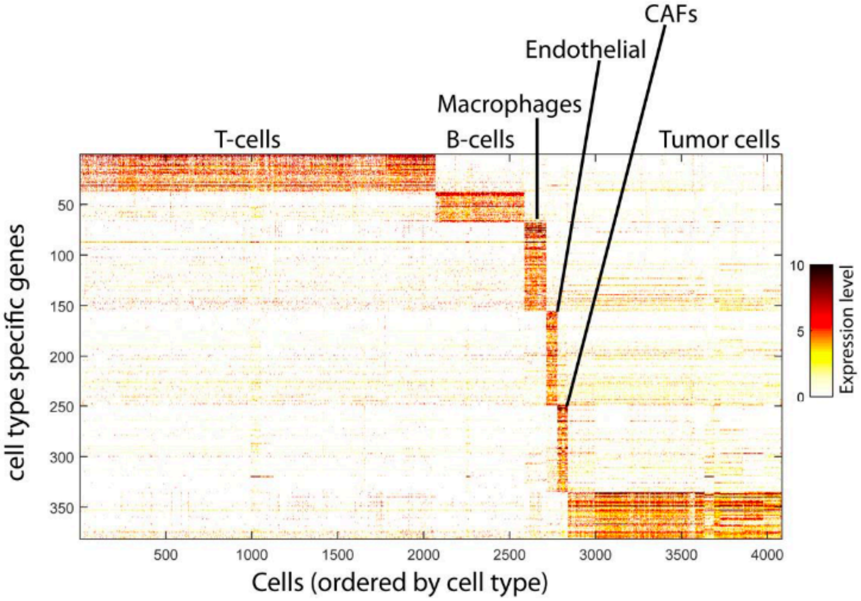


Figure S13 | Identification of cell-type specific genes in melanoma tumors. Shown are the cell- type specific genes (**rows**) as chosen from single cell profiles, sorted by their associated cells cell type, and their expression levels (log2(TPM/10+1)) across non-malignant and malignant tumor cells, also sorted by type (**columns**).

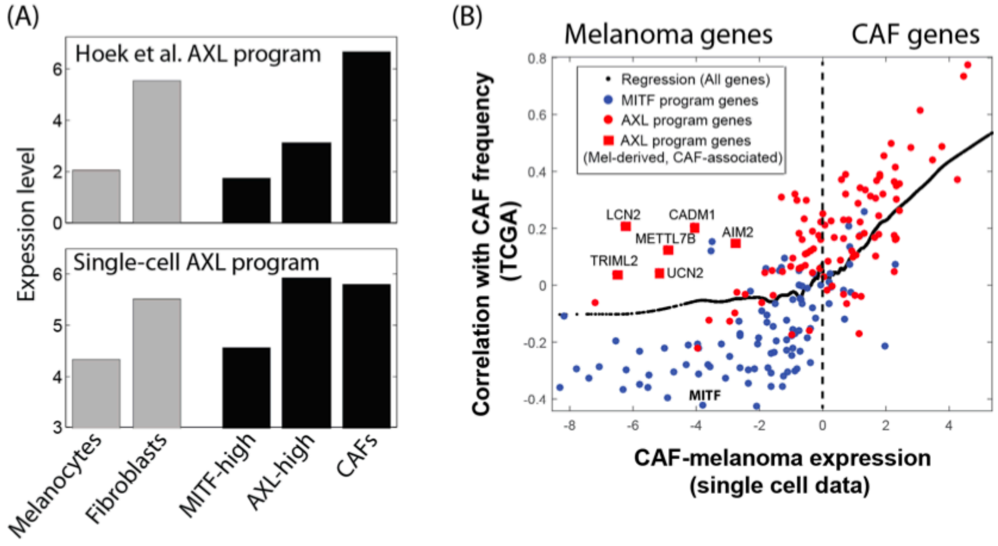


Figure S14 | Association between a malignant AXL program and CAFs. (A) Average expression (log2(TPM+1)) of the AXL program (Y-axis) as defined here (bottom) and by Hoek et al. (top, ref. 30) in CAFs and melanoma cells from our tumors (this work, black bars) and in foreskin melanocytes and primary fibroblasts from the Roadmap Epigenome project (grey bars). Melanoma cells were partitioned to those from AXL-high and MITF-high tumors as marked in Chapter 2, Figure 3A. **(B)** CAF expression correlates with higher AXL program than MITF program expression in melanoma malignant cells. Scatter plot shows for each gene (dot) from the MITF (blue) or AXL (red) programs (as defined based on single- cell transcriptomes) the correlation of its expression with inferred CAF frequency across bulk tumors (Y-axis, from TCGA transcriptomes), and how specific its expression is to CAFs vs. melanoma malignant cells (X-axis, based on single-cell transcriptomes). Black dots indicate the expected correlations at each value of the horizontal axis as defined by a LOWESS regression over all genes. The average correlation values of MITF program genes are significantly lower than those of all genes and the correlation values of AXL program genes are significantly higher than those of all genes, even after restricting the analysis to melanoma-specific genes (X-axis < -2, P < 0.01, t-test). A subset of AXL-program genes are specifically expressed in melanoma cells (but not CAFs) based on the single cell expression profiles, but associated with CAF abundance in bulk tumors (marked by red squares and gene names). MITF is negatively correlated with CAF abundance (R=-0.42) and is also indicated by gene name.

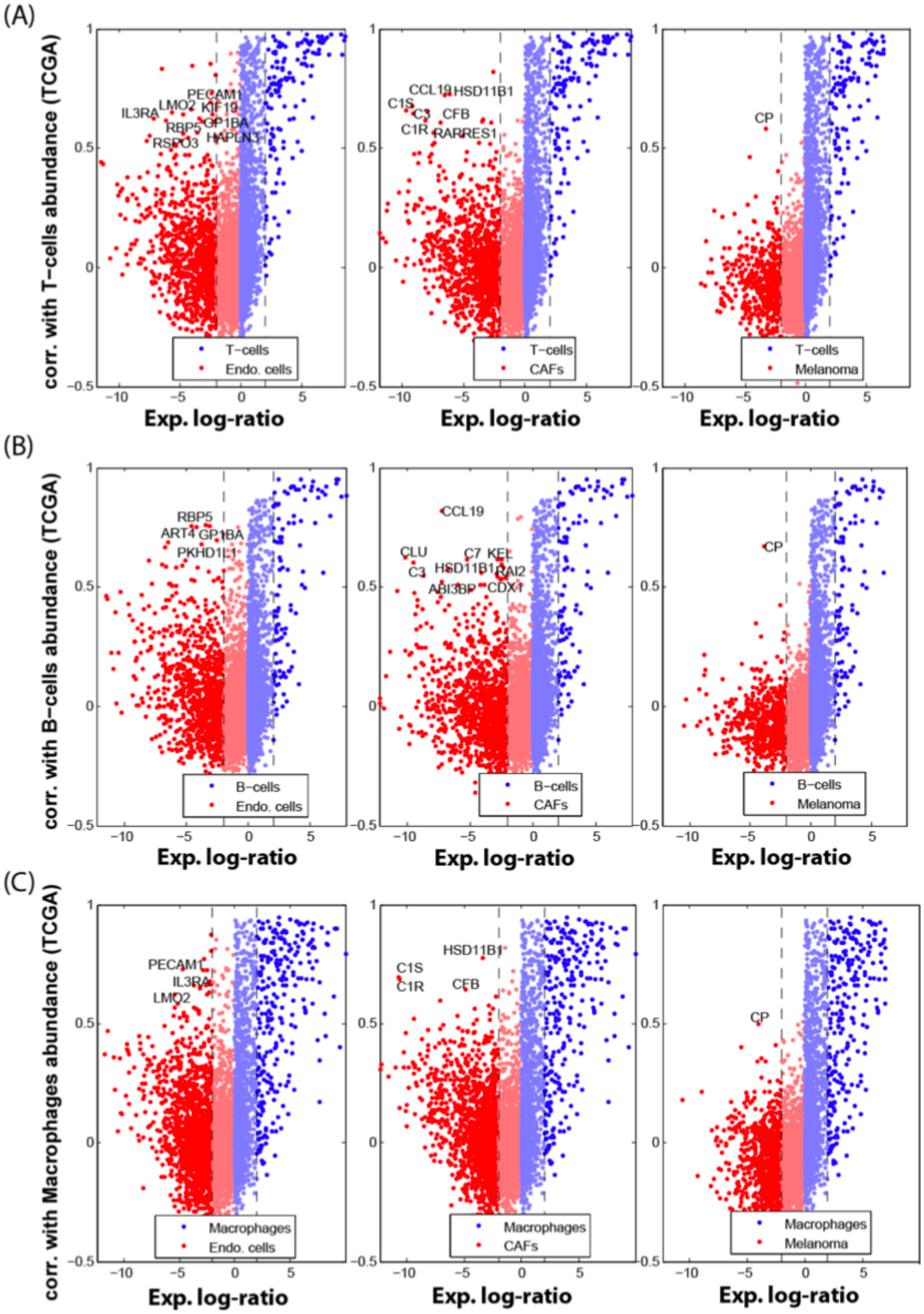

(A)
corr. with T−cells abundance (TCGA)
1
0.5
0
−0.5
−10
−5
0
5
Exp. log-ratio
PECAM1
LMO2
KIF19
IL3RA
RBP5
GP1BA
RSPO3
HAPLN3
T−cells
Endo. cells
CCL19
HSD11B1
C1S
C3
CFB
C1R
RARRES1
T−cells
CAFs
CP
T−cells
Melanoma
(B)
corr. with B−cells abundance (TCGA)
RBP5
ART4
GP1BA
PKHD1L1
B−cells
Endo. cells
CCL19
CLU
C7
KEL
C3
HSD11B1
RAI2
ABI3BP
CDX1
B−cells
CAFs
CP
B−cells
Melanoma
(C)
corr. with Macrophages abundance (TCGA)
PECAM1
IL3RA
LMO2
Macrophages
Endo. cells
HSD11B1
C1S
C1R
CFB
Macrophages
CAFs
CP
Macrophages
Melanoma

Figure S15 | Identification of putative genes underlying cell-to-cell interactions from analysis of single cell profiles and TCGA samples. We searched for genes that underlie potential cell-to-cell interactions, defined as those that are primarily expressed by cell type M (as defined by the single cell data) but correlate with the inferred relative frequency of cell type N (as defined from correlations across TCGA samples). For each pair of cell types (M and N), we restricted the analysis to genes that are at least four-fold higher in cell type M than in cell type N and in any of the other four cell types. We then calculated the Pearson correlation coefficient (R) between the expression of each of these genes in TCGA samples and the relative frequency of cell type N in those samples, and converted these into Z-scores. The set of genes with Z > 3 and a correlation above 0.5 was defined as potential candidates that mediate an interaction between cell type M and cell type N. **(A)** Of all the pairwise comparisons we identified interactions only between immune cells (B, T, macrophages) and non-immune cells (CAFs, endothelial cells, malignant melanoma) cells, such that the expression of genes from non- immune cells correlated with the relative frequency of immune cell types. Each plot shows a single pairwise comparison (M vs. N), including interactions of non-immune cell types (endothelial cells: left; CAFs: middle; malignant melanoma: right) with each of T-cells **(A)**, B-cells **(B)** and macrophages **(C)**. Each plot compares for each gene (dot) the relative expression of genes in the two cell types being compared (M–N) and the correlations of these genes' expression with the inferred frequency of cell type N across bulk TCGA tumors. Dashed lines denote the four-fold threshold. Genes that may underlie potential interactions, as defined above, are highlighted.

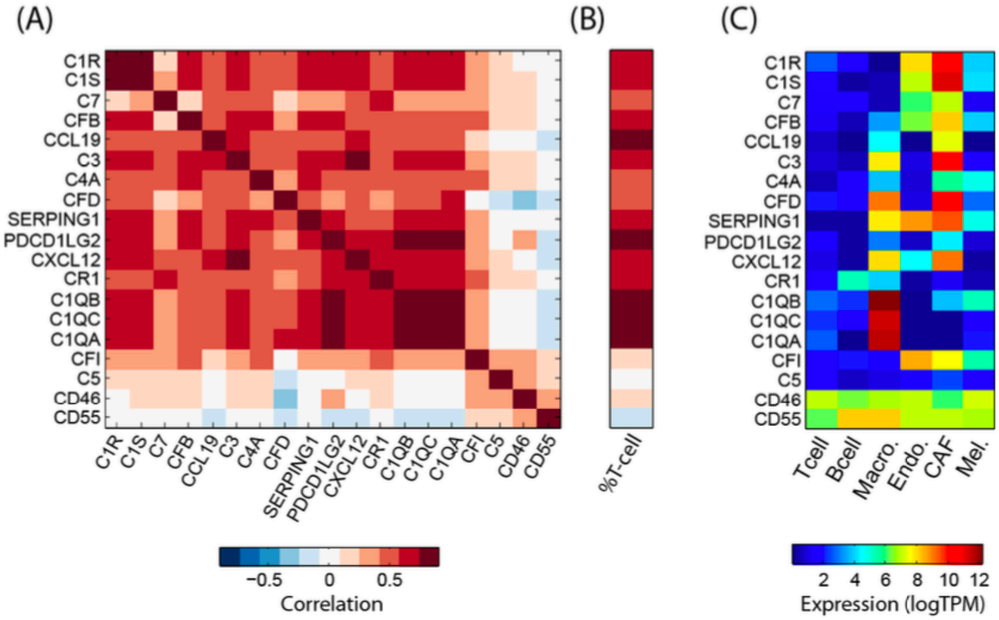


Figure S16 | Immune modulators expressed by CAFs and macrophages. (A) Pearson correlation coefficient (color bar) across TCGA melanoma tumors between the expression level of each of the immune modulators shown in Fig. 4B and additional complement factors with significant expression levels. **(B)** Correlations across TCGA melanoma tumors between the expression level of the genes shown in **(A)** and the average expression levels of T cell marker genes. **(C)** Average expression level (log2(TPM+1), color bar) of the genes shown in **(A)** in the single cell data, for cells classified into each of the major cell types we identified. These results show that most complement factors are correlated with one another and with the abundance of T cells, even though some are primarily expressed by CAFs (including C3) and others by macrophages. In contrast, two complement factors (CFI, C5) and the complement regulatory genes (CD46 and CD55) show a different expression pattern.

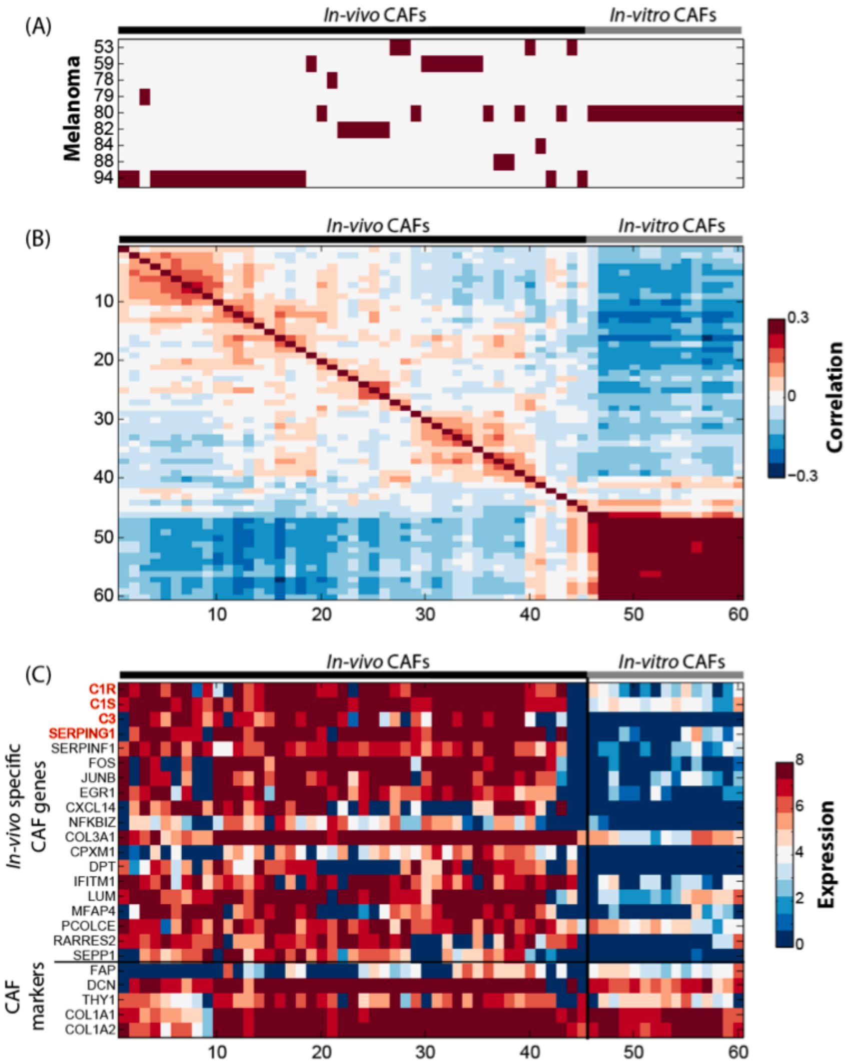


Figure S17. Unique expression profiles of in vivo CAFs. (A-B) Distinct expression profiles in in vivo and in vitro CAFs. Shown are Pearson correlation coefficient between individual CAFs isolated in vivo from seven melanoma tumors, and CAFs cultured from one tumor (melanoma 80). Hierarchical clustering shows two clusters, one consisting of all in vivo CAFs, regardless of their tumor-of-origin (marked in **(A)**), and another of the in vitro CAFs. **(C)** Unique markers of in vivo CAFs include putative cell-cell interaction candidates. Left: Heatmap shows the expression level (log2(TPM+1)) of CAF markers (bottom) and the top 14 genes with higher expression in in-vivo compared to in-vitro CAFs (t-test). Right: average (bulk) expression of the genes in the in-vivo CAFs, in-vitro CAFs, and primary foreskin fibroblasts from the Roadmap Epigenome project. Potential interacting genes from Chatper 2, Figure 4B are highlighted in bold red.

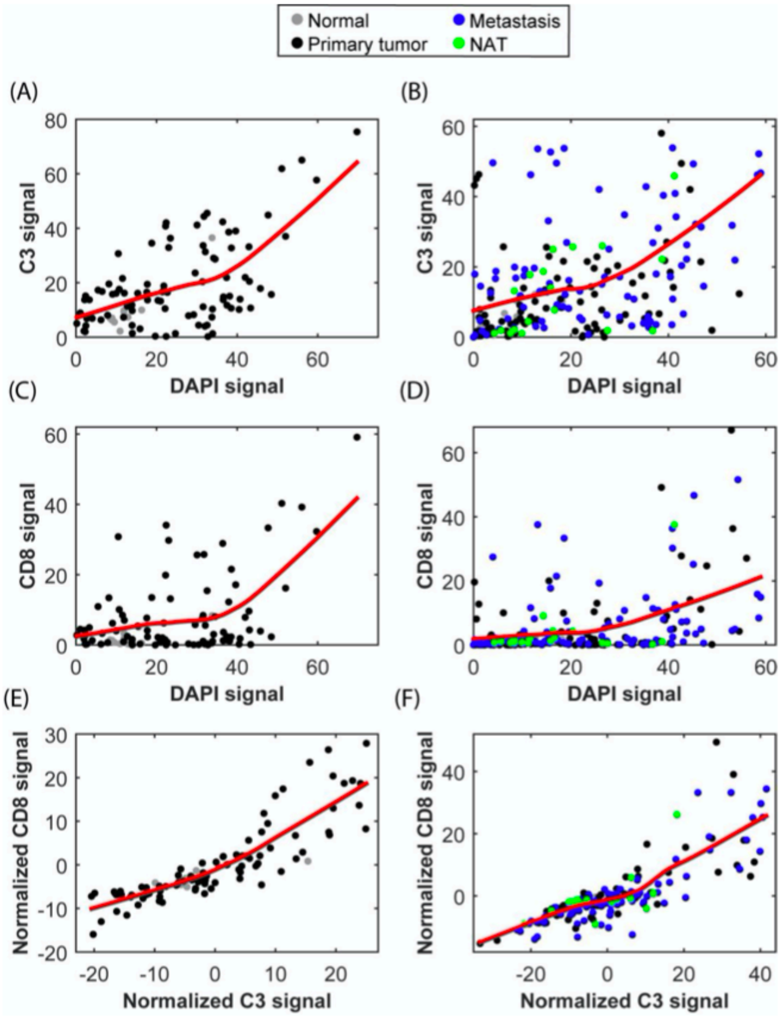

● Normal
● Primary tumor
● Metastasis
● NAT
(A)
(B)
(C)
(D)
(E)
(F)
C3 signal
CD8 signal
DAPI signal
Normalized CD8 signal
Normalized C3 signal
0
20
40
60
80

Figure S18 | TMA analysis of complement factor 3 association with CD8+ T-cell infiltration, and control staining. Two TMAs (CC38-01 and ME208, shown in A, C, E and B, D, F, respectively) were used to evaluate the association between complement factor 3 (C3) and CD8 across a large number of tissues obtained by core biopsies of normal skin, primary tumors, metastatic lesions and NATs (normal skin with adjacent tumor). In both TMAs with a total of 308 core biopsies, we observed high correlation between C3 and CD8 (R > 0.8, shown in Chapter 2, Figure 4C for one TMA). To verify that this correlation is not due to technical effects in which some tissues stain more than others irrespective of the stains examined (e.g., due to variability in cellularity or tissue quality), we normalized the values (%area, Methods) for both C3 and CD8 by those of DAPI staining. Indeed, we found a non-random yet non-linear association between DAPI stains and either C3 **(A, B)**, or CD8 **(C, D)**, which were removed by subtracting a LOWESS regression, shown as red curves in panels A-D. The normalized C3 and CD8 values were not correlated with DAPI levels, yet maintained a high correlation with one another **(E, F)**. R = 0.86 and 0.74 for primary and normal skin in panel E (TMA CC38-01), and R = 0.78, 0.86, 0.63 and 0.31 for primary melanomas, metastasis, NATs and normal skin in panel F (TMA ME208), respectively.

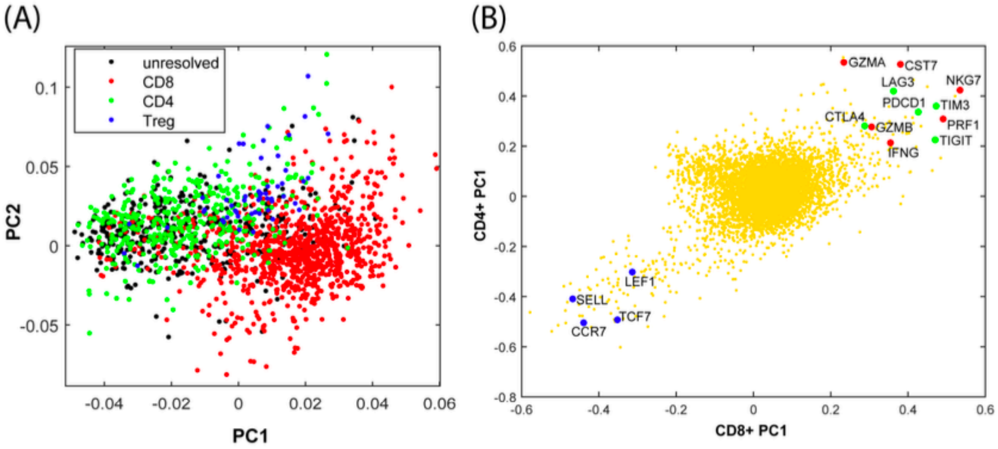


Figure S19 | Cytotoxic and naïve expression programs in T cells. (A) Cell scores from a combined PCA of all T cells. Cells are colored as CD8+ (red), CD4+ (green), T-regs (blue) and unresolved (black) based on expression of marker genes (**Chapter 2**, **Figure 5A**). **(B)** Gene scores for PC1 from a PCA of CD8+ cells (x-axis) and PC2 from a PCA of CD4+ cells (Y-axis). Selected marker genes are highlighted, including genes known to be associated with cytotoxic/active (red), naïve (blue) and exhausted (green) T cell states.

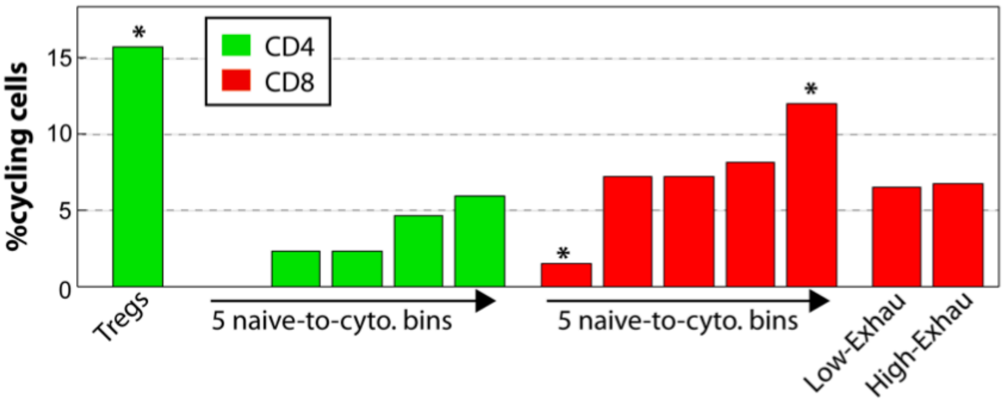


Figure S20 | Frequency of cycling cells in different subsets of T-cells. Shown is the frequency of cycling T cells (as identified based on the expression of G1/S and G2/M gene-sets; Methods) for different subsets of T cells, including Tregs, CD4+ cells separated into five bins of increasing activation (**arrow below green bars**), CD8+ cells separated into five bins of increasing activation (**arrow below red bars**), and active/cytotoxic CD8+ further partitioned into those with relatively high or low exhaustion, as shown in Chapter 2, Figure 5D. Asterisks denote subsets with significant enrichment or depletion of cycling cells across all cells from the same subset of CD4+ or CD8+ cells as defined by $P < 0.05$ in a hypergeometric test. Cell cycle frequency is associated with activation state of CD8+ T-cells, as the first bin is significantly depleted and the fifth bin is significantly enriched. A similar trend is observed in CD4+ T-cells (no cycling cells in the first bin and highest frequency in fifth bin), although none of the CD4 bins was significantly depleted or enriched. Exhaustion was not associated with significant differences in cell cycle frequency ($P = 0.34$, Chi-square test).

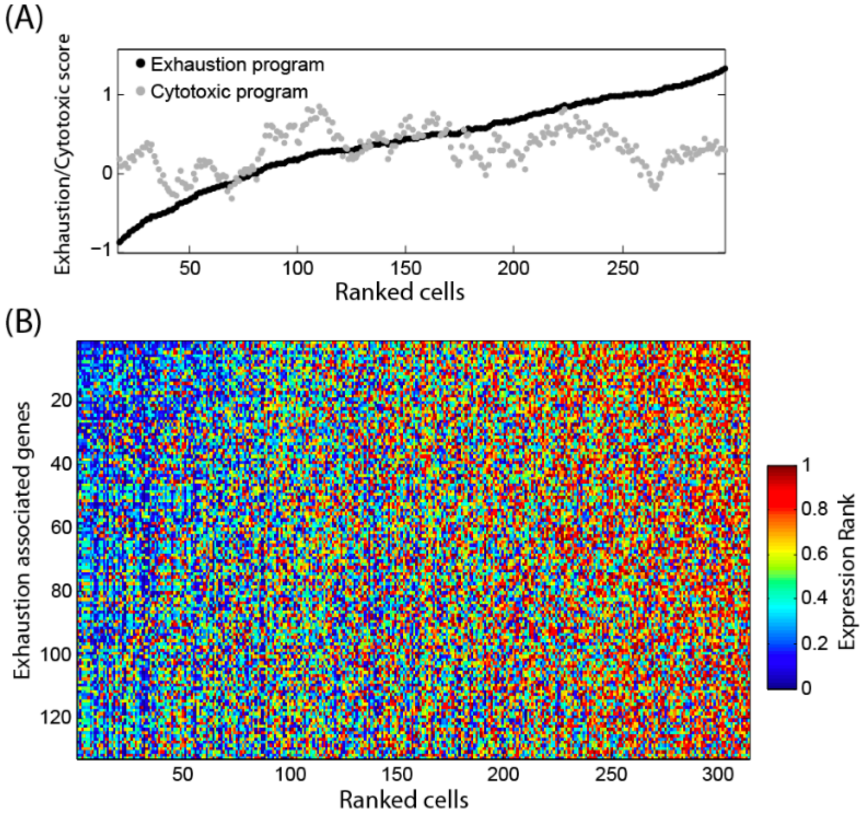


Figure S21 | Exhaustion program in Mel75. PCA of 314 CD8 T-cells from Mel75 identified an exhaustion program in which the top scoring genes for PC1 included the five co-inhibitory receptors shown in Chapter 2, Figure 5B as well as additional exhaustion-associated genes (e.g., BTLA, CBLB). We defined PC1-associated genes based on a correlation p–value of 0.01 (with Bonferroni correction for multiple testing, see Table S13). Cells were then ranked by the residual between average expression of these PC1-associated genes (referred to as the exhaustion program) and average expression of the cytotoxic genes shown in Chapter 2, Figure 5B (referred to as the cytotoxic program) using a LOWESS regression, as shown in Chapter 2, Figure 5D. Finally, for each gene, we ranked its expression levels across the CD8 T-cells from Mel75 and converted these to rank scores between 0 and 1 such that the i highest-expressing cell received a rank score of i/314, where 314 represents the number of CD8 T cells from Mel75. **(A)** Exhaustion and cytotoxic program scores for ranked Mel75 CD8 T-cells, after applying a moving average with windows of 31 genes. **(B)** The heatmap shows expression ranks of PC1-associated genes across the CD8 T- cells from Mel75 cells, ranked as described above.

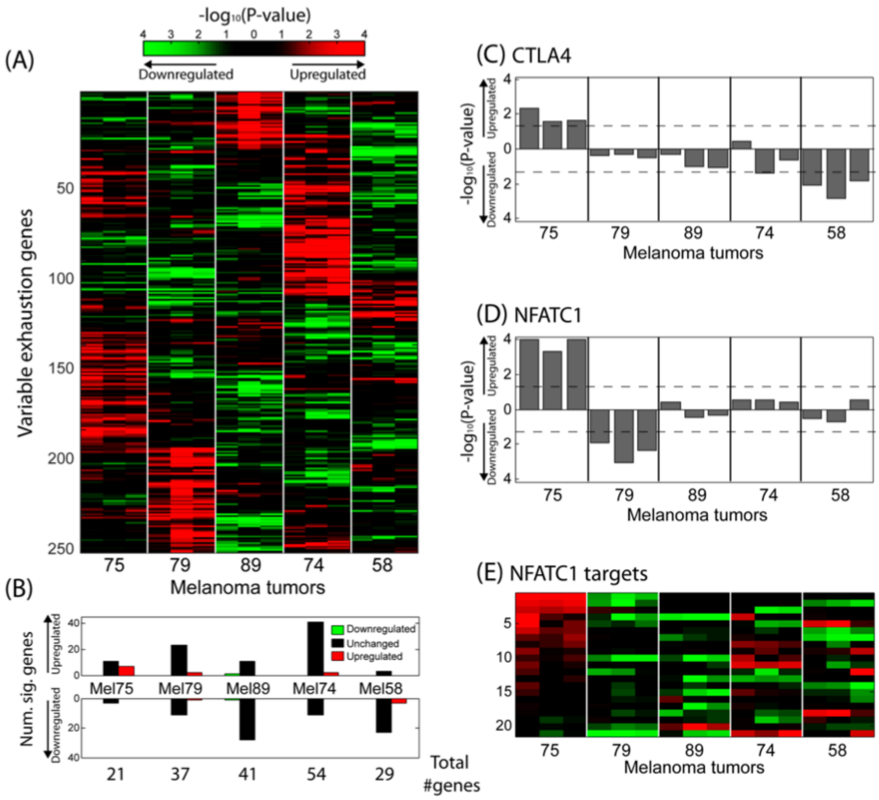

(A)
-log10(P-value)
4 3 2 1 0 1 2 3 4
Downregulated
Upregulated
Variable exhaustion genes
50
100
150
200
250
75 79 89 74 58
Melanoma tumors
(B)
Num. sig. genes
Upregulated
Downregulated
40
20
0
Downregulated
Unchanged
Upregulated
Mel75 Mel79 Mel89 Mel74 Mel58
21 37 41 54 29
Total #genes
(C) CTLA4
-log10(P-value)
Upregulated
Downregulated
4
2
0
2
4
75 79 89 74 58
Melanoma tumors
(D) NFATC1
-log10(P-value)
Upregulated
Downregulated
75 79 89 74 58
Melanoma tumors
(E) NFATC1 targets
5
10
15
20
75 79 89 74 58
Melanoma tumors

Figure S22 | Tumor-specific exhaustion programs. (A) Heatmap shows the significance (–log10(P-value)) of tumor-specific variation in exhaustion gene scores (log-ratio in high vs. low exhaustion cells) comparing each tumor to all other tumors combined, for the same genes (and the same order) as shown in Figure 5F. The sign of significance values reflects the direction of change (positive values shown in red reflect higher exhaustion values compared to other tumors while negative values shown in green reflect lower exhaustion values compared to other tumors). Three values are shown for each tumor, corresponding to exhaustion scores based on the exhaustion gene-sets derived from Mel75 analysis (Figure S22), from Wherry et al., and from Baitsch et al. respectively. **(B)** Tumor-specific associations with the exhaustion program, detected by co- expression across single cells, are not detected by the overall (bulk) tumor-specific expression in CD8 T-cells. Genes with significant tumor-specific up- or down-regulation in high-exhaustion cells (FDR < 0.05 in each tumor, based on median of the three exhaustion scores), were divided to three classes (bars) based on the differences in their overall expression level in CD8 T-cells among the different tumors (green: genes lower in the respective tumor by at least two fold. Red: genes higher in the respective tumor by at least two-fold. Black: genes with less than two-fold difference). This demonstrates that most differences across tumors in exhaustion co-expression are in genes whose overall expression is similar in the different tumors and thus their distinct association with co-expression could not have been identified in bulk level analysis of the CD8 T-cells. **(C–D)** Bar plots showing the significance of tumor-specific variation, as in (A), for CTLA4 (C) and NFATC1 (D). Dashed lines indicate significance thresholds that correspond to P < 0.05. **(E)** Heatmap (as in Chapter 2, subfigure A) for the target genes of NFATC1.

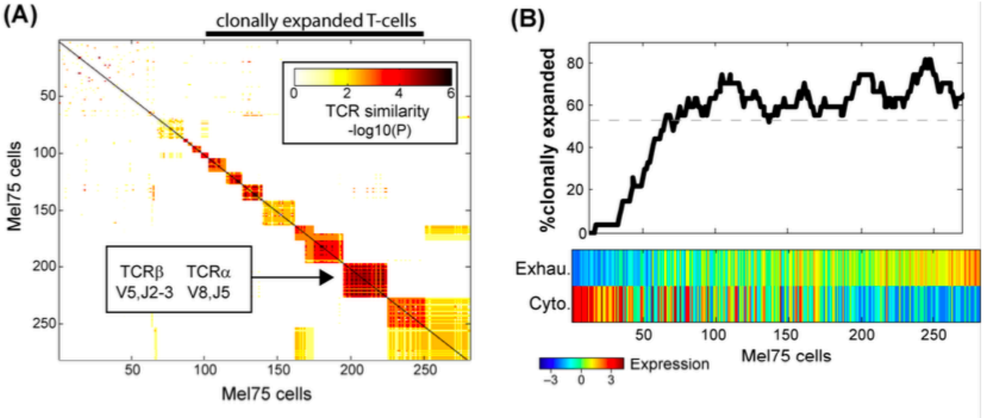


Figure S23 | Detection of Mel74 expanded T-cell clones by TCR sequence. (A) Clustering of Mel75 cells by their TCR segment usage. TCR Similarity was defined as zero for any pair with at least one inconsistent allele (i.e., resolved in both cells but distinct among the two cells), and as –log10(P) for any pair without inconsistent alleles, where P reflects the estimated probability of randomly observing this or a higher degree of segment usage similarity. P is equal to the product of the probabilities for the four TCR segments, P(i,j)=Pβv(i,j)*Pβj(i,j)*Pαv(i,j)*Pαj(i,j). For each segment, the probability equals one if segment usage is unresolved in at least one of the cells of the pair, and otherwise (i.e., if the two cells have the same allele) the probability is 1/N, where N is the number of distinct alleles that were identified for that segment. The TCR usage of one exemplary cluster is indicated. **(B)** Mel75 cells were ordered by the average relative expression of Exhaustion and Cytotoxic genes, as shown in Chapter 2, Figure 5B, and the percentage of clonally expanded cells (i.e., belonging to the clusters indicated in A) is shown with a moving average of 20 cells, demonstrating the depletion of expanded T cells among cells with high cytotoxic and low exhaustion expression. Dashed line indicates the overall frequency of clonally expanded cells. Note that the top and bottom panels are aligned but that due to the use of a 20-cell moving average, the top panel can only start at the 11th cell and end at the 11th cell from the end.

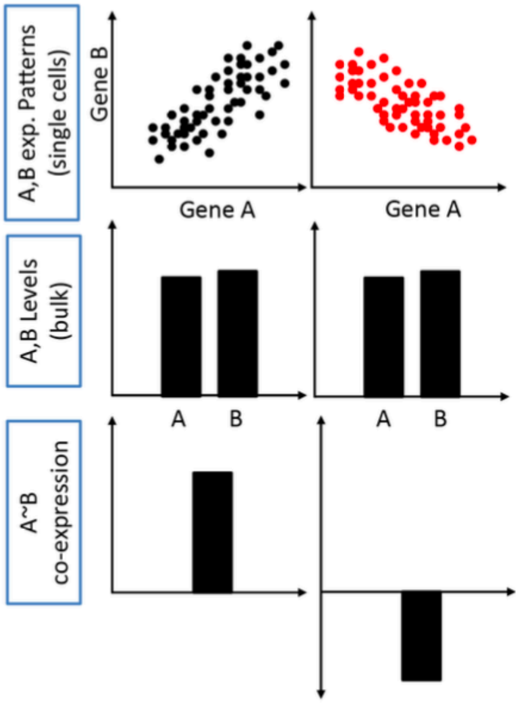


Figure S24 | Identification of distinct co-expression programs may require single cell analysis. Schematic depicting how single-cell RNA-seq can distinguish two scenarios that are indistinguishable by bulk profiling. Across individual tumor cells (**top**), genes A and B are either positively (**left**) or negatively (**right**) correlated. In bulk tumor (**middle**), the average expression of A,B cannot distinguish the two scenarios, whereas co-expression estimates from single cell RNA-seq (**bottom**) do so.

Table S1. Characteristics of patients and samples included in this study

Sample ID	Age/sex	Mutation status	Pre-operative treatment	Site of resection	Post-op. treatment	Alive/ deceased
Melanoma_53	77/F	Wild-type	None	Subcutaneous back lesion	None	Alive
Melanoma_58	67/F	Wild-type	Ipilimumab	Subcutaneous leg lesion	None	Alive
Melanoma_59	80/M	Wild-type	None	Femoral lymph node	Nivolumab.	Deceased
Melanoma_60	69/M	BRAF V600K	Trametinib, ipilimumab	Spleen	None	Alive
Melanoma_65	65/M	BRAF V600E	None	Paraspinal intramuscular	Neovax	Alive
Melanoma_67	58/M	BRAF V600E	None	Axillary lymph node	None	Alive
Melanoma_71	79/M	NRAS Q61L	None	Transverse colon	None	Alive
Melanoma_72	57/F	NRAS Q61R	IL-2, nivolumab, ipilimumab + anti-KIR-Ab	External iliac lymph node	None	Alive
Melanoma_74	63/M	n/a	Nivolumab	Terminal Ileum	None	Alive
Melanoma_75	80/M	Wild-type	Ipilimumab + nivolumab, WDVAX	Subcutaneous leg lesion	Nivolumab	Alive
Melanoma_78	73/M	NRAS Q61L	WDVAX, ipilimumab + nivolumab	Small bowel	None	Deceased
Melanoma_79	74/M	Wild-type	None	Axillary lymph node	None	Alive
Melanoma_80	86/F	NRAS Q61L	None	Axillary lymph node	None	Alive
Melanoma_81	43/F	BRAF V600E	None	Axillary lymph node	None	Alive
Melanoma_82	81/M	Wild-type	None	Axillary lymph node	None	Alive
Melanoma_84	67/M	Wild-type	None	Acral primary	None	Alive
Melanoma_88	54/F	NRAS Q61L	Tremelimumab + MEDI3617	Cutanoues met	None	Alive
Melanoma_89	67/M	n/a	None	Axillary lymph node	None	Alive
Melanoma_94	54/F	Wild-type	IFN, ipilimumab + nivolumab	Iliac lymph node	None	Alive

Table S2. Number of cells classified to each cell type in each tumor.

Table S3. Cell type specific genes.

Table S4. PCA. The top 50 correlated genes and the top MsigDB enrichments of those genes for the first six PCs.

Table S5. Core signature of cell cycle genes expressed in cycling malignant cells from both low-cycling and high-cycling tumors.

Table S6. Differentially regulated genes in Region 1.

Table S7. List of genes included in the MITF-program.

Table S8. List of genes included in the AXL-program.

Table S9. Expression data for pre-treatment and post-relapse samples.

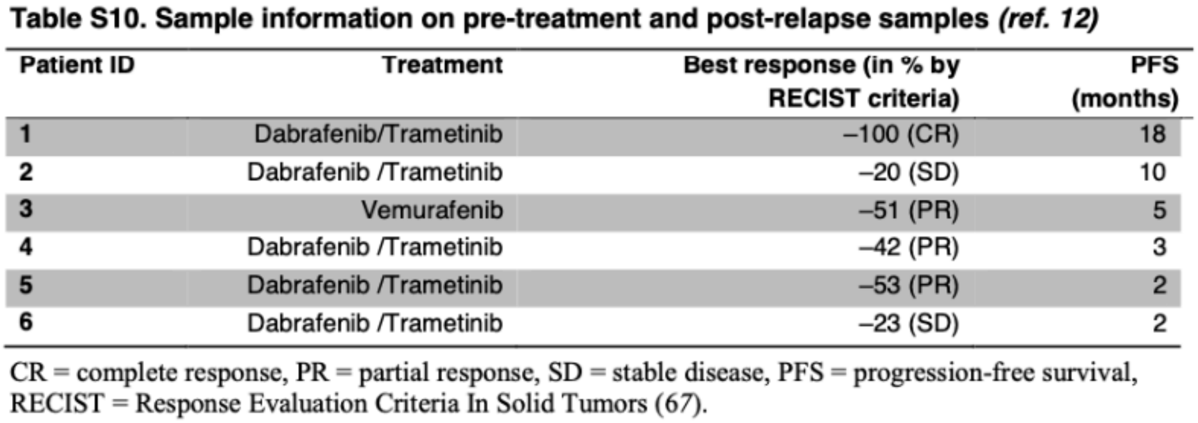


Table S10. Sample information on pre-treatment and post-relapse samples *(ref. 12)*

Patient ID	Treatment	Best response (in % by RECIST criteria)	PFS (months)
1	Dabrafenib/Trametinib	–100 (CR)	18
2	Dabrafenib /Trametinib	–20 (SD)	10
3	Vemurafenib	–51 (PR)	5
4	Dabrafenib /Trametinib	–42 (PR)	3
5	Dabrafenib /Trametinib	–53 (PR)	2
6	Dabrafenib /Trametinib	–23 (SD)	2

CR = complete response, PR = partial response, SD = stable disease, PFS = progression-free survival, RECIST = Response Evaluation Criteria In Solid Tumors (*67*).

