A Single-Cell Immune Map of Normal and Cancerous Breast Reveals an Expansion of Phenotypic States Driven by the Tumor Microenvironment

Ambrose J. Carr

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2018

ABSTRACT

A Single-Cell Immune Map of Normal and Cancerous Breast Reveals an Expansion of Phenotypic States Driven by the Tumor Microenvironment

Ambrose J. Carr

Knowledge of the phenotypic states of immune cells in the tumor microenvironment is essential to understand immunological mechanisms of cancer progression, responses to cancer immunotherapy, and the development of novel rational treatments. Yet, this knowledge is opaque to traditional bulk sequencing methods, and novel single-cell RNA sequencing (scRNA-seq) methods which could potentially address these questions introduce complex patterns of error into data that are poorly characterized. This dissertation describes a computational framework, SEQC, built to facilitate rapid and agile analysis of scRNA-seq approaches that utilize molecular barcodes. It combines SEQC with a clustering and normalization method, BISCUIT, and approaches to examine phenotypic diversity and gene variation. These methods are applied to address the unique computational challenges inherent to analysis of single-cell RNA-seq data derived from multiple patients with diverse phenotypes. This dissertation describes an experiment comprising scRNA-seq of over 47,000 immune cells collected from primary breast carcinomas, matched normal breast tissue, peripheral blood, and using these computational approaches. This atlas revealed significant similarity between normal and tumor tissue resident immune cells. However, it also describes continuous tumor-specific phenotypic expansions driven by distinct environmental cues. These results argue against discrete activation states in T cells and the polarization model of macrophage activation in cancer, and have important implications for characterizing tumor-infiltrating immune cells.

Contents

Acknowledgements

Before beginning my PhD, I had never written a single line of code. My studies had woven through Neuroscience and Molecular Biology, but my contributions were exclusively made at the lab bench. To this day I have no idea what Dana saw in me that suggested that I would be successful in a computational PhD, but I am indebted to her for the patience she has shown and mentoring she has provided. Always available to provide advice and right my course when I digressed, these projects would not have been possible without Dana's support, insight, and connections.

I also want to thank Prof. Aviv Regev and now-professors Dr. Rahul Satija and Dr. Alex Shalek, who, through generosity of time and expertise, both helped shape my impressions of nascent single-cell sequencing, and also opened my eyes to the open questions, motivating much of the work contained in this dissertation.

Like many dissertation, this one summarizes work that contains contributions from many researchers. None of these experiments would be possible if not for the hard work of my tireless wet-lab colleagues: Dr. Linas Mazutis, Juozas Nainys, and Vaidotas Kiseliovas, who carried out all of the sequencing experiments described, were instrumental in providing feedback on the computational approaches I developed, and implemented the modifications to the barcodes and

InDrop procedures discussed in Chapter 2. Similarly, Thank you to Dr. Manu Setty, Dr. The Phuong Dao, and Kristy Choi for their programming contributions to SEQC, and to George Plitas, the surgeon who obtained all of the samples for these experiments, and who provided an abundance of biological insight. Thanks to Sandhya Prabhakaran and Elham Azizi, who transformed my simple ideas about how to combine data from disparate scRNAseq patient samples into a rich Bayesian hierarchical model that I would never have been able to construct myself. Their contributions made much of the work discussed in Chapter 3 possible. Finally, thank you to Elham Azizi and Andrew Cornish for their help with the analysis of the processed Breast Cancer Immune Atlas, described in Chapter 4.

As no experiments can proceed in a vacuum, I am grateful for the insight and criticism provided by my committee members Prof. Harmen Bussemaker and Prof. Peter Sims, and for Prof. Rahul Satija and Prof. Alexander Rudensky, for serving on my defense committee. I am also grateful to the Howard Hughes Medical Institute, who generously funded me during my PhD at Columbia University, and to all of my collaborators: Prof. Peter Sims, Prof. Alexander Rudensky, Prof. Frederic Geissmann, Dr. George Plitas, and Kasia Konnopacki for pushing my research in directions I would not have explored on my own. Similarly, I am grateful to Dr. Elham Azizi, Dr El-ad David Amir, Dr. Jacob Levine, Dr. Sandhya Prabhakaran, Andrew Cornish, Dr. Manu Setty, Cassandra Burdziak, Roshan Sharma, and Kristy Choi for their help editing this dissertation and for their insight and friendship, which made coming to work easy.

Finally, thank you to my family, and my wife, Sharonmoyee Goswami, for their support throughout my PhD.

Introduction

Tumors result from aberrant uncontrolled growth of human cells. However, it is understood that individual cells within a tumor have uneven responsibility for patient mortality. A large fraction of the cells are often terminally differentiated—unable to continue dividing. As a result, limited numbers of cancer cells are capable of moving outside their tissue of origin, or embedding themselves in other tissues. Heterogeneity of this type presents a challenge to the personalization of cancer treatment, as until recently there were no tools to find these rare but dangerous cells within the larger population of mostly-benign cancer cells without already knowing surface markers that specifically identify them. As a result, scientists and clinicians were limited to profiling cell types in aggregate, which produced an artificial, blurred average of all the cells in a tumor. These results, which reflect no true cell state, proved only weakly predictive of cancer outcomes.

The deficiency of experimental approaches to characterize novel cell phenotypes has significantly hampered the development of promising cancer treatments such as immunotherapies, a category of biologic drugs that relieve immunosuppression. Currently, bio-markers that predict treatment success are the presence of highly mutated surface proteins on a patient's tumor cells, abundant tumor infiltrating immune cells (TILs), and the evidence of active immunosuppression, driven by the pharmacologically targeted pathways. Unfortunately, these bio-markers have low predictive power and cohorts selected for trials based on these characteristics display polarizing results: a small subset of patients displays an extremely exciting complete remission, but the great majority of patients still fail treatment. There is good reason to suspect that some of these remissions may be permanent, suggesting that for a small number of patients, the immune system holds the keys to curing them of cancer.

Yet, the majority of patients fail treatment, and the reasons for this failure are not clear. As a result, these trials are ongoing across cancers, and rare success cases across many cancer types hint that the immune system harbors an unharnessed and systemic capability to recognize, eliminate, and provide lasting immunity against cancer. However, a significantly uneven ratio of success and failure suggests that there is hidden complexity in the tumor-immune ecosystem that extends beyond the systems that are currently being targeted by drugs.

Because cancer treatments all kill human cells at some low rate, they carry significant toxicity. Thus, to maximize patient survival, it is imperative that clinicians pair patients with rational treatments based on the molecular characteristics of their tumors, as opposed to the current standard which often involves cocktails of drugs designed to, on average, have the best effect across patients.

Better characterization of immunosuppression systems may enable this kind of personalization for immunotherapy. Unfortunately, sequencing bulk tumor or immune isolates cannot identify the patterns of immune suppression acting on individual cells, and as a result, has not been an effective tool for personalizing cancer therapies against individual tumors. To resolve this problem, researchers originally turned to cell sorting strategies that partition cells based on surface proteins to identify and understand individual immune cells and immune cell populations.

While powerful, these strategies are limited to deciphering a-priori known cell types or states, which are defined by relatively simple combinations of surface proteins. In contrast, single-cell RNA-seq (scRNA-seq) merges many of the strengths of fluorescence cell sorting and bulk RNA-seq. scRNA-seq retains the unbiased measurement of all expressed genes, but adds the ability to resolve phenotypes of individual cells. This unbiased, transcriptome-wide analysis approach has accelerated cell state discovery and enabled, for the first time, unbiased measurement of large-scale population interactions. However, the technology is fraught with new technical and computational challenges, including a low signal to noise ratio and a low probability of capturing and observing mRNA in cells. These problems have prevented its easy application to critical biological questions.

This dissertation will describe, in 4 chapters, technical and computational development to support scRNA-seq and the application of scRNA-seq to better understand how variance in the states of TILs may explain the clinically observed variability in treatment results. Chapter one will review the literature of immuno-oncology and single-cell transcriptomics to contextualize the application of single-cell technologies to TILs. Chapter two will examine droplet-based scRNA-seq in detail, highlighting SEQC, a framework developed to control technical variances and produce a cleaner view of TIL biology, while enabling rapid iteration of library construction to improve data quality. Chapter three will address how new algorithms can help us generalize to an understanding of phenotypic states of populations of cells, despite the data's high, 95% sparsity and its derivation from multiple patients. Chapter four will combine the methods from chapters two and three and apply them to a large set of more than 45,000 breast-carcinoma infiltrating immune cells, characterizing TILs, but also distinguishing them from normal tissue-resident and blood-resident immune phenotypes. Finally, the dissertation will conclude with a discussion of the outstanding challenges surrounding single-cell analysis of immuno-oncology and the single-cell technologies themselves, and will point to promising directions for new research.

Chapter 1

Review of Single-Cell Immuno-Oncology

1.1 Immune-Mediated Cancer Cell Killing

The adaptive immune system evolved to combat exogenous threats such as viruses and bacteria. However, it is increasingly thought to also play an important role in cancer surveillance because, for a tumor to form and proliferate, it must evade recognition by the immune system (Corthay, 2014). In a normal immune response, the primary surveillance cells that are tasked with identifying foreign threats are called Dendritic cells. Dendritic cells warn other immune cells of detected threats by presenting antigens—small pieces of foreign molecules—using a system of surface receptors called major histocompatibility complex I and II (MHC-I, MHC-II). Any antigen presented on an MHC molecule of an active Dendritic cell is presumed to identify a molecule or cell that should be eliminated by the immune system (Joffre et al., 2009). In contrast, antigens presented by inactive Dendritic cells are considered markers of human molecules or cells, and suppress immune responses against cells carrying molecules that match the presented antigen. Thus, it is of critical importance that (1) any tumor antigen presented on a Dendritic cell specifically mark a *foreign* cell type, as antigens shared by normal cells risk turning the immune system on healthy human cells, and (2) that Dendritic cells bearing tumor antigens are activated (Darrasse-Jèze et al., 2009; Steinman et al., 2000).

In cancer, specific tumor antigens can be mutated proteins, products of non-mutated proteins that have much higher over-expression in tumor cells, or differentiation antigens associated with tissue of origin; for example melanosome proteins in melanoma (Boon et al., 2006; Segal et al., 2008). However, the latter case will produce an auto-immune reaction in addition to destroying cancer cells because the antigen is partially shared with normal cells. To prevent widespread auto-immunity, Dendritic cell activation is carefully controlled. Activation signals are plentiful, including many signaling molecules that are generated during inflammation or innate immune responses. These signals include CD40, IFN-α Toll-like receptor stimulation, or GM-CSF (Lippitz, 2013). Dying cells, another byproduct of an innate immune response, are thought to release additional signals that result in Dendritic cell activation and maturation, including high mobility group proteins, ATP, and ER proteins like calreticulin (Zitvogel and Kroemer, 2009).

Once a Dendritic cell has been exposed to a foreign antigen and activated, the second step of an immune response is the migration of activated Dendritic cells to lymphoid organs. In a lymphoid organ, a tumor-antigen loaded Dendritic cell interacts with and activates CD8-expressing T cells that carry a T cell receptor (TCR) that recognizes the antigen presented by the Dendritic cell. Once a T cell is activated, it is primed to recognize the antigen that was presented to it by the Dendritic cell, and is designed to interact with such cells and induce them to die. As a result, T cell activation is also under tight control. Activation relies upon a host of pro- and anti-activation interactions. The positive signals include CD28, CD137/4-1BB, OX40, HVEM, GITR, IL-2, and IL12 (Franciszkiewicz et al., 2012; Lippitz, 2013), while the negative signals include CTLA-4, TIM-3, VISTA, and LAG-3 (Sharma et al., 2017). The combination of these systems with diverse activation signals produces a complex set of possible responses and many functionally different T cell states.

Finally, active, cancer-specific T cells must enter the tumor to find and destroy tumor cells. Chemokines CX3CL1, CXCL9, CXCL10, CCL5 all encourage trafficking of T cells to tumors (Franciszkiewicz et al., 2012). At the tumor site, regulatory T cells are produced by interacting with inactive, antigen-presenting Dendritic cells. These cells must be present in lower frequencies than effector T cells for cancer cell killing to proceed. The goal of anti-tumor response is to produce a self-sustaining immune cycle wherein the killing of cancer cells produces antigens and activation signals for Dendritic cells, which in turn provoke the subsequent stages of the immune cycle (Chen and Mellman, 2013).

How then do tumors that express large numbers of antigens avoid immune-mediated cell killing? It has become clear that immunosuppression occurs at each stage of the immune response, and that there are a variety of actors that play a part during each stage.

1.2 Malignant Immunosuppression

The adaptive immune response begins with Dendritic cells, whose activation can be inhibited by factors in the tumor microenvironment (Michielsen et al., 2012; Chevalier et al., 2017). As discussed above, inactive Dendritic cells that present tumor antigens can actively suppress immune responses. In addition to being incapable of activating effector T cells, They recruit T Regulatory cells, which suppress T cell mediated cell killing (Ohta et al., 2006; Curiel et al., 2004). However, because adjuvants can usually be found that will activate Dendritic cells, more research has focused on the later stages of the immune response: T cell activation and T cell killing.

For normal T cell activation to proceed, the Dendritic cell must signal that it is active, which is communicated by cell-surface presentation of B-7 family ligands. These B-7 ligands complex with CD80 and CD86, which when activated alongside T cell receptor recognition of the antigen presented by the Dendritic cell, signals that the antigen recognized by the T-cell should be the target of an immune response. CTLA-4 is a competitive inhibitor for B-7 ligands, and when present will out-compete CD80 and CD86 (Qureshi et al., 2011). This competitive inhibition suppresses T cell activation (Wing et al., 2008) and may bias T cells towards immunosuppressive regulatory cell states. Conversely, the elimination of the CTLA-4 receptor in mice causes a lethal lympho-proliferative disorder, suggesting that it unleashes unrestrained clonal T cell expansion (Tivol et al., 1995). Thus, CTLA-4 functions as an immune checkpoint.

Since the discovery of CTLA-4, myriad additional immune checkpoints have been uncovered including but probably not limited to TIM-3, BTLA, VISTA, and LAG-3. Each of these proteins has been observed on cancer cells, and when present, produces complex patterns of T cell suppression. As a result, each protein is being investigated as a possible drug target (Sharma et al., 2017).

In addition to the more general CTLA-4 pathway, numerous other mechanisms exist that have been observed to suppress T cell activation in certain contexts. Release of adenosine by tumor cells, triggered by hypoxic conditions, suppresses activation and recruits T Regulatory cells (Ohta et al., 2006). VEGF-A and EDNRB/ETBR, molecules secreted by tumor cells to control tumor vasculature and oxygen availability, may reduce T cell homing and adhesion, excluding them from the tumor environment (Bouzin et al., 2007). Additionally, VEGF-A can induce IL-10 secretion, which suppresses IFN-γ, a critical T cell activation signal. IL-10 is one of many mechanisms by which myeloid derived suppressor cells, immune cells which are functionally characterized by their ability to suppress anti-tumor immune responses, exert their effects. Other mediators of suppression include the production of reactive oxygen species and the secretion of NO, arginase, prostaglandin E2, or TGF-β, several of which are clinically prognostic indicators of suppression (Muller and Scherle, 2006).

At later stages of an immune response, the most prominent inhibitor of T cell mediated cell-killing is PD-1, which is expressed on the cell surface of activated T cells. The activity of PD-1 functions similarly to CTLA-4: when a T-cell complexes with its target, it will induce the cell to die. However, if PD-1 becomes activated, it is interpreted by the T-cell as a signal that the target cell was mistakenly identified, and causes the T-cell to die instead (Keir et al., 2008). The ligands for PD-1, PD-L1 and PD-L2, are expressed on tumor and Dendritic cells, and this pathway is therefore thought to serve an immuno-regulatory function. Unlike CTLA-4 knockout mice, PD-1 -/- mice survive, but display chronic autoimmune phenotypes (Nishimura et al., 1999; Nishimura et al., 2001).

Tumor cells are capable of expressing the majority of the immune-inhibitory markers and secreting many of the immunosuppressive compounds discussed above. Because of the variety of mechanisms through which tumor cells are capable of inducing this effect, it is logical to expect a large variance in the druggability of each pathway across patients. This fact may partially explain the frequent treatment failures observed in immunotherapy trials. Even our limited knowledge of the functionality of these pathways reveals a system of tremendous complexity. It is designed to carefully balance immunity against foreign pathogens against the risk of self-recognition and auto-immunity. Because of this balance, cancer is often able to avoid detection by co-opting systems the body uses to prevent auto-immunity.

1.3 Immune Therapies for Cancer

Over the past decade, some of the most promising advances in oncology have come from vaccines that paint the cancer as an immune target or drugs that disrupt cancer's ability to masquerade as "self." The idea behind cancer vaccines is to stimulate the immune system to recognize a tumor as a pathogen and eliminate it. However, early attempts at vaccines quickly taught that simply exposing the body to cancer antigens was not enough—if Dendritic cells are exposed to an antigen in absence of an activation signal, they tend to suppress, instead of activate, the immune response (Rosenberg, Yang, and Restifo, 2004). In contrast, follow-up studies that added adjuvant treatments designed to stimulate Dendritic cells had more positive effects: co-administration of IL-2, a Dendritic cell activation signal, with gp100, a melanocyte differentiation antigen improved melanoma outcomes (Eberlein, 2012). Similar pairings have been effective in intraepithelial neoplasia, B-cell lymphomas, and non-small cell lung cancer. (Eberlein, 2012; Stockman, 2011), There have also been attempts to use viral vectors encoding tumour antigens to exploit the naturally strong antiviral immune response to elicit reactivity against cancer antigen, but these attempts have missed phase 3 trial targets (Bavarian-Nordic, 2017).

A second type of vaccine involves ex-vivo culturing of Dendritic cells with tumor antigens. This hyper-personalized treatment selects for Dendritic cells that show reactivity against the patient's tumor, which can then be injected back into the patient alongside an activation adjuvant to foment an immune response. Unfortunately, the complexity of cell isolation, ex-vivo manipulation and re-infusion has made this approach very costly. Consequently, it has been unpopular with pharmaceutical companies and has seen limited commercialization (Mellman, Coukos, and Dranoff, 2011). However, unlike the culturing of Dendritic cells, engineered T cells have been successfully commercialized to treat acute lymphoblastic leukemia (Yescarta, Kite Pharma) and large B cell lymphoma (Kymriah, Novartis). These approaches identify immunogenic antigens in a patient's tumor and engineer chimeric T cell receptors to recognize and respond to them.

While cancer vaccines have shown some success, there are still several challenges that have prevented it from evolving into a broadly effective treatment strategy. First, there is confusion about which tumor antigens are adequately immunogenic to activate Dendritic cells, and which are adequately expressed across cells in the tumor to enable pan-cancer targeting. Second, safe adjuvant therapies for Dendritic cell activation are not yet apparent; most induce moderate to strong autoimmune side effects. Finally and most critically, though the conditions for immunization eventually may be optimized, effectiveness can be dampened by immunosuppression mechanisms acting at any of the stages identified above: Dendritic cell priming and activation, T cell activation, and T cell mediated cell killing. As a result of the linear nature of this activation process, therapies that enter earlier in the chain, such as activated Dendritic cells must pass additional immune checkpoints relative to those that enter towards the end, such as activated T cells in CAR-T therapies, and as such often have lower success rates. However, it is not yet clear how difficult each immune checkpoint is to pass, how commonly cancer co-opts each checkpoint to block immune activation, and the what effect the interaction of multiple mechanisms has at each checkpoint.

In addition to cancer vaccines, there have been attempts to deplete immune-regulatory cells such as T Regs from the tumor environment. Hampering this approach, no specific surface marker of T Reg cells has yet been identified that is not also expressed on effector T cells, although some proteins such as GITR and OX40 may be transiently expressed (Ito et al., 2006; Cohen et al., 2010). As a result, depletion methods tend to remove both T Regulatory and CD8+

effector T cells. The best attempt so far may be the use of anti-CD25 antibodies, which preferentially deplete T Regulatory cells, at least following short-term therapy, and may help increase the efficacy of active immunization (Golovina and Vonderheide, 2010).

These problems have driven researchers to investigate the use of drugs for targeting specific molecular mechanisms underlying immune suppression and activation. Like with vaccines, there have been efforts to characterize immune checkpoints at each stage of the adaptive immune response. Motivated by mouse research showing that CTLA-4 knockouts induced lethal auto-immune reactions, it was theorized that a weaker, temporary blockade of CTLA-4 signaling may be effective in unleashing T cell responses in patients for which the main barrier is T cell activation. Thus, a drug was developed targeting the CTLA4 pathway, aiming to increase the effectiveness of Dendritic-cell based activation of CD8+ T cells. This was attempted first in mice and then in a series of trials for Melanoma (approved), prostate, lung, and bladder cancer. These trials have succeeded not because of a high response rate—relatively few patients respond—but rather because those patients that do respond appear to obtain durable and long-lasting recoveries (Robert et al., 2011; Hodi et al., 2010). However, because of significant on- and off-target inflammatory toxicities, anti CTLA-4 drugs are primarily targeted to late-stage patients who do not respond to front line therapies.

Similar to vaccines, drugs targeting later-acting immune checkpoints appear to have superior specificity. Mirroring the result in rodents, anti-PD-1 drugs seem safer than Ipilimumab (Brahmer et al., 2010) and consistently show durable responses in subsets of patients (Hamid et al., 2013; McDermott et al., 2014). These clinical results suggest that anti-tumor immunity is functional up to but not including T cell mediated cell killing in some patients.

In addition to these two pathways, there are numerous other co-stimulatory or co-inhibitory pathways that are believed to be involved in modulation of the anti-tumor responses. Stimulatory pathways that might be activated include CD28, OX40, GITR, CD137, CD27, and HVEM, and inhibitory pathways that could be blocked include TIM-3, BTLA, VISTA, and LAG-3. It is hoped that part of the reason for the limited patient response to PD-1 and CTLA-4 drugs is that aberrations in these other pathways are responsible for immunosuppression in those patients. By better characterizing and personalizing treatment, it is hoped that the response rate of patients to immune-based treatments may be improved. Initial investigation into this with combination therapies has been promising: Ipilimumab (CTLA-4) + Nivolumab (PD-1) appear to enhance immune activity over either therapy alone (Wolchok et al., 2013).

The results of this research and the clinical trials that followed have revealed an extremely complex web of overlapping mechanisms governing immune-cancer interactions. Each identified immunosuppression mechanism has been shown to have a critical role in preventing autoimmune disease, outlining the need for rational therapy designs, and where possible, targeted delivery. However, later checkpoints unleash increasingly specific responses. If it were possible to identify the particular pathway that is primarily responsible for inhibition of an immune mechanism, this would minimize the off-target autoimmune effects. Similarly, improvements allowing therapies to be delivered directly to the tumor, at least for T cell homing and T cell mediated cell killing, should in theory also serve to minimize off-target effects. Taken in combination, these improvements may allow greater doses of drugs to be brought to bear, in cases where patient toxicity would otherwise prevent a dose that facilitates complete penetrance. However, this research does not benefit individuals whose immune system is suppressed by checkpoint pathways outside the regime of existing drugs, or patients who are given a drug that opens the wrong checkpoint. Thus, there is a significant and pressing clinical need to better characterize the immune phenotypes present in cancer patients at the level of individuals. Only recently have tools capable of probing these complex questions been invented. This dissertation deals, in part, with the methodological and computational development of one such technology and its application to characterize complex immune phenotypes.

1.4 Single-Cell Technologies

Invention of new genomics technologies have triggered rapid change in biological research by allowing researchers to ask new categories of questions. For example, sequencing the human genome provided biologists with the complete blueprint of human cells. This provided a context in which to place the previously haphazard identification of individual expressed genes. The combination of that knowledge with targeted approaches to identify RNA molecules—expressed sequence tagging—has taught us how the genome is functionally expressed. This information was combined to start the transcriptomics revolution with the creation of DNA microarrays, which for the first time provided a relatively unbiased functional readout of the genes expressed by cells isolated from a tissue.

Microarrays increased the amount of data generated in a single assay by 1000-fold and enabled scientists to make predictions about responses to treatment in breast and prostate cancer that previously required much more laborious investigation (Glinsky et al., 2004; Van't Veer et al., 2002). They were also used to predict response to early immune therapies in melanoma (Monsurrò et al., 2004) and identify pan-cancer signatures of immune infiltration (Chifman et al., 2016).

RNA-sequencing (RNA-seq) expanded on these capabilities by eliminating bias inherent in microarray technology, reducing the required amount of RNA input, and for the first time allowed an absolute measurement of the number of each RNA in a sample isolate. RNA-seq was effectively applied to generate, among other projects, The Cancer Genome Atlas (McLendon et al., 2008; Network, 2011). This Atlas has been extensively mined and has contributed prognostic signatures of immune therapy success (Şenbabaoğlu et al., 2016). RNA-seq has since been utilized to interrogate the binding patterns of transcription factors (CHIP-seq, Johnson et al., 2007), determine nucleosome occupancy (DNAse-seq, Boyle et al., 2008), refine transcription factor binding preferences (SELEX-seq, Riley et al., 2014), map the 3D spatial organization of DNA in the nucleus (Hi-C, Belton et al., 2012), and create many more experimental paradigms. RNA-seq has spawned large studies on the functionality of the genome, such as the ENCODE (Consortium, 2012) and ROADMAP epigenomics projects (Kundaje et al., 2015), which revealed that a large portion of the genome is responsible for functional differences in gene expression, even if it is not itself transcribed.

However, the utility of RNA-sequencing is limited because it requires a large number of input cells to achieve the concentration of DNA necessary to run the sequencer. Because millions of cells were needed to provide the required microgram of DNA input (Wilhelm and Landry, 2009), RNA-sequencing cannot tell the difference between 50% of the cells expressing two copies of an RNA and 100% of cells expressing one copy. Yet, this distinction is critical to understanding population level immune variance, and also cancer antigen variation.

Single cell approaches, in contrast, can answer these types of questions. Flow cytometry is a two step procedure wherein cells are first exposed to antibodies that are bound to fluorophores, and then shot through a fluorescence detector at high speed. This approach is able to characterize 8-17 proteins, limited by the overlap of fluorophore emission and excitation spectra. (Perfetto, Chattopadhyay, and Roederer, 2004).

Advances in cytometry by time-of-flight (Cytof) replaces fluorophores with metal ions, increasing the number of measurable proteins in a single cell to 35-50. This is still typically limited to surface proteins, and has a large experimental lead time as high-quality, specific antibodies must be developed for each protein target. Despite these drawbacks, Cytof's antibody based measurements produce relative but reliable continuous estimates of protein abundance and measure individual cells. In addition, it is one of the highest throughput technologies available for assaying single cells, capable of easily measuring millions of cells per sample, although these numbers are rarely necessary. Finally, the individual cell measurements made by Cytof are typically quite reliable; antibodies have good binding affinities and as a result, there is high confidence that if cell surface markers are present, Cytof will detect them.

As a result, Cytof has been very effectively applied to better characterize immune cell states in cases where the marker combinations are already known: Cytof identified previously unknown signaling mechanisms in the otherwise well-understood hematopoiesis system (Bendall et al., 2011), helped decompose CD8+ T cells states and display combinatorial cytokine producing subtypes (Newell et al., 2012), enabled by trajectory-finding approaches, found novel early human B-cell populations (Bendall et al., 2014), stratified Macrophage and T cell phenotypes in renal cell carcinoma (Chevrier et al., 2017), and characterized t- and myeloid-cell dysfunction in lung cancer (Lavin et al., 2017).

Unfortunately, Cytof (and FACS) requires prior knowledge or guesswork to identify cell surface proteins that define cell types before experimention begins. In addition, it has a limited ability to resolve intracellular states because cells must be permeabilized to allow antibody entrance, and therefore targeting intracellular proteins does not always produce effective staining. As a result, it is a powerful tool for teasing apart cell populations based on previously identified surface markers, but is not a capable tool for characterization of unknown phenotypes.

1.5 Single-Cell RNA-Sequencing (scRNA-seq)

scRNA-seq is the first opportunity to make an unbiased measure of more than a few proteins or transcripts, allowing populations of seemingly homogeneous cells to be deconvolved into their component parts. As early as 2009, transcriptomes of the largest single cells—oocytes—and multi-cell blastomeres had been amplified and measured singly in tubes by adapting a protocol originally designed for microarrays (Tang et al., 2009). Multiplexing of this method for multiple cells was accomplished shortly thereafter with a cell barcoding strategy called CEL-seq wherein the poly-A capture primers were modified to contain short designed nucleotide sequences ("cell barcodes") that differed between cells and uniquely marked them (Islam et al., 2011).

Developed contemporaneously and released shortly following CEL-seq, SMART template-switching chemistry was introduced to allow isolation of full-length transcripts (Ramsköld et al., 2012). It accomplished this by leveraging the Nextera transposase reagent from bulk sequencing assays, which incorporates itself randomly into the captured RNA fragments, carrying an Illumina index. Each cell is incubated with Nextera using different index pairs, the combination of which uniquely tag each cell. Therefore, when the indices were read off the sequencer, it allowed reads from up to 96 cells to be sorted according to the cell that generated them per sequencing lane. This technological innovation was important because it allowed splicing events to be observed in single cells for the first time. scRNA-seq, applied in this fashion, can measure any cell of any size that can pass through a FACS sorter

CEL- and SMART-seq enabled reading of an estimated 7-10 thousand molecules of RNA in each cell. However, these data are not without problems. While 10,000 molecules is many more than proteomics approaches allow, it is a relatively small fraction of the mRNA available in a eukaryotic cells. Because many of the features of interest, such as intracellular transcription factors, are present at low copy number, capture-based stochasticity may drive their expression as much or more than biological state.

Second, because cells contain so little starting material, and sequencing devices are designed for bulk genomes, large amounts of PCR (15+ cycles) or linear amplification (overnight) is necessary to generate enough material to run a sequencer. Thus, both CEL- and SMART-seq utilize amplification, and this interacts with the capture rate to produce odd mixture distributions over captured RNA molecules.

The Central Limit Theorem describes how the mean values of a series of samples extracted independently from a population tend towards a Normal or Gaussian distribution. Because of the tendency for Gaussian distributions to naturally arise, there has been abundant work to analyze data that follow these distributions, and as such it is a desirable property. In bulk sequencing, this is approximately achieved, as each technical or biological replicate samples many cells, each of which contains many mRNA. After correcting for amplification by taking the log of the observed counts, the expression of most genes across replicates can usually be fit to Gaussian distributions.

Single cell data is not so well behaved. The shallow sampling of the transcriptomes causes some cells to miss particular genes, which ”drop out” of analysis. Thus, when the gene’s expression is examined across cells, data is shared between a zero category or “drop-out” and a Log-Normal component which represents amplification over the captured molecules. Because capture is so sparse, biases in amplification and other steps in library construction can contribute as significantly to the ratio of cells in the ”drop-out” and ”continuous” components as do the numbers of molecules originally present in cells (Zheng, Chung, and Zhao, 2011; Dohm et al., 2008). Indeed, capture in these technologies can be so variable that popular analysis visualizations such as ”Dot plots” treat the fraction of cells that detect a transcript as carefully as they do the magnitude of expression in a detected gene (Shekhar et al., 2016). As a result, only large effects in molecule number are reliably detectable.

To overcome these technical biases, additional experimental controls were necessary to make the data more interpretable. The most significant advance was the inclusion of Unique Molecular Barcodes (UMIs). Like cell barcodes, these are added to capture primers. Unlike cell barcodes, molecular barcodes contain random sequences. When the barcodes are long enough, they probabilistically provide a unique marker for each molecule. This allows a computational scientist to resolve the reads obtained in an experiment at molecular resolution. Inclusion of molecular barcodes allows computational scientists to exchange the complex mixture distributions for well characterized Poisson statistics—the statistics that describe the rate of rare sampling events (Shiroguchi et al., 2012). The effectiveness of this molecular barcoding is demonstrated with exogenous spike-in control reagents. RNA of known concentration are added, and the accuracy of population estimates were improved with inclusion of UMIs (Grün, Kester, and Van Oudenaarden, 2014).

Amplification methods have an interaction with cell and molecular barcodes. Traditional polymerase chain-reaction (PCR) amplification has a relatively small error rate, but the output of PCR is also valid input. Each round of PCR adds new substrate to the pool to be amplified. Therefore, any errors introduced into the cell barcode are propagated into the reaction, producing error trees wherein branches inherit errors from their trunks. In contrast, linear amplification through in-vitro transcription (Eberwine et al., 1992), as used in CEL-seq, takes cDNA as input and generates RNA output. Thus, unlike PCR, the product of the reaction cannot act as substrate, and any errors that occur do not propagate or compound. In addition, CEL-seq2/C1 and MARS-seq have a steeper slope at low sequencing depths than both Drop-seq and SMART-seq, potentially due to a less biased amplification by in vitro transcription (Ziegenhain et al., 2017).

These technological advances—capture primers, amplification improvements, cell and molecular barcoding—form the basis of modern single-cell sequencing. The small cell counts of initial experiments were adequate to ask very specific questions of well controlled systems. Often, the cell type had already been isolated by FACS and studies were limited to determining if the isolated cell population displayed a single phenotype with variation, or if there were modes hidden in the population that represented distinct states (Shalek et al., 2013). To scale beyond these experiments, robotics was used to optimize plate loading for both SMART-seq (Shalek et al., 2014) and CEL-seq (Jaitin et al., 2014) chemistries, allowing 96 cells to be processed at a time with reduced hands-on time. More recently, droplet-based microfluidics approaches to CEL-seq (Klein et al., 2015) and SMART-seq (Macosko et al., 2015) were developed which enabled thousands of cells to be generated at once[1].

These scRNA-seq approaches have revealed several interesting characteristics of the techniques used in sequencing library preparation. First, the capture primers used to extract the mRNA from the cell have quite low efficiency—early methods had as low as 5% capture rate, and cutting edge approaches have 30-35% capture rates. As a result, it can be difficult to measure low-expression transcripts such as transcription factors, which are often present at very low copy number in cells. Second, because some gene sequences are better substrates for PCR amplification than others, the (much larger) amplification necessary to create libraries for scRNA-seq with

[1] Near the end of this dissertation, 10x Genomics (Zheng et al., 2017b) provided a commercial application that pulls from both approaches and by focusing on tight control of the bead construction process. It is thought to currently produce the highest quality data, albeit at significantly increased cost relative to non-commercial droplet technologies.

adequate concentration to load the sequencer produces significant variation in gene abundances. Together, these combine to produce very sparse libraries – bulk RNA-seq could be expected to identify over 20,000 genes in a sample. In contrast, scRNA-seq captures between 1000 and 5000, depending on the size of the cells, abundance of their RNA, and lack of cellular stress. In spite of these disadvantages, the capability of sequencing individual cells has taught us much about the development and function of immune cells.

1.6 Single-Cell Approaches to Characterize Immune Populations

Studies predating scRNA-seq had long identified that cellular variation in expression of immune marker genes can have a functional influence on clinical outcomes. For example, a 2010 study using FACS discovered variable expression of the IL-2R in a population of T cells during an immune response correlates with T Effector and T Regulatory cell survival (Feinerman et al., 2010).

scRNA-seq has been applied to great effect to improve our understanding of how immune cells develop, and the triggers that are necessary to induce the differentiation of precursors into mature cell types. Jaitin et al., 2014 classified splenic cells into known immune cell types (B, Natural Killer, Macrophages, Monocytes, and Plasmacytoid Dendritic cells), and characterized variation in Dendritic cell responses to lipopolysaccharide, a bacterial protein, identifying universal IFN response genes, and transcriptionally separating them into sub-populations. Björklund et al., 2016 profiled CD127+ innate lymphoid cells in tonsil and small intestine. They identified previously characterized Innate Lymphoid cell 1, 2, 3, and NK cells based on surface marker expression, but also identified novel transcriptional signatures suggestive of possible subpopulations within the identified groups. Gury-BenAri et al., 2016 examined helper-like Innate Lymphoid cells in the small-intestine, finding new populations defined by expression of NKp46, retinoic acid receptor related orphan receptor-γ-t, and IFNγ or IL-2 and CCL22. Schlitzer et al., 2015 examined lineage commitment in conventional Dendritic cells, while Drissen et al., 2016 improved the resolution of early myeloid lineage branching. Lönnberg et al., 2017 and Paul et al., 2015 helped map out how surface marker based cell states map to transcriptional changes in early bone marrow differentiation, and Nestorowa et al., 2016 carried out a similar study on hematopoietic stem cells. Finally, Villani et al., 2017 demonstrates the power of deeply sequencing a relatively homogeneous population by examining 2400 HLA-DR+ cells, revealing 6 Dendritic cell sub-populations.

Similarly, scRNA-seq has been effective at teasing apart how immune cell activation functions in well controlled model systems. The first major application of SMART-seq examined bone-marrow derived Dendritic cells subjected to a bacterial lipopolysaccharide activation signal to study variation in gene expression and splicing patterns among Dendritic cells in response to infection (Shalek et al., 2013). This system was selected in part because it presents an interesting biological question, but also because Dendritic cell activation is known to induce temporal phasing, and because activated Dendritic cells are post-mitotic, therefore the majority of variation was expected to be biological. The study revealed a correlated component of gene expression including Stat2 and Irf7, that drive the antiviral response. Lönnberg et al., 2017 examined variation in t-helper differentiation states in response to malaria. A series of knockout experiments targeting the IFN-R pathway (Gaublomme et al., 2015) were able to identify heterogeneity of TH-17 cells in CNS and lymph node at peak of autoimmune encephalomyelitis. Finally, recent studies demonstrated how tumor-associated cells can recruit Macrophages that suppress immunity by sequencing individual Macrophage cells and identifying hippo pathway Yes-associate protein

(YAP) as the critical checkpoint for Macrophage recruitment (Guo et al., 2017)

These studies demonstrate how profiling single cells enables functional characterization of immune populations, both in natural states and when stimulated by foreign pathogens. Yet, at the time this research began, the study of immune infiltration of cancer had not been broached by large scRNA-seq experiments. The following chapter will describe how a technology was selected from the set of approaches reviewed above, adapted to sequencing immune cells, and benchmarked for sources of variation It will then introduce a statistical toolkit that was developed to correct the technical problems introduced by the process of transforming individual cells into a library which, after sequencing, describes the phenotypes of thousands of immune cells per tumor.

1.7 Towards an Atlas of Tumor Immune Phenotypes

Recent scRNA-seq studies highlight several critical questions in cancer immuno-biology: namely, how do individual immune cells react to cancer, and what can we learn about population-level differences between patients? How do the expression of individual marker genes, or pairs of marker genes, that are used to defined cell states correlate with cellular phenotypes? What do these transcriptional states imply about individual patient's amenability to drug treatments?

To answer these questions, we would need to sequence a very large number of cells. We reasoned that cancer exists as a natural perturbation, and that the unique nature of each patient's tumors should provoke significant variation between patients. However, recent demonstrations have shown that tissue residence alone exerts a significant effect on immune phenotypes. Therefore, we rationalized that it would be critical to first characterize the cellular states within the healthy tissue, thus we will hierarchically characterize blood, tissue, and tumor-infiltrating immune cells. Previous studies and retrospectively, ones that were carried out in parallel, have demonstrated that small numbers of cells from large numbers of patients are not capable of capturing meaningful heterogeneity across immune cell types; they simply recover common cell states (Chung et al., 2017). Thus, shallow sampling of TILs cannot determine which cells are present across tumors at variable levels, versus the states that are patient-specific.

Therefore a successful study would deeply sample multiple patients, but also multiple tissues within each patient. This introduces a second problem: early studies focused on technical demonstrations, often within model organisms or cell lines, and eschewed statistical approaches that are necessary to examine variances across multiple samples of variable sources. Very few studies had dealt with human patient effects in immune data–most studies used mouse models with intentionally limited genetic variation. As a result, limited work had been done to address statistical problems like normalization and adjustment for patient-patient differences, and the ability to ask large-scale questions in humans was therefore limited. Yet, the ability to generalize across patients is crucial to interpret the relative importance of any states that are discovered.

Thus, before any attempts could be made to describe immune phenotypes, several technological and computational challenges needed to be addressed. The approaches for solving these challenges will be detailed in the next two chapters. Chapter 2 will discuss the selection of the appropriate technology to assay immune cells at a scale that was at the time unprecedented, and the computational tools that were developed to ensure that the resulting data highlighted biological variation and minimized technical effects introduced by the construction of the sequencing library. Chapter 3 will then discuss the statistical models and tools that were developed to reason about the relationships between cells from different patients whose different tumors introduced huge biological variation that was not always cleanly separable from technical variation. Finally, Chapter 4 will describe an experiment made possible by these tools, wherein a broad range of immune cell states are uncovered and characterized, highlighting how breast tumors dramatically expand the range of observed immune cell phenotypes relative to those observed in healthy tissue.

Chapter 2

Constructing a Flexible Framwork to Maximize scRNA-seq Data Quality

2.1 Droplet-based Sequencing Enables Deep Profiling of Immune Cells

Expanding our knowledge of the functionality of tumor infiltrating immune cells requires examination of multiple tissues. Because peripheral blood is the most accessible source of immune data, most knowledge of human immune phenotypes comes from experiments done on circulating immune cells from blood. Therefore, without measuring circulating blood, new findings in tumor immune cells could be difficult to compare with prior experiments. In addition, it is established that when immune cells transition from circulation into tissues, they are subjected to different stimuli that cause shifts in gene expression (Fan and Rudensky, 2016). As a result, without also measuring immune cells in healthy tissue, it would not be possible to distinguish between tumor infiltrating immune phenotypes that result from cancer from those caused by tissue residence. Therefore, we reasoned that to effectively study tumor infiltrating immune cells we would also need to characterize immune cells in healthy tissue and blood.

These requirements, combined with the complexities of single-cell sequencing, lead to complex experimental designs that require many cells. For example, It is important to separate phenotypes that are commonly observed across patients from those that result from specific microenvironments of individual tumors. Thus, multiple patients must be sequenced to determine the relative frequency and generalizability of observed phenotypic states. Additionally, scRNA-seq is a new technology with significant uncharacterized technical variability. To verify that observed cell states result from biological differences and not technical changes induced by the experimental protocol, it is critical to measure each sample multiple times to identify what fraction of observed variability is technical. Each of these variances increase the number of observable cell states, and consequently, we expected to need approximately 1000 cells per replicate to measure each sample.

To sequence, in triplicate, immune cells from blood, healthy tissue, and tumor, would require 9,000 cells. Using the Smart-seq 2 on the Fluidigm C1 and devoting 1 million sequencing reads per cell would cost approximately $12 per cell, for a total cost of $108,000. Such an experiment would nearly double the largest Smart-seq 2 experiment to current date for just a single patient (Zheng et al., 2017a). Fortunately, InDrop and Drop-seq, microfluidics approaches capable of sequencing tens of thousands of cells at low cost were announced shortly after this experiment was conceived (Klein et al., 2015; Macosko et al., 2015).