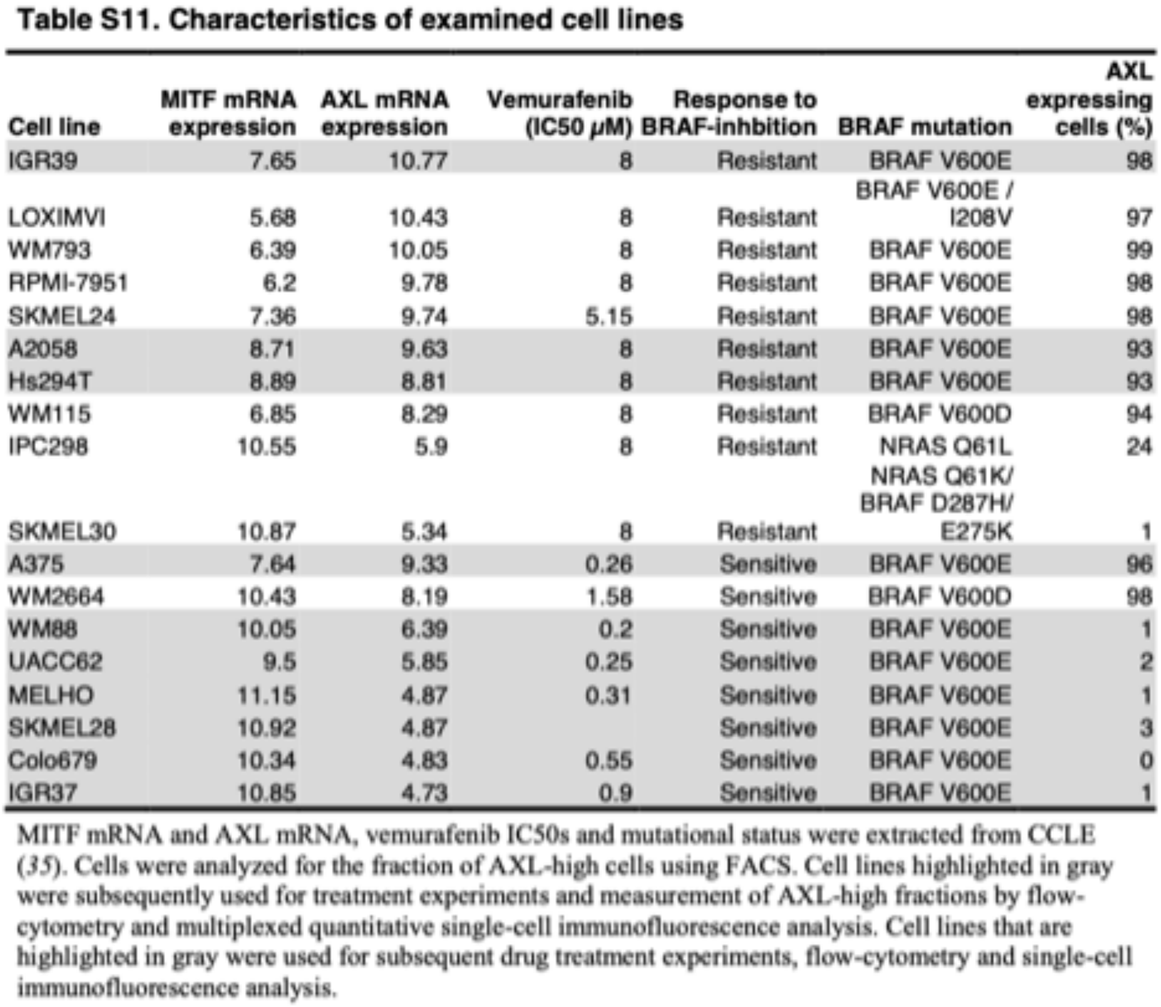


Table S11. Characteristics of examined cell lines

Cell line	MITF mRNA expression	AXL mRNA expression	Vemurafenib (IC50 μM)	Response to BRAF-inhbition	BRAF mutation	AXL expressing cells (%)
IGR39	7.65	10.77	8	Resistant	BRAF V600E	98
LOXIMVI	5.68	10.43	8	Resistant	BRAF V600E / I208V	97
WM793	6.39	10.05	8	Resistant	BRAF V600E	99
RPMI-7951	6.2	9.78	8	Resistant	BRAF V600E	98
SKMEL24	7.36	9.74	5.15	Resistant	BRAF V600E	98
A2058	8.71	9.63	8	Resistant	BRAF V600E	93
Hs294T	8.89	8.81	8	Resistant	BRAF V600E	93
WM115	6.85	8.29	8	Resistant	BRAF V600D	94
IPC298	10.55	5.9	8	Resistant	NRAS Q61L	24
SKMEL30	10.87	5.34	8	Resistant	NRAS Q61K/ BRAF D287H/ E275K	1
A375	7.64	9.33	0.26	Sensitive	BRAF V600E	96
WM2664	10.43	8.19	1.58	Sensitive	BRAF V600D	98
WM88	10.05	6.39	0.2	Sensitive	BRAF V600E	1
UACC62	9.5	5.85	0.25	Sensitive	BRAF V600E	2
MELHO	11.15	4.87	0.31	Sensitive	BRAF V600E	1
SKMEL28	10.92	4.87		Sensitive	BRAF V600E	3
Colo679	10.34	4.83	0.55	Sensitive	BRAF V600E	0
IGR37	10.85	4.73	0.9	Sensitive	BRAF V600E	1

MITF mRNA and AXL mRNA, vemurafenib IC50s and mutational status were extracted from CCLE (*35*). Cells were analyzed for the fraction of AXL-high cells using FACS. Cell lines highlighted in gray were subsequently used for treatment experiments and measurement of AXL-high fractions by flow-cytometry and multiplexed quantitative single-cell immunofluorescence analysis. Cell lines that are highlighted in gray were used for subsequent drug treatment experiments, flow-cytometry and single-cell immunofluorescence analysis.

Table S12. List of genes preferentially expressed in Tregs.

Table S13. Genes associated with Mel75 exhaustion signature.

Table S14. Association of genes with exhaustion signature across five tumors.

Table S15. CAF-derived genes that correlate with abundance of T-cells.

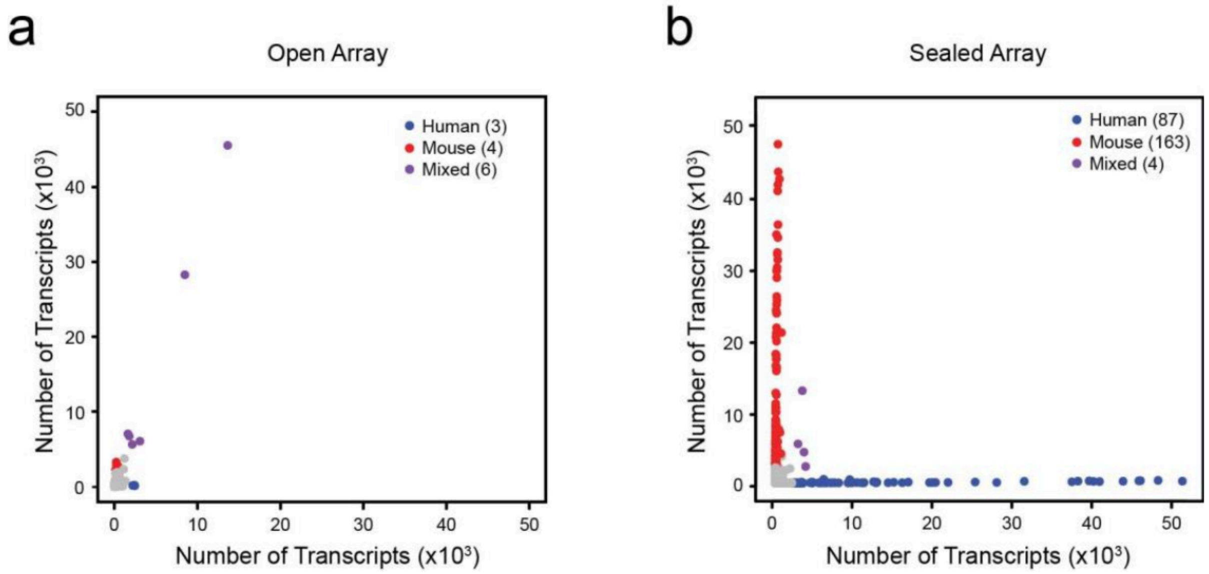
Table S16. Curated list of housekeeping genes used for QC.

Appendix B - Supplemental Information for Chapter 3: Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput

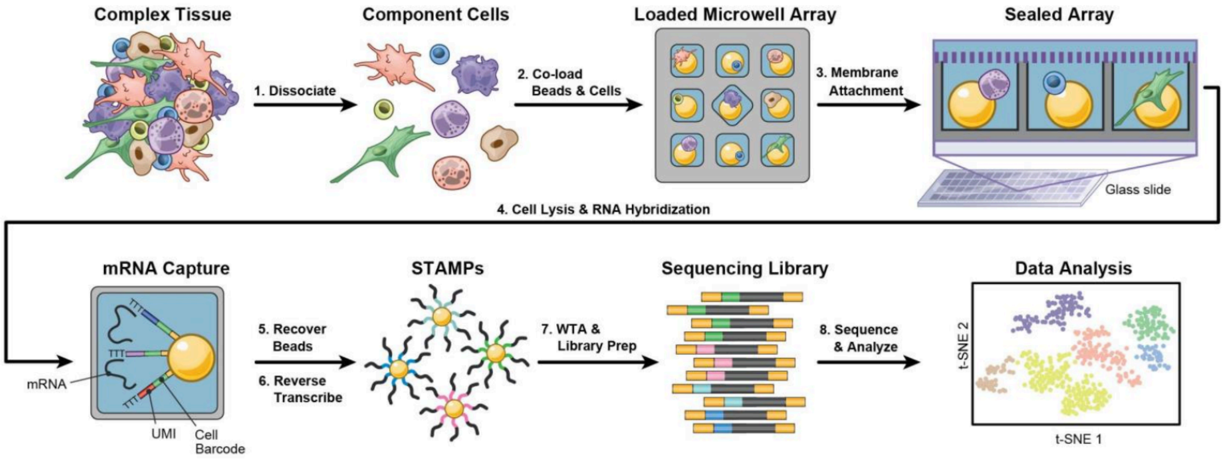
This chapter is adapted in accordance with Nature's open access policy from the following article published in Nature Methods:

Gierahn, T.M.*, Wadsworth II, M.H.*, Hughes, T.K., Bryson, B.D., Butler, A., Satija, R., Fortune, S., Love, J.C., and Shalek, A.K. "Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput," *Nature Methods*, 14, 395-398 (2017).

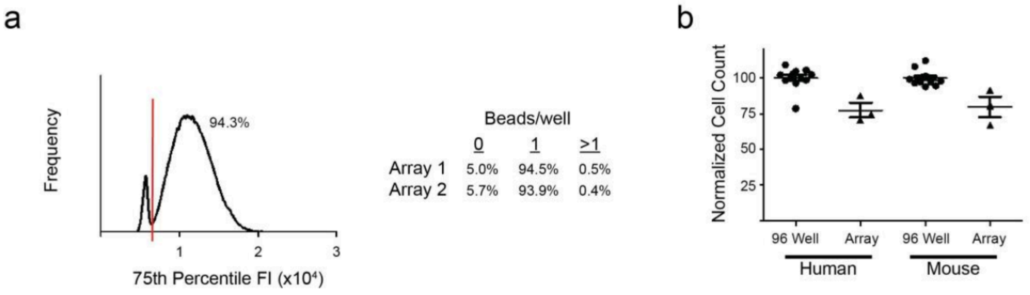
** Denotes equal authorship*



Supplementary Figure 1 | Open Array Gene and Transcript Capture. **(a)** An open array format results in decreased gene and transcript capture, and increased cross-contamination, relative to the membrane sealing implemented inSeq-Well. **(b)** Species mixing experiments with reversible membrane sealing using Seq-Well provides increased gene/transcript capture and improved single-cell resolution.



Supplementary Figure 2 | Seq-Well Experimental Workflow. Cells are obtained from complex tissues or clinical biopsies and digested to form a single-cell suspension. Barcoded mRNA capture beads are added to the surface of the microwell device, settling into wells by gravity, and then a single-cell suspension is applied. The device is sealed using a semi-permeable membrane that, upon addition of a chemical lysis buffer, confines cellular mRNAs within wells while allowing efficient buffer exchange. Liberated cellular transcripts hybridize to the bead-bound barcoded poly(dT) primers that contain a cell barcode (shared by all probes on the same bead but different between beads) and a unique molecular identifier (UMI) for each transcript molecule. After hybridization, the beads are removed from the array and bulk reverse transcription is performed to generate single-cell cDNAs attached to beads. Libraries are then made by a combination of PCR and tagmentation, and sequenced. After, single-cell transcriptomes are assembled *in silico* using cell barcodes and UMIs.



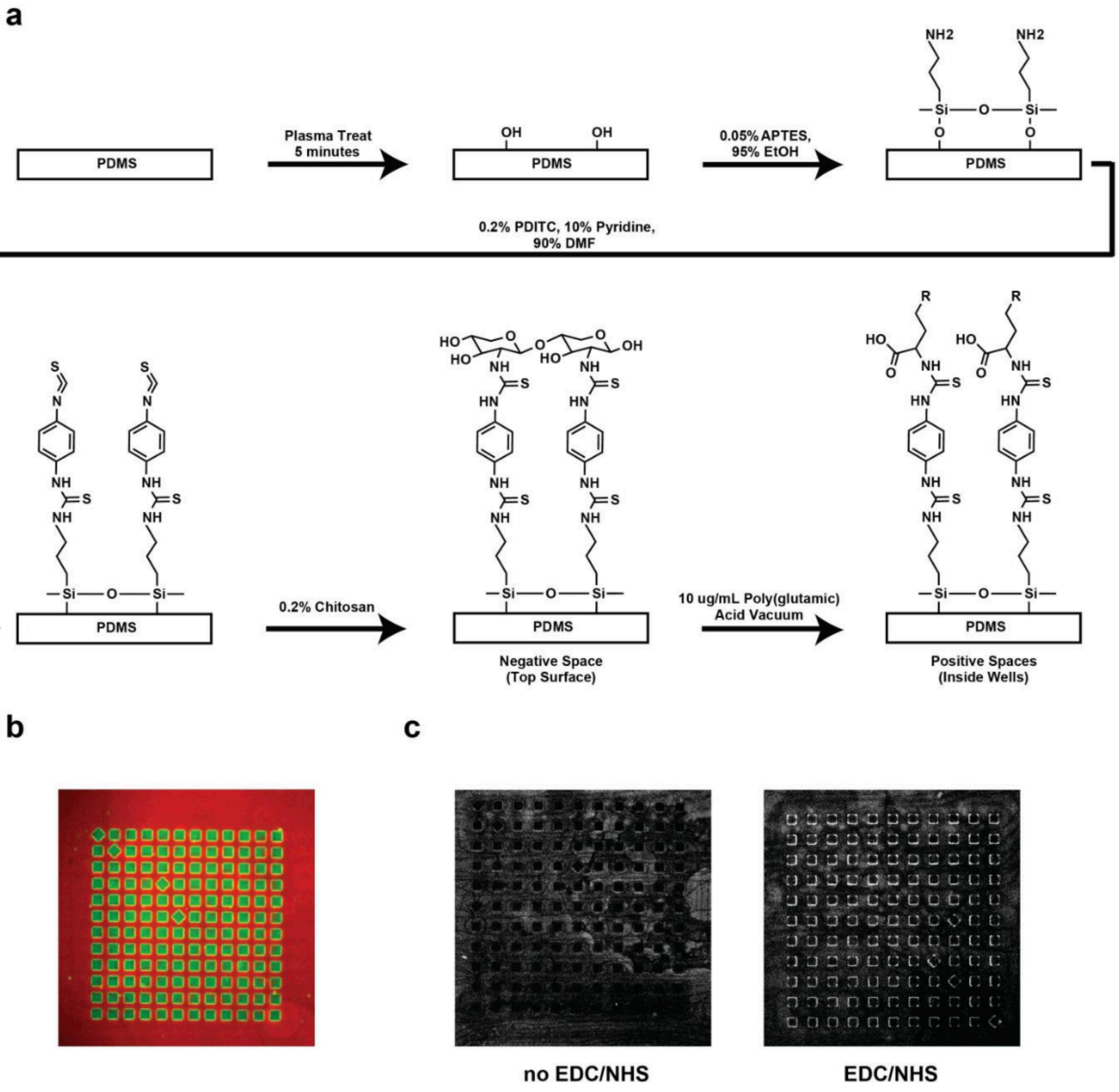
Beads/well

	0	1	>1
Array 1	5.0%	94.5%	0.5%
Array 2	5.7%	93.9%	0.4%

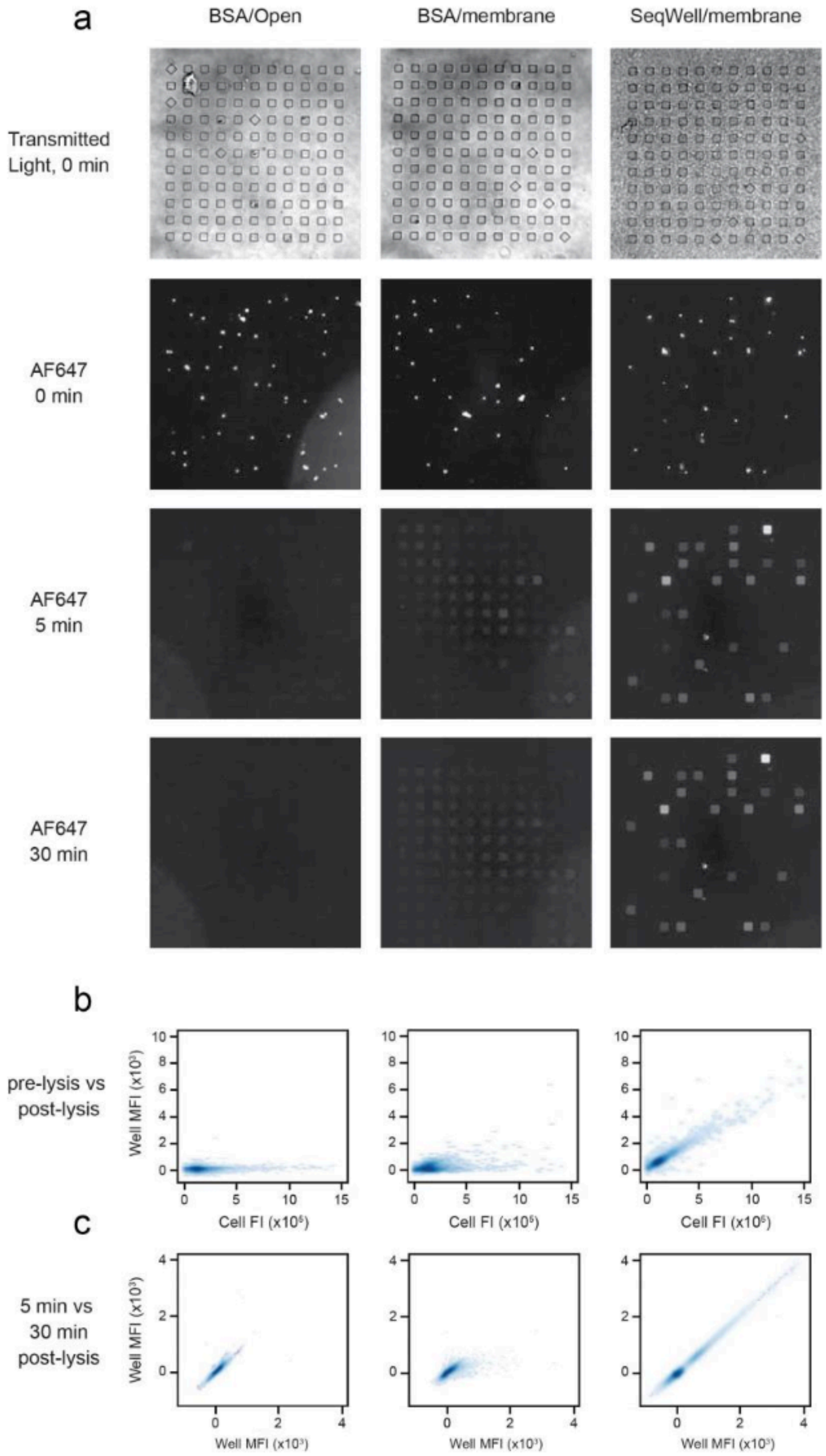
c

Loading Density	Empty	1 Human	1 Mouse	2+ Human	2+ Mouse	Human & Mouse Multiplet	Total Multiplet	Human & Mouse Multiplet Freq. (%)	Total Multiplet Freq. (%)
200	83,200 ± 293	134 ± 23	177 ± 53	0.33 ± 0.6	10.3 ± 17.9	0.7 ± 0.6	11.3 ± 19	0.1 ± 0.2	2.7 ± 4.4
2,000	81,100 ± 320	1,310 ± 150	1,080 ± 170	9.7 ± 3.2	23.3 ± 13	10.3 ± 3.1	43.3 ± 11	0.4 ± 0.1	1.8 ± 0.5
20,000	60,700 ± 560	13,100 ± 340	9,190 ± 250	976 ± 100	598 ± 23	1,100 ± 160	2,603 ± 250	4.8 ± 0.6	11.4 ± 0.9

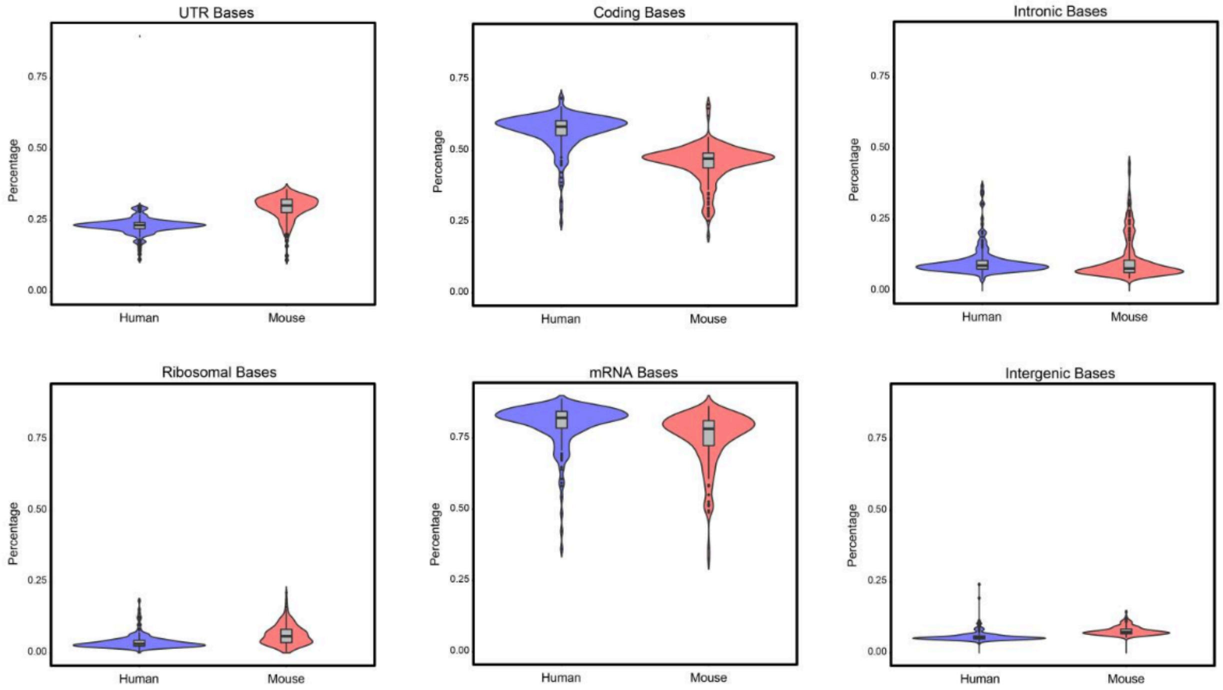
Supplementary Figure 3 | Bead and Cell Loading Efficiency. (a) Two arrays were loaded with barcoded beads through intermittent rocking. After washing, arrays were imaged in transmitted light and AF488 channel to capture bead autofluorescence. A plot of the frequency of the 75th percentile AF488 well intensity across the array (Panel 1) and the frequency of wells containing zero, one and multiple beads is displayed (Panel 2). **(b)** 200 L of a 1:1 mix of fluorescently labeled human (HEK 293) and mouse (3T3) cell solution was loaded into 3 arrays and 12 wells of a 96 well plate. The number of cells loaded into each array and well as enumerated by fluorescent imaging is plotted, normalized to the average number of cells/well in the 96 well plate. Mean and standard error are denoted by line and error bars respectively. **(c)** 2x102, 2x103, and 2x104 total cells of a 1:1 mixture of fluorescently labeled HEK 293T and 3T3 cells were loaded onto three functionalized arrays each. All arrays were fluorescently imaged to enumerate the number of each cell line in each array microwell. The mean ± standard deviation of the number of empty, single and multiple occupancy wells across the three replicate arrays for each loading density is displayed along with the mean ± standard deviation of the percentage of occupied wells containing a cell from each species.


a
PDMS
Plasma Treat
5 minutes
OH
OH
PDMS
0.05% APTES,
95% EtOH
NH2
NH2
Si
O
Si
O
O
PDMS
0.2% PDITC, 10% Pyridine,
90% DMF
S
N
NH
S
NH
Si
O
Si
PDMS
0.2% Chitosan
HO
HO
O
O
HO
OH
NH
S
HN
NH
S
NH
PDMS
Negative Space
(Top Surface)
10 ug/mL Poly(glutamic)
Acid Vacuum
R
HO
O
NH
S
HN
NH
S
NH
PDMS
Positive Spaces
(Inside Wells)
b
c
no EDC/NHS
EDC/NHS

Supplementary Figure 4 | PDMS Surface Chemistry Functionalization Protocol and Differential Functionalization of Microwell Arrays. (a) The surface of the PDMS device is initially treated with an air plasma under mild vacuum, terminating the surface in hydroxyls. This PDMS surface is aminated using (3-Aminopropyl)triethoxysilane (APTES). The amine surface is then activated with PDITC to create an isothiocyanate surface. The isothiocyanate on the top surface of the array (negative space) is covalently linked to chitosan polymers through their amine group. The hydrophobicity of the isothiocyanate surface prevents solvation of the microwells with the aqueous chitosan solution, preventing chitosan from reacting with the inner well surfaces (positive space). These surfaces are subsequently reacted with the free amine of poly(glutamic) acid polymers under vacuum to drive the solvation of the wells. **(b)** The top surface of a PDITC-activated array was coated with streptavidin-PE (red) and the inner well surfaces were coated with streptavidin-AF488 (green) using same method used to functionalize with chitosan and poly(glutamate). **(c)** Two chitosan/poly(glutamate) bifunctionalized arrays were submerged in MES buffer without (Panel 1) or with (Panel 2) 100 µg/mL EDC and 10 µg/mL NHS for 10 minutes. The arrays were washed and then submerged in PBS solution containing 1 µg/mL AF568-labeled antibody overnight. After washing, arrays were imaged for AF568 fluorescence.

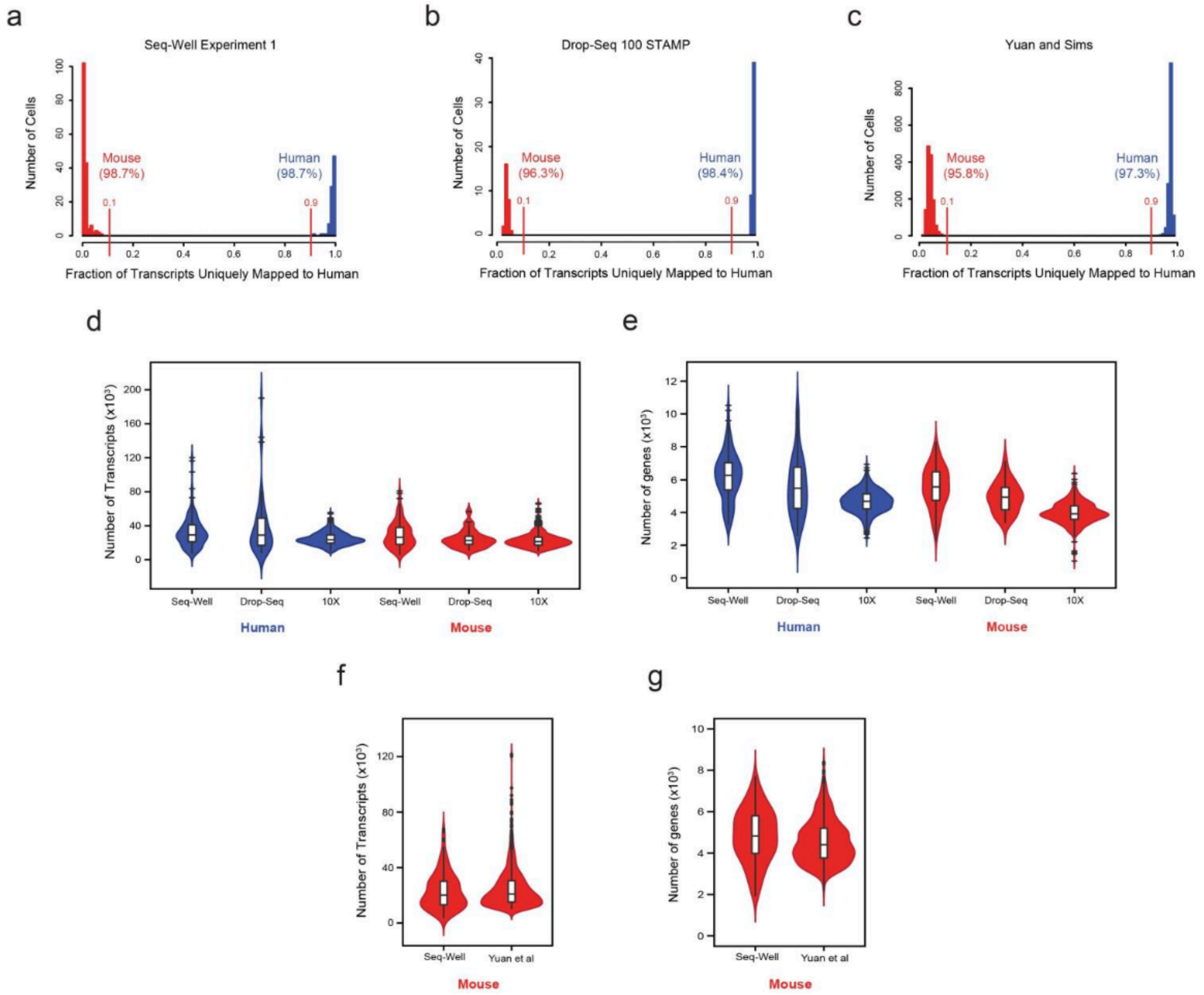

a
BSA/Open
BSA/membrane
SeqWell/membrane
Transmitted Light, 0 min
AF647 0 min
AF647 5 min
AF647 30 min
b
pre-lysis vs post-lysis
Well MFI (x10³)
0
2
4
6
8
10
Cell FI (x10⁶)
0
5
10
15
c
5 min vs 30 min post-lysis
Well MFI (x10³)
0
2
4
Well MFI (x10³)

Supplementary Figure 5 | Microwell Sealing With Semipermeable Membrane. PBMCs labeled with CD45-AF647 were loaded into two BSA-blocked arrays and one array functionalized with chitosan and poly(glutamate). A semipermeable membrane was attached to one of the BSA-blocked arrays and the chitosan:polyglutamate functionalized array prior to addition of lysis buffer. **(a)** Example images of transmitted light and AF647 fluorescence of the arrays before, and 5 and 30 minutes after addition of lysis buffer are displayed for each array. **(b)** The total fluorescence intensity (FI) of all pixels associated with cells within a well is plotted against the median fluorescent intensity (MFI) of the volume of the same well 5 minutes after lysis for 12,100 wells from each array. **(c)** The MFI of the well volume 5 minutes after lysis is plotted against the MFI of the volume of the same well 30 minutes after lysis for the same 12,100 wells from each array.

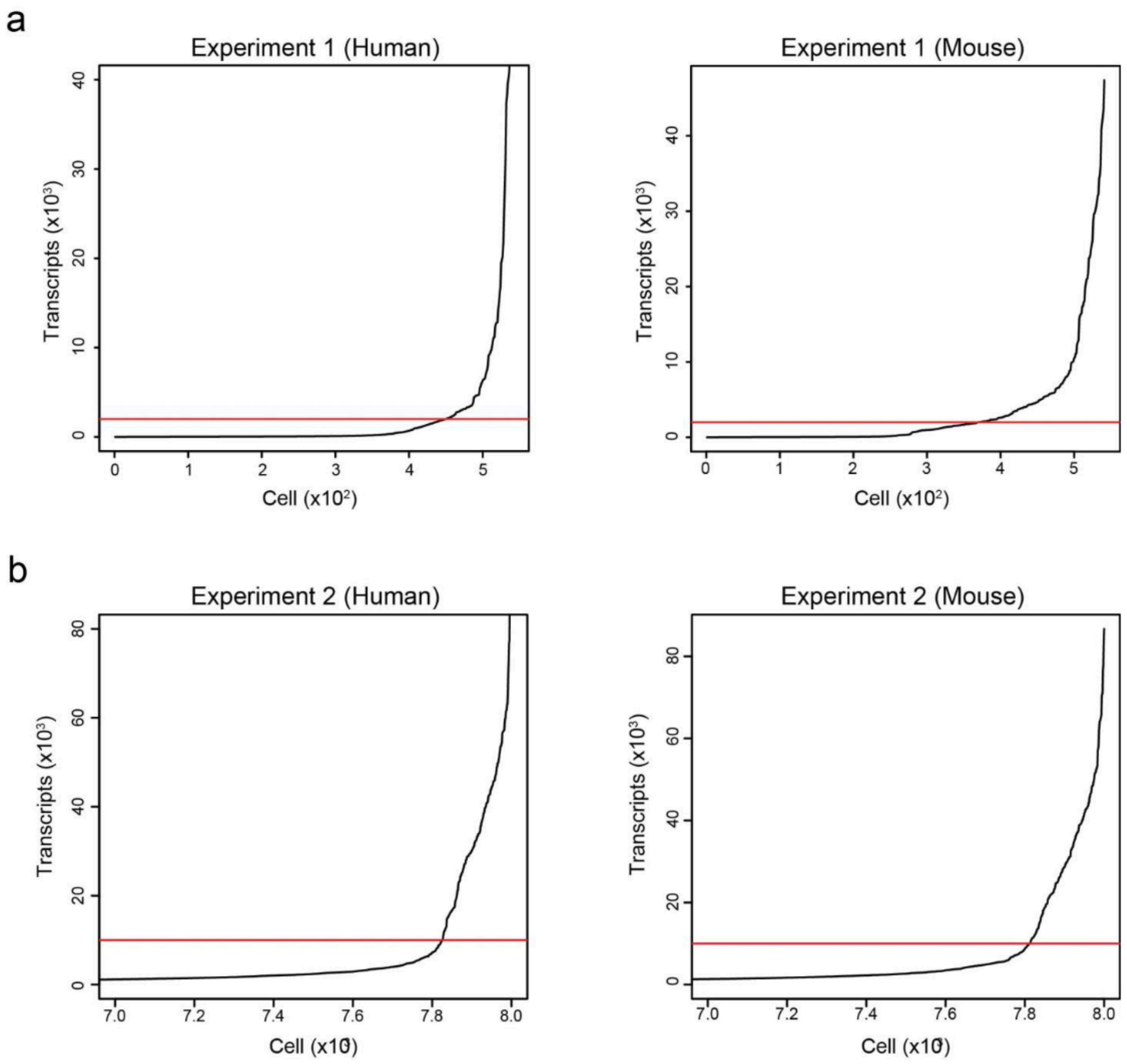


Species	Ribosomal Bases	mRNA Bases	Intergenic Bases	Intronic Bases	Coding Bases	UTR Bases
Human	3.86%	80.0%	5.79%	10.4%	56.9%	23.1%
Mouse	6.32%	75.2%	7.68%	10.9%	45.6%	29.6%

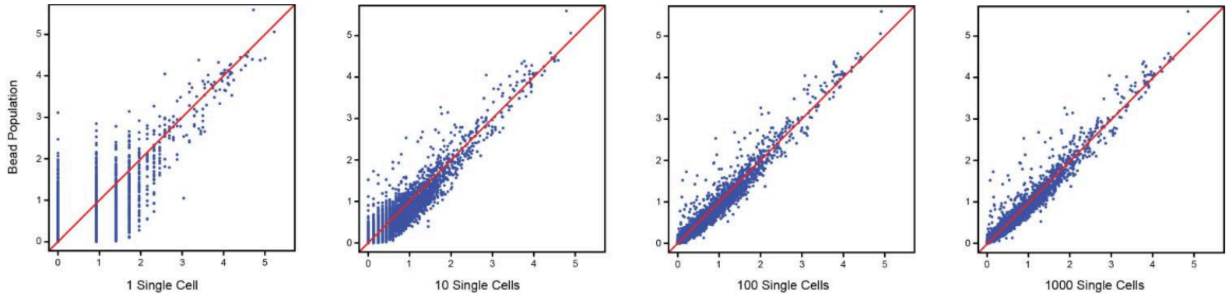
Supplementary Figure 6 | Read Mapping Quality. Read mapping quality matrices were generated for each sample for human (blue) and mouse (red) cells, aligned to hg19 and mm10, respectively. High quality samples had relatively higher percentages of annotated genomic (genic) and exonic transcripts and low percentages of annotated intergenic and ribosomal transcripts (Center-line: Median; Limits: 1st and 3rd Quartile; Whiskers: +/-1.5 IQR; Points: Values > 1.5 IQR).


a
Seq-Well Experiment 1
Number of Cells
100
80
60
40
20
0
Mouse
(98.7%)
Human
(98.7%)
0.1
0.9
0.0
0.2
0.4
0.6
0.8
1.0
Fraction of Transcripts Uniquely Mapped to Human
b
Drop-Seq 100 STAMP
Number of Cells
40
30
20
10
0
Mouse
(96.3%)
Human
(98.4%)
0.1
0.9
0.0
0.2
0.4
0.6
0.8
1.0
Fraction of Transcripts Uniquely Mapped to Human
c
Yuan and Sims
Number of Cells
800
600
400
200
0
Mouse
(95.8%)
Human
(97.3%)
0.1
0.9
0.0
0.2
0.4
0.6
0.8
1.0
Fraction of Transcripts Uniquely Mapped to Human
d
Number of Transcripts (x10³)
200
160
120
80
40
0
Seq-Well
Drop-Seq
10X
Seq-Well
Drop-Seq
10X
Human
Mouse
e
Number of genes (x10³)
12
10
8
6
4
2
0
Seq-Well
Drop-Seq
10X
Seq-Well
Drop-Seq
10X
Human
Mouse
f
Number of Transcripts (x10³)
120
80
40
0
Seq-Well
Yuan et al
Mouse
g
Number of genes (x10³)
10
8
6
4
2
0
Seq-Well
Yuan et al
Mouse

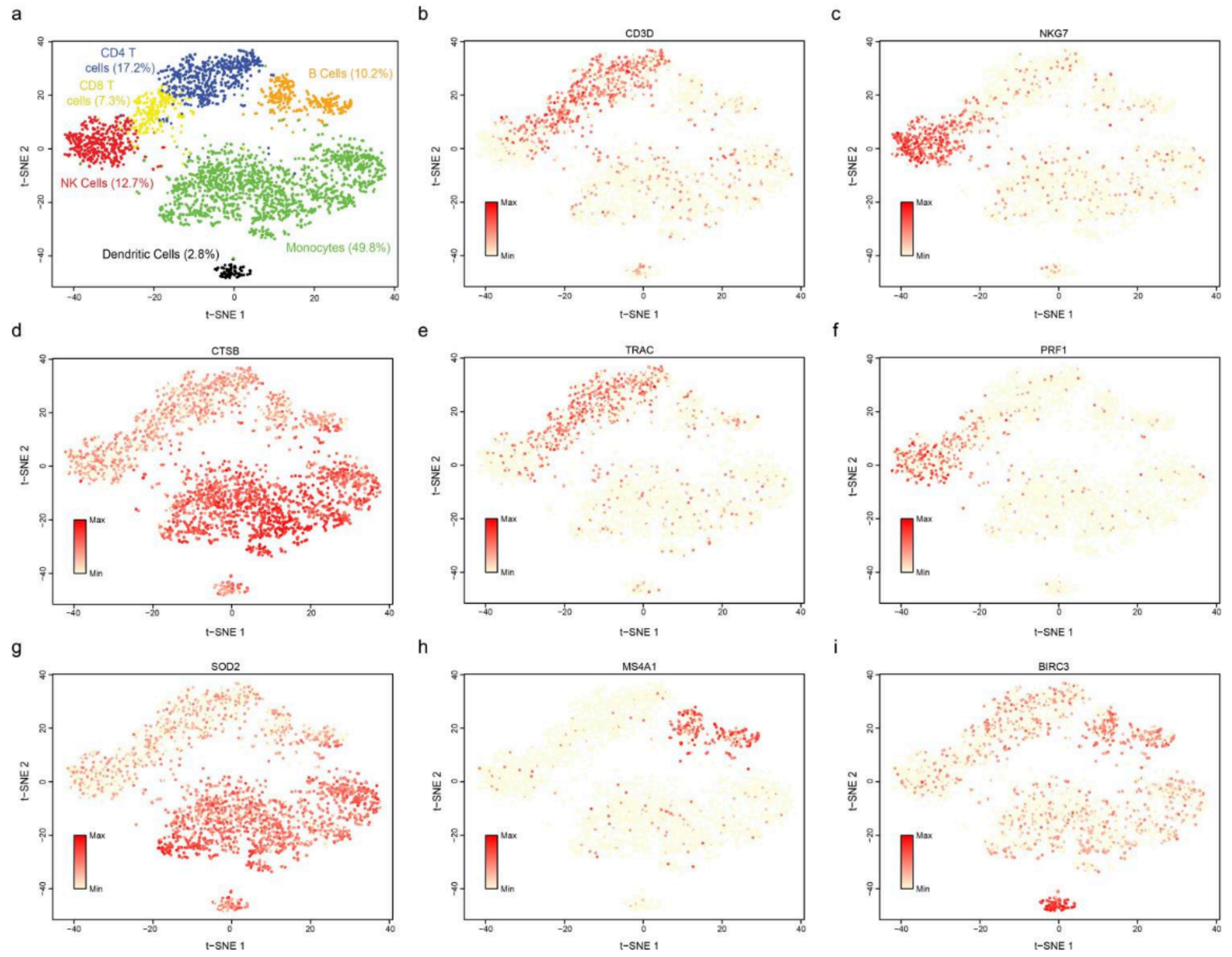
Supplementary Figure 7 | Comparison of Gene and Transcript Capture and Percent Contamination Among Massively-Parallel scRNA-Seq Methods Using Mouse and Human Cell Lines. Histograms of the percent cross-species contamination in **(a)** Seq-Well, **(b)** Drop-Seq, and **(c)** Yuan and Sims. In each plot, cells with greater than 90% of human transcripts are displayed in blue and cells with less than 10% human transcripts are displayed in red. **(d)** Transcript capture in human (blue) and mouse (red) cell lines across three massively-parallel, bead-based single-cell sequencing platforms (Seq-Well, Drop-Seq, and 10x Genomics, with downsampling to an average read-depth of 80,000 reads per cell, consistent with 10x genomics data (Center-line: Median; Limits: 1st and 3rd Quartile; Whiskers: +/-1.5 IQR; Points: Values > 1.5 IQR). We detect an average of 32,841 human transcripts and 29,806 mouse transcripts using Seq-Well compared to an average of 39,400 human transcripts and 24,384 mouse transcripts using Drop-Seq, an average of 24,751 human transcripts and 22,971 mouse transcripts using 10X Genomics (available from http://support.10xgenomics.com/single-cell/datasets/hgmm). **(e)** Gene detection across human and mouse cell lines across the same three single-cell sequencing platforms with down-sampling to the average read-depth of 80,000 reads per cell, consistent with 10x genomics (Center-line: Median; Limits: 1st and 3rd Quartile; Whiskers: +/-1.5 IQR; Points: Values > 1.5 IQR). We detect an average of 6,174 human genes and 5,528 mouse genes using Seq-Well, an average of 5,561 human genes and 4,903 mouse genes using Drop-Seq and an average of 4,655 human genes and 3,950 mouse genes using 10X Genomics. **(f)** Downsampling to an average of 42,000 reads per cell consistent with data published in Yuan and Sims 2016, results in average detection of 23,061 mouse transcripts using Seq-Well compared to an average of 24,761 mouse transcripts using the Yuan and Sims platform (Center-line: Median; Limits: 1st and 3rd Quartile; Whiskers: +/-1.5 IQR; Points: Values > 1.5 IQR). **(g)** Downsampling to an average of 42,000 reads per cell results in average detection of 4,827 mouse genes using Seq-Well compared to an average of 4,569 mouse genes using the Yuan and Sims platform(Center-line: Median; Limits: 1st and 3rd Quartile; Whiskers: +/-1.5 IQR; Points: Values > 1.5 IQR).



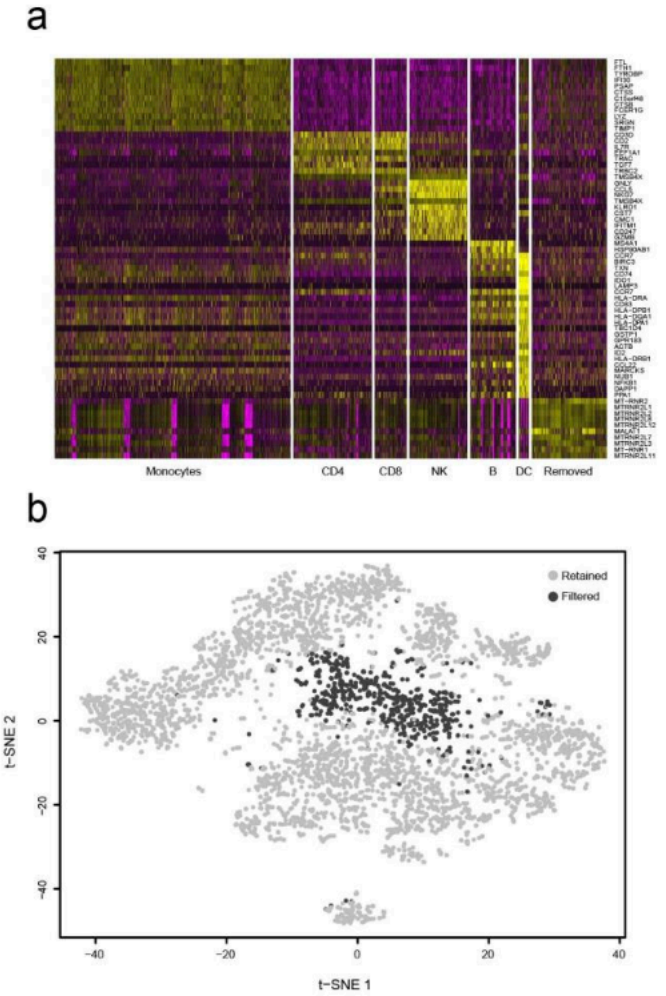
Supplementary Figure 8 | Transcript Cutoff For Species-Mixing Validation. We sequenced two arrays (**a & b**) to confirm single-cell resolution and minimal cross-contamination between mouse and human cells. We called cells by plotting the cumulative distribution of transcripts and making a cutoff at the elbow in the curve. In the first experiment **(a)**, which was used to validate our single-cell resolution, we shallowly sequenced the array and made the cutoff at 2,000 transcripts. In the second experiment **(b)**, where we sequenced the array deeply to allow a competitive comparison to Drop-Seq, we made our cutoff at 10,000 transcripts.



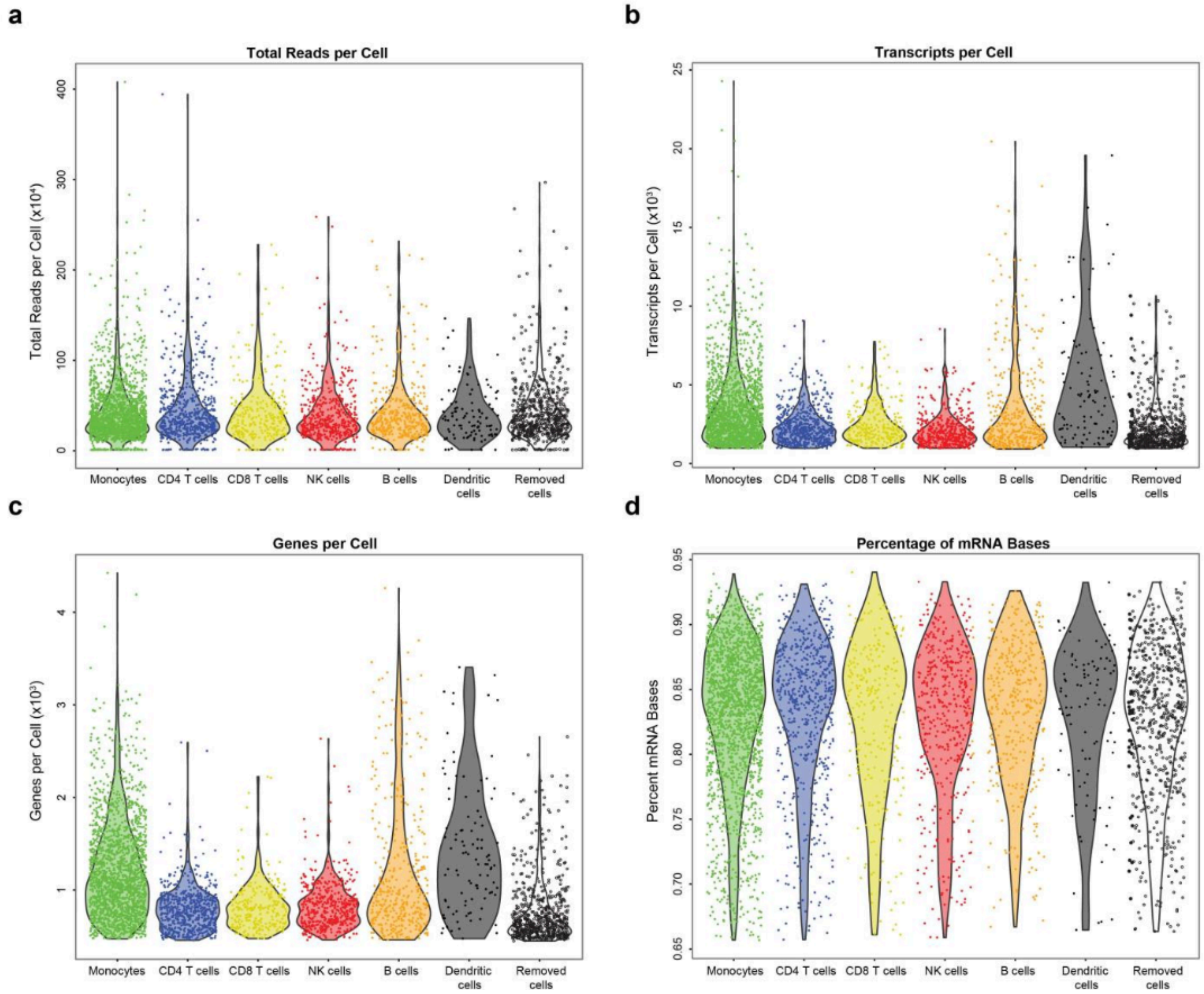
Supplementary Figure 9 | Comparison of In-Silico HEK293 Populations With Bulk Populations. Scatterplots showing the correlation between gene expression estimates from bulk populations (40,000 HEK cells and 40,000 mRNA capture) and populations generated in-silico from 1, 10, 100, and 1,000 randomly-sampled single HEK293 cells (1 Cell: R = 0.751 ± 0.0726; 10 Cells: R = 0.952 ± 0.008; 100 Cells: R = 0.980 ± 0.0006; 1000 Cells: R = 0.983 ± 0.0001).



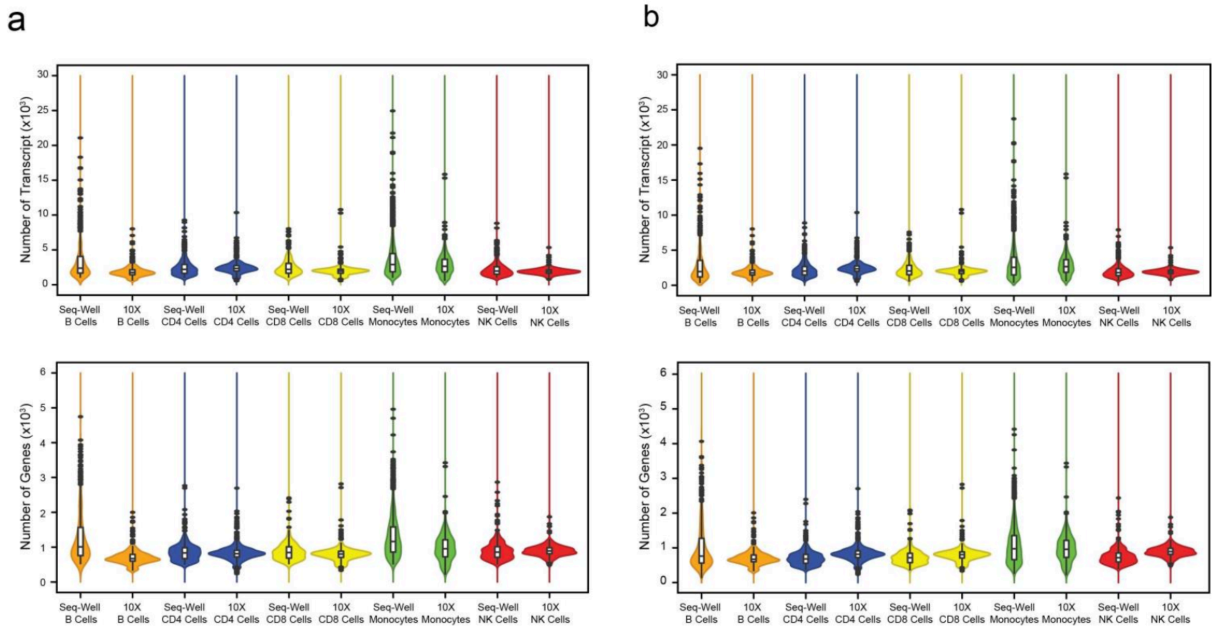
Supplementary Figure 10 | Mapping Lineage Defining Transcripts to PBMC Clusters. (a) Clusters identified through graph-based clustering (Chapter 3, Methods) correspond to major immune cell populations. **(b, e)** CD4 T cells are characterized by expression of CD3D and T-cell receptor expression without pronounced expression of cytoxic genes NKG7 and PRF1. **(c, f)** CD8 T cells are defined by expression of NKG7 and PRF1. **(d, g)** Monocytes are defined by expression of cathepsin B (CTSB) and SOD2. **(e)** Natural killer cells are characterized by expression of cytotoxic genes in the absence of T cell receptor expression. **(h)** B cells are marked by elevated expression of MS4A1 (CD20) transcripts. **(i)** Dendritic cells are enriched for expression of BIRC3.



Supplementary Figure 11 | Heatmap Of PBMCs. (a) Genes enriched in each cluster were identified using an “ROC” test in Seurat, comparing cells assigned to each cluster to all other cells. A heatmap was constructed using enriched genes found to define each cluster. One cluster of 602 cells that demonstrated exclusive enrichment of mitochondrial genes was removed as these likely represent low-quality or dying cells. **(b)** We generated a t-SNE projection of 4,296 cells with greater than 10,000 reads, 1,000 transcripts, 500 genes, and 65% transcript mapping. We removed a total of 602 cells from the final analysis found to be strongly enriched for expression of mitochondrial genes. The remaining 3,694 cells form distinct clusters enriched for lineage-defining that distinguish cells types from one another.



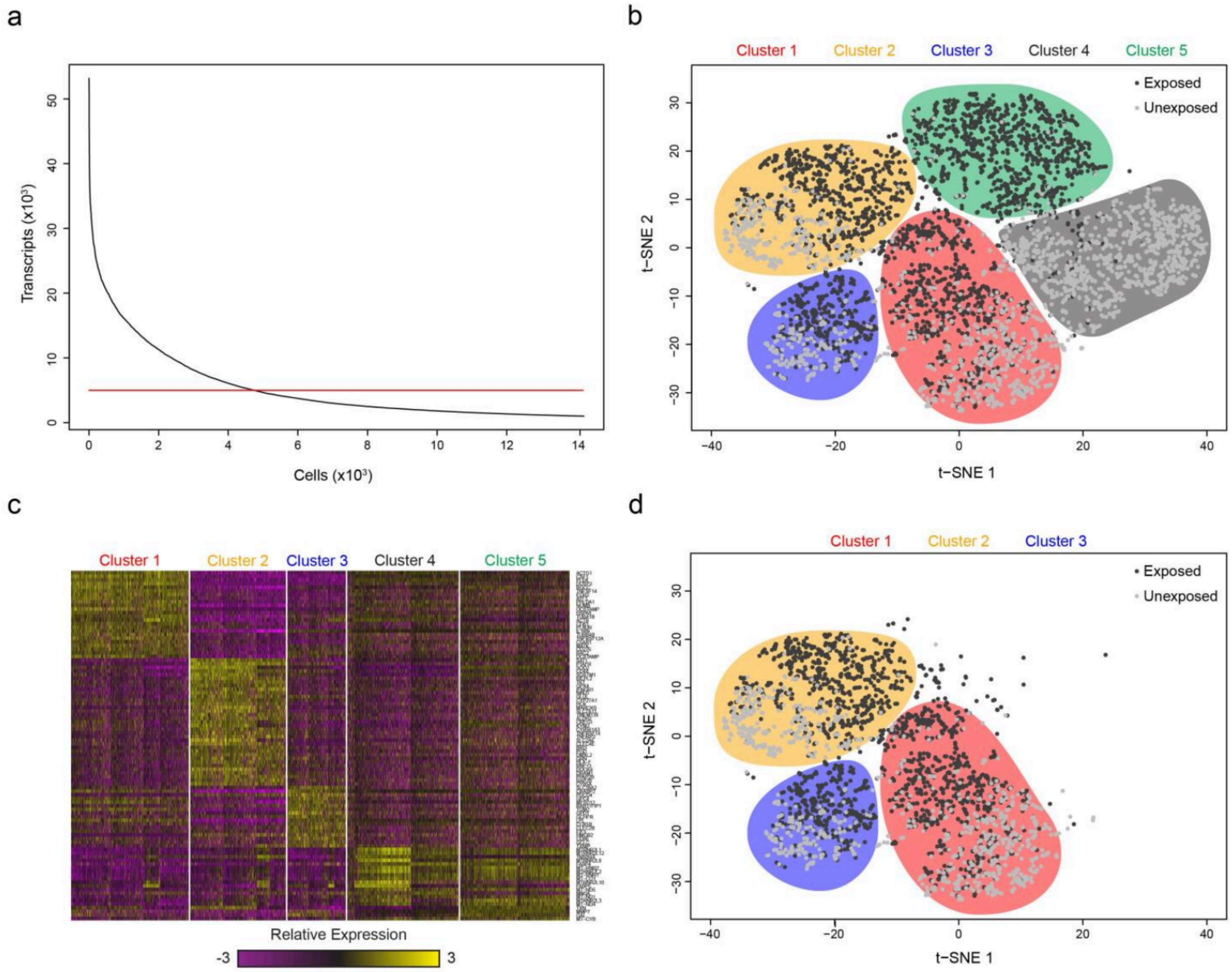
Supplementary Figure 12 | Read Mapping Quality in PBMCs. (a-c) Violin plots depicting **(a)** reads, **(b)** transcripts, and **(c)** genes per cell, separated by cell type. **(d)** Percent mRNA bases per cell, separated by cell type.



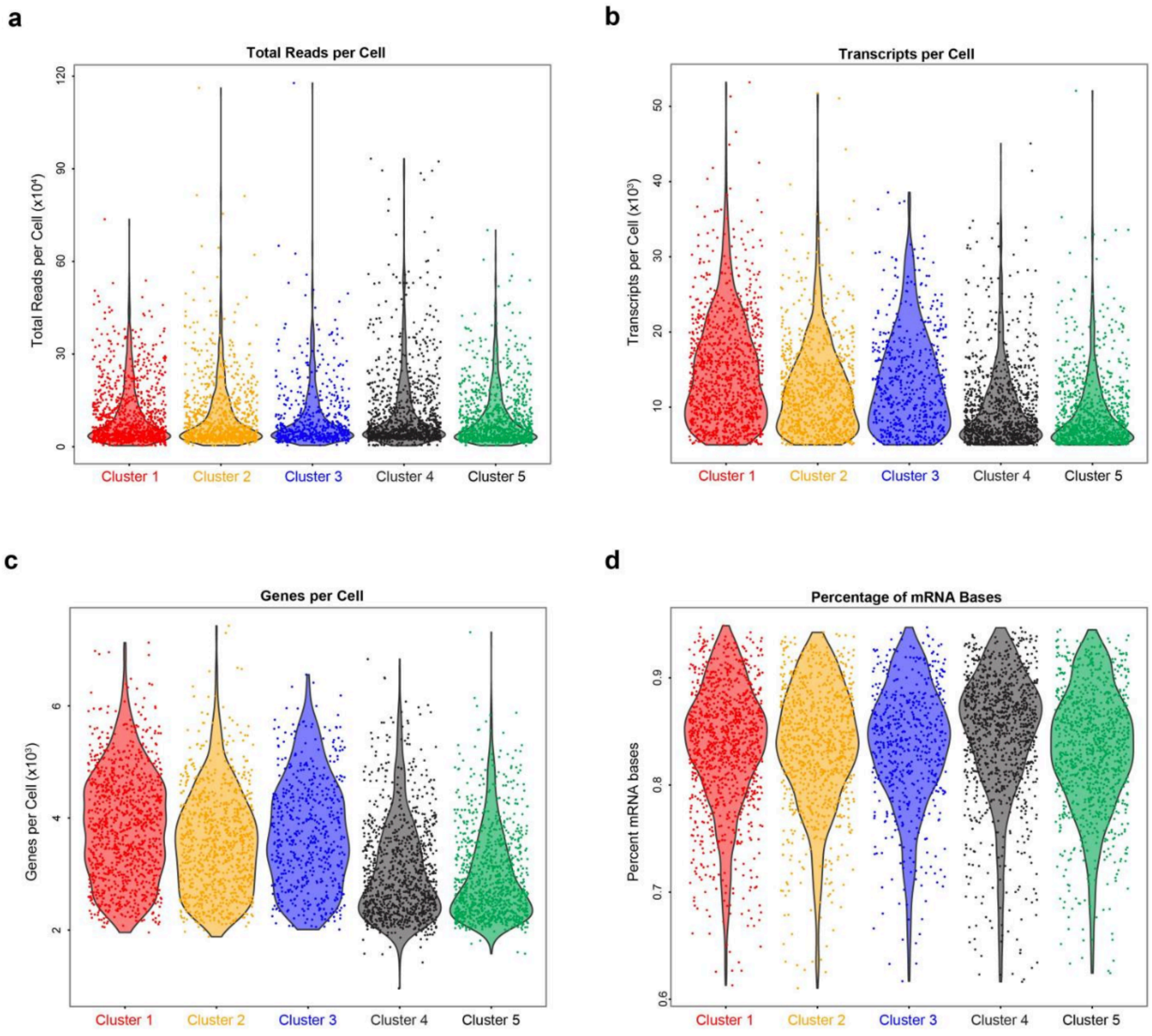
Supplementary Figure 13 | Comparison of Human PBMC Gene and Transcript Capture with Other Massively-Parallel scRNA-Seq Methods. (a) Comparison of transcript capture (**top**) and gene detection (**bottom**) between Seq-Well and 10XGenomics within PBMC cell types prior to downsampling (colored as in **Figure 2**; Center-line: Median; Limits: 1st and 3rd Quartile; Whiskers: +/-1.5 IQR; Points: Values > 1.5 IQR). Among B cells (orange), an average of 1,315 genes and 3,632 transcripts were detected using Seq-Well and an average of 710 genes and 1,910 transcripts were detected in 10X Genomics data. Among CD4 T cells (blue), an average of 861 genes and 2,444 transcripts were detected using Seq-Well and an average of 815 genes and 2,370 transcripts were detected in 10X Genomics data. Among CD8 T cells (yellow), an average of 885 genes and 2,574 transcripts were detected using Seq-Well and an average of 809 genes and 2,029 transcripts were detected in 10X Genomics data. Among Monocytes (green), an average of 1,288 genes and 3,568 transcripts were detected using Seq-Well and an average of 974 genes and 2,835 transcripts were detected in 10X Genomics data. Among NK cells (red), an average of 902 genes and 2,338 transcripts were detected using Seq-Well and an average of 907 genes and 1,943 transcripts were detected in 10X Genomics data. **(b)** Transcript capture (**top**) and gene detection (**bottom**) upon downsampling of Seq-Well data to an average read depth 69,000 reads per cell (Center-

(Continues on the next page)

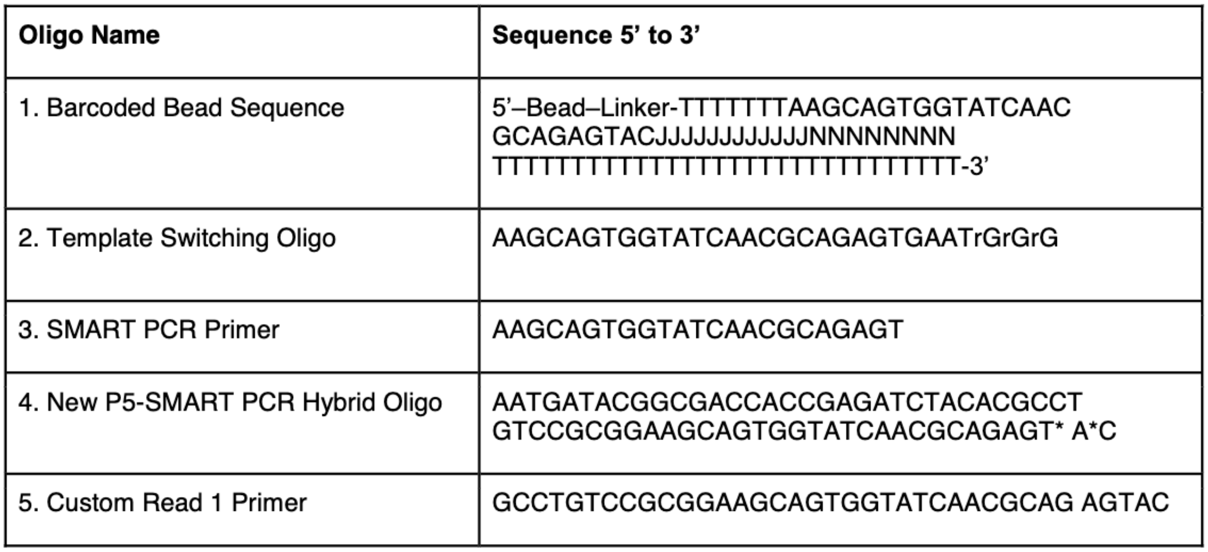
line: Median; Limits: 1st and 3rd Quartile; Whiskers: +/-1.5 IQR; Points: Values > 1.5 IQR). Upon downsampling, in Seq-Well, an average of 1,048 genes and 3,103 transcripts were detected among B cells, 735 genes and 2,221 transcripts among CD4 T cells, 763 genes and 2,353 transcripts among CD8 T cells, 1,052 genes and 3,105 transcripts among monocytes, and 789 genes and 2,041 transcripts among NK cells.



Supplementary Figure 14 | T-SNE Visualization Of Exposed And Unexposed Macrophages Using A 5,000 Transcript Cutoff. (a) Using a threshold of 5,000 detected transcripts, we identified 4,638 macrophages. **(b)** Among these 4,638 cells, we identified 5 distinct clusters of macrophages by performing graph-based clustering over 5 principal components (377 variable genes). **(c)** Clusters 1-3 are defined by unique gene expression signatures, while Clusters 4 and 5 are defined by expression of mitochondrial genes, suggesting low-quality cells. **(d)** Following removal of cells within Clusters 4 and 5, there remain a total of 2,560 cells in Clusters 1-3.



Supplementary Figure 15 | Quality By Cluster Among TB Macrophages. (**a-c**) Violin plots depicting **(a)** reads, **(b)** transcripts, and **(c)** genes per cell, separated by cluster. **(d)** Percent mRNA bases per cell, separated by cluster.



Oligo Name	Sequence 5’ to 3’
1. Barcoded Bead Sequence	5’–Bead–Linker-TTTTTTTAAGCAGTGGTATCAAC GCAGAGTACJJJJJJJJJJJJNNNNNNNN TTTTTTTTTTTTTTTTTTTTTTTTTTTTTT-3’
2. Template Switching Oligo	AAGCAGTGGTATCAACGCAGAGTGAATrGrGrG
3. SMART PCR Primer	AAGCAGTGGTATCAACGCAGAGT
4. New P5-SMART PCR Hybrid Oligo	AATGATACGGCGACCACCGAGATCTACACGCCT GTCCGCGGAAGCAGTGGTATCAACGCAGAGT* A*C
5. Custom Read 1 Primer	GCCTGTCCGCGGAAGCAGTGGTATCAACGCAG AGTAC

Supplementary Table 1 | Oligo Sequences. Sequences of oligos used in Seq-Well. **(1)** The barcoded bead sequence is constructed on the surface of the bead, and cell barcodes are generated through split and pool synthesis. **(2)** The template switching oligo (TSO) is used to tag the 5’ end of captured mRNA using a reverse transcriptase enzyme with terminal transferase activity. **(3)** Sequence for PCR primer used to perform whole-transcriptome amplification (WTA) PCR reaction following reverse transcription and ExoI digestion. **(4)** Sequence that selectively primes the bead-specific SMART sequence during the post-tagmentation step-out PCR, which appends a P5 sequencing adapter. (**5**) Primer used during sequencing that selectively primes the bead-specific primer site to initiate sequencing of the barcode and UMI in Illumina Read 1.

Name	Excitation, nm	Emission, nm	Exposure, ms	Intensity, %	Gain, abs.
Transmitted Light	-	-	50	5	10
CerCP710	485	725/40	100	100	50
PECy5.5	560	725/40	100	100	20
APCCy7	650	775LP	100	100	100
PECy7	560	775LP	100	100	20
PECy5	560	680/42	100	100	10
PerCP650	485	680/42	100	100	30
AF647	650	680/42	100	100	10
AF568	560	607/36	100	100	10
AF488	485	525/39	100	100	10

Supplementary Table 2 | Microscope Settings. The excitation light wavelengths, emission filters, exposure times, light source intensity and camera gain settings used tocapture the fluorescence of the indicated fluorophore are displayed.

Supplementary Table 3 | Gene Expression Matrix for PBMCs. UMI count matrix for the 4,296 PBMCs, labeled by array, that had at least 10,000 reads, 1,000 transcripts, and 500 genes, with at least 65% bases mapping to the transcriptome.

Supplementary Table 4 | PBMC Cluster Enrichments. Lists of genes enriched within each PBMC cluster (B cells, CD4 T Cells, CD8 T cells, Dendritic cells, Monocytes, NK Cells) based on a likelihood-ratio test in which members of each cluster are compared to members of all other clusters.

Supplementary Table 5 | Gene Expression Matrix for Mtb-Exposed Monocyte-Derived Macrophages and Unexposed Control Cells. UMI count matrix for the 4,638 monocyte-derived macrophages, labeled by exposure, that had at least 5,000 mapped transcripts.

Supplementary Table 6 | TB Cluster Enrichments. We examined sets of enriched sets of genes among all cells (irrespective of TB exposure) within Cluster 1, 2 and 3 using the find.markers function in Seurat, which implements a 'roc' test to identify relative expression differences. **(a)** From this analysis, we identified sets of genes exclusively enriched in each cluster but not the others. For each of the identified gene sets, we performed gene set enrichment analysis in DAVID and GSEA. For analysis in DAVID, we compared each gene list for enrichment among GO terms and curated pathways against a background list of 9045 genes contained in the DAVID database that were detected in at least 5% of cells. For analysis using GSEA, we compared each gene list to the complete database of gene sets contained within GSEA. **(b, c)** Within Cluster 1, we observe unique enrichment of 134 genes related to TNF-alpha signaling, inflammation, immune response and LPS response among 1099 cells. **(d, e)** In Cluster 2, we observe exclusive enrichment of 251 genes among 904 cells that distinguish monocytes from dendritic cells in culture and characterize TNF-alpha signaling. **(f, g)** Finally, in Cluster 3, we observe unique enrichment of 118 genes among 557 cells related to specifically to hypoxia, LPS stimulation, TNF-alpha signaling and apoptosis.

Supplementary Table 7 | Cluster Enrichments between Exposure Groups. Initially, we separately identified enriched genes among exposed and unexposed cells within Clusters 1, 2 and 3 using the find.markers function in Seurat. In total, we identified 18 genes with conserved enrichment among TB exposed cells within Clusters 1,2 and 3. For each cluster, we identified the set of genes enriched among exposed cells and unexposed cells within each cluster. **(a)** We identified 28 enriched genes that were conserved among exposed cells across clusters and 31 conserved genes among unexposed, of which 18 were conserved between exposed and unexposed. We identified 38 genes uniquely enriched among exposed cells in Cluster 1 and 54 genes uniquely enriched among unexposed cells, of which 5 were conserved between exposed and unexposed cells within Cluster 1. We identified 134 genes unique to Cluster 2 among exposed cells and 200 genes unique to Cluster 2 among unexposed cells, of which 35 were conserved between exposed and unexposed cells within Cluster 2. In Cluster 3, we identified 43 genes unique to cluster 3 among exposed cells and 44 genes unique to Cluster 3 among unexposed cells, of which 9 were conserved between exposed and unexposed cells within Cluster 3. We performed gene set enrichment analyses for single gene list using DAVID and GSEA. In DAVID, we specified a background list of 9045 genes and examined enrichments within GO terms and curated pathways. For the analysis in GSEA, we made comparisons of gene lists to the complete database of gene sets within GSEA. **(b, c)** Among the 18 genes conserved across clusters in both exposed and unexposed cells, we observed strong enrichment for LPS response, TNF-alpha signaling, phagosome formation and macrophage activation. **(d, e)** Among the 5 genes unique to Cluster 1 conserved between exposed and unexposed cells, we observed enrichment of PI3K-Akt signaling and immune activation. **(f, g)** Among the 35 genes unique to Cluster 2 conserved between exposed and unexposed cells, we observed enrichment of genes related to monocyte culture and the coagulation cascade. (**h, i**) Among the 9 genes unique to Cluster 3 conserved between exposed and unexposed, we observed enrichment of genes up-regulated by HGF and apoptosis.

Supplementary Table 8 | Differentially Expressed Genes between TB Exposed and Unexposed Cells within Each Cluster. We performed a likelihood ratio test to identify genes differentially expressed between TB exposed and unexposed cells within each cluster. **(a)** Differential expression results between 673 TB-exposed and 426 unexposed cells in Cluster 1. **(b)** Differential expression results between 627 TB-exposed and 277 unexposed cells in Cluster 2. **(c)** Differential expression results between 386 TB-exposed and 171 unexposed cells in Cluster 3.

Supplementary Table 9 | TB Infection by Cluster Enrichments. Initially, we performed LRT within each cluster to identify genes differentially expressed between TB exposed and unexposed cells. For each cluster, we created lists of genes differentially expressed with p-values less than 5.0x10-6within each cluster (**Figure 3c**). **(a)** We then compared these lists to identify genes that are differentially expressed genes exclusively within each cluster. We also identified a list of 37 genes detected as differentially expressed across all clusters. We then performed gene set enrichment analysis in DAVID and GSEA to examine functional enrichment of the identified gene sets (i.e. Genes conserved across and unique to each cluster). We performed analysis in DAVID, comparing genes 37 conserved genes, 22 genes unique to Cluster 1, 142 genes unique to Cluster 2, and 40 genes unique to Cluster 3 to a background list of 9,381 genes expressed in at least 5% of filtered cells (Methods). Using GSEA, we made comparisons of the above gene list to the complete list of curated gene sets within the GSEA database (MSigDB v5.1: http://software.broadinstitute.org/gsea/msigdb/index.jsp). **(b,c)** Within Cluster 1, we observed unique enrichment of genes related to growth, proliferation and cell cycle. **(d, e)** Within Cluster 2, we observed enrichment of genes that identify monocyte and dendritic cell culture in addition to proliferation. **(f, g)** Within Cluster 3, we observed unique enrichment of genes related to hypoxia, oxidative stress and oxygen homeostasis.

Supplementary Table 10 | GSEA Comparisons of Exposed and Unexposed Cells within Each Cluster. We performed comparisons between M. tuberculosis exposed and unexposed cells within each cluster using GSEA. For each cluster we created .gct files containing normalized expression data for every cells within each cluster and assigned phenotypes (i.e. TB exposed vs. unexposed) to each cell using a .cls file. We then performed gene set enrichment analysis for each cluster across the complete gene set database in GSEA with 1000 permutations of assigned phenotype. **(a,b)** In cluster 1, we observed enrichment of dendritic cell maturation, monocytes in culture, response to L. donovani, and TNF-alpha signaling among 673 TB exposed cells and relative enrichment of ribosomal genes and protein synthesis among 426 unexposed cells. **(c,d)** In Cluster 2, we observed enrichment of LPS response, dendritic cell maturation, IL1 stimulation and response to TGF-beta among 627 TB exposed cells and relative enrichment of housekeeping functions, ribosomal genes and translation among 277 TB unexposed cells. **(e,f)** In Cluster 3, we observed enrichment of among delayed response to LPS (48 response), TLR7/8 stimulation, inflammatory response, intracellular infection and TNF signaling among 386 TB exposed cells and relative enrichment of housekeeping functions (ribosome, translation, actin) among 171 unexposed cells. **(g,h)** In Cluster 4, we observed enrichment of mitochondrial gene signatures, oxidative phosphorylation, hypoxia response and interferon response among 74 TB exposed cells and enrichment of ribosomal genes and translation among 975 unexposed cells. **(i, j)** In Cluster 5, we observed enrichment of LPS stimulation, TNF signaling, sepsis and dendritic cells maturation among 988 TB exposed cells and enrichment of ribosomal proteins and translation among 41 unexposed cells.

Appendix C - Supplemental Information for Chapter 4: Highly Efficient, Massively-Parallel Single-Cell RNA-Seq Reveals Cellular States and Molecular Features of Human Skin Pathology

This chapter is adapted in accordance with Cold Spring Harbor Laboratory's open access policy from the following article published in BioRxiv:

Travis K Hughes*, Marc H Wadsworth II*, Todd M Gierahn*, Tran Do , David Weiss , Priscilla R. Andrade , Feiyang Ma , Bruno J. de Andrade Silva , Shuai Shao , Lam C Tsoi , Jose Ordovas-Montanes, Johann E Gudjonsson , Robert L Modlin, and Alex K Shalek

** Denotes equal authorship*

STAR* Methods

KEY RESOURCES TABLE

Reagent of Resource	Source	Identifier
Maxima H-RT and Buffer	ThermoFisher Scientific	EP0751
dNTPs	New England Biolabs	N0447L
Polyethylene Glycol 8000	Fisher Scientific	BP233-1
SUPERase*In RNase inhibitor	ThermoFisher Scientific	AM2696
Exonuclease I and Buffer	New England Biolabs	M0293S
1M Tris-HCl, pH 8.0	ThermoFisher Scientific	15568025
Klenow Fragment (3'→5' exo-)	New England Biolabs	M0212L
KAPA 2x HiFi HotStart PCR mix	Kapa Biosystems	KK2602
Nextera XT Kit	Illumina, Inc	FC-131-1096
UltraPure DNase/Rnase-Free Distilled Water	ThermoFisher Scientific	10977015
TWEEN 20	Fisher Scientific	BP337-100
Sodium Dodecyl Sulfate (SDS) Solution	Sigma	71736-100mL
TE Buffer	ThermoFisher Scientific	12090015

Primers

Template-Switching Oligo	IDT	AAGCAGTGGTATCAACGCAGAG TGAATrGrGrG
SMART PCR Primer	IDT	AAGCAGTGGTATCAACGCAGAGT
S^3 Randomer	IDT	AAGCAGTGGTATCAACGCAGAGTGANNNGGNNNB
P5-SMART Hybrid Oligo	IDT	AATGATACGGCGACCACCGAGATCTACACGCCTGTCCGCG-GAAGCAG TGGTATCAACGCAGAGT*A*C
Custom Read 1 Primer	IDT	GCCTGTCCGCGGAAGCAGTGGTATCAACGCAGA-GTAC

Biological Samples

Skin Biopsies	Clinical Biopsies	UCLA
PBMCs	Patient Blood Draw	MGH
HEK293	Cell Lines	ATCC
NIH/3T3s	Cell Lines	ATCC

Critical Commercial Assays

mRNA Capture Beads	Chemgenes Corp.	MACOSKO-2011-10B
KAPA 2x HiFi HotStart PCR mix	Kapa Biosystems	KK2602
NextSeq500	Illumina	Ragon Institute
Nova-Seq S2	Illumina	Broad Institute

Software and Algorithms

Seurat	Satija et al, 2015	http://satijalab.org/seurat/
SCANPY	Wolf et al, 2018	http://github.com/theislab/Scanpy
UMAP	Becht et al, 2018	http://github.com/lmcinnes/umap/
t-SNE	Van der Maaten et al, 2008	http://lvdmaaten.github.io/tsne/

Immunofluorescence Antibodies

IL-17RA	LS-C359381	Lifespan Bioscience
IL-17RC	LS-400522	Lifespan Bioscience
APOBEC3A	LS-C98892-400	Lifespan Bioscience
FOSL	A03927	Boster
IL-36G	sc-80056	Santa Cruz Biotechnology
TNFAIP3	ab74037	Abcam

Human Protein Atlas Skin (FLG):

https://www.proteinatlas.org/ENSG00000143631-FLG/tissue/skin+2#img

Human Protein Atlas Skin (KRT14):

https://www.proteinatlas.org/ENSG00000186847-KRT14/tissue/skin+1#img

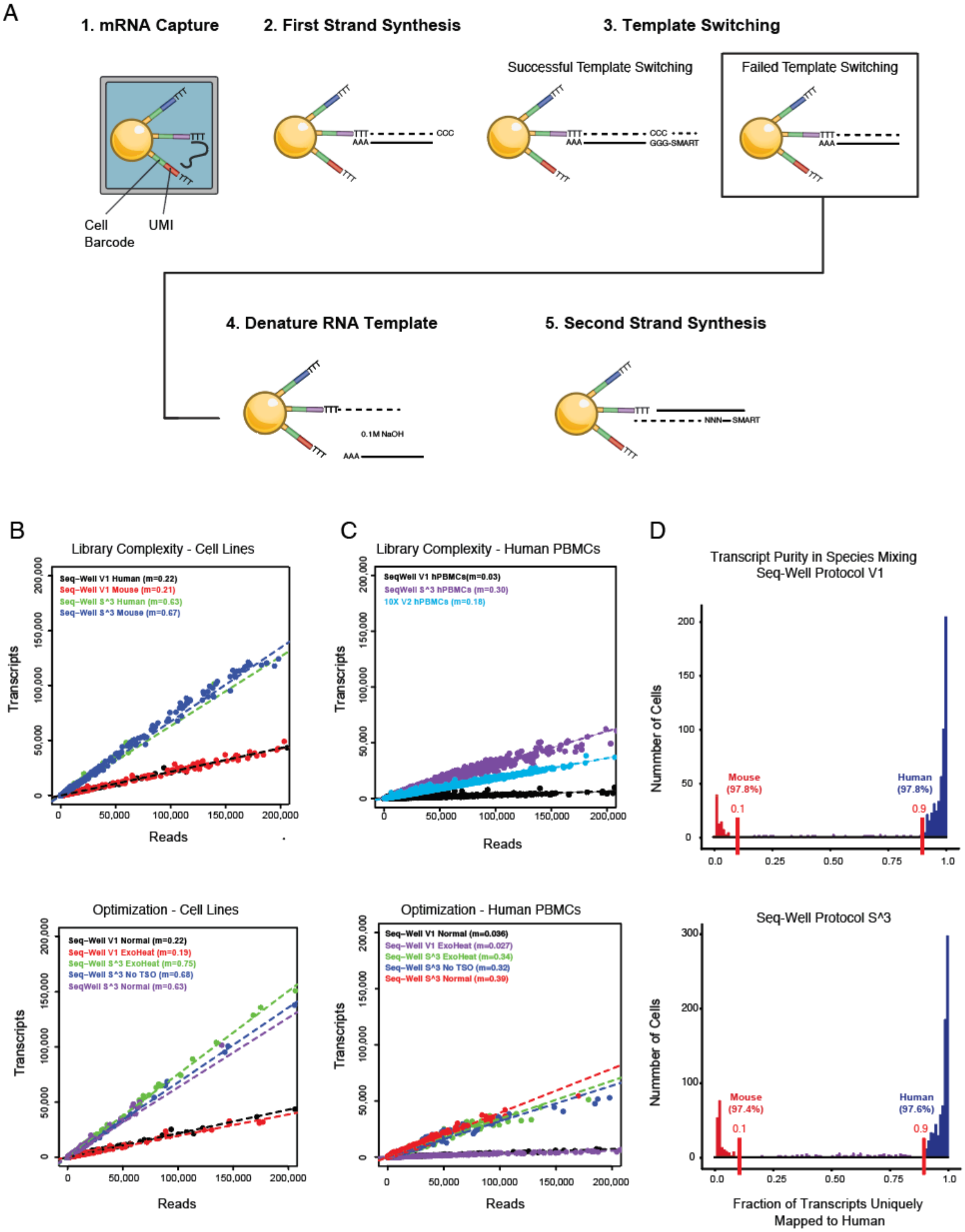

A
1. mRNA Capture
2. First Strand Synthesis
3. Template Switching
Successful Template Switching
Failed Template Switching
TTT
AAA
CCC
GGG-SMART
Cell Barcode
UMI
4. Denature RNA Template
5. Second Strand Synthesis
0.1M NaOH
NNN—SMART
B
Library Complexity - Cell Lines
Seq–Well V1 Human (m=0.22)
Seq–Well V1 Mouse (m=0.21)
Seq–Well S^3 Human (m=0.63)
Seq–Well S^3 Mouse (m=0.67)
Transcripts
Reads
0
50,000
100,000
150,000
200,000
C
Library Complexity - Human PBMCs
SeqWell V1 hPBMCs(m=0.03)
SeqWell S^3 hPBMCs (m=0.30)
10X V2 hPBMCs (m=0.18)
Transcripts
Reads
D
Transcript Purity in Species Mixing
Seq-Well Protocol V1
Nummber of Cells
0
50
100
150
200
Mouse
(97.8%)
0.1
Human
(97.8%)
0.9
0.0
0.25
0.50
0.75
1.0
Optimization - Cell Lines
Seq–Well V1 Normal (m=0.22)
Seq–Well V1 ExoHeat (m=0.19)
Seq–Well S^3 ExoHeat (m=0.75)
Seq–Well S^3 No TSO (m=0.68)
SeqWell S^3 Normal (m=0.63)
Transcripts
Reads
Optimization - Human PBMCs
Seq–Well V1 Normal (m=0.036)
Seq–Well V1 ExoHeat (m=0.027)
Seq–Well S^3 ExoHeat (m=0.34)
Seq–Well S^3 No TSO (m=0.32)
Seq–Well S^3 Normal (m=0.39)
Transcripts
Reads
Seq-Well Protocol S^3
Nummber of Cells
300
200
100
0
Mouse
(97.4%)
0.1
Human
(97.6%)
0.9
Fraction of Transcripts Uniquely Mapped to Human

Figure S1 | Second-Strand Synthesis Overview, related to Figure 1.

A. Illustration of the second strand synthesis procedure: (1) mRNA is captured via poly-T priming of poly-adenylated mRNA; (2) First strand synthesis is performed to generate single-stranded cDNA template on bead-bound sequences; (3) Successful template switching: The use of enzymes with terminal transferase activity generates a 3' overhang of 3 cytosines. Template switching utilizes this overhang to append the SMART sequence to both ends of the cDNA molecule during first strand synthesis. Failed Template Switching: If template switching fails, this results in loss of previously primed and reverse transcribed mRNA molecules; (4) mRNA template is chemically denatured using 0.1M NaOH; (5) Second strand synthesis is performed using a random-octamer with the SMART sequence in the 5' orientation; and, (6) Following second strand synthesis, PCR amplification, library preparation and sequencing are performed to generate data.