By exchanging plates for microfluidic encapsulation flow cells, both platforms were capable of preparing tens of thousands of cells in hours, a feat that would have required either expensive mechanization or nearly two week's work for a trained technician using existing plate approaches. This advancement enables the scale of sequencing necessary to deeply profile large numbers of immune cells from the multiple tissues of multiple patients. The next section will discuss how droplet-based sequencing makes this possible.

2.2 Technical Characteristics of Droplet-based Sequencing

Droplet-based sequencing allows thousands of cells to be sequenced using the same number of sequencing reads previously used to characterize 96 cells using plates, or 3-5 bulk samples. This is accomplished adding barcodes to each sequencing read which allows them to be matched back to the cell and molecule they originated from (Klein et al., 2015; Macosko et al., 2015). By barcoding molecules, only a single read is necessary to identify a captured mRNA. This allows the output of a sequencing experiment to be transformed from read abundances into a ”count” matrix populated by molecules, and confers an added benefit of collapsing PCR outliers that amplify well into single observations, yielding more accurate counts (Grun and Oudenaarden, 2016). As as result, fewer reads must be spent to characterize each molecule.

Second, the addition of cell barcodes allowed the number of cells included in each reaction to be increased from 1 per well to many thousands. This had several practical benefits. First, the RNA in each cell serves as the substrate for the initial amplification round, and with 50x more cells than plate-based approaches, the enzymes are exposed to more substrate in a smaller volume, which makes them more efficient Second, the increase in initial substrate allows the number of PCR cycles to be decreased by 5-6 without decreasing the output cDNA concentration. This reduced cellular ”jackpotting”, where one read or gene accrues additional copies relative to its cellular abundance due to favorable amplification. Reducing jackpotting therefore leads to more uniform sequencing coverage across cells, and increases their average molecule count. Together, these advancements reduced the per-cell cost to approximately $0.4.

However, these improvements did not fully offset the decrease in sequencing depth. Instead of spending 250,000 to 4 million or more reads per cell (Shalek et al., 2013; Shalek et al., 2014; Jaitin et al., 2014), droplet technologies measure phenotypes using $\frac{1}{40}th$ of the sequencing reads, as few as 20-50,000 per cell (Klein et al., 2015; Macosko et al., 2015). This meant sparser cell data; instead of 5-8,000 genes per cell, InDrop and Drop-seq observe 1-3,000 genes, implying that more than half of the low expression genes that were captured by already-sparse SMART-seq 2 technologies are missed by droplet approaches. The dramatic decrease in information that is attributed to each cell and the critical role played by DNA barcodes both lead to technical challenges that must be addressed to maximize the utility of droplet-based sequencing.

First, the addition of molecular barcodes revealed that the molecular capture rate of these technologies is actually starkly lower than expected, in the 5-20% range (Shah et al., 2016). Combined with the reduced numbers of reads associated with each cell, this leads to a phenomenon called ”drop out” wherein the read-out for a cell may fail to capture any of the molecules of a given gene, causing it to incorrectly masquerade as unexpressed. This produces a significant problem, as many canonical genes used to mark cell types code for stable proteins with low resting transcription. Thus, many of these important genes have low capture rates, and may drop out, which eliminates the historically most effective mechanism of identifying cell types. This phenomena is also computationally damaging, as the random drop-out effect causes cell-cell distances to improperly increase, making it more difficult to group cells into classes based on similarity.

Second, while barcodes enable increased multiplexing, they are susceptible to errors, which add substantial noise to the sequencing experiments. Compared to bulk sequencing, scRNA-seq requires a much greater number of enzymatic reactions to create a sequencing library. Additional amplification requires more polymerase reactions, while adding barcodes and sequencing adapters requires additional ligation. Each of these reactions introduce error into the sequences, which occur in the cell and molecule barcodes at a rate of approximately 1%[1]. Errors in these barcodes make cellular data appear to do perplexing things: errors in molecule barcodes make cells appear to express two molecules instead of one, and errors in cell barcodes make cells appear to express molecules that they do not. This last characteristic is particularly confounding: while drop-out makes us question the meaning of zeros, cell barcode errors make the values we do observe less reliable.

The third and most critical error source in scRNA-seq experiments is ambient contamination. Cells dislike being dissociated and sorted, two protocols that are often necessary to transform tissues into the single cell solution necessary for droplet sequencing. As a result of these stressful protocols, some cells will respond by lysing, dumping their mRNA loads into solution. These mRNA then find their way into emulsion droplets, sometimes with other cells, and sometimes in droplets containing only a barcoding bead. The result of this effect is that despite loading no more than 5-6000 cells, it is extremely unusual to observe fewer than 200,000 cell barcodes with associated sequencing reads.

Compounding this problem are amplification biases that can cause barcodes that were paired with real cells to amplify badly, or empty droplets paired only with contamination to amplify well. This blurs the line between the two types of data, making classification of cells a difficult problem, and decreasing the yield of sequencing experiments by allowing higher amplification of contamination. These technical effects, in combination, make the design of data processing methods a critical part of scRNA-seq analysis.

[1] details discussed in 2.12

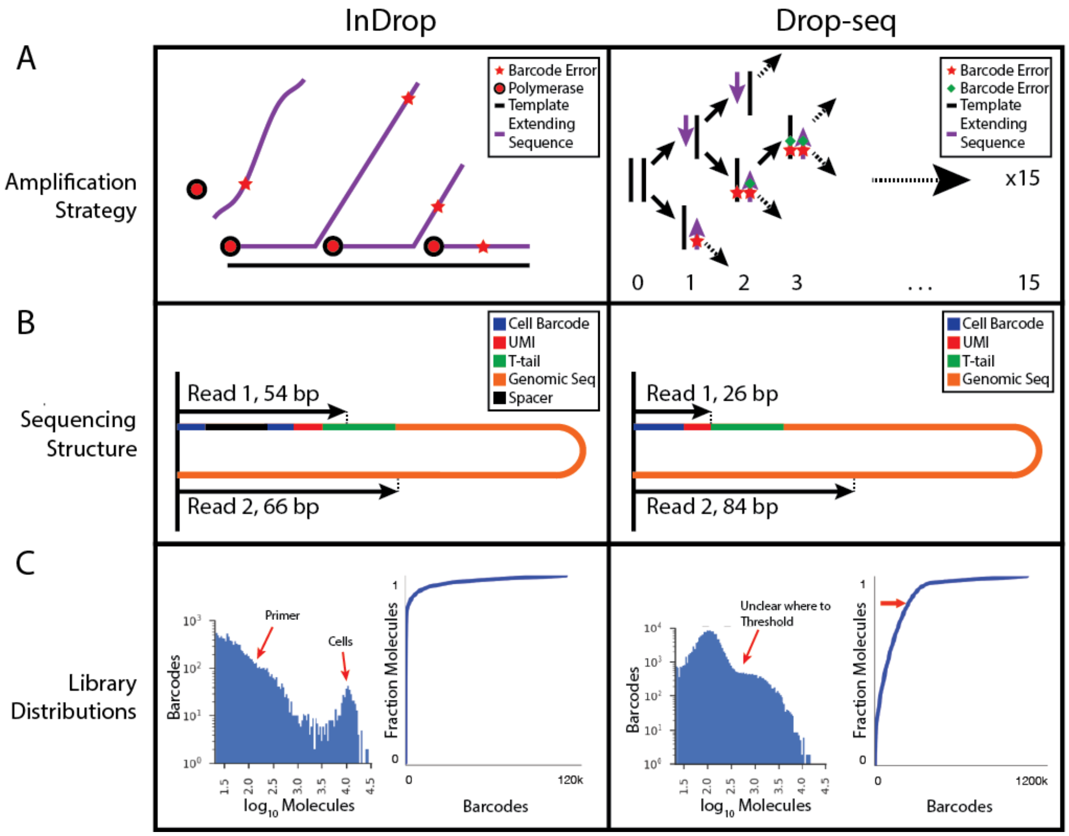


Figure 2.1: Demonstrative of major differences between InDrop and Drop-seq. (A) InDrop uses linear amplification while Drop-seq uses PCR. Linear amplification introduces more errors, but rarely has more than one error per barcode. Drop-seq can introduce chains of errors through PCR. (B) Summary of InDrop and Drop-seq primer and sequencing structure. InDrop uses a 54bp forward read containing two 8-11bp and 8bp cell barcode fragments, an 8bp UMI, and 5 bases of the poly-T capture primer. Drop-seq has a 26 bp forward read containing a 16bp cell barcode and a 10bp UMI. (C) empirical cumulative density function over molecules in an experiment. Each step upwards increments by the number of molecules in a cell (largest first) and each step right increments by a cell. Intuitively, faster movement upwards indicates larger concentration of molecules within individual cells, while movement right indicates relatively few molecules spread over very many cells.

2.3 Selecting a Droplet Sequencing Assay

When we began designing single-cell sequencing experiments, there were no computational analysis methods applicable to either InDrop or Drop-seq. In addition, while it was obvious that plate-based sequencing would be economically infeasible, choosing between droplet-based approaches was more difficult. While both technologies leveraged droplet-based encapsulation approaches, an examination of the technologies' chemistries reveal extensive differences.

Drop-seq used the SMART-seq approach leveraged in the Fluidigm C1 and plate-seq technologies, which typically captures more genes than than the CEL-seq approach used by InDrop (Ziegenhain et al., 2017). However, Drop-seq also had a higher relative difference between its cell and bead flow rates, causing it to capture only about 1% of cells, compared to InDrop's 25% capture rate. Thus, while drop-seq might enjoy a better capture rate, it had the disadvantage of requiring a much larger number of input cells, one that we thought could be difficult to obtain from tumor samples with variable immune infiltrate.

Another critical difference was that while InDrop's beads have designed cell barcodes with known sequences, Drop-seq's cell barcodes are randomly generated using synthetic combinatorial chemistry. These random barcodes have over 100x the number of possibilities, which allows Drop-seq to enjoy a lower theoretical doublet rate, but aren't designed with error correcting codes. Therefore, drop-seq has no guarantee that errors in their barcodes will be detectable.

This problem is exacerbated by the use of PCR, which propagates errors from early sequencing rounds, since the product of PCR also serves as substrate (Figure 2.1 A). InDrop, in contrast, uses a linear amplification approach based on in-vitro transcription (IVT). IVT has a higher error rate than PCR, but errors don't propagate, this results in the majority of reads containing at most 1 error, a state that is easy to correct through the use of error correcting codes.

Unfortunately, there were no controlled experiments which would allow these two technologies to be quantitatively benchmarked. In addition, the technologies were demonstrated on very different biological systems: Drop-seq was run on a human-mouse cell line mixture and retinal neurons, whereas InDrop was demonstrated on cultured induced-pluripotent stem-cells. Because these cells have different sizes and stress responses, it made direct comparison of their results impossible.

Therefore, to differentiate between Drop-seq and InDrop, we carried out an in-house comparison of using an acute myeloid leukemia cancer model, reasoning that this system would be closer to the eventual tumor infiltrating immune cells that we would measure in our experiments than the published technologies (Figure 2.1 C). We then examined the resulting data for the experimental yield of molecules and cells, the extent of cell contamination, and the feasibility of differentiating cells from non-cellular contamination.

In our hands, InDrop produced data wherein the larger-count cell barcodes were more clearly separable than Drop-seq, and the overall yield of the InDrop experiment was higher. In addition, the fraction of data concentrated in large-count cell barcodes was much higher in InDrop (Figure 2.1 C, second panels). This is an important observation, as it suggests that the overall ambient contamination was lower in the InDrop system, and therefore that the observed molecule counts in InDrop cells had a higher signal to noise ratio. Combined with InDrop's better internal technical controls like error correcting cell barcodes and it's ability to capture more cells from rare samples, we were steered to utilize the InDrop assay for our lab's single-cell sequencing experiments.

2.4 Improvements to InDrop Barcodes Increase Molecular Yield and Error Correction Capacity

When we began working with InDrop, it had only been applied to well-behaved cell lines, and displayed worse performance on our clinical samples. Therefore, before exploring computational solutions to the error patters described above, we wondered if there were experimental changes that would improve the baseline performance of the InDrop assay.

One reason we favored InDrop over Drop-seq was that it had a clearer separation of cells from non-cellular contamination, at least partially due to a lower ambient RNA contamination level. However, we reasoned that common asystematic biases might also blur the boundaries between cell-bead and contamination-bead distributions. If those biases could be removed, we might further separate signal from noise, and make cell detection more feasible.

We began by measuring the "GC content" of the capture primers. GC content is the percentage of nucleotides in a DNA or RNA polymer that are guanine or cytosine, and it is established that sequences with 50-55% GC content are the best substrates for enzymatic reactions like PCR (Mamedov et al., 2008). Therefore, we reasoned that imbalances in GC content could cause some contamination-barcode pairs to amplify well, and cell-barcode pairs to amplify poorly, blurring the boundary between them.

When we measured the GC constant across cell barcodes, we observed that it was highly variable, with extreme high and low values of 20 and 70%. When we compared the number of molecules associated with barcodes of different GC contents, we observed that cell barcodes with "balanced" GC content of 50% were paired with more molecules in pilot experiments (Figure 2.2 A), and in published InDrop data (Klein et al., 2015)[2]. To quantify this phenomena, we correlated the deviance d from balanced GC content (50%) for each barcode $d = 1 - |GC - 0.5|$, and calculated the correlation of this statistic with the number of detected molecules for each barcode. We observed a small but very significant effect of $r^2 = 0.23, p < 10^{-45}$ suggesting that barcodes with 50% GC content obtained higher molecule count than those with higher or lower GC fraction.

[2] This phenomena was also observed in Drop-seq, but because their barcodes are not designed, it cannot be addressed for that platform

Since cells and barcodes were randomly assigned, these results implied that cells receiving GC-balanced beads were optimally amplified, and the presence of variable GC content was introducing a sampling variance over our sequencing libraries. Because there are a fixed number of reads to assign to each sample, increasing the variance of the number of molecules detected in each cell increases the molecule count of the largest cells, but decreases the average molecule count (see Figure 2.2 C). Since the eventual statistical analyses expect cells to be identically distributed—or be transformed to be identically distributed—extra sampling of a small number of cells does not provide any experimental benefit. Therefore, balancing GC content across our barcodes would decrease variance across our libraries, reducing the size of high molecule-count cells, and improving data quality.

Next, we thought about ways to improve the error correction capability of InDrop. Compared to Drop-seq, InDrop libraries have a high probability of containing sequencing errors. On average, we observe that one in 50 cell barcodes contains a single-base substitution error[3].

Because InDrop sequences a very large number of cells, it needs an even larger number of cell barcodes. However, it must pack those barcodes into DNA sequences of limited length, each base of which must be one of A, C, G, and T. As a result, there is a trade-off between the number of barcodes of a given length and the number of substitution errors needed to convert one barcode into another, also called the barcode's Hamming distance.

If the cell barcodes are too similar, substitution errors that convert one barcode into another can result in a molecule being mistakenly associated with the wrong cell. This is a critical problem as miss-assignment of marker genes can disrupt type identification, since they are used by biologists to label the type or lineage of each cell. It also frustrates detection of cell doublets, the

[3] This is likely because the T7 polymerase used to amplify InDrop libraries does not have proofreading capability.

rare events where two cells are encapsulated with the same barcode, as gene miss-assignment can cause cells to masquerade as doublets by associating two markers of different lineages with the same barcode. We observed that the originally published barcode sequences had a minimum Hamming distance of 2, which is adequate to identify but not to correct single-base errors (see Figure 2.2 B). Because single base errors are common in In-drop, we reasoned that this introduced unnecessary data loss.

Finally, a careful examination of cell barcode error rates showed that the most common error was that the first base of the second cell barcode would be converted to A at high rate (10%) from any other nucleotide. We reasoned that this substitution was the result of an extension process that occurred during barcode construction.

To address these shortfalls, we redesigned a cell barcode set so that all barcodes had balanced GC content, with Hamming distance of ≥ 3 (mean $D_h = 13.3$), excluding the first base of the second cell barcode, which was made a constant A to eliminate the observed error profile. This was done by performing a constrained optimization over barcodes of the variable lengths required by InDrop, obtained from Edittag (Faircloth and Glenn, 2012). Comparing libraries from before and after the redesign of our barcoding beads, showed that scRNA-Seq libraries generated with new DNA barcoding hydrogel beads obtained 5.3% improved yield as measured by molecules/million sequencing reads.

2.5 Improvements to the InDrop Assay Reduce Non-cellular RNA Contamination

A second problem that was observed during pilot experiments on patient samples is that long encapsulation time can allow cells to execute cell-stress or cell-death programs. Because sample

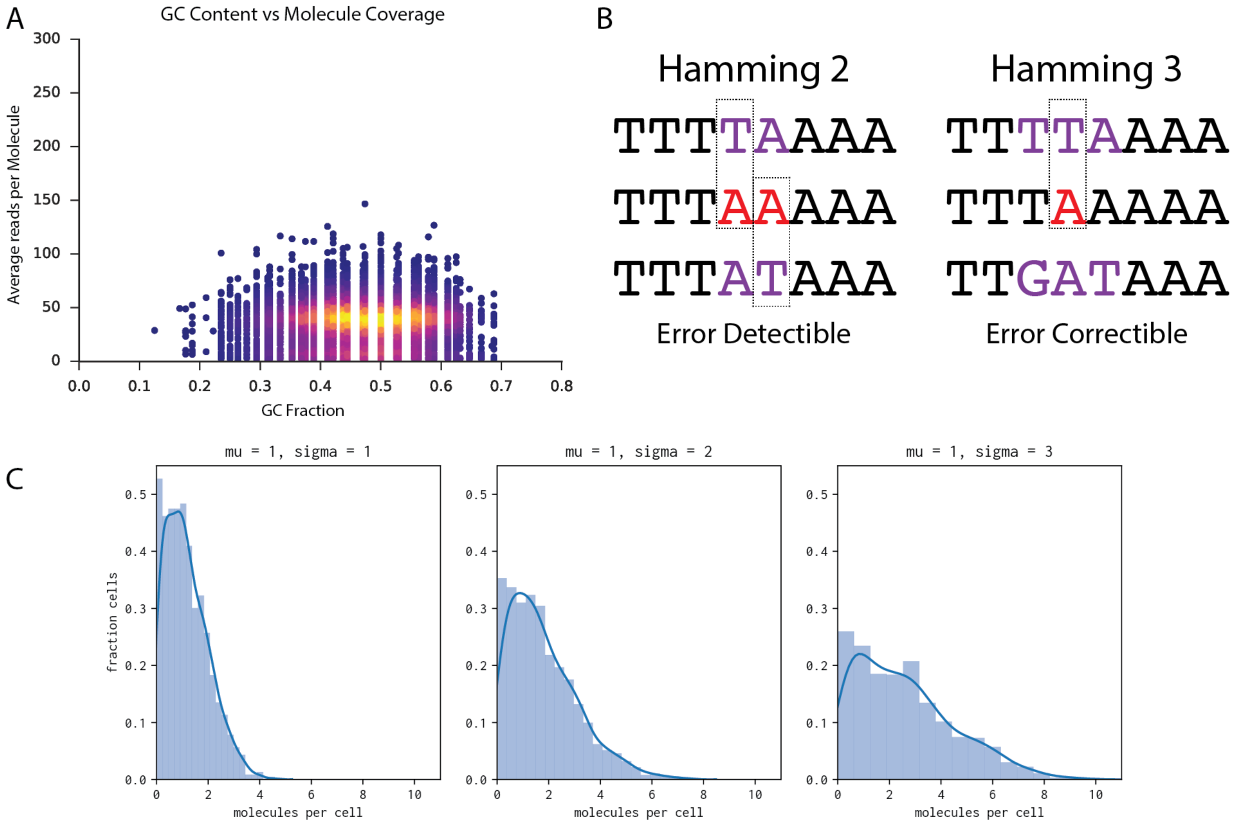


Figure 2.2: (A) Visualization of cell barcode GC content (percentage, x-axis) versus cell barcode read coverage (y) displaying higher coverage at 50% GC content. Color scale represents density of cell barcodes. Yellow is high, purple is low. (B) Two example barcode pairs where top and bottom represent expected barcodes and the middle, with possible errors highlighted in red, represents an observed cell barcode sequence. For the case of Hamming distance of 2 (left), the observed barcode may have been generated from a single T->A mutation in either the fourth position (true barcode is top) or fifth position (true barcode is bottom). If instead a hamming 3 set is used, every single base substitute error can be corrected—the only single-base error that could convert an expected barcode to the observed is a T->A mutation in the fourth position. (C) Toy data displaying the effect of increasing the variance while holding constant the mean and number of drawn samples from a standard Gaussian distribution truncated at zero, where values below zero indicate capture failures and are redrawn.

preservation techniques that ”freeze” cell phenotypes during transport or storage of cells have not yet been adapted to single-cell, it is critically important to rapidly prepare cells for sequencing. This implies that experiments should proceed within minutes but not hours, in the same city, or ideally, in the same institute. Unfortunately, InDrop has a single flow channel per device, and therefore multiple technical samples per patient must be run sequentially. Unlike the cell lines used to demonstrate InDrop's capabilities which can be dissociated and sorted between each run of the InDrop device, patient samples are dissociated contemporaneously, and have different time-lags until cell lysis, at which point apoptosis and RNA degradation are halted.

For one sample where we had abundant cells, we ran 7 sequential technical replicates in series and measured, for each cell, the fraction of molecules that came from mitochondria against the total number of observed molecules (Figure 2.3). We observed that time from extraction correlated with mitochondrial RNA content ($r^2 = 0.98, p < 1e-4$), implying that MT-RNA made up increasing fractions of the cells as time progressed. This suggested that equalizing time from sample extraction to processing is an important technical consideration, and that increasing the speed of in-drop encapsulation would allow us to measure more cells at higher molecule count with smaller MT-RNA-related stress responses.

Given that sample extraction, dissociation and sorting was expected to take approximately 3-4h, we reasoned that if we could increase the speed of encapsulation, we could minimize data variance attributable to differences in encapsulation latency. To increase the cell isolation throughput, we developed a new cell barcoding chip (V2; Droplet Genomics) and adjusted the flow rates for cell suspension at 250 µl/hr, for RT/lysis mix at 250 µl/hr, and for barcoded hydrogel beads at 75 µl/hr. The flow rate for droplet stabilization oil was 550 µl/hr. These flow speeds generated approximately 40,000 droplets an hour, a 250% increase, which allowed us to barcode each sample

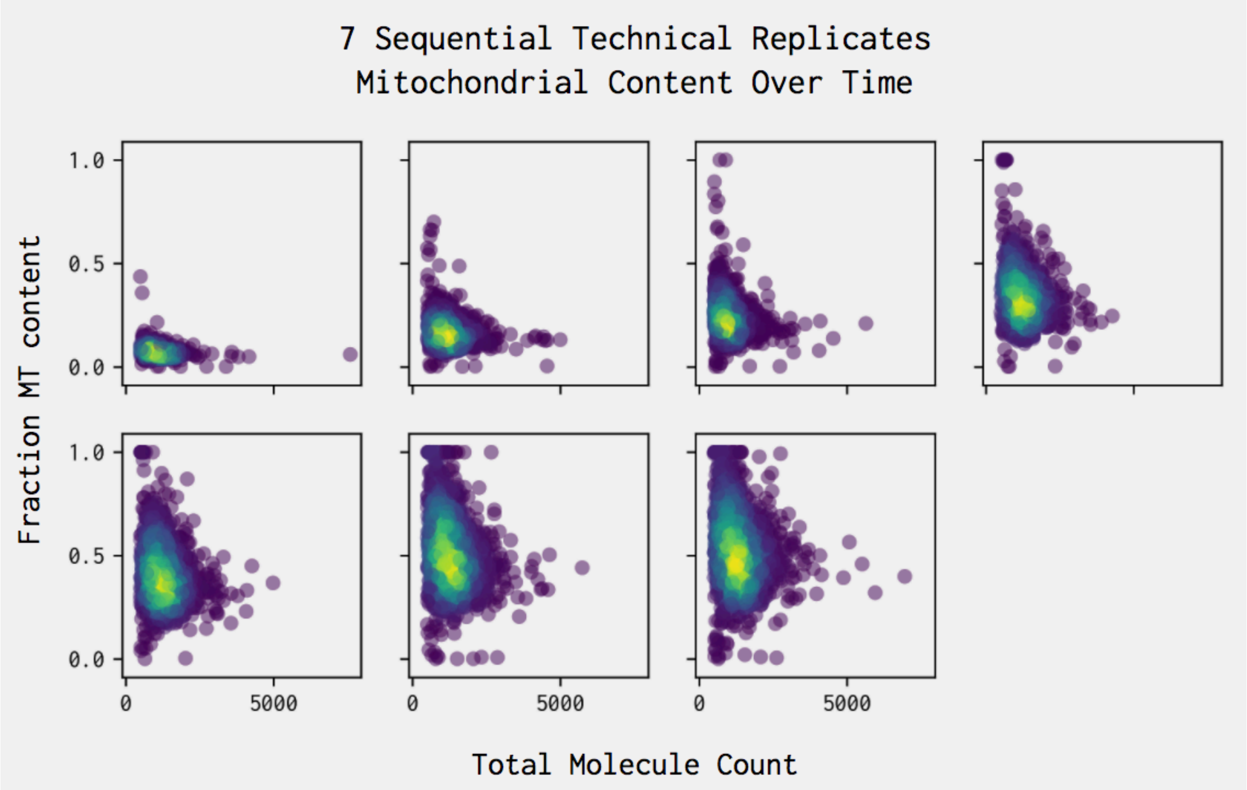


Figure 2.3: Mitochondrial content (y axis) vs library size (x axis) of seven technical replicates for an early InDrop experiment. Color scale represents cell density (yellow is high, blue is low).

in as little as 30 minutes. Thus, if we sequenced each sample in triplicate, and we assumed a fast transport and sample preparation time of 3h, the last sample would take at most 28% longer to process than the first, a 100% improvement[4].

Together, the improvements to the InDrop chip, redesign of the library, and troubleshooting and optimization of cell to reagent ratios transformed InDrop into a sequencing platform well suited to comparing immune phenotypes within and between multiple patients and tissues.

[4] 10x Genomics now provides a device that can encapsulate 8 samples in parallel. This can be a superior approach for samples that are significantly perturbed by temporal effects.

2.6 Data Preprocessing: SEQC

Despite improvements to both the microfluidics and barcodes, InDrop data retains many of the fundamental problems described in Section 2.2 that demand computational solutions: Correction of barcoding errors, removal of non-cellular contamination, and discrimination of cells from ambient contamination. Unfortunately, at the time of data collection, published data processing pipelines were specifically tailored to the library construction methods they were designed to process. In addition, the rapid pace of technology development has induced computational approaches to be constructed with similar haste; most novel computational methods were bash scripts (Shalek et al., 2013; Shalek et al., 2014), unpublished R scripts (Jaitin et al., 2014), tools published without source code (Macosko et al., 2015) or written descriptions without software (Klein et al., 2015).

Given the fast rate of technological evolution, we believed that we and others would benefit from a modular data processing package capable of rapid adaptation to changes in data generation from multiple technologies. To address this deficiency, we developed SEquence Quality Control (SEQC), a general purpose python package to build a count matrix from single cell sequencing reads which is able to process data from InDrop, Drop-seq, 10X, and Mars-Seq2 technologies, but more critically, a package that incorporates cutting edge analysis methods to maximize signal:noise in scRNA-seq data. The SEQC package, which takes Illumina Fastq or BCL files, the standard sequencing data formats, and generates a count matrix that is carefully filtered for errors and biases; the SEQC package is outlined in Figure S2.4.

Briefly, SEQC begins by extracting the cell barcode and UMI from the forward read and storing these data in the header of the reverse read. This produces a single fastq file containing alignable

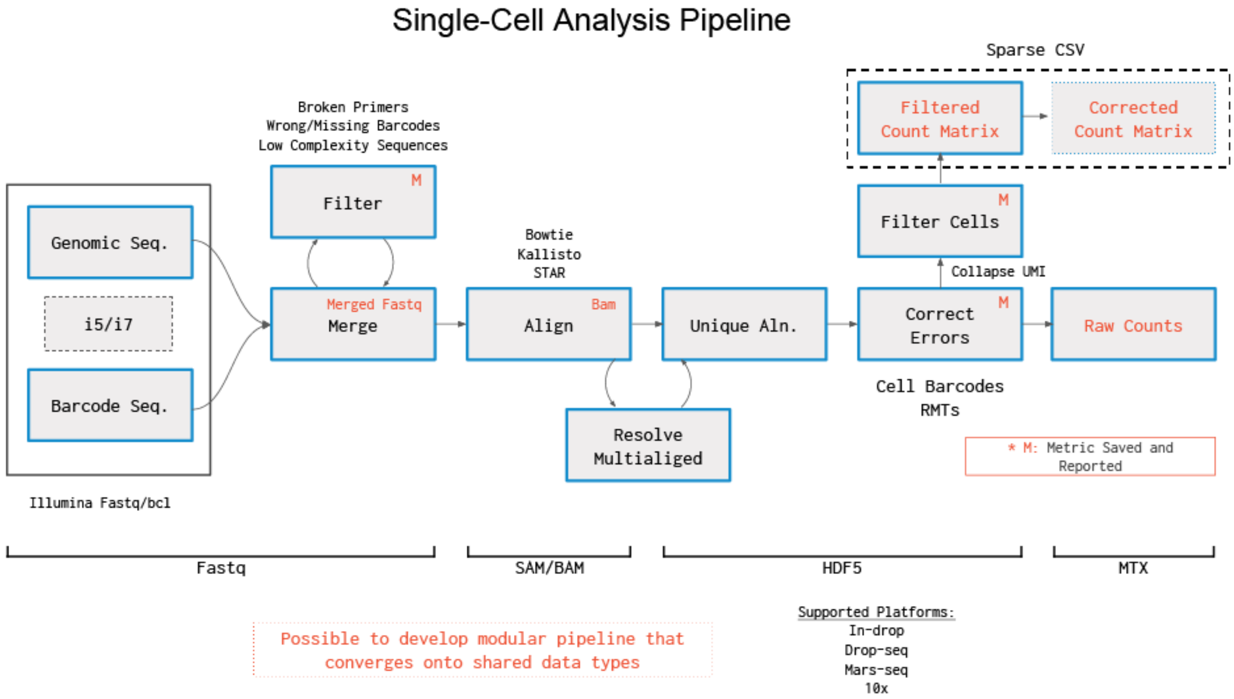


Figure 2.4: Schematic of the SEQC package. Data analysis proceeds from right to left through modules, following the directed arrows.

sequence and all relevant metadata. Merged fastq files are aligned against the genome with STAR (Dobin et al., 2013), a high performance and community-standard aligner. After alignment, minimal representations of sequencing reads are translated into an Hdf5 `ReadArray` object, where cell barcodes and UMIs are represented in reduced 3-bit coding. Reads are annotated with a reduced set of exon and gene ids representing gene features—only the ones that are possible to detect with poly-A capture based droplet RNA sequencing—and SEQC attempts to resolve reads with multiple equal-scoring alignments. The Hdf5 `ReadArray` object is efficiently indexed and is an ideal data structure for in-memory filtering of cell barcode substitution errors, broken barcodes, and low-complexity polymers to flag errors early in the pipeline, saving analysis cost (Alted and Vilata, 2002–).

In cases where genomic and transcriptomic alignments are present, the transcriptomic alignments are retained. Unique alignments from the previous step are corrected for errors using an enhancement of the method designed in Jaitin et al., 2014, with an additional probability model to constrain the false positive rate. The error-reduced, uniquely-aligned data are grouped by cell, molecule, and gene annotation, and compressed into count matrices containing (1) reads and (2) molecules.

This matrix is thresholded in a similar manner to what has been previously described (Macosko et al., 2015; Zheng et al., 2017b). These data matrices contain cells as rows and genes as columns, where the entries in the matrix represent the number of molecules of a given gene observed in a cell. Consequently, row vectors represent the observed frequencies of each gene in a cell, similar to the read-out of a bulk sequencing experiment, while column vectors summarize the distribution of gene observation frequencies across the experiment. These count matrix data structures serve as the basis for most analyses of single-cell RNA-seq, and are the major deliverable of any data processing pipeline.

Finally, SEQC outputs a series of QC metrics in an HTML archive that can be used to evaluate the quality of the library and the success of the run. SEQC is fully modular, which facilitated easy adaptation to use with drop-seq, 10x, and mars-seq data. In addition, it can be configured either to run on a local high-performance cluster, or can automatically initiate runs on Amazon Web Services compute platforms, for those without access to local compute servers. The SEQC code is free and open-source, and can be found at https://github.com/ambrosejcarr/seqc.git, licensed under the MIT license. A public Amazon machine image with SEQC pre-installed is available at AMI id: `ami-8927f1f3` and a docker image of SEQC that can be used to launch experiments against a user's AWS account is available at `ambrosejcarr/seqc:1.0.0` The following sections describe each SEQC module in detail.

2.7 Data Complexity Requires Flexible Optimization Strategies

To solve the classification problems that plague scRNA-seq data, like discrimination of cells from ambient contamination, one would usually design an experiment that would allows an experimenter to label the cells. Given this labeled data, one would search for data features that differentiate the two conditions (cell and contaminant). However, the experiments that generate this type labeled data for droplet-based sequencing were too easy, and didn't generalize to use on human tissue.

For example, both InDrop and Drop-seq technologies carried out experiments to show very low contamination when human and mouse cell lines are mixed, measured by the number of cells that had both human and mouse DNA. Unfortunately, this experiment is both favorable to the assays, since cultured cells don't lyse as frequently as clinical samples, and underpowered, since human and mouse genomes are similar enough that a large number of potentially-contaminating fragments can't be specifically assigned. As a result, when we attempted to extrapolate from this cell line experiment to a more complex human tissue sample, the features we learned from the cell lines co-varied in much more complex ways in the tissue sample, and failed to accurately predict doublets or contaminated cells. Thus, this type of control experiment failed to generalize to clinical data.

As a result, as the individual algorithms of SEQC are described, they will primarily be evaluated based on their ability to optimize data characteristics like the removal of barcoding errors or recovery of additional molecules. Assumptions, when made, will be stated clearly. However, the Chapter will conclude by examining the aggregate impact of the SEQC methods, showing that taken together, these methods were critical to uncovering the biological structure of the data, and that without SEQC, the high inherent noise in scRNA-seq data made it impossible to group cells of similar types or states.

2.8 Fastq Demultiplexing

The file formats for sequence data were designed for bulk sequencing data, wherein all sequences are expected to contain genomic information, not barcodes. As a result, there is no standard approach for storing the non-genomic barcodes with the genomic sequence in a way that is compatible with alignment algorithms. This has produced numerous different, and incompatible, methods, that either involve format conversion (Macosko et al., 2015) or inclusion of unicode text tags in the fastq name field (Jaitin et al., 2014). Both approaches incur significant computation or storage cost. However, 3' scRNA-seq approaches all share characteristics that facilitates a common specification: each technology uses one or more barcode to define a cell, and contains a UMI. Even complex cases such as Mars-seq, which has an additional "pool" barcode that defines the plate of origin, can have the cell and pool barcodes concatenated to define a unique cell in a multi-plate experiment. Thus, if there were a standard abstraction for cell barcodes and UMIs, it would facilitate rapid analysis of diverse scRNA-seq data types.

The first stage of SEQC addresses this shortcoming by taking input fastq files containing genomic information and barcoding metadata spread arbitrarily across multiple sequencing files, and merges that information into a single fastq file. Sequence data is stored prepended to the first read name, delimited by a colon, and separated from the original read name with a semicolon. For InDrop, which contains cell and molecule barcodes, plus a number of T nucleotides, the name field of a record in the merged fastq appears as follows: `@<CELLBARCODE>:<UMI>:<#T>;R2 READ NAME`. No additional tag information is stored, and the sequence found in the name field is

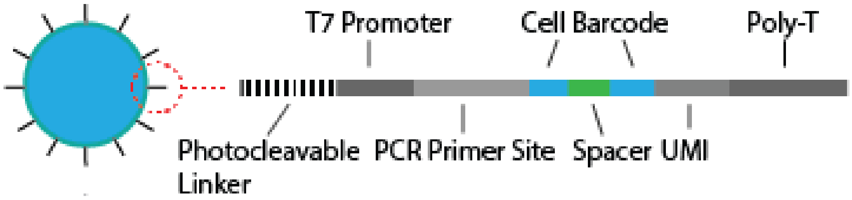


Figure 2.5: Schematic of capture primer displaying amplification machinery, cell barcodes, UMIs, and poly-T capture site.

efficiently compressed by the zlib compression library (Gailly and Adler, 2004).

To remain general, SEQC implements a platform class comprising the locations of the cell barcodes and UMIs, the type of barcode and UMI correction to be run, the number of T-nucleotides that are expected to be read from the capture primer, and a merge function that indicates how to extract barcodes and construct the standardized merge fastq file. Thus, the platform contains the complete information required to specify where to find the barcodes that define a read's provenance, but also the algorithms that must be run on a particular library construction method to generate optimal scientific data. This allows us to produce a single file format, a merged fastq file, that losslessly represents all known types of 3' RNA-seq data.

The merged fastq file contains genomic, alignable sequence in the sequence field, and has read metadata prepended to the name field, separated by colons. This step can be adjusted for novel sequencing approaches by adding a new platform class, often with only 10 lines of code[5]. This allows the complete SEQC pipeline to be rapidly tested on iterations of InDrop, or novel technologies.

InDrop has a more complex library construction process that required us to devise a custom fastq merging solution. InDrop constructs its cell barcodes from two pools of 384 cell barcode fragments which hybridize in a constant "spacer" sequence (see Figure 2.5). Illumina sequencers

[5] see Figure 1 for the 10-line platform that allowed us to adapt SEQC to 10x v2 chemistry when it was released

cannot read constant sequences, as observing the same base simultaneously at all points on the chip saturates the fluorescence sensor and prevents localization of base calls to individual read ”spots” (see Metzker, 2010 for a review of sequencing technology chemistry and limitations). To prevent this, InDrop’s first cell barcode fragment has 4 lengths, which causes the spacer to have 4 different offsets, produced a library that is easy to sequence.

However, this organization required us to localize the spacer sequence on the fly for each sequencing read. The original InDrop publication accomplished this with exact pattern matching, however this is computationally expensive An alternative that can identify errors in the spacer sequence is fuzzy-matching, but this approach extended the run time of SEQC by several hours, and thus is computationally prohibitive. The latter problem was significant, as we would occasionally see sequencing experiments with a single failed ”N” cycle in the middle of the fastq file. In these cases, the data was 100% viable, as there were no failures in the barcodes or genomic sequence, but the existing approach would fail all reads.

SEQC addresses this problem by identifying a 4-base window within the spacer that is unique at all four spacer offsets, and hashing the observed windows to the cell barcode fragment lengths that they correspond to:

```
GAGTGATTGCTTGTGA|CGCC|TT---
-GAGTGATTGCTTGTG|ACGC|CTT--
--GAGTGATTGCTTGT|GACG|CCTT-
---GAGTGATTGCTTG|TGAC|GCCTT
```

If the sequence fails to match, then a fuzzy pattern match defined in a fast C-extension is run against the failing read, identifying spacers with up to 3 substitution errors. Once the spacer is identified, the cell barcodes, UMI, and the number of sequenced T-nucleotides from the capture primer’s tail are all stored for downstream processing in the fastq name field. The generated fastq file has the following format, where R2 refers to the read carrying the genomic sequence, while the barcode sequences come from R1[6]:

```
@<CELLBARCODE>:<UMI>:<#T>;R2 READ NAME
R2 SEQUENCE
+
R2 QUALITY
```

2.9 Alignment

Data collected from the sequencer consists of mRNA fragments. To draw biological conclusions about a dataset, fragments must be matched to the part of the genome the mRNA were transcribed from. This process is carried out by assembly algorithms when the genome is unknown (Haas et al., 2013), and alignment algorithms when there is a reference genome that can be compared with. Most model systems examined with scRNA-seq have known genomes, so SEQC was built against aligners by default. However, because SEQC does not utilize any custom tags generated by the aligners or assemblers, it is compatible with any method that takes data from multiple cells in fastq format and outputs a BAM file[7]. We selected STAR as the default aligner because it is a fast, highly parallel, cloud-scalable aligner that benchmarks well against existing aligners[8] (Ilicic et al., 2016). We note that STAR automatically trims bases as necessary to find alignments, and as such no pre-trimming of reads based on quality is carried out. Alignment parameters used are as follows:

[6] Extraction of barcodes is parameterized, and supports chemistries where genomic sequence lies on R1 such as Mars-seq. Additionally, conversations with the author of STAR have prompted them to add support for the alignment of reads in BAM format, a significantly more flexible format with better support for the inclusion of cell barcode and UMI tags. Future iterations of SEQC will move towards a merged format that utilizes BAM instead of Fastq format files.

[7] Kallisto requires the user to determine cell assignment before alignment. Support for Kallisto is in-process.

[8] Since the design of SEQC, Hisat2 (Kim, Langmead, and Salzberg, 2015), an algorithm based on the Bowtie2 burrows-wheeler strategy (Langmead and Salzberg, 2012), was released which promises higher speed and lower memory usage. We are benchmarking this aligner for possible replacement of STAR.

```
    -outFilterType BySJout,
    -outFilterMultimapNmax 100,
    -limitOutSJcollapsed 2000000,
    -alignSJDBoverhangMin 8,
    -outFilterMismatchNoverLmax 0.04,
    -alignIntronMin 20,
    -alignIntronMax 1000000,
    -readFilesIn fastqrecords,
    -outSAMprimaryFlag AllBestScore,
    -outSAMtype BAM Unsorted
```

This module thus takes as input a fastq file and produces a BAM file containing up to 20 multiple alignments per input fastq record and with all unaligned reads contained in the same file.

2.10 Annotation Construction

Aligners identify the best match of each sequencing fragment to the genome, finding the chromosome, and the position on that chromosome, for each fragment. A critical step after the alignment of reads is to determine the gene that overlaps the chromosome coordinates the aligner assigned to the fragment. Gene location information is summarized by a genome annotation, a set of metadata including exons, introns, transcripts, genes, and untranslated regions, that are matched to genomic coordinates. Bulk sequence alignment recommends the use of the complete genome annotation, and this recommendation has been applied to scRNA-seq data without modification (Shalek et al., 2014; Jaitin et al., 2014; Klein et al., 2015; Macosko et al., 2015). However, because the genome annotation is designed to be a comprehensive compendium of information about an organism, it contains many features that are theoretically undetectable by InDrop and other 3' sequencing technologies.

Two characteristics of InDrop limit its ability to capture certain gene biotypes. First, it employs poly-A capture, and thus will not detect non-polyadenylated transcripts. Second, it uses SPRIselect beads at several stages to deplete primers from reaction media. These beads carry out size selection, preferentially depleting primers but also small RNA species such as snoRNA, miRNA, and snRNA. Thus, libraries are expected to contain only transcribed, polyadenylated RNA of length > ~200 nt. Examining gene biotypes, this meant retaining protein coding and lncRNA biotypes, and excluding others. We hypothesized that the reduction in reference features would result in a concentration of alignments in biologically relevant genes by depleting non-specific features, and that there would be many drop-out events where genes would be detected in the complete reference, but not the subset. Two methods exist to address this problem, but we find that neither method is appropriate for 3' sequencing data.

Cell Ranger, the most commonly used pipeline to process 10x Genomics data, carries out an extreme version of this redesign: it removes any gene that is not protein coding. We believe that this is too harsh: it excludes numerous transcribed pseudogenes and lncRNA which have been previously shown to be expressed, have biological functionality, and to be detectable in scRNA-seq.

Alternatively, alignment can be restricted exclusively to transcriptomic features. Several methods implement this approach, including Kallisto (Bray et al., 2016) and Tophat2 (Kim et al., 2013). However, 3' scRNA-seq data typically contains between 10-30% genomic contamination, as identified by reads aligning intergenic alignments. When we aligned directly against the transcriptome using TopHat2, we found that approximately 1% of intergenic reads were mistakenly aligned to exonic locations despite having higher alignment scores to genetic regions (data not shown). Without knowledge of the genome, these reads would be mistakenly counted as gene alignments, introducing significant error into the count matrix.

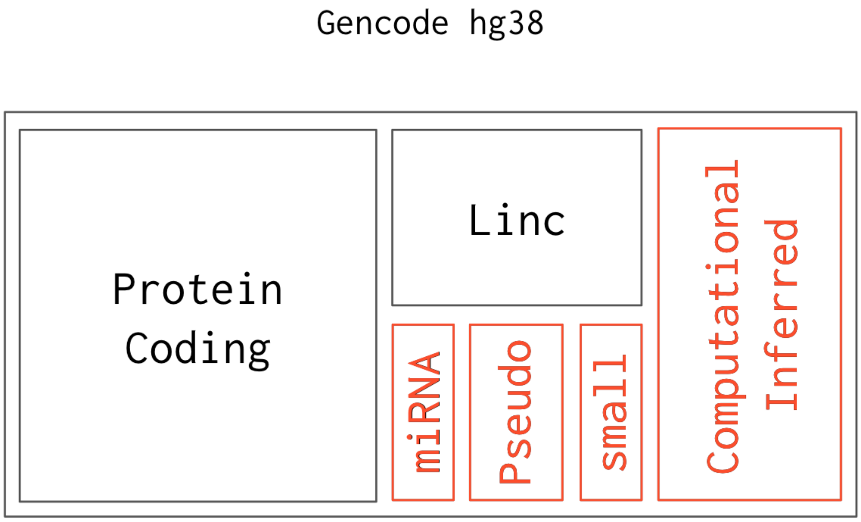
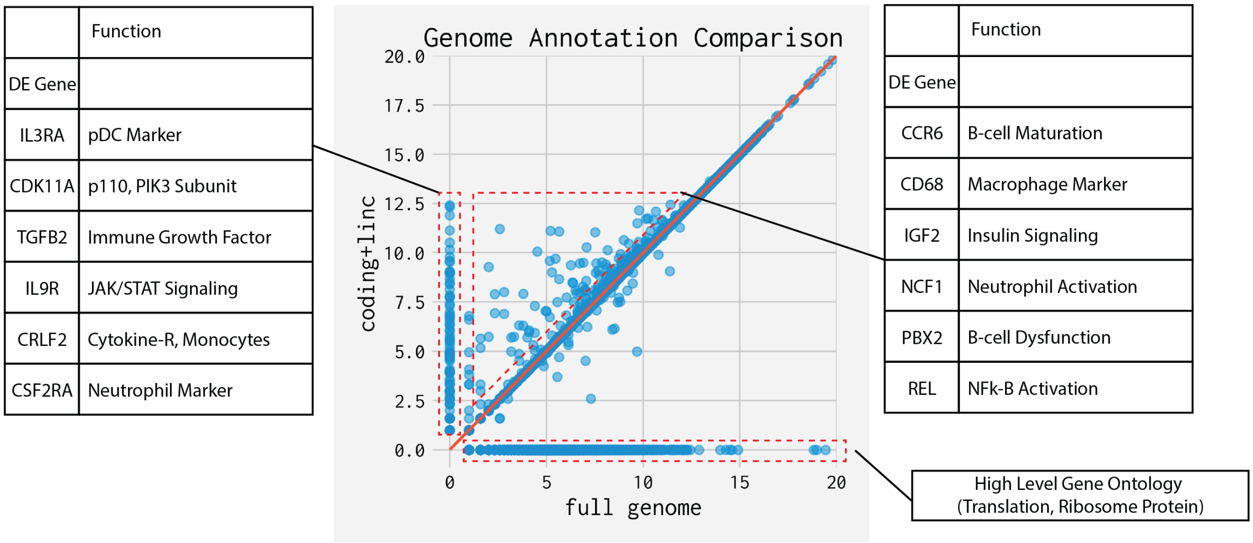


Figure 2.6: Not-to-scale schematic of the major components of the GENCODE genome annotation. Transcripts from categories in red should not be observed by SEQC, either due to beads which remove small molecules, or a lack of poly-a tails, which are used by SEQC to capture RNA.

To address this, we constructed a custom annotation by starting with the current GENCODE genome and GTF file and removing all feature annotations that are not theoretically detectable by InDrop (Figure 2.6). We then align to the full genome, but prefer transcriptomic alignments in cases where there are equivalent genomic and transcriptomic alignments.

To determine the impact of this change of reference on our data, we aligned the same single-cell immune dataset against the full reference and the reduced reference described above. We constructed a "pseudo-bulk" dataset for each reference by summing the molecules of each gene across all cells, producing an expression vector that contained the total number of molecules of each gene detected by each annotation. We then examined the correlation, and discrepancies, between the two references. (Figure 2.7).

The overall r^2 value between the references is 0.94, with 93% of genes holding the exact same values in both reference alignments. In addition, information is concentrated in 35% fewer features, despite losing only 8% of the total molecules. There is also a large drop-out contingent



	Function
DE Gene	
IL3RA	pDC Marker
CDK11A	p110, PIK3 Subunit
TGFB2	Immune Growth Factor
IL9R	JAK/STAT Signaling
CRLF2	Cytokine-R, Monocytes
CSF2RA	Neutrophil Marker

	Function
DE Gene	
CCR6	B-cell Maturation
CD68	Macrophage Marker
IGF2	Insulin Signaling
NCF1	Neutrophil Activation
PBX2	B-cell Dysfunction
REL	NFk-B Activation

Figure 2.7: Comparison of complete GENCODE annotation against a reduced annotation containing only GENCODE-annotated lncRNA and protein coding RNA. Displaying drop-out events occurring on x-axis as well as masking events on y-axis.

present only when aligned against the complete reference. Gene ontology enrichment against this reference revealed high-level biologically agnostic enrichments, such as “protein coding,” “translation,” and other enrichments, which suggest a random sampling of high-expression genes.

Surprisingly, there was also a contingent of genes present only in the reduced alignment. These genes were highly enriched for immunological pathways, including JAK/STAT signaling, cytokine production, cytokine receptors, and immune growth factors, and further included critical immune genes such as IL3RA, a plasmacytoid dendritic cell marker (Figure 2.7). This suggests that they are likely to represent true annotations for genes in this dataset, and that reducing the annotation produces a gain in specificity. We reasoned that these genes were uncovered in the reduced annotation because there are features in the complete set, such as pseudogenes, which have high homology to transcribed genes.

Including these annotations, which should not be detectable, produces illogical multi-alignment to multiple genetic locations. When such multi-alignment cannot be resolved, most pipelines (including this one) exclude those multi-aligned reads, losing valuable signal. Given these results, we believe that the 8% reduction in molecules cited above that occurs from switching to the reduced reference is the result of correctly discarding low-complexity alignments that were spuriously assigned a low-quality transcriptomic feature. Thus, this change allows 3' sequencing to detect more genes than strategies that do not utilize this approach.

2.11 An In-Memory Hdf5 Read Store Allows Efficient Computation over Single-Cell Sequencing Data

Sequencing data formats, Fastq (Cock et al., 2009) and BAM (Li et al., 2009), were designed with bulk sequence data in mind. Each file is designed to house a single sample, and is efficiently indexed for random access by genomic coordinate. This is a critical capability for genome sequencing, where the full dataset is far too large to fit in memory, and tasks commonly center around detecting genomic variants which exist at defined chromosome positions (McKenna et al., 2010) It is much less useful for scRNA-seq data, where most computational methods require random access to the data associated with a given cell or molecule.

To address this problem, we designed a ReadArray data structure that summarizes the information in an aligned BAM file that is critical to scRNA-seq analysis. The ReadArray is built on top of the `Hdf5` platform (The HDF Group, 1997-2018) using the `pytables` package, which confers three advantages: First, Hdf5 is a columnar data format that supports arbitrary multi-column indices. This allows indexes to be built for both cells and molecules. Second, it is a numeric format that can be efficiently compressed (Alted, 2010) to have a smaller disk footprint than the BAM format. Finally, since most scRNA-seq experiments devote at most two lanes of sequencing to each set of cells, and replicates are processed independently, the complete data format fits into

10Gb memory, which allows for rapid querying with decreased computational cost.

Several changes to the data representation were made to shrink the in-memory footprint of the ReadArray. First, sequence information for the cell barcode and UMI are stored in a 3-bit encoding[9] and the nucleotides are concatenated to fit into a 64bit integer (cell barcode) and 32 bit integer (UMI). Second, information that is summarized multiple times in the BAM format, like the genomic sequence information and chromosome and alignment position, are summarized by the minimal representation that confers adequate knowledge. In this case, the chromosome and position are retained. Third, the results of each filter are stored as binary status flags, storing analysis results concisely in a way that is extremely fast to filter over. Finally, information that is extraneous to scRNA-seq analysis, such as custom BAM tags and sequence quality scores, are excluded completely.

The resulting ReadArray specification is broken up into two parts, a core of status, `cell, rmt`, and `n_poly_t` which have a fixed disk representation, and `gene, position, chromosome`, and `strand`, which are initially represented on disk as JaggedArrays, a flexible representation where each array index may support multiple alignments. Once Alignments have been disambiguated, they are converted to columns to reduce memory usage. Regardless of the stage of processing, the interface to access the fields remains constant. The complete specification is as follows:

```
_dtype = [
('status', int16), # if > 16 tests, use int32
('cell', int64),
('rmt', int32),
('n_poly_t', int8)
('gene', int32), # initially empty
('position' uint32), # variable on-disk implementation
```

[9] It is possible to encode A, C, G, and T in 2-bits to further compress the representation, but we elected to use 3-bits to support N-nucleotides, as otherwise N nucleotides must be randomized to one of A, C, G, or T

```
    ('chromosome' int8), # variable on-disk implementation
    ('strand', int8)]
```

Thus, a single record fits in 25 bytes, and 400M sequencing reads, the equivalent of two illumina lanes, will fit into 10Gb memory. Adjustments to the ReadArray format are relatively simple to make for Fixed or Variable representation fields. The need only add a field name and numerical type to the above specification, and define an extraction method in the `ReadArray.from_samfile()` constructor

One field that is conspicuously absent from the ReadArray specification are the sequencer quality scores. Some pipelines, such as 10x Genomics' Cell Ranger, posit that sequencing error is the major source of substitution mutations in 3' sequencing data (not enzymatic error during library construction), and thus is predicted by barcode quality scores. If this were true, quality scores could be used to help correct barcode errors.

Our InDrop data does not support this view[10]. In InDrop, each read contains a 16-19 bp cell barcode selected from a whitelist of known barcodes. By examining barcodes for single base mutations, we estimated a positional, nucleotide-specific error rate for each sample (Table S1). E.g. to calculate the probability of a conversion from adenosine to cytosine, where $A \rightarrow C$ denotes this nucleotide conversion: $P_{A\rightarrow C} \ = \frac{1}{n\cdot m} \left\{ 1 \ if \ x_{ij \, : \, A\rightarrow C} \ else \ 0 \ \right\}$ where $x_j$is a barcode, j $\in \{1,\ldots,m\}$ and each barcode has $n$bases. The average observed per-barcode error rates are 4%, a number far in excess of the abundance reported by the Illumina sequencer, which can be reliably calculated from errors in phiX included in sequencing runs (mean error rate 0.2% ∓ 0.1%) (Manley, Ma, and Levine, 2016); a 4% error rate is more in line with aggregate error rates of the

[10] An analysis of 10x data found similar results to those described for InDrop.

enzymes used in the preparation of sequencing libraries (Zilionis et al., 2017).

To verify that quality scores do not predict error rates, we tested the correlation between the error state of the cell barcode (1 if the base contains an error and 0 otherwise) with Illumina quality scores. If quality were predictive of substitution errors, we would expect to observe strong negative correlations, suggesting that low quality implies high error probability. However, we observed no relationship (mean r^2=0.04, max r^2=0.06; 'C' errors) on either InDrop or 10x data. In contrast, mutations to N bases produce the expected relationship, with base quality negatively correlating with → N substitutions (r^2=-0.87). However, N base errors made up less than 1 / 100,000 of the observed errors in our experiment, and we conclude (1) that base quality is not meaningfully predictive of error rates, and (2) that most sequenced error is derived from upstream library construction steps.

2.12 Barcode Sequencing Errors Arise In Library Construction and Are Correctable

Proper assignments of reads to the molecule and cell they were captured in is a critical step in scRNA-seq analysis. Under ideal circumstances, the combination of the UMI and cell describe the cell of origin. However, there are two major sources of error that are introduced during library construction: primer fracturing and barcode substitution errors. These errors confuse this association by disrupting the matching of observed barcodes with the barcodes that were present on primers during mRNA capture.

Cell barcode errors in InDrop (Figure 2.8 C) are easy to detect by design: we have a whitelist of 147,456 barcodes, each with Hamming distance >= 3. Thus, any single base substitution error is resolved by creating a lookup table for all barcodes and all single substitutions. If found in the table, the barcode is corrected. If not, it is discarded. As estimated above, the probability of a cell barcode containing an error is ~2%, and thus the expected rate of barcodes accruing 2 errors in a barcode is 1 / 2500. A 2-error lookup table has a very large memory footprint and would significantly increase computational cost of processing each experiment. Alternative algorithms have greater complexity and would increase run time. Thus, we accept this low rate of loss and proceed to correct single base errors, recovering approximately 2% additional data for each sample.

This error rate, while high, is easy to correct and results in minimal data loss. Although a 4% barcode error rate is higher than the error rate observed by other technologies, it directly results from the use of linear amplification. If errors are independent, then the probability of obtaining 2 or more errors, which would produce an uncorrectable barcode, is 1 - cdf beginning at 2 for a Poisson distribution with $\lambda = 0.04$. Given the observed error rate, we estimated that only 0.035% of barcodes would be uncorrectable. This allows SEQC to correct errors using a fast hash-based strategy. This is in contrast to PCR-based amplification approaches which propagate errors that occur in early cycles, requiring more complex, graph-based correction methods, and larger Hamming buffers (see https://github.com/vals/umis).

In contrast to cell barcodes, UMIs are random, and correction cannot proceed by the same strategy, so we devote a section later in the pipeline to the detection of UMI errors after the gene and mapping position of a fragment are identified.

Another source of error in scRNA-seq experiments, including InDrop, cel-seq, mars-seq, and likely drop-seq and 10x genomics, is the fracturing and random-priming of capture primers (Figure 2.8 A, B) (Jaitin et al., 2014). We often observe cell-barcode prefixes followed by randomers. When fragmentation occurs at the cell barcode level, we can remove the fragments using the

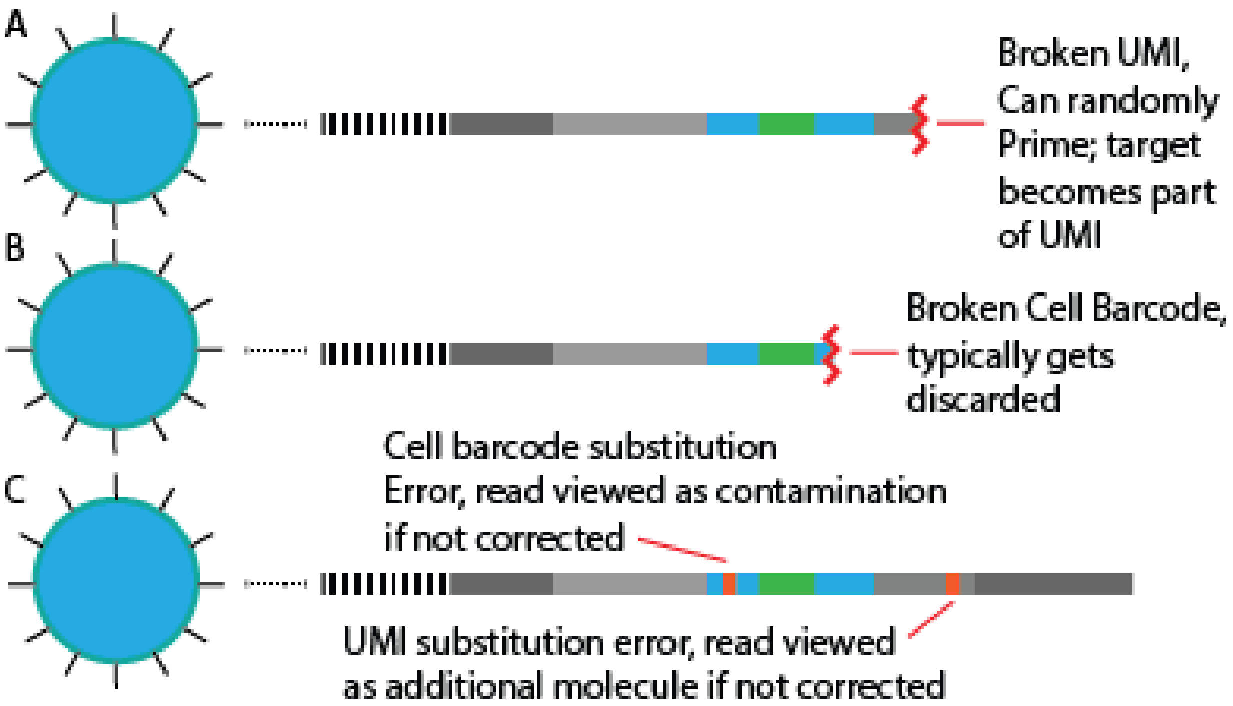

A
Broken UMI,
Can randomly
Prime; target
becomes part
of UMI
B
Broken Cell Barcode,
typically gets
discarded
Cell barcode substitution
Error, read viewed as contamination
if not corrected
C
UMI substitution error, read viewed
as additional molecule if not corrected

Figure 2.8: This figure displays a schematic that describes the types of barcoding errors that can occur in InDrop data, but also other approaches that utilize 3' or 5' capture by poly-A primers. These error sources are: (A) the barcode fragments within the UMI sequence; these barcodes may randomly prime, if broken before encapsulation, or produce incorrect UMIs or fragment-fragment hybridization events, if this occurs during library preparation. (B) A barcode that breaks within the cell barcode. Because InDrop has a set of valid barcodes, these errors are easily excluded, but the process that produces these errors is the same as in (A). In (C) we display barcode substitution errors, which may cause barcodes to aberrantly manifest as separate cells or molecules, depending on which barcode they occur in.

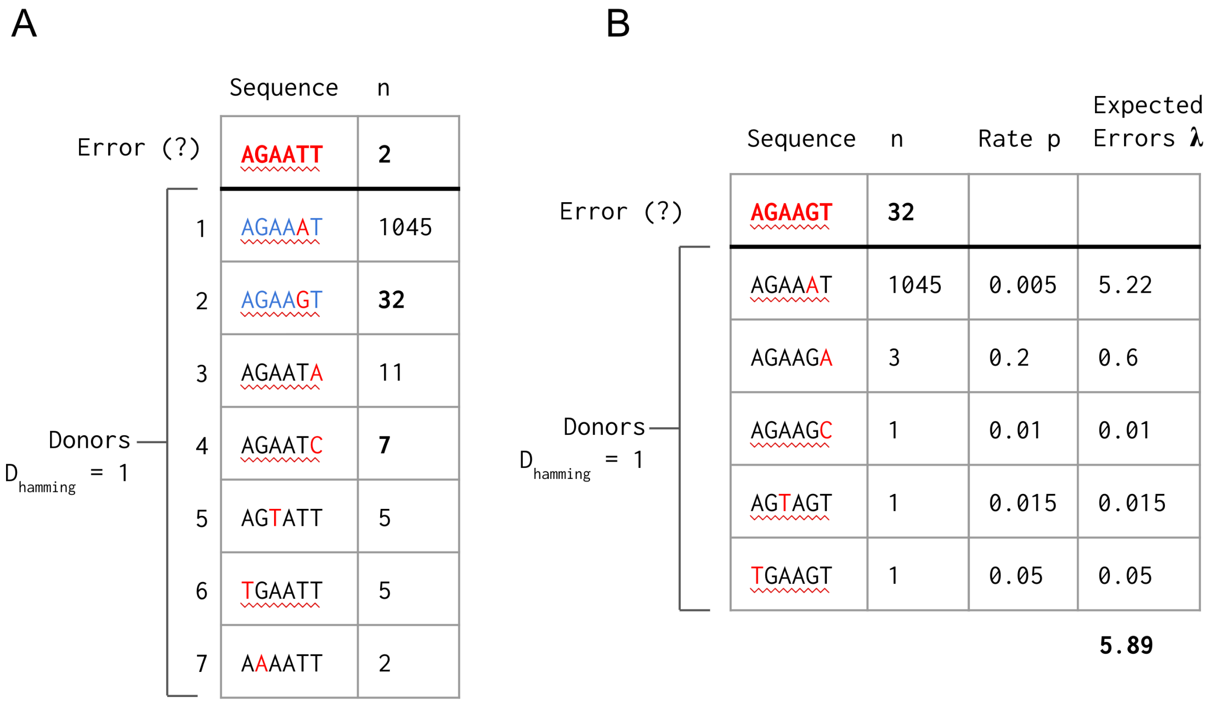
whitelist approach above. To remove barcodes that break in the UMI, we determined that we would sequence 5 bases into the poly-T tail of the primer, which we expect to be all T-nucleotides. By excluding reads with more than 1 non-T nucleotide, we are able to exclude most broken UMIs.

The second source of error are barcode fractures, observed in both CEL- and SMART-seq chemistry. A barcode fracture occurs wherein some prefix of the Cell, and UMI, and in the case of InDrop, also the spacer and poly-T tail, is observed, but the remainder of the read corresponds to non-barcode information. Barcodes that break in cell barcode sequences will be excluded by cell barcode correction, as described above. However, UMIs are not *a priori* known; if an aligned read breaks inside the UMI sequence during amplification, it will manifest as a new molecule despite having a proper, full-length UMI that it should be associated to. This will result in inflation of UMI counts for the matched gene.

To test for the presence of these types of errors, we used a `trie` data structure to efficiently count the largest hamming-corrected cell barcode prefix observed in each of our *aligned* sequencing reads. We used the hamming-corrected barcodes because we reasoned that substitution errors would be the most common error type, and wanted to exclude those from analysis, as they are corrected through other methods, described above. The largest cell barcode prefix is most often a complete cell barcode, owing to the high quality of InDrop data. However, for 4.7% of our sequencing reads, the prefix is a partial barcode.

These partial barcodes could arise from multiple sources. One option is an insertion or deletion in the barcode. Errors of this type would produce a frame shift in the barcode. A second option is that the barcode has broken, and the broken end acted as a randomer, an alternative capture strategy to poly-A capture. To differentiate between these cases, we calculated, based on the list of known InDrop barcodes, which suffixes match single base insertions or deletions (indels). We then determined whether there was an existing barcode that explained each broken primer.

Cases where an indel explains the observed barcode prefix were very rare (approximately 1/4000), and most prefixes did not contain the expected poly-A tail at any offset. Thus the more likely explanation is random priming. As a result, we assume that reads missing the poly-A tail may have fractured within the UMI, and those reads are flagged for filtering. In aggregate, the filters in this section remove an average of 36% of reads (sd = 9.3%), depleting the count matrix of


A
B
Donors
$D_{hamming} = 1$
Donors
$D_{hamming} = 1$

A

		Sequence	n
Error (?)		**AGAATT**	**2**
Donors	1	AGAAAT	1045
	2	AGAAGT	**32**
	3	AGAATA	11
	4	AGAATC	**7**
	5	AGTATT	5
	6	TGAATT	5
	7	AAAATT	2

B

	Sequence	n	Rate p	Expected Errors λ
Error (?)	**AGAAGT**	**32**		
Donors	AGAAAT	1045	0.005	5.22
	AGAAGA	3	0.2	0.6
	AGAAGC	1	0.01	0.01
	AGTAGT	1	0.015	0.015
	TGAAGT	1	0.05	0.05
				5.89

Figure 2.9: A mock-up of the error correction query for a single molecule (red, top) using the Jaitin method (left) and the SEQC approach (right). Each table displays all of the molecules with a cell barcode that matches the top sequence that have a molecular barcode (1st column) within 1 edit distance of the query barcode (red) and their abundances (second column). The Jaitin approach (A) will discard any molecule when another molecule is observed to align against the same gene in the same cell. In this example, each of the bolded barcodes (0, 2, 4) would be discarded. The SEQC approach (B) Builds a probability model to estimate the expected rate that each barcode would convert into the query barcode.

spurious molecules (see Table S1 for detailed values). These values are consistent with the results of running SEQC on drop-seq or MARS-seq datasets (data not shown).

2.13 Molecular Identifier Correction

Errors in molecular identifiers are well-known to introduce noise in sequencing experiments (Jaitin et al., 2014), since undetected errors induce spurious increases in molecule counts. SEQC utilizes information in the ReadArray to identify errors in UMIs, and replace them with their corrected value. The most common approach, published in (Jaitin et al., 2014) for MARS-seq, does a very good job of detecting and removing molecule errors in InDrop (due to the similar CEL-seq protocol used in both technologies). This approach deletes any UMI for which a higher-abundance donor UMIs can be identified that (1) lie within a single base error (2) have higher count (3) and contain all observed alignment positions of the recipient RNA. This results in removal of approximately 20% of observed UMIs. However, we observed that this model can be overly stringent, correcting UMIs when the donor molecule has as few as one read count higher than the recipient (Figure 2.9 A).

We apply a modified version of the (Jaitin et al., 2014) approach, where we replace errors with corrected barcodes instead of deleting them, and where we only eliminate errors when we have adequate statistical evidence (Figure 2.9 B). To accomplish this, we utilize the spacer and cell-barcode whitelist to empirically estimate a per-base error UMI error rate of approximately 0.2% per base. E.g. to calculate the probability of a conversion from adenosine to cytosine, where $A \rightarrow C$ denotes this nucleotide conversion:

$$P_{A\rightarrow C} \ = \frac{1}{n \cdot m} \left\{ 1 \ if \ x_{ij \ : \ A\rightarrow C} \ else \ 0 \ \right\}$$

where $x_j$is a barcode, j $\in \{1, \ldots, m\}$ and each barcode has n bases. To calculate the probability that a target read was generated in error from a specific donor molecule, we calculated the product of the errors that could potentially convert a donor into the observed molecule. To convert, for example, ACGTACGT into TTGTACGT, having one $A \rightarrow T$and one $C \rightarrow T$ conversion: $e = \{P_{A\rightarrow T}, P_{C\rightarrow T}\}$ Because there are multiple potential donors for each molecule, we calculated the conversion probability for each molecule. Assuming errors are randomly distributed, they can be modeled by a Poisson process, and Poisson rate term can be estimated from the data:

$\lambda \ = \ n_{\text{donor}} \times P_{\text{conversion}}$ where n_{donor} is the number of observations (reads) attributed to the donor molecule in the data. Since the sum of multiple independent Poisson processes is itself Poisson, the rate of conversion from each donor could be combined into a single rate λ_{agg} for each target molecule. The set of conversions s that we consider for each target molecule were all conversions that could occur with two or fewer nucleotide substitutions, in other words, all molecules within a Hamming distance $D_h \ \leq \ 2$, where $D_h$is a matrix of pairwise Hamming distances between barcodes. $s = \ \{\lambda_{j \to i}$ if $D_{h, \ (i,j)} \ \leq 2\}$ $\lambda_{\text{agg}} = \ \lambda_i$ Finally, given the probability of a molecule being observed via the substitution errors that are corrected by the Jaitin method, we could calculate the probability that $n$observations of a specific molecule x were generated via the Poisson process with rate λ_{agg}: $P \ = \ \frac{{\lambda_{\text{agg}}}^{x} e^{-\lambda_{\text{agg}}}}{x!}$

Only cases with a probability $p > 0.05$ were corrected. For InDrop experiments, this resulted in a recovery of an additional 3-5% of molecules in the data that were otherwise error-corrected without adequate evidence. We note that this model is not applicable to all data; It was useful in this instance because we had relatively high coverage (10 reads / molecule) that allowed us to evaluate our confidence in molecule observations. For lower-coverage data containing fewer than 3 observations per molecule, it may be difficult to accurately estimate the Poisson error rates, For such data, it may be appropriate to err towards removing molecules instead of retaining them.

This dynamic suggests a corollary: while spending more reads on each molecule reduces an experiments theoretical yield, since the maximum yield is 1 read : 1 molecule, higher molecular coverage may yield data that more accurately reflects the cellular phenotypes. Thus, datasets that capture more fragments of each molecule

2.14 Disambiguation of Multiply-Aligned Reads Recovers Substantial Sequencing Data

Alignment algorithms like STAR aim to identify the unique portion of the genome that was transcribed to generate the read that is being aligned. In some cases, this unique source cannot be identified, and in these cases multiple possible sources are reported. These are commonly termed "multi-alignments", and because 3' ends of genes have higher homology than other parts of the genome, multi-alignments are more common in 3' sequencing data than in approaches that cover the full-transcriptome, such as Smart-seq2. Despite the increased frequency, most 3' pipelines discard multi-alignments and deal exclusively with unique genes. SEQC is designed to resolve all multiple alignments, producing an output that contains resolved (now unique) alignments, or alignments that are flagged for exclusion because a unique source could not be determined.

There are two main preexisting approaches to resolving multi-alignments. The most common approaches are transcriptomic pseudo-alignment, wherein a sequencing read is broken up into smaller pieces and the pieces are aligned to the transcriptome (Patro et al., 2017; Bray et al., 2016), and expectation-maximization approaches where information that could arise from multiple transcripts is shared across each of the possible sources (Li and Dewey, 2011). However, both methods are too lenient, and propagate errors from InDrop sequencing data into the final count matrix.

Low-coverage 3' sequencing data contains too much uncertainty for expectation-maximization to function properly. RSEM, which was designed for full-length bulk data, passes this uncertainty directly into the count matrix, because it expects the data, in aggregate, to contain enough coverage of each gene for errors to average out. This uncertainty is normally removed by UMI-aware count based methods, and analyses have shown that the inclusion of UMIs significantly improves data accuracy (Grun and Oudenaarden, 2016). As such, RSEM is not an appropriate approach for 3' data.

A second problem is that due to memory constraints, both expectation-maximization and transcriptome pseudo-alignment only consider matches to the transcriptome. This causes a small but significant fraction of reads from genomic sources to be miss-aligned transcriptomic positions (approximately 1%), producing inflated and spurious alignments for low-homology genes.

Of the high-throughput droplet-based approaches, InDrop is unique in combining linear amplification and UMIs, which produces high fragment coverage per UMI. Although individual reads are often ambiguously aligned to more than one location, it is often possible to examine the *set* of fragments assigned to an UMI and to identify a unique gene that is compatible with all the observed fragments. Here we implement an efficient method to find the unique genes that generate each fragment set. When a fragment set cannot be attributed to a specific gene, it is discarded.

Starting with all reads attributed to a cell, we begin by grouping reads according to their UMI, producing "fragment sets" S. Typically, these fragment sets represent trivial problems, such as $s_1 = \{A,\ A,\ AB\}$, a set with two unique alignments to gene A and a third ambiguous alignment to genes A and B. In this case all three observations support the gene A model, while only one observation supports the gene B model.

In cases of UMI collisions, where two mRNA molecules were captured by different primers that happen to share the same UMI sequence, this can lead to problems wherein reads from these merged fragment sets are mistakenly discarded as multi-aligning. However, because the probability of two genes sharing significant homology is low, it is usually possible to recover these molecules by first separating fragment sets into disjoint sets. For example, if a fragment set

s_2 is observed to be associated with an UMI in a single cell, it can be resolved into two disjoint sets, and the second set s_4 can be uniquely assigned to gene E:

$$s_2 = \{A,\ AB,\ B,\ B,\ C,\ CD,\ ABC,\ E,\ EF,\ EF\}$$

$$s_2 = s_3 \cup s_4;\ \ s_3 \cap s_4 = \varnothing, \text{ where :}$$

$$s_3 = \{A,\ AB,\ B,\ B,\ C,\ CD,\ ABC\} \text{ and}$$

$$s_4 = \{E,\ EF,\ EF\}$$

This is biologically reasonable, as molecule collisions are the only way to reasonably obtain a group of molecules that covers two non-overlapping gene annotations. To calculate disjoint sets efficiently, we utilize a Union-Find data structure (Aho and Ullman, 1983), which finds disjoint sets in $O(log(n))$ time. Pseudo-code is as follows:

```
# cell and umi are sequences stored 2-bit encoded in long int
def int count, cell, umi, gene
alignments <- Map[(cell, umi): list[list[gene], count]
alignments <- sorted(alignments, reverse=True)  # inverse numerical sort
for c in cell:
  for u in umi:
    disjoint_sets <- UnionFind(alignments[(c, u)])
    for s in disjoint_sets:
    s <- sort(s, key=len(s))
    alignment[s] = 0
```

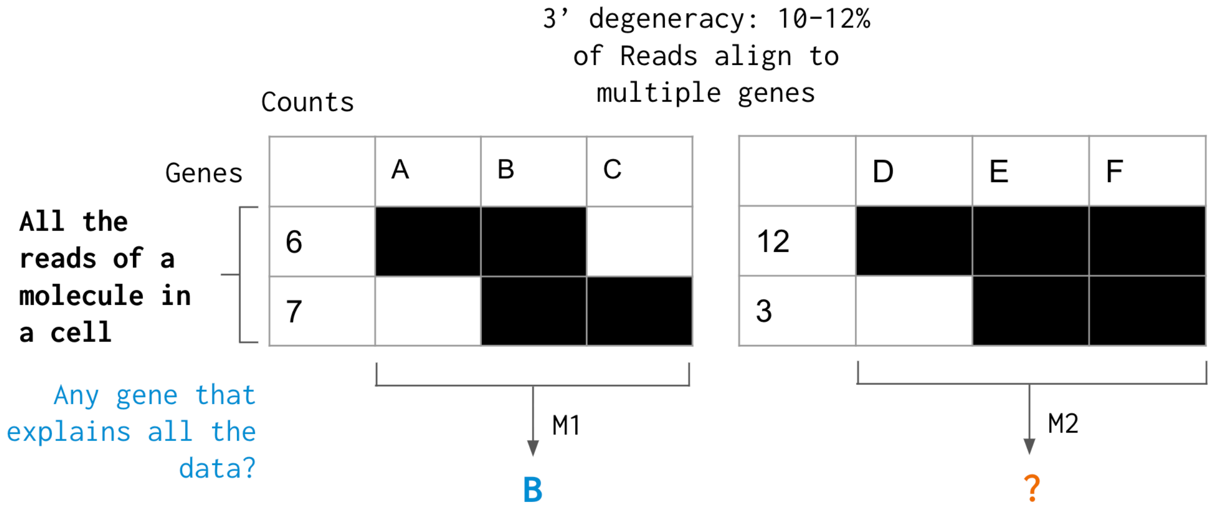


Figure 2.10: A cartoon displaying two examples in which a molecule, comprised of alignments to multiple genes, can be resolved (left) or is ambiguous (right). This figure displays a set of 13 alignments against genes A, B, and C, and an additional 15 against genes D, E, and F. All alignments share the same cell and molecular barcodes. The alignments can first be split, since no alignment bridges A, B, C, and D, E, F (e.g. there is no alignment, for example, to A and D which links the groups). The alignments to A, B, C are then uniquely explained by A and can be resolved, whereas the alignments to D, E, and F are jointly explained by E and F, and are ambiguous.

```
for g in s:
  if g in all s:
    alignment[s] = 1  # mark alignment resolved
```

This algorithm is summarized by Figure 2.10.

By resolving multialignments, we can more accurately identify the alignment rates for each gene, build better error models for barcode correction, and recover cases where reads align multiply to the same gene. More critically, it confers the ability to recover fragments that would otherwise not be resolvable due to sequence homology, and these improved fragment counts per molecules act as significant predictors of molecule likelihood and UMI quality. We note that a similar strategy has since been published (Klein et al., 2015) and a comparable logic underlies the concept of transcript compatibility in Kallisto (Bray et al., 2016). Multi-alignment resolution typically resolves approximately 1M reads per hiseq lane[11]. The result of this module is a BAM or h5 file containing only unique alignments to gene features.

2.15 Cell Selection and Filtering

The preceding sections focus, to the extent possible, on cleaning the data of rational sources of error that have been detected in 3' sequencing data. Once errors have been depleted, the next task is to identify which cell barcodes represent real, high quality cells that warrant biological investigation. There are several potential sources of technical and biological variation that exist in scRNA-seq data that might motivate a researcher to exclude a barcode from analysis.

The most prominent technical source of variation is ambient RNA. Because barcoding beads are loaded into InDrop at higher rates than cells in order to ensure that a high fraction of cells are encapsulated with exactly one bead. As a result, the raw count matrix contains a mixture of barcoded beads that were encapsulated with cells and barcoded beads that were encapsulated alone, but may nevertheless capture some ambient mRNA molecules that float in solution due to premature lysis or cell death in the cell solution. We want to retain barcodes that contain a large number of specific RNA molecules but deplete for cells that are dominated by Ambient RNA.

SEQC accomplishes this by finding the saddle point in the distribution of total molecule counts per barcode and excluding the mode with lower mean. In practice, we accomplish this by constructing the empirical cumulative distribution of cell sizes and finding the minimum of the sec-

[11] We had previously created a model wherein disjoint sets with more than 1 common gene could also be disambiguated by calculating the probability of gene-gene multi-alignments from their homology. This was accomplished by comparing gene sequences using a Suffix Array built from the final 1000 bases of each gene. With this strategy, we could estimate the relative probability that genes were generated from each potential candidate molecule shared across all reads in the fragment set. However, the relative rarity of such events ($< 1\%$ of data) combined with the additional run-time complexity of this method caused us to omit it from the production version of SEQC.

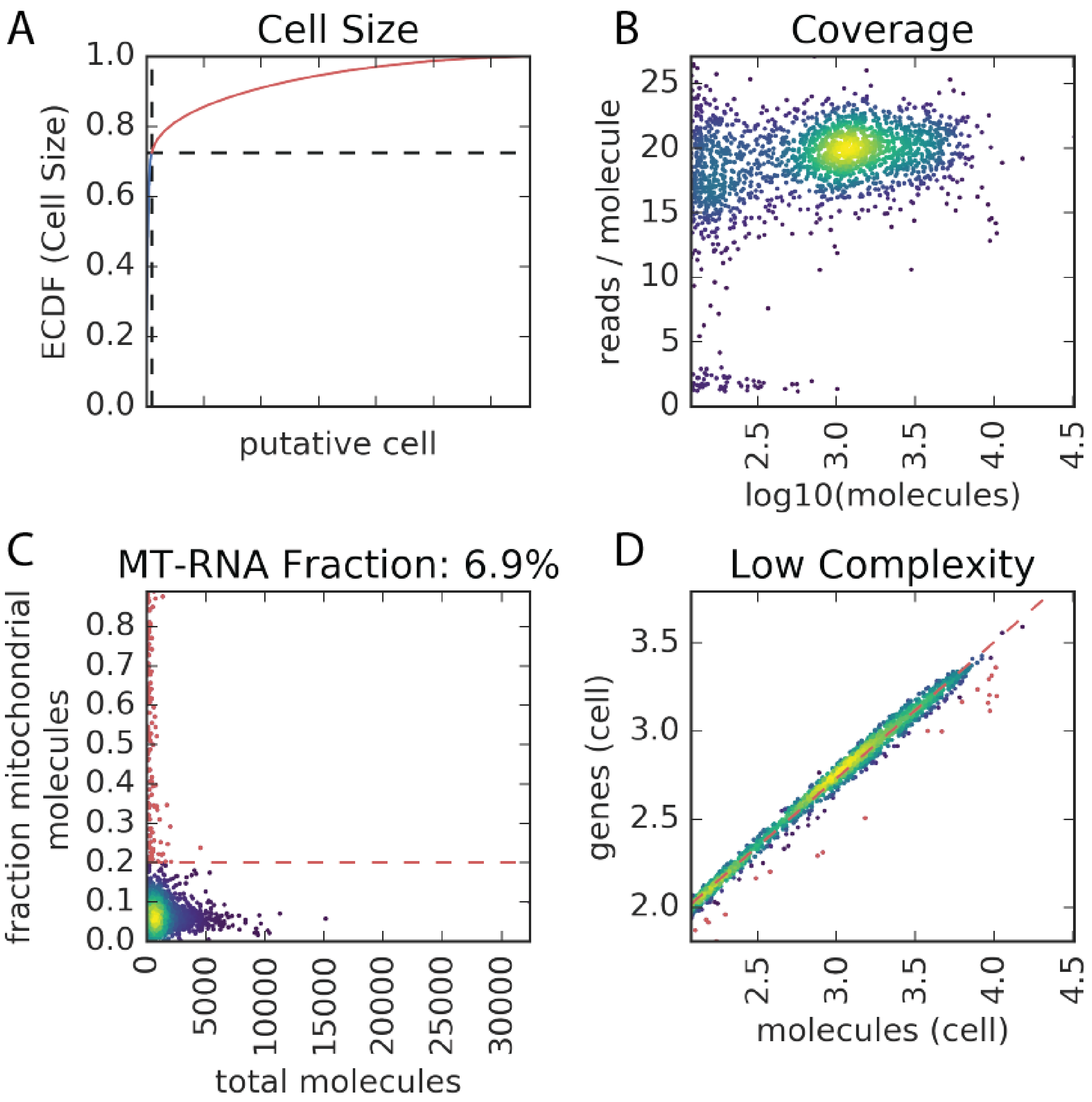


Figure 2.11: (A) Example cell filtering plot showing the empirical cumulative density of molecules (y- axis) per cell barcode (x-axis). Note that a small number of cell barcodes contain most of the molecules in the experiment. Dashed black lines represent cut-off points after which cell barcodes are considered to consist of contamination. Red barcodes are excluded. (B) Coverage plot comparing the total molecules in each cell (x axis) against the average coverage in each cell (y axis). Densities of cells with aberrantly low coverage such as those with lower than 5 reads / molecule are considered likely errors and are discarded. (C) Mitochondrial (MT) RNA fraction plot displaying the total number of molecules in each cell vs the fraction of those molecules that come from mitochondrial sources. Cells in red consist of more than 20% MT-RNA and are considered to be likely dying cells. These cells (red) are discarded. (D) Cell complexity plot. Each point is a cell, and the x axis measures the number of molecules and the y measures the number of genes. Cells with unexpectedly low numbers of genes relative to their molecule count are marked in red and filtered out.

ond derivative (Figure 2.11 A) of the distribution[12]. For typical InDrop runs, this results in the elimination of over 95% of the cell barcodes, but retains as many as 95% of the molecules.