B. Scatterplots show the relationship between transcript detection (y-axis) and number of aligned reads per cell (x-axis) for an initial experiment (**top**) series of optimization conditions using HEK293 and NIH-3T3 cell lines (**botttom**).

C. Scatterplots that illustrate the relationship between number of transcripts detected (y-axis) and number of aligned reads per cell (x-axis) between Seq-Well V1 and Seq-Well S^3 in sequencing experiments for an initial experiment (**top**) and a series of optimization experiment using human PBMCs (**bottom**).

D. Histograms that show the fraction of transcripts uniquely mapped to the human genome for each cell for Seq-Well V1 (**Top**) and Seq-Well S^3 (**Bottom**). Colors indicate species classification for cells with at least 90% purity of human (blue) or mouse (red) mapping.

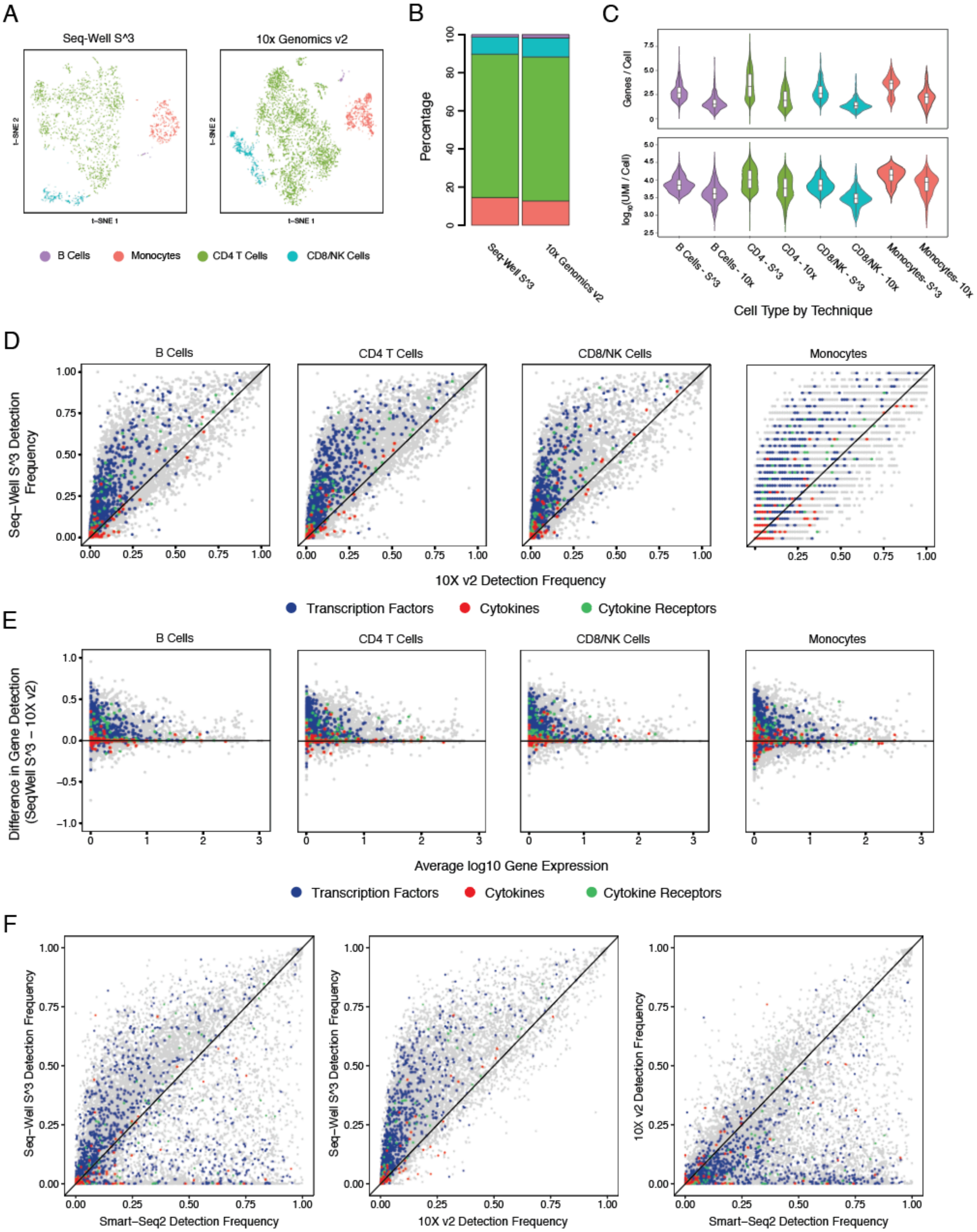

A
Seq-Well S^3
10x Genomics v2
t-SNE 2
t-SNE 1
B Cells
Monocytes
CD4 T Cells
CD8/NK Cells
B
Percentage
0
20
40
60
80
100
Seq-Well S^3
10x Genomics v2
C
Genes / Cell
0
2.5
5.0
7.5
log10(UMI / Cell)
2.5
3.0
3.5
4.0
4.5
5.0
B Cells - S^3
B Cells - 10x
CD4 - S^3
CD4 - 10x
CD8/NK - S^3
CD8/NK - 10x
Monocytes- S^3
Monocytes- 10x
Cell Type by Technique
D
B Cells
CD4 T Cells
CD8/NK Cells
Monocytes
Seq−Well S^3 Detection Frequency
0.00
0.25
0.50
0.75
1.00
10X v2 Detection Frequency
Transcription Factors
Cytokines
Cytokine Receptors
E
B Cells
CD4 T Cells
CD8/NK Cells
Monocytes
Difference in Gene Detection (SeqWell S^3 − 10X v2)
−1.0
−0.5
0.0
0.5
1.0
0
1
2
3
Average log10 Gene Expression
Transcription Factors
Cytokines
Cytokine Receptors
F
Seq−Well S^3 Detection Frequency
Smart−Seq2 Detection Frequency
Seq−Well S^3 Detection Frequency
10X v2 Detection Frequency
10X v2 Detection Frequency
Smart−Seq2 Detection Frequency
0.00
0.25
0.50
0.75
1.00

Figure S2 | PBMC Methods Comparisons, related to Figure 1.

A. t-SNE plot showing detected cell-types among PBMCs including CD4$^+$ T cells (green), CD8/NK Cells (blue), B cells (purple), and Monocytes (red) using 10X v2 and Seq-Well S^3. Cells recovered using Seq-well are colored with darker shades.

B. Stacked barplots show the proportion of cell types recovered using Seq-Well S^3 (**left**) and 10X v2 (**right**).

C. Top: Violin plots (boxplots median +/- quartiles) showing the distribution of per cell gene detection from Seq-Well S^3 (**left**) and 10X v2 (**right**). Bottom: Violin plots (boxplots median +/- quartiles) showing the distribution of per cell-gene detection from Seq-Well S^3 (**left**) and 10X v2 (**right**).

D. Scatterplots showing a comparison of gene detection frequencies between Seq-Well S^3 (y-axis) and 10x v2 (x-axis) for each cell type.

E. Scatterplots showing the difference in gene detection between Seq-Well S^3 and 10X v2 (y-axis) as a function of average normalized expression (x-axis).

F. Scatterplots showing a comparison of gene detection frequencies among sorted CD4$^+$ T cells between **(Left)** Seq-well S^3 (y-axis) and 10x v2 (x-axis), (**Middle)** Seq-Well S^3 (y-axis) and Smart-Seq2 (x-axis), and **(Right)** 10x v2 (y-axis) and Smart-Seq2 (x-axis).

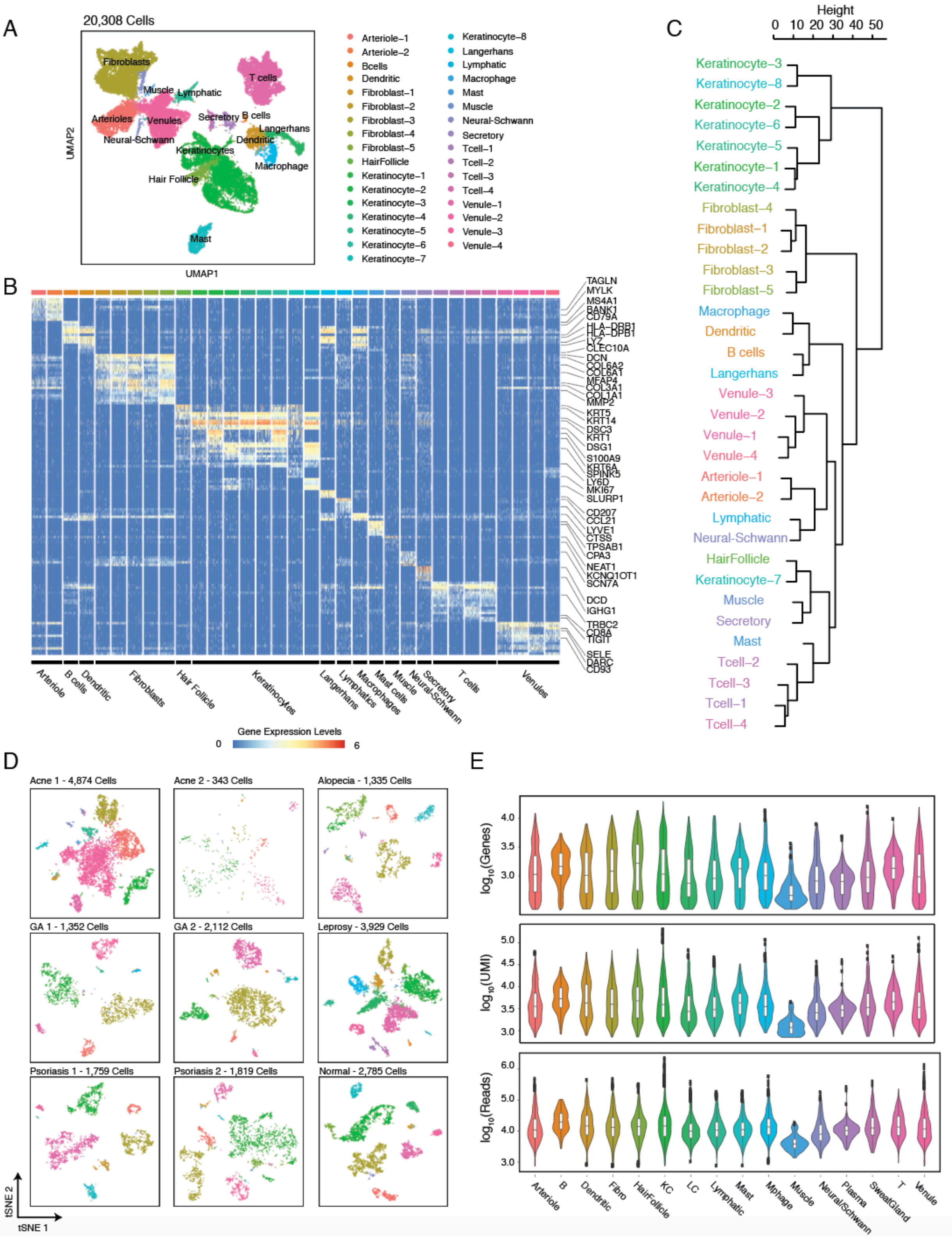

A
20,308 Cells
Fibroblasts
Muscle
Lymphatic
T cells
Arterioles
Venules
Secretory
B cells
Langerhans
Neural-Schwann
Dendritic
Keratinocytes
Macrophage
Hair Follicle
Mast
UMAP2
UMAP1
Arteriole–1
Arteriole–2
Bcells
Dendritic
Fibroblast–1
Fibroblast–2
Fibroblast–3
Fibroblast–4
Fibroblast–5
HairFollicle
Keratinocyte–1
Keratinocyte–2
Keratinocyte–3
Keratinocyte–4
Keratinocyte–5
Keratinocyte–6
Keratinocyte–7
Keratinocyte–8
Langerhans
Lymphatic
Macrophage
Mast
Muscle
Neural-Schwann
Secretory
Tcell–1
Tcell–2
Tcell–3
Tcell–4
Venule–1
Venule–2
Venule–3
Venule–4
B
TAGLN
MYLK
MS4A1
BANK1
CD79A
HLA–DRB1
HLA–DPB1
LYZ
CLEC10A
DCN
COL6A2
COL6A1
MFAP4
COL3A1
COL1A1
MMP2
KRT5
KRT14
DSC3
KRT1
DSG1
S100A9
KRT6A
SPINK5
LY6D
MKI67
SLURP1
CD207
CCL21
LYVE1
CTSS
TPSAB1
CPA3
NEAT1
KCNQ1OT1
SCN7A
DCD
IGHG1
TRBC2
CD8A
TIGIT
SELE
DARC
CD93
Arteriole
B cells
Dendritic
Fibroblasts
Hair Follicle
Keratinocytes
Langerhans
Lymphatics
Macrophages
Mast cells
Muscle
Neural-Schwann
Secretory
T cells
Venules
Gene Expression Levels
0
6
C
Height
0 10 20 30 40 50
Keratinocyte–3
Keratinocyte–8
Keratinocyte–2
Keratinocyte–6
Keratinocyte–5
Keratinocyte–1
Keratinocyte–4
Fibroblast–4
Fibroblast–1
Fibroblast–2
Fibroblast–3
Fibroblast–5
Macrophage
Dendritic
B cells
Langerhans
Venule–3
Venule–2
Venule–1
Venule–4
Arteriole–1
Arteriole–2
Lymphatic
Neural-Schwann
HairFollicle
Keratinocyte–7
Muscle
Secretory
Mast
Tcell–2
Tcell–3
Tcell–1
Tcell–4
D
Acne 1 - 4,874 Cells
Acne 2 - 343 Cells
Alopecia - 1,335 Cells
GA 1 - 1,352 Cells
GA 2 - 2,112 Cells
Leprosy - 3,929 Cells
Psoriasis 1 - 1,759 Cells
Psoriasis 2 - 1,819 Cells
Normal - 2,785 Cells
tSNE 2
tSNE 1
E
log10(Genes)
4.0
3.5
3.0
log10(UMI)
5.0
4.5
4.0
3.5
3.0
log10(Reads)
6.0
5.0
4.0
3.0
Arteriole
B
Dendritic
Fibro
HairFollicle
KC
LC
Lymphatic
Mast
Mphage
Muscle
Neural/Schwann
Plasma
SweatGland
T
Venule

Figure S3 | Overview of Samples, related to Figure 2.

A. UMAP plot for 20,308 cells colored by 33 cell type cell type clusters (Louvain Resolution: 2.0).

B. Heatmap showing the relative expression of cell-type defining gene signatures across 20,308 cells (**Table S3**).

C. Dendrogram of hierarchical clustering shows similarity of cell type clusters among top 25 cluster-defining genes (**Appendix D, Figure S3B**).

D. t-SNE plots for each of the nine skin biopsies colored by generic cell type.

E. Violin plots show the distribution of per-cell quality metrics displayed in UMAP embedding of 20,308 cells colored by colored generic cell-type classification (**Figure 2B**).

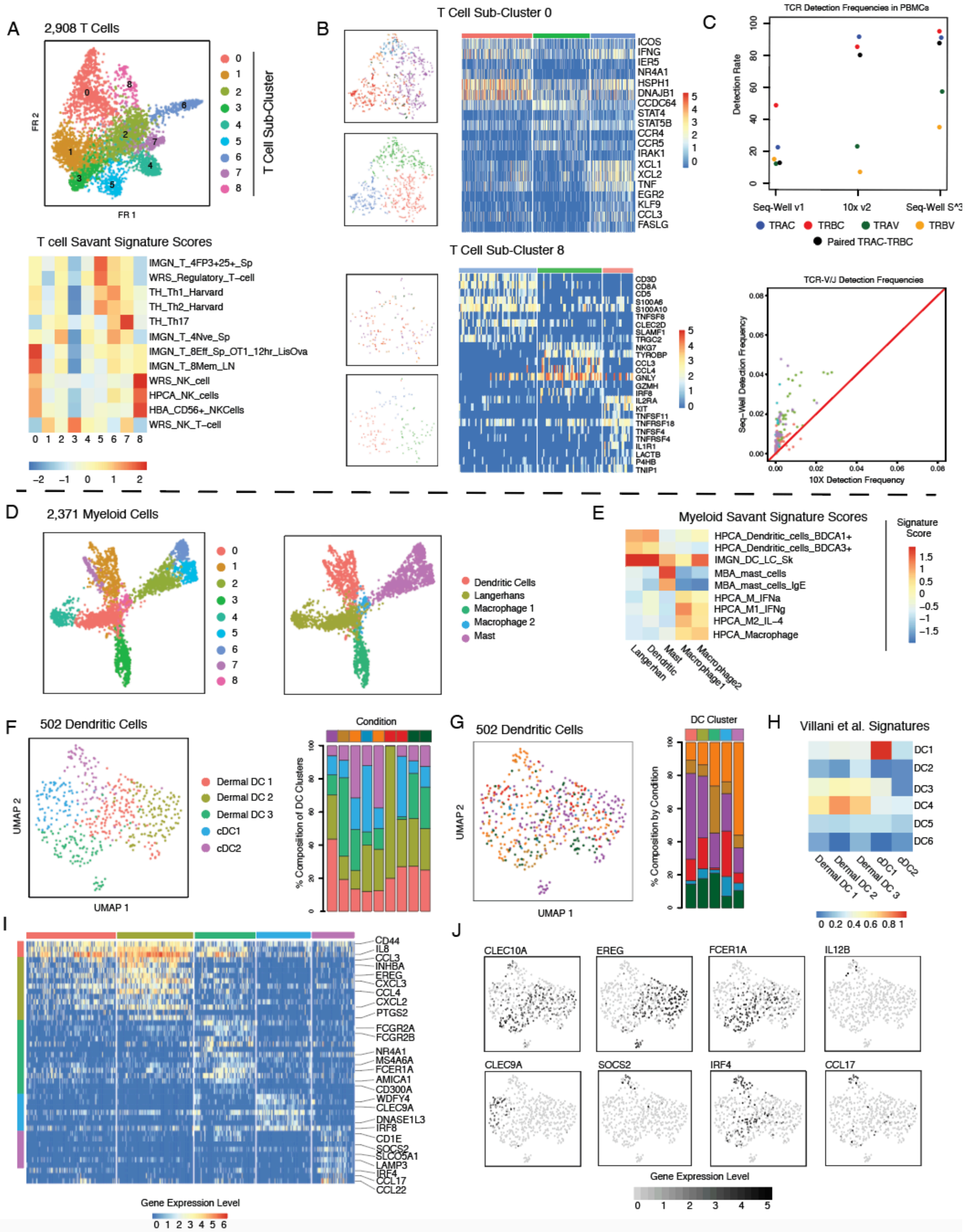

A
2,908 T Cells
FR 2
FR 1
0
1
2
3
4
5
6
7
8
T Cell Sub-Cluster
T cell Savant Signature Scores
IMGN_T_4FP3+25+_Sp
WRS_Regulatory_T-cell
TH_Th1_Harvard
TH_Th2_Harvard
TH_Th17
IMGN_T_4Nve_Sp
IMGN_T_8Eff_Sp_OT1_12hr_LisOva
IMGN_T_8Mem_LN
WRS_NK_cell
HPCA_NK_cells
HBA_CD56+_NKCells
WRS_NK_T-cell
0 1 2 3 4 5 6 7 8
−2 −1 0 1 2
B
T Cell Sub-Cluster 0
ICOS
IFNG
IER5
NR4A1
HSPH1
DNAJB1
CCDC64
STAT4
STAT5B
CCR4
CCR5
IRAK1
XCL1
XCL2
TNF
EGR2
KLF9
CCL3
FASLG
5 4 3 2 1 0
T Cell Sub-Cluster 8
CD3D
CD8A
CD5
S100A6
S100A10
TNFSF8
CLEC2D
SLAMF1
TRGC2
NKG7
TYROBP
CCL3
CCL4
GNLY
GZMH
IRF8
IL2RA
KIT
TNFSF11
TNFRSF18
TNFSF4
TNFRSF4
IL1R1
LACTB
P4HB
TNIP1
5 4 3 2 1 0
C
TCR Detection Frequencies in PBMCs
Detection Rate
0 20 40 60 80 100
Seq-Well v1
10x v2
Seq-Well S^3
● TRAC ● TRBC ● TRAV ● TRBV
● Paired TRAC-TRBC
TCR-V/J Detection Frequencies
Seq-Well Detection Frequency
0.00 0.02 0.04 0.06 0.08
10X Detection Frequency
0.00 0.02 0.04 0.06 0.08
D
2,371 Myeloid Cells
0
1
2
3
4
5
6
7
8
Dendritic Cells
Langerhans
Macrophage 1
Macrophage 2
Mast
E
Myeloid Savant Signature Scores
HPCA_Dendritic_cells_BDCA1+
HPCA_Dendritic_cells_BDCA3+
IMGN_DC_LC_Sk
MBA_mast_cells
MBA_mast_cells_IgE
HPCA_M_IFNa
HPCA_M1_IFNg
HPCA_M2_IL-4
HPCA_Macrophage
Langerhan
Dendritic
Mast
Macrophage1
Macrophage2
Signature Score
1.5
1
0.5
0
−0.5
−1
−1.5
F
502 Dendritic Cells
UMAP 2
UMAP 1
Dermal DC 1
Dermal DC 2
Dermal DC 3
cDC1
cDC2
Condition
% Composition of DC Clusters
0 20 40 60 80 100
G
502 Dendritic Cells
UMAP 2
UMAP 1
DC Cluster
% Composition by Condition
0 20 40 60 80 100
H
Villani et al. Signatures
DC1
DC2
DC3
DC4
DC5
DC6
Dermal DC 1
Dermal DC 2
Dermal DC 3
cDC1
cDC2
0 0.2 0.4 0.6 0.8 1
I
CD44
IL8
CCL3
INHBA
EREG
CXCL3
CCL4
CXCL2
PTGS2
FCGR2A
FCGR2B
NR4A1
MS4A6A
FCER1A
AMICA1
CD300A
WDFY4
CLEC9A
DNASE1L3
IRF8
CD1E
SOCS2
SLCO5A1
LAMP3
IRF4
CCL17
CCL22
Gene Expression Level
0 1 2 3 4 5 6
J
CLEC10A
EREG
FCER1A
IL12B
CLEC9A
SOCS2
IRF4
CCL17
Gene Expression Level
0 1 2 3 4 5

Figure S4 | Immune Cell Heterogeneity, related to Figures 3 and 4.

A. (**Top**) Force-directed graph of 2,903 T cells colored by T cell sub-cluster. (**Bottom**) Heatmap of gene-set enrichment scores based on comparison of T cell phenotypic sub-clusters to a curated list of reference signatures in the Savant database.

B. Sub-grouping results for (**top**) T cell sub-cluster 0 and (**bottom**) T cell sub-cluster 8. For each analysis, t-SNE plots colored by inflammatory skin condition (**top-left**) and sub-cluster (**bottom-left**) are shown. For each clusters, heatmaps show gene expression patterns across T and NK cells sub-types (**right**).

C. (**Top**) Detection rates for TCR genes for PBMCs in Seq-Well v1, 10x v2. and Seq-Well S^3. (**Bottom**) Detection frequency of TCR V-J (e.g. TRAV/J and TRBV/J) genes in $CD4^+$ T cells from peripheral blood between Seq-Well S^3 (y-axis) and 10x v2 (x-axis). Colors correspond to TRAJ (red), TRAV (green), TRBJ (blue), and TRBV (purple) genes.

D. Force-directed graph of 2,371 myeloid cells colored by myeloid phenotypic sub-clusters.

E. Heatmap of gene-set enrichment scores based on comparison of myeloid phenotypic sub-clusters to a curated list of reference signatures in the Savant database.

F. (**Left**) UMAP plot for 502 dendritic cells from human skin colored by phenotypic sub-grouping. (**Right**) Stacked barplot showing composition of dendritic cells within each of nine skin biopsies by DC sub-cluster.

G. (**Left**) UMAP plot for 502 dendritic cells from human skin colored by inflammatory skin condition. (**Right**) Stacked barplot showing contribution of inflammatory skin conditions to each dendritic cell sub-grouping.

H. Heatmap showing average signature score across 5 dermal DC populations based on dendritic cell signatures from *Villani et al. Science 2017*.

I. Heatmap showing the distribution of normalized gene expression levels for cluster-defining genes across dermal DC subpopulations.

J. UMAP plots colored by normalized expression levels for DC sub-grouping-defining genes.

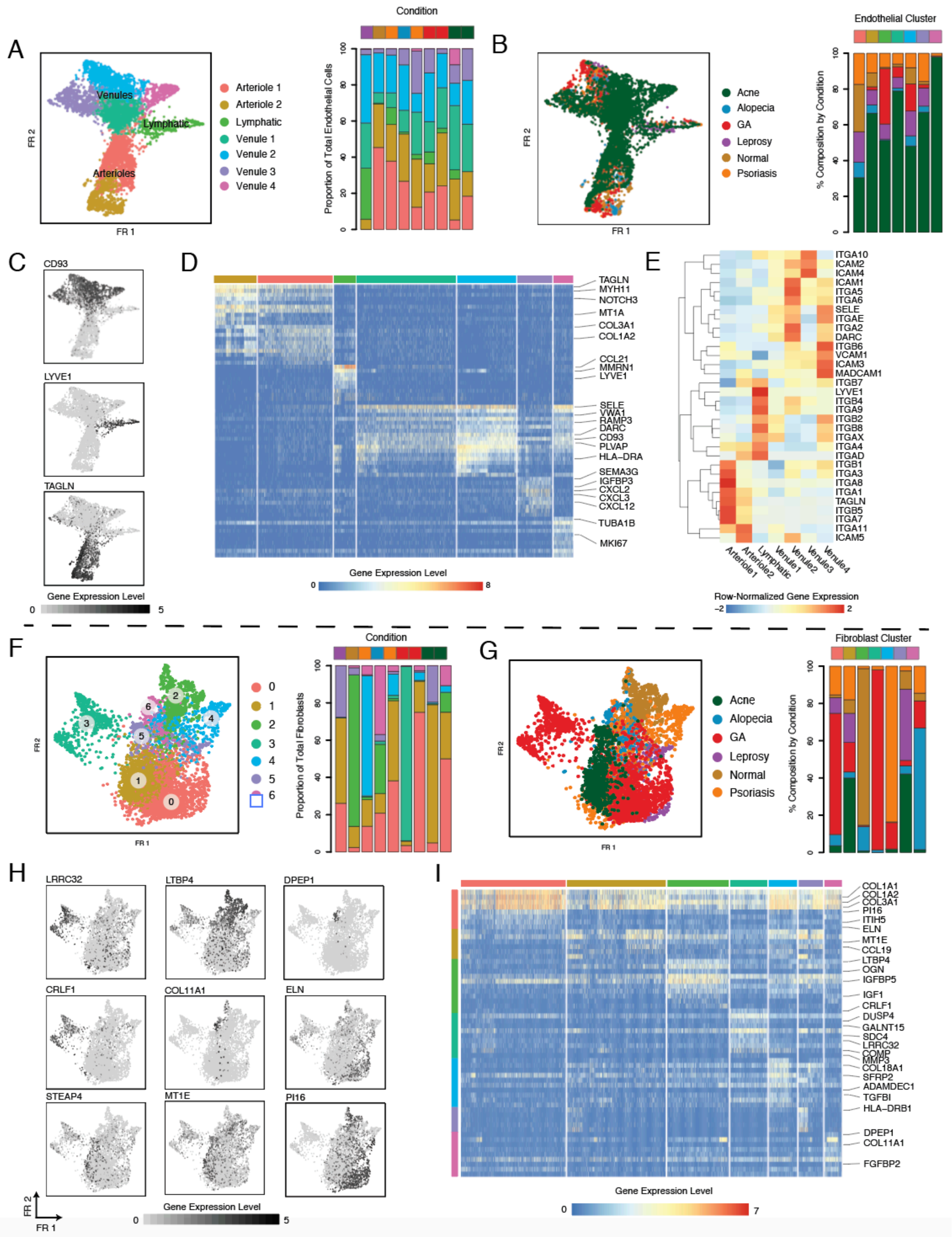

A
Venules
Lymphatic
Arterioles
FR 2
FR 1
Arteriole 1
Arteriole 2
Lymphatic
Venule 1
Venule 2
Venule 3
Venule 4
Condition
Proportion of Total Endothelial Cells
0
20
40
60
80
100
B
FR 2
FR 1
Acne
Alopecia
GA
Leprosy
Normal
Psoriasis
Endothelial Cluster
% Composition by Condition
0
20
40
60
80
100
C
CD93
LYVE1
TAGLN
Gene Expression Level
0
5
D
TAGLN
MYH11
NOTCH3
MT1A
COL3A1
COL1A2
CCL21
MMRN1
LYVE1
SELE
VWA1
RAMP3
DARC
CD93
PLVAP
HLA–DRA
SEMA3G
IGFBP3
CXCL2
CXCL3
CXCL12
TUBA1B
MKI67
Gene Expression Level
0
8
E
ITGA10
ICAM2
ICAM4
ICAM1
ITGA5
ITGA6
SELE
ITGAE
ITGA2
DARC
ITGB6
VCAM1
ICAM3
MADCAM1
ITGB7
LYVE1
ITGB4
ITGA9
ITGB2
ITGB8
ITGAX
ITGA4
ITGAD
ITGB1
ITGA3
ITGA8
ITGA1
TAGLN
ITGB5
ITGA7
ITGA11
ICAM5
Arteriole1
Arteriole2
Lymphatic
Venule1
Venule2
Venule3
Venule4
Row-Normalized Gene Expression
−2
2
F
0
1
2
3
4
5
6
FR2
FR 1
Condition
Proportion of Total Fibroblasts
0
20
40
60
80
100
G
FR2
FR 1
Acne
Alopecia
GA
Leprosy
Normal
Psoriasis
Fibroblast Cluster
% Composition by Condition
0
20
40
60
80
100
H
LRRC32
LTBP4
DPEP1
CRLF1
COL11A1
ELN
STEAP4
MT1E
PI16
FR 2
FR 1
Gene Expression Level
0
5
I
COL1A1
COL1A2
COL3A1
PI16
ITIH5
ELN
MT1E
CCL19
LTBP4
OGN
IGFBP5
IGF1
CRLF1
DUSP4
GALNT15
SDC4
LRRC32
COMP
MMP3
COL18A1
SFRP2
ADAMDEC1
TGFBI
HLA–DRB1
DPEP1
COL11A1
FGFBP2
Gene Expression Level
0
7

Figure S5 | Stromal Cell Diversity.

A. Force-directed plots for 4,996 endothelial cells colored by phenotypic sub-cluster (**left**) and stacked barplot showing the distribution of endothelial phenotypic sub-clusters across samples (**right**).

B. Force-directed plots for 4,996 endothelial colored by inflammatory skin condition (**left**) and stacked barplot (**right**) showing the contribution of each inflammatory skin condition to endothelial phenotypic sub-clusters.

C. Forced-directed plot colored by normalized expression level of genes that mark endothelial cell types: (**Left**) CD93, venules, (**Middle**) TAGLN, arterioles, (**Right**) LYVE1, lymphatics.

D. Heatmap showing patterns of gene expression across 7 clusters of endothelial cells.

E. Heatmap showing row-normalized expression levels of vascular addressins across phenotypic sub-clusters of endothelial cells.

F. Force-directed plots for 4,189 fibroblasts colored by phenotypic sub-cluster (**left**) and stacked barplot showing the distribution of fibroblast phenotypic sub-clusters across samples (**right**).

G. Force-directed plots for 4,189 fibroblasts colored by inflammatory skin condition (**left**) and stacked barplot (**right**) showing the contribution of each inflammatory skin condition to fibroblast phenotypic sub-clusters.

H. Force-directed graphs highlighting fibroblast cluster defining genes.

I. Heatmap showing the normalized gene expression values of fibroblast cluster-defining genes.

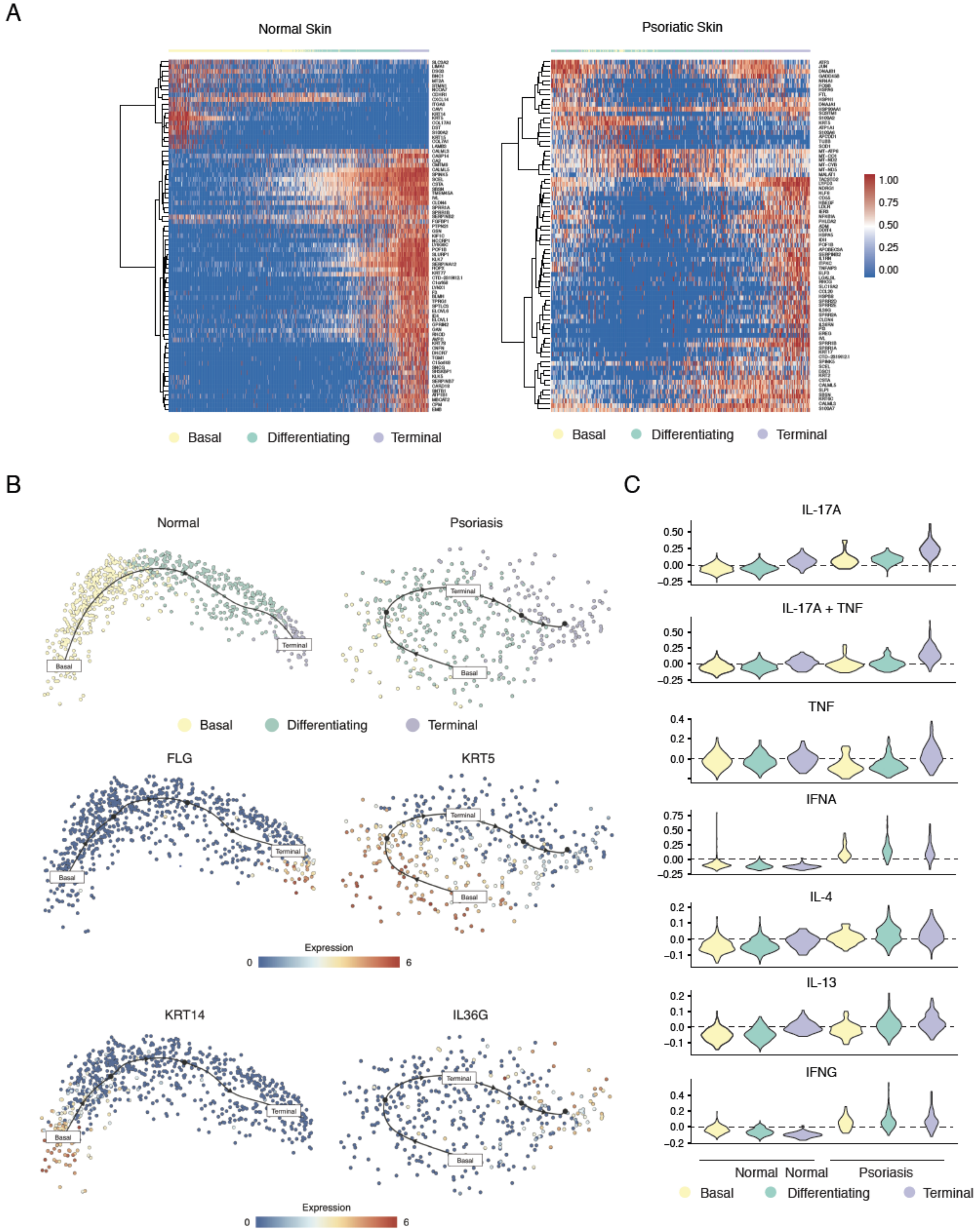

A
Normal Skin
Psoriatic Skin
1.00
0.75
0.50
0.25
0.00
Basal
Differentiating
Terminal
Basal
Differentiating
Terminal
B
Normal
Psoriasis
Basal
Terminal
Basal
Differentiating
Terminal
FLG
KRT5
Expression
0
6
KRT14
IL36G
Expression
0
6
C
IL-17A
IL-17A + TNF
TNF
IFNA
IL-4
IL-13
IFNG
Normal
Normal
Psoriasis
Basal
Differentiating
Terminal

Figure S6 | Keratinocyte Differentiation Trajectories, related to Figure 5.

A. (**Left**) Heatmap showing enrichment of genes along pseudo-temporal trajectories for normal keratinocytes. (**Right**) Heatmap showing enrichment of genes along pseudo-temporal trajectories among psoriatic keratinocytes.

B. Differentiation trajectories for Normal (**left**) and Psoriatic (**right**) keratinocytes.

C. Violin plots showing localization of cytokine response signatures in basal, differentiating and terminal keratinocytes for Normal (**left**) and Psoriatic (**right**) keratinocytes.

SUPPLEMENTARY TABLES

Table S1. Gene detection frequencies for transcription factors, cytokines and cytokine receptors (**Figure S2**).

Table S2. Patient Information

Demographic and condition information for all patients included in this study.

Donor	Visit Date	Lesion Type	Gender	Age	Ethnicity	Race	Biopsy
301	10/23/2017	Psoriasis	F	56	Non-Hispanic	White	1
303	10/23/2017	Alopecia	F	27	Non-Hispanic	Asian, White	1
304	10/23/2017	Psoriasis	F	63	Non-Hispanic	White	1
305	10/23/2017	GA	F	61	Non-Hispanic	White	1
306	10/23/2017	GA	F	58	Non-Hispanic	White	1
307	10/24/2017	Acne	F	46	Hispanic	White	1
308	10/24/2017	Acne	F	29	Hispanic	White	1

Table S3. Gene Signatures. Gene signatures for generic cell-type clusters (**Figure 2B**).

Table S4. T cell sub-cluster-defining genes.

Table S5. Enriched genes across CD8 T cell sub-groupings (within sub-cluster 0).

Table S6. Enriched genes across cytotoxic T cell/NK cell subsets (sub-cluster 8).

Table S7. Myeloid cub-cluster-defining genes.

Table S8. Genes differentially expressed between Langerhan's cells from Leprosy and normal skin.

Table S9. Dendritic cell sub-cluster defining genes.

Table S10. Endothelial cell sub-cluster defining genes

Table S11. Fibroblast sub-cluster defining genes.

Table S12. Differential Expression Results between Psoriatic and Normal Keratinocytes

Table S13. Diffusion Pseudo-time Values for Normal and Psoriatic Keratinocytes.

Per-cell Diffusion Pseudotime values for normal and psoriatic keratinocytes.

Table S14. Keratinocyte Cytokine response signatures. Gene expression signatures generated following cytokine exposure of keratinocyte *in vitro*.

Appendix D: Seq-Well S^3 Master Protocol

As outlined in:

Highly efficient, massively-parallel single-cell RNA-Seq reveals cellular and molecular features of human skin pathology

Travis K Hughes[1,2,3,4,5,8], Marc H Wadsworth II[1,3,4,5,8], Todd M Gierahn[5,8], Feiyang Ma[6], Tran Do[6], David Weiss[6], Priscilla Andrade[6], Bruno Andrade[6], Shuai Shao[7], Lam C Tsoi[7], Johann E Gudjonsson[7], Robert L Modlin[6], J Christopher Love[1,3,4,5,9], and Alex K Shalek[1,2,3,4,5,9]

Affiliations:

[1] Institute for Medical Engineering & Science (IMES) and Department of Chemistry, MIT, Cambridge, Massachusetts, USA

[2] Department of Immunology, Harvard Medical School, Boston, Massachusetts, USA

[3] Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA

[4] Ragon Institute of MGH, MIT and Harvard, Cambridge, Massachusetts, USA

[5] Koch Institute for Integrative Cancer Research, MIT, Cambridge, Massachusetts, USA

[6] Division of Dermatology and Department of Microbiology, Immunology and Molecular Biology, David Geffen School of Medicine, UCLA, Los Angeles, California, USA

[7] Department of Dermatology, University of Michigan, Ann Arbor, Michigan, USA

[8] These authors contributed equally to this work

[9] These senior authors contributed equally to this work

[9] To whom correspondence should be addressed: shalek@mit.edu (AKS), clove@mit.edu (JCL)

*For the latest protocol see Shalek Lab website (www.shaleklab.com/Seq-Well)

Table of Contents

In-Depth Protocol

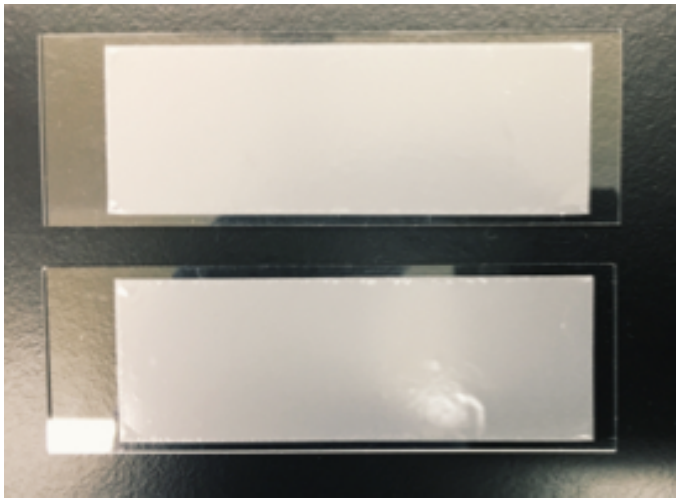
Sub-Appendices

Membrane Preparation

1. Carefully place a pre-cut (22 x 66 mm) polycarbonate membrane onto a glass slide using a gloved finger and tweezers to separate the membrane and paper.

   Note 1: Make certain the shiny side of the polycarbonate membrane is facing up to be in contact with the oxygen plasma and eventually the surface of the array.

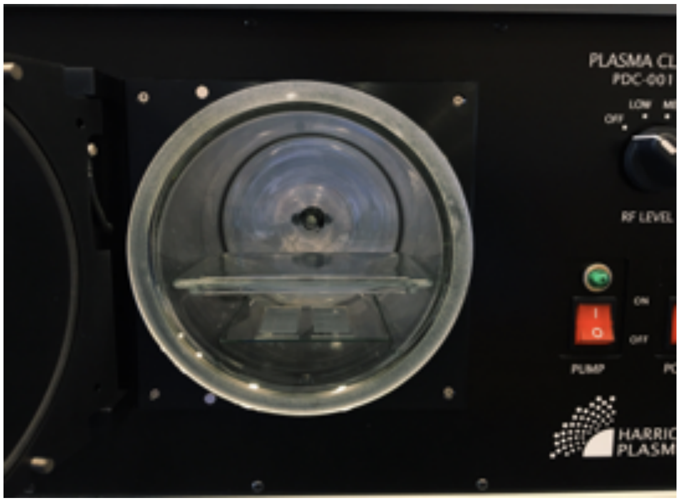
   Note 2: Discard any membranes that have creases or other large-scale imperfections.



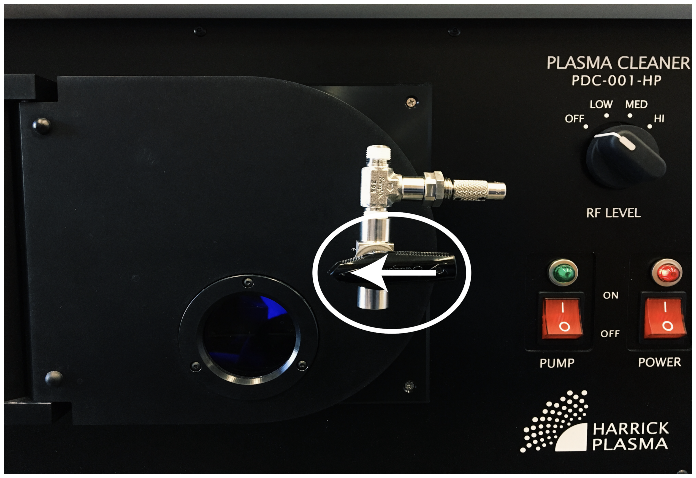
2. Place membranes onto a shelf in the plasma cleaner.

   Note 1: Shelves are not provided, but any piece of glass will do.

   Note 2 (optional): If you have two shelves, place membranes on the bottom shelf to reduce risk of them flying after vacuum is removed.