Molecule size alone is not adequate to remove all barcodes that were not paired with real cells. Some barcodes appear to aggregate higher numbers of errors, and as such we often see a bimodal distribution of molecule coverage: a higher mode that represents real cells, and a smaller mode that represents aggregated errors (Figure 2.11 B). We remove the low-count density by fitting 2-component and 1-component Gaussian mixture models to each axis and comparing their relative fits using the Bayesian information criterion. When the 2-component model's log-likelihood is at least 5% larger than the 1-component model, we exclude the density with the smaller mean (Figure 2.11 B).

We score cells for mitochondrial RNA content, which is widely used as a proxy for cell death in scRNA-seq. We observe that a small fraction of cells contain a higher abundance of molecules annotated by this signature, as much as 20–95% of their RNA. Since InDrop does not lyse mitochondria, we reason that these are likely to be cells dying due to stress imposed on them by the InDrop procedure or prior sorting, and remove them from further analysis. This filter may be turned off for studies where apoptosis is a relevant phenotype (Figure 2.11 C).

Finally, we regress the number of genes detected per cell on the number of molecules contained in that cell. We observe that there are sometimes cells whose residuals are significantly negative, indicating a cell which detects many fewer genes than would be expected given its number of molecules. We exclude these cells whose residual genes per cell are more than 3 standard deviations below the mean (Figure 2.11 D).

[12] Recent advances may improve upon this approach by leveraging additional features to build a cell/non-cell classifier that integrates additional information (Petukhov et al., 2017).

To create a digital expression matrix, the uniquely-aligned, error-corrected Hdf5 read store is made non-redundant by counting unique groups of reads with the same UMI, cell barcode, and gene annotation. A single molecule then replaces each set, and those molecules are summed to create a cells x genes matrix. scRNA-seq count matrices are often over 95% sparse, and thus are stored in matrix market format and operated on as coordinate sparse or compressed sparse row matrices. We call these count matrices “raw” count matrices because they contain all barcodes observed in an experiment.

2.16 Information Storage & Run Time

While the scientific quality of data generated by an analysis pipeline is its most important characteristic, the cost and speed of an approach are also important. Faster analysis means faster technological iteration, while lower cost allows for additional data production, which may increase experimental power to answer biological hypotheses. SEQC is optimized with cost and time in mind. scRNA-seq generates large volumes of data whose storage can be costly and onerous, thus we store only aligned, barcode-tagged BAM files which losslessly retain all information from the original multiplex fastq files in small storage space. SEQC supports reprocessing of these files, and backwards conversion into fastq files, if users desire the ability to process their data on other platforms or reprocess with updated versions of SEQC. Additional metadata files take up nominal space, and generated count matrices are stored in matrix market sparse format in light of the sparsity of the data.

SEQC requires approximately 8 hours to run on a standard 32 GB / 16 core Amazon c4.4xlarge, and costs $5.84 on on-demand or $0.88 on pre-emptible (spot) instances to process an InDrop, Drop-seq or 10x genomics experiment. The lower memory usage of 32GB supported by SEQC also makes it much cheaper, easier and more flexible to run on local and remote compute clusters than 10x Cell Ranger, which recommends 128GB RAM and costs twenty times as much on Amazon, given the difficulty of procuring spot instances for high-memory virtual machines[13]. Finally, SEQC is programmed to run on a local machine, on a high performance compute cluster, or on AWS.

2.17 SEQC compares favorably with other pipelines

Near the end of SEQCs development, we ran a head-to-head comparison with a pipeline that had just been open-sourced by the original InDrop computational author `https://github.com/AllonKleinLab/SPRING`. We ran two samples of sorted mouse T-regulatory cells on both pipelines, using a standard index provided by the Klein lab. We were shocked to discover that the Klein pipeline recovered nearly double the number of molecules, suggesting a much higher sensitivity than SEQC (Figure 2.12 A). However, when compared each pipeline to a bulk control sample, we discovered that the Klein pipeline detected many genes in the single-cell data that were not detected in the bulk dataset (Figure 2.12 B, C).

Bulk data has a much larger number of input cells, and as a result, is theorized to proceed with higher efficiency. In addition, because bulk data is full-length instead of 3' localized, it has a greater chance of detecting a larger number of genes. For these reasons, it is unlikely that the genes observed in the single-cell data were true positives. If the single-cell specific observations are removed, the total number of detected molecules is reduced to a comparable, although still higher number.

[13] Replacement of STAR with Kallisto or Hisat2, both methods that are currently being benchmarked, would further reduce the cost of analysis and increase SEQC's speed.

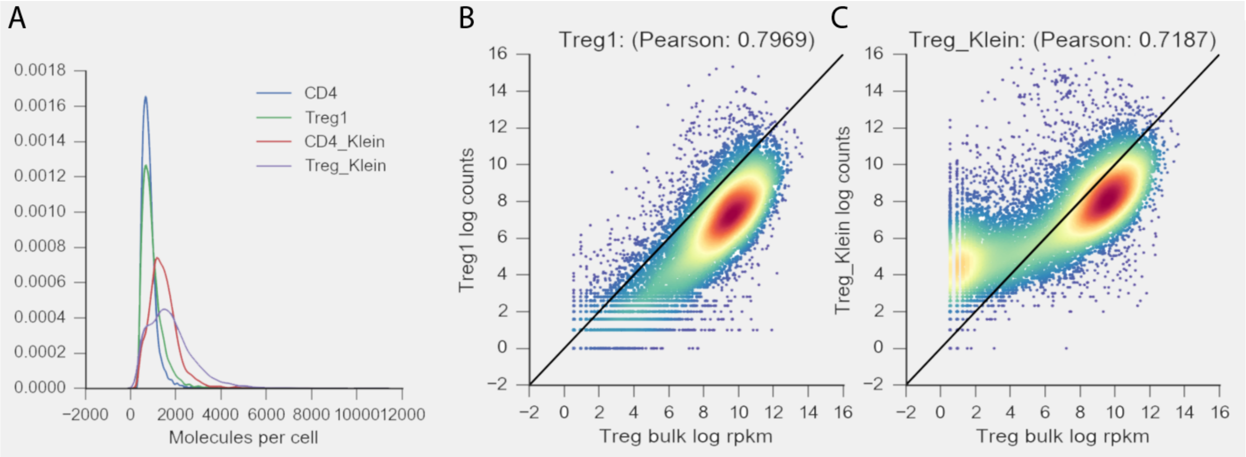


Figure 2.12: (A) Stacked histogram of molecules per cell for two biological samples: CD4+ and T Reg cells, processed using SEQC (blue, green) and the Klein pipeline (red, purple). Comparison of SEQC (B) or Klein pipeline (C) ”pseudo-bulk” sum of gene expression across cells vs. bulk Truseq sequencing of a second aliquot of the same T Reg cells processed from the same biological specimen.

We eventually tracked down the problem in the Klein pipeline, which derived from the analysis of multiply-aligned reads. In cases where a unique alignment could not be identified, the alignment was randomized to any of the “best-match” genes, producing a large increase in spurious molecule and gene observations. As a result of this analysis, this problem in the Klein pipeline was rectified. However, it demonstrates the importance of attention to detail in single cell analysis and the impact that small computational changes can have on data quality. Finally, it demonstrates that in scRNA-seq experiments, it is more important to identify the *right* molecules, than simply the largest number.

2.18 SEQC Identifies Biological Structure by Removing Noise from scRNA-seq Data

Each algorithm discussed above improves the quality of the data, either by removing erroneous reads, correcting errors, removing ambiguity from alignments, restricting alignments to observable features, or removing low complexity aggregations of ambient contamination. How-

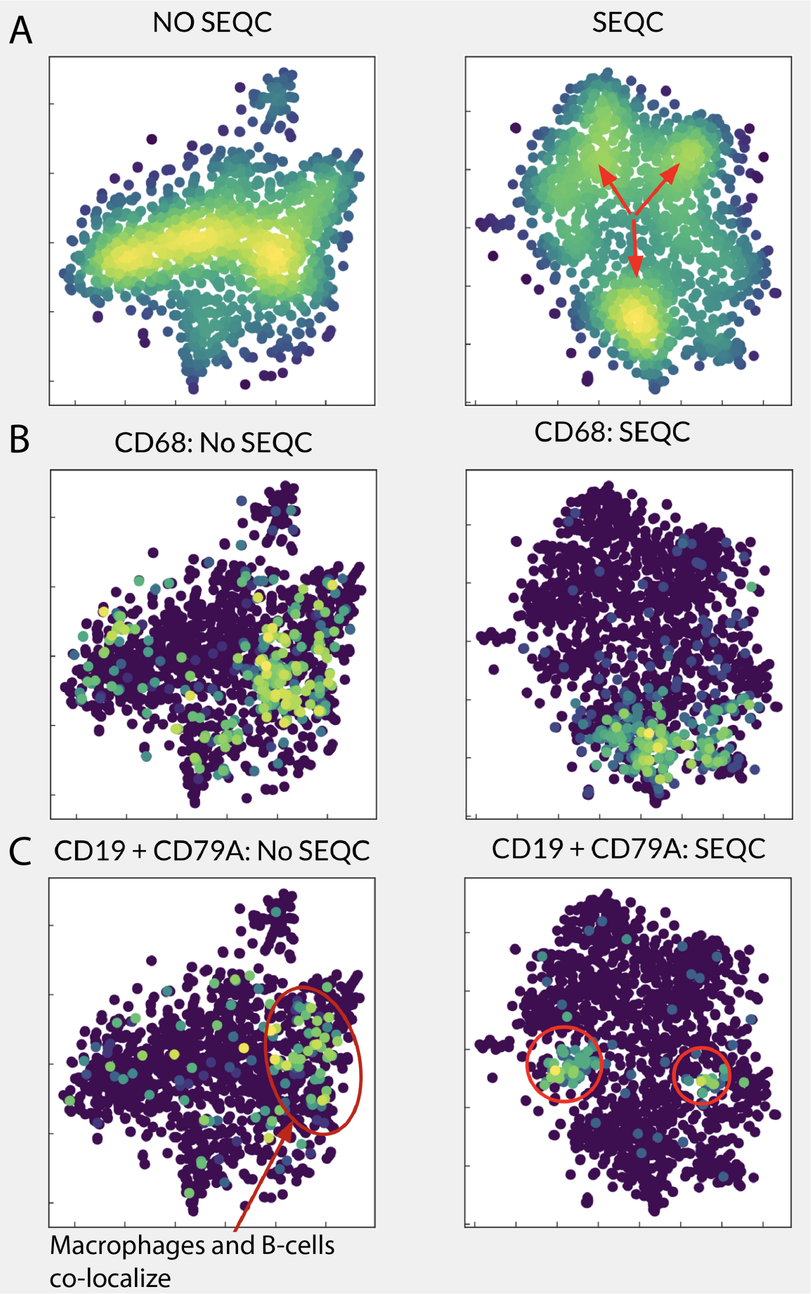


Figure 2.13: All panels display features on tSNE projections constructed from the top 10 principle components from the same set of cells processed with a pipeline constructed from publicly available tools (left) and SEQC (right). Each point represents a cell. Panel A displays a density projection where yellow is higher and blue is lower. Panel B displays the expression of CD68, a marker for macrophage cells. Panel C displays the sum of CD19 and CD79A expression which marks B-cells.

ever, I have not yet demonstrated how critical these methods were to enabling biological reasoning over complex input data from clinical samples.

To display the impact of these methods, I compare a single pilot tumor sample analyzed with SEQC with a pipeline constructed from publicly available components. This pipeline does not contain SEQC's filters, uses the STAR aligner with a standard GENCODE annotation, applies the Jaitin error correction method, requires unique alignments, and selects cells using SEQC approach, so that the same cells could be compared.

These data were then median-normalized, dimensionality reduced with PCA, and projected with tSNE. Examination of the density projections reveals that the comparison pipeline identifies a single high-density region and two smaller densities (Figure 2.13 A, left). In contrast, SEQC recovers a much more structured data projection, consistent with expectations that multiple cell types would be discovered in a clinical immune sample (Figure 2.13 A, right).

To avoid conflating the output of SEQC with clever algorithms for visualizing or clustering data, I will simply display marker genes that are known to identify specific populations to test each pipeline's ability to resolve cell types. Examining the expression of CD68, a macrophage marker, and CD19 and CD79A, B cell markers, reveals that the locations of B-cells somewhat disordered by the comparison pipeline, with the largest densities for both populations found in the middle right. This cell type co-localization suggests that given the comparison pipeline's data, standard analysis algorithms like PCA and tSNE are unable to differentiate between these very different cell types. This implies that non-biological signal may be dominating the expression matrices of the comparison pipeline.

In contrast, SEQC identifies exactly what we would wish to see: three well circumscribed and separated populations: one macrophage and two B cell (Figure 2.13 B, C). This highlights the importance of rational algorithm design to remove errors from data prior to exploratory data analysis, and highlights the importance of good data processing approaches. Without SEQC, it would not have been possible to analyze patient data from human tissue.

2.19 Conclusion

SEQC addresses the most critical data quality problems with single-cell sequencing. It corrects errors introduced by enzymes during library construction, by filtering fractured barcodes and correcting barcode substitution errors. It provides high-quality alignment by limiting read annotation to gene features that are detectable by scRNA-seq, and resolves multi-alignments by aggregating data at the molecule level. However, when aligning, SEQC uses the genome to filter out contamination from genomic sources, an approach overlooked by other pipelines. SEQC then aggregates the cleaned sequencing reads, producing a count matrix of genes x cells which is carefully examined and depleted of cells that display biological or technical hallmarks of low quality. With these approaches, SEQC provides high-quality data faster, or at lower cost, than contemporaneously developed, data-specific pipelines, and compares favorably with them scientifically.

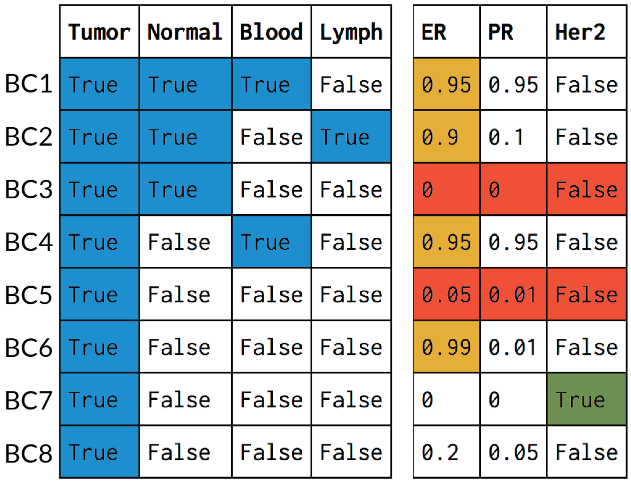
SEQC also provides the first data-type agnostic platform for the analysis of single-cell sequencing data. SEQC is fully open source, and completely modular, allowing us to rapidly test methods from other laboratories that may improve upon our initial computational approaches. As a result, we expect to be able to maintain SEQC as a high-quality analysis tool for scRNA-seq data for some time.

To encourage user adoption, we constructed ready-to-run docker and AWS installations of SEQC, allowing it to be used, without requiring configuration or installation, on any operating system or cloud provider. These characteristics, combined with its low cost and high reliability, make it an important addition to the field of single-cell sequencing. Additionally, these characteristics caused SEQC to be selected for use as the 3' sequencing prototype for the Human Cell Atlas, a recently launched project that aims to process what will likely amount to more than 1 billion human cells.

This chapter began by stating that the critical event that enabled the tumor atlas project was the publication of new droplet-based sequencing methods. As such, an important characteristic of SEQC or any analysis pipeline is that it enable and encourage technological development. SEQC has enabled our group to rapidly iterate on the InDrop and other technologies.

We are able to return fully analyzed sequencing experiments including a complete clustering and QC analysis of a sample to the biologist on the same day that the sample is submitted for sequencing (MiSeq) or within 8 hours of the completion of fastq generation (HiSeq). This rapid turn-around has enabled us to produce 4 versions of the InDrop chemistry, each improving upon the previous method by reducing the number of unnecessary bases that are sequenced, thus saving on cost, improving the barcode libraries, thus increasing data quality, and experiment with contemporaneous enrichment of target genes with paired full-transcriptome sequencing in the same cells, improving scRNA-seq's power to detect rare but important transcripts.

SEQC has allowed us to compare and contrast InDrop with other technologies. It has been used to compare over 10 different chemistries, including, most recently, the processing of nucleus-sequencing data. In addition, when technical disparities between single-cell approaches take up less dominant fractions of data variation, it may enable us to compare or pool data across experiments done by other labs using other chemistries—a feat not yet attempted, to our knowledge. SEQC is currently the data analysis platform used in Memorial Sloan Kettering Institute's



	Tumor	Normal	Blood	Lymph	ER	PR	Her2
BC1	True	True	True	False	0.95	0.95	False
BC2	True	True	False	True	0.9	0.1	False
BC3	True	True	False	False	0	0	False
BC4	True	False	True	False	0.95	0.95	False
BC5	True	False	False	False	0.05	0.01	False
BC6	True	False	False	False	0.99	0.01	False
BC7	True	False	False	False	0	0	True
BC8	True	False	False	False	0.2	0.05	False

Figure 2.14: Summary of samples obtained and patient metadata. Tumor, Normal, Blood and Lymph describe whether or not tissue of each type was extracted from each patient. ER, PR, and Her2 summarize the fraction of a tumor that stained positive for the ER and PR, and whether or not the Her2 gene was amplified.

single-cell data processing platform, and to date has processed over 250 individual sequencing experiments, resulting in 6 publications.

Finally, SEQC enabled us to build a high-quality atlas of the cellular phenotypes of tumor infiltrating breast leukocytes. We chose breast over other cancer models because patients suffering from breast cancer often elect to undergo bilateral mastectomies. This confers a rare opportunity to sequence truly matched healthy tissue, devoid of inflammation effects or pre-neoplastic aberrations which are often found in tumor-adjacent healthy tissue, the standard for tissue matching in other cancer types. To ensure that we recovered a variety of immune cell states responding to variable microenvironments, we included patients with breast tumors of varying type and grade, and devoted the majority of our sequencing to TILs (8 patients). In order to determine what phenotypic differences could be accounted for by tissue residence, we took matching normal (3) and peripheral blood mononuclear cells (PBMCs) (2) samples from the same patients when those patients elected to undergo prophylactic mastectomies (Figure 2.14). We also extracted one involved lymph node to determine how TILs might act in a metastatic context.

These samples were profiled by 61 sequencing experiments with at least two technical replicates per sample and produced over 100,000 cells, each of which was covered by an average of 22,000 reads. Cells contained on average 15 reads per molecule, and cell saturation was 91% across all samples and replicates. After running SEQC on each sample, filtering for complexity, stress responses, apoptosis, low transcript abundance, low gene detection, and non-leukocyte cell types, we retained over 47,000 high quality cells which can be interrogated about the tissue or environmental stimuli that shape their expression profiles. When aggregated by replicates, each group displayed high within-sample correlations (min$r^2 = 0.92, \mu = 0.97, \sigma = 0.02$) and significant between-sample variability ($\mu = 0.72$) (Figure 2.15).

These results suggest that SEQC recovers high quality, low-noise data, and also that there is significant variability between our samples, most of which is biological. The next chapter will discuss how samples from diverse patients, cleaned of technical noise by SEQC, was integrated into a single, cohesive atlas that we used to form biological hypotheses about tumor immunology.

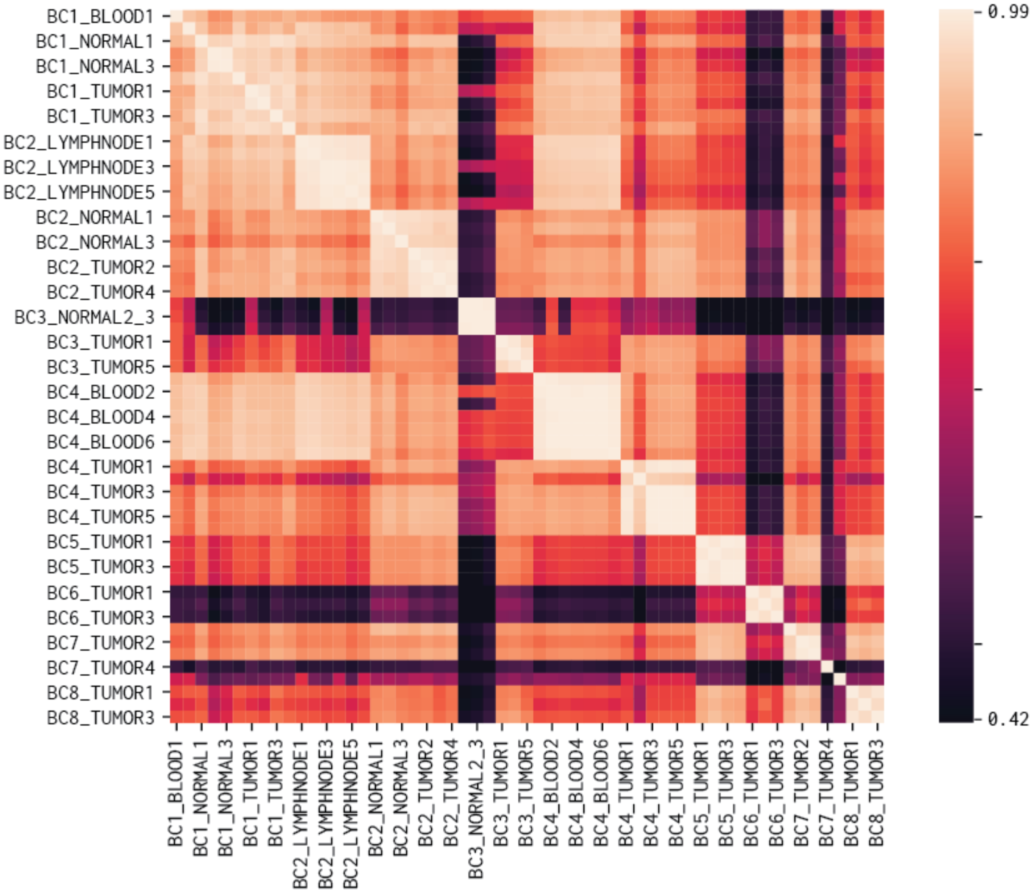


Figure 2.15: Heatmap of pairwise pseudo-bulk sample-sample correlations (r^2) across all samples and replicates in the experiment.

Chapter 3

Algorithms for Analysis of Multi-Patient scRNA-seq Experiments

3.1 Introduction

The previous chapter describes an effort to generate a deep transcriptional map of immune cell states in human breast cancer. Using fluorescence-assisted cell-sorting and single-cell sequencing, we constructed an atlas of the tumor immune ecosystem microenvironment comprising 47,016 CD45$^+$ cells collected from 8 primary breast carcinomas from treatment naive patients. The extracted tumors had multiple types, including estrogen receptor (ER$^+$) and progesterone receptor (PR$^+$) positive, human epidermal growth factor receptor 2 amplified (Her2$^+$), and triple negative (TNBC) cancers. Through careful modeling of error sources and extensive data filtering, we confirmed that variation in the observed cells stems primarily from biological, and not technical, factors.

At the time, this dataset was the largest that had been generated using InDrop, and the only

InDrop dataset generated from multiple human patients. Our experimental design was substantially more ambitious than the datasets preceding it, which had assayed well characterized model systems such as induced pluripotent stem cell differentiation (Klein et al., 2015) or retinal cells (Macosko et al., 2015) from mice with identical genetic backgrounds and growth conditions[1]. Our experiment also included multiple genetic backgrounds, tumor types, tissues, and cell types. Consequently, the analysis of scRNA-seq data subject to numerous and varying stimuli was a significant and unanswered challenge.

This chapter describes efforts to address this challenge. While the last chapter demonstrated approaches to verify the technical quality of the data, this chapter begins by verifying its biological quality by performing some sanity checks on the data using established statistical approaches. Specifically, it will analyze each tumor sample independently to confirm that each of the cell types that are expected to be present in TIL isolates are recovered in each patient in the proportions indicated by FACS. The second part of this chapter describes an approach to merge data from different patients through iterative normalization and clustering.

3.2 Individual Tumor Samples Capture Complete Human Immune Systems

To characterize the immune cells extracted from patients, we began by analyzing samples independently to identify their cellular composition and cell type abundances (Figure 3.1,S3.3). We reasoned that there would be fewer technical effects that influence cells within a sample than across large numbers of samples. Thus, we limited our initial analyses to characterize the cell types within individual tumors.

[1] Other clinical datasets were generated in parallel which examined data from multiple patients(Tirosh et al., 2016), however we are aware of no study that sequenced samples at equivalent depth.

To discover cell types in single cell data, a standard approach is to group cells by similarity into ”clusters” before comparing the average expression profile of each cluster to a previously identified type. The best matching type for each cluster can then be considered a good candidate for the type of the cluster. There are numerous ways of accomplishing clustering, and there is no clear best approach. However, PhenoGraph (Levine et al., 2015), a method that was originally developed to cluster Cytof data, has been adapted to scRNA-seq and appears to have gained community support as the best-practice for clustering cells in a single sample Shekhar et al., 2016; Butler and Satija, 2017.

For a distance between two cells to be accurately quantified, cells must first be transformed to approximate “independent and identitically distributed” data, meaning, practically, that each observed cell expression profile has an equal chance of observing a molecule if it were present in the physical cell. This is manifestly untrue for our data, as the vastly different molecule counts achieved for different cells (and different cell types!) obliterate the “identically distributed” requirement.

The standard approach to address this problem is median library size normalization[2], an approach inspired by bulk RNA-seq approaches (Robinson and Oshlack, 2010). Median library size normalization is a linear scaling approach that sets the total number of molecules in a library to median molecule sum across all observed cells. Since the earlier pseudo-bulk analysis of technical replicates showed that the replicates were highly correlated ($\bar{r^2} = 0.96$,Figure 2.15), we combined the replicates and normalized them together. If each replicate $X_i \in X_1, X_2, \dots X_m$ is

[2] In the case of droplet-based single-cell data “normalization” is a misnomer, as the data, when independently sampled, is better approximated by a negative binomial distribution. However, transformation to the same scale is the important part, rather than the distribution of the data, and so this “normalization” method is an appropriate one to this task.

composed of N cells, and all replicates measure the same set of P genes, then samples N_i can be scaled according to the median ”library size” or median of the total number of molecules in each cell, m, by:

$$m = \operatorname*{median}_{\sum_{p=1}^{P} X_{i,p}} N_i = m * \frac{X_i}{\sum_{p=1}^{P} X_{i,p}}$$

This transformation allows samples which receive different depths of sequencing (different numbers of molecules), to be examined as if they lie on the same scale. If the observed differences in distribution resulted only from differences in sampling rates between cells, then this transformation is adequate to produce data that are *approximately* identically distributed[3].

The transformed data, while statistically suitable for distance calculations, contains over 20,000 genes (described computationally as “features” or “dimensions”), which produces a significant computational burden, as many algorithms scale slowly with increases in features. However, transcription is controlled through the binding of transcription factors to DNA, each of which control many genes. As a result, cells responses’ to stimuli tend to simultaneously modify or “co-regulate” the expression of modules of genes.

The modularity of transcription implies that high-order correlation structures exist in the gene features of our cells, and as a result, that the data actually lie on relatively low-dimension manifolds within the space of observed features. As a result, it is possible to reduce the dimensionality of biological data by collapsing it into correlated components without significant loss of information (Segal et al., 2004; Hartwell et al., 1999).

To identify a low dimensional representation of the data, we apply Randomized Principal

[3] Median library size normalization does not, however, provide a solution for normalizing heterogeneous data from multiple cell types, or enabling comparisons between cells with extremely different sampling rates(Anders and Huber, 2010)

Component Analysis (rPCA) (Halko, Martinsson, and Tropp, 2009; Rokhlin, Szlam, and Tygert, 2009) prior to carrying out subsequent algorithmic steps. The normal PCA method uses singular value decomposition of the data covariance matrix to identify a number of orthogonal components of variation equal to the number of genes in the data. A user then examines the fraction of the original variance explained by each component, and selects some number of components K, to retain. This is typically done either by selecting components until a certain fraction of the total variation is retained (often 75, 95, or 99%) or by searching for a "knee point", beyond which each components explained variation drops off precipitously. Randomized PCA differs accelerates the PCA method bootstrapping over several decompositions of low-rank approximations (with dimension M) of the full data covariance matrix, which has dimension P, the number of gene features with non-zero observations in at least one cell ($M << P$) (Halko, Martinsson, and Tropp, 2009). Randomization can introduce error into low-variation component estimates ($M_i, i > 1000$), however, in scRNA-seq data, most variation is compressed into fewer than 30 components, which rPCA estimates with high accuracy[4]. Thus, use of this algorithm produces a large improvement in speed in what is typically the slowest step of single-sample analysis.

A second, and less appreciated, benefit of using PCA to reduce data dimensionality is that it depletes random variation from the data. By grouping coherent variation into the largest principal components ($k_i, i <= K$), a large proportion of discordant or random variation is pushed into components $k_j, j > K$, which are excluded from analysis. Thus, pre-processing with PCA serves both a practical purpose of reducing computation time but also improves data quality. Because biological data is sparse, and sparse data naturally lie on or close to a low-dimensional manifold,

[4] In practice, M is set equal to approximately $2K$, the number of components that are expected to be retained. For contemporary 3' scRNA-seq data, approximately 25 components are retained, and thus M can be safely set to 50.

this transformation can be achieved without a significant loss of information.

Therefore, for each replicate, we applied rPCA to the normalized data N and selected K using the "knee point" method (Valle, Li, and Joe Qin, 1999), as described above. Because our data was derived from different biological conditions and had different sampling rates, the knee point varied across our samples. We observed that retention of $K = 6 - 11$ Principal components per sample produced optimal results, but some iteration over the subsequent clustering and analysis steps was required to determine the correct value. The final number of retained PCs in each sample correlated with the pre-normalization library size of the samples, as expected ($r^2 = 0.82$).

The dimension-reduced PCA projection was used as the input to PhenoGraph (Levine et al., 2015), which was applied with default parameters (k=30 nearest neighbors). The same principal components were used to generate tSNE projections (Maaten and Hinton, 2008), which were generated with barnes-hut tSNE, implemented in the bhtsne package *https://github.com/lvdmaaten/bhtsne* (Figure 3.1,S3.3).

3.3 Gross Cell Type Annotation

Although our immune atlas consists of considerable variability due to the genetic background of the patient, type of tumor, and the tumor microenvironment, it is nevertheless reasonable to expect that high-quality cell profiles should correlate better with cells of their own lineage than those of other immune lineages. Therefore, we collected, to our knowledge, all previously generated bulk gene expression profiles of sorted immune cells in humans (Novershtern et al., 2011; Jeffrey et al., 2006). These two studies comprised 37 and 32 microarray experiments taken from sorted normal human immune cells and the same cell populations stimulated by bacterial antigens to provoke immune activation.

Consistent with our hypothesis that lineage is a stronger source of variance than immune activation or microenvironment, cells of the same lineage clustered together within the microarray experiments regardless of activation state. Because the bulk data was generated with microarrays, the data is a complex function of library preparation, but also probe capture efficiency. The magnitude of variation in probe capture efficiency is such that microarray analyses are typically limited to making assertions about the relative abundance of genes across samples, rather than measuring the absolute abundance of a gene in a sample (Tusher, Tibshirani, and Chu, 2001). As such, in addition to normalizing by library size, microarray data is additionally translated into units of variance by "Z-scoring" to remove the effect of abundance on downstream computation.

This presents a problem for comparison with single-cell sequencing data, as scale is an important predictor of sample quality. As discussed in the previous section, the ambient RNA contamination in single cell experiments means that low-abundance genes are often the result of non-specific RNA diffusion, whereas high-abundance genes are likely specifically expressed in the cells they are detected in. As a result, Z-scoring scRNA-seq data prior to analysis significantly degrades the quality of the results.

To address this problem, genes were stringently filtered such that they must be expressed at an average of at least 1 count in at least one cluster in order to be considered for comparison with the bulk data. However, for clusters with low RNA expression or capture, such as naive T-cells, a floor of no fewer than 1000 genes was set to guarantee robust comparisons.

While a 1-count threshold may sound lenient, it implies that that on average, each cell in a cluster detected the gene. Thresholding by cluster was important to avoid biasing the comparisons towards genes that were only present in large cell clusters, which would reduce our power to determine the types of rare populations. Both the scRNA-seq and microarray data was then

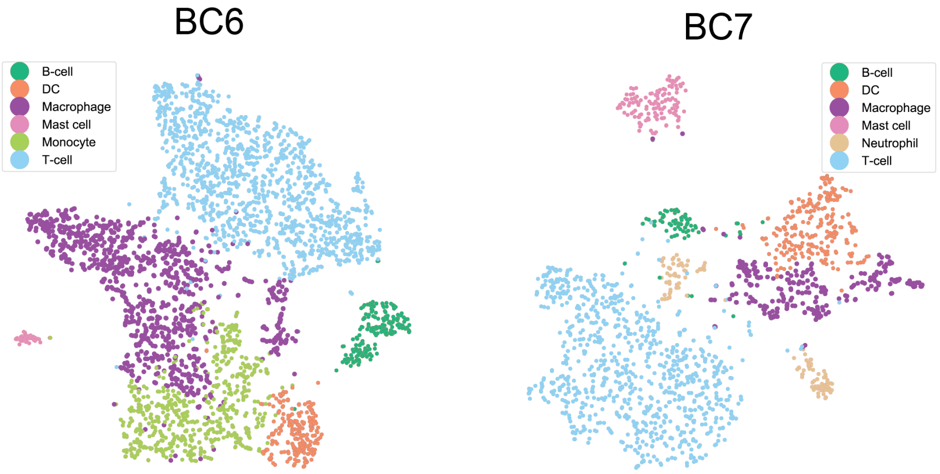


Figure 3.1: t-SNE projection of complete immune systems from two example breast cancer tumors. scRNA-seq data for each tumor is processed with pipeline described in Chapter 2, library size-normalized, PCA-reduced, and clustered with PhenoGraph. Each dot represents a single-cell colored by its cluster label, and clusters are labeled by cell types, inferred through similarity with bulk profiles. Additional tumors are presented in Figure S2

normalized and Z-scored, and the PhenoGraph cluster centroids were correlated with the microarray profiles. The correlations were then averaged across lineages and the highest scoring lineage was used to assign each cluster a type. The types are displayed in (Figure 3.1), and a heatmap of the correlations is demonstrated in Figure 3.11.

Our first attempts to annotate clusters with cell types revealed several clusters with low correlations with all bulk immune datasets. We reasoned that low correlation was most likely due to one of two possibilities. First, low-correlation clusters could be composed of low-quality cells dominated by ribosomal, mitochondrial, or other housekeeping transcripts, which would imply that a cell that should have been filtered out had improperly been retained by SEQC, or it could result because the clusters were not composed of immune cells,. Alternatively, low-correlation clusters could be composed of stromal or tumor cell contamination in our samples, allowed to pass through the sort due to cell autofluorescence or low αCD45 antibody specificity.

To differentiate between these possibilities we examined library sizes of each cluster, rea-

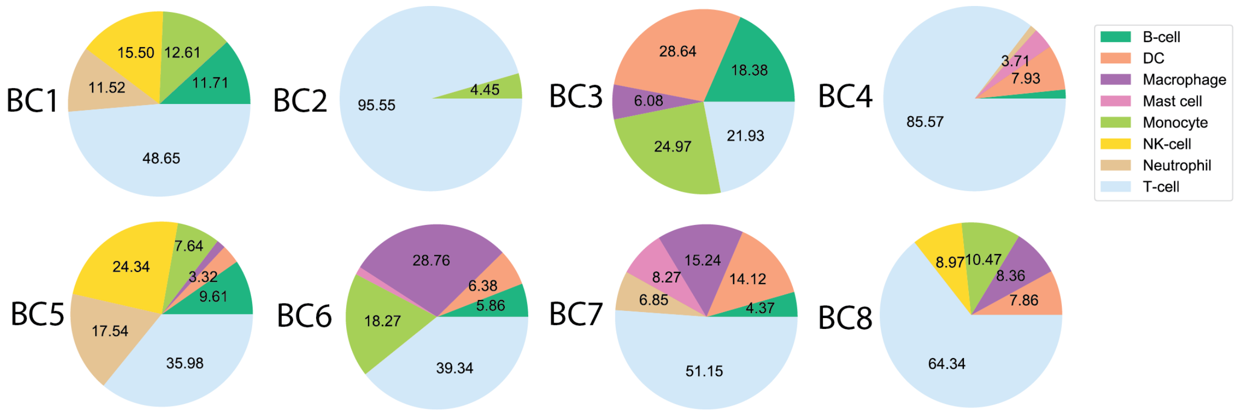


Figure 3.2: Pie charts showing cell type fractions in each patient tumor.

soning that low-quality cells should have smaller libraries than high-quality counterparts. We found no association between library size and quality of immune cell correlation. We next reran the annotation after including several cell lines from epithelial and mesenchymal lineages as negative controls, reasoning that if the cell types were indeed stromal or cancer cells, they should appear closer in phenotype to these populations. This allowed us to identify several populations of fibroblasts, epithelial populations. Post-hoc analyses of the epithelial populations by our surgeon allowed us to further differentiate the epithelial populations into malignant and non-malignant clusters, providing additional support for their separate clustering. Because these populations came primarily from only two patient samples, statistical power would be too low to derive meaningful conclusions about them, and they were excluded from downstream analyses.

3.4 Variation Between Individual Tumor Immune Microenvironments

Having filtered out non-immune types and positively identified immune cells in each patient, we next examined the relative frequency of immune types across tumors. In agreement with the mass cytometry analyses introduced in Chapter 1 (Chevrier et al., 2017; Lavin et al., 2017) and

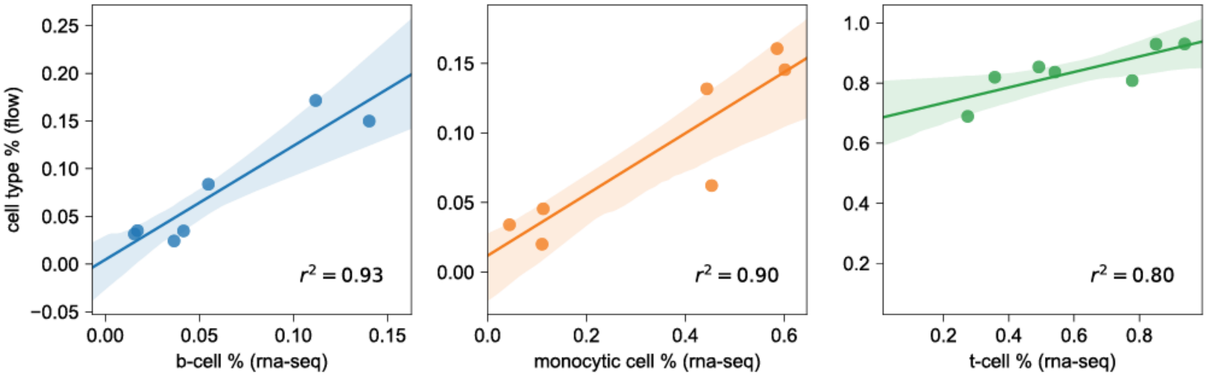


Figure 3.3: Regression of flow cytometry cell type percentages in each patient against RNA-seq cell type percentages for B cells (blue), Monocytic cells (orange), and T cells (green).

prior clinical observations, we found a large degree of variation in the immune cell composition of each tumor (Figure 3.2). For example, the fraction of T cells varied between 21%-96% and the fraction of myeloid cells varied between 4-55%. To determine the reliability of InDrop's sampling of these heterogeneous populations of immune cells, we compared the proportions of cell types as measured by flow cytometry and InDrop scRNA-seq. Although a comparison of the relative representation of major immune cell types identified by scRNA-seq to those measured by FACS revealed a significant bias towards Monocytic lineage cell subsets relative to expected input ratios, we observed high correlation between cell type frequencies across all patient samples ($r^2 > 0.8$, Figure S3.3). The observed bias, likely due to the larger cytoplasmic volume and higher RNA yield of Monocytic/myeloid cells vs. T cells, was systemic and did not adversely affect our analyses. As a result, we concluded that we were able to identify the majority of immune cell types expected to be present in human tumors, including Monocytes, Macrophages, Dendritic cells, T cells, B cells, Mast cells, and Neutrophils (Figure 3.1,S3.3) (Jeffrey et al., 2006; Novershtern et al., 2011). Thus, we were able to capture a comprehensive representation of the immune ecosystem from each individual tumor.

3.5 Integration of Data Across Multiple Tumors

To enable an unbiased systematic comparison across patients, we attempted to merge the data from all tumors to create a map of tumor-infiltrating immune cells. However, we observed that the normalization approaches applied to individual tumors were not adequate for data derived from multiple patients. Cells from the same patient, of different types, were often more similar than cells of the same lineage from another patient (Figure 3.6). Figure 3.6 (left) shows scRNA-seq data from 9K immune cells from 4 breast cancer patients after normalization of cells to median library size, suggesting large differences between patients. Moreover, the tSNE projection did not suggest a diversity of subpopulations beyond two main lymphoid and myeloid lineages. Since biological lineage should, in most cases, produce larger differences in transcript abundances than external signaling from the microenvironment, and we had already confirmed the presence of finer structure within individual patients, we believed that this phenomena was most likely due to technical effects acting across the different InDrop runs. However, we also observed that activated immune cells contained higher numbers of mRNA molecules, a phenomena which has been previously reported (Blackinton and Keene, 2016; Cheadle et al., 2005; Marrack et al., 2000; Singer et al., 2016). Specifically, our analyses showed a gradient of activation of CD8 T cells in tumors as compared to normal- or blood-resident T-cells, where the most pronounced T-cell activation occurred in a TNBC tumor (BC3), which agrees with reports from clinical trials suggesting that TNBC tumors are the most immunogenic (Figure 3.4) (Dushyanthen et al., 2015; García-Teijido et al., 2016). Thus, our data displays technical and biological factors that both influence molecule abundance in individual samples.

The tendency of samples to co-cluster highlights the complexity of analyzing scRNA-seq data

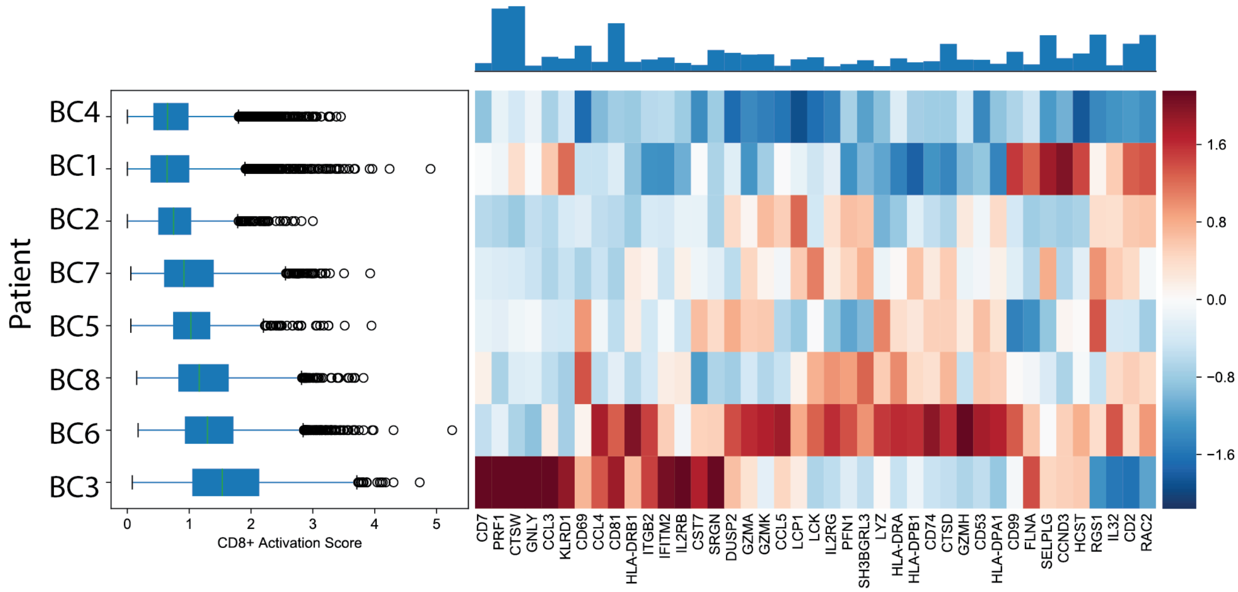


Figure 3.4: Left: Boxplots showing expression of CD8 T cell activation signature (defined as the normalized mean expression of genes in the activation signature listed in Table S4) across immune cells from each patient. Right: heatmap displaying z-scored mean expression of genes in activation signature. Top: Bar plot showing total expression of each gene indicated in the heatmap across all patients. Expression of T cell activation signature shows variability across patients and increased expression in patients BC6 and BC3.

from multiple patients. There are various reasons, both technical and biologically stochastic, that result in co-clustering of samples. First, because the observed data is only one small sample set from the transcriptome of the cell (the full range of mRNAs that the cell expresses to support its phenotype) there is a high chance of missing low-expression genes. In addition, the depth of sampling strongly correlates with the number of features that are observed in each cell, and the sampling depth varies significantly across cells and across samples. The sparsity of scRNA-seq measurements mean that, in our data, the average gene is detected with only a single count. As a result, drop-out is very common, and drop-out is not recoverable by median library size normalization, as any number multiplied by zero remains zero. Thus, in small cells, detected genes are scaled up far more than they should be, producing spurious differential expression

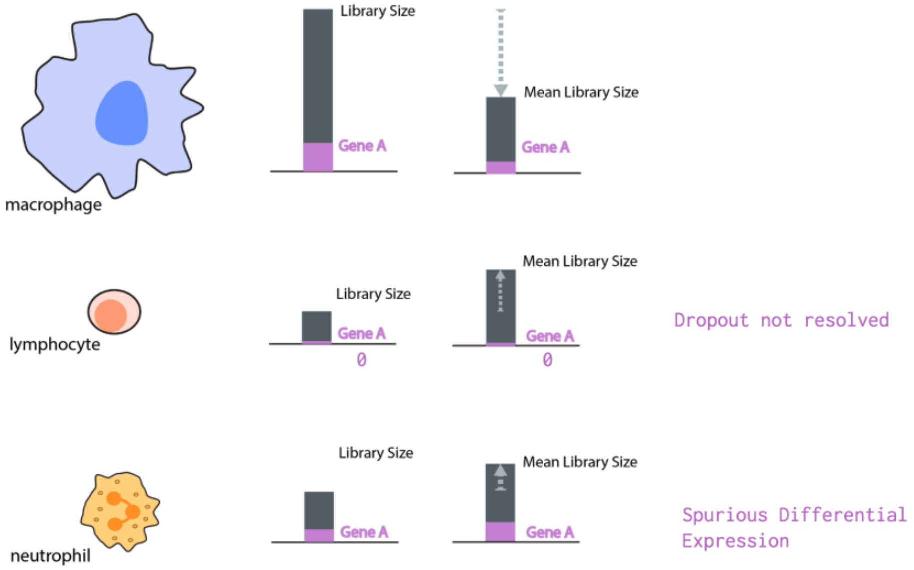


Figure 3.5: Toy example showing the impact of normalizing across cells of different sizes. Large cells, e.g. macrophages, have a larger number of input RNA, and therefore sampling from them produces the most representative single-cell libraries, with minimal drop-out. Smaller cells, such as neutrophils and lymphocytes, are subject to more stochasticity as it is possible that mRNA molecules of a gene are present, but will not be captured. As a result, the disparity between capture/non-capture can produce misleading differential expression results at the single-cell level.

relative to better sampled cells of the same state (Figure 3.5).

Technical factors include differences in sequencing machine, enzyme activity, lysis efficiency or experimental protocol. These samples were also subject to operational variation during the clinical resection, transport, and handling. These factors all impact cell viability, which in turn affects the single cell RNA-seq library preparation, in particular molecular capture rate and sampling. Because molecular capture is a binary event, and the capture rates are very low, these technical variations often determine whether a given gene feature is observed in the data for a given cell.

These more technical artifacts, particularly in capture rate, are confounded with biological differences. This is particularly challenging in the case of immune cells, where activated cells have substantially heightened transcription rates, and therefore if sampling efficiency is constant, we expect to capture more molecules (Blackinton and Keene, 2016; Cheadle et al., 2005; Singer et al., 2016). Therefore, the sampling rate is affected by biological as well as technical processes. Indeed, we see large differences in the number of activated T-cells across patients (Figure 3.4), with more activated T-cells in the Triple Negative subtype as expected (Dushyanthen et al., 2015). Hence, normalizing by library size will likely remove these biological variations.

Both the technical and biological effects tend to average out at the population level, but distance metrics do not share information across cells, as bulk approaches were able to.. As a result, distance metrics tend to be very sensitive to differences in sampling, which can lead to spurious differential expression or removal of biological stochasticity specific to each cell type, both of which induce improper clustering and characterization of latent cell types. Therefore, cell type-specific normalization is especially crucial in experiments involving vast subtype diversity, such as immune cells ranging from large Macrophages to much smaller lymphocytes (Lun, Marioni, and Bach, 2016; Vallejos et al., 2017), wherein the sampling rate contains biological information.

Unfortunately, cell's types are defined by the clusters they fall in, and thus cannot be not determined a priori. Thus, the transformation and clustering of scRNA-seq is a chicken-and-egg problem wherein it is illogical to start from either step.

3.6 Biscuit Clustering and Normalization Corrects for Technical Effects Across Samples

To solve address the convolution of biological and technical effects, we developed and applied the method "Biscuit" (short for Bayesian Inference for Single-cell ClUstering and ImpuTing) to simultaneously cluster cells and normalize according to their assigned clusters (Prabhakaran et

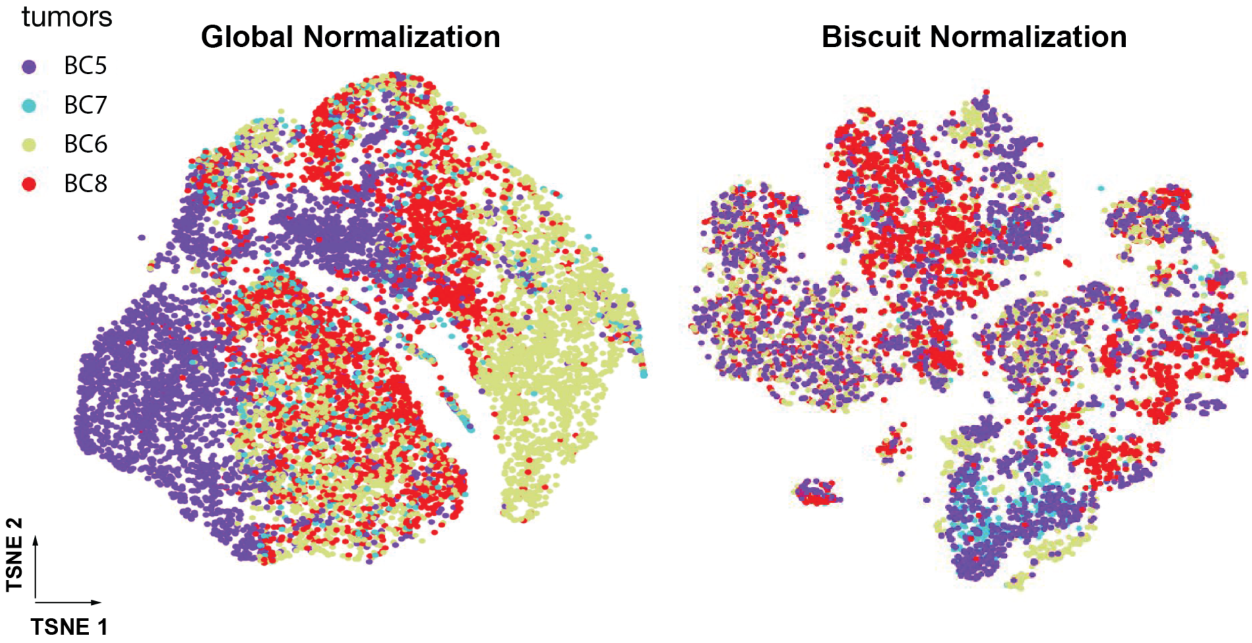


Figure 3.6: T-SNE projection of tumor-infiltrating immune cells from 4 breast cancer patients after library-size normalization (left panel) and Biscuit normalization and imputation (right panel). Cells are colored by tumor (patient). Less mixing of tumors indicates either batch effects or patient-specific cell states.

al., 2016) (Figure 3.7). This is accomplished through incorporating parameters denoting cell-specific technical variation into a Hierarchical Dirichlet Process Mixture Model (HDPMM) (Görür and Rasmussen, 2010) (Figure 3.7 A). This allows for inference of cell clusters based on similarity in gene expression as well as in co-expression patterns, while identifying and accounting for technical variation per cell (Figure 3.7 B, C). Two key ideas that power Biscuit are the use of gene co-expression as a more robust means to identify cell types, and the normalization of each cell type separately to better account for cell type-specific effects on technical variation. The main idea behind the use of co-expression is that cell types not only share similar mean expression, but also share similar co-expression patterns (covariance) between genes. While mean expression can be more sensitive to capture efficiency, covariation is more robust to such effects. This similarity in co-variation can be used to improve normalization and in turn improve the clustering, through the learning of cluster specific parameters.

By jointly performing normalization and clustering, we retain biological heterogeneity and avoid biases that result from independent clustering and normalization, and instead are able to match cells to clusters of the same cell type from different patients which may have very different sampling rates. Figure 3.6(right) shows the same data from 4 tumors after normalization with Biscuit. Note that Biscuit does not use any information on sample IDs in the normalization, and normalization is only driven by cluster assignments. The Biscuit-normalized data shows that the differences in library-size normalized data were largely artifacts of normalization and batch effects. We therefore applied Biscuit to data from all 8 tumors to infer the full diversity of immune cell types in the breast tumors, which identified 67 clusters indicating significant diversity in both lymphoid and myeloid cell types (Figure 3.8).

This transformation also imputes dropouts in each cell by sampling dropped-out genes from the posterior distributions for the cluster that a cell is assigned to. The use of covariance parameter in the model ensures that intra-cluster heterogeneity is preserved after imputing. We show a systematic evaluation of the algorithm performance (on synthetic and real single cell data), its robustness, as well as the ability of this method to impute dropouts in (Prabhakaran et al., 2016).

To formalize and quantify Biscuit's ability to correct batch effects across data from all eight tumors (Figure 3.8) and match immune subtypes across the tumors, we devised an entropy-based metric that quantifies the "mixing" of the normalized data across samples. The entropy-based metric is computed as follows: We constructed a k-NN graph (k=30) on the normalized data using Euclidean distance and computed the distribution of patients (tumors) $m = 1, \ldots, 8$ in the neighborhood of each cell j, denoted as ${q_j}^m$. Then we computed Shannon entropy $H_j = -{q_j}^m$ log q${}_j{}^m$ as a measure of mixing between patients, resulting in one entropy value $H_j$per cell

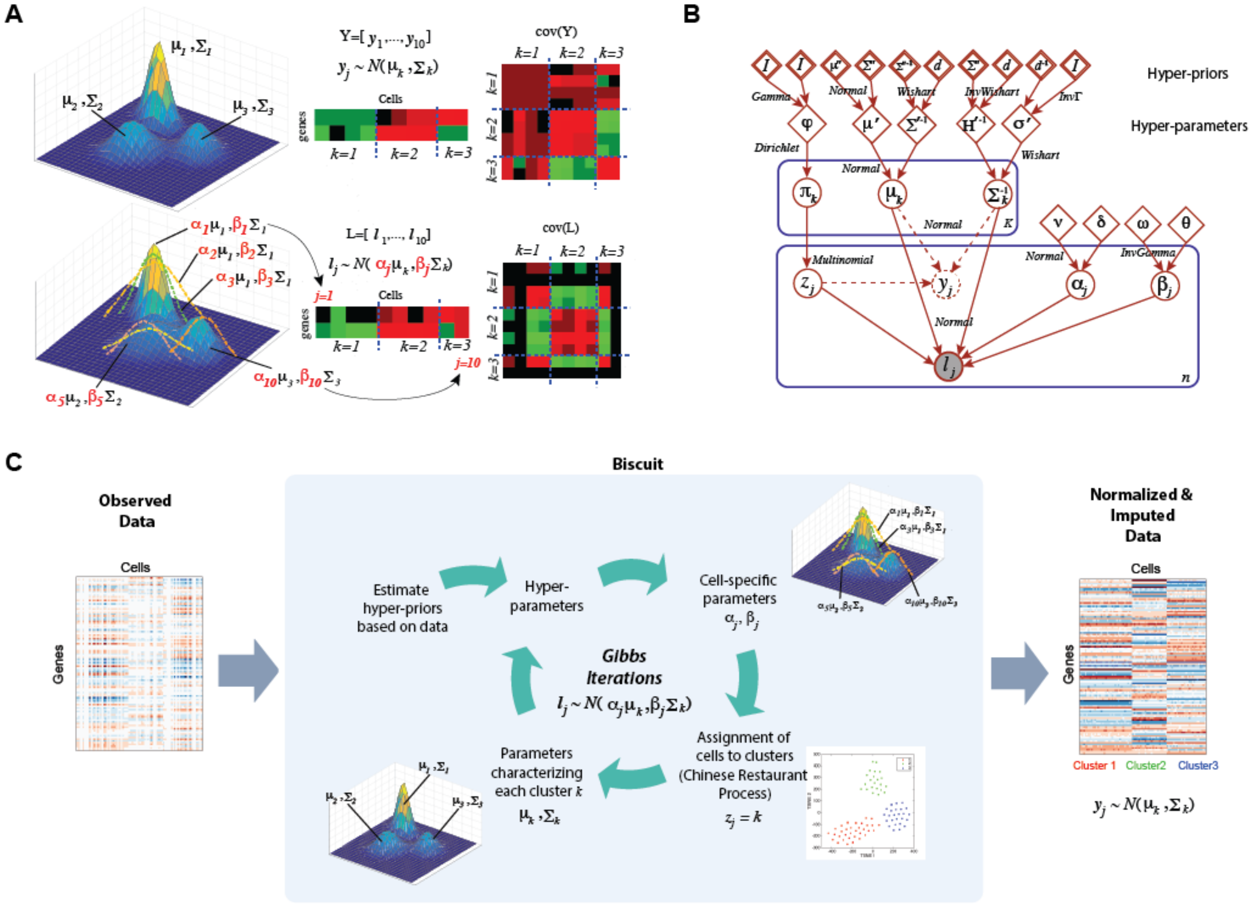


Figure 3.7: (A) Stochastic data generative process for Biscuit illustrated with a toy example. Top panel: Left: shows 3 multivariate Gaussian densities with no technical variation. Middle: An ideal cell (y_j) is simulated as a random draw from any of these 3 Gaussians. Right: The covariance matrix across 10 such randomly-drawn cells showing 3 block covariances across the diagonal corresponding to three clusters. Bottom panel: Left: shows 3 multivariate Gaussian densities with means and covariances scaled using (j, j) to handle cell-specific variations. Middle: A cell (l_j) is simulated as a random draw from any of these 3 scaled Gaussians. Right: The covariance matrix across 10 such randomly-drawn cells showing loss of signal in the 3 block diagonal covariances. We assume the model for l_j captures real single-cell measurements and the goal is to normalize data by converting it to follow the model for y_j. (B). Finite state automata for Biscuit. The shaded circle denotes l_j, which is observed gene expression for cell j, white circles show latent variables of interest, rectangles depict the number of replications at the bottom right corner, diamonds are hyper-parameters, and double diamonds are hyper-priors obtained empirically. Inference equations are obtained by inverting the date generative process. (C) Left panel: Input count matrix to Biscuit. Middle panel: Inference algorithm with Gibbs iterations are depicted where cell-specific (j, j) and cluster-specific (k, k) parameters are iteratively inferred leading to cell assignments to clusters. Right panel: Output from Biscuit, which is the normalized and imputed count matrix.

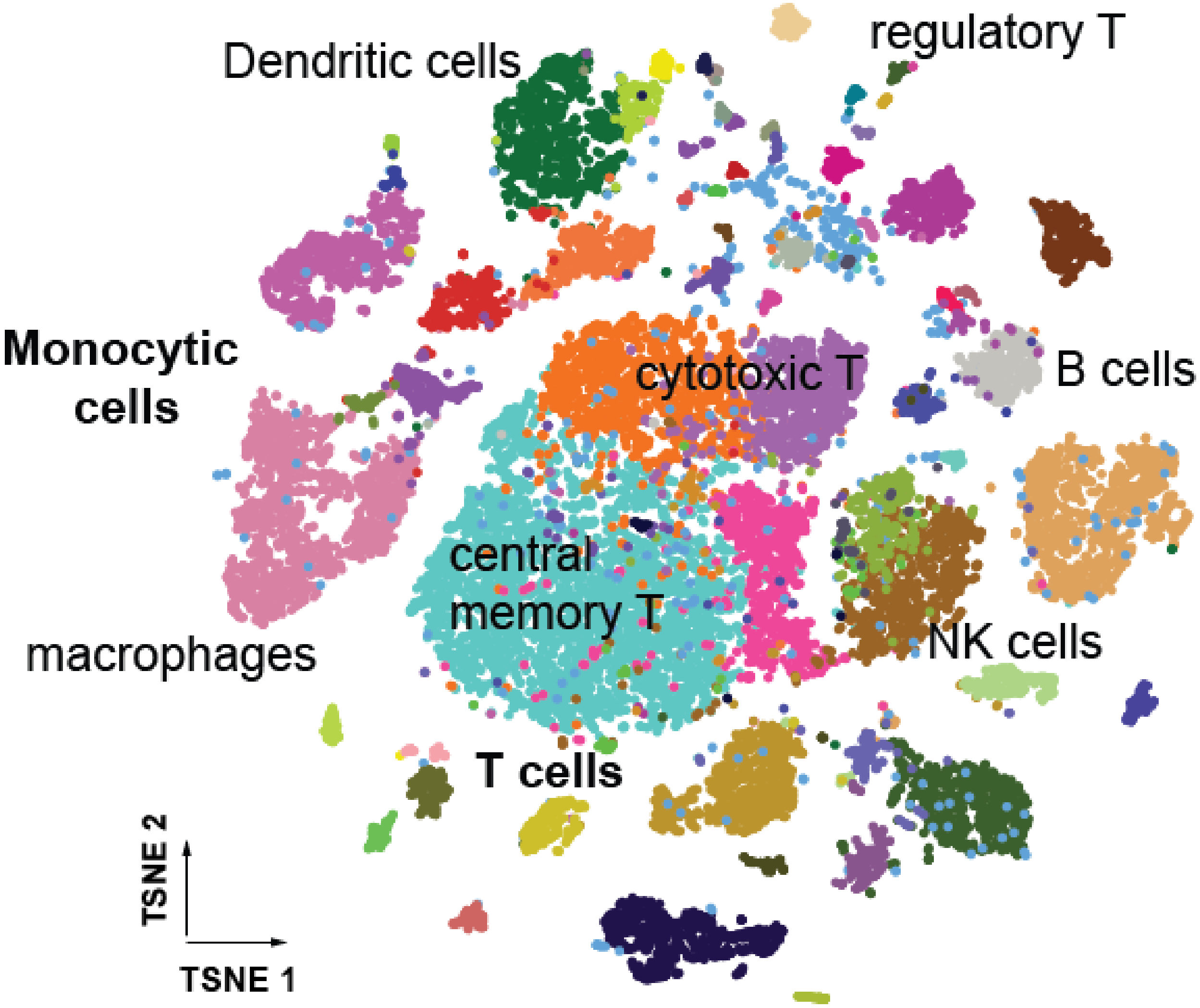


Figure 3.8: T-SNE map of breast tumor-infiltrating immune cells from all 8 patients after Biscuit normalization and imputation showing rich structure and diverse cell types. Cells colored by Biscuit clusters and labeled with inferred cell types.

j. High entropy indicates that the most similar cells come from a well mixed set of additional tumors, whereas low entropy indicates that the most similar cells largely come from the same tumor. Prior to Biscuit, most cells in the data had low entropy values, with 40% of the cells residing in a neighborhood of cells purely from the same tumor. We compare the distribution of entropies across all cells from all 8 tumors, before and after Biscuit, which reveals that the median of entropy shifts significantly towards higher mixing of samples after processing with

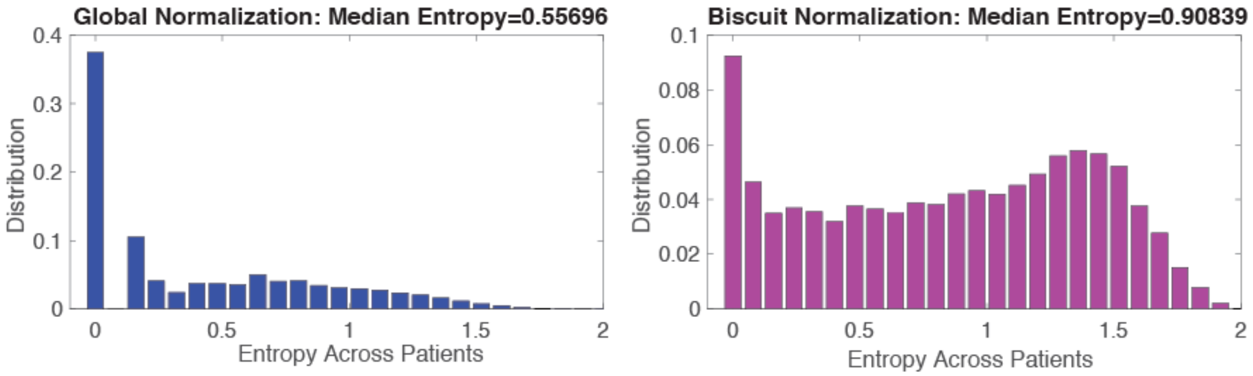


Figure 3.9: Histogram depicting entropy of the patient distribution as a measure of sample mixing. Entropy is computed per cell, based on the distribution of patients in (30-NN) local cell neighborhoods after library-size normalization (left panel) as compared to Biscuit (right panel).

Biscuit (Mann-Whitney U-test: U=1.7721e+09, p=0; Figure 3.9). Thus, we conclude that Biscuit substantially corrected batch effects in this data.

To generate a global atlas of immune cell types, we combined samples from all patients and tissues by applying Biscuit to the full set of $n = 62024$ cells and $d = 14875$ genes, resulting in a global atlas of $K = 95$ clusters (Table S2) in which $n = 57143$ cells had statistically significant cluster assignments. The remainder of cells had low library size and were hence removed from further analysis. A subset of these clusters were identified as probable cancer or stromal populations through correlation with bulk gene expression datasets and marker gene expression. While these non-immune clusters may be of significant interest in their own right, they were beyond the scope of this paper and were therefore excluded from downstream analysis, leaving 47,016 cells in 83 clusters (Table S2).

While biscuit improved mixing across patients overall, We observed that individual clusters displayed differing amounts of mixing between samples (Figure S6). To further quantify the exact degree of mixing (between patients) in each cluster, we defined an entropy-based metric. We used bootstrapping to correct for cluster size (which ranged from over 8900 cells to just over

30 cells), such that we uniformly sampled 100 cells with replacement from each cluster, computed the distribution of patients across these cells, and then computed the Shannon entropy for this distribution. We repeated this procedure 100 times for each cluster, to achieve a range of entropy values per cluster. Figure S9 shows box plots for entropy values for each cluster, with the order of clusters based on their mean entropy. Clusters with entropy of 0 denote entirely patient-specific clusters. Figure S9 shows that there is a continuous range of entropies, and thus a full range of sample specificity versus mixing, across clusters. These results suggest that our experiment succeeded in observing both general immune cell states, and also tumor-specific states that may result from specific microenvironments.

3.7 Cluster Robustness

To evaluate cluster robustness, we performed 10-fold cross-validation, independently clustering and normalizing on random subsets of data. For each of 10 subsets, we ran Biscuit to obtain a set of clusters. To compare the results across the 10 subsets, we computed the confusion matrix, which indicates the probability of each pair of cells j, j' being assigned to the same cluster: $P(z_j = z_{j'})$, where z_j is the j^{th} cell. Figure S7 illustrates box plots for the probabilities of co-clustering (across 10 subsets) for every pair of cells that are assigned to the same cluster in the analysis of the full dataset. The average co-clustering probability in each cluster ranges between 92%-100%, showing remarkable robustness of clusters.

3.8 Distances Between Clusters

The distances between BISCUIT clusters can be directly computed from the posterior probability distributions of each cluster. While Euclidean distances are defined for vectorial objects

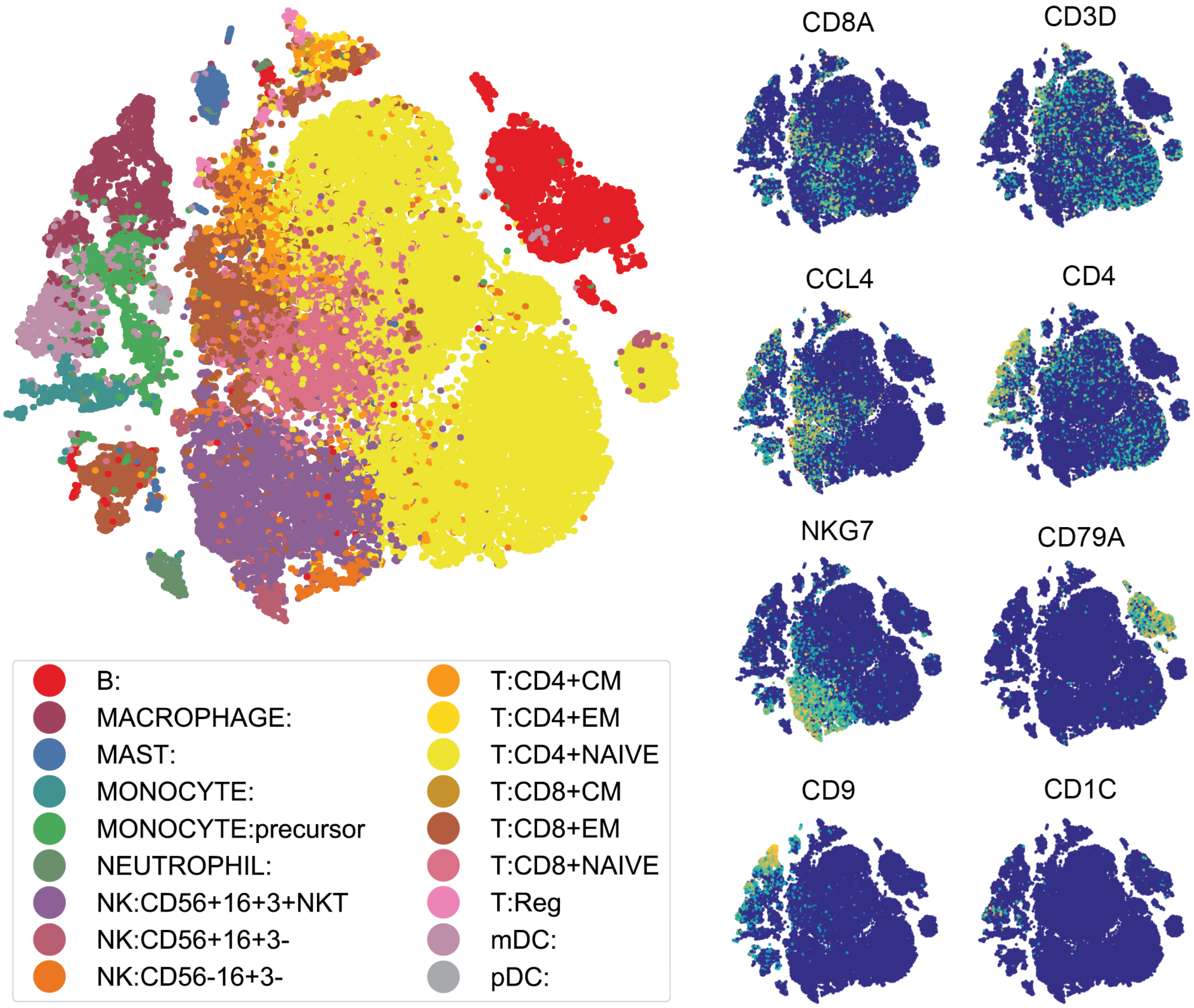


Figure 3.10: T-SNE projection of complete atlas of immune cells, post-Biscuit normalization, from all patients and all tissues including tumor, blood, lymph, and contra-lateral normal tissue,labeled by inferred cell type (left panel) and normalized expression of 8 immune cell markers (right panel). Figure S2 presents further details on inferred clusters with complete annotations in Table S2.

and operate under a Cartesian coordinate system, Euclidean distance with non-vectorial objects such as probability distributions requires embedding them in Euclidean space. Such embeddings are non-unique and lead to loss of information. It is therefore advisable to use the non-vectorial objects as is and to work with the objects' pairwise similarities or distances instead. One such distance metric, which is effective at comparing pairwise probability distributions, is the Bhattacharyya distance (BD) (Bhattacharyya, 1990).

We defined distances between each pair of clusters k, k' with distributions p_k and $p_{k'}$ as $BD = -log(BC(p_k, p_{k'}\ ))$ where BC is the Bhattacharyya coefficient measuring similarity (overlap) of the distributions. We use the BD to compute distances between pairs of inferred clusters' moments to create the Bhattacharyya kernel. The Bhattacharyya kernel has closed forms for any exponential distribution including the (multivariate) Gaussian distribution (Jebara, T., Kondor, R., & Howard, A., 2004), which is Biscuit's underlying data-generation distribution. For the case of multivariate normal distributions: $p_k \sim N(\overrightarrow{\mu_k}, \Sigma_k)$and $p_{k'} \sim N(\overrightarrow{\mu_{k'}}, \Sigma_{k'})$:

$D_B\ = \frac{1}{8}(\overrightarrow{\mu_k} - \overrightarrow{\mu_{k'}})^{\mathrm{T}}\, \Sigma^{-1}(\overrightarrow{\mu_k} - \overrightarrow{\mu_{k'}})\ +\ \frac{1}{2}log(\frac{\det \Sigma}{})$ where $\Sigma = (\Sigma_k + \Sigma_{k'})/2$.

Figure S10 shows the heatmap of pairwise distances between all pairs of clusters.

A geometric interpretation of BD is that, via its cosine formulation, the distance subsumes a full hypersphere and the centre of the hypersphere is the centroid (mean) of the cluster, whereas the Euclidean distance only covers a quarter of the hypersphere with the center at the origin.

3.9 Contribution of Covariance in Defining Clusters

We used the above Bhattacharyya distance (BD) metric to study the contribution of Biscuit's covariance parameters to characterizing clusters of different cell types. First, we computed the BD between pairs of clusters of the same type (T, Monocytic, NK, B) and compared these to distances between pairs of clusters of different cell types (e.g. a T cell cluster and a Monocytic cluster). Figure S2E shows violin plots for distances between pairs of clusters with dots (overlaid on violins) representing cluster pairs; violins are sorted based on median distance. As reference, we also split each cluster into two halves and computed the empirical BD between two splits (shown at the left end in Figure S11). We observe that, overall, pairs of clusters of different types are more distant than pairs of clusters from the same type, as expected.

We then computed these same pairwise distances while removing the contribution of mean parameters for each cluster, via setting $\overrightarrow{\mu_k} - \overrightarrow{\mu_{k'}} = 0$ and computing the distance only based on covariance parameters of the pair of clusters $\Sigma_k, \Sigma_{k'}$ (Figure S11, right). We observed that pairs of T cell clusters or Monocytic clusters still show prominent distances, and therefore covariance parameters have a crucial role in defining these clusters.

In the case of Biscuit clusters (Figure 3.11), mean parameters for each cluster were correlated with bulk profiles. Each of the bulk profiles was marked as having derived from one of several major cell types: b-cells; T-cells (naive, central memory, cytotoxic, T-regulatory); Monocytic cells (monocytes, dendritic cells, macrophages); Mast cells; Neutrophils; or NK-cells. The highest scoring bulk profile for each centroid was used to categorize each cluster by its type, and types were split for downstream analysis.

Cells were also typed by examining expression of known marker genes. In this analysis, cells were scored as detecting a marker gene if the cell contained a non-zero molecule count for that gene. Each cell was corrected for its detection rate (the fraction of total genes detected in that cell) and the marker detection rate was then averaged across cells of a cluster. Markers used to assign classical types cells included NCAM1, NCR1, NKG2 (NK-cells), GNLY, PFN1, GZMA, GZMB, GMZM, GZMH (cytotoxic T-cell, NK), FOXP3, CTLA4, TIGIT, TNFRSF4, LAG3, PDCD1

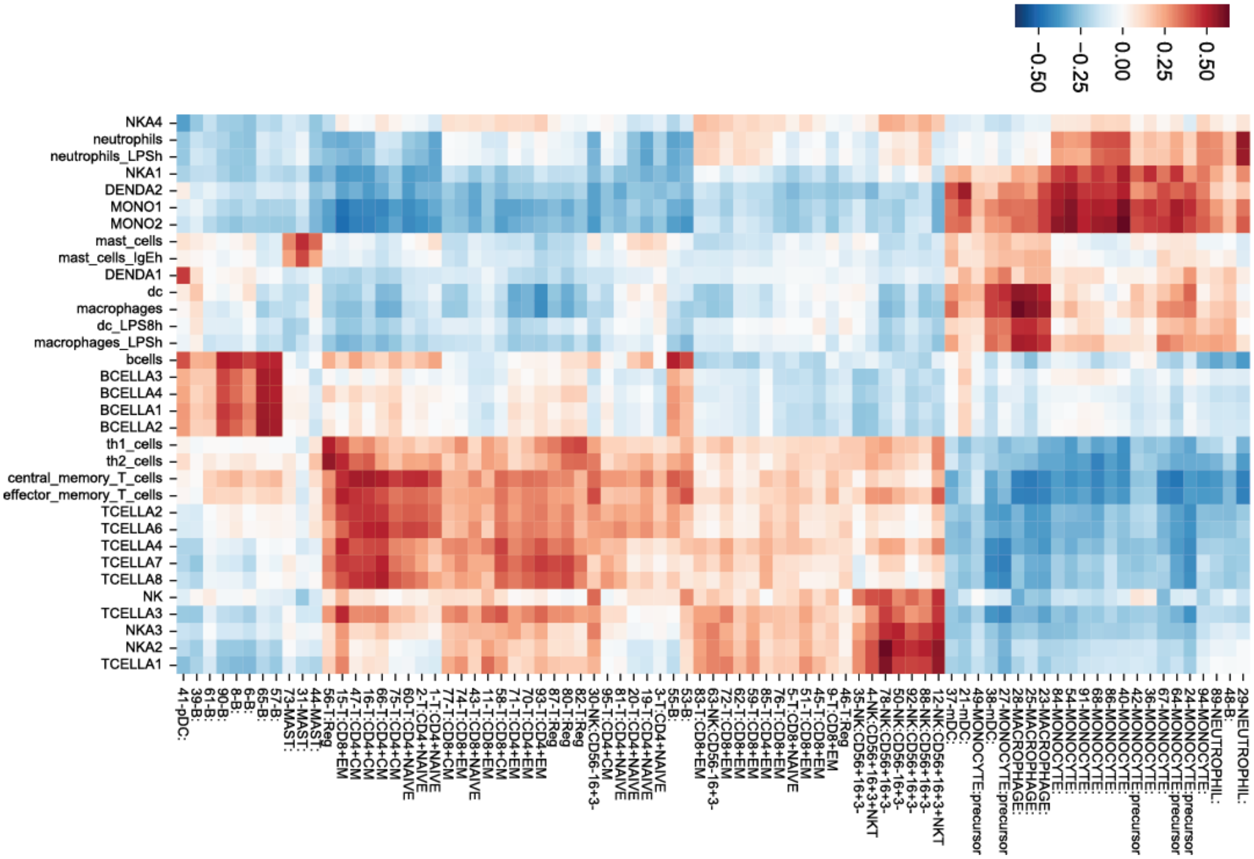


Figure 3.11: Pearson correlations of cluster expression centroids to bulk RNA-seq data from purified immune populations (from Jeffrey 2006 and Novershtern 2011). Scale bar displays r-values. (G) Histogram of frequency of patients contributing to each cluster showing that 19 clusters (out of 95) are present in all 8 patients and 10 clusters are patient-specific.

(exhausted T-cell, T-regulatory cell), CD8, CD3, CD4 (T-cells), IL7R (naive T-cells), CD19 (b-cells), ENPP3, KIT (Mast cells), IL3RA, LILRA4 (Plasmacytoid Dendritic cells), HLA-DR, FCGR3A, CD68, ANPEP, ITGAX, CD14, ITGAM, CD33 (Monocytic lineage). For all retained clusters, the two typing methods agreed (Figure 3.12).

3.10 Gene Signature Summarization

To interpret the observed cell states we made extensive use of gene signature enrichment. However, in addition to testing for heightened expression of genes in the signature across cells in the cluster, we also examined signatures in terms of their variation. Specifically, we examine

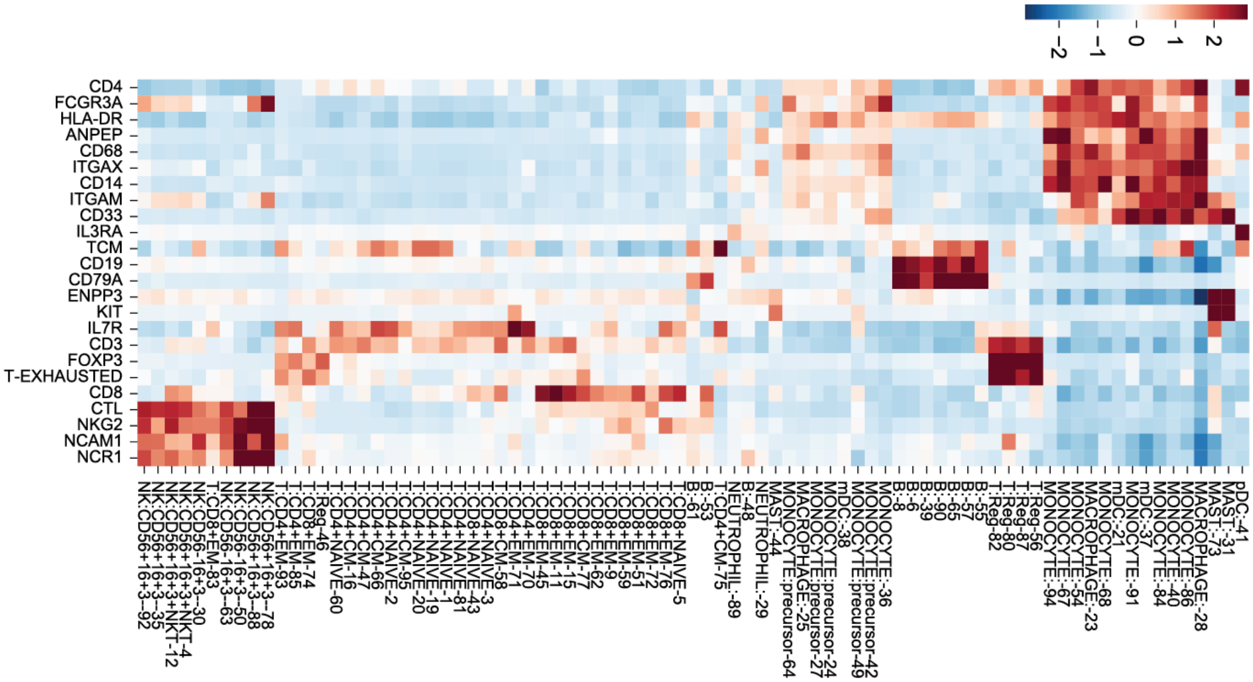


Figure 3.12: Expression of canonical and cell type markers across clusters, z-score normalized across clusters. T-exhausted denotes the mean expression of terminal differentiation signature listed in Table S4.

the marginal distribution of cell loadings across the signature and the relative contribution of each gene.

Therefore, when examining signature expression across patients we began by constructing a bar plot of the counts for each gene in the signature, corrected for cellular observation rate (the total number of genes observed with molecule count > 1). This displays the contribution of each gene to the signature (top panel in Figure 3.13,S3). The normalized values for each signature, per cell, are then summarized as a box plot to display the variation of cells in each patient (left panels). Finally, the cluster median of each gene is taken per patient, and the cluster medians are z-scored across patients. The z-scored values are plotted as a heatmap (center-right panel in Figure 3.13,S3), facilitating a comparison of signatures across patients[5].

[5] To create these lists we broadly surveyed the literature and manually curated consensus lists of genes to be included. The relevant literature that these signatures were derived from includes:

The genome-wide view allowed us to assess system-level differences between immune cell consortia in individual patients in, for example, metabolic signatures, including hypoxia (Figure 3.13). It is interesting to note that while all tumors expressed a similar average degree of a hypoxia signature, patients differed considerably in expression at the level of individual genes included in the signature. Similar variation was observed in fatty acid metabolism, glycolysis, and phosphorylation (Figure S3).