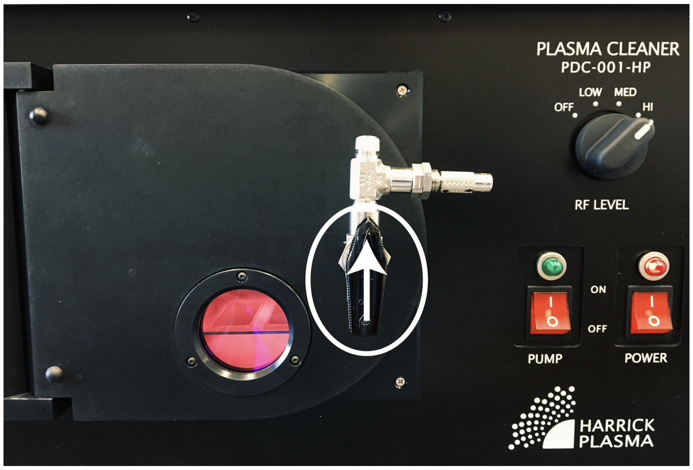


3. Close the plasma cleaner door, then turn on the main power and pump switch. To form a vacuum, ensure that the 3-way valve lever is at the 9:00 position as shown below and that the door is completely shut.



4. Allow vacuum to form for 2-3 minutes. Once the vacuum has formed, simultaneously turn the valve to 12:00 while turning the power to the Hi setting (shown below).

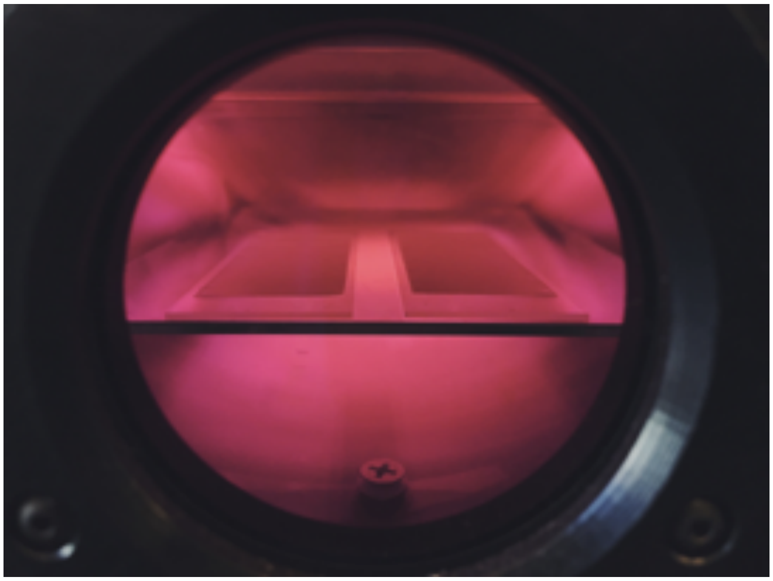
   Note: The plasma should be a bright pink. If not, adjust the air valve to increase or decrease the amount of oxygen entering the chamber.



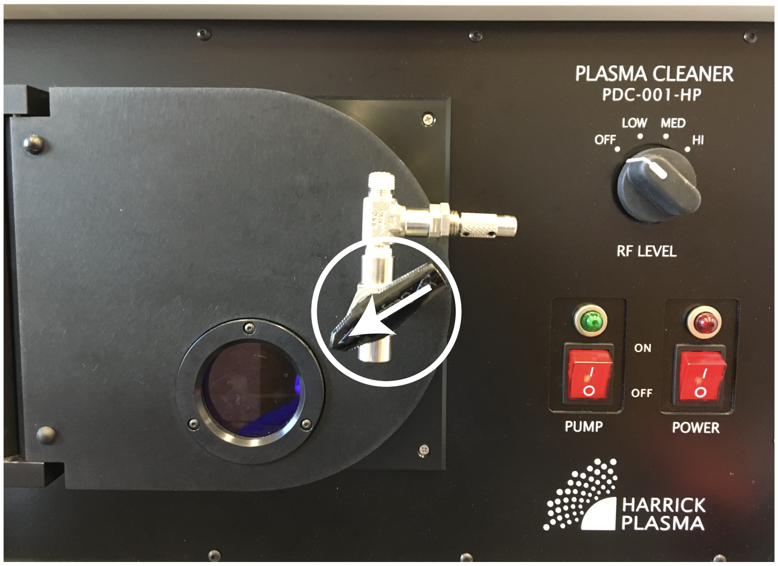
5. Treat membranes with plasma for 5-7 minutes.

   Note: We treat membranes for 7 minutes, but treatment times can vary.

Experiment Notes



6. **Critical** – After treatment, in the following order: **(1)** turn the RF level valve from HIGH to OFF, **(2)** turn the air valve from the 12:00 position to the 9:00 position, and **(3)** then turn off the power followed by turning off the vacuum. Then *slowly open* the valve until air can be heard entering the chamber (approximate valve position shown below). Leave until door opens (~5 min).

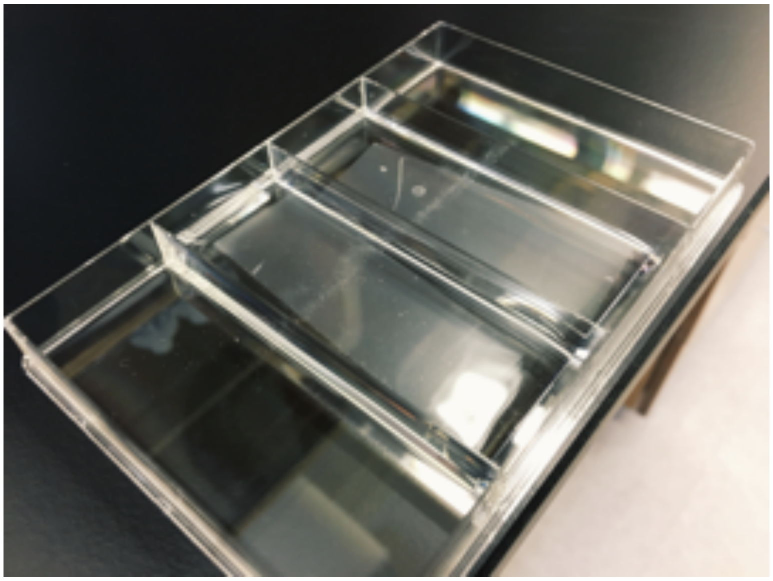


7. Remove slides (with membranes) from the oven and transfer to a 4-well dish.

   Note 1: If membranes have slightly folded over, slowly flip the membrane back using needle nose tweezers.

   Note 2: If membranes have blown off the slide entirely, repeat above procedure to ensure you know which side was exposed to plasma.

8. Using a P1000 pipette, gently hydrate one end of the membrane with a single drop of 1xPBS so that it adheres to the slide before dispensing the entire volume. Once the membrane is hydrated, continuing add 1xPBS until you reach 5 mL (use either a serological pipette or P1000 pipette to complete hydration).



9. Remove any air bubbles underneath the membrane using wafer forceps or a pipette tip.
10. Membranes are now functionalized and ready for use.

    Note 1: Membranes solvated with 1xPBS should be used within **48 hours.**

    Note 2: If transporting solvated membranes (e.g. between buildings), remove all but ~1 mL of 1xPBS to prevent membranes from flipping within the dish.

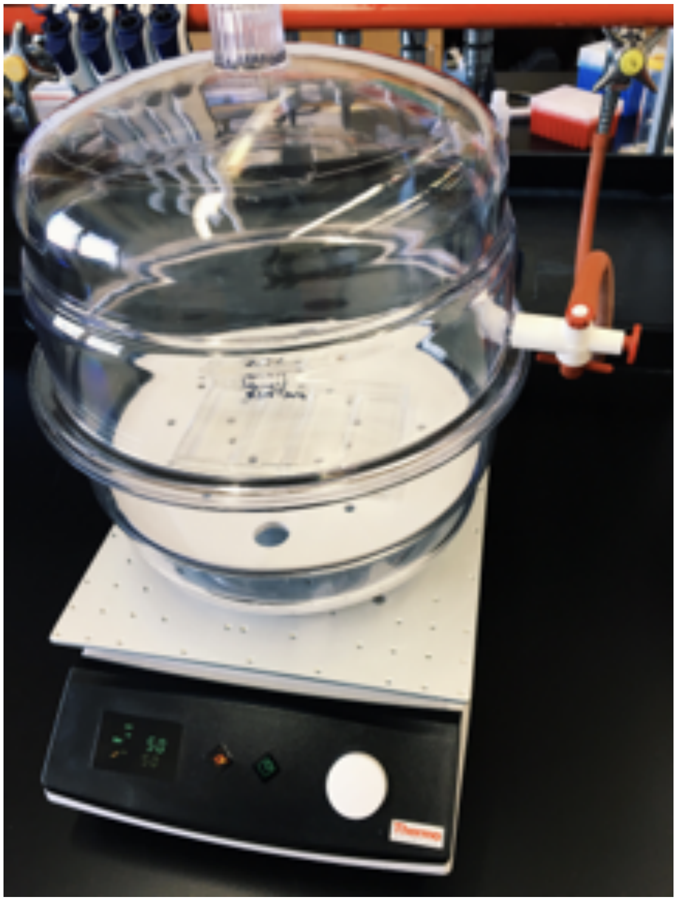
    Note 3: Alternatively, membranes initially solvated in 1xPBS can be dried and stored for 4 weeks at room temperature. To dry them out, carefully remove membranes, keeping them on their glass slides, from the 1xPBS solution, transfer the membranes to the benchtop, cover them with a tip box, and let them dry for 15-20 minutes. As the membranes dry they'll become opaque which is normal.

    Note 4: Before use the membranes should be rehydrated with 5 mL of 1xPBS. Drying out membranes is helpful when traveling or when running seq-well in a laboratory without access to a plasma cleaner.

Experiment Notes

Bead Loading

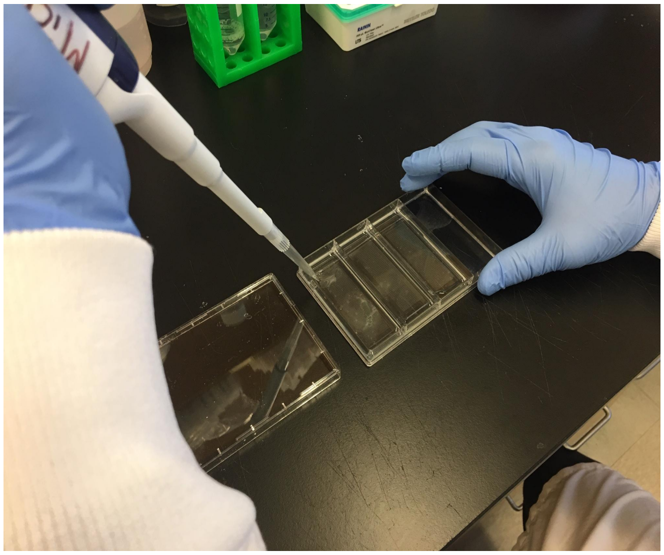
1. Aspirate storage solution and solvate each array with 5 mL of bead loading buffer (BLB; **See Sub-Appendix D: Buffers Guide**).
2. Place array(s) under vacuum with rotation (50 RPM) for 10 minutes to remove air bubbles in wells. **Note:** Rotation is optional



3. Aliquot ~110,000 beads from stock into a 1.5 mL tube and spin on a tabletop centrifuge for 15 seconds to form a pellet.
4. Aspirate storage buffer and wash beads twice in 500 uL of BLB.
5. Pellet beads, aspirate BLB, and resuspend beads in 200 uL of BLB.

   Note: For each array, it's recommended to load ~110,000 beads.
6. Before loading beads, thoroughly aspirate BLB from the dish containing the array(s), being careful not to aspirate or dry the PDMS surface of the array(s).

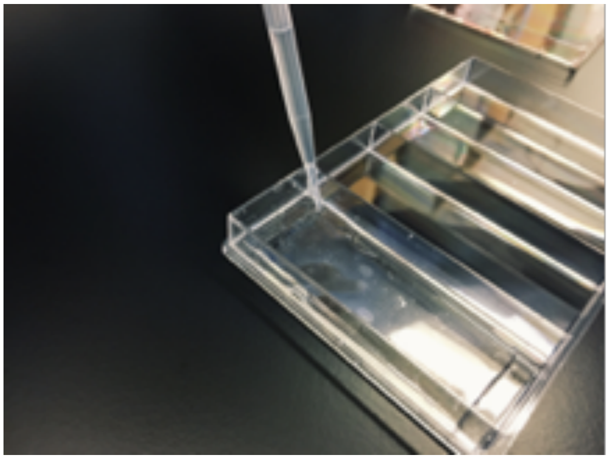
7. Using a P200 pipette, apply 200 uL containing 110,000 beads, in a drop-wise fashion, to the surface of each array (see image below and Bead Loading Diagram on page 10).

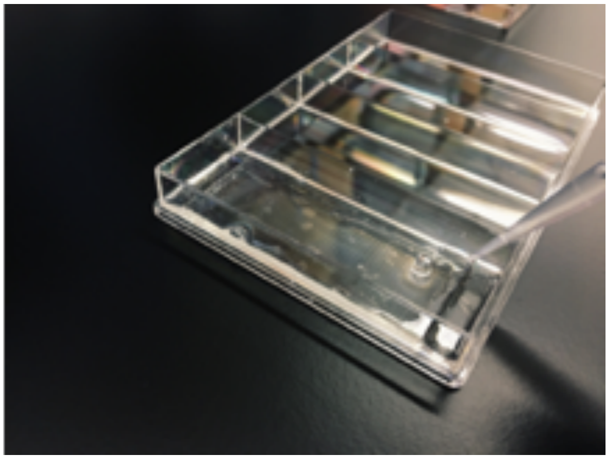


8. Allow the arrays to sit for 5 minutes, rocking them intermittently in the x & y direction.

   Pro-Tip: This step can be extended to 10 minutes to allow the beads more time to settle. However, make sure to monitor the surface of the array so that it doesn't dry out.

9. Thoroughly wash array(s) to remove excess beads from the surface. For each wash:
   1. Position each array so that it sits in the center of the 4-well dish.
   2. Dispense 500 uL of BLB in the upper right corner of each array and 500 uL in the bottom right corner of each array. Be careful not to directly pipette onto the microwells, as it can dislodge beads.
   3. Using wafer forceps or a pipette tip, push each array against the left side of the 4-well dish to create a capillary flow; this will help remove beads from the surface.
   4. Aspirate the liquid, reposition each array, and repeat on the other side.





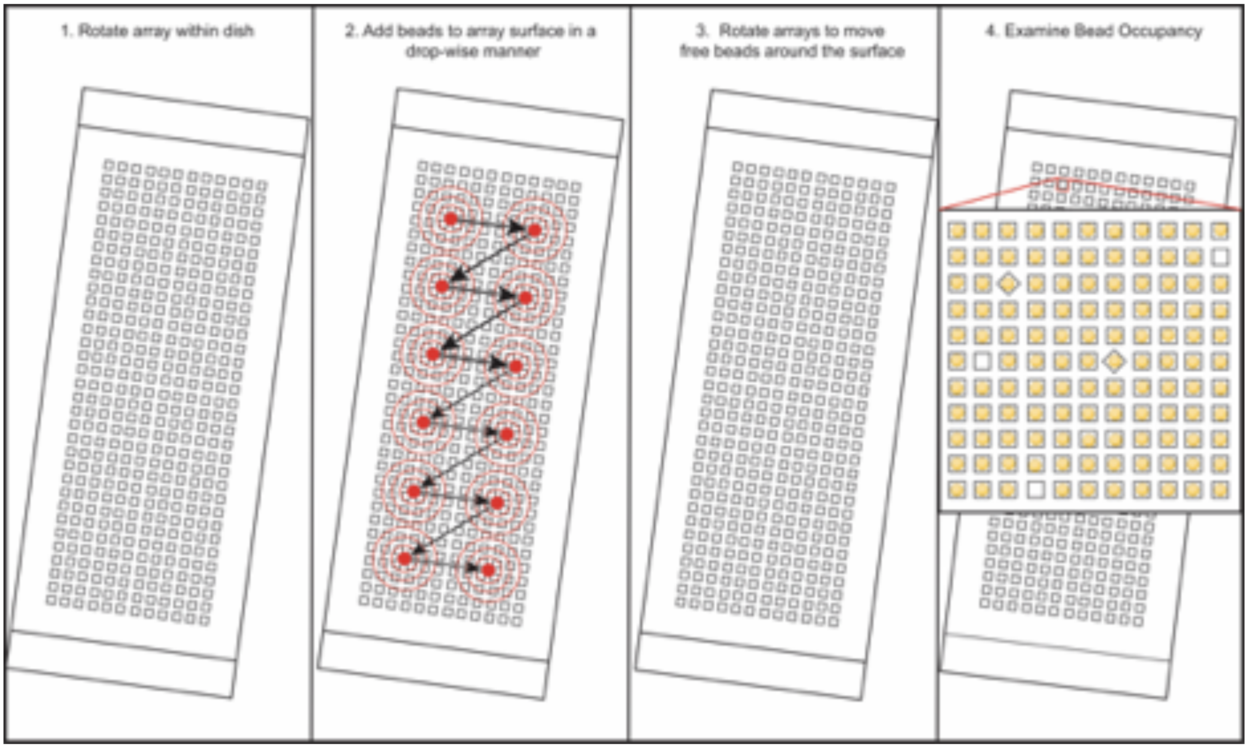
10. Repeat step 9 as necessary. Periodically examine the array(s) under microscope to confirm that no loose beads are present on the surface, as this will interfere with membrane attachment. Usually it takes 4 washes/side to thoroughly remove excess beads (this depends on your original loading density).
11. Once excess beads have been removed from the surface, solvate each array with 5 mL of BLB and proceed to cell loading.

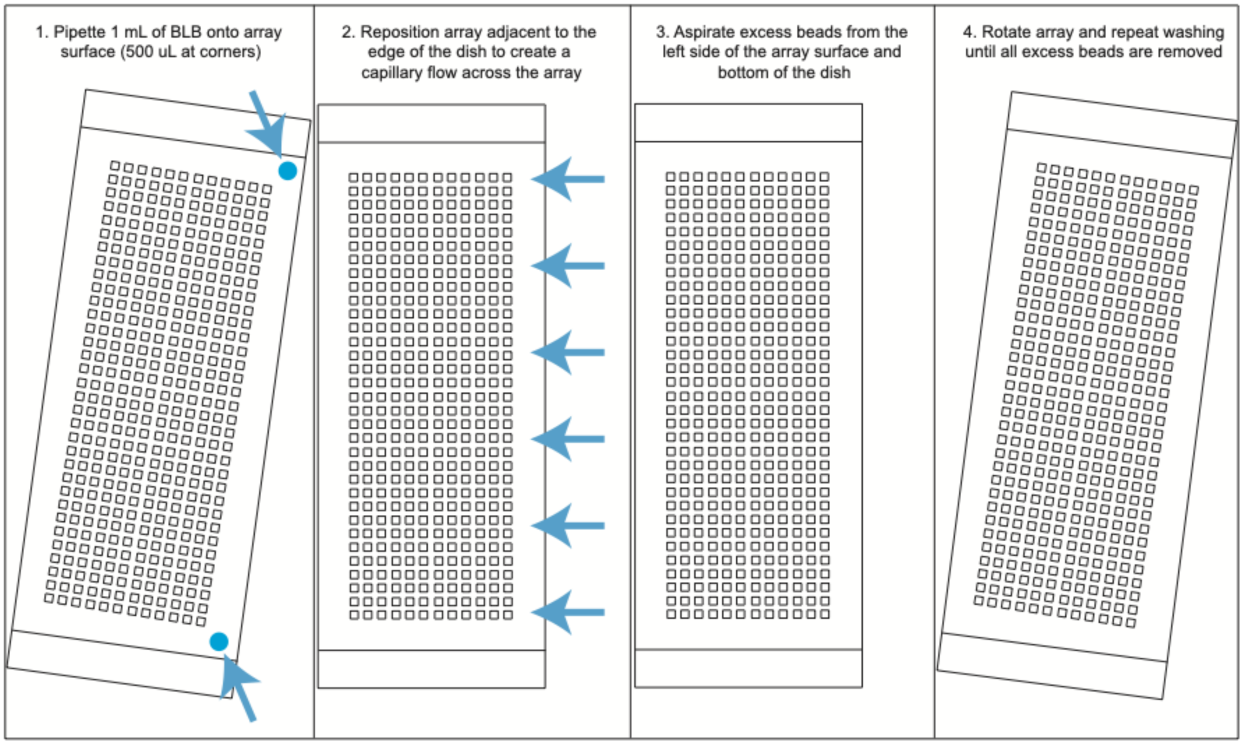
    Notes:

    1. If continuing to cell loading immediately (i.e., within 6 hours), loaded arrays should be stored in 5 mL of BLB.
    2. If you are not going to use the arrays on the day they're loaded, remove the BLB buffer, rinse the arrays once with 5 mL of 1xPBS, and then solvate the arrays with 5 mL of quenching buffer. Arrays can be stored in quenching buffer for 10 days (**See Sub-Appendix D: Buffers Guide**).

Experiment Notes

Bead Loading Diagrams


1. Rotate array within dish
2. Add beads to array surface in a drop-wise manner
3. Rotate arrays to move free beads around the surface
4. Examine Bead Occupancy


1. Pipette 1 mL of BLB onto array surface (500 uL at corners)
2. Reposition array adjacent to the edge of the dish to create a capillary flow across the array
3. Aspirate excess beads from the left side of the array surface and bottom of the dish
4. Rotate array and repeat washing until all excess beads are removed

Cell Loading (Without Imaging)

If you want to image cells in the array, please refer to Sub-Appendix F

1. At this point, your array should be loaded with beads and sitting in 5 mL of BLB.
2. Obtain the cell or tissue sample and prepare a single cell suspension using an optimized protocol for tissue dissociation.
3. While preparing your single-cell suspension, aspirate the BLB from each array (or quenching buffer) and rinse the array twice in 5 mL of 1xPBS to bring the solution in the four-well dish to physiological pH.
4. After the second wash, aspirate the 1xPBS and soak the loaded array in 5 mL of RPMI +10% (RP-10) FBS for 5 minutes.

   Note 1: This step is performed to mitigate non-specific adhesion of cells to primary amines on the top surface of the array.

   Note 2: Any supplemented media can be used in place of RP-10.
5. After obtaining a single-cell suspension, count cells using a hemocytometer and make a new solution of 10,000-15,000 cells in 200 uL of RP-10.

   Note 1: You can use your preferred media for prepping the cell loading solution.

   Note 2: Be sure to not use automated cell counters, particularly following tissue dissociation. This can provide an inaccurate cell count, compromising the experiment.
6. Thoroughly aspirate the RP-10/supplemented media (to ensure the array will not move during cell loading).
7. Center your array in the well and then apply the cell loading solution onto the surface in a dropwise fashion (similar to how beads were applied in the previous section).
8. Allow cells to settle for 10 minutes, intermittently rock the array in the x & y direction.
9. Wash array 4x with 5 mL of 1xPBS to remove the serum. For each wash, gently rock the array in the x & y direction, and then aspirate the 1xPBS. Once you have aspirated the 1xPBS out of the dish, gently tilt the 4-well dish toward you and

aspirate directly off the bottom border of the array; this will help to completely remove the excess serum on the surface of the array

Note: These washes are critical to remove excess serum which can interfere with successful membrane attachment.

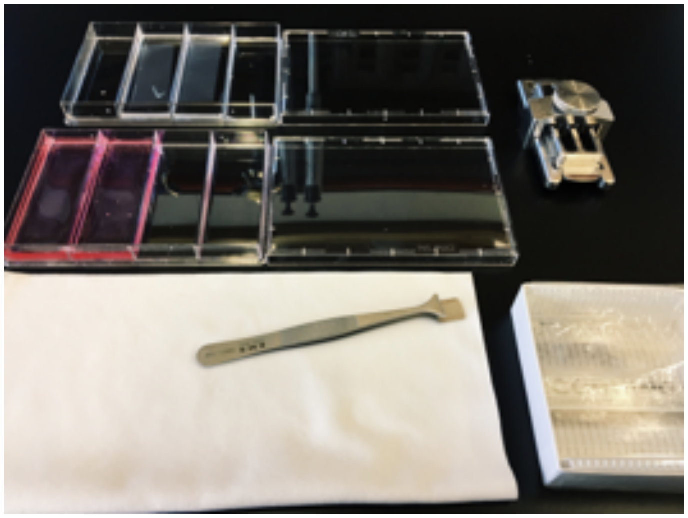
10. Aspirate the final 1xPBS wash and replace with 5 mL of RPMI media ***without*** FBS.

    Note: You can use any media here as long as it **does not** contain serum.

Experiment Notes

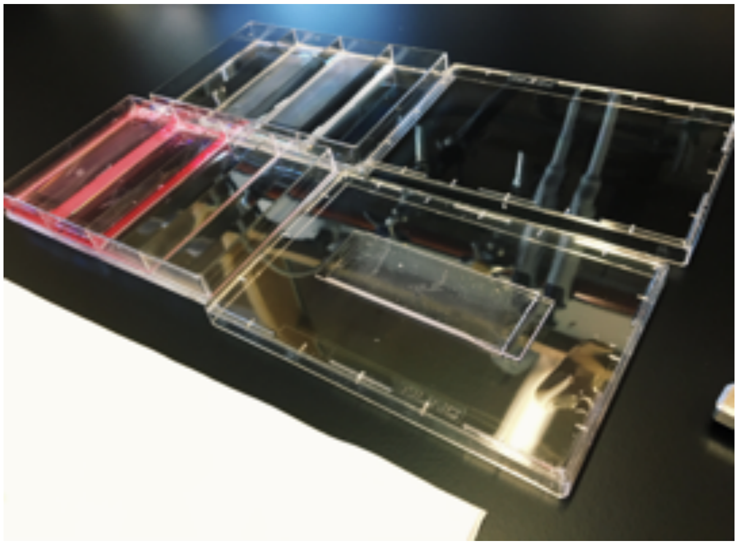
Membrane Sealing

1. Gather the following materials before sealing the array(s):

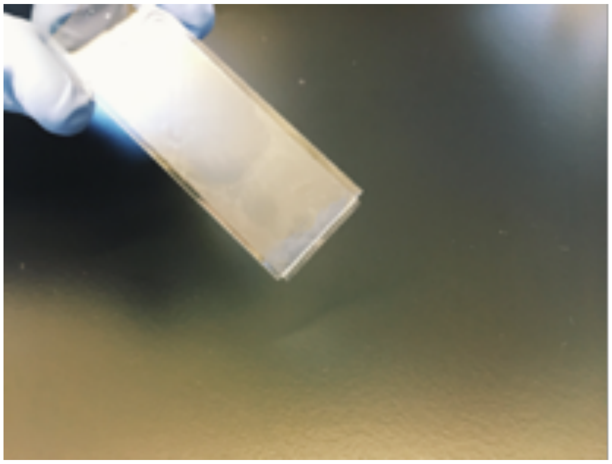


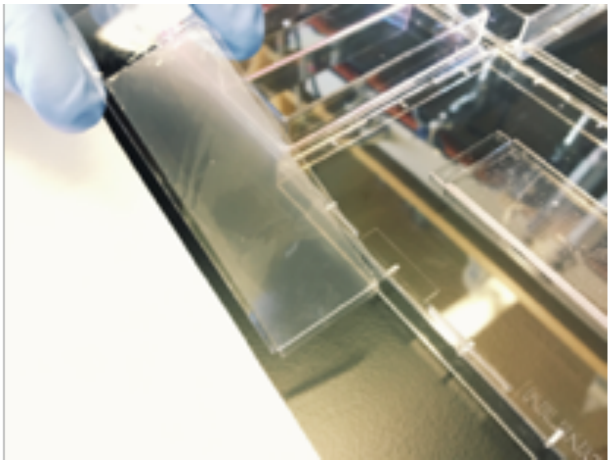
- Array loaded with beads and cells (See Bead/Cell Loading)
- Pre-treated membrane (See Membrane Preparation)
- Wafer forceps (or P1000 pipette tip)
- Paper towels
- Agilent clamp
- Clean microscope slides

2. Use the wafer forceps to transfer the array from media to the lid of a 4-well dish, being careful to ensure that the array is not tilted.

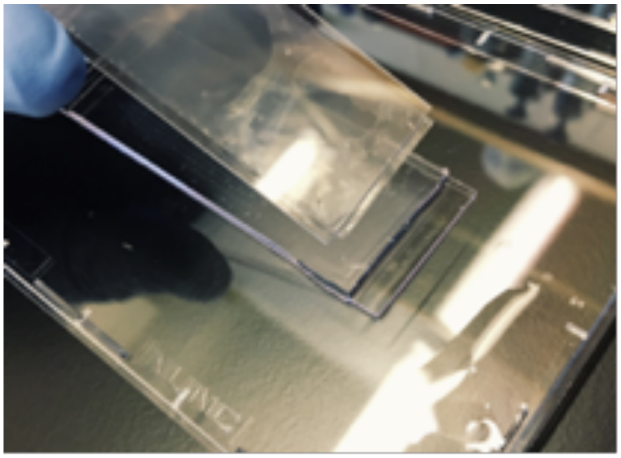


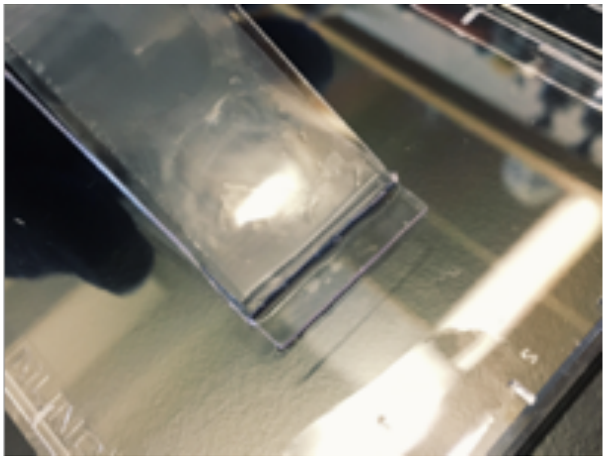
3. Once the array is positioned on the lid of a 4-well dish, carefully aspirate excess liquid from around the edge of the array and the exposed surface of the glass slide. (**Note**: Be careful not to aspirate directly from the PDMS surface).
4. Using wafer forceps or a pipette tip, remove a pre-treated membrane from the 4-well dish.
5. Gently dab away moisture from the glass slide on the paper towel until the membrane does not spontaneously change position on the glass slide.
6. Carefully position the membrane on the center of the microscope slide, leaving a small membrane overhang (2-3 mm) beyond the edge of slide.





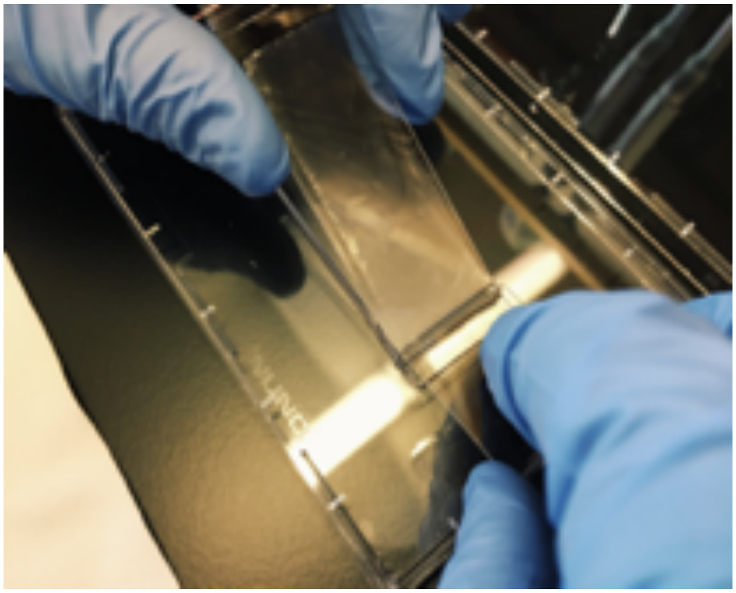
7. Holding the membrane in your left hand, invert the microscope slide so that the treated surface of the membrane is facing down.





Experiment Notes

8. Place the overhang of the membrane in contact with the PDMS surface of the array just beyond the boundary of the microwells.



9. Using a clean slide held in your right hand, firmly hold down the overhang of the membrane against the PDMS surface of the array.

10. **Critical Step:** While maintaining pressure with your right hand to hold the membrane in place, gently apply the membrane.

    Note 1: For optimal results, use only the weight of the slide to apply the membrane with the left hand.

    Note 2: Attempts to manually seal the microwell device using excess pressure result in a 'squeegee' effect, effectively removing moisture from the membrane while fixing membrane creases in place.

    Note 3: As you apply the membrane you should see a fluid interface form and expand as direct, uniform contact between the slide and the array will naturally remove some of the media as the membrane is applied.

    Note 4: You can use either your left or right hand for membrane-sealing (most people use their dominant hand to apply the membrane). Please practice this step before the actual experiment to figure out which hand you're most comfortable with.

11. After applying the membrane, carefully pry the array and membrane from the surface of the lid and transfer to an Agilent clamp.

12. After transferring the sealed array to the clamp, place a glass slide on top of the sealed array.

13. Close the clamp and tighten to the point of resistance, then place it in a 37C incubator for 30-40 minutes.

    Note: This time is flexible and depends on the incubator. If you want to decrease this incubation time, please optimize on cell lines before proceeding with precious samples.

14. Repeat membrane-sealing protocol procedure if running multiple arrays.

Experiment Notes

Cell Lysis & Hybridization

1. Remove the clamp from the incubator, and then remove the array from the Agilent clamp. (Note: At this point, the glass slide will be attached to the array and membrane).
2. Submerge the array, with top slide still attached, in 5 mL of complete lysis buffer (**See Sub-Appendix D: Buffers Guide**).
3. Gently rock the array in lysis buffer until the top glass slide spontaneously detaches.

   Note 1: Do not pry the top slide off as this can reverse membrane sealing. The time necessary for detachment of the top slide varies (10 seconds – 10 minutes).

   Note 2: If the top slide does not release after 10 minutes, gently pry the top slide off using wafer forceps or a pipette tip. Just be careful.
4. Once the top slide has detached, place the arrays on a horizontal rotator for 20 minutes at 50-60 rpm.
5. After 20 minutes, remove the lysis buffer and wash each array with 5 mL of hybridization Buffer (**See Sub-Appendix C: Buffers Guide**).

   Note 1: Use a separate waste container for lysis buffer because guanidine thiocyanate can react with bleach in TC traps to create cyanide gas.

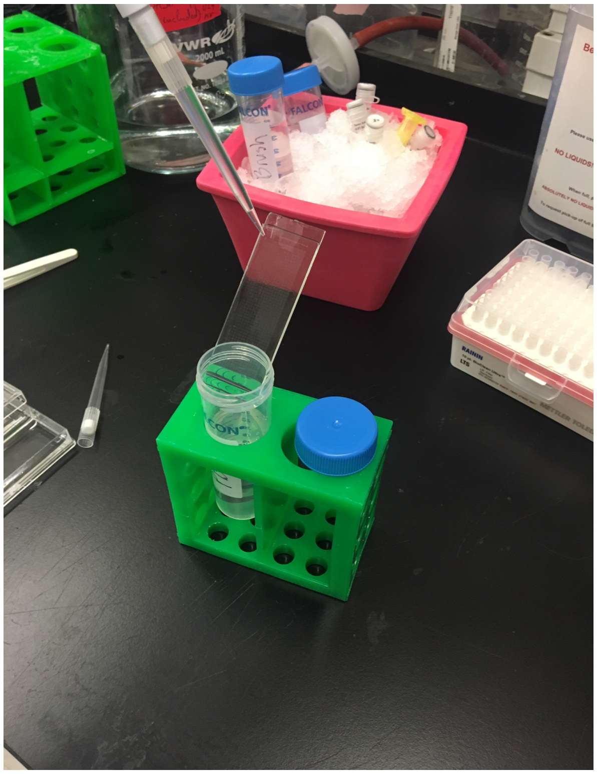
   Note 2: The hybridization buffer used to wash the array post-lysis may contain trace amounts of guanidine thiocyanate and should, therefore, be disposed of in the lysis buffer waste container.
6. Aspirate hybridization buffer and add another 5 mL of hybridization buffer to each array and rotate for 40 minutes at 50-60 rpm.
7. While the arrays are rocking in hybridization buffer, prepare RT master mix. (See **Reverse Transcription & Exonuclease Digestion**)

Experiment Notes

Bead Removal Method 1

1. After the arrays have rocked in hybridization buffer for 40 minutes, carefully peel back each membrane using fine-tipped tweezers.
2. Place array into a 50 mL conical containing 30-40 mL of Wash 1 solution.
3. Holding the array above the 50mL conical (shown below), repeatedly dispense approximately 1 mL of Wash 1 solution from the conical across the surface of the array to dislodge beads (**See Sub-Appendix D: Buffers Guide**).

   Note: Vigorously dispense Wash 1 buffer to remove beads.
4. Repeat these 10 times, periodically checking to see if beads are dislodging.



5. After repeatedly rinsing the array from top to bottom, use a clean glass slide to ***gently*** scrape the array to remove any beads that remain in the array.

   Note: At this point it is possible to visually inspect the array to assess bead removal.
6. Once you are satisfied with bead removal, place the empty array back in the 4-well disk, cap the 50 mL conical, and pellet beads for 5 minutes at 1000xg.

   Note 1: You can visually inspect the success of your bead removal by looking at the arrays under a light microscope. *(continues on the next page)*

Note 2: Where possible, use a swinging bucket centrifuge to collect beads. The use of a fixed-rotor centrifuge can lead to the formation of a bead pellet on the elbow rather than the bottom of the conical tube, which can lead to inefficient recovery.

7. After centrifugation, aspirate all but ~1 mL of excess Wash Buffer, collect the beads using a P1000 pipette, and transfer beads suspended in wash buffer to a separate 1.5 mL eppendorf tube for each array.

Experiment Notes

Reverse Transcription & Exonuclease Digestion

Reverse Transcription (RT)

1. Prepare the following RT mastermix during the hybridization step:

40 uL	H_2O
40 uL	Maxima 5X RT Buffer
80 uL	30% PEG8K
20 uL	10 mM dNTPs (Clontech)
5 uL	RNase Inhibitor (Lucigen)
5 uL	100 uM Template Switch Oligo
10 uL	Maxima H-RT

Note: Add the Maxima H-RT enzyme to the mastermix immediately before adding to beads.

2. Centrifuge eppendorf tubes containing collected beads for 1 minute at 1000xg.
3. Remove supernatant and resuspend in 250 uL of 1X Maxima RT Buffer and centrifuge beads for 1 minute at 1000xg.
4. Aspirate 1X Maxima RT Buffer and resuspend beads in 200 uL of the RT mastermix.
5. Incubate at room temperature for 30 minutes with end-over-end rotation. After 30 minutes, incubate at 52C for 90 minutes with end-over-end rotation.

Note: The reverse transcription reaction can proceed overnight, if necessary.

7. Following the RT reaction, wash beads once with 500 uL of TE-SDS, and twice with 500 uL of TE-Tween (TE-TW). **Following Reverse Transcription, beads can be stored at 4C in TE-TW.**

Exonuclease I Treatment

1. Prepare the following Exonuclease I Mix:

20 uL	10x ExoI Buffer
170 uL	H_2O
10 uL	ExoI

2. Centrifuge beads for 1 minute at 1000xg and aspirate the TE-TW solution.

3. Resuspend in 500 uL of 10 mM Tris-HCl pH 8.0.
4. Centrifuge beads again, remove supernatant and resuspend beads in 200 uL of exonuclease I mix.
5. Incubate at 37C for 50 minutes with end-over-end rotation.
6. Wash the beads once with 500 uL of TE-SDS, twice with 500 uL TE-TW.

   Beads can be stored at 4C in TE-TW.

Experiment Notes

Second Strand Synthesis & PCR

Second Strand Synthesis (*Beginning after 2nd wash of TE-TW after Exo treatment)*

1. Prepare the following 2nd strand synthesis mix:

40 uL	Maxima 5X RT Buffer
80 uL	30% PEG8000
20 uL	10 mM dNTPs (Clontech)
2 uL	1 mM dN-SMRT oligo
5 uL	Klenow Enzyme
53 uL	H_2O

   Note: Add the Klenow enzyme immediately before adding to beads.

2. After aspiration of 2nd TE-TW wash, resuspend beads in 500 uL 0.1 M NaOH.

   Note: Make the 0.1 M NaOH solution fresh for each experiment.

3. Rotate tube for 5 min at room temp, then spin (800xg for 1 minute) and aspirate supernatant.
4. Wash once with 500 uL of TE-TW, and once with 500 uL 1xTE
5. Resuspend beads in 200 uL 2nd strand synthesis reaction and rotate end-over-end at 37C for 1 hr.
6. Wash beads twice with 500 uL TE-Tween and once with 500 uL TE
7. Proceed directly with the PCR protocol.

PCR (Whole Transcriptome Amplification (WTA))

1. Prepare the following PCR mastermix:

25 uL	2X KAPA HiFi Hotstart Readymix
14.6 uL	H_2O
0.4 uL	100 uM SMART PCR Primer
40 uL	per reaction

2. Wash beads once with 500 uL of water, pellet beads, remove supernatant and resuspend in 500 uL of water.

   Note 1: If you do not want to count the beads then after the 500 uL water wash in step 2, resuspend the beads in 240 uL of water and proceed to step 6.

Note 2: If you choose this path, prepare mastermix for 24 PCR reactions for each array being processed.

3. Mix well (do not vortex) to evenly resuspend beads and transfer 20 uL of beads to a separate 1.5 mL tube to count the beads.

   Note: Don’t vortex beads as this can result in bead fragmentation.

4. Pellet the small aliquot of beads, aspirate the supernatant, and resuspend in 20 uL of bead counting solution (10% PEG, 2.5 M NaCl).

   Note: The bead counting solution aids in even dispersion of beads across a hemocytometer.

5. Count the beads using a hemocytometer.
6. Add 40 uL of PCR mastermix per reaction to 96-well plate.
7. Add 1,500 – 2,000 beads per reaction in 10 uL of water for a total volume of 50 uL per PCR reaction, making certain to PCR the entire array.
8. Use the following cycling conditions to perform whole-transcriptome amplification:

Start:	
95C	3 minutes
4 Cycles:	
98C	20 seconds
65C	45 seconds
72C	3 minutes
9-12 Cycles:	
98C	20 seconds
67C	20 seconds
72C	3 minutes
Final Extension:	
72C	5 minutes
4 C	Infinite hold

   Note: The total number of PCR cycles necessary for amplification depends on the cell type used.

   - 13 cycles are optimal for cell lines or larger cells (e.g. macrophages)
   - 16 cycles are optimal for primary cells

Purification of PCR products and analysis on the BioAnalyzer or Agilent TapeStation

1. Pool PCR products from between 6 and 8 PCR reactions in a 1.5 mL microcentrifuge tube so that you have 10-12,000 beads/1.5 mL microcentrifuge tube.
2. Purify PCR products using Ampure SPRI beads and the following protocol:

   Note: Please refer to the Ampure SPRI bead official protocol for more details.

   A. Spri at 0.6x volumetric ratio.

   B. Allow the tubes to sit on the tube-rack off the magnet for 5 minutes, and then place
   the rack on the magnet for 5 minutes.

   C. Perform 3 washes with 80% ethanol (**Note:** At each wash step rotate each tube 180 degrees 6 times to allow beads)
   to pass through the ethanol solution to the opposite side of the tube.

   D. After the third wash, remove the 80% ethanol wash solution. Further, use a P200 with fresh tips to remove any residual ethanol and allow beads to dry for 10-15 minutes. (**Note**: Beads will have a cracked appearance once dry). Remove the rack from the magnet, elute dried beads in 100 uL, place the rack on the magnet and then transfer the 100 uL supernatant which contains eluted DNA to a new 1.5 mL microcentrifuge tube or 96-well plate.

   E. Spri the 100 uL at 1.0x volumetric ratio and repeat steps b and c

   F. After the third wash, allow the beads to dry for 15 minutes, remove the rack from the magnetic, elute the beads in 15 uL, place the rack back on the magnet and then transfer the 15 uL to a new 1.5 mL microcentrifuge tube or 96-well plate.
3. Run a BioAnalyzer High Sensitivity Chip or Agilent D5000 High Sensitivity Screentape according to the manufacturer's instructions. Use 2 uL of the purified cDNA sample as input (**Note:** Your WTA library should be fairly smooth, with an average bp size of 0.7-2 kbps).
4. Proceed to library preparation or store the WTA product at 4C (short-term) or -20C (long-term).

Library Preparation

Tagmentation of cDNA with Nextera XT

1. Ensure your thermocyclers are setup for Tagmentation (step 5) & PCR (step 9).
2. For each sample, combine 1000 pg of purified cDNA with water in a total volume of 5 uL. It's ideal to dilute your PCR product in a separate tube/plate so that you can add 5 uL of that for tagmentation.

   Example: For 1000 pg reactions, dilute PCR product, in a new plate, to 200 pg/uL, then you can add 5 uL of this to a reaction tube for a 1000 pg reaction.

   Note 1: We typically perform Nextera reactions in duplicate for WTA product from each pool of 6-8 PCR reactions. For example, if you recover 3 pools/array, you would run a total of 6 nextera reactions.

   Note 2: These volumes can be reduced by half to reduce reagent costs, if desired.
3. To each tube, add 11 uL of Nextera TD buffer, then 4 uL of ATM buffer (the total volume of the reaction is now 20 uL).
4. Mix by pipetting ~5 times. Centrifuge plate at 1000x g for 10-15 seconds.
5. Incubate at 55C for 5 minutes.
6. Add 5 uL of Neutralization Buffer. Mix by pipetting ~5 times. **Note:** Bubbles are normal.
7. Incubate at room temperature for 5 minutes.
8. Add to each PCR tube:

15 uL	Nextera PCR mix
8 uL	H_2O
1 uL	10 uM New-P5-SMART PCR hybrid oligo
1 uL	10uM Nextera N7XX oligo

(continues on the next page)

9. After sealing and centrifuging (1 minute at 1000xg) the PCR plate, run the following PCR program:

Start:	
72C	3 minutes
95C	30 seconds
12 cycles:	
95C	10 seconds
55C	30 seconds
72C	30 seconds
Final Extension:	
72C	5 minutes
4C	Infinite hold

Purification of PCR products and analysis on the BioAnalyzer or Agilent TapeStation

1. If you performed Nextera reactions in duplicates, please pool duplicates before proceeding with step 2. If you ran a single Nextera reaction for each pooled WTA, proceed directly to step 2.
2. Purify PCR products using Ampure SPRI beads and the following protocol:

   Note: Please refer to the Ampure SPRI bead official protocol for more details.

   A. Spri at 0.6x volumetric ratio.

   B. Allow the tubes to sit on the tube-rack off the magnet for 5 minutes, and then place the rack on the magnet for 5 minutes.

   C. Perform 3 washes with 80% ethanol (**Note:** At each wash step rotate each tube 180 degrees 6 times to allow beads).
   to pass through the ethanol solution to the opposite side of the tube.

   D. After the third wash, remove the 80% ethanol wash solution. Further, use a P200 with fresh tips to remove any residual ethanol and allow beads to dry for 10-15 minutes. (**Note**: Beads will have a cracked appearance once dry).

   (continues on the next page)

Remove the rack from the magnet, elute dried beads in 100 uL, place the rack on the magnet and then transfer the 100 uL supernatant which contains eluted DNA to a new 1.5 mL microcentrifuge tube or 96-well plate.

E. Spri the 100 uL at 1.0x volumetric ratio and repeat steps b and c

F. After the third wash, allow the beads to dry for 15 minutes, remove the rack from the magnetic, elute the beads in 15 uL, place the rack back on the magnet and then transfer the 15 uL to a new 1.5 mL microcentrifuge tube or 96-well plate.

3. Run a BioAnalyzer High Sensitivity Chip or Agilent D1000 High Sensitivity Screentape according to the manufacturer's instructions.

- Use 1 uL of the purified cDNA sample as input.
- Your tagmented library should be fairly smooth, with an average bp size of 400-800 bp.
- Smaller-sized libraries might have more polyA reads
- Larger libraries may have lower sequence cluster density and cluster quality.

Note: We have successfully sequenced libraries from 400-800bp.

5. Proceed to sequencing.

Sequencing

Once your sequencing library has passed the proper quality controls, you're ready to proceed to sequencing. For a detailed loading protocol, please consult the Illumina website for a step-by-step manual. (https://support.illumina.com/downloads.html)

NextSeq500 – Shalek Lab protocol

1. Make a 5 uL library pool at 4 nM as input for denaturation.
2. To this 5 uL library, add 5 uL of 0.2 N NaOH (make this solution fresh).
3. Flick to mix, then spin down and let tube sit for 5 minutes at room temperature.
4. After 5 minutes, add 5 uL of 0.2 M Tris-HCl pH 7.5.
5. Add 985 uL of HT1 Buffer to make a 1 mL, 20 pM library (solution 1).
6. In a new tube (solution 2), add 165 uL of solution 1 and dilute to 1.5 mL with HT1 buffer to make a 2.2 pM solution – this is the recommended loading concentration.

   Note: Optimal loading concentration is 1.8-2.5 pM
7. Follow Illumina's guide for loading a NextSeq500 Kit

Sequencing specifications for the MiSeq or NextSeq:

Read 1: 20 bp *

Read 2: 50 bp

Read 1 Index: 8 bp ← *only necessary if you are multiplexing samples*

Custom Read 1 primer

Sequencing specifications for the Nova-Seq:

Read 1: 20 bp *

Read 2: 50-80 bp

Read 1 Index: 8 bp

Read 2 Index: 8 bp (optional, but recommended)

Custom Read 1 primer

Note 1: If you’re loading on a Nova-Seq you’ll want to use dual-indexing to mitigate index switching.

Note 2: Read 1 can sometimes be 21 base pairs; this depends on the company and bead lot you are ordering from. Please consult with your bead provider to determine which read length to use.

NextSeq 500:
(http://support.illumina.com/content/dam/illumina-support/documents/documentation/system_documentation/nextseq/nextseq-custom-primers-guide-15057456-01.pdf)
(Follow Illumina’s guide for custom primers)
MiSeq:
(http://support.illumina.com/content/dam/illumina-support/documents/documentation/system_documentation/miseq/miseq-system-custom-primers-guide-15041638-01.pdf)

Sub-Appendix A: Array Synthesis

Day 0: Pouring PDMS Arrays

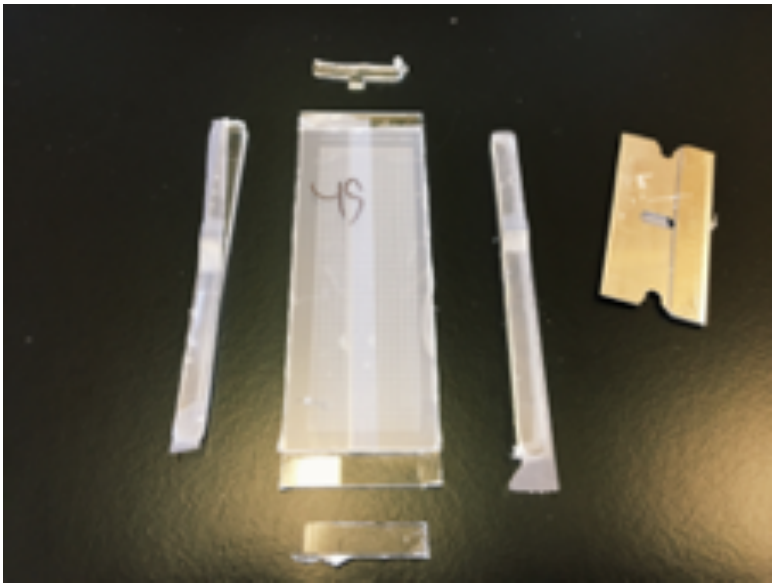
Note: If you need to mount your master, please refer to **appendix F**

1. Combine Sylgard crosslinker with Sylgard base at a 1:10 ratio and mix vigorously for 5 minutes to create a PDMS master mix.
2. Once mixing is complete, put your PDMS master mix under vacuum for 20 minutes to remove any air bubbles.
3. Use a 10 mL syringe to inject 6-10 mL of PDMS master mix into molds with mounted PDMS masters.
4. Incubate at 70C for 2.5 hours.

Day 1: Array Functionalization Part 1

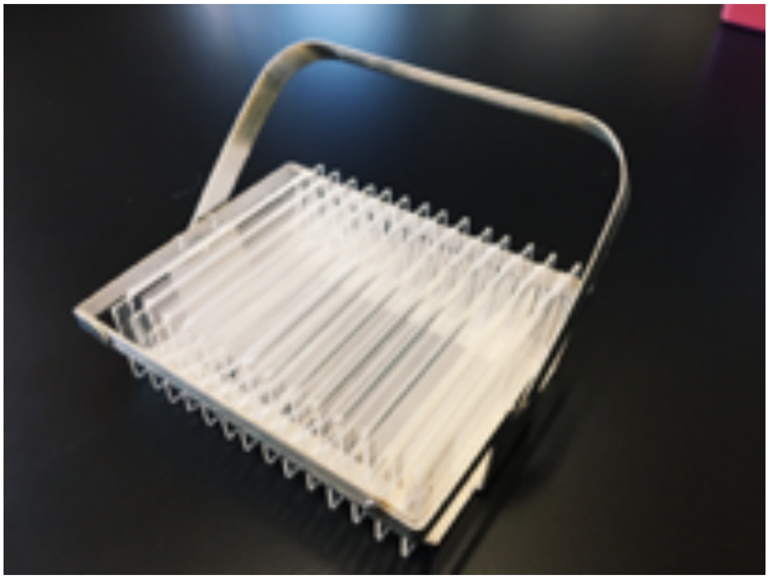
Note: For this section, make all solutions fresh!

1. Remove excess PDMS from edges of the glass slide.

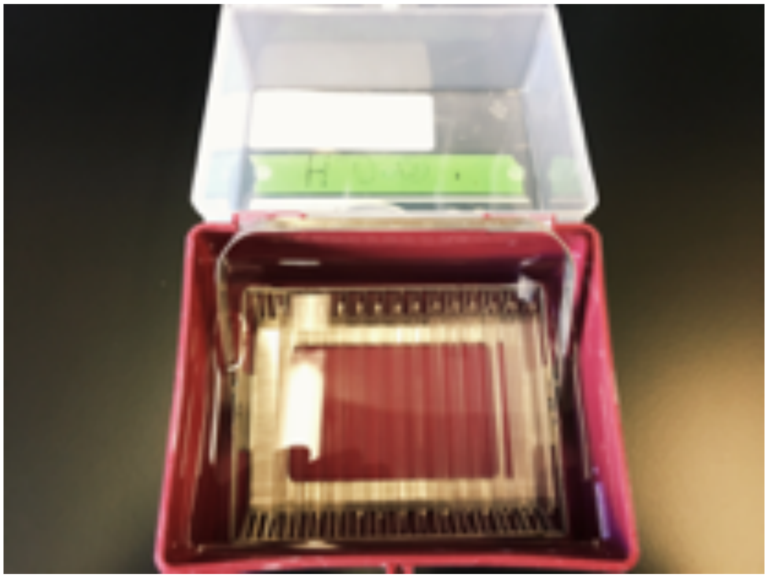


2. Use scotch tape to remove excess PDMS from the surface of the array and the glass slide.
3. Place clean arrays into a metal slide basket

(continues on the next page)



4. Rinse arrays in 100% ethanol for 5 minutes, then let dry at room temperature (RT) for 15 minutes.



5. Plasma treat arrays on high for 5-7 minutes.

   Note 1: Adjust the air valve so that the plasma is pink.

6. Following plasma treatment, immediately submerge arrays in 350 mL of 0.05% APTES in 95% ethanol for 15 minutes.

7. Spin dry arrays **(500 RPM** for 1 minute).

   Note: Our rotor model is TX-10000 75003017 (Thermo) with a rotor radius of 209 mm. 500 RPM on this instrument is ~ 60xg.

8. Incubate at room temperature for 10 minutes.

9. Submerge in 300 mL of acetone and rock until all bubbles are out of the wells; this typically takes approximately 5 minutes.

10. Place in 350 mL of 0.2% PDITC/10% pyridine/90% DMF solution in a glass chamber (or polypropylene tip box) for 2 hours at room temperature.
    Note: While this is rocking, prepare your chitosan solution (See **Sub-Appendix D**)

11. After the PDITC soak, wash arrays briefly in two boxes of 300 mL DMF.
    Note: For each brief wash, simply dunk the arrays in the solution 5-10 times and then transfer to the new solution.

12. Dunk and wash the arrays in 300 mL of acetone.

13. Move to a fresh 350 mL of acetone and rock for 20 minutes.

14. Spin dry arrays (**500 RPM** for 1 minute).

15. Place arrays at 70C for 2 hours.

16. Remove from oven and let sit at room temperature for 20 minutes.

17. Submerge arrays in 350 mL of 0.2% chitosan solution (pH 6.0-6.1; See **Sub-Appendix D**) and incubate at 37C for 1.5 hours.

18. Wash arrays 4x in separate 300 mL distilled water baths.

19. Submerge in 350 mL of 20 ug/mL aspartic acid, 2 M NaCl, and 100 mM sodium carbonate solution (pH 10.0).
20. Place in vacuum chamber and apply house vacuum.

    Note: You should see bubbles form indicating the solvation of wells.
21. Place vacuum chamber (still connected to house vacuum) on a rocker and rock (50-70 RPM) overnight at room temperature.

Day 2: Array Functionalization – Part 2

1. The following morning, remove arrays from vacuum and rotate at 50-60 RPM for 3 hours at room temperature.
2. Place arrays at 4C and soak 24 hours before use.

   Note: Arrays can be stored in the aspartic acid solution for 3 months at 4C.

EXPERIMENTAL NOTES

Sub-Appendix B: Synthesis Protocol Checklist

Date:

Synthesizer:

Number of Arrays:

Start time / End time:

Before you start:

1. Pull the PDITC from the fridge (this takes ~1hr to come to room temperature)
2. Make certain you have enough boxes for the various incubations
3. Clean a 1L bottle, add stir bar, and dissolve 1 gram of chitosan in 500mL of DI water.

Step 1: Plasma treatment of the arrays

1. Soak the arrays in 300mL of 95% ethanol for 5 minutes (50 rpm)
2. Dry the arrays for 5 minutes @ 500 rpm (60xg)
3. Plasma treat **two** trays at a time
   A. Form seal for **3 minutes**
   B. Plasma treat for **5-7 minutes**.
   C. What color was the plasma? (circle one):
   No Color Light purple Light pink
   D. While the arrays are being treated, prep the APTES solution
   APTES Solution: 180uL of APTES stock in 350mL of 95% ethanol
4. Proceed with protocol

Step 2: PDITC Soak

**Autoclave the chitosan solution after starting the PDITC incubation

1. Volume of PDITC solution you're prepping:___________liters (Standard: 350 mL)
2. Mass of PDITC added:________________________grams (Standard: 0.72 grams)
3. Volume of pyridine added:__________________________liters (Standard: 35 mL)
4. Length of incubation:_____________________________hours (Standard: 2 hours)
5. Number of DMF washes:______________________________(Standard: 2 washes)
6. Number of acetone washes:__________(Standard: 1 wash, and then transfer to new acetone box for a **20-minute** soak)
7. Proceed with protocol

Step 3: Oven incubation and chitosan preparation

1. After **20-minute soak**, remove arrays from acetone and spin down (1 min. @ 500 rpm)
2. Length of 70C incubation:______________________________hrs (Standard: 2 hrs)
3. Chitosan protocol checklist: Did you… (Y / N responses)
    - A. Autoclave chitosan (40-minute sterilization, 20-minute drying):___________
    - B. Let solution come to room temperature (can also do this @ 4C):_________
    - C. Calibrate the pH Meter with appropriate buffers:______________________
    - D. Add 4 mL of glacial acetic acid (solution should be on stir plate):_________
    - E. Let solution stand for **5 minutes**:___________________________________
    - F. Add 50 mL of 5 M NaCl:____
    - G. Titrate with 5 M NaOH:____
    - H. Achieve pH of 6.0 – 6.1:____
    - I. Remember, it is **critical** to make certain the you achieve a pH of 6.0 – 6.1 and that it holds. Parts A-H should be completed **before** the completion of the 2 hr 70C incubation.
4. Length of room temperature incubation:_________minutes (Standard: 20 minutes)
    - A. Second check of chitosan pH:
5. Length of chitosan incubation:___________________________(Standard: 1.5 hrs)
    - A. Temperature = 37C, Rotation = 70 rpm

Step 4: In-well functionalization

1. Number of DI water rinses:___________________________________(Standard: 4)
2. pH of aspartic acid solution:______________________________(Standard: pH 10)
3. Length of overnight incubation:_________________________(Standard: 12-16 hrs)
4. Day 2: length of room temperature incubation @ 50 rpm:______(Standard: 4 hrs)

Array Lot: <Your initials>_<Synthesis Date>_<Box Number>

Sub-Appendix C: Master Mounting Protocol

1. Mix and degas PDMS in normal 1:10 ratio
2. While PDMS degases, use sandpaper to gently score back of silicon master and base plate to improve adhesion. Careful – silicon masters are brittle.
3. Wash back of master and base plate with 95% ethanol until no more dust is removed when wiping surface clean with paper towel.
4. Use gloved finger to spread vacuum grease on bottom of BasePlate2 around square holes where the nanowell arrays will be cast. You want a relatively thick layer, even on skinny parts between array holes, to make sure there is a seal between master and plate.
5. Carefully lower BasePlate2 onto the array side of the master making sure to not touch the array area with any of the greased surface. 4 array masters should fit into the 4 square holes. Gently slide plate against master to center the arrays.
6. Place Base Plate 1 on paper towels to catch PDMS running off plate.
7. Pour ~30 mL of mixed PDMS in center of Base Plate 1.
8. Place master/BasePlate2 sandwich on top of the PDMS.
9. Gently apply pressure in the center of the master while making circular motions to push PDMS out from between layers. You want to see PDMS coming out of all sides to ensure a complete coat.
10. Screw 6/32 screws into respective holes on base plate very gently. Too much pressure too fast may crack master. Do not fully tighten. Do your best to make screws even – look at width of crack between base plates on all sides and make equal.
11. Place both top plates on top.
12. Screw 10/24 screws into their holes just enough such that they catch. Again, do not fully tighten.
13. Place in 90C oven for 3 hours.
14. May need to do one dummy round of arrays to remove any PDMS or grease that got onto the nanowell features.

Sub-Appendix D: Buffers Guide

CellCover10

Reagents

- CellCover (Anacyte Art. No. 800-125)
- FBS (Thermo Fisher Scientific Cat. No. 10437028)
- Sodium Carbonate (Sigma Cat. No. 223530-500G)

Working Concentrations

- 10% FBS
- 100 mM Sodium Carbonate

Bead Loading Buffer

Reagents

- Sodium Carbonate (Sigma Cat No. 223530-500G)
- BSA (Sigma Cat No. A9418-100G)
- Water (Thermo Fisher Scientific Cat No. 10977023)

Quick Preparation Guide (50 mL)

1. 2.5 mL 2 M Sodium Carbonate
2. 42.5 mL H_2O
3. Add 5 mL BSA (100 mg/mL)
4. Titrate with glacial acetic acid to achieve a pH of 10.0

Working Concentrations

- 100 mM Sodium Carbonate
- 10% BSA

Complete Lysis Buffer

Reagents

- Pre-lysis buffer
- 10% Sarkosyl (Sigma Cat No. L7414)
- 100% 2-Mercaptoethanol (Sigma Cat No. M3148-25ML)

(continues on the next page)

Quick Preparation Guide (50 mL)

1. 47.25 mL Pre-Lysis Buffer
2. 250 uL 10% Sarkosyl
3. 500 uL BME

Working Concentrations

- 5 M Guanidine Thiocyanate
- 1 mM EDTA
- 0.50% Sarkosyl
- 1.0% BME

Hybridization Buffer

Reagents

- 5 M NaCl (Thermo Fisher Scientific Cat No. 24740011)
- 1x PBS (Thermo Fisher Scientific Cat No. 10010023)
- 8% (v/v) PEG8000 (Sigma Cat No. 83271-500ML-F)

Quick Preparation Guide (50 mL)

1. 20 mL 5 M NaCl
2. 26 mL of PBS
3. 4 mL PEG8000

Working Concentrations

- 2 M NaCl

Wash Buffer

Reagents

- 5 M NaCl (Thermo Fisher Scientific Cat No. 24740011)
- 1 M $MgCl_2$ (Sigma Cat No.63069-100ML)
- 1 M Tris-HCl pH 8.0 (Thermo Fisher Scientific Cat No. 15568025)
- Water (Thermo Fisher Scientific Cat No. 10977023)
- 8% (v/v) PEG8000 (Sigma Cat No. 83271-500ML-F)

(continues on the next page)

Quick Preparation Guide (50 mL)

1. 20 mL 5 M NaCl
2. 150 uL 1 M $MgCl_2$
3. 1 mL 1 M Tris-HCl pH 8.0
4. 24.85 mL H_2O
5. 4 mL PEG8000

Working Concentrations

- 2 M NaCl
- 3 mM $MgCl_2$
- 20 mM Tris-HCl pH 8.0

Array Quenching Buffers

Reagents

- Sodium Carbonate (Sigma Cat No. 223530-500G)
- 1 M Tris-HCl pH 8.0 (Thermo Fisher Scientific Cat No. 15568025)
- Water (Thermo Fisher Scientific Cat No. 10977023)

Quick Preparation Guide (50 mL)

1. 2.5 mL 2 M Sodium Carbonate
2. 500 uL 1 M Tris-HCl pH 8.0
3. 47 mL H_2O

Working Concentrations

- 100 mM Sodium Carbonate
- 10 mM Tris-HCl pH 8.0

0.2% Chitosan Solution

Reagents

- Chitosan (Sigma Cat No. C3646-100G)
- Water (Thermo Fisher Scientific Cat No. 10977023)

(continues on the next page)

Quick Preparation Guide

1. Add 1 gram of chitosan to 500 mL of DI water
2. Autoclave solution (40 minutes sterilization, 20 minutes dry)
3. Allow chitosan solution to come to room temperature, and then add 2-3 mL of glacial acetic acid.

Note: The chitosan will not start dissolving until the pH is acidic, and even then it will not fully dissolve. This is ok.

4. Add 50 mL 5 M NaCl, then titrate the chitosan solution with NaOH to bring the pH to 6.2.

TE - Tween Storage Solution

- 10 mM Tris pH 8.0 + 1 mM EDTA
- 0.01% Tween-20

Quick Preparation Guide (50 mL)

1. 49.95 mL H_2O
2. 5 uL Tween-20

TE - SDS Solution

- 10 mM Tris pH 8.0 + 1 mM EDTA
- 0.5% SDS

Quick Preparation Guide (50 mL)

1. 49.75 mL H_2O
2. 250 uL SDS

Sub-Appendix E: Bead Removal Method 2 ("Spin-Out")

1. Remove membrane and place array into an empty 50 mL conical tube.
2. Ensure that the array is angled within the tube as shown below.

   Note: The array might move around at this point, which isn't something to worry about.
3. Add 48-50 mL of Wash 1 solution (See Buffers Guide)
4. Place the insert so the array is secured angled as shown in the image below.
5. Secure the lid and seal with parafilm, if necessary.
6. Put the sealed conical in a centrifuge, making certain the PDMS surface of the array is facing away from the rotor arm (See Diagram Below).
7. Centrifuge at 2000 x g for 5 minutes to remove the beads.
8. At this point you should see a small, but visible, pellet of beads at the bottom of the tube.
9. Aspirate 5 - 10 mL of Wash 1 solution to enable easier removal of the array.
10. Remove the array and carefully position it over the top of the 50 mL tube.
11. Repeatedly wash any remaining beads from the surface of the array over the surface of the 50 mL falcon tube using 1 mL of Wash 1 remaining in the tube.
12. Spin again at 2000 x g for 5 minutes to pellet beads.
13. Aspirate all wash 1 solution except for ~ 1mL.

    Note: Be careful to not disturb the pellet of beads.
14. Transfer beads to a 1.5 mL centrifuge tube and proceed to reverse transcription.

Sub-Appendix F: Imaging in Array

1. When pre-imaging cells, cells should be loaded first as beads will obstruct view of many cells and beads autofluorescence can interfere with the signal.
2. Obtain a cell or tissue sample and prepare a single cell suspension using your preferred protocol.
3. Count cells using a hemocytometer and resuspend 10,000 cells in 200 uL of cold CellCover (Anacyte).
4. Incubate cells in at 4C for 1 hour.
5. After the cells have been fixed, perform antibody staining at 4C.

   Note: Some epitopes may no longer be available as a result of the fixation process.
6. Wash cells twice with 1x PBS, resuspend in 200 uL of CellCover10 buffer (pH 10 + 10% FBS; **See Sub-Appendix D: Buffers Guide**), and place on ice.

   Note: CellCover != CellCover10.
7. Obtain empty functionalized array(s), aspirate storage solution and soak each array in 5 mL of CellCover10 buffer (**See Sub-Appendix D: Buffers Guide**).
8. Aspirate media and load your fixed cells onto each array in a dropwise format.
9. Gently rock the array(s) in the x & y direction for 5 minutes.
10. Wash each array twice with 5 mL of CellCover10 (pH 10 + 10% FBS), then solvate each them in 5 mL of CellCover (No FBS).
11. Place a lift slip on each array, then image with a microscope.
12. After imaging, wash each array in 5 mL of CellCover10 media.
13. Immediately load beads using the bead loading protocol provided above.

    Note: In the protocol provided above, beads are washed and loaded in BLB. When loading cells first, you will replace BLB with CellCover10 for all steps. After beads are loaded and sufficiently washed, you will wash the array 4x with CellCover10 without FBS and solvate arrays in CellCover.
14. Proceed with membrane sealing.

Sub-Appendix G: Shopping List

Device Manufacturing

- Dow Corning Sylgard 184 Silicone Encapsulant Clear 0.5 kg kit (Part No. 184 SIL ELAST KIT 0.5 PG)
- Protolabs Custom Array Molding Plates (Please refer to www.shaleklab.com/seq-well
  - Make out of aluminum and make sure to tap holes only on base plateBasePlate1 v3.1 (Bottom plate you mount the wafer to)
- BasePlate2 v3.1 (Divider for arrays)
- TopPlate1 v3.1 (Plate that holds the glass slides)
- TopPlate2 v3.1 (Top plate)
- 45 micron Silicon Master Wafer Size (Please refer to www.shaleklab.com/seq-well)
- Master, pre-silanized – (FlowJem, Inc. Toronto, Canada)
- Corning 72x25 Microscope Slides (Corning Life Sciences Cat. No. 2947)
- 6/32 ¼" Hex Screws
- 5/8" Hex 10/24 Screws
- Hex Screwdriver
- Vacuum grease
- 80 grit sandpaper
- 95% ethanol in spray bottle

Array Functionalization

Equipment

- Plasma Oven (Harrick Plasma PDC-001-HP)
- 2x 30-slide rack slotted (VWR Cat No. 25461-014)
- 16x20 cm staining dish (VWR Cat No. 25461-018)
- Vacuum Desiccator (VWR Cat No. 24988-164)
- Sterile 4-well dishes (Thermo Fisher Scientific Cat No. 267061)

(continues on the next page)

Reagents

- 200 proof ethanol (VWR Cat No. 89125-188)
- (3-Aminopropyl)triethoxysilane (Sigma Cat No. A3648)
- Acetone (Avantor Product No. 2440-10)
- p-Phenylene Diisothiocyanate (PDITC) (Sigma Cat No. 258555-5G)
- Pyridine (Sigma 270970-1L)
- Dimethylformamide (DMF) (Sigma Cat No. 227056-1L)
- Chitosan (Sigma Cat No. C3646-100G)
- Poly(L-glutamic) acid sodium solution (Sigma Cat No. P4761-100MG)
- 5M NaCl (Sigma Cat No. S6546-1L)
- Sodium Carbonate (Sigma Cat No. S2127-500G)

Buffer Reagents

Bead Loading Buffer

- Sodium Carbonate (Sigma Cat No. 223530-500G)
- BSA (Sigma Cat No. A9418-100G)
- Water (Thermo Fisher Scientific Cat No. 10977023)

Complete Lysis

- Guanidine Thiocyanate, (Sigma Cat No. AM9422)
- 0.5 M EDTA (Thermo Fisher Scientific Cat No. 15575020)
- Water (Thermo Fisher Scientific Cat No. 10977023)
- 10% Sarkosyl (Sigma Cat No. L7414)
- 100% 2-Mercaptoethanol (Sigma Cat No. M3148-25ML)

Hybridization Buffer

- 5 M NaCl (Thermo Fisher Scientific Cat No. 24740011)
- 1x PBS (Thermo Fisher Scientific Cat No. 10010023)
- PEG-8K (50%) (Fisher Scientific Cat No. BP337-100ML)

(continues on the next page)

Wash Buffer

- 5 M NaCl (Thermo Fisher Scientific Cat No. 24740011)
- 1 M $MgCl_2$ (Sigma Cat No.63069-100ML)
- 1 M Tris-HCl pH 8.0 (Thermo Fisher Scientific Cat No. 15568025)
- Water (Thermo Fisher Scientific Cat No. 10977023)

Array Quenching Buffer

- Sodium Carbonate (Sigma Cat No. 223530-500G)
- 1 M Tris-HCl pH 8.0 (Thermo Fisher Scientific Cat No. 15568025)
- Water (Thermo Fisher Scientific Cat No. 10977023)

RT Reagents

- UltraPure Distilled Water (Thermo Fisher Scientific Cat No. 10977023)
- Maxima 5x RT Buffer/Maxima H-RT (Thermo Fisher Scientific Cat No. EPO0753)
- 20% Ficoll PM-400 (Sigma Cat No. F5415-50mL)
- 10 mM dNTPs (New England BioLabs Cat No. N0447L)
- RNAse Inhibitor (Thermo Fisher Scientific Cat No. AM2696)
- Template Switching Oligo (Order from IDT)

Exonuclease Reagents

- Exonuclease I (*E. coli*) (New England Biolabs Cat No. M0293S)

Second Strand Synthesis Reagents

- Maxima 5x RT Buffer/Maxima H-RT (Thermo Fisher Scientific Cat No. EPO0753)
- 10 mM dNTPs (New England BioLabs Cat No. N0447L)
- dN-SMART Oligo (Order from IDT)
- UltraPure Distilled Water (Thermo Fisher Scientific Cat No. 10977023)
- Klenow Exo- (New England BioLabs Cat No. M0212S)
- 30% PEG8000 (Sigma-Aldrich 89510-1KG-F)

PCR Reagents

- IS PCR Primer (Order from IDT)
- KAPA HiFi Hotstart Readymix PCR Kit (Kapa Biosystems Cat No. KK-2602)

Nextera Reagents

- Nextera XT DNA Library Preparation Kit (96 samples) (Illumina FC-131-1096)
- New-P5-SMART PCR Hybrid Oligo (Order from IDT)
- Nextera N70X Oligo (Order from Illumina)

Operating Equipment

- Polycarbonate (PCTE) 0.01 micron 62x22 mm precut membranes, 100 count (Sterlitech Custom Order)
- mRNA Capture Beads (Chemgenes Cat No. MACOSKO-2011-10)
- Lifter Slips, 25x60mm (Electron Microscopy Science Cat No. 72186-60)
- Agilent Clamps (Agilent Technologies Cat No. G2534A)

Sequences

Barcoded Bead SeqB:

5'–Bead–Linker--TTTTTTTAAGCAGTGGTATCAACGCAGAGTAC-
JJJJJJJJJJJJNNNNNNNNTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT--3'

Template Switching Oligo (TSO):
AAGCAGTGGTATCAACGCAGAGTGAATrGrGrG

dN-Smart Randomer (dN-SMRT):
AAGCAGTGGTATCAACGCAGAGTGANNNGGNNNB

Smart PCR Primer (TSO_PCR):
AAGCAGTGGTATCAACGCAGAGT

New-P5-SMART PCR Hybrid Oligo (P5-TSO_Hybrid):

AATGATACGGCGACCACCGAGATCTACACGCCTGTC-
CGCGGAAGCAGTGGTATCAACGCAGAGT*A*C

Custom Read 1 Primer (Read_1_Custom_SeqB):
GCCTGTCCGCGGAAGCAGTGGTATCAACGCAGAGTAC

Appendix E: Loss of DNA methyltransferase activity in primed human ES cells triggers increased cell-cell variability and transcriptional repression

This appendix is adapted in accordance with The Company of Biologist's open access policy from the following article published in Development:

Tsankov, A.M.*, Wadsworth II, M.H.*, Akopian, A., Charlton, J., Allon, S.J., Arczewska, A., Mead, B.E., Drake, R.S., Smith, Z.D., Mikkelsen, T.S., Shalek, A.K., Meissner, A., "Loss of DNA methyltransferase activity in primed human ES cells triggers increased cell-cell variability and transcriptional repression," *Development*, 146, (2019).

** Denotes equal authorship*

Abstract

Maintenance of pluripotency and specification towards a new cell fate are both dependent on precise interactions between extrinsic signals and transcriptional and epigenetic regulators. Directed methylation of cytosines by the de novo methyltransferases DNMT3A and DNMT3B plays an important role in facilitating proper differentiation, whereas DNMT1 is essential for maintaining global methylation levels in all cell types. Here, we generated single-cell mRNA expression data from wild-type, DNMT3A, DNMT3A/3B and DNMT1 knockout human embryonic stem cells and observed a widespread increase in cellular and transcriptional variability, even with limited changes in global methylation levels in the de novo knockouts. Furthermore, we found unexpected transcriptional repression upon either loss of the de novo methyltransferase DNMT3A or the double knockout of DNMT3A/ 3B that is further propagated upon differentiation to mesoderm and ectoderm. Taken together, our single-cell RNA-sequencing data provide a high-resolution view into the consequences of depleting the three catalytically active DNMTs in human pluripotent stem cells.

INTRODUCTION

Appendix E.1 Background

Isogenic populations of cells can exhibit substantial phenotypic variation, which can, in turn, play an important role in development and in adapting to changing external conditions.[1] Variation in gene expression, due to stochastic bursting and asymmetric division of key molecular drivers of cellular identity, accounts for a large amount of observed cell-to-cell (cell-cell) variability within a given cell type.[2] Cellular heterogeneity has historically been measured using microscopy and fluorescent labeling of key markers. These techniques have high spatial and cellular resolution but rely on prior knowledge and a limited number of markers, making it difficult to assay cellular differences comprehensively. The advent of single- cell genomic methods now enables profiling of transcriptional, genetic and epigenetic variation between individual cells on a global scale that depends less on a priori hierarchies and predefined markers.[3]

Single-cell RNA-sequencing (scRNA-seq), in particular, has led to remarkable advances in defining and refining the myriad cell states[4,5], cell types[6-8] and progenitors[9,10] that are present during mammalian development and differentiation.[11-14] This has been aided by computational advances in clustering and pseudotemporal ordering of single cells that have enabled accurate inference of cell states and developmental trajectories, respectively[15-17]. From a biological perspective, scRNA-seq has allowed the role of transcriptional heterogeneity to be explored. For example, single-cell profiling of mouse embryonic stem (ES) cells has revealed sporadic expression of polycomb targeted lineage regulators and less heterogeneity among pluripotency-associated genes in 2i versus serum growth conditions.[18] These results suggest a model whereby mouse ES cells are afforded the opportunity to access lineage specification programs through stochastic expression of pluripotency factors and lineage regulators typically repressed by H3K27me3.

DNA methylation also plays an important role in maintenance of and exit from pluripotency. Variation in DNA methylation modulates metastable switching in mouse ES cells between ZFP42 low and high states.[19] Three catalytically active DNA methyltransferases (DNMTs) are responsible for maintenance (DNMT1) and de novo DNA methylation (DNMT3A/3B) in mammals, and all three are essential for normal development.[20] DNA methylation by DNMT3A/3B plays a particularly important role during development and ES cell differentiation[21,22], and both catalytically active enzymes are highly expressed in undifferentiated cells. Bulk experiments have shown a limited global impact of DNMT3A/3B knockout on the global DNA methylation landscape in human ES cells.[23] This limited effect may be, in part, a consequence of bulk measurements, and it remains unknown how these epigenetic regulators affect transcriptional variation at the single-cell level, including how this may bias differentiation to new cell fates. To study this, we utilized previously generated knockout cell lines[23] in the undifferentiated and differentiated states to investigate the effects of these mutations on transcription at single-cell resolution.

RESULTS

Appendix E.2 Increased cellular variation in ES cells lacking DNMT3A and DNMT3A/3B

To explore the role of DNMTs in transcriptional regulation within individual cells, we used Smart-Seq2-based scRNA-seq[24] to profile three HUES64 human ES cell lines – wild type (WT), with homozygous catalytic disruption of DNMT3A (3AKO), and with double knockout of both DNMT3A/3B (DKO).[23] Although the global decrease in methylation levels in the DKO cells is limited (**Figure 1A**), they have 10-fold more differentially methylated regions than 3AKO relative to WT.[23] Dimensionality reduction showed that WT, 3AKO and DKO cells mostly cluster by cell line (**Figure 1B**). We found that 3AKO and DKO undifferentiated cells were equally dissimilar to WT ES cells (**Figure 1C**, **top**), which was unexpected given the much greater similarity in the global methylation landscape between WT and 3AKO bulk samples.[23] Interestingly, we noticed a significantly higher intra-sample cell-cell distance in the DKO and 3AKO populations relative to WT ($P<10^{-15}$, Wilcoxon signed rank test, **Figure 1C**, **bottom**).

To control for the effect of background differentiation on our measure of cellular heterogeneity, we classified all cells as pluripotent, endoderm (dEN), mesoderm (dME) and ectoderm (dEC) using previously reported germ layer markers.[21,25,26] We observed

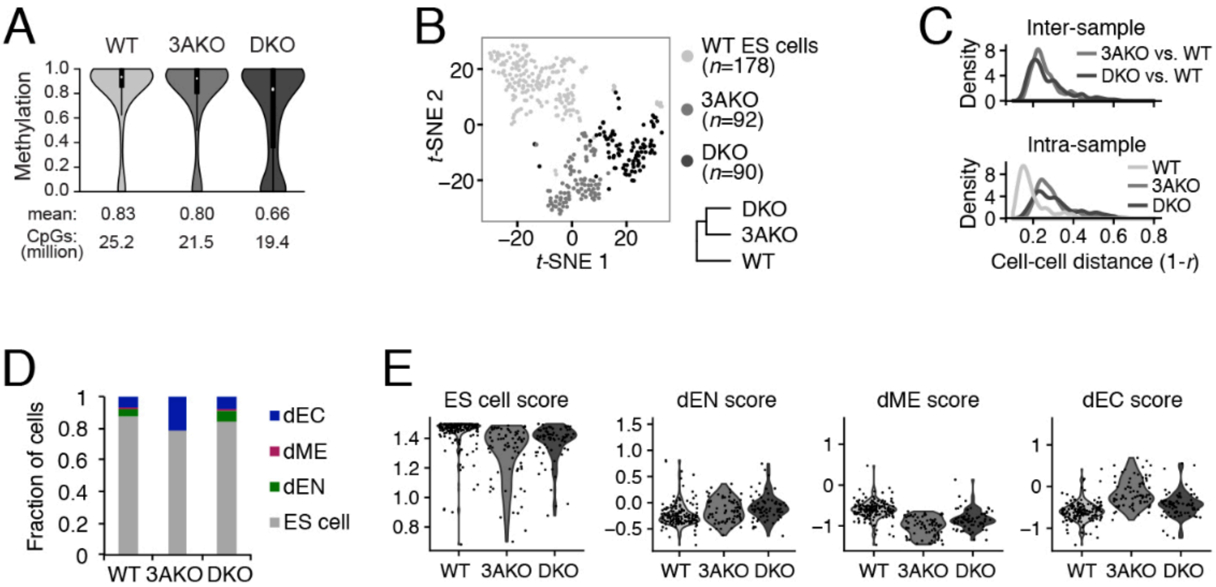


Figure 1 | Increased cellular variation in DNMT3A and DNMT3A/3B knockout ES cells. (A) Violin plot of CpG methylation for wild-type (WT),DNMT3A−/−(3AKO)andDNMT3A/3B−/−(DKO) cells averaged across two replicates. Mean methylation level and number of CpGs per sample are shown at the bottom and black boxes within the violin plots represent the interquartile range. Data were obtained from Liao et al. (2015) and are available at GEO under accession numberGSE63281. **(B)** Dimensionality reduction of WT, 3AKO and DKO single ES cells (dots) using t-distributed stochastic neighbor embedding (t-SNE) and hierarchical clustering (**bottom right**) of the averaged expression profiles for sorted ES cells from each cell line. Samples 3AKO and DKO were more similar to each other than to WT ES cells. **(C)** Inter-sample (**top**) and intra-sample (**bottom**) density distribution of all pairwise cell-cell distances for WT, 3AKO and DKO cells. **(D)** Fraction ofWT, 3AKO and DKO cells classified into four categories: ES cell; endoderm (dEN); mesoderm (dME); and ectoderm (dEC). **(E)** Violin plots of ES cell, dEN, dME and dEC scores for WT, 3AKO and DKO samples: each dot represents anin silico-sorted undifferentiated cell.

an increase in differentiated cells in DNMT mutant cells and a distinct bias towards ectoderm in 3AKO cells (**Figure 1D**).

To control for the effect of background differentiation on our measure of cellular heterogeneity, we classified all cells as pluripotent, endoderm (dEN), mesoderm (dME) and ectoderm (dEC) using previously reported germ layer markers.[21,25,26] We observed an increase in differentiated cells in DNMT mutant cells and a distinct bias towards ectoderm in 3AKO cells (**Figure 1D**). We then in silico sorted for all undifferentiated cells and found that the intra-sample cell-cell distance using only cells classified as pluripotent was also significantly higher in the mutant cell lines relative to WT ($P<10^{-15}$, Wilcoxon signed rank test; **Figure S1A**). Our results were unchanged when repeating this analysis using three different cell-cell distance metrics (Euclidean, Manhattan, Spearman correlation; see Materials and Methods) and after controlling for data quality by focusing the analysis on the highest-quality ES cells (**Figure S1B, C**). Among undifferentiated 3AKO and DKO cells, we also found increased variation in pluripotency, ectoderm and endoderm scores (**Figure 1E**). Taken together, these results suggest increased cell- cell transcriptional variation that may affect the differentiation potential in 3AKO and DKO cells.

Appendix E.3 DNA methylation and transcript variation in DNMT3A/3B knockouts

To further examine whether disruption of the de novo methyltransferases also increases global transcriptional variability, we computed the dispersion – log(variance/mean) – and standard deviation in expression for every gene within each sample population (**Figure 2A; Figure S2A left**). DKO and 3AKO showed a significant increase in transcript variation at all genes relative to WT using both metrics ($P<10^{-15}$, Wilcoxon signed rank test; **Figure 2A**) that associated with a corresponding decrease in mean promoter methylation level (**Figure 2B**). To control for the impact of technical dropouts on our measurements of transcript variation, we explicitly modeled three parameters for the expression of each gene: the fraction of cells with no detectable expression (α), and the mean (μ) and standard deviation (σ) of expression among only detectably expressing cells.[5,27] As an example, the transcriptional variation of CTCFL increased in DKO versus WT cells as measured both by dispersion and by σ, whereas α decreased (**Figure 2C**). As observed using dispersion, we also confirmed a global increase in transcriptional variation using σ in the DNMT mutants relative to WT cells (**Figure S2A, right**). Globally, we found that difference in dispersion was highly correlated with difference in σ (r=0.75) but not with difference in α (r=−0.04, Fig. 2D), indicating that changes in the fraction of cells with detectable expression between samples does not associate with the global increase in transcriptional variation we report among DNMT mutant and WT ES cells. Consistent with the increased variation in pluripotency scores (**Figure 1E**), we also observed a significant increase in standard deviation of expression across all cells and cells with detectable expression (σ) at pluripotency gene markers ($P<10^{-6}$, Wilcoxon signed rank test; **Figure S2A**), indicating that DNMT disruption leads to more variable expression of key pluripotency genes. We further quantified the relationship between changes in dispersion and average expression. We found 4740 and 1139 genes with a higher dispersion (difference greater than 1.5) in the DNMT mutants and WT cells, respectively; of those, 92% and 97%, respectively, also displayed a higher mean expression in the sample with higher dispersion (**Figure S2B**). We confirmed the trends in transcriptional variation that we observed in the scRNA-seq data from WT and 3AKO cells for ZFP42, MAP4K4 and RAD51 using RNA fluorescence in situ hybridization (FISH; **Figure 2E,F; Figure S2C**). The standard deviation of gene expression for ZFP42 using RNA FISH was slightly higher in WT versus 3AKO, whereas the difference in transcriptional variation was more pronounced between the two conditions for MAP4K4 and RAD51. In summary, we find increased transcriptional variation in undifferentiated 3AKO and DKO cells at genes that predominantly increase in mean expression; however, this increase in transcript variation is uncorrelated with dropout rate in our scRNA-seq data.