3.11 Breast Tumor Immune Cell Atlas Reveals Substantial Diversity of Cell States

Together, these analyses produce a tumor immune atlas that can be interrogated to understand tumor, tissue, and patient dependent differences in immune phenotypes. A complete map of the experimental procedure described in this chapter and the one preceding it, from sample extraction to the end of BISCUIT clustering, is displayed in Figure 3.14. After applying Biscuit to the data from all tumors (Figure 3.4), we found 67 clusters covering various T cell, macrophage, monocyte, B cell, and NK cell clusters. We first asked whether individual cells tended to be most similar to cells from their own samples or if the resulting cell profiles were well mixed using an entropy measure (STAR Methods). For each cell, this measure considers the neighborhood of

For the M1 and M2 macrophage polarization signatures we merged gene lists from (Sica and Mantovani, 2012); (Biswas and Mantovani, 2010); (Bronte et al., 2016) (Ugel et al., 2015) (Gabrilovich, 2017). For other myeloid-specific signatures we used (Villani et al., 2017) (pDCs, AXL/SIGLEC6 DCs, CD141/CLEC9A DCs, CD11C_A DCs, CD1C-/CD141- DCs, CD1C_B DCs, New Monocytes 1, New Monocytes 2, CD14+CD16- Classical Monocytes, and CD14+CD16+ Non-Classical Monocytes); and (Gesta, Tseng, and Ronald Kahn, 2007), (Perera et al., 2006), (Farmer, 2006; Lefterova and Lazar, 2009) (Lipid Mediators).

For T-cell-specific signatures we used (Wherry and Kurachi, 2015), (Wherry, 2011), (Schietinger et al., 2012) (Exhaustion and Anergy); (Glimcher et al., 2004) (Cytolytic Effector Pathway); and (Smith-Garvin, Koretzky, and Jordan, 2009), (Chtanova et al., 2005), and (Adam Best et al., 2013) (T-cell Activation).

For gene signatures used across cell types we used (Mantovani et al., 2008), and (Grivennikov, Greten, and Karin, 2010) (Pro and Anti-Inflammatory); (Platanias, 2005) (Type I and II Interferon Responses); (Ho et al., 2015) (glucose deprivation); (Benita et al., 2009; Makino et al., 2003) (Hypoxia/HIF Regulated); (Moreno-Sánchez et al., 2009), (Caton et al., 2010; Funes et al., 2007; Mues et al., 2009), (Beale, Harvey, and Forest, 2007) (Glycolysis, Gluconeogenesis, TCA Cycle, Pentose Phosphate Pathway, and Glycogen Metabolism), and (Whitfield et al., 2002) (G1/S).

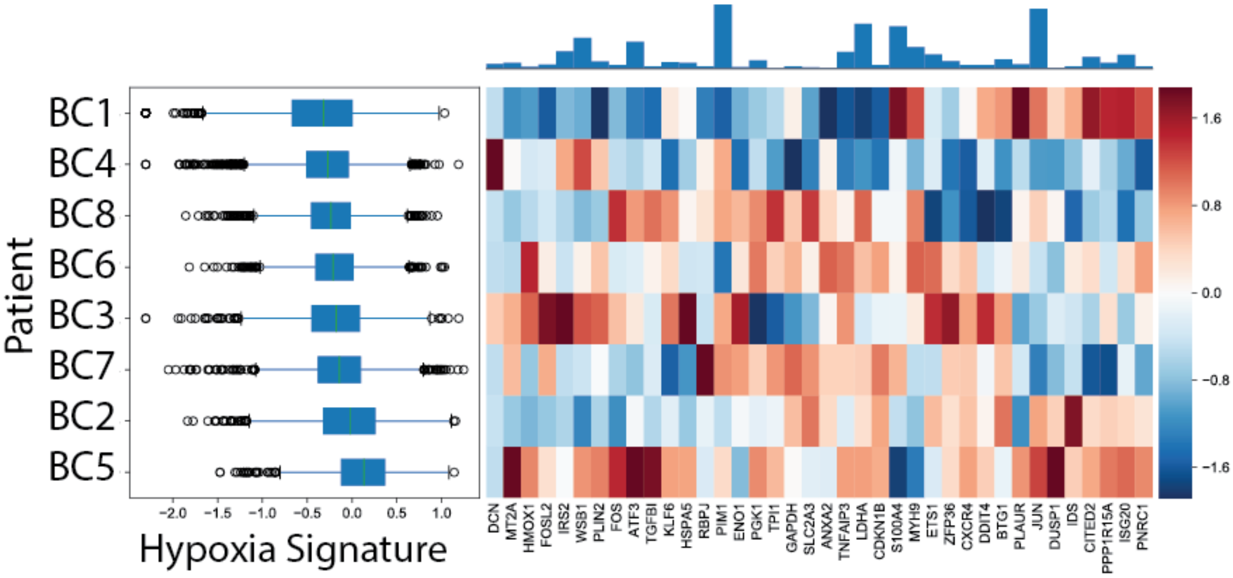


Figure 3.13: Left: Boxplots showing expression of Hypoxia signature (defined as the mean normalized expression of genes in the hypoxia signature listed in Table S4) across immune cells from each patient. Right: Heatmap displaying z-scored mean expression of genes in hypoxia signature. Top: Barplot showing total expression of each gene indicated in the heatmap, across all patients. See Figure S3 for additional signatures.

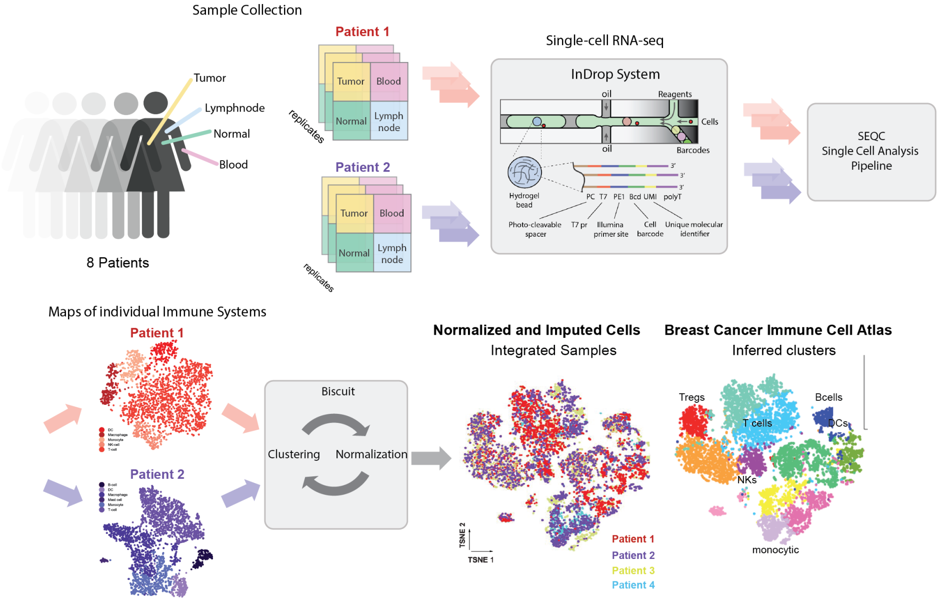


Figure 3.14: Flow chart displaying experimental design and analysis strategy.

its most similar cells and evaluates the entropy of the sample distribution in each such neighborhood. Low entropy indicates that most neighbors come from the same sample, whereas high entropy indicates that the neighbors (most similar cells) are well distributed across the different samples. Indeed, while cells were most similar within individual samples before normalization, this was corrected after Biscuit normalization with significantly improved mixing of cells across patients when compared against standard normalization methods (Figure 3.6,3.9). (U=1.7721e+09, p=0). Using this approach, we successfully retained information on immune cell activation while stabilizing differences in library size, and uncovered a rich and robust structure in imputed data, suggesting diversity in immune cell subtypes (Figure 3.6,3.8).

To construct a global atlas of immune cells, enabling characterization of the impact of environment on immune cell states, we merged data from 47,016 cells across all tissues and patients revealing a diverse set of 83 clusters, each identifying a cell type or state (Figure 3.10,3.11). This unexpectedly large number of clusters prompted us to test their robustness using cross-validation on subsets of the data (STAR Methods), finding assignments of cells to clusters were robust for most clusters (Figure S7). Most clusters were shared across multiple patients, indicating similar immune states across patients, with only 10 being patient-specific (Figure S6). We used entropy as a more stringent metric for patient mixing within clusters and found that the clusters span a range of different mixing levels (Figure S6).

We assigned each cluster to its associated cell type by comparing cluster mean expression to bulk RNA-seq as described above (Figure 3.10,3.11) and found 38 T cell clusters, 27 myeloid lineage clusters, 9 B cell clusters, and 9 NK cell clusters (Table S2). By examining the expression of canonical markers in immune cell clusters, we were able to confirm and build upon predictions made by the preceding analysis (Figure 3.12). Of the T cell clusters, we identified 15 CD8+ T cell clusters and 21 CD4+ T cell clusters, which were together split into 9 naive, 7 central memory, 15 effector memory, and 5 Treg clusters. We were additionally able to divide the myeloid lineage clusters into 3 macrophage, 3 mast cell, 4 neutrophil, 3 dendritic cell, 1 plasmacytoid dendritic cell, and 13 Monocytic clusters. Finally, we identified 9 B cell clusters, 3 CD56^{--} NK cell clusters, and 6 CD56^{++} NK cell clusters, of which 2 of which are likely NKT cells. These clusters can be distinguished by their differential expression patterns (Figure 3.15).

Since our characterization identified multiple clusters with the same cell type "label" based on surface markers and prior characterization of the corresponding peripheral blood cell phenotypes, e.g. 15 effector memory T-cell clusters (Figure 3.11), we wanted to confirm that all these clusters were indeed distinct. The distributions defined by the Biscuit parameters identified differentially expressed genes between clusters, including canonical immune genes, and defined multiple subpopulations within each major cell type (Table S3). Moreover, we observed a prominent effect of covariance in defining the T cell clusters by comparing similarity of pairs of clusters with and without the effect of mean expression (Figure S11); large differences between most clusters remained even after mean gene expression was equalized. Thus, our approach robustly identified cell states that were distinct from one another and shared across multiple tumor microenvironments. As T-cell and myeloid cells represent the most abundant and diverse, and arguably most biologically significant, immune cell subsets in the tumor microenvironment, we focused our subsequent in-depth analyses on these two major cell types. This investigation is detailed in the next, final chapter of the dissertation.

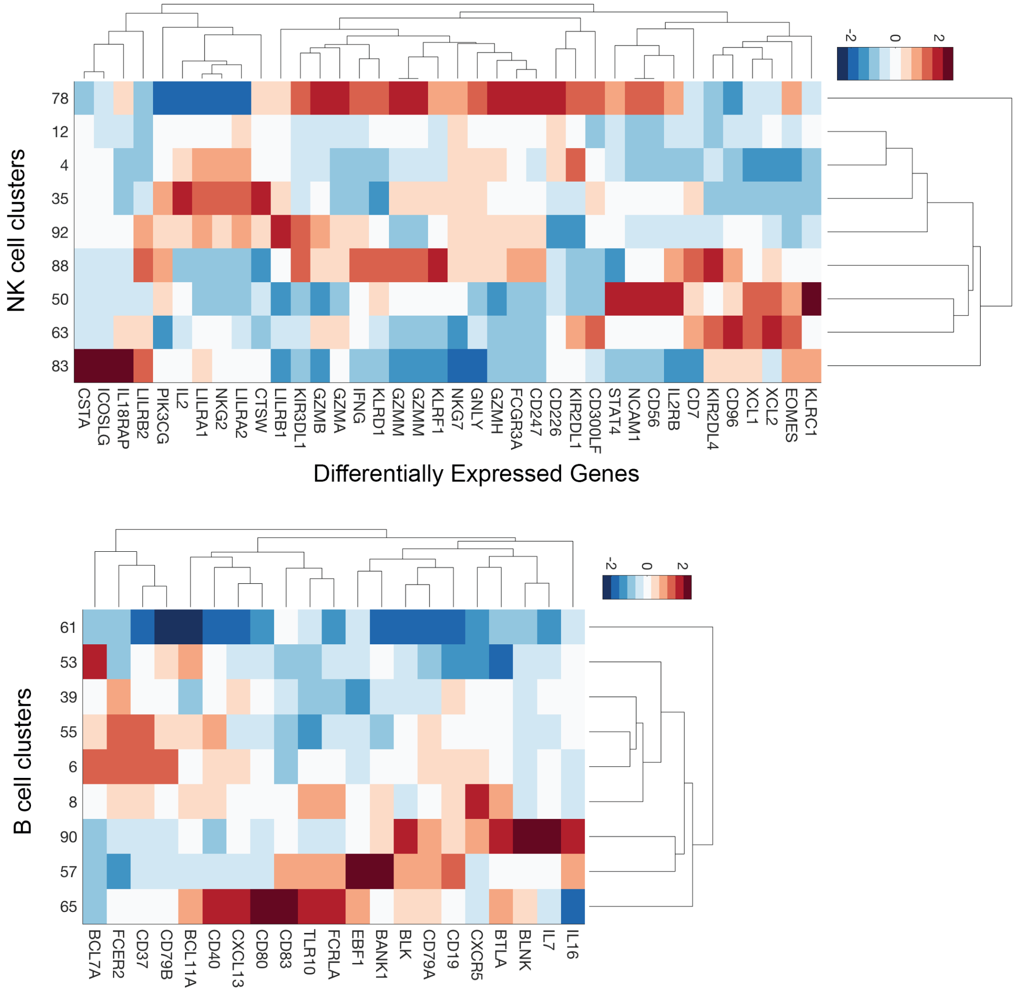


Figure 3.15: Differentially expressed genes in b-cells (top) and NK-cells (bottom) standardized by z-scores within cell type. As an example, the expression of CD19 is standardized across all B cell clusters to highlight clusters with higher or lower expression of the marker compared to the average B cell cluster, but is highly expressed in nearly all B cell clusters (refer to Table S3 for all DEGs in these and other clusters).

Chapter 4

Quantifying Tissue- and Microenvironment-Induced Immune Cell Variation

4.1 Tissue Residence has a Strong Effect on the Diversity of Immune Phenotypic States

A key goal of this study was to quantify the extent to which variation in immune cell phenotypes is driven by their tissue of residence, i.e. cancerous vs. normal breast tissue, using peripheral blood or the lymph node cells as references. To gain a qualitative understanding of phenotypic overlap between tissues, we carried out tSNE co-embedding (Maaten and Hinton, 2008) of the merged dataset annotated by clusters. This analysis showed that T cells in blood and lymph node were dramatically dissimilar to cancerous or normal breast tissue resident T cells, which in contrast, displayed many shared phenotypes (Figure 4.1). We observed that gene expression of

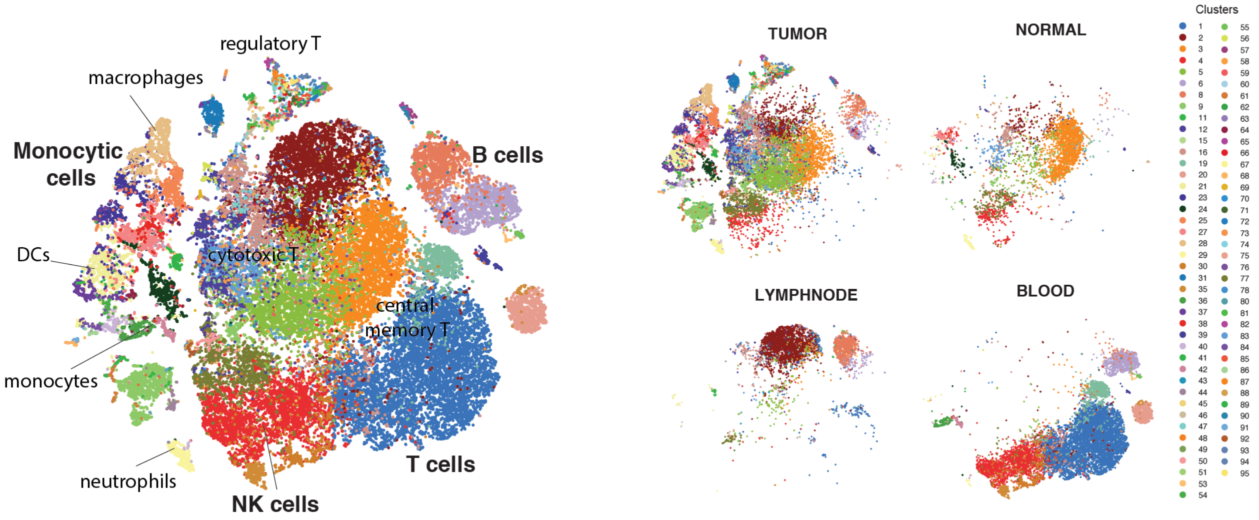


Figure 4.1: Breast immune cell atlas inferred from combining all patient samples and tissues, presented after Biscuit and projected with t-SNE. Each dot represents a cell and is colored by cluster label; major cell types are marked according to Figure 2F, H. Right: Subsets of immune atlas t-SNE projection in showing cells from each tissue presented separately on the same coordinates as right to highlight the differences between tissues compartments.

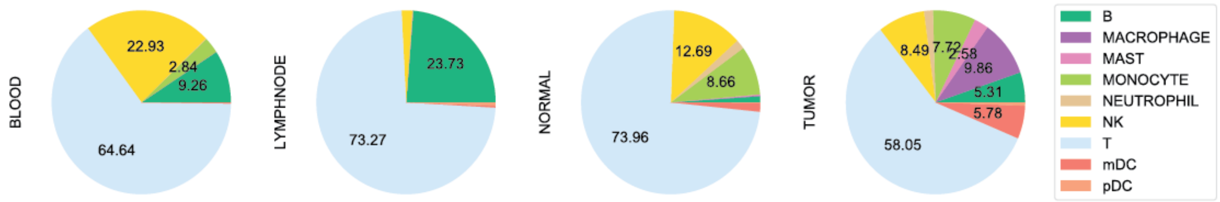


Figure 4.2: Proportions of cell types across tissue types in pie charts.

T cells dramatically differed between blood and tissue resident cells, with a large blood-derived cluster of cells being phenotypically distinct from T cells in normal and tumor tissue (shown in blue). In contrast, Both T cells and myeloid lineage cells exhibited considerable phenotypic overlap between tumor and normal tissue samples. Of the two classes of tissue resident cells, tumor cells displayed greater phenotypic heterogeneity, appearing to expand populations observed in the normal breast (Figures 4.1, 4.2 summarize distributions of cell types across tissues.)

Next, we quantifed the above observation that immune cells from tumor and normal tissue are more similar to one another than to other tissues. To accomplish this, we constructed a 10-nearest neighbor graph over 15 PCA components summarizing a uniformly selected subset of n=3000 cells from each tissue. We reasoned that a cell's closest neighbors in this low-dimensional embedding are the cells with the closest phenotypes. We then examined the overlap between each pair of tissues u and v: $o_{u,v} = \frac{1}{n}\{1 \; if \; \omega_i = u$ and $\omega_j = \; v \; else \; 0\}$, where n is the number of cells in the subset, k is the number of neighbors, and $u, v \in \{$tumor, normal, lymph node, blood$\}$. with ω_i denoting the tissue for cell i and $j = 1, \ldots, k$ denotes the neighbors of cell i. Examining all the pairwise shared-neighbor relationships, we confirmed that tumor and normal have the highest frequency of being co-identified as neighbors.

To determine if this enrichment was significant, we built a null distribution from all overlaps between all pairs of tissues, and we calculated the z-score of $o_{tumor,\ normal}$ compared to the distribution of all pairwise overlaps (z=2.68), for which a z-test confers a p-value of p=1.4e-4. This suggests that similarity between tissue resident cells is a positive outlier, compared with similarities between other pairs of tissues. Consequently, this result highlights that tissue of residence is a significant determinant of phenotypes of human cells of hematopoietic origin, and that states or biomarkers identified from blood immune cells may not necessarily extend to tissue embedded immune populations.

Finally, we confirmed that the cell types and states observed in these data comport with our prior understanding of the structure and function of the immune system using χ^2 enrichment testing between cell types and tissues. We began by transforming the data for each tissue to have equal cell count and created a 2-factor contingency table of cell types versus tissues. We then calculated χ^2 enrichments for each tissue type. We confirmed that naive T cells were strongly enriched in three blood-specific clusters (χ^2=361.4, df=1, p=3e-80), while B cells were most prevalent in the lymph node than in other tissues (χ^2=1737.1, df=1, p=0.0). A subset of T cell clusters were present in both tumor and normal tissue, but the cytotoxic T cell clusters (χ^2=93.7, df=1, p=3e-25) and T reg cells (χ^2=336.0, df=1, p=5e-91) were more abundant in tumor, as expected, given that tumor should be the target for the immune response. Similarly, some myeloid clusters were shared between normal and tumor tissue, whereas clusters of more activated monocytes and tumor-associated Macrophages (TAMs) were specific to tumor (χ^2=2420.6, df=1, p=0.0). Overall, these observations confirm that our atlas is composed of rationally generated data, consistent with expectations of normal immune functionality.

4.2 Tumor Microenvironment Drives an Expansion of Immune Cell Phenotypic Space

BISCUIT uncovered a large number of normal breast tissue resident cell states, manifested by 13 myeloid and 19 T cell clusters that were not observed in circulating blood or in the secondary lymphoid tissue. Furthermore, our data showed that the set of clusters found in normal breast tissue cells represented a subset of those observed in the tumors; 14 myeloid and 17 T cell clusters were only found in the tumor, doubling the number of observed clusters of these cell types relative to normal tissue, and there were no clusters specific to normal tissue. This increased diversity of cell states correlates with a significant increase in the variance of gene expression in tumor compared to normal tissue (Figure 4.3),

To better understand whether the increase in variance of gene expressions in tumor tissue is due to activation or additional phenotypes that are independent from those found in normal tissue, we sought to define a metric for the "phenotypic volume" occupied by cells. Given that the volume of an N-dimensional matrix can be expressed as the absolute value of the determinant

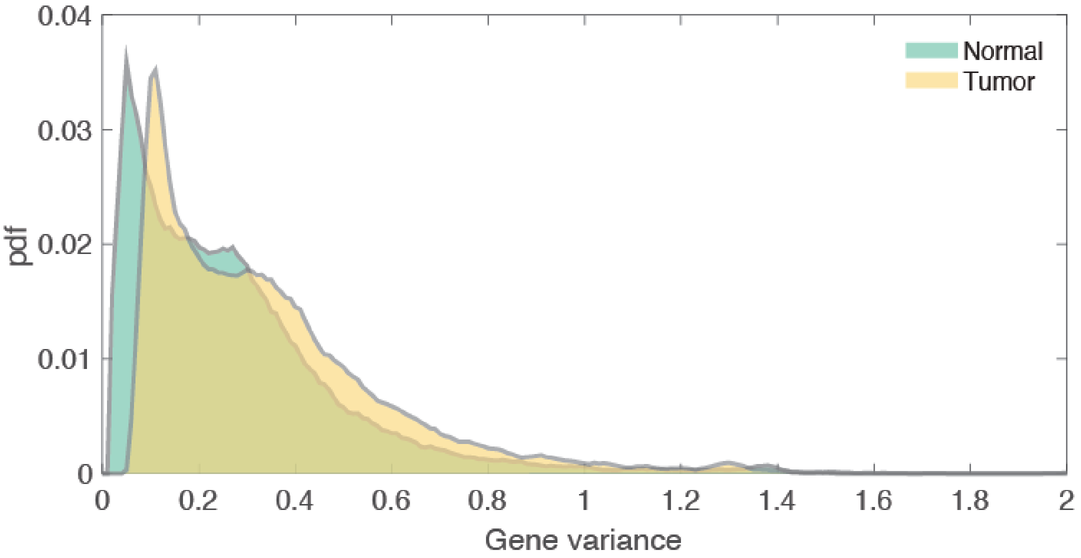


Figure 4.3: Distribution of variance of normalized expression computed for each gene across all immune cells (all patients) from tumor tissue compared to that in normal breast tissue.

of the matrix, we reasoned that we could leverage this relationship to calculate the phenotypic volume of our data matrices.

We therefore defined "phenotypic volume" (V) for a subpopulation of cells as the determinant of the gene expression covariance matrix in that subpopulation, which considers covariance between all gene pairs in addition to their variance. The (symmetric) covariance matrix can be written as $\Sigma = [\overrightarrow{s_1}, \ldots, \overrightarrow{s_d}]$ where $\overrightarrow{s_i}$ for $i = 1, \ldots, d$ is a vector containing covariance between gene i and all other genes. Its determinant $det\ (\Sigma)$ is equal to the volume of a parallelepiped spanned by vectors of the covariance matrix (Tao and Vu, 2005).

For example, if the covariance values between a gene i and other genes is very similar to the covariance of another gene i' and other genes, such that $\overrightarrow{s_i}, \overrightarrow{s_{i'}}$ are dependent, gene i' does not add to the volume. Extending this to all genes, we sought to evaluate whether the increase in expression variances (Figure 4.3) are associated with phenotypes activated in tumor that are independent from those in normal tissue, i.e. are novel independent phenotypes observed in tumor that suggest additional mechanisms and pathways being activated in tumor.

Applied to a simplified case with only two phenotypes, the determinant, which is equal to the area of the parallelogram spanned by two vectors representing the phenotypes, is larger if the phenotypes are independent, but would be equal to zero if they are dependent. With more than two phenotypes, we are then interested in measuring the volume of the parallelepiped spanned by these phenotypes. The (pseudo-)determinant can also be more rapidly computed as the product of nonzero eigenvalues of the covariance matrix:

$$V = det(\Sigma) \;=\; \lambda_{\mathrm{e}} \;= \lambda_1 \lambda_2 \ldots \lambda_E$$

To quantify the change in phenotypic volume from normal to tumor, we computed this volume metric for each major cell type of T, monocytic, and NK cells. To correct for the effect of differences in the number of cells across cell types and tissues, we uniformly sampled 1000 cells with replacement from each cell type per tissue and computed the empirical covariance between genes based on imputed expression values for that subset of cells. This was followed by singular-value decomposition (SVD) of each empirical covariance matrix and computation of the product of nonzero eigenvalues as stated in the equation above. B cells were not included in this comparison due to the very small number of B cells in normal tissue.

Given the high number of dimensions (genes), the volumes were normalized by the total number of genes (d). For robustness, this process was repeated 20 times to achieve a range of computed volumes for each cluster in each tissue, which are summarized with box plots (Figure 4.4,S12) showing statistically significant expansions of volume in tumor compared to normal in all three cell types. The fold change in volume was 7.39e4 in T cells, 1.18e14 in myeloid cells and 6.08e4 in NK cells (Mann-Whitney U, p=0.0, for all three tests), indicating a massive increase in phenotypic volume in tumor compared to normal tissue. These data confirm that we observe

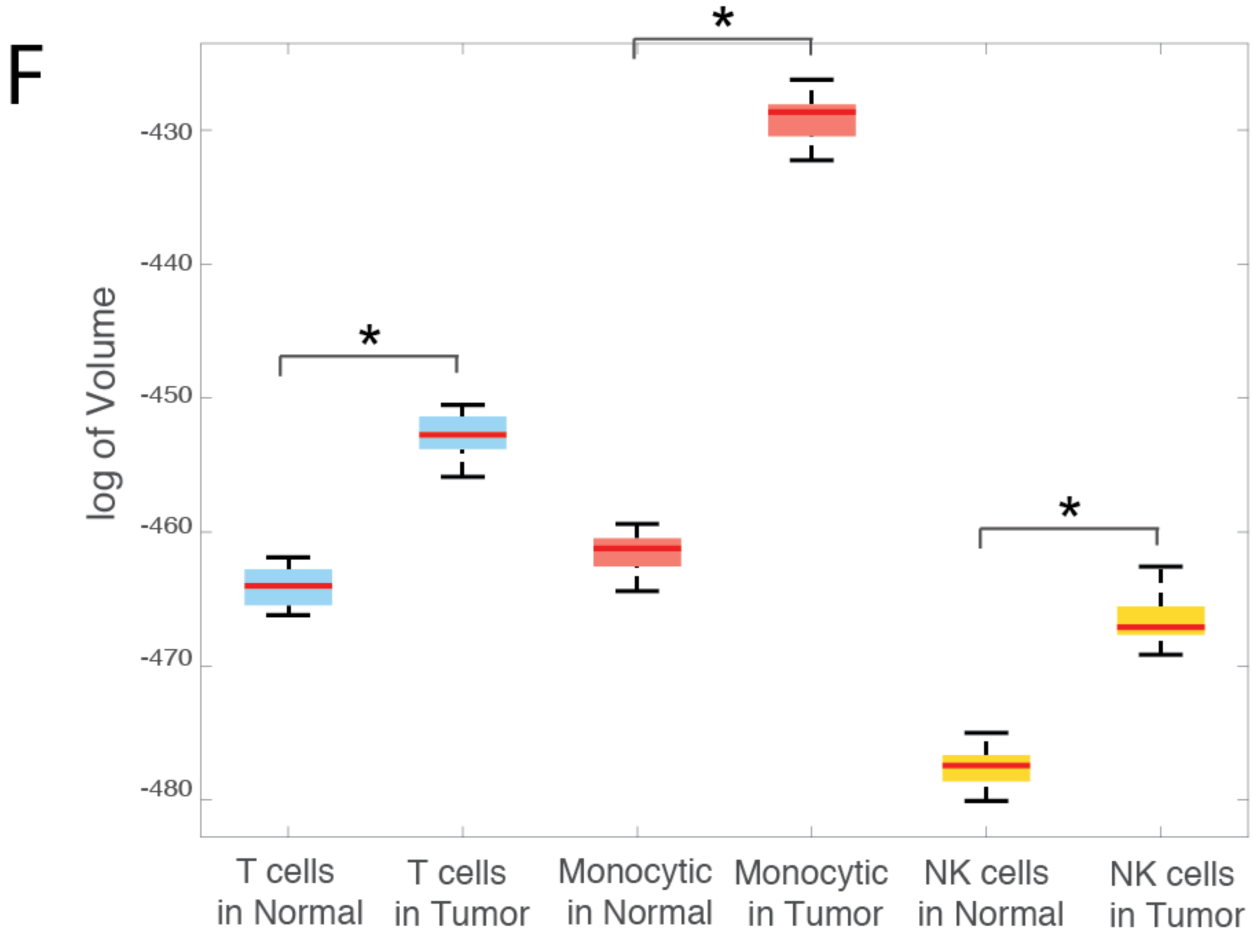


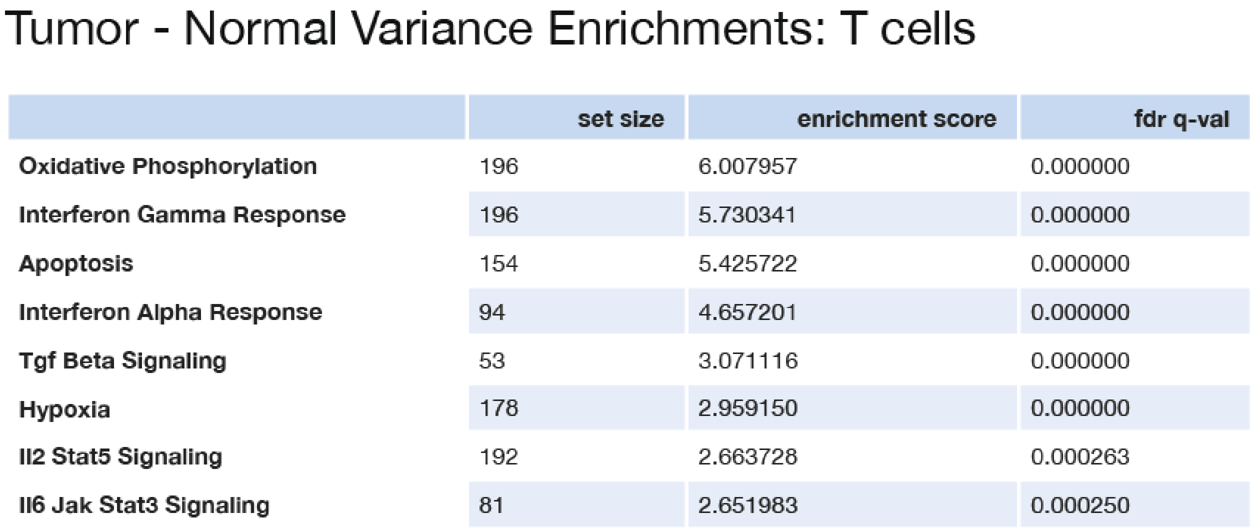
Figure 4.4: Phenotypic volume in log-scale (defined as determinant of gene expression covariance matrix, detailed in STAR methods) of T cells, monocytic cells, and NK cells, comparing tumor immune cells and normal breast immune cells after correcting for differences in number of cells. Massive expansion of volume spanned by independent phenotypes active in tumor compared to normal tissue is shown for all three major cell types.

an expansion of cell states in the tumor in comparison to the normal tissue. The volume analysis also alleviated concerns that technical factors may drive the increased diversity of immune cell phenotypes in tumor that were highlighted in the previous section.

We were motivated to undertake volume analysis in part because we observed higher variance in tumor cells, as quantified by both by greater variance explained by the top 10 PCs, and larger, more disperse clusters in tSNE. However, as highlighted earlier, immune cells increase their mRNA expression when activated, a signal that BISCUIT does not explicitly remove. Thus we would expect to see additional signal within tumor immune cells. In addition, because our experiment focused on tumor infiltrating lymphocytes, we observed many more of them than their tissue-resident cognates.

None of our previous approaches explicitly control for the total number of input observations. However, the volume analysis downsampled each tumor and normal tissue both in terms of molecules and cells, confirming our observations with a much stricter normalization method. Thus, volume analysis allowed us to confirm that the heightened variation observed in tumor was the result of a richer complement of biological stimuli, rather than variation induced by sampling effects stemming from the higher coverage of tumor.

To determine possible sources of the observed increase in phenotypic volume, we performed Gene Set Enrichment Analysis (GSEA) (Subramanian et al., 2005) on the genes with the largest differences in variance between tumor and normal immune cells. This revealed heightened variation in targets of key immune signaling molecules, including type I (IFNα) and II interferons (IFNγ), TNFα, TGFβ, and IL6/JAK/STAT signaling (Figure 4.5, S12). These results suggest that the heightened variation observed in tumor immune cells is likely due at least in part to variation in the local concentrations of signaling factors designed to elicit immune reactivity against cancer and other foreign pathogens, consistent with previous findings that tumor microenvironments differ significantly in their extent of inflammation, hypoxia, expression of ligands for activating and inhibitory receptors, and nutrient supply (Finger and Giaccia, 2010; Jiménez-Sánchez et al., 2017).



Tumor - Normal Variance Enrichments: T cells

	set size	enrichment score	fdr q-val
Oxidative Phosphorylation	196	6.007957	0.000000
Interferon Gamma Response	196	5.730341	0.000000
Apoptosis	154	5.425722	0.000000
Interferon Alpha Response	94	4.657201	0.000000
Tgf Beta Signaling	53	3.071116	0.000000
Hypoxia	178	2.959150	0.000000
Il2 Stat5 Signaling	192	2.663728	0.000263
Il6 Jak Stat3 Signaling	81	2.651983	0.000250

Figure 4.5: Hallmark GSEA enrichment results on genes with highest difference in variance in tumor T cells vs normal tissue T cells. See Figure S3 for enrichment in monocytic and NK-cells. Most significant results are shown; full lists of enrichments are presented in Table S5.

4.3 Intra-tumoral T cells Display Continuous Phenotype Gradients

To explore further explore the most significant sources of variation in T-cell immune states, we carried out unbiased analyses by decomposing the gene expression using diffusion maps (Coifman et al., 2005; Haghverdi, Buettner, and Theis, 2015; Haghverdi et al., 2016; Moignard et al., 2015; Setty et al., 2016). Diffusion maps is a nonlinear dimensionality reduction technique to find the major non-linear components of variation across cells. It can be thought of at a high level as a non-linear analogue of PCA, and is often applied as such.

We computed diffusion components in each cell type separately using the Charlotte Python package, which implements diffusion maps as described in (Coifman et al., 2005). To account for differences in cell density and cluster size, we used a fixed perplexity Gaussian kernel with perplexity 30, with symmetric Markov normalization and $t = 1$ diffusion steps. We selected $t = 1$ because, in our data, this approximates diffusion of information for each cell through its 20 nearest neighbors. Put another way, when t is low, diffusion maps function more to identify components of non-linear variation. When t is high, diffusion maps function to spread information, which can be useful for imputing missing values, by filling in missing information from other similar cells.

We selected a conservative value because we wanted to ensure that information did not diffuse beyond the borders of our smallest cluster (30 cells). Equally important, we wanted to ensure that claims made about continuity of phenotypic space could not be driven by Diffusion Maps themselves. Given that we observe over 25,000 T-cells of various types, and that diffusion does not exceed each cells 20-nearest neighbors, we can confidently claim that the global manifold is unaffected by these changes.

When we examined the components produced by diffusion maps, we observed that while some components distinguished discrete clusters, the majority of components defined gradual trends of variation across T cell clusters (Figure 4.6 left,S13). However, the first two diffusion components identified two isolated clusters, owing to their strong dissimilarity (Figure 4.1). The first was cluster 9, which is the most distinct T cell cluster as measured by Bhattacharyya distance (Figure S10) and shows characteristics similar to NKT cells (Table S3) and the second was cluster 20, which is a blood-specific naive T cell cluster predominantly from one patient (Table S2).

Since these two clusters were very distant from other T cell clusters according to a variety of comparison metrics, the two components corresponding to them function more like classifiers, and so were ignored as we wished to focus on studying "continuous" components that quantify heterogeneity across multiple clusters. The top 3 continuous components correlate, respectively, with signatures for immune cell activation, terminal differentiation, and hypoxia.

The most informative component of variation, labeled as "activation", was highly correlated with gene signatures of T cell activation and progressive differentiation (p=0.0), along with IFNγ signaling (p=0.0). The mean expression of the activation signature steadily increases along the component (Figure 4.6, top right), with a concomitant gradual increase in expression of activation-related genes (Figure 4.6, bottom right). The next components were labeled as T cell activation, Terminal Exhaustion, and Hypoxia (Figure 4.6), respectively as they were most highly correlated with the corresponding gene signatures. The subsequent component is labeled as Tissue Specificity, as it separates cells primarily on the basis of their tissue of origin and helps explain heterogeneity in T cells across tissues.

When we examined the localization of different cell types along the activation component, we found that intra-tumoral T cell populations are enriched at the positive end of the component relative to T cells found in healthy tissue (t-test p=0.0, Figure 4.6,4.7). Specifically, tumor-resident effector memory T cells and T reg cells compose the most activated end, while naïve T cells from peripheral blood congregate at the inactive terminus, consistent with their quiescent cell state (t-test p=0.0, Figure 4.7). However, while the mean expression levels of clusters vary gradually along the component, there is also a wide range of activation states within each cluster (Figure 4.7). Examining the individual genes most correlated with the component reveals a diverse set of genes whose expression is well documented to increase upon T cell activation and progressive differentiation. These included genes encoding cytolytic effector molecules granzymes A and K (GZMA and GZMK), pro-inflammatory cytokines (IL-32), cytokine receptor subunits (IL2RB), chemokines (CCL4, CCL5), and their receptors (CXCR4, CCR5) (Figure 4.6, bottom right).

The next most informative component of variation was labeled terminal differentiation (Figure 4.8). The genes most correlated with it include co-stimulatory molecules (CD2, GITR, OX40, and 4-1BB) as well as co-inhibitory receptors (CTLA-4 and TIGIT) (Figure S14). This set also

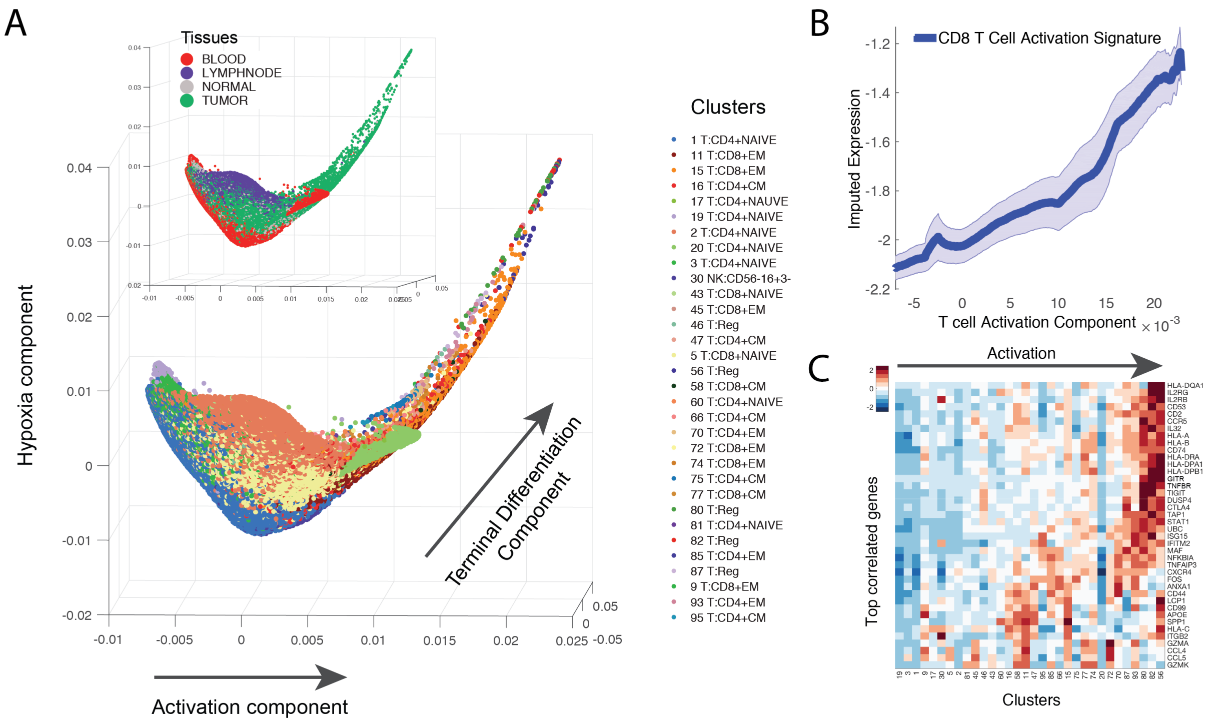


Figure 4.6: (left) Visualization of all cells from T Cell clusters using first, second, and third informative diffusion components (two uninformative components denoting isolated NKT and blood-specific clusters were removed from further analysis). Each dot represents a cell colored by cluster, and by tissue type in insert. The main trajectories are indicated with arrows and annotated using the signature most correlated with each component. See Figure S15 for additional components. (top right) Traceplot of CD8 T cell activation signature (defined as mean expression across genes in signature listed in Table S4) for all T cells along first informative diffusion component. Cells are sorted based on their projection along the diffusion component (x-axis), and the blue line indicates moving average over normalized and imputed expression, using a sliding window of length equal to 5% of total number of T cells; shaded area displays standard error (y-axis). (bottom right) Heatmap showing expression of immune-related genes with the largest positive correlations with activation component, averaged per cluster and z-score standardized across clusters; columns (clusters) are ordered by mean projection along the component.