Changes in DNA methylation variability have been linked to cancer risk markers, higher order chromatin organization and variability in gene expression across cancer patients.[28-30] DNA methylation variability can be measured in phase at the individual read level using bisulfite sequencing, whereby each read can be considered to derive from a different cell. Globally, an increase in the percentage of reads displaying discordant methylation states was reported for DKO, but not for 3AKO, relative to WT ES cells.[23] In addition, we measured the normalized methylation entropy (NME)[29], an alternative approach based on statistical physics and information theory, and found that NME increased slightly in 3AKO and drastically in DKO relative to WT as mean methylation level decreased (**Figure 2G**). Density scatter plots showed that the increase in NME was largely due to a shift of high methylation level CpGs with low entropy in WT to intermediate methylation level CpGs with high entropy in DKO (**Figure 2H**). As the DKO-specific increase in NME does not appear to be proportional to the increase in transcriptional variation observed in both 3AKO and DKO, it suggests that genome-wide DNA methylation variability does not fully explain global transcript variation.

To further explore the relationship between DNA methylation and transcriptional dispersion at a single-cell level, we focused our analysis on gene promoters. We initially performed promoter epigenetic state enrichment analysis in ES cells[21] for the most and least transcriptionally variable genes in WT, 3AKO and DKO samples, and found an

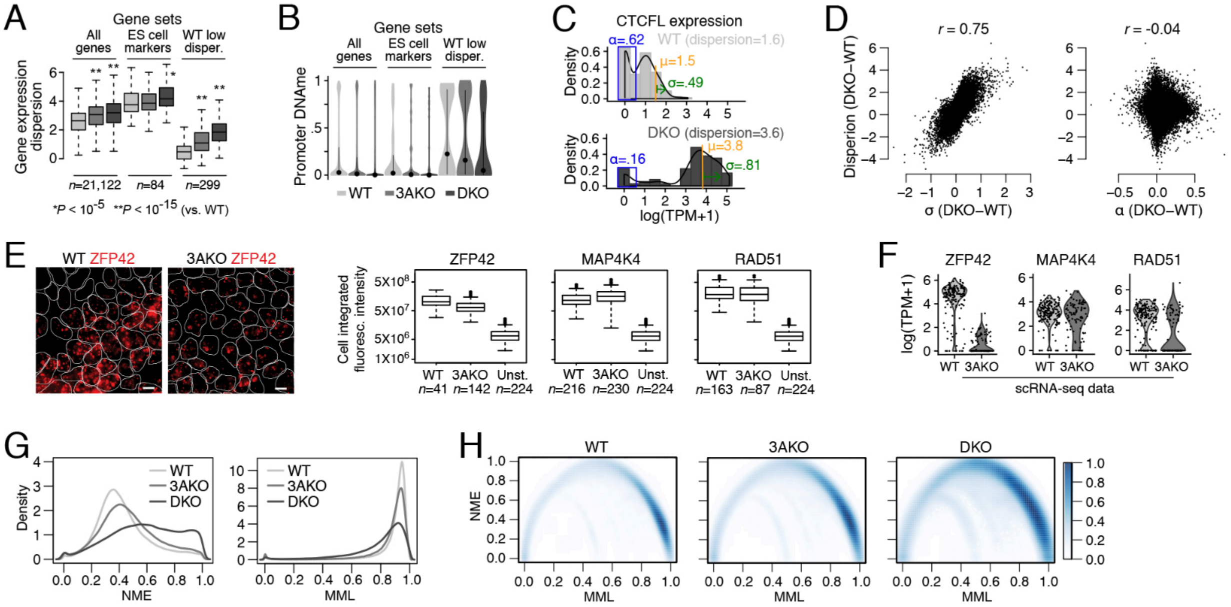


Figure 2 | Relationship between DNA methylation level, mean methylation entropy and transcript variation in DNMT3A and DNMT3A/3B knockout. (A) Boxplots of gene expression dispersion distribution, log(variance/mean), for all genes, ES cell markers and WT low dispersion genes for WT, 3AKO and DKOES cells. **(B)** Violin plots of promoter mean CpG methylation for all genes, ES cell markers and WT low dispersion genes from WT, 3AKO and DKO ES cell bulk samples, averaged across two replicates. Dots represent the mean and lines extend at most one standard deviation from the mean. **(C)** Histograms of CTCFL expression in WT (top) and DKO (bottom) cells binned at intervals of 0.5 log(TPM+1) expression levels and normalized to the total cell counts. The three parameters that are estimated for CTCFL gene expression distribution (α,μ and σ) are shown in blue, orange and green, respectively. Dispersion for CTCFL increases and coincides with an increase in σ and as well as a decrease inα. **(D)** Scatter plots of the difference in dispersion,log(variance/mean) and parameters σ(left) andα(right) in DKO versus WT cells. We observe a high correlation between dispersion and σ difference (r=0.75) but not between dispersion and α difference (r=−0.04). **(E)** Left: representative images of RNA FISH with probes targeting ZFP42 (red) in WT (left) and 3AKO (right) ES cells. Cell segmentation is shown using white outlines. Scale bars: 10μm. Right: box plots of ZFP42 (left), MAP4K4 (middle) and RAD51 (right) integrated probe intensity summed over the volume of the cell for WT, 3AKO and unstained (Unst.) ES cells. **(F)** Violin plots of log(TPM+1) gene expression for ZFP42 (left), MAP4K4(middle) and RAD51 (right) in WT and 3AKO ES cell scRNA-seq data show similar trends in transcript variation as the RNA FISH experiment for these three genes. **(G)** Normalized methylation entropy (NME; left) and mean methylation level (MML; right) measured using WT, 3AKO and DKO ES cell whole genome bisulfite sequencing data across all chromosome 21 and 22 CpGs using the approach in Jenkinson et al. (2017). **(H)** Smoothed scatter plot with color intensity showing density of all chromosome 21 and 22 CpGs NME versus MML data. For WT, most CpGs have high MML and low NME (dark blue, bottom right). DKO CpGs with high NME spread across a lower MML (middle top of DKO plot; intensity gets darker), consistent with the global loss of methylation in the DKO sample. Box plots: boxes display the interquartile range, horizontal line within the box shows the median, whiskers extend to the most extreme data point that is no morethan 1.5 times the length of the interquartile range.

association between low dispersion genes and H3K9me3-enriched or highly methylated promoters in WT cells (**Figure S2D**). In contrast, whereas genes with H3K9me3-enriched promoters also correlated with genes of lowest transcriptional variance in 3AKO and DKO cells, high methylation promoters showed low to no correlation (**Figure S2D**). Consistent with this result, for the least variable genes in WT ES cells, we found a significant increase in mean expression and transcriptional dispersion in 3AKO and DKO samples ($P<10^{-15}$, Wilcoxon signed rank test; **Figure 2A, right**) and a concomitant decrease in DNA methylation at the corresponding promoters (**Figure 2B**). This implies that the expression of low dispersion genes in WT is regulated, in part, by the DNA methylation level. Although mean methylation and transcript dispersion levels appear to correlate at these genes, the same correlation is not apparent when comparing promoter mean NME and transcriptional variation globally, as measured by gene dispersion or σ (**Figure S2E,F**). Together, this suggests that the increase in transcript variability observed after loss of DNMT3A/3B associates with loss of methylation at a subset of promoters but not globally.

Appendix E.4 Widespread transcriptional repression and super-enhancer misregulation

To better understand the regulatory changes that underlie the observed transcriptional dynamics in the DNMT3A/3B mutants, we identified all three-way differentially expressed genes between WT, 3AKO and DKO sorted samples (**Figure 3A**). We found that the vast majority of genes were repressed (1964) rather than activated (470) relative to WT, which was somewhat unexpected given that loss of methylation is typically more associated with gene activation. Among the most downregulated genes in human ES cells, we observed a number of zinc fingers and important pluripotency transcription factors (TFs), including ZFP42, PRDM14, NANOG, POU5F1 and MYC (**Appendix B, Table S1**). Interestingly, the latter three TFs showed lower expression in 3AKO than DKO despite the DKO being generated through a DNMT3B deletion in the DNMT3A knockouts.[23] We also found a number of housekeeping genes with reduced expression, including those encoding actins, heterogeneous nuclear ribonucleoproteins (HNRNPs) and proteasome genes.

We next performed a comprehensive search for promoter enrichment against published DNA methylation, histone modification and TF binding data from matched samples (**Figure 3B**) to explore the potential underlying mechanism.[21,23,25,26] We found a significant association between loss of promoter methylation and expression increase, as illustrated at the CTCFL locus (**Figure 3C, top**). Surprisingly, we also identified 152 and 82 promoters that increased in DNA methylation (e.g. ZFP42; **Figure 3C, bottom**) in

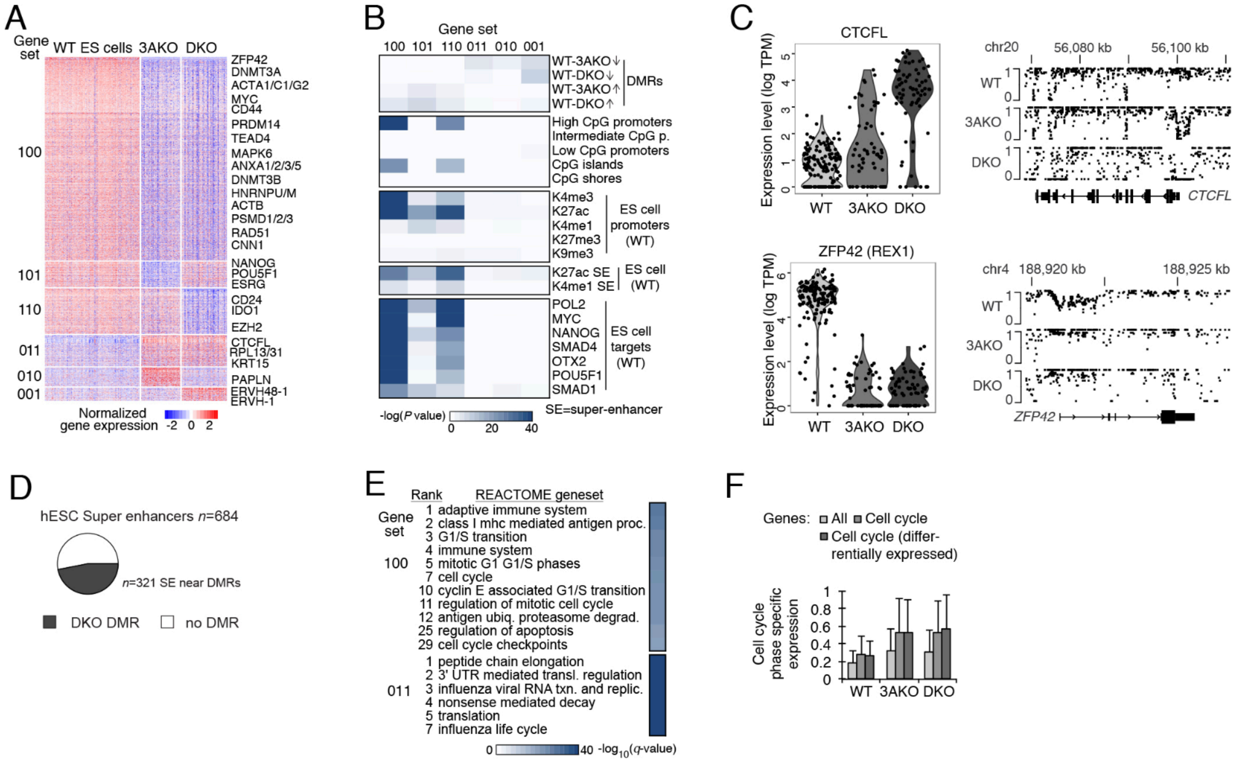


Figure 3 | Global transcriptional repression and altered regulation in DNMT3A and DNMT3A/3B knockout ES cells. (A) Differentially expressed genes (right;rows) for sorted populations of WT, 3AKO and DKO ES cells (columns). Genes are separated into six gene sets [left: 100 (n=1443), 101 (n=191), 110 (n=330), 011(n=229), 010 (n=143) and 001 (n=98)], where 1 or 0 indicates high or low expression for the respective condition (order: WT, 3AKO, DKO). **(B)** Genomic enrichment analysis for gene sets (columns) defined in panel A against CpG density features, epigenetic and TF binding data collected in matching WT ES cells (Gifford et al., 2013; Tsankov et al., 2015b). **(C)** Top: distribution (dots indicate individual cells) of CTCFL expression (left) and the corresponding CpG methylation levels at the CTCFL locus for WT, 3AKO and DKO ES cells. Bottom: ZFP42 cellular expression (left) and promoter methylation (right) as described above. **(D)** Of all 684 H1 ES cell super-enhancers (Hnisz et al., 2013), 321 (47%) are located within 1 kb of a DKO DMR (displayed in black). In total, 734 DKO DMRs (of44,244 total) were associated with super-enhancers, and are defined as regions with difference in methylation>0.6 relative to WT, withP<0.01 (F-test). **(E)** Functional enrichment analysis for the gene sets defined in panel A against the REACTOME database. **(F)** Distribution of cell cycle phase-specific expression for sorted WT, 3AKO and DKO ES cells considering all genes, cell cycle annotated genes and differentially expressed cell cycle annotated genes. Error bars indicate one standard deviation. DMR, differentially methylated region; K, lysine on histone 3; me3, tri-methylation; ac, acetylation; me1, mono-methylation.

3AKO and DKO, respectively, which overlapped significantly with decreased expression in the mutants (**Figure 3B**). Genes repressed in the knockouts frequently had high CpG-dense promoters that were enriched for active histone modification in WT cells (H3K27ac and, to a lesser degree, H3K4me3 and H3K4me1; **Figure 3B**). DNMT3A and DNMT3B have previously been shown to occupy active enhancers, and knockdowns of the de novo methyltransferases reduced super- enhancer activity and disrupted homeostasis in epidermal stem cells.[31] In our dataset, repressed genes (e.g. NANOG, POU5F1) associated significantly with upstream H3K27ac and H3K4me1 super-enhancers, suggesting a similar role for the de novo methyltransferases at super-enhancers in human ES cells. We also found that nearly half of human ES cell super-enhancers[32] showed drastic changes in methylation levels in DKO (**Figure 3D; Figure S3A**), suggesting that DNMT3A/3B shape the methylation landscape near super-enhancers. We further identified high enrichment for in vivo binding of a number of key pluripotency associated TFs upstream of 3AKO and DKO repressed genes, including MYC, NANOG, and POU5F1. As these factors occupy 76% of downregulated gene promoters in WT ES cells, a large fraction of the repressed phenotype may be mediated by their decreased expression and/or activity in the mutants. Taken together, loss of DNMT3A and DNMT3A/3B appears to interfere with normal super-enhancer activity upstream of pluripotency associated regulators, leading to downregulation of these core ES cell TFs and their downstream targets.

Appendix E.5 Loss of DNMT3A/3B alters cell cycle gene expression

We next performed gene set enrichment analysis and observed that genes upregulated in the 3AKO and DKO mutants included those encoding a number of ribosomal proteins (e.g. RPL13/31) that are associated with the influenza life cycle and viral RNA transcription and replication (**Figure 3E, bottom**). Combined with the observation that ERVH48-1 and ERVH-1 are also upregulated in the DKO, these changes in expression point to increased activity of endogenous retroviral elements. Interestingly, we also found that downregulated genes associated with a number of cell cycle categories, including gene sets related to G1/S transition and the establishment of checkpoints (**Figure 3E, top**).

To investigate possible cell cycle alterations in the DNMT3A/3B mutants, we identified all differentially expressed cell cycle annotated genes in the WT, 3AKO and DKO ES cell samples (**Figure S3B**). We found decreased expression relative to WT ES cells in a number of key cell cycle genes (e.g. TP53, MCM2/3/4/5/6 and ORC1/2/5; **Figure S3B**) that were also downregulated during normal differentiation.[21] We also observed downregulation of CDK4/6 and upregulation of CCND1 in the 3AKO, which has previously been observed during normal ectoderm differentiation.[21] Although the proportion of cells in different phases of the cell cycle is similar for all three samples (**Figure S3C**), we found a global shift from constant to phase-specific cell cycle expression in the DNMT3A/3B mutants at all genes and, especially, at ones annotated to have cell cycle function (**Figure 3F**). Taken together, our data show a global change in expression of cell cycle-associated genes upon DNMT3A/3B loss with increases in phase-specific cell cycle expression that suggest the establishment of a regulated G1/S transition and cell cycle checkpoints relative to WT human ES cells.

<u>Appendix E.6 Aberrant expression following ES cell differentiation of DNMT3A/3B knockouts</u>

To investigate whether and how the observed transcriptional changes in the knockouts affect cellular specification, we differentiated all three cell lines for 5 days towards dME and dEC followed by scRNA-seq. Dimensionality reduction showed that cells clustered primarily by cell type and sample identity (**Figure 4A**). We observed a similar proportion of dME- and dEC-positive cells between knockout and WT samples following differentiation (Fig. 4B). The spread of dEC scores was similar across WT, 3AKO and DKO dEC samples, whereas variation in dME scores was slightly greater in the 3AKO dME sample relative to WT and DKO (**Figure S4A**). Population averaged transcriptomes for all samples clustered by cell type (**Figure 4C**) and showed that dEC samples were more similar to ES cells than dME samples, which is consistent with the inherent dEC bias we noted in 3AKO and DKO ES cells (**Figure 1D, E**). In line with our ES cell results, we found that the DKO dME/dEC cells were slightly more similar to WT dME/dEC than 3AKO dME/dEC cells (**Figure 4D, left**). We also observed an increase in intra- sample cell distance in the knockouts versus WT for both dME and dEC (**Figure 4D, right**), although the difference was not as pronounced as in ES cells (**Figure 1C**).

We then identified all differentially expressed genes in WT, 3AKO and DKO dME and dEC samples and compared them with the ES cell populations. We found that in dME 36% and in dEC 34% of differentially expressed genes had a similar change in expression in the ES cell mutants, including 59% in dME and 42% in dEC of the genes repressed in both ES cell knockouts (**Figure 4E**). These genes included ZFP42 (**Figure S4B**), actin family genes and proteasome genes (**Figure 4F,G**, gene set 100.1), and associated with some of the same functional categories as repressed genes in the ES cell knockouts (e.g. protein metabolism, immune system, apoptosis; **Figure 4H**). In dME, a number of genes associated with extracellular matrix organization and collagen formation (BMP1, COL4A1/2/6) were downregulated uniquely in dME DNMT mutants and not in ES cells (**Figure 4H**, gene set 100.0). We saw a similar enrichment for translation and viral response for upregulated genes in both ES cell and dME knockouts versus WT (**Figure 4H**, gene set 011.1, e.g. RPS19/27/28), and enrichment for mRNA processing and splicing pathways for dME-specific knockout activated genes (**Figure 4H**, gene set 011.0).

To investigate the underlying mechanisms that may explain the transcriptional changes in the dME and dEC knockouts, we overlapped the promoter epigenetic state with differentially expressed gene categories. We found that genes that gain methylation in the three germ layers are also upregulated in DKO but not in 3AKO dME, suggesting that DNMT3B may compensate for DNMT3A loss at these lineage-specific targets (**Figure S4C**). This trend was also notable in dEC but to a lesser degree (**Figure S4D**). Further, we observed enrichment of high-CpG promoters (HCP) and CpG islands for inherited repressed genes (100.1) but not for dME- or dEC-specific repressed genes (100.0). Finally, we found an enrichment for genes downstream of super-enhancers being misregulated in dEC knockouts relative to WT, including FGFR1 (gene set 101), SOX11 (011) and NR6A1 (010). In dME, we found an association between dME upstream super-enhancers and dME- specific gene repression (gene set 100.0), including COL4A1/2, KRT8, CD99, and the TF HAND1, which may point to a cell type- specific role of DNMT3A/3B at dME super-enhancers. Moreover, downregulation of HAND1 may mediate further downstream repression at its dME targets, as we observed for core TFs POU5F1 and NANOG in undifferentiated DNMT3A/3B knockouts.

We also found a number of TFs with important roles in developmental processes and oncogenesis to be aberrantly expressed in the dEC and dME DNMT mutants relative to WT. In dEC, genes encoding key TFs associated with ectoderm lineage development

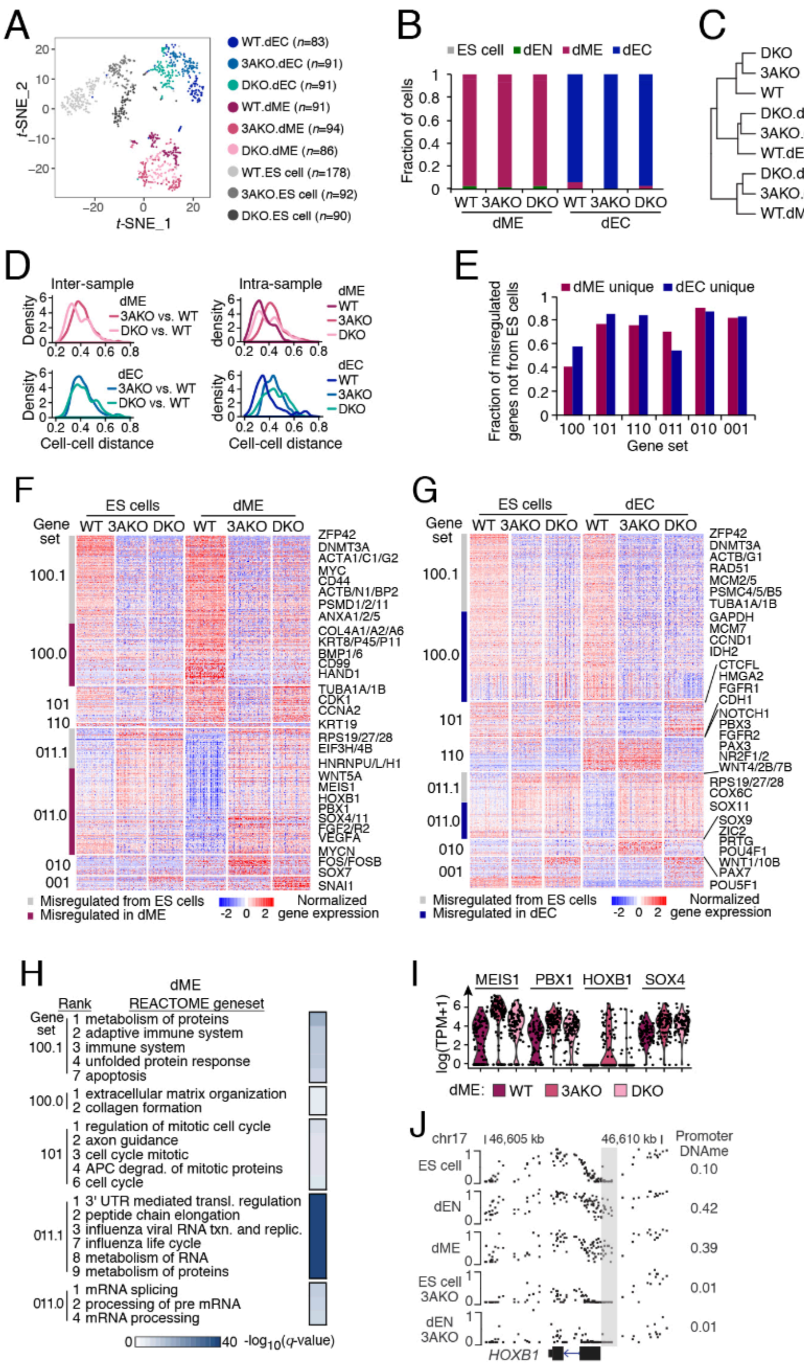


Figure 4 | Transcriptional changes and mis-regulation in DNMT3A/3B knockout cells during ES cell differentiation. (A) Dimensionality reduction of wild-type (WT), DNMT3A knockout(3AKO) and DNMT3A/3B knockout (DKO) single ES, mesoderm (dME) and ectoderm (dEC) cells (dots) using t-distributed stochastic neighbor embedding (t-SNE). Number of cells is shown in parentheses. **(B)** Fraction of WT, 3AKO and DKO mesoderm (left)and ectoderm (right) cells classified into four cell types (ES cell, dEN, dME, dEC). **(C)** Hierarchical clustering of the averaged expression profiles for all sorted samples. **(D)** Inter-sample (**left**) and intra-sample (**right**) density distribution of all pairwise cell-cell distances (1–Pearson correlation coefficient) for WT, 3AKO and DKO dME (**top**) and dEC (**bottom**) cells. **(E)** Fraction of differentially expressed genes in dME (red) and dEC (blue) that were not already present in ES cells, or are dME/dEC unique. Gene sets are defined in the legend for F. **(F)** Differentially expressed genes **(right; rows)** for sorted population of WT, 3AKO and DKO ES and mesoderm cells (columns). Genes are separated into eight gene sets (left: 100.1, 100.0, 101, 110, 011.1, 011.0, 010 and 001), for which 1 or 0 indicates high or low expression, respectively, for each condition (order: dME WT, 3AKO, DKO). Suffix .1 indicates inherited from ES cells whereas .0 indicates dME unique. **(G)** Differentially expressed genes **(right; rows)** for sorted population of WT, 3AKO and DKO ES and ectoderm cells (**columns**). Genes are separated into eight gene sets as described above. **(H)** Functional enrichment analysis for the dME gene sets defined in F against the REACTOME database. **(I)** Distribution of gene expression, log(TPM+1), for selected TFs aberrantly expressed in 3AKO and DKO dME cells, relative to WT. Dots represent cells. **(J)** CpG methylation levels at theHOXB1locus for ES, dEN, dME, 3AKO ES and 3AKO dEN cells. TheHOXB1promoter is highlighted with a gray bar and the mean promoter methylation level is listed on the right.

were specifically downregulated in DKO (e.g. PAX3, NR2F1/2), upregulated in both mutants (e.g. SOX11) or upregulated in 3AKO relative to WT (e.g. SOX9, ZIC2, POU4F1, PAX7; **Figure S4E**). Moreover, key pluripotency TFs, such as POU5F1, with a promoter that is focally methylated during differentiation, and PRDM1, along with POU domain TFs POU5F2 and POU2F2, were specifically upregulated in the DKO dEC sample relative to WT (**Figure S4E, bottom**). This was accompanied by an increase in the median and standard deviation of ES cell scores observed in dEC DKO cells versus WT (**Figure S4F**). In dME, TFs MEIS1, PBX1, HOXB1 and S OX4 were upregulated in the 3AKO cells relative to WT (**Figure 4I**). As the promoter methylation of HOXB1 increases drastically during ES cell differentiation towards dEN and dME, and this gain in dEN depends on the catalytic activity of DNMT3A (Fig. 4J), it is likely that the HOXB1 promoter methylation is misregulated in a similar manner in dME cells lacking DNMT3A. Although we do not observe a change in promoter methylation for the genes encoding TFs MEIS1 and SOX4, their expression is correlated with HOXB1 (r=0.2, P value<0.05; Pearson) implying that these TFs are co-regulated as part of the same gene expression program. Finally, we observed aberrant expression for a number of cell cycle annotated genes and key transcriptional regulators in the knockouts after dME and dEC differentiation. In dME mutants, and especially in 3AKO, we observed downregulation of mitosis-associated genes CDK1 and CCNA2 (**Figure S4G**) as well as cell cycle-associated genes linked to lineage choice (CCND1, CCND3, CDKN1A). In dEC mutants, a number of S-phase genes were downregulated (MCM2/5/7) as well as CCND1 (**Figure 4G**), which acts to block endoderm formation in late G1 phase[33] and promotes neuroectoderm cell fate.[34] In both dME and dEC, we observed a similar proportion of cells in different phases of the cell cycle (**Figure S4H**) and levels of phase-specific expression (**Figure S4I**).

Appendix E.7 Loss of DNMT1 triggers increased transcript variation and differentiation

To complement our results from the de novo DNA methyltransferases, we explored the effects of loss of the maintenance enzyme DNMT1. Loss of DNMT1 results in a global loss of DNA methylation rather than the limited dynamics we find in the DNMT3 knockouts.[23] Specifically, we utilized our previously established doxycycline-inducible downregulation of DNMT1 system and collected live cells every day for 8 days for single cell methylation profiling and at day 0, 2 and 8 following doxycycline treatment for scRNA-seq. We observed global loss of methylation in all the profiled cells beginning at day 2, which plateaued at a minimum at around day 6 to 8 (**Figure 5A**). Dimensionality reduction of the scRNA-seq revealed a high similarity between most day 0 and day 2 cells, with a gradual departure from WT for some day 2 cells and all day 8 cells (**Figure 5B**). Quantifying cell type identity (**Figure 5C**) showed an increase in cells exiting pluripotency at day 8, with a preference of escape towards ectoderm, as observed in 3AKO and DKO. We found that both the population average similarity between in silico-sorted undifferentiated samples (**Figure 5B, bottom right)** and the inter-sample ES cell distance versus day 0 (**Figure 5D, top**) increased with time after doxycycline induction. We also observed an increase in intra-sample cell-cell distance at day 2 and day 8 compared with day 0 (**Figure 5D, bottom**), and note that the heterogeneity at day 8 exceeds that found in the DNMT3A/3B knockout ES cells (**Figure 1C**). Variation in gene expression also increased at day 2 and day 8 for all genes, pluripotent markers and the least variable WT genes ($P<10^{-8}$, Wilcoxon signed rank test; **Figure 5E**). Our results were consistent after controlling for differences in data quality and sequencing depth between samples. We again found an association between genes with the lowest expression dispersion and high methylation promoter occupancy at day 0, and this enrichment gradually decreased at day 2 and day 8, with downregulation of DNMT1 and concurrent global loss of methylation (**Figure S5A**), as we observed for 3AKO and DKO versus WT ES cells.

To gain insight into the functional changes induced by downregulation of DNMT1, we identified all differentially expressed genes in ES cells collected at day 0, 2 and 8 (**Figure 5F**). The majority (1638 of 2631; 62%) of day 8 differentially expressed genes were repressed relative to day 0, as observed in 3AKO and DKO ES cells. We found downregulation as early as day 2 of a number of ribosomal protein genes (e.g. RPS24/15, RPL12/19) associated with influenza life cycle (**Figure 5G**, gene set 100). At day 8, we observed a small downregulation of POU5F1 and other pluripotency-associated genes (e.g. CD24, DPPA4) and concomitant activation of NODAL signaling genes, including NODAL, CER1, LEFTY1/2 and downstream TF PITX2. We also note a shift in expression from glycolysis genes (e.g. GAPDH, PFKP/M) at day 0/2 to lipid metabolism at day 2 (e.g.

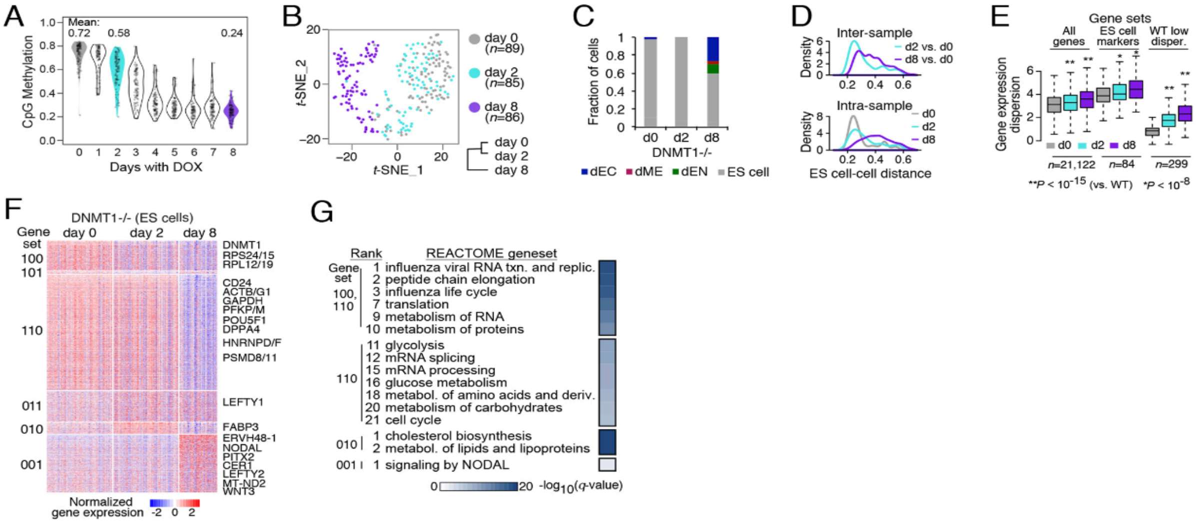


Figure 5 | Increased transcript variation and differentiation upon loss of DNMT1. (A) Violin plot of single cell methylation data, where each dot represents the average CpG methylation level per cell. Cells were collected for scRNA-seq after 0, 2 and 8 days of doxycycline treatment. **(B)** Dimensionality reduction of day 0, 2and 8 single cells (dots) using t-distributed stochastic neighbor embedding (t-SNE) and hierarchical clustering (bottom right) of the averaged expression profiles forin silico-sorted ES cell populations. **(C)** Fraction of cells classified into four categories (ES cell, dEN, dME, dEC) following 0, 2 and 8 days of doxycycline treatment. **(D)** Inter-sample (top) and intra-sample (bottom) density distribution of all pairwise cell-cell distances for in silico-sorted ES cells at day 0, 2 and 8. **(E)** Box plots of gene expression dispersion distribution at all genes, ES cell markers, and WT low dispersion genes for sorted ES cell populations at day0,2 and 8. **(F)** Differentially expressed genes (right; rows) for sorted population of ES cells at day 0, 2 and 8 (columns). Genes are separated into six gene sets[left: 100 (n=337), 101 (n=36), 110 (n=1301), 011 (n=349), 010 (n=139) and 001 (n=644)], where 1 or 0 indicates high or low expression for the respective condition (day 0, 2 and 8). **(G)** Functional enrichment analysis for the gene sets defined in F against the REACTOME database.

FABP3, FADS2), to oxidative phosphorylation genes (MT-ND2, MT-ND4L) at day 8 (**Figure 5F,G**). Finally, we observed changes in cell cycle regulation for ES cells that survived loss of methylation, including an increase in fraction of G2/M cells (**Figure S5B**) and increase in cell cycle phase-specific expression (**Figure S5C**). These changes might reflect a longer G2/M phase needed for methylation maintenance fidelity and compacting of chromosomes. Taken together, we observe repression at most differentially expressed genes and an increase in differentiation, as well as cellular and gene expression variation in ES cells upon loss of DNMT1.

Appendix E.8 Discussion and Conclusions

Pluripotent stem cells are a powerful model to explore the targets and role of epigenetic regulators. We have previously generated knockout human ES cell lines for the three catalytically active DNA methyltransferases.[23] With the advance of single- cell technologies, we wanted to explore the effects of these knockouts within individual cells to better understand how the subtle changes in the undifferentiated state translate to substantial disruptions upon exit from pluripotency.[22] Using our scRNA-seq approach, we observed a global increase in cellular and gene expression variation for all DNMT mutants. As variability has been linked to the ability of a cell to evolve and adapt to a changing environment[1], our results suggest that disruption of DNMTs may increase cellular plasticity. It would therefore be interesting in the future to explore the effects of this by tracking individual cells using molecular barcoding.[36]

We also found two somewhat unexpected effects in the double knockout ES cells. First, we found widespread repression in gene expression upon loss of DNMT3A and DNMT3A/3B in the undifferentiated cells, particularly at genes associated with CpG islands and with H3K27ac super-enhancers. In epidermal stem cells, knockdown of the de novo methyltransferases triggers a reduction of super-enhancer activity[31] and this may occur through a similar mechanism in human ES cells, albeit at different loci. In support of our findings, in epidermal stem cells we also observe 7765 genes that are downregulated versus 2136 upregulated (1.4 fold difference) in the DNMT3A knockdown versus control.[31] Secondly, we do observe a gain of DNA methylation at selected sites in the 3AKO and DKO ES cells. As the latter are derived from the 3AKO this may be a consequence of DNMT3B activity. Known DNMT3B targets include germline genes and it will be interesting to explore how and why these additional loci are targeted in the mutant ES cells.

Differentiation of 3AKO and DKO towards mesoderm and ectoderm showed that the knockout repressed genes were largely inherited from ES cells. We also observed that dME DNMT mutant repressed genes associated with super-enhancers in a mesoderm-specific manner. As core TFs (NANOG, POU5F1 in ES cells; HAND1 in mesoderm) are associated with super-enhancers, we provide evidence that DNMT3A/3B disruption may lead to decreased expression of key cell identity TFs and their downstream targets. Furthermore, we find upregulation of a number of key developmental and oncogenic TFs in 3AKO mesoderm (e.g. MEIS1/2, PBX1, HOXB1, SOX4) and 3AKO ectoderm (SOX11, SOX9, ZIC2, POU4F1, PAX7). DNMT3A is often mutated in human tumors[37], has been shown to act as a first hit mutation[38], and its loss in hematopoietic stem cells and the epidermis promotes leukemia and squamous cell carcinoma formation, respectively.[39,40] It will be interesting in the future to further explore the possible role of increased transcriptional variability in tumor initiation and progression. Taken together, we show that combining scRNA-seq and genetic perturbations presents a powerful tool for dissecting the role of epigenetic regulators in development and disease.

MATERIALS AND METHODS

Appendix E.9 Human ES cell culture

Cell culture was carried out as reported previously.[26] Briefly, we chose the National Institutes of Health-approved, male human ES cell line HUES64 because it has maintained a stable karyotype over many passages and is able to differentiate well into mesoderm and ectoderm. The cells are frequently tested for mycoplasma and identity for the knockout cell lines was confirmed through genotyping PCR. ES cells were maintained on ~15,000 cells/cm^2 irradiated murine embryonic fibroblasts (MEFs, MTI-GlobalStem) and cultured in 20% KnockOut Serum Replacement (KSR, Life Technologies), 200 mM Glutamax (Life Technologies), 1X Minimal Essential Medium (MEM) Non-essential Amino Acids Solution (Life Technologies), 10 µg/ml basic fibroblast growth factor (bFGF, Millipore), 55 µM β-mercaptoethanol in Knockout Dulbecco's Modified Eagle Medium (KO DMEM, Life Technologies). ES cells were passaged every 4-5 days using 1 mg/ml Collagenase IV (Life Technologies). All human ES cell work has been approved by the Harvard University ESCRO (#E00021).

Appendix E.10 Directed differentiation of human ES cells towards mesoderm and ectoderm

When human ES cells reached 60-70% confluency on MEFs, the cells were plated as clumps on 6-well plates coated with Matrigel (Life Technologies) in mTeSR1 basal medium (Stemcell Technologies). We maintained the cells for 3 days in feeder-free culture and then induced directed differentiation towards mesoderm and ectoderm. For the first 24h of mesoderm differentiation, cells were cultured in DMEM/F12 medium supplemented with 100 ng/ml Activin A (R&D Systems), 10 ng/ml bFGF (Millipore), 100 ng/ml BMP4 (R&D Systems), 100 ng/ml VEGF (R&D Systems), 0.5% fetal bovine serum (Hyclone), 200 mM GlutaMax (Life Technologies), 0.2× MEM Non-essential Amino Acids Solution (Life Technologies) and 55 μM β-mercaptoethanol. From 24 to 120 h of mesoderm differentiation, Activin A was removed from the culture. To induce ectoderm differentiation, cells were cultured for 5 days in DMEM/F12 differentiation media supplemented with 2 μM TGFβ inhibitor (Tocris, A83-01), 2 μM WNT3A inhibitor (Tocris, PNU-74654), 2 μM Dorsomorphin BMP inhibitor (Tocris), 55 μM β-mercaptoethanol, 1× MEM Non-essential Amino Acids Solution (Life Technologies) and 15% KOSR (Life Technologies). Media were changed daily. Before inducing differentiation, we manually removed the differentiated cell clumps.

Appendix E.11 Cell collection and fluorescence-activated cell sorting

Cells were treated with StemPro Accutase (Life Technologies, #A1110501) for 5 min, quenched in MEF medium and pelleted using centrifugation [5 min, 1000 rpm (94 g)]. Media was aspirated and cell pellets were washed once in PBS. RNA was immediately stabilized by resuspending the cells in RNAprotect Cell Reagent (~100 μl per 100,000 cells, Qiagen, #76526) and 1μl of RNaseOUT Recombinant Ribonuclease Inhibitor (Life Technologies, #10777-019). Before sorting, cells in RNAprotect Cell Reagent were diluted in ~1.5 ml PBS (pH 7.4; no calcium, no magnesium, no phenol red; Life Technologies, #10010-049). Also, 5 μl of lysis buffer, composed of a 1/500 dilution of Phusion HF buffer (New England Biolabs, #B0518S) was aliquoted in Eppendorf 96-well skirted plates (VWR, #95041-430). Cells were sorted individually in each well of 96-well plates using the FACSAria II flow cytometer (BD Biosciences), avoiding doublets and cell debris. After sorting, plates were immediately sealed, spun down, frozen on dry ice and stored at −80°C.

Appendix E.12 Cell culture, fixation and FISH

Human ES cells were dissociated to single cells using Accutase (Life Technologies, A11105-01), and 30,000 cells were plated per well of 96-well imaging plate coated with Geltrex (Gibco) in mTeSR1 media. The culture media were changed daily and fixed on the third day when cells were ~90% confluent. Before fixing they were stained with 2uM CFSE for 20 min in the incubator. The cells were fixed in 4% formaldehyde solution while covered with aluminum foil for 30 min at room temperature and then dehydrated in 50%, 70% and 100% ethanol for 2 min each concentration. The plates were stored in 100% ethanol in a −20°C chest freezer.

Following fixation, expression levels of three different mRNA transcripts were measured in situ using RNA-FISH probes (Thermo Fisher Scientific) as previously described.[4] Briefly, the ViewRNA ISH Cell Assay Kit (Invitrogen) was performed to stain cells according to the manufacturer's recommendations. Following staining, cells were imaged on an Olympus IX83 inverted microscope using 405 nm excitation for the DAPI stain and 647 nm excitation for the RNA-FISH probes. To quantify RNA expression, single cells were segmented using CellProfiler, and their total probe content was summed over the volume of the cell. Integrated probe intensity box plots were generated to confirm qualitative agreement between RNA-FISH and scRNA-seq.

Appendix E.13 scRNA-seq

Following sorting, 96-well plates of single cells were whole-transcriptome amplified using a Smart-Seq2-based approach, as previously described.[41] Cell lysates were first cleaned with 2.2× volume AMPure XP SPRI beads (Beckman Coulter). Reverse transcription and PCR were then performed on the samples. Following whole-transcriptome amplification, PCR products were cleaned with 0.9× volume SPRI beads and eluted into 20 µl of water. Concentration of cDNA in the resulting solution was determined using a Qubit 3.0 Fluorimeter (Thermo Fisher Scientific) and analyzed using a high sensitivity DNA chip for

BioAnalyzer (Agilent Technologies). Whole-transcriptome amplification products were diluted to a concentration of 0.1 to 0.4 ng/µl and tagmented and amplified using Nextera XT DNA Sample preparation reagents (Illumina). Tagmentation was performed according to the manufacturer's instructions, modified to use one quarter of the recommended volume of reagents, extended tagmentation time to 10 min and extended PCR time to 60 s. PCR primers were ordered from Integrated DNA Technologies. Primer sequences: 3′ SMART CDS Primer IIA: 5′ AAGCAGTGGTATCAACGCAGAGTACT(30)VN; SMARTer II A oligonucleotide: 5′ AAGCAGTGGTATCAACGCAGAGTACATrGrGrG; IS PCR primer: 5′ AAGCAGTGGTATCAACGCAGAGT. Nextera products were then cleaned with 0.9× volume of SPRI beads and eluted in water. The library was quantified using Qubit and analyzed using a high-sensitivity DNA chip. The library was diluted to 2.2 pM and sequenced on a NextSeq 500 (Illumina).