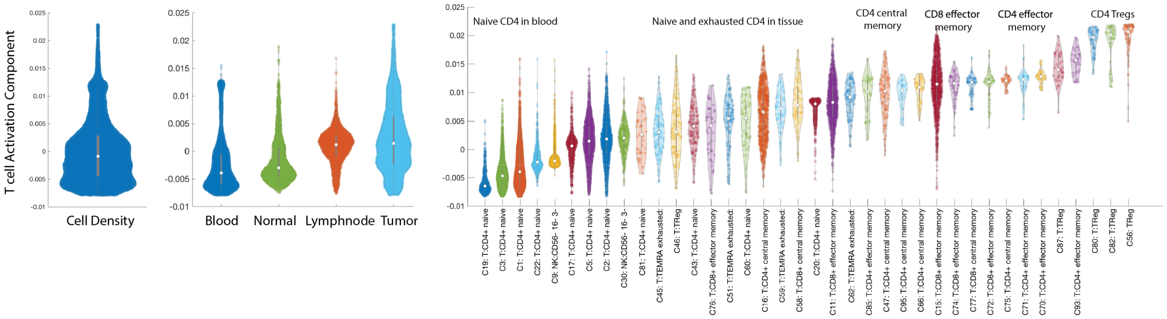


Figure 4.7: Violin plot showing the projection of T-cells along activation component aggregated by total density (left), tissue type (middle), and cluster (right). See Figure S4 for violin plots for additional components. Number of dots inside each violin are proportional to number of cells.

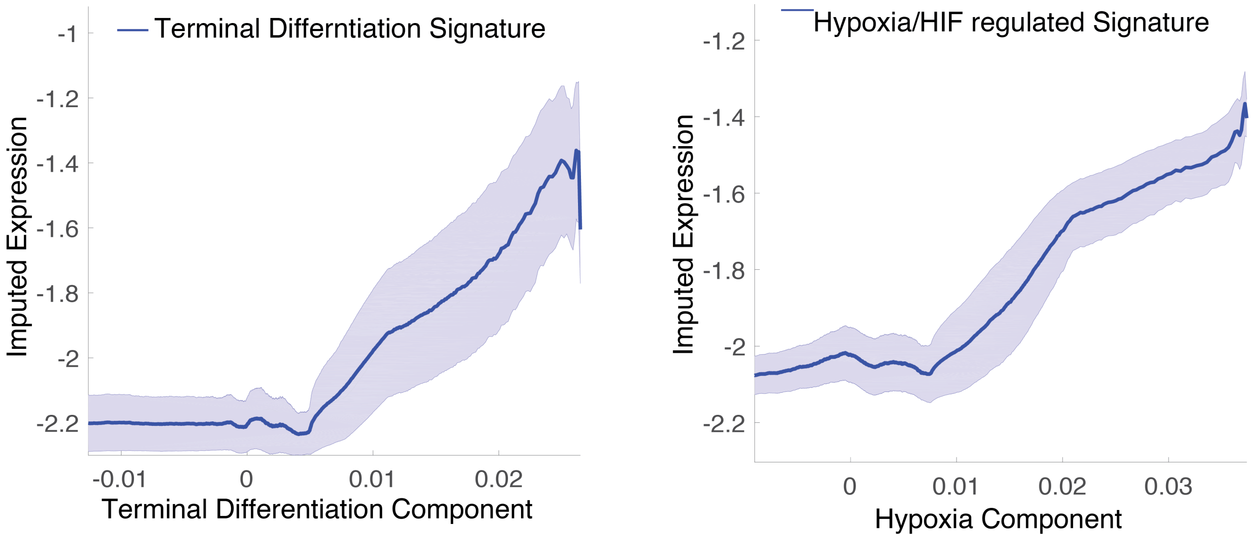


Figure 4.8: Trace-plots (as in B) of (left) terminal differentiation signature along second informative component and (right) hypoxia signature along third informative component, labeled respectively as terminal differentiation and hypoxia components. List of genes associated with signatures are presented in Table S4.

included Foxp3, IL2RA, and Entpd1 (CD39), genes whose high expression is characteristic of T reg cells (Josefowicz, Lu, and Rudensky, 2012). The same primarily T reg clusters reside at the very terminal end of both the activation and terminal differentiation components, and there is a moderate degree of overlap in the genes most correlated with the two (Figure 4.6 left, bottom right; S14). However, there are also important exceptions—including the markers of exhaustion listed above—and crucially, the two trajectories traverse different paths through the remaining clusters (Figure 4.9). Indeed, some clusters—notably T cells from the lymph node (e.g. cluster 16)—express higher levels of activation than terminal differentiation (t-test p=0.0; Figure 4.6,4.7), consistent with the idea that T cell exhaustion and terminal differentiation largely occurs in non-lymphoid tissues and not in the draining lymph node.

Interestingly, visualizing the T cell activation and terminal differentiation components together revealed remarkable continuity, in essence representing a single continuous trajectory of T cells towards a terminal state (Figure 4.6 left,S13). Thus, our observations suggest that T cells reside along a broad continuum of activation, and that their conventional classification into relatively few discrete activation or differentiation subtypes may grossly oversimplify the phenotypic complexity of T cell populations resident in tissues.

4.4 Response to Diverse Environmental Stimuli Define Intra-Tumoral T-Cell States

Noting that only a few of the clusters were well delineated by the strongest components of variation, we sought to understand the variation driving the observed clustering. We examined the expression of gene signatures for response to environmental stimuli in each T cell cluster and found that while most clusters were arranged in a continuous fashion along the activation

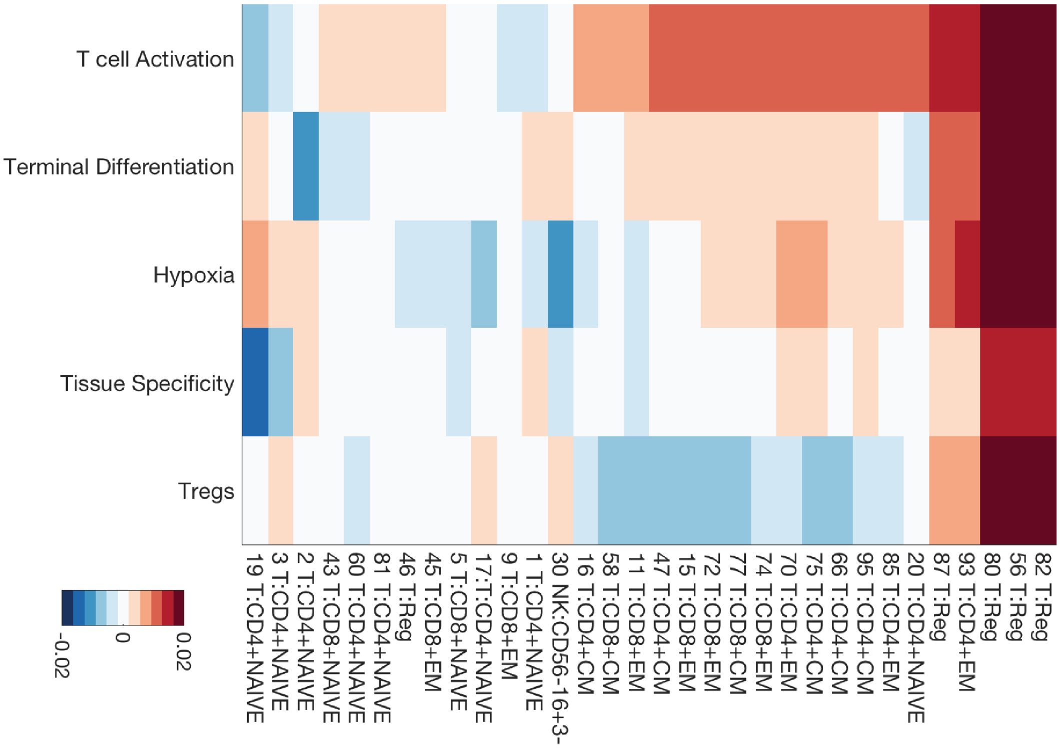


Figure 4.9: Heatmap of cells projected on each diffusion component (rows) averaged by cluster (columns).

component, each cluster appeared unique when looking across multiple components and signatures in a combinatorial fashion. We were interested to know whether cells show continuity as opposed to defined cell states along various diffusion components. For example, we wanted to know whether T cells exhibit defined states with different activation levels. For this, we computed the distribution of cells projected on each diffusion component and then used Hartigan's Dip Test (Hartigan and Hartigan, 1985) to test whether the distribution of cells is unimodal (broad continuum of cells) or alternatively multimodal (representative of multiple defined states) with $p < 0.01$.

In Supp. Figure S13, we observe that in the case of the T Cell Activation component, the null hypothesis of unimodality is not rejected, indicating that the distribution of cells is similar to a broad unimodal distribution as opposed to a multimodal distribution with defined states. In contrast, other components (such as the Tissue Specificity Component) exhibit multimodal distributions with distinct modes implying distinct states (in this case corresponding to various tissues)[1].

Our data show that CD4 effector and central memory clusters (Figure 4.10) exhibit variable levels of expression of genes contributing to signatures for Type I and II interferon response (F-test, p=1e-54 and 0.008 respectively), Hypoxia (F-test, p=4e-64), and Anergy anergy (F-test p=4e-69). Moreover, different CD8 effector and central memory clusters (Figure 4.10) have different expression levels of activation (F-test p=2e-114), pro-inflammatory (F-test p=1e-39), and cytolytic effector pathways related genes (F-test p=6e-32). These examples suggest that in a heterogeneous tumor microenvironment, differing in degree of inflammation, hypoxia and nutrient

[1] In the case of myeloid cells, the null hypothesis of unimodality is rejected in all diffusion components, indicating that myeloid cells lie in distinct states along all major components explaining variation across cells that were analyzed (Supp. Figure S19).

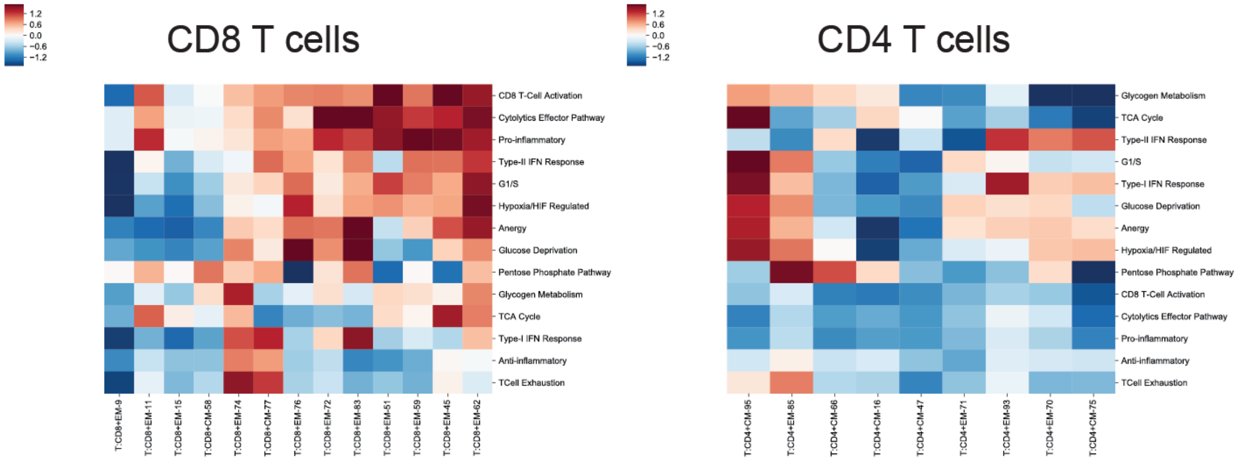


Figure 4.10: Heatmaps showing normalized and imputed mean expression levels for a curated set of transcriptomic signatures (rows) important to T Cells (listed in Table S4) for (A) CD4 memory clusters, (B) CD8 memory clusters, and (C) T Regulatory clusters. Only signatures with high expression in at least one T cell cluster are shown. Signature expression values are z-scored relative to all T cell clusters but only shown for clusters of the same cell type for ease of visualization.

availability, subpopulations of T cells either sense different environmental stimuli or respond differently to these stimuli. While many of these responses (e.g. activation or hypoxia) create phenotypic continuums, their different combinations can result in more discrete behaviors.

In contrast to effector T cells, T reg clusters displayed less variation in expression across these gene signatures: the majority of these clusters featured comparable patterns for anti-inflammatory activity, exhaustion, hypoxia, and metabolism gene sets (Figure 4.10). To identify features distinguishing the T reg clusters, we examined the Biscuit parameters that differ between them. We found that beyond mean expression levels, covariance parameters varied significantly between clusters, and drove the observed differences. Specifically, two marker genes may exhibit similar mean expression in two different clusters (e.g. highly expressed in both), while the clusters show opposite sign in covariances in these genes. This occurs due to the genes typically being co-expressed in the same cells in one cluster (i.e. positive covariance), while being expressed in the other cluster in a mutually exclusive manner (i.e. negative covariance) (Fig 4.11). It is note-

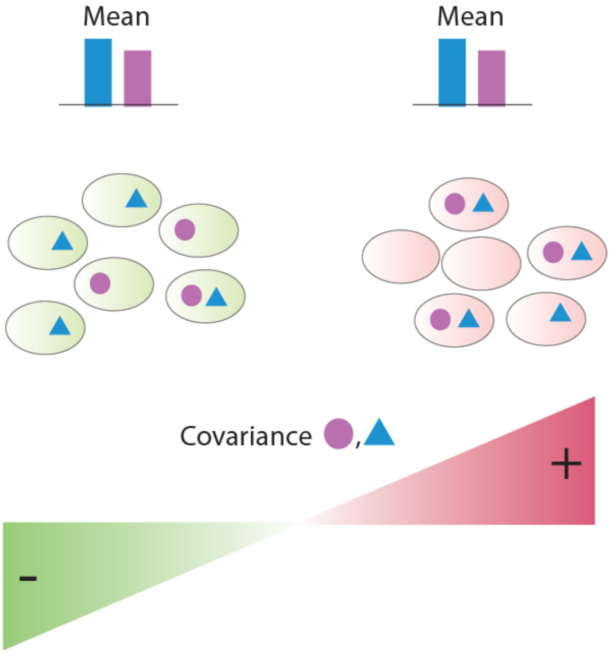


Figure 4.11: Cartoon illustration of two clusters of cells showing similar mean expression for two example marker genes but opposite covariance between the same two genes.

worthy that clusters were inferred based on the expression of over 14,000 genes; hence, negative covariance between two specific genes does not necessarily imply the existence of sub-clusters.

As an example, our analysis showed that the CTLA-4 gene, which encodes a prototypical inhibitory checkpoint receptor that is highly expressed in T regs and activated T cells, exhibited rich covariance patterns with other mechanistically related genes (Figure 4.12,4.13; S16,S17). CTLA-4 co-varied strongly with TIGIT and co-stimulatory receptor GITR in T reg clusters 46, 56, and 87; with CD27 in clusters 46 and 80; and with co-stimulatory receptor ICOS only in cluster 80 (Figure 4.12,4.13); We observed considerable differences in covariance patterns between numerous pairs of other checkpoint genes across T reg clusters. Additionally, covariance between other key immune genes in T reg clusters exhibited modular structures, with groups of genes

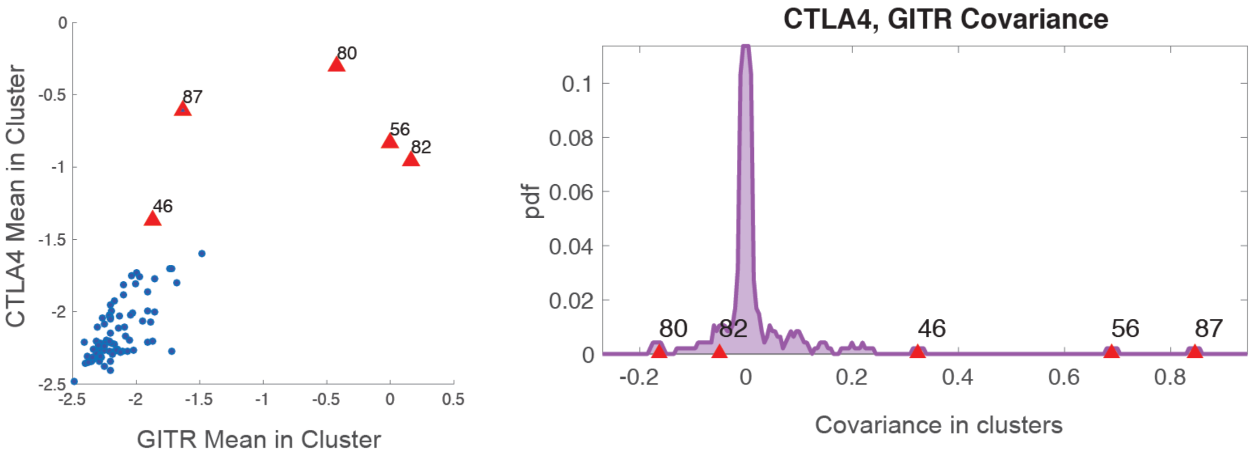


Figure 4.12: Scatter plot showing mean expression of GITR vs. CTLA-4 for each T cell cluster (represented by a dot). T reg clusters, labeled in red, have high mean expression levels of both genes. Distribution of covariance between GITR and CTLA-4 across all T cell clusters (purple), with values for T reg clusters labeled in red. Note that T reg cluster covariance values are present as both positive (46, 56, 87) and negative (80) outliers, exhibiting differences in covariance despite sharing high mean expression levels. See Figure S16 for similar computation on the raw, un-normalized, and un-imputed data, verifying the result.

co-expressed together, suggesting co-regulation and potential involvement in similar functional modalities (Figure 4.13).

Since varied proportions of T reg clusters were observed in individual patient samples, the differences in gene co-expression were present across patients as well as clusters within a given patient (Figure 4.14). We observed that the majority of patients did not have all 5 subtypes of T reg cells, and in fact most were dominated by only one subtype (cluster). It must be noted that we also observed similar differences in co-variation patterns across activated T cell clusters, even if not playing as essential a role in their delineation (Figure S18). Thus, co-variation of genes has a role in defining T cell clusters, in particular T reg clusters (Figure 4.12)

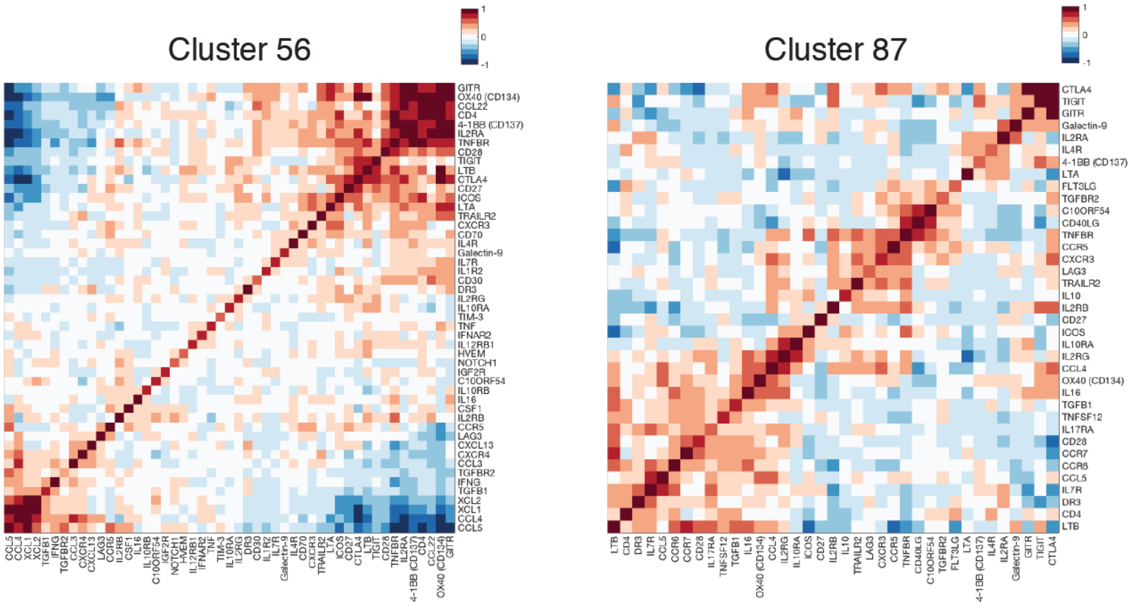


Figure 4.13: Heatmaps showing covariance between immune genes in T reg clusters 56 (left panel), and 87 (right panel). Note different modules of covarying genes.

4.5 Significance of Differences in Covariances of Raw Data Drive Biscuit Clustering

To verify that the differing covariance patterns in Figures 5 and 7 were not the result of computational modeling decisions, we tested the difference in covariance in raw median library size normalized data, categorizing the raw data using the BISCUIT cluster labels. As the raw data involves significant amount of dropouts, co-expression patterns and their signs cannot be easily visualized or inferred. Hence, we performed hypothesis testing accounting for the level of dropouts by comparing the observed empirical covariance between a pair of genes $i,\ i'$ to a null distribution for the gene pair in which co-expression patterns are removed. We assume the null hypothesis to be the case where covariance between a specific gene pair for a given cluster is the same across all clusters.

Specifically, to test whether $cov(\overrightarrow{x_i}, \overrightarrow{x_{i'}})$ in a cluster $k,$ with $\overrightarrow{x_i}, \overrightarrow{x_{i'}}$ being expressions of genes

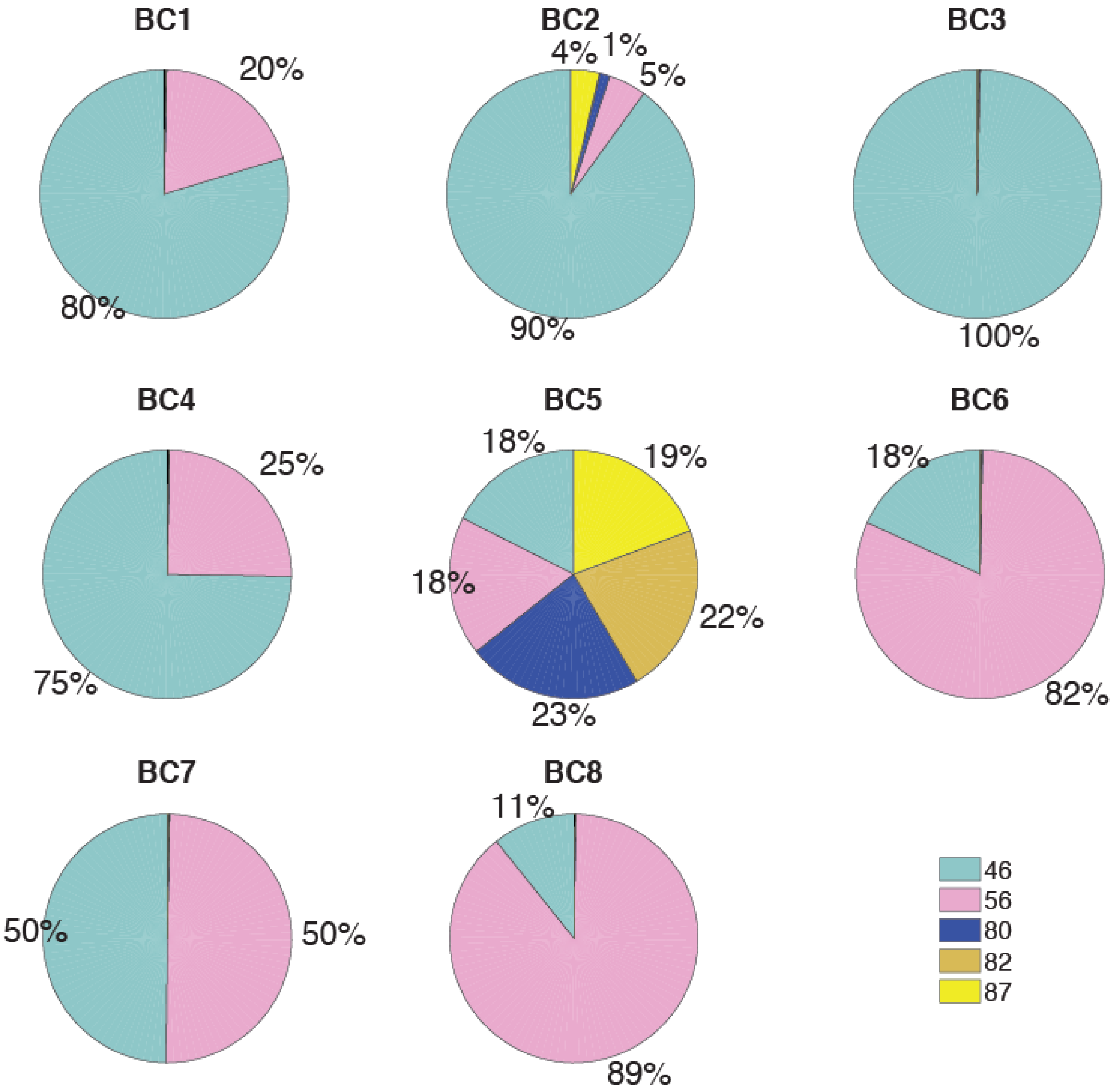


Figure 4.14: Pie charts showing proportion of the five T reg clusters in each patient, indicating that differences in covariance patterns between clusters also translate to patients.

i, i' across cells assigned to cluster k, is significantly different from that in all other clusters, we used bootstrapping and permutation testing as follows: We started by generating a null distribution for the covariance between a pair of genes by first uniformly sampling a subset of cells from all clusters, with the subset being the same size as cluster k. Then, to further remove existing structures of co-expression in cells, we permuted the cell labels for gene i' (while retaining labels for gene i) and computed empirical covariance between the two genes in this subset of "scrambled" cells. We repeated this on 10,000 subsets to achieve a null distribution of $cov(\overrightarrow{w_i},\ \overrightarrow{w_{i'}})$ where $w_i, w_{i'}$ are the expressions of gene i, i' in the sets of scrambled cells. We then compared the observed $cov(\overrightarrow{x_\mathrm{i}}, \overrightarrow{x_{i'}})$ (marked with a star in Figure S5A, S7A) to the null distribution, which was rejected for that pair of genes if p-value<0.05 indicating that the covariance is significantly different in cluster k compared to all other clusters.

We concluded that the signal is also apparent in raw un-normalized data for all the aforementioned clusters and we observe a range of covariance values with different signs between GITR and CTLA4 across T reg clusters (Figure S16), and similarly different values in covariance between MARCO and CD276 in TAM clusters (Figure S22).

4.6 Components of Variation of Intra-tumoral Myeloid Cells

Although myeloid lineage cells are commonly thought to be highly diverse and able to markedly influence the state of the tumor microenvironment and, thereby, impact clinical outcomes, the heterogeneity of intra-tumoral monocytes and Macrophages remains insufficiently characterized (Campbell et al., 2011; De Henau et al., 2016; Engblom, Pfirschke, and Pittet, 2016; Eppert et al., 2011; Gholamin et al., 2017; Pyonteck et al., 2013). A broad survey of the major monocytic subsets suggests the existence of both gradual and abrupt phenotypic shifts (Fig-

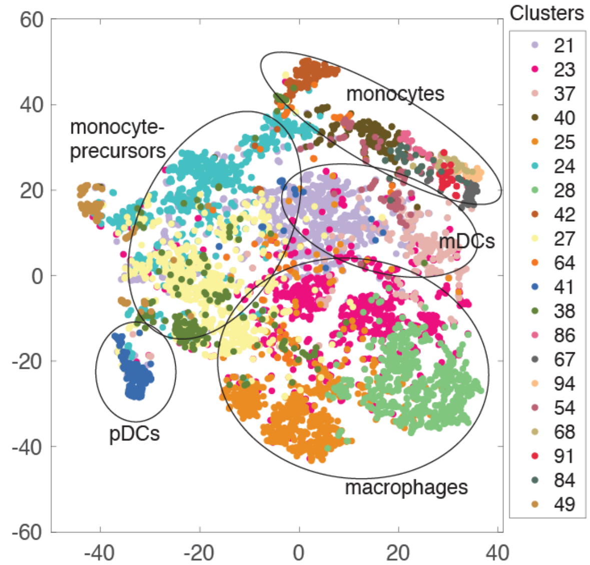


Figure 4.15: t-SNE map projecting only myeloid cells across all tissues and patients. Cells are colored by Biscuit cluster and cell types are circled and labeled based on bulk RNA-seq correlation-based annotations.

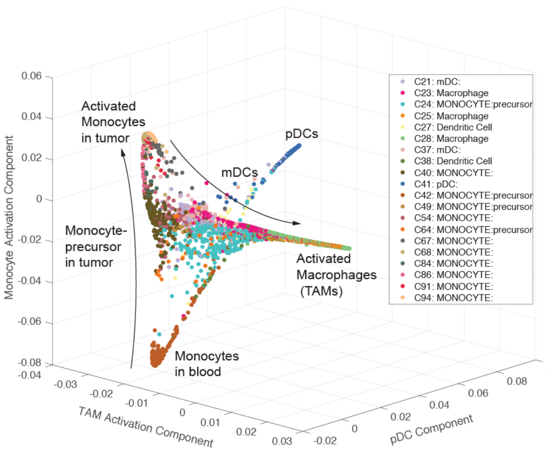


Figure 4.16: Projection of cells in myeloid clusters on Macrophage activation, pDC, and monocyte activation (first, second, and fourth) diffusion components. Cells are colored by cluster

ure 4.15).

As with the T cells above, we employed diffusion maps to assess heterogeneity in and across these monocytic populations, excluding neutrophils and mast cells, which formed separated clusters and were therefore better assessed through other techniques (Figure 4.16). This analysis revealed four major branches that displayed clearer segregation of cell states, and moderately less continuity, than the analogous T cell maps (Figure S19).

The first branch almost entirely comprises intra-tumoral Macrophages from three clusters (23, 25, and 28) (Figure4.17). Among the top genes correlated with the branch were are Macrophage activation-associated genes APOE, CD68, TREM2, and CHIT1 (Figure S20); the branch, thus, likely reflects progression towards a distinct state resulting from the differentiation and activation of either recruited or tissue-resident Macrophages in the tumor microenvironment (TME) (4.18.

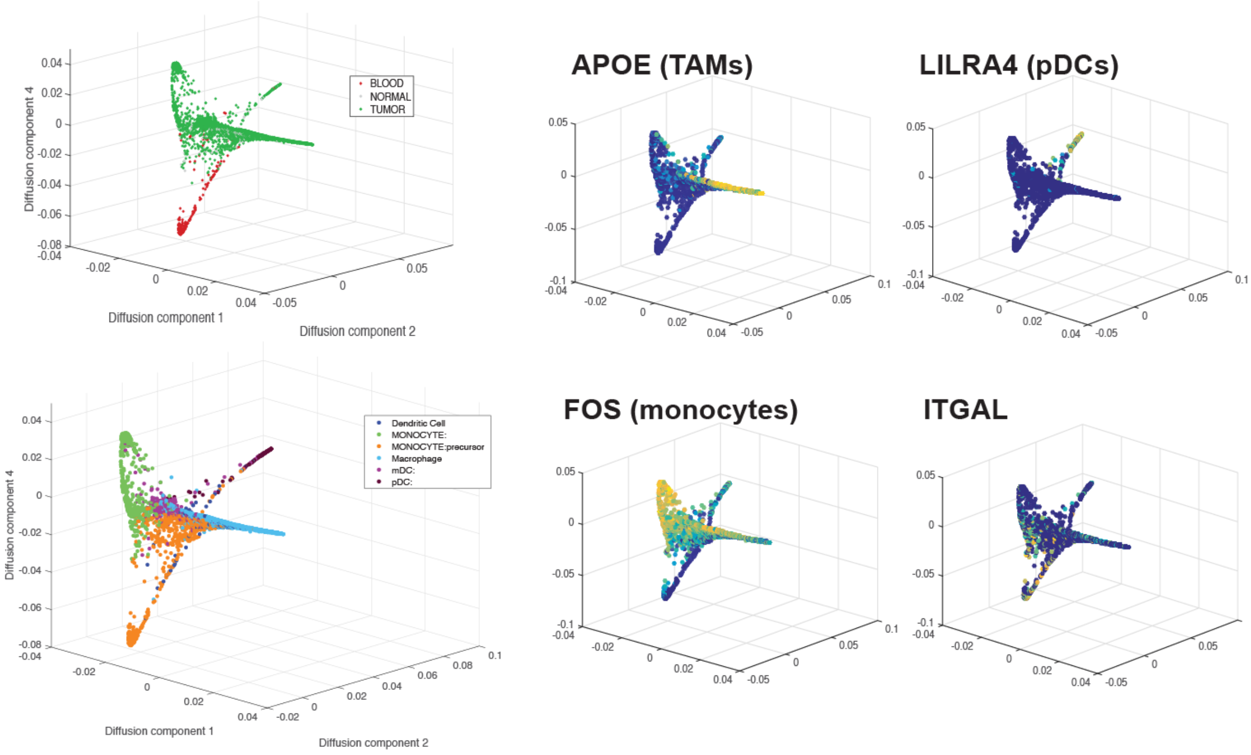


Figure 4.17: Projection of cells in myeloid clusters on Macrophage activation, pDC, and monocyte activation (first, second, and fourth) diffusion components. Cells are colored by (B) cluster, (C) tissue type, (D) cell type (as explained in STAR Methods), and (E) expression of example lineage demarcating genes. The main trajectories are indicated with arrows and labeled in (B).

Additionally, expression of genes typically implicated in a polarization model of tissue-reparative and immunosuppressive M2 Macrophage activation, including scavenger receptor MARCO, extracellular matrix component FN1, pro-angiogenic receptor NRP2, SPP1 (osteopontin), and inhibitory molecule B7-H3 (CD276), increased along this branch (Figure S20). Concomitantly, pro-inflammatory and immunostimulatory genes, including chemokine CCL3 (MIP-1a), typically associated with M1 Macrophages likewise increased along the branch.

Quite strikingly, we found that M1 and M2 gene signatures were positively correlated in the myeloid populations (Figure 4.19). These findings support the idea that Macrophage activation is markedly impacted by the tumor microenvironment in a manner that does not comport with

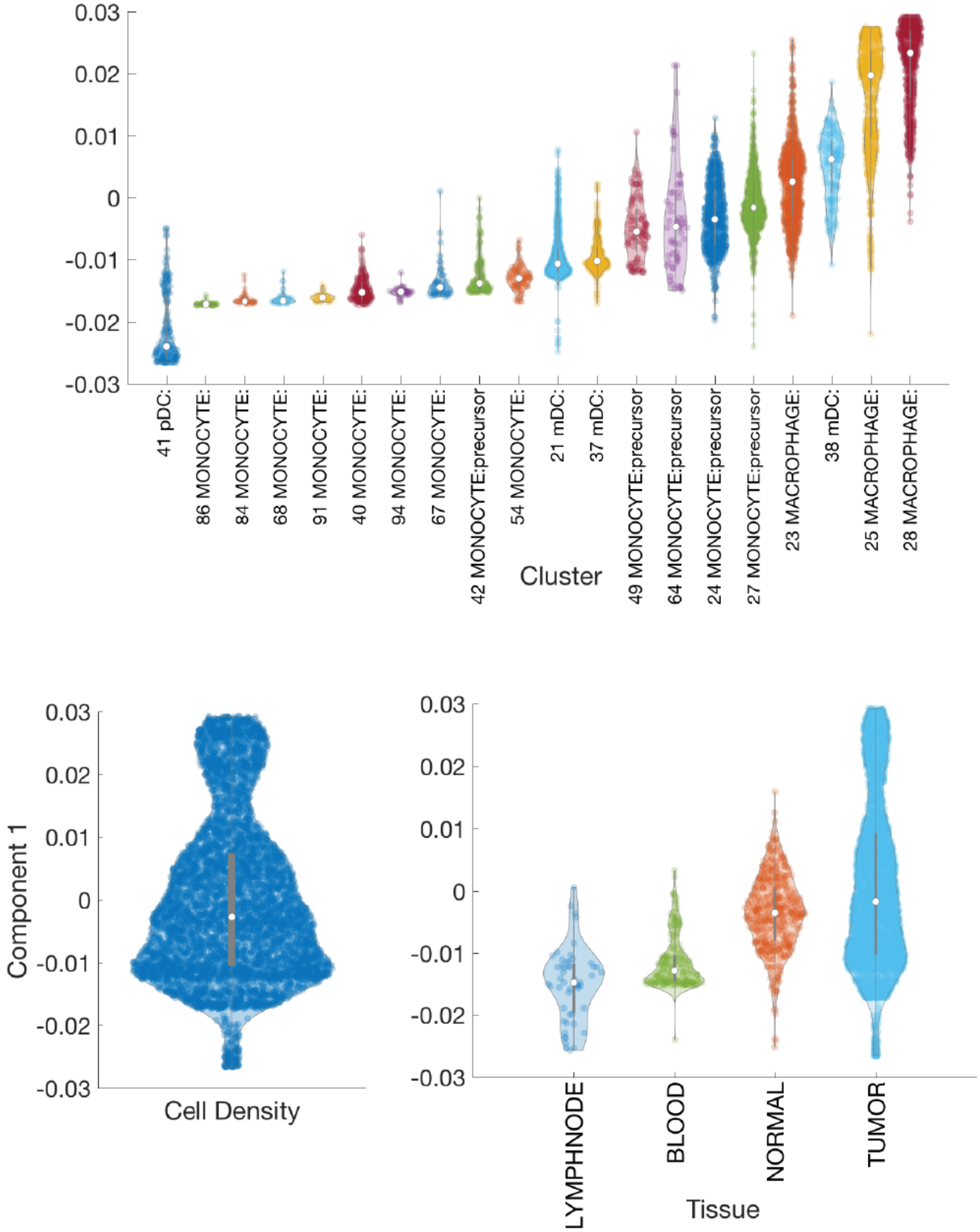


Figure 4.18: Violin plots showing the density of cells along Macrophage activation component and organized by overall density (left panel), tissue type (middle panel), and cluster (right panel).

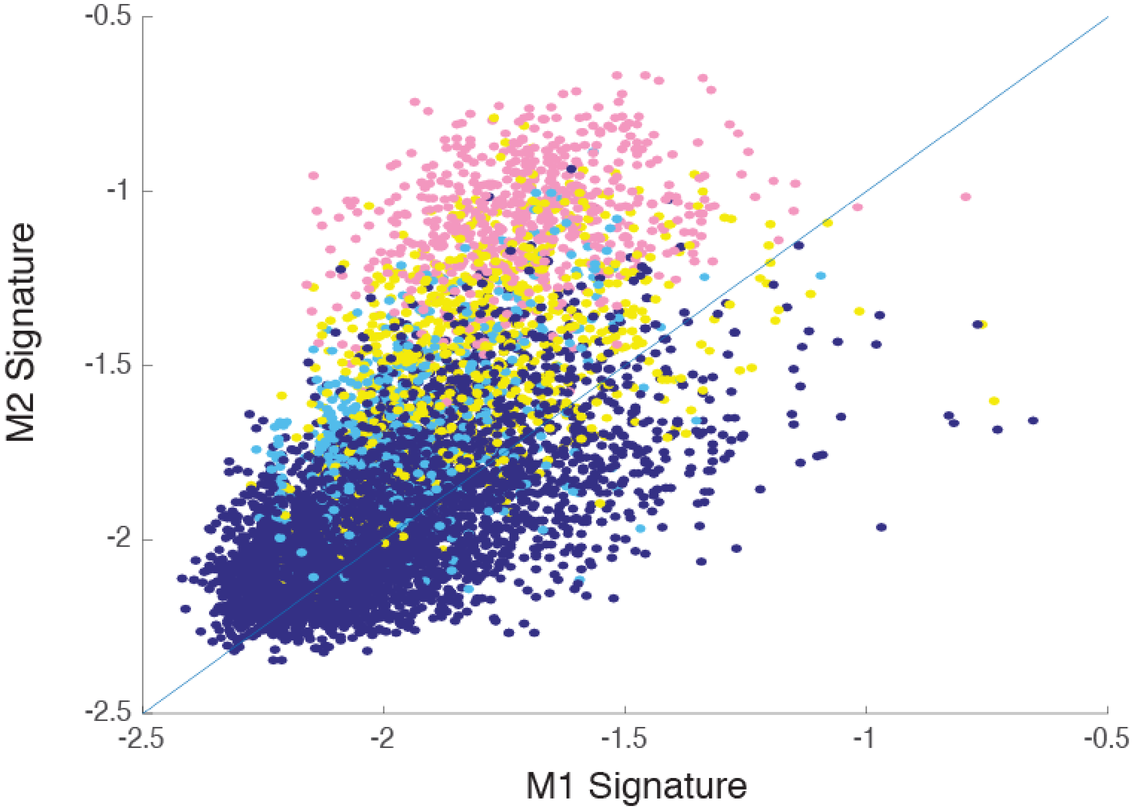


Figure 4.19: Scatter plot of normalized mean expression of M1 and M2 signatures per cell (dark blue); cells assigned to 3 TAM clusters have been highlighted by cluster (light blue, pink, yellow); each dot represents a cell and cells are plotted in randomized order.

the polarization model, either as discrete states or along a spectrum of alternative polarization trajectories.

The second and third branches together captured a more gradual trajectory from blood monocytes (mainly cluster 42, 97.5% present in blood) to intra-tumoral monocytes (clusters 67, 91, 68 and 94) (Figure 4.17). The "blood terminus" of the trajectory correlated with expression of co-stimulatory gene ITGAL, but also with several tumor growth-promoting genes, i.e. fibroblast and epidermal growth factors, as well as IL-4 (Figure S20). The latter has been proposed to support the M2 type of Macrophage activation (Mantovani and Locati, 2013; Mills et al., 2000; Murray et al., 2014). The other end of the trajectory, populated by intra-tumoral monocytes, was characterized by high expression of activation and antigen presentation-related genes encoding CD74 and HLA-DRA, but also an IFN-inducible gene encoding ISG15, which has been described to be secreted by TAMs and enhance stem-like phenotypes in pancreatic tumor cells (Figure S20) (Sainz et al., 2014).

The fourth branch correlated with canonical plasmacytoid dendritic cell (pDC) markers such as LILRA4, CLEC4C (CD303), and IL3RA. The most discrete of the myeloid components, this branch separated the lone pDC cluster (41) from the other myeloid-monocytic cell clusters (Figure 4.16,4.17,4.18,S21) This subset was also the only monocytic cluster common between the tumor and the lymph node; it featured high levels of granzyme B (GZMB) (Figure S20), which has been proposed to be a means, by which pDCs may suppress T cell proliferation in cancer (Jahrsdörfer et al., 2010; Swiecki and Colonna, 2015). These results highlight how diffusion maps can be used to uncover major sources of variance in heterogeneous data, and how analysis of those components can inform us of the biological signals of greatest importance.

4.7 Covariance Patterns Help Distinguish TAM Subpopulations

While the TAM clusters projected to a distinct region in the diffusion component, separating them from other monocytic cells, they appeared very similar to one another (Figure 4.20, S19). This similarity was supported at the genomic scale by shared pattern of differentially expressed genes (Table S3) and short pairwise distances (Figure S10). However, similarly to the intra-tumoral T reg cells, co-variation patterns defined distinctions between intra-tumoral myeloid cell subsets. Specifically, co-variation of canonical genes for M1 or M2 Macrophages distinguished the TAM clusters. All three of the TAM populations, particularly clusters 23 and 28, were among

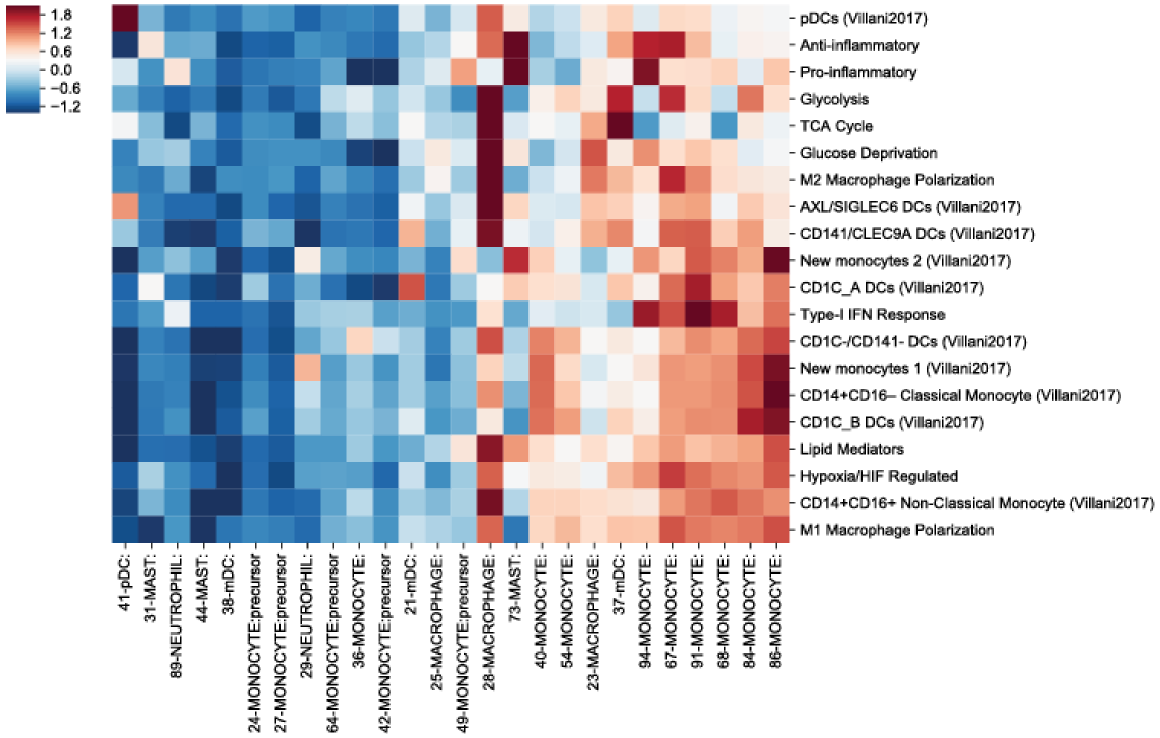


Figure 4.20: Heatmap showing imputed mean expression levels in myeloid clusters for a curated set of transcriptomic signatures important to myeloid cells (listed in Table S4), z-score normalized per signature. See also Figure S6 for additional violin plots and Heatmaps representing the other components.

the monocytic lineage clusters that exhibited the most similarity to the canonical M2 signature (Figure4.19). However, both of these clusters also expressed high levels of the M1 signature genes, and significant expression of the two signatures was often coincident (Figure 4.19, 4.20).

We observed pronounced inter-cluster differences in co-expression patterns in TAM clusters. One example among many was co-expression of two M2-type markers, MARCO and B7-H3. In an unexpected manner, while TAM clusters 23, 25, and 28 all expressed high levels of both genes, they co-varied positively in clusters 23 and 25, but negatively in 28 ($p = 0, p = 5e - 06, p = 0$, respectively; (Figure 4.21, 4.22;S22). The differing covariance patterns were not an artifact of

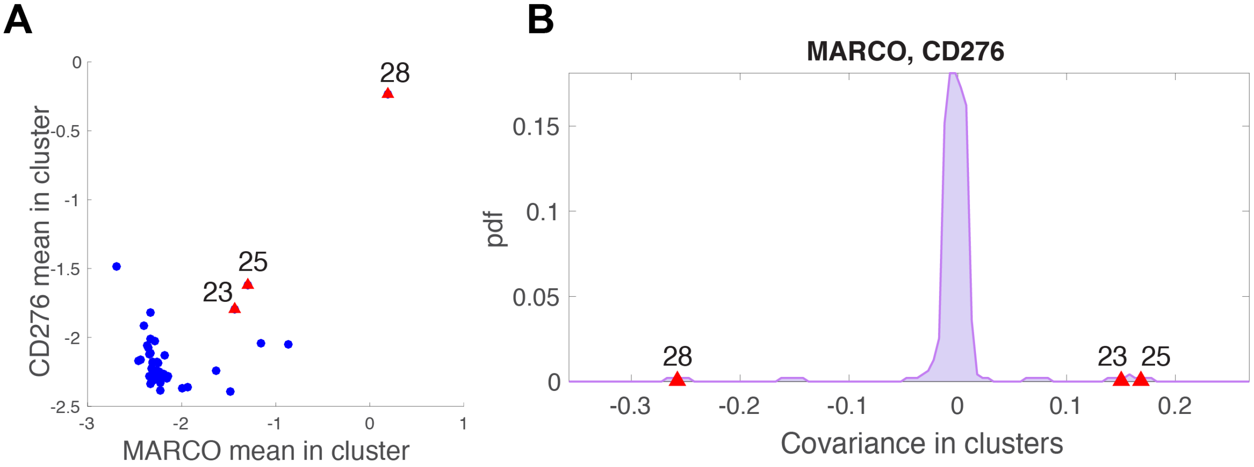


Figure 4.21: Scatterplot of mean expression of MARCO and CD276 in each myeloid cluster; each dot represents a cluster. Average expression levels for the three TAM clusters (23, 25, and 28) are marked in red, indicating high expression of both markers in Macrophage clusters. Distribution of covariance between MARCO and CD276 across all myeloid clusters. TAM clusters(23, 25, and 28) are marked in red and present substantial outliers. See Figure S7A for similar computation on the raw, un-normalized, and un-imputed data, verifying the result.

modeling as they were also significant in raw un-normalized data (Figure S22).

The degree of co-expression of genes associated with M1 and M2 signatures also varied widely within clusters in a manner not fitting the functional M1/M2 annotation. For example, in cluster 23 expression of CD64 exhibited varying degrees of positive co-variance with FN1, MMP14, MSR1, cathepsins, MARCO, and VEGFB, but co-varied slightly negatively with chemokine CCL18 (Figure 4.22). Taken together, these findings demonstrate that co-variation patterns define TAM clusters, and further highlight the lack of mutual exclusivity between the proposed prototypical M1 and M2 states.

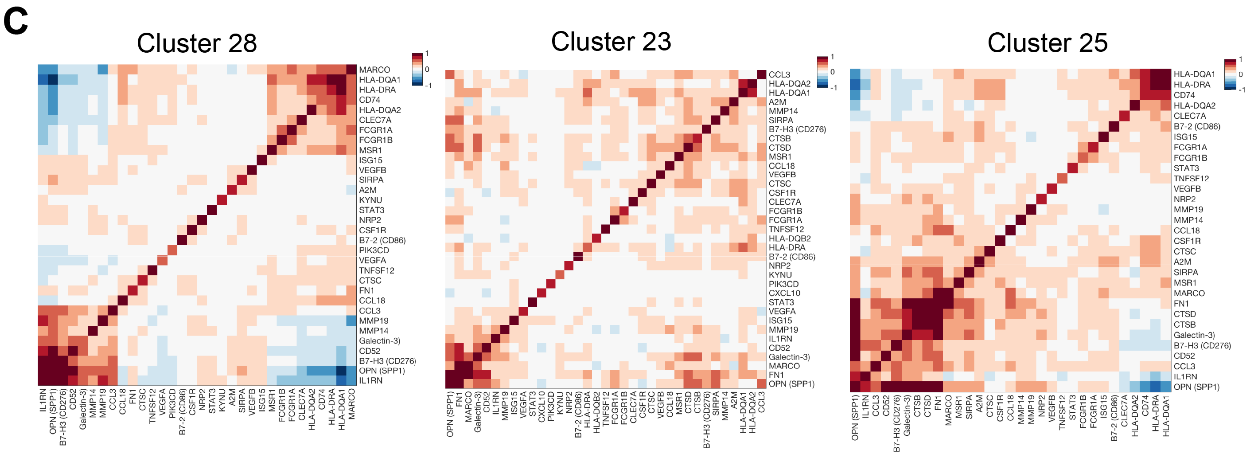


Figure 4.22: Heatmaps showing covariance patterns of M1 and M2 Macrophage polarization marker genes (including many current or potential drug targets) in 3 TAM clusters (23, 25, and 28).

Chapter 5

Discussion and Perspectives

5.1 Implications of the Breast Tumor Immune Atlas

Despite major clinical advances in cancer immunotherapy, our ability to understand its mechanisms of action or predict its efficacy is confounded by the complex, heterogeneous, and poorly understood composition of immune cells within tumors. Since cancer is generally a disease that affects older, post-reproductive individuals, with the exception of inherited genetic predisposition, it is unlikely that specialized mechanisms of the adaptive or innate immunity evolved to facilitate tumor immune surveillance. It seems reasonable to suggest that immune mechanisms affecting tumor progression must also operate in non-cancerous tissues to maintain organismal integrity and tissue function in the face of infection, stress, inflammation, and injury. A corollary to this notion is that features of immune cells in tumors must, by and large, resemble features of cells in non-cancerous tissues. A recent population-level RNA-seq analysis of T Reg cells and effector CD4 T cells in breast cancer and normal breast tissue identified a high level of phenotypic similarity between tissue and tumor-resident T cells, thus providing experimental support for this idea (Plitas et al., 2016). A similar RNA-seq study focusing primarily on T Reg cell analysis in colorectal and lung cancer suggested that cancer-resident T Reg cells differ considerably from those found in the normal tissue (De Simone et al., 2016). Despite seeming differences in conclusions, distinguishing features of intra-tumoral T Reg cells as compared to normal tissue-resident ones detected in these two reports were associated with their heightened activation and thus can not distinguish between differences in immune states themselves or differences in immune state proportions. Thus, the averaging of gene expression features in the bulk cell population analyses and the lack of the assessment of a broad spectrum of immune cell subsets do not allow for a definitive investigation of specific effects of the tumor environment on immune cells.

To address this question, we characterized the available single-cell approaches and selected and adapted InDrop, the best-suited single-cell method to assaying rare populations of immune cells. With InDrop, we undertook an unbiased comparative single-cell RNA-seq analysis of all tumor versus normal tissue-resident immune cell subsets and constructed an comprehensive immune atlas in breast carcinomas, combining immune cells isolated from normal and cancerous breast tissue, as well as peripheral blood and the lymph node. Our analysis was empowered by a suite of novel computational tools for single-cell RNA-seq data, including a data processing pipeline more sensitive in its ability to detect immune molecules, a powerful clustering and normalization algorithm, and new metrics for volume of the phenotypic space. These secondary analysis methods allowed us to overcome significant technical artifacts, correct for amplification biases, eliminate spurious molecules generated by library construction, recover molecules that would otherwise be hidden by multiple alignment, and select and eliminate problematic cells that were undergoing apoptosis, expressing transcripts consistent with technical stressors, or whose profiles represented undesirable cell types that had escaped flow sorting. Our novel tertiary methods facilitated clustering and normalization in the face of the strong batch effects typical of clinical samples that would have otherwise dominated the signal and obscured the identification of shared cell states across tumors. Additionally, the BISCUIT model facilitated an in-depth characterization of variance within and between identified cell states, a goal that had been at best partially realized in single-cell analysis.

The constructed atlas revealed vast diversity in the repertoire of immune cells representative of both the adaptive and innate immune systems. Our examination of hematopoietic nucleated cells from treatment-naïve human breast cancer and normal breast tissue across different patients revealed that the biggest change to the immune cells was linked to the tissue environment, resulting in cell states that are substantially different than those present in the blood and lymph node. Interestingly, immune cell subpopulations in normal tissue were observed to be a subset of those found in tumor tissue, an observation that could not have been found with bulk gene expression measurements. Furthermore, the diversity of cell states significantly expanded between normal tissue and tumor, as quantified by the "phenotypic volume" occupied by immune cell states. We observed tremendous expansion of the immune phenotypic space occupied by all major cell types in breast tumors as compared to normal breast tissue. It seems reasonable to speculate that the majority, if not all immune cell states found in cancer can be found in corresponding non-cancerous tissues in response to different stresses such as infection, wound healing, or inflammation.

The observation of an expanding continuous T cell "phenotypic space" in the tumor argues against the view of activated T cells rapidly traversing through sparse transitional cell states towards a few predominant, discrete, and stable states, including T Reg, effector, memory, and exhausted T cells. Three major components contributed to this phenotypic expansion in tumor tissue that helped explain the heterogeneity of T cells contributing to this phenotypic expansion in tumor tissue, we identified three major components that help explain the heterogeneity of T cells, including T cell activation, terminal differentiation, and hypoxic response. The strongest of these components is a predominant trajectory of progressive T cell activation and differentiation across 38 T cell clusters, including T Reg and terminally differentiated T cell clusters, found at the extreme activation terminus. One obvious explanation for the "continuity" of intra-tumoral T cell activation is the presence of increasingly diverse environments defined by a multitude of gradients including growth, pro-inflammatory, and tissue repair factors, as well as oxygen, nutrient, and metabolite gradients which exist to a lesser extent in healthy breast tissue (Buck et al., 2017). Indeed, we found groups of genes within the corresponding signaling pathways, most prominently immune activation (IFN/IL6/JAK/STAT) and hypoxia, to be differentially expressed across T cell clusters.

A non-mutually exclusive possibility is that the wide range of TCR signal strengths afforded by a diverse repertoire of T cell receptors (TCR) accounts for the continuous spectrum of T cell activation, obscuring the transitional states. The latter may also be accounted for by asynchrony in polyclonal T cell activation or heterogeneity in the types of antigen-presenting cells, their activation status, and their anatomical distribution. Unlike polyclonal T cell populations, activation of a monoclonal T cell population with a "fixed" specificity for tumor "self" or neo-antigen may yield sparse discontinuous "phenotypic" spaces reflecting discrete functional T cell states. In support of the latter possibility, recent bulk gene expression and chromatin accessibility analyses showed that cognate tumor neo-antigen recognition by TCR transgenic T cells results in an orderly progression of activated T cells through a reversible dysfunctional intermediate state towards an irreversible dysfunctional terminal state (Philip et al., 2017). Additionally, diverse TCR specificities are known tocan contribute to spatial distribution of T cells and, therefore, facilitate their exposure to the distinct environments ("mini-niches") discussed above.

While T cells of various cell types exhibit continuous levels of activation, our inferred subsets further show variable levels of responses to environmental stimuli, and the combinations of these environmental exposures jointly define the identity of discrete CD4+/CD8+ T cell subsets. We also identified 5 T Reg subsets that showed similar responses to environmental pressures and shared differentially expressed genes, but exhibited drastic differences in gene covariance patterns. Particularly noteworthy was co-expression of checkpoint receptor genes in some T Reg subpopulations as compared to mutually exclusive expression of the same genes in other T Reg clusters. In this regard, co-variant expression of CTLA-4, TIGIT and co-stimulatory receptor GITR and other co-receptors in multiple T Reg cell clusters suggests that these T Reg cell populations may occupy different functional niches; CTLA-4 and TIGIT co-expressing cells have been demonstrated to selectively inhibit pro-inflammatory Th1 and Th17, but not Th2 responses promoting tissue remodeling (Joller et al., 2014). The observed co-expression of functional cell surface and signaling molecules by intra-tumoral T Reg cells may enable targeted modulation of T Reg cell activity in the tumor microenvironment using combinatorial therapeutic approaches (Mantovani and Locati, 2013). This finding has implications in the way we describe and interrogate the tumor immune response. It is also noteworthy that the discrete cell states that are commonly utilized to describe immune responses are largely defined from highly polarizing conditions such as infection and tissue injury.

Our analyses appear to offer a more nuanced view of tumor and normal tissue-resident myeloid lineage cells, in comparison to T cells, in terms of continuity vs. separation of cell states. Unlike T cells, which primarily displayed continuous activation transitions, we observed sharper state delineations in myeloid populations. This difference between T cells and myeloid cells was likely due to a markedly less appreciated developmentally established myeloid cell heterogeneity, whose understanding has started to emerge only recently (Perdiguero and Geissmann, 2016). Indeed, the phenotypic expansion in myeloid cells was associated with activation of macrophages and monocytes and emergence of pDC subsets distinct from cDCs. However, our analyses also showed common features to those in T cells, including gene expression covariance identifying cell clusters, and an expansion of immune phenotypic space in breast tumor as compared to normal breast tissue.

Similarly to T cells, we have not observed discrete states of myeloid cell activation/differentiation such as M1 or M2 macrophages or myeloid derived suppressor cells. In contrast, we found both M1 and M2 associated genes frequently expressed in the same cells, positively correlated with one another and following the same activation trajectory. Furthermore, we found that covariance patterns between gene markers associated with the M1 and M2 model show rich diversity, and help distinguish the three TAM clusters. These results challenge the prevailing model of macrophage activation, wherein M1 and M2 activation states either exist as mutually exclusive discrete states or macrophages reside along a spectrum between the two states with a negatively correlated expression of M1 and M2-associated genes. Our findings solidify and reinforce previous reports from the bulk analysis of tumor-associated macrophages in mouse models of oncogene-driven breast cancer and analysis of myeloid cells in lung and kidney cancer using mass-cytometry (Chevrier et al., 2017; Franklin et al., 2014; Lavin et al., 2017). Notably, we observed more patient-specific variation in myeloid lineage cells than in T cells, with the frequency of the former ranging from just over 10% to over 50% in individual patients. Individual clusters similarly exhibited ranges of patient specificity. The large patient effect in myeloid cells suggests that attempts at generalized targeting or reprogramming of suppressive myeloid cell populations are not likely to yield uniform responses and personalization at the patient level may be needed.

Thus, our findings show that studying average gene expression across groups of cells fails to characterize heterogeneity in co-expression of genes, and by extension their potential suitability as therapeutic co-targets. Single-cell RNA-sequencing analyzed using Biscuit, as shown here, allows for inference of accurate and meaningful covariance parameters — indeed, the algorithm takes into consideration these covariance values when defining clusters. This makes it possible to query in a precise manner how numerous functionally and therapeutically important immune markers are co-expressed at the level that matters: that of individual cells. Our characterization of the immune cell subsets inhabiting primary solid tumor and the corresponding normal tissue, and their heterogeneity within a given patient and between different patients revealed expansions of a continuous "phenotypic space" as a principal feature of the two main cellular targets of cancer immunotherapy - T cells and myeloid cells. These observations, along with the resulting extensive immune single- cell RNA-seq datasets and the comprehensive analytical platform, will facilitate better knowledge understanding of potential mechanisms behind immune cell contributions to promoting and opposing tumor progression.

5.2 Validation & Follow-up Experimentation

The logical follow-up of this atlasing project is to confirm that the results observed in our experiments exclusively result from biological signals and not technical aspects of scRNA-seq. Our observations at the cellular level of the correspondence of the FACS-sorted populations to the observations in scRNA-seq are suggestive that our analyses mirror the biology represented in the tissues, albeit with reduced representation of small-volume cells. While not all studies carry out independent validations, additional confirmations of the intra-cellular covariance observations are warranted. Previous studies (Shalek et al., 2013; Shalek et al., 2014) have focused on microsocopy-based confirmations, using either single-molecule fluorescence in-situ hybridization (FISH), immuno-histochemistry, or immuno-fluorescence approaches to measure RNA or protein abundances. These approaches are problematic to apply to our data. Unlike many previous studies, we are working with rare immune isolates from complex patient tissues, rather than tumors, cell lines, or mouse models, each of which contain plentiful cells of the type under study. Our most interesting hypotheses, about T Regulatory cell co-expression, also occur in the rares cell types under study (Figure 3). These problems combine to make it un-economical to detect an adequate number of cells through microscopy-based approaches. Instead, since the majority of our findings center on surface proteins, we believe that FACS sorting of T Regulatory cells based upon CTLA-4, TIGIT, and GITR, the markers with differential covariance, would be a suitable approach to confirm the functional nature of our discoveries. As such, we are in the process of profiling several additional patients with FACS, and expect to observe patient-specific co-expression differences: while all patients are expected to display T Regulatory cells with CTLA-4, TIGIT, and GITR, we expect that co-variance will differ across patients.

A second question raised by this experiment is the cause of the observed T cell diversity. If the observed diversity is indeed correlated with TCR repertoire, the phenotypic volume of T cells should increase with TCR diversity. Within weeks of the completion of the studies described in this dissertation, 10x genomics released a kit that allows the simultaneous profiling of the 5' transcriptome and the TCR. We will apply this kit to the FACS-sorted cells from the above patients, and will be interested to see if the TCR diversity correlates with phenotypic volume, or if our observations were indeed instead driven by the diversity of signaling molecules and gradients present within the tumor microenvironment. In other experiments, we are characterizing two TCRs with known epitope targets from a mouse cancer model using scRNA-seqthese experiments should provide convergent evidence: if significant differences in expression are observed between the epitope-specific TCR clones, this will support the idea that TCR functionality may be responsible for siginficant variation in T cell mRNA expression profiles.

Taken holistically, these experiments highlight the incredible diversity of tumor microenvironments; tumors bombard immune cells with diverse panels of cytokines, chemokines, and growth factors. Our observations strongly suggest that combinations of these stimuli are responsible for the diversity of observed phenotypic profiles in TILs. Thus these results suggest that prediction of phenotypic profiles and checkpoint expression, and therefore druggability of T Regulatory or t-effector cells, depends extensively upon characterization of T cell responses to complex cocktails of stimuli. The diversity of cancers would require profiling of a very large number of tumors, and may thus place this predictive goal out of reach for some time. Nevertheless, it may still be possible to profile target cells, and by observing their checkpoint expression, enable immune functionality despite not understanding how the cells came to express the particular set of checkpoint markers they are presenting. Because droplet-based scRNA-seq requires cells be dissociated to flow through the encapsulation devices, these experiments cannot directly observe the microenvironment of the individual cells that we infer to contribute significantly to their phenotypes. However, an advantage of scRNA-seq is that we are able to measure many thousands of cells per experiment, and as such, if environmental characteristics of a cell could be quantified, and cells could be stratified in terms of similar environments, it would create an immediate and powerful method through which to quantify the effect of tissue microenvironment on each cell type under study. Perhaps more critically, while we were able to confirm the presence of significant immune infiltrate in each of our patients using immuno-histochemistry, we cannot distinguish cells that were directly in contact with tumor cells from those restricted to the periphery of a tumor or those in contact only with stromal or other immune cells. If differences in the phenotypes of these different spatial contexts and the abundance of these cells could be posed as independent variables, it would be of great interest to correlate with drug response outcomes in clinical trials. There are emerging approaches capable of spatially profiling cells' transcriptomes and proteomes, and I believe that large-scale application of these technologies to patient tumors is both a logical next step, and an exciting opportunity to make rapid progress in understanding cancer immunology after controlling for these factors.

5.3 Implications of the SEQC Framework and Future Directions

In addition to providing the basis for this immune atlas, SEQC also served as the processing framework for several other published studies. Its filters were used to generate interpretable data in a Fluidigm C1-based study wherein cells were cultured inside the C1 device, enabling single-cell live imaging of NF-κB activation dynamics following LPS stimulation (Lane et al., 2017). SEQC was also used to generate the data that was used to develop the MAGIC (Dijk et al., 2017) BISCUIT (Prabhakaran et al., 2016; Azizi et al., 2017) and Wishbone (Setty et al., 2016) algorithms. Early iterations of the processing methods were used to analyze microplate-based single-cell sequencing data (Bose et al., 2015), and it is used in active production by Memorial Sloan Kettering Cancer Center to process the institute's droplet-based InDrop, 10x, and Nucleus-sequencing data. In addition, it has been used in at least one instance to process Drop-seq, 4 iterations of InDrop chemistry, and 3 versions of Mars-seq chemistry. In total, SEQC has processed over 250 datasets from diverse chemistries, tissues and multiple organisms. Finally, SEQC's demonstrated modularity and flexibility provoked the Human Cell Atlas (Regev et al., 2017) to adopt the SEQC framework as the first draft of the 3' analysis pipeline that it will use in the secondary analysis of droplet-based sequencing data. The Human Cell Atlas project is expected to generate data at petabyte scale, a task made simple by SEQC's ability to scale using the Cloud.

The development of SEQC, and its ability to deliver clean cellular phenotypes in spite of significant experimental noise, was instrumental in enabling he constribution of the tumor atlas. However, it is important to remember that the atlas would not have been possible at all without droplet-based sequencing. Underlying the data generated in this dissertation were 2 major barcode redesigns, made possible in part by analysis metrics generated by InDrop providing clear feedback to experimenters about what aspects of the technology required improvement. The Pe'er lab has since iterated through 3 major changes to the InDrop chemistry, and we expect more will follow.

Before the SEQC framework was mature, the ability to process data would often lag weeks behind the development of new library construction approaches, significantly retarding our ability to make adjustments that would facilitate analysis of primary immune cells. Initially, the technologies were changing much more quickly than the computational approaches. With the mature SEQC pipeline, we are now normally able to iterate and produce appropriate computational approaches for changes in library construction within a few weeks. In cases where appropriate algorithms exist, we can often make necessary changes in hours.

Also symptomatic of this problem, novel technologies are often published with data processing tools that are inadequate, or at least sub-optimal. For example, sNuc-Seq (Habib et al., 2016) and DroNc-seq (Habib et al., 2017) are exciting methods that sequence nuclei instead of complete cells. This difference promises to enlarge the scope of samples that can be processed with scR-

NAsq, since nuclear membranes are more robust than cell membranes, frozen sample or samples with more degradation can be sequenced with these methods.

However, the data generated by these approaches were processed using tools designed for full-cell sequencing, ignoring the fact that many RNA in nuclei are found in pre-spliced "pre-mRNA" form and would therefore contain intronic reads, which are discarded by full-cell analysis pipelines. This introduces a data loss of 15-30%, and strongly indicates that the technologies are still iterating much more rapidly than the computational methods. This highlights that in addition to a pipeline for processing data, the field would benefit from a framework for rapidly mixing and matching algorithms that consume and produce standard sequencing data types and file formats.

To improve computational iteration speed, SEQC and other frameworks like it must be made more portable and trivial to use. Because different labs and institutions have back-end compute server frameworks that can be idiosyncratic and utilize queueing architectures that are often incompatible with one another, this is currently difficult to achieve. While SEQC has unprecedented flexibility, it is still limited to being run on single physical or virtual machines, and can only easily be run on laptops, desktops, local compute servers or on Amazon web services. There are technologies being developed which will solve these problems, and improve method portability not just for processing frameworks like SEQC, but also increase the portability of analysis algorithms like BISCUIT, MAGIC, and Wishbone.

The first of these technologies are container services like Docker[1]. Docker is a relatively mature scripting framework for constructing a lightweight software image that contains all of the installed dependencies for an particular software package, including the OS layer, and the software package itself. Docker images support versioning, meaning that a properly constructed

[1] see https://github.com/docker/docker-ce

docker image should go on working forever, regardless of what changes may happen to the greater ecosystem of the programming language or advancements in packages it depends upon. Docker is used extensively by companies to package software to make it more robust, reliable, and easier to ship to customers.

However, this approach has not made much of an inroad in academia, perhaps because of a higher focus on publication than eventual usability. SEQC was initially programmed to follow a strategy similar to docker, but with some limitations. We first programmed an Amazon Machine Image, which is like an Amazon-specific docker container. As a result, it only worked on Amazon's Elastic Cloud, and we needed to use separate installation instructions for local installations.

Docker, by contrast, can run on Amazon, but also on Google cloud, and on mac and PC. For these reasons, SEQC has been made available in a docker container, which can be pulled from the dockerhub repository at ambrosejcarr/seqc:1.0.0. As a result, SEQC can now be used on any operating system, without needing expert knowledge of their operating system to install any dependencies.

Because containers allow a developer to easily produce ready-to-run container, if methods consume and produce standard data types, such as BAM and FASTQ formats, then multiple methods that accomplish the same task can be benchmarked and interchanged, and these pieces can be woven together through workflow languages. Two workflow languages, Common Workflow Language (CWL) and Workflow Design Language (WDL)[2] are being developed to serve this purpose, and these frameworks have been adopted by the Broad Institute, University of Santa Cruz, European Bioinformatics Institute, Chan Zuckerberg Initiative, and Human Cell Atlas, among others, to serve as a framework for building open software that can benefit the community.

[2] see http://www.commonwl.org/ and https://github.com/openwdl/wdl

A framework based upon SEQC is presently being written in WDL, and once complete, its computational approaches will serve as a standard against which alternative algorithms may be benchmarked. Through creative use of control datasets, I hope to discover the optimal combination of computational approaches that produce the highest quality sequencing data for, at first, 3' droplet-based sequencing approaches. Nucleus sequencing, 10x Genomics, and other approaches all utilize a common set of core methods, and as such, the InDrop backbone can be adapted to suit other technologies. It is my hope that by refining this pipeline and pairing it with any number of front-end suites of analysis tools, we can speed up the experiment, analyze, refine loop for advances in single cell sequencing, enabling faster technological development, opening doors to generate many more cell atlases like the one described here, each with clear clinical implications.

Bibliography

Adam Best, J et al. (2013). “Transcriptional insights into the CD8+ T cell response to infection and memory T cell formation.” en. In: *Nat. Immunol.* 14.4, p. 404.

Aho, A V and J D Ullman (1983). “Data structures and algorithms.” In:

Alted, Francesc (2010). “Why modern CPUs are starving and what can be done about it.” In: *Computing in Science & Engineering* 12.2.

Alted, Francesc, Ivan Vilata, et al. (2002–). *PyTables: Hierarchical Datasets in Python.* URL: http://www.pytables.org/.

Anders, Simon and Wolfgang Huber (2010). “Differential expression analysis for sequence count data Genome Biology, 11.” In: *R106.*

Azizi, Elham et al. (2017). “Single-Cell Immune Map of Breast Carcinoma Reveals Diverse Phenotypic States Driven by the Tumor Microenvironment.” In: *bioRxiv*, p. 221994.

Bavarian-Nordic (2017). *A Randomized, Double-blind, Phase 3 Efficacy Trial of PROSTVAC-V/F +/- GM-CSF in Men With Asymptomatic or Minimally Symptomatic Metastatic Castrate-Resistant Prostate Cancer (Prospect). Clinical Trial Identifier: NCT01322490.*

Beale, Elmus G, Brandy J Harvey, and Claude Forest (2007). “PCK1 and PCK2 as candidate diabetes and obesity genes.” en. In: *Cell Biochem. Biophys.* 48.2-3, pp. 89–95.

Belton, Jon-Matthew et al. (2012). “Hi–C: a comprehensive technique to capture the conformation of genomes.” In: *Methods* 58.3, pp. 268–276.

Bendall, Sean C et al. (2011). “Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum.” en. In: *Science* 332.6030, pp. 687–696.

Bendall, Sean C et al. (2014). “Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development.” en. In: *Cell* 157.3, pp. 714–725.

Benita, Yair et al. (2009). “An integrative genomics approach identifies Hypoxia Inducible Factor-1 (HIF-1)-target genes that form the core response to hypoxia.” In: *Nucleic Acids Res.* 37.14, pp. 4587–4602.

Bhattacharyya, A (1990). “On a Geometrical Representation of Probability Distributions and its use in Statistical Inference.” In: *Calcutta Statist. Assoc. Bull.* 40.1-4, pp. 23–49.

Biswas, Subhra K and Alberto Mantovani (2010). “Macrophage plasticity and interaction with lymphocyte subsets: cancer as a paradigm.” en. In: *Nat. Immunol.* 11.10, ni.1937.

Björklund, Åsa K et al. (2016). “The heterogeneity of human CD127(+) innate lymphoid cells revealed by single-cell RNA sequencing.” en. In: *Nat. Immunol.* 17.4, pp. 451–460.

Blackinton, Jeff G and Jack D Keene (2016). “Functional coordination and HuR-mediated regulation of mRNA stability during T cell activation.” In: *Nucleic acids research* 44.1, pp. 426–436.

Boon, Thierry et al. (2006). “HUMAN T CELL RESPONSES AGAINST MELANOMA.” In: *Annu. Rev. Immunol.* 24.1, pp. 175–208.

Bose, Sayantan et al. (2015). “Scalable microfluidics for single-cell RNA printing and sequencing.” In: *Genome biology* 16.1, p. 120.

Bouzin, Caroline et al. (2007). “Effects of vascular endothelial growth factor on the lymphocyte-endothelium interactions: identification of caveolin-1 and nitric oxide as control points of endothelial cell anergy.” en. In: *J. Immunol.* 178.3, pp. 1505–1511.

Boyle, Alan P et al. (2008). “High-resolution mapping and characterization of open chromatin across the genome.” In: *Cell* 132.2, pp. 311–322.

Brahmer, Julie R et al. (2010). “Phase I study of single-agent anti-programmed death-1 (MDX-1106) in refractory solid tumors: safety, clinical activity, pharmacodynamics, and immunologic correlates.” en. In: *J. Clin. Oncol.* 28.19, pp. 3167–3175.

Bray, Nicolas L et al. (2016). “Near-optimal probabilistic RNA-seq quantification.” en. In: *Nat. Biotechnol.* 34.5, pp. 525–527.

Bronte, Vincenzo et al. (2016). “Recommendations for myeloid-derived suppressor cell nomenclature and characterization standards.” en. In: *Nat. Commun.* 7, p. 12150.

Butler, Andrew and Rahul Satija (2017). “Integrated analysis of single cell transcriptomic data across conditions, technologies, and species.” In: *bioRxiv*, p. 164889.

Campbell, Michael J et al. (2011). “Proliferating macrophages associated with high grade, hormone receptor negative breast cancer and poor clinical outcome.” In: *Breast cancer research and treatment* 128.3, pp. 703–711.

Caton, P W et al. (2010). “Metformin suppresses hepatic gluconeogenesis through induction of SIRT1 and GCN5.” In: *J. Endocrinol.* 205.1, pp. 97–106.

Cheadle, Chris et al. (2005). “Control of gene expression during T cell activation: alternate regulation of mRNA transcription and mRNA stability.” In: *BMC genomics* 6.1, p. 75.

Chen, Daniel S and Ira Mellman (2013). “Oncology meets immunology: the cancer-immunity cycle.” en. In: *Immunity* 39.1, pp. 1–10.

Chevalier, Mathieu F et al. (2017). “Immunoregulation of Dendritic Cell Subsets by Inhibitory Receptors in Urothelial Cancer.” en. In: *Eur. Urol.* 71.6, pp. 854–857.

Chevrier, Stéphane et al. (2017). “An immune atlas of clear cell renal cell carcinoma.” In: *Cell* 169.4, pp. 736–749.

Chifman, Julia et al. (2016). “Conservation of immune gene signatures in solid tumors and prognostic implications.” In: *BMC cancer* 16.1, p. 911.

Chtanova, Tatyana et al. (2005). “Identification of T Cell-Restricted Genes, and Signatures for Different T Cell Responses, Using a Comprehensive Collection of Microarray Datasets.” en. In: *The Journal of Immunology* 175.12, pp. 7837–7847.

Chung, Woosung et al. (2017). “Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer.” In: *Nature Communications* 8.

Cock, Peter JA et al. (2009). “The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants.” In: *Nucleic acids research* 38.6, pp. 1767–1771.

Cohen, Adam D et al. (2010). “Agonist anti-GITR monoclonal antibody induces melanoma tumor immunity in mice by altering regulatory T cell stability and intra-tumor accumulation.” en. In: *PLoS One* 5.5, e10436.

Coifman, Ronald R et al. (2005). “Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps.” In: *Proceedings of the National Academy of Sciences of the United States of America* 102.21, pp. 7426–7431.

Consortium, ENCODE Project (2012). “An integrated encyclopedia of DNA elements in the human genome.” In: *Nature* 489.7414, p. 57.

Corthay, Alexandre (2014). “Does the immune system naturally protect against cancer?” en. In: *Front. Immunol.* 5, p. 197.

Curiel, Tyler J et al. (2004). “Specific recruitment of regulatory T cells in ovarian carcinoma fosters immune privilege and predicts reduced survival.” en. In: *Nat. Med.* 10.9, pp. 942–949.

Darrasse-Jèze, Guillaume et al. (2009). “Feedback control of regulatory T cell homeostasis by dendritic cells in vivo.” en. In: *J. Exp. Med.* 206.9, pp. 1853–1862.

De Henau, Olivier et al. (2016). “Overcoming resistance to checkpoint blockade therapy by targeting PI3Kγ in myeloid cells.” In: *Nature* 539.7629, pp. 443–447.

Dijk, David van et al. (2017). “MAGIC: A diffusion-based imputation method reveals gene-gene interactions in single-cell RNA-sequencing data.” In: *BioRxiv*, p. 111591.

Dobin, Alexander et al. (2013). “STAR: ultrafast universal RNA-seq aligner.” en. In: *Bioinformatics* 29.1, pp. 15–21.

Dohm, Juliane C et al. (2008). “Substantial biases in ultra-short read data sets from high-throughput DNA sequencing.” en. In: *Nucleic Acids Res.* 36.16, e105.

Drissen, Roy et al. (2016). “Distinct myeloid progenitor-differentiation pathways identified through single-cell RNA sequencing.” en. In: *Nat. Immunol.* 17.6, pp. 666–676.

Dushyanthen, Sathana et al. (2015). “Relevance of tumor-infiltrating lymphocytes in breast cancer.” In: *BMC medicine* 13.1, p. 202.

Eberlein, T J (2012). “gp100 Peptide Vaccine and Interleukin-2 in Patients with Advanced Melanoma.” In: *Yearbook of Surgery* 2012, pp. 350–352.

Eberwine, J et al. (1992). “Analysis of gene expression in single live neurons.” en. In: *Proc. Natl. Acad. Sci. U. S. A.* 89.7, pp. 3010–3014.

Engblom, Camilla, Christina Pfirschke, and Mikael J Pittet (2016). “The role of myeloid cells in cancer therapies.” In: *Nature Reviews Cancer* 16.7, pp. 447–462.

Eppert, Kolja et al. (2011). “Stem cell gene expression programs influence clinical outcome in human leukemia.” In: *Nature medicine* 17.9, pp. 1086–1093.

Faircloth, Brant C and Travis C Glenn (2012). “Not all sequence tags are created equal: designing and validating sequence identification tags robust to indels.” en. In: *PLoS One* 7.8, e42543.

Fan, Xiying and Alexander Y Rudensky (2016). “Hallmarks of tissue-resident lymphocytes.” In: *Cell* 164.6, pp. 1198–1211.

Farmer, Stephen R (2006). “Transcriptional control of adipocyte formation.” en. In: *Cell Metab.* 4.4, pp. 263–273.

Feinerman, Ofer et al. (2010). “Single-cell quantification of IL-2 response by effector and regulatory T cells reveals critical plasticity in immune response.” en. In: *Mol. Syst. Biol.* 6, p. 437.

Finger, Elizabeth C and Amato J Giaccia (2010). "Hypoxia, inflammation, and the tumor microenvironment in metastatic disease." In: *Cancer and Metastasis Reviews* 29.2, pp. 285–293.

Franciszkiewicz, Katarzyna et al. (2012). "Role of chemokines and chemokine receptors in shaping the effector phase of the antitumor immune response." en. In: *Cancer Res.* 72.24, pp. 6325–6332.

Funes, Juan M et al. (2007). "Transformation of human mesenchymal stem cells increases their dependency on oxidative phosphorylation for energy production." en. In: *Proc. Natl. Acad. Sci. U. S. A.* 104.15, pp. 6223–6228.

Gabrilovich, Dmitry I (2017). "Myeloid-Derived Suppressor Cells." en. In: *Cancer Immunol Res* 5.1, pp. 3–8.

Gailly, Jean-loup and Mark Adler (2004). "Zlib compression library." In:

García-Teijido, Paula et al. (2016). "Tumor-infiltrating lymphocytes in triple negative breast cancer: the future of immune targeting." In: *Clinical Medicine Insights. Oncology* 10.Suppl 1, p. 31.

Gaublomme, Jellert T et al. (2015). "Single-Cell Genomics Unveils Critical Regulators of Th17 Cell Pathogenicity." en. In: *Cell* 163.6, pp. 1400–1412.

Gesta, Stephane, Yu-Hua Tseng, and C Ronald Kahn (2007). "Developmental Origin of Fat: Tracking Obesity to Its Source." In: *Cell* 131.2, pp. 242–256.

Gholamin, Sharareh et al. (2017). "Disrupting the CD47-SIRPα anti-phagocytic axis by a humanized anti-CD47 antibody is an efficacious treatment for malignant pediatric brain tumors." In: *Science translational medicine* 9.381, eaaf2968.

Glimcher, Laurie H et al. (2004). "Recent developments in the transcriptional regulation of cytolytic effector cells." In: *Nat. Rev. Immunol.* 4.11, pp. 900–911.

Glinsky, Gennadi V et al. (2004). "Gene expression profiling predicts clinical outcome of prostate cancer." In: *Journal of Clinical Investigation* 113.6, p. 913.

Golovina, Tatiana N and Robert H Vonderheide (2010). "Regulatory T cells: overcoming suppression of T-cell immunity." en. In: *Cancer J.* 16.4, pp. 342–347.

Görür, Dilan and Carl Edward Rasmussen (2010). "Dirichlet Process Gaussian Mixture Models: Choice of the Base Distribution." In: *J. Comput. Sci. Technol.* 25.4, pp. 653–664.

Grivennikov, Sergei I, Florian R Greten, and Michael Karin (2010). "Immunity, inflammation, and cancer." en. In: *Cell* 140.6, pp. 883–899.

Grün, Dominic, Lennart Kester, and Alexander Van Oudenaarden (2014). "Validation of noise models for single-cell transcriptomics." In: *Nature methods* 11.6, p. 637.

Grun, Dominic and Alexander van Oudenaarden (2016). "Design and Analysis of Single-Cell Sequencing Experiments." In: *Cell* 163.4, pp. 799–810. ISSN: 0092-8674. DOI: 10.1016/j.cell.2015.10.039. URL: http://dx.doi.org/10.1016/j.cell.2015.10.039.

Guo, Xiaocan et al. (2017). "Single tumor-initiating cells evade immune clearance by recruiting type II macrophages." In: *Genes & Development* 31.3, pp. 247–259.

Gury-BenAri, Meital et al. (2016). "The spectrum and regulatory landscape of intestinal innate lymphoid cells are shaped by the microbiome." In: *Cell* 166.5, pp. 1231–1246.

Haas