Appendix E.14 Processing of scRNA-seq data

RNA-seq reads were first trimmed using Trimmomatic.[42] Trimmed reads were aligned to the RefSeq hg38 genome and transcriptome (GRCh38.2) using Bowtie2[43] and TopHat[44], respectively. The resulting transcriptome alignments were processed using RSEM to estimate the abundance of RefSeq transcripts[45], in transcripts per million reads mapped (TPM). All cells with fewer than 2000 detectable transcripts (TPM>1) were removed from further analysis. Expression levels for gene i in sample j were quantified as $E_{i,j}$=log(TPM$_{i,j}$+1). Relative expression level for gene i was computed within each subpopulation S as $Er_{i;Sj}$ 1⁄4 $E_{i;Sj}$ –$E^{\wedge}{}_{i;S}$, where $E^{\wedge}{}_{i;S}$ 1⁄4 average1⁄2$E_{i;S1...Sn}$ or the mean expression of that gene across all cells within subpopulation S.

Appendix E.15 Unsupervised dimensionality reduction

To visualize cells in 2-dimensional space, we first performed principal component analysis (PCA) using the Seurat R package version 2.0 as previously described[46] using highly variable genes of mean expression ≥1. We then determined the statistically significant principal components by calculating 1000 random permutations of 1% of genes in the data. We used all significant principal components ($P<10^{-10}$) as input to non-linear dimensionality reduction via t-distributed stochastic neighbor embedding (t-SNE).

Appendix E.16 Classification of cells into ES cells, endoderm, mesoderm and ectoderm

We calculated ES cell, endoderm (dEN), mesoderm (dME) and ectoderm (dEC) scores for all cells by using the AddModuleScore function in Seurat with default parameters for the top 50 most uniquely expressed markers for the ES cell, dEN, dME and dEC purified populations[21] that were also present in the scRNA-seq data. Uniqueness was defined as in previous studies.[25] Cells were then classified into one of four cell types, based on the maximal ES cell, dEN, dME or dEC score. We obtained in silico-sorted populations of ES cells by filtering out all cells collected at day 0 that had an ES cell score ≥max [dEN score, dME score, dEC score]. We defined dEN, dME and dEC in silico-sorted populations similarly.

Appendix E.17 Hierarchical clustering of sorted samples

Relative expression values for all genes was averaged across all ES or dME cells for the defined subpopulations (WT ES cells, 3AKO, DKO, DNMT1$^{-/-}$ day 0, 2, 8). The mean relative expression values were then clustered using hierarchical clustering, average linkage, and 1−Pearson correlation coefficient (r) of all non-zero values as a distance metric.

Appendix E.18 Inter- and intra-sample cell-cell distance

Inter-sample cell-cell distance was computed by comparing all pairs of cells between two samples, using 1−Pearson correlation coefficient (r) of all non- zero values as a distance metric. Intra-sample cell-cell distance was computed by comparing all pairs of cells within a sample using the same distance metric. Cells in each comparison were in silico sorted to contain only ES (**Figures 1, 2, 3 and 5**), dME or dEC cells (**Figure 4**). Before computing the distance, all cells were quantile normalized to control for the total number of transcripts detected per cell. The same approach was applied for other distance metrics (Euclidean, Manhattan and 1-Spearman correlation).

Appendix E.19 Gene expression dispersion analysis

To assay the level of transcriptional variation per gene, we first quantile normalized TPMs for all cells within each sample population and then computed the dispersion or log(variance/mean) for all genes. We also performed quantile normalization before computing other measures of transcript variation.

Appendix E,20 Normalized methylation entropy analysis

To compare the variability of methylation levels at the individual CpG and read level, we used 'informMe', an information-theoretic approach that uses the Ising model of statistical physics to generate mean methylation levels and normalized methylation entropy per CpG. We ran informME for WT, 3AKO and DKO data, for all CpGs located on chromosomes 21 and 22, and used R to plot the respective levels of mean methylation level (MML) and normalized methylation entropy (NME).

Appendix E.21 Three-way differential expression analysis

Differential expression was tested across all possible pairwise comparisons (100, 101, 110, 011, 010 and 001) of three samples, where 1 or 0 indicates high or low expression for the respective sample (e.g. WT, 3AKO and DKO ES cells). To measure differential expression, we used the likelihood-ratio test for single-cell gene expression[27] as implemented in the Seurat R package, requiring a P value$\leq 10^{-8}$ and 1.22-fold change. For ease of visualization, differentially expressed genes were then combined into gene sets representing all possible three-way comparisons (100, 101, 110, 011, 010 and 001) and gene expression was row normalized across cells. Genes were only included in one gene set that had the highest P value in differential expression. The same analysis was performed to compare WT, 3AKO and DKO dME/dEC cells and day 0, 2 and 8 DNMT1-depleted ES cells.

Appendix E.22 Genomic region enrichment analysis

We assessed the significance of overlap of any gene set against a number of predefined genomic regions that can be mapped to their nearest downstream gene. Significance was calculated using the hypergeometric distribution ith Bonferroni correction for multiple hypotheses testing. The resulting P value was −log() transformed and displayed for a number of genomic regions (rows), including CpG density features, epigenetic, and TF binding data collected in matching WT, ES, or dME cells.[21,26] This analysis was performed for gene sets predefined using the three-way differential expression analysis as well as for high and low dispersion set of genes for all samples.

Appendix E.23 Gene set enrichment analysis

Gene sets enrichment analysis (http://software.broadinstitute.org/gsea/) was performed on defined gene sets above selecting only for common pathways from the REACTOME database (http://www.reactome.org/).

Appendix E.24 Cell cycle differential expressed genes and phase classification

To show differentially expressed cell cycle annotated genes, we performed the three-way differential expression analysis as described above solely for genes related to the cell cycle.[47,48] We used a less stringent threshold for displaying cell cycle annotated differentially expressed genes of P value$\leq 10^{-4}$. For visualization, cells (columns) within each sample were ordered according to progress in the cell cycle, as previously described[49], starting with M/G1 cells on the left and ending with G2/M cells on the right. Expression values were averaged using a 20-cell window.

To assign cells according to cell cycle phase, we used a similar approach to that previously described.[50] Briefly, we defined cell cycle phase-specific markers for G1/S, S, G2/M for ES, dME and dEC cells separately, keeping only genes in each predefined cell cycle phase gene set[45] if they had a correlation r≥0.3 with the average gene set expression. The most predictive markers for M/G1 phase cells were key markers with a low expression in the other phases (G1/S, S, G2/M). Cells were quantile normalized in expression, which preserves the order in expression levels between genes within a cell. We then measured the cell cycle phase score of each cell as the average relative expression Er_i, S_j of the selected cell cycle phase markers, where the M/G1 score was multiplied by −1, as it consisted of lowly expressed cell cycle markers for other phases.

We used these scores to assign single cells to phases of the cell cycle, according to their maximal score for the four cell cycle phases.

Appendix E.25 Cell cycle phase-specific expression

Phase-specific expression for each gene i that peaked in expression in phase j (for example j=M/G1) was defined as $E_{i,j} = E_{i,M/G1} = \hat{E}_{i,M/G1}$ – average[$\hat{E}_{i,G1/S}$,$\hat{E}_{i,S}$,$\hat{E}_{i,G2/M}$], where $\hat{E}_{i,j}$ represents the average expression of gene in all cells classified as phase j of the cell cycle for a given sorted population of cells. This analysis was repeated for all genes that peaked in expression in one of four possible phases of the cell cycle (M/G1, G1/S, S and G2/M). Bar plots for cell cycle phase-specific expression in different cell types display the mean phase specific expression for a given gene set (all genes, cell cycle genes or differentially expressed cell cycle genes); error bars represent one standard deviation.

3.26 References

1. Heitzler, P. and Simpson, P. (1991). The choice of cell fate in the epidermis of Drosophila. Cell 64, 1083-1092. doi:10.1016/0092-8674(91)90263-X
2. McAdams, H. H. and Arkin, A. (1997). Stochastic mechanisms in gene expression. Proc. Natl. Acad. Sci. USA 94, 814-819. doi:10.1073/pnas.94.3.814
3. Tanay, A. and Regev, A. (2017). Scaling single-cell genomics from phenomenology to mechanism. Nature 541, 331-338. doi:10.1038/nature21350
4. Shalek, A. K., Satija, R., Adiconis, X., Gertner, R. S., Gaublomme, J. T., Raychowdhury, R., Schwartz, S., Yosef, N., Malboeuf, C., Lu, D. et al. (2013). Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. Nature 498, 236-240. doi:10.1038/nature12172
5. Shalek, A. K., Satija, R., Shuga, J., Trombetta, J. J., Gennert, D., Lu, D., Chen, P., Gertner, R. S., Gaublomme, J. T., Yosef, N. et al. (2014). Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. Nature 510, 363-369. doi:10.1038/nature13437
6. Jaitin, D. A., Kenigsberg, E., Keren-Shaul, H., Elefant, N., Paul, F., Zaretsky, I., Mildner, A., Cohen, N., Jung, S., Tanay, A. et al. (2014). Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. Science 343, 776-779. doi:10.1126/science.1247651
7. Shekhar, K., Lapan, S. W., Whitney, I. E., Tran, N. M., Macosko, E. Z., Kowalczyk, M., Adiconis, X., Levin, J. Z., Nemesh, J., Goldman, M. et al. (2016). Comprehensive classification of retinal bipolar neurons by single-cell transcriptomics. Cell 166, 1308-1323.e30. doi:10.1016/j.cell.2016.07.054
8. Montoro, D. T., Haber, A. L., Biton, M., Vinarsky, V., Lin, B., Birket, S. E., Yuan, F., Chen, S., Leung, H. M., Villoria, J. et al. (2018). A revised airway epithelial

hierarchy includes CFTR-expressing ionocytes. Nature 560, 319. doi:10.1038/ s41586-018-0393-7

9. Treutlein,B., Brownfield,D.G.,Wu,A.R.,Neff,N.F.,Mantalas,G.L.,Espinoza, F. H., Desai, T. J., Krasnow, M. A. and Quake, S. R. (2014). Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. Nature 509, 371-375. doi:10.1038/nature13173
10. Olsson, A., Venkatasubramanian, M., Chaudhri, V. K., Aronow, B. J., Salomonis, N., Singh, H. and Grimes, H. L. (2016). Single-cell analysis of mixed-lineage states leading to a binary cell fate choice. Nature 537, 698-702. doi:10.1038/nature19348
11. Petropoulos, S., Edsgard, D., Reinius, B., Deng, Q., Panula, S. P., Codeluppi, S., Reyes, A. P., Linnarsson, S., Sandberg, R. and Lanner, F. (2016). Single-cell RNA-seq reveals lineage and X chromosome dynamics in human preimplantation embryos. Cell 167, 285. doi:10.1016/j.cell.2016.08.009
12. Tang, F., Barbacioru, C., Bao, S., Lee, C., Nordman, E., Wang, X., Lao, K. and Surani, M. A. (2010). Tracing the derivation of embryonic stem cells from the inner cell mass by single-cell RNA-Seq analysis. Cell Stem Cell 6, 468-478. doi:10. 1016/j.stem.2010.03.015
13. Scialdone, A., Tanaka, Y., Jawaid, W., Moignard, V., Wilson, N. K., Macaulay, I. C., Marioni, J. C. and Gö ttgens, B. (2016). Resolving early mesoderm diversification through single-cell expression profiling. Nature 535, 289-293. doi:10.1038/nature18633
14. Klein, A. M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D. A. and Kirschner, M. W. (2015). Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell 161, 1187-1201. doi:10.1016/j.cell.2015.04.044
15. Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N. J., Livak, K. J., Mikkelsen, T. S. and Rinn, J. L. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat. Biotechnol. 32, 381-386. doi:10.1038/nbt.2859
16. Haghverdi, L., Buettner, F. and Theis, F. J. (2015). Diffusion maps for high-dimensional single-cell analysis of differentiation data. Bioinformatics 31, 2989-2998. doi:10.1093/bioinformatics/btv325
17. Street, K., Risso, D., Fletcher, R. B., Das, D., Ngai, J., Yosef, N., Purdom, E. and Dudoit, S. (2018). Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. BMC Genomics 19, 477. doi:10.1186/s12864-018-4772-0
18. Kumar, R. M., Cahan, P., Shalek, A. K., Satija, R., Daleykeyser, A. J., Li, H., Zhang, J., Pardee, K., Gennert, D., Trombetta, J. J. et al. (2014). Deconstructing transcriptional heterogeneity in pluripotent stem cells. Nature 516, 56-61. doi:10.1038/nature13920
19. Singer, Z. S., Yong, J., Tischler, J., Hackett, J. A., Altinok, A., Surani, M. A., Cai, L. and Elowitz, M. B. (2014). Dynamic heterogeneity and DNA methylation in embryonic stem cells. Mol. Cell 55, 319-331. doi:10.1016/j.molcel.2014.06.029
20. Smith, Z. D. and Meissner, A. (2013). DNA methylation: roles in mammalian development. Nat. Rev. Genet. 14, 204-220. doi:10.1038/nrg3354

21. Gifford, C. A., Ziller, M. J., Gu, H., Trapnell, C., Donaghey, J., Tsankov, A., Shalek, A. K., Kelley, D. R., Shishkin, A. A., Issner, R. et al. (2013). Transcriptional and epigenetic dynamics during specification of human embryonic stem cells. Cell 153, 1149-1163. doi:10.1016/j.cell.2013.04.037

22. Ziller, M. J., Ortega, J. A., Quinlan, K. A., Santos, D. P., Gu, H., Martin, E. J., Galonska, C., Pop, R., Maidl, S., Di Pardo, A. et al. (2018). Dissecting the functional consequences of de novo DNA methylation dynamics in human motor neuron differentiation and physiology. Cell Stem Cell 22, 559-574.e9. doi:10.1016/j.stem.2018.02.012

23. Liao, J., Karnik, R., Gu, H., Ziller, M. J., Clement, K., Tsankov, A. M., Akopian, V., Gifford, C. A., Donaghey, J., Galonska, C. et al. (2015). Targeted disruption of DNMT1, DNMT3A and DNMT3B in human embryonic stem cells. Nat. Genet. 47, 469-478. doi:10.1038/ng.3258

24. Picelli, S., Faridani, O. R., Bjö rklund, A. K., Winberg, G., Sagasser, S. and Sandberg, R. (2014). Full-length RNA-seq from single cells using Smart-seq2. Nat. Protoc. 9, 171-181. doi:10.1038/nprot.2014.006

25. Tsankov, A. M., Akopian, V., Pop, R., Chetty, S., Gifford, C. A., Daheron, L., Tsankova, N. M. and Meissner, A. (2015a). A qPCR ScoreCard quantifies the differentiation potential of human pluripotent stem cells. Nat. Biotechnol. 33, 1182-1192. doi:10.1038/nbt.3387

26. Tsankov, A. M., Gu, H., Akopian, V., Ziller, M. J., Donaghey, J., Amit, I., Gnirke, A. and Meissner, A. (2015b). Transcription factor binding dynamics during human ES cell differentiation. Nature 518, 344-349. doi:10.1038/nature14233

27. Mcdavid, A., Finak, G., Chattopadyay, P. K., Dominguez, M., Lamoreaux, L., Ma, S. S., Roederer, M. and Gottardo, R. (2013). Data exploration, quality control and testing in single-cell qPCR-based gene expression experiments. Bioinformatics 29, 461-467. doi:10.1093/bioinformatics/bts714

28. Hansen, K. D., Timp, W., Bravo, H. C., Sabunciyan, S., Langmead, B., Mcdonald, O. G., Wen, B., Wu, H., Liu, Y., Diep, D. et al. (2011). Increased methylation variation in epigenetic domains across cancer types. Nat. Genet. 43, 768. doi:10.1038/ng.865

29. Jenkinson, G., Pujadas, E., Goutsias, J. and Feinberg, A. P. (2017). Potential energy landscapes identify the information-theoretic nature of the epigenome. Nat. Genet. 49, 719. doi:10.1038/ng.3811

30. Teschendorff, A. E. and Widschwendter, M. (2012). Differential variability improves the identification of cancer risk markers in DNA methylation studies profiling precursor cancer lesions. Bioinformatics 28, 1487-1494. doi:10.1093/bioinformatics/bts170

31. Rinaldi, L., Datta, D., Serrat, J., Morey, L., Solanas, G., Avgustinova, A., Blanco, E., Pons, J. I., Matallanas, D., von Kriegsheim, A. et al. (2016). Dnmt3a and Dnmt3b associate with enhancers to regulate human epidermal stem cell homeostasis. Cell Stem Cell 19, 491-501. doi:10.1016/j.stem.2016.06.020

32. Hnisz, D., Abraham, B. J., Lee, T. I., Lau, A., Saint-André, V., Sigova, A. A., Hoke, H. A. and Young, R. A. (2013). Super-enhancers in the control of cell identity and disease. Cell 155, 934-947. doi:10.1016/j.cell.2013.09.053

33. Pauklin, S. and Vallier, L. (2013). The cell-cycle state of stem cells determines cell fate propensity. Cell 155, 135-147. doi:10.1016/j.cell.2013.08.031

34. Pauklin, S., Madrigal, P., Bertero, A. and Vallier, L. (2016). Initiation of stem cell differentiation involves cell cycle-dependent regulation of developmental genes by Cyclin D. Genes Dev. 30, 421-433. doi:10.1101/gad.271452.115

35. Yoshioka, H., Meno, C., Koshiba, K., Sugihara, M., Itoh, H., Ishimaru, Y., Inoue, T., Ohuchi, H., Semina, E. V., Murray, J. C. et al. (1998). Pitx2, a bicoid-type homeobox gene, is involved in a lefty-signaling pathway in determination of left-right asymmetry. Cell 94, 299-305. doi:10.1016/S0092-8674(00)81473-7

36. Chan, M. M., Smith, Z. D., Grosswendt, S., Kretzmer, H., Norman, T. M., Adamson, B., Jost, M., Quinn, J. J., Yang, D. and Jones, M. G. (2019). Molecular recording of mammalian embryogenesis. Nature 570, 77-82. doi:10. 1038/s41586-019-1184-5

37. Kim, M.S., Kim, Y.R., Yoo, N.J., and Lee, S.H. (2013). Mutational analysis of DNMT3A gene in acute leukemias and common solid cancers. *APMIS* **121**, 85-94. Doi: 10.1111/j.1600-0463.2012.02940.x

38. Shlush, L. I., Zandi, S., Mitchell, A., Chen, W. C., Brandwein, J. M., Gupta, V., Kennedy, J. A., Schimmer, A. D., Schuh, A. C., Yee, K. W. et al. (2014). Identification of pre-leukaemic haematopoietic stem cells in acute leukaemia. Nature 506, 328. doi:10.1038/nature13038

39. Rinaldi,L.,Avgustinova,A.,Martıń,M.,Datta,D.,Solanas,G.,Prats,N.and Benitah, S. A. (2017). Loss of Dnmt3a and Dnmt3b does not affect epidermal homeostasis but promotes squamous transformation through PPAR-γ. eLife 6, e21697. doi:10.7554/eLife.21697

40. Yang, L., Rodriguez, B., Mayle, A., Park, H. J., Lin, X., Luo, M., Jeong, M., Curry, C. V., Kim, S.-B., Ruau, D. et al. (2016). DNMT3A loss drives enhancer hypomethylation in FLT3-ITD-associated leukemias. Cancer Cell 29, 922-934. doi:10.1016/j.ccell.2016.05.003

41. Trombetta, J. J., Gennert, D., Lu, D., Satija, R., Shalek, A. K. and Regev, A. (2014). Preparation of single-cell RNA-seq libraries for next generation sequencing. Curr. Protoc. Mol. Biol. 107, 4.22.1-4.22.17. doi:10.1002/ 0471142727.mb0422s107

42. Bolger, A. M., Lohse, M. and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114 -2120. doi:10.1093/bioinformatics/btu170 .

43. Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357-359. doi:10.1038/nmeth.1923

44. Trapnell, C., Pachter, L. and Salzberg, S. L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25, 1105-1111. doi:10.1093/ bioinformatics/btp120

45. Li, B. and Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA- Seq data with or without a reference genome. BMC Bioinformatics 12, 323. doi:10. 1186/1471-2105-12-323

46. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. and Regev, A. (2015). Spatial reconstruction of single-cell gene expression data. Nat. Biotechnol. 33, 495-502. doi:10.1038/nbt.3192

47. Whitfield, M. L., Sherlock, G., Saldanha, A. J., Murray, J. I., Ball, C. A., Alexander, K. E., Matese, J. C., Perou, C. M., Hurt, M. M., Brown, P. O. et al. (2002). Identification of genes periodically expressed in the human cell cycle and their expression in tumors. Mol. Biol. Cell 13, 1977-2000. doi:10.1091/mbc.02-02- 0030

48. Kanehisa, M. and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 28, 27-30. doi:10.1093/nar/28.1.27

49. Kowalczyk, M. S., Tirosh, I., Heckl, D., Rao, T. N., Dixit, A., Haas, B. J., Schneider, R. K., Wagers, A. J., Ebert, B. L. and Regev, A. (2015). Single-cell RNA-seq reveals changes in cell cycle and differentiation programs upon aging of hematopoietic stem cells. Genome Res. 25, 1860-1872. doi:10.1101/gr.192237.115

50. Tirosh, I., Izar, B., Prakadan, S. M., Wadsworth, M. H., II, Treacy, D., Trombetta, J. J., Rotem, A., Rodman, C., Lian, C., Murphy, G. et al. (2016). Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science 352, 189-196. doi:10.1126/science.aad0501

Supplemental Information

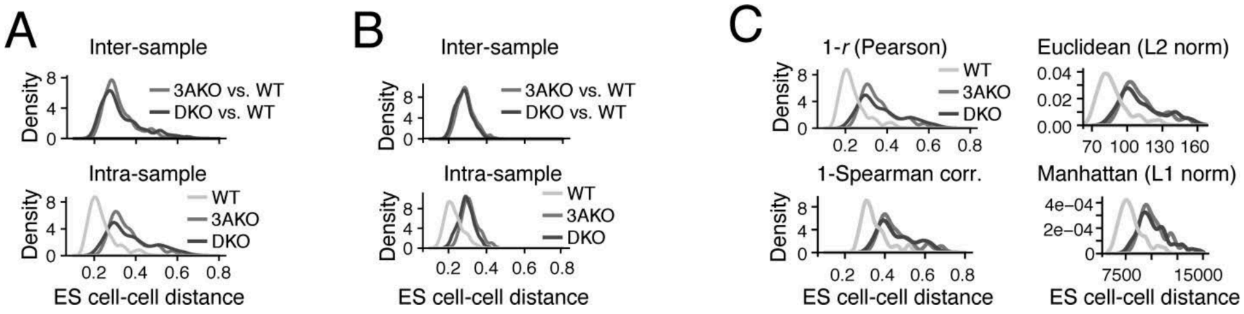


Figure S1 supporting Figure 1 | Increased cellular variation in DNMT3A and DNMT3A/3B knockout ES cells. (A) Inter-sample (**top**) and intra-sample (**bottom**) density distribution of pairwise cell-cell distances (1-Pearson correlation coefficient) for in silico sorted undifferentiated WT (n = 162), 3AKO (n = 74), and DKO cells (n = 74). **(B)** Inter-sample (**top**) and intra-sample (**bottom**) density distribution of pairwise cell-cell distances (1-Pearson correlation coefficient) for only the highest quality cells (number of genes detected > 7,000) for wildtype (n = 149), 3AKO (n = 56), and DKO (n = 58) ES cells. **(C)** Intra-sample density distribution of pairwise cell-cell distances in *in silico* sorted undifferentiated WT (n = 162), 3AKO (n = 74), and DKO cells (n = 74) using four different distances: 1- Pearson correlation coefficient (**top left**), 1- Spearman rank correlation (**bottom left**), Euclidean L2 norm (**top right**) and Manhattan L1 norm (**bottom right**).

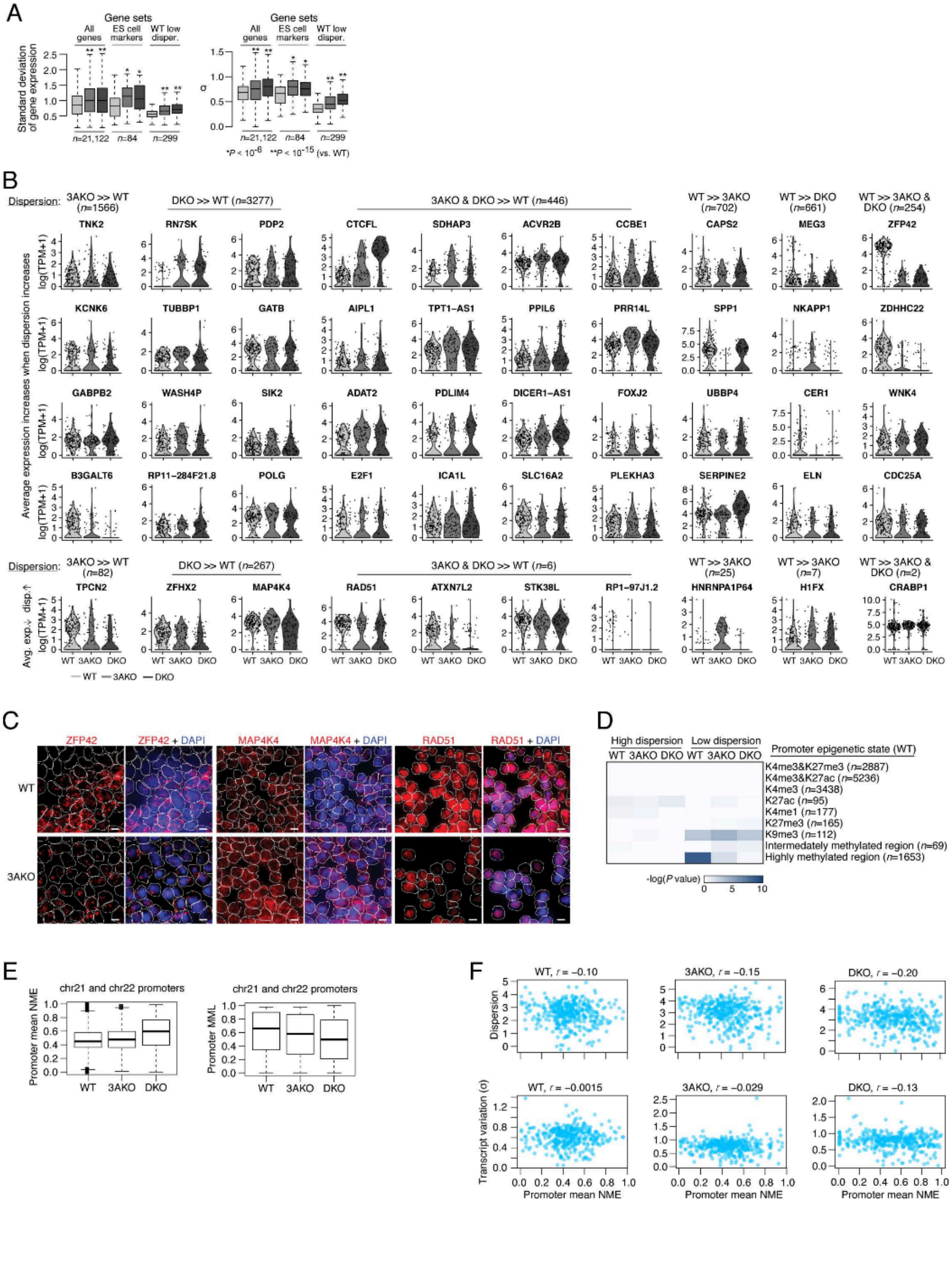

A
Gene sets
All genes
ES cell markers
WT low disper.
Standard deviation of gene expression
n=21,122
n=84
n=299
σ
*P < 10^-6 **P < 10^-15 (vs. WT)
B
Dispersion: 3AKO >> WT (n=1566)
DKO >> WT (n=3277)
3AKO & DKO >> WT (n=446)
WT >> 3AKO (n=702)
WT >> DKO (n=661)
WT >> 3AKO & DKO (n=254)
TNK2
RN7SK
PDP2
CTCFL
SDHAP3
ACVR2B
CCBE1
CAPS2
MEG3
ZFP42
KCNK6
TUBBP1
GATB
AIPL1
TPT1−AS1
PPIL6
PRR14L
SPP1
NKAPP1
ZDHHC22
GABPB2
WASH4P
SIK2
ADAT2
PDLIM4
DICER1−AS1
FOXJ2
UBBP4
CER1
WNK4
B3GALT6
RP11−284F21.8
POLG
E2F1
ICA1L
SLC16A2
PLEKHA3
SERPINE2
ELN
CDC25A
Average expression increases when dispersion increases
log(TPM+1)
Dispersion: 3AKO >> WT (n=82)
DKO >> WT (n=267)
3AKO & DKO >> WT (n=6)
WT >> 3AKO (n=25)
WT >> 3AKO (n=7)
WT >> 3AKO & DKO (n=2)
TPCN2
ZFHX2
MAP4K4
RAD51
ATXN7L2
STK38L
RP1−97J1.2
HNRNPA1P64
H1FX
CRABP1
Avg. exp.↓ disp.↑
WT 3AKO DKO
— WT — 3AKO — DKO
C
ZFP42
ZFP42 + DAPI
MAP4K4
MAP4K4 + DAPI
RAD51
RAD51 + DAPI
WT
3AKO
D
High dispersion
Low dispersion
WT 3AKO DKO WT 3AKO DKO
Promoter epigenetic state (WT)
K4me3&K27me3 (n=2887)
K4me3&K27ac (n=5236)
K4me3 (n=3438)
K27ac (n=95)
K4me1 (n=177)
K27me3 (n=165)
K9me3 (n=112)
Intermedately methylated region (n=69)
Highly methylated region (n=1653)
-log(P value)
0 5 10
E
chr21 and chr22 promoters
Promoter mean NME
chr21 and chr22 promoters
Promoter MML
WT 3AKO DKO
F
WT, r = −0.10
3AKO, r = −0.15
DKO, r = −0.20
Dispersion
WT, r = −0.0015
3AKO, r = −0.029
DKO, r = −0.13
Transcript variation (σ)
Promoter mean NME

Figure S2 supporting Figure 2 | Relationship between DNA methylation level, mean methylation entropy and transcript variation in DNMT3A and DNMT3A/3B knockouts. (A) Box plots of gene expression standard deviation computed across all cells (**left**) and only among cells with detectable gene expression (σ, **right**) for gene sets composing of all genes, ES cell markers, and WT low dispersion genes for WT, 3AKO, and DKO ES cells. Boxes display the interquartile range while the bold line shows the median and whiskers extend to the most extreme data point that is no more than 1.5 times the interquartile range. **(B)** Violin plots of log gene expression level, log(TPM+1), for 50 selected genes that have a difference in dispersion greater than 1.5 between two samples, where the samples being compared are annotated using column headers on the top and the overall number of genes present in each category is shown in parentheses. The change in average expression relative to dispersion is annotated along rows on the left. The majority of genes (>90%) that increase in dispersion also increase in average expression. TPM = transcripts per million fragments mapped. **(C)** Representative images of RNA FISH experiment showing staining for DAPI (blue) and red fluorescent probes targeting ZFP42 (**left**), MAP4K4 (**middle**) and RAD51 (**right**) in WT (**top**) and 3AKO (**bottom**) ES cells. Cell segmentation is shown using white outlines. White bar in bottom right corner of each panel indicates a distance of 10 microns. **(D)** Genomic enrichment analysis for high (left) and low (right) transcript dispersion genes in WT, 3AKO, and DKO sorted ES cells overlapped with the promoter epigenetic state of matching WT ES cells.[21,26] We observe a high enrichment of highly methylated promoter regions at low dispersion WT genes but this enrichment decreases for low dispersion 3AKO and DKO genes. **(E)** Boxplot of the promoter mean normalized methylation entropy (NME; left) and mean methylation level (MML; **right**) measured for WT, 3AKO, and DKO ES cell WGBS data for all chromosome 21 and 22 promoters using the approach in (Jenkinson et al., 2017). Boxes display the interquartile range while the bold line shows the median and whiskers extend to the most extreme data point that is no more than 1.5 times the interquartile range. **(F)** Correlation scatter plots of transcriptional variation measured in terms of dispersion (top) and standard deviation (σ) of detectable transcripts (**bottom**) versus promoter mean normalized methylation entropy for all WT (**left**), 3AKO (**middle**) and DKO (**right**) promoters on chromosomes 21 and 22.

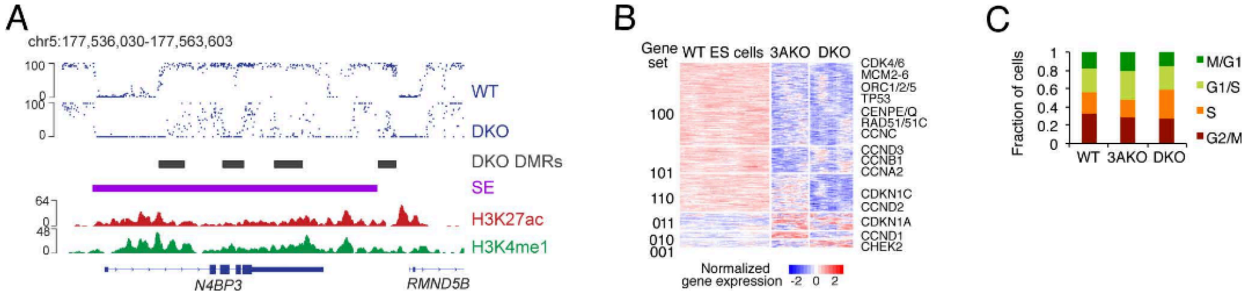


Figure S3 supporting Figure 3 | Widespread transcriptional repression and changes in cell cycle gene expression in DNMT3A and DNMT3A/B knockout ES cells. (A) Browser tracks display methylation levels for WT and DKO cells over a 28kb region on chromosome 5. Grey bars highlight DKO-specific differentially methylated regions (DMRs; difference > 0.6, P < 0.01). An ES cell super-enhancer (Hnisz et al., 2013) is highlighted in purple with ENCODE ChIP-seq data for H3K27ac and H3K4me1 in H1 ES cells displayed below. CpGs located within the super-enhancer region lose substantial methylation upon loss of DNMT3A and 3B. **(B)** Differentially expressed cell cycle annotated genes (**right; rows**) for sorted population of WT, 3AKO, and DKO ES cells (**columns**) ordered by progressin the cell cycle. Gene sets (**left**) are defined in Figure 2A. **(C)** Fraction of cells in M/G1, G1/S, S, and G2/M phase for in silico sorted WT, 3AKO, and DKO ES cell populations.

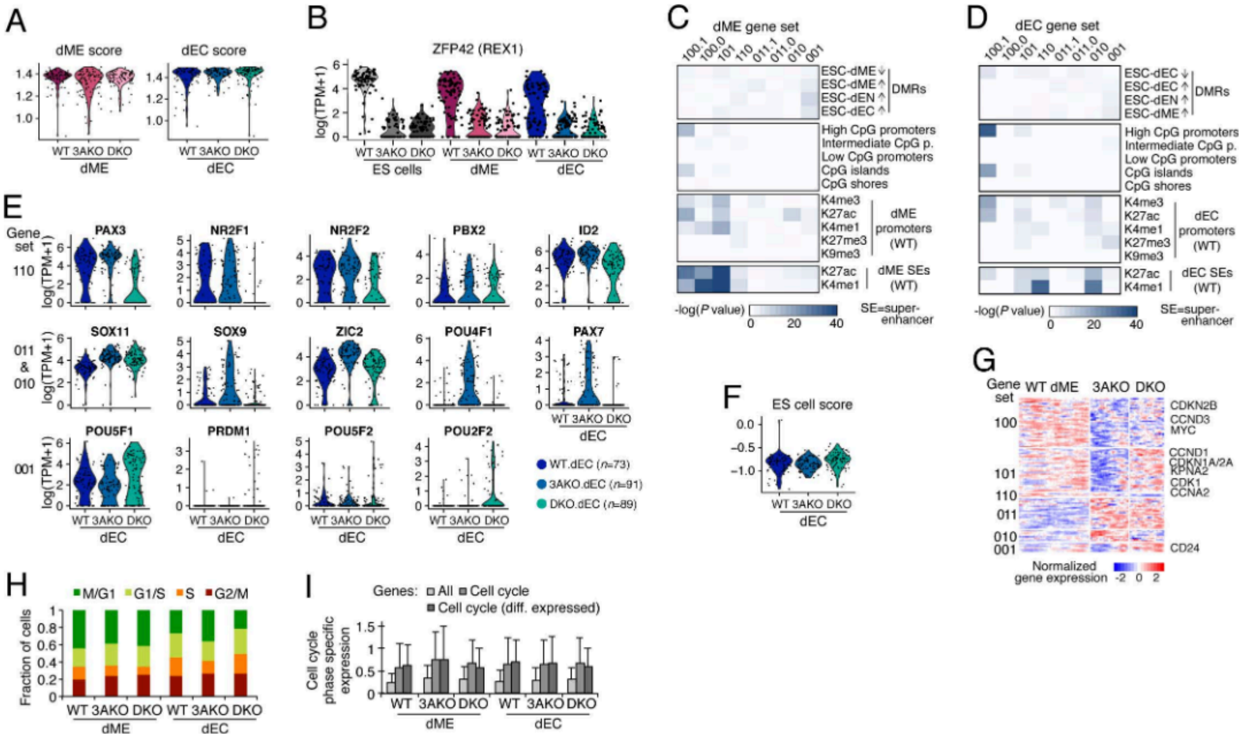

A
dME score
dEC score
WT 3AKO DKO
dME
dEC
B
ZFP42 (REX1)
log(TPM+1)
WT 3AKO DKO
ES cells
dME
dEC
C
dME gene set
100,1 100,0 101 110 011,1 011,0 010 001
ESC-dME↓
ESC-dME↑
ESC-dEN↑
ESC-dEC↑
DMRs
High CpG promoters
Intermediate CpG p.
Low CpG promoters
CpG islands
CpG shores
K4me3
K27ac
K4me1
K27me3
K9me3
dME promoters (WT)
K27ac
K4me1
dME SEs (WT)
-log(P value)
0 20 40
SE=super-enhancer
D
dEC gene set
ESC-dEC↓
ESC-dEC↑
ESC-dEN↑
ESC-dME↑
dEC promoters (WT)
dEC SEs (WT)
E
Gene set
110
011 & 010
001
log(TPM+1)
PAX3
NR2F1
NR2F2
PBX2
ID2
SOX11
SOX9
ZIC2
POU4F1
PAX7
POU5F1
PRDM1
POU5F2
POU2F2
WT 3AKO DKO
dEC
WT.dEC (n=73)
3AKO.dEC (n=91)
DKO.dEC (n=89)
F
ES cell score
G
Gene set
WT dME
3AKO
DKO
100
101
110
011
010
001
CDKN2B
CCND3
MYC
CCND1
CDKN1A/2A
KPNA2
CDK1
CCNA2
CD24
Normalized gene expression
-2 0 2
H
M/G1
G1/S
S
G2/M
Fraction of cells
WT 3AKO DKO WT 3AKO DKO
dME
dEC
I
Genes:
All
Cell cycle
Cell cycle (diff. expressed)
Cell cycle phase specific expression

Figure S4 supporting Figure 4 | Transcriptional misregulation in DNMT3A/B knockout cells following mesoderm differentiation. (A) Violin plot of mesoderm (**left**) and ectoderm (**right**) scores for WT, 3AKO, and DKO cells following 5 days of differentiation towards mesoderm and ectoderm, respectively. Each dot represents a cell. **(B)** Distribution of ZFP42 expression for in silico sorted WT, 3AKO, and DKO ES (**left**), mesoderm (**middle**), and ectoderm (**right**) cells. **(C)** Genomic enrichment analysis for gene sets (**columns**) defined in Figure 4F against DNA methylation, CpG density features, and chromatin data collected in matching WT dME cells.[21,26] DMR = differentially methylated region; K = lysine histone 3; me3 = tri-methylation; ac = acetylation; me1 = mono-methylation. **(D)** Genomic enrichment analysis for gene sets (**columns**) defined in Figure 4G against DNA methylation, CpG density features, and chromatin data collected in matching WT dEC cells. **(E)** Violin plots of log(TPM+1) gene expression for key developmental and oncogenic TFs misregulated in dEC 3AKO and/or DKO mutants. TFs displayed were either downregulated in DKO (top row; gene set 110), upregulated in 3AKO (**middle row**; gene sets 011 & 010), or upregulated in DKO (**bottom row**; gene set 001). **(F)** Violin plot of ES cell scores for WT, 3AKO, and DKO cells following 5 days of differentiation towards ectoderm. DKO dEC sample has a higher median and standard deviation in ES cell scores. **(G)** Differentially expressed cell cycle annotated genes (right; rows) for sorted population of WT, 3AKO, and DKO dME cells (**columns**) ordered by progress in the cell cycle. Gene sets (**left**) are defined in panel Figure 4F. **(H)** Fraction of cells in M/G1, G1/S, S, and G2/M phase for sorted WT, 3AKO, and DKO dME (**left**) and dEC (**right**) cell populations. **(I)** Distribution of cell cycle phase specific expression for sorted WT, 3AKO, and DKO dME (**left**) and dEC (**right**) cells considering all genes, cell cycle annotated genes, and differentially expressed cell cycle annotated genes. Error bars indicate one standard deviation.

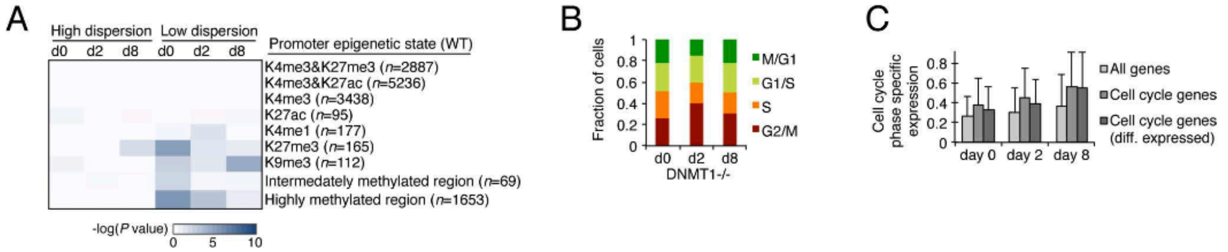


Figure S5 supporting Figure 5 | Loss of DNMT1 triggers increased transcript variation and differentiation. (A) Genomic enrichment analysis for high (**left**) and low (**right**) transcript dispersion genes at day 0, 2, and 8 sorted ES cells following DOX treatment overlapped with the promoter epigenetic state of WT HUES64 ES cells.[21,26] We observe a high enrichment of highly methylated promoter regions at day 0 low dispersion genes but this enrichment gradually decreases for low dispersion day 2 and day 8 genes while the enrichment at H3K9me3 promoters remains. **(B)** Fraction of cells in M/G1, G1/S, S, and G2/M phase for in silico sorted ES cells at day 0, 2, and 8. **(C)** Distribution of cell cycle phase specific expression for day 0, 2, and 8 sorted ES cells considering all genes, known cell cycle associated genes, and known, differentially expressed cell cycle annotated genes. Error bars indicate one standard deviation.

Table S1 | Differentially expressed genes in wildtype (WT), DNMT3A-/- (3AKO) and DNMT3A/B-/- (DKO) ES cells. Three-way differentially expressed genes (rows in spreadsheet “Markers”) for sorted population of WT, 3AKO, and DKO ES cells, displayed in Figure 2A. Genes are separated into 6 clusters (100, 101, 110, 011, 010, and 001), where 1 or 0 indicates high or low expression for the respective condition (order: WT, 3AKO, DKO). Spreadsheets “100” to “001” contain functional enrichment analysis for genes in each cluster from spreadsheet “Markers” against the REACTOME database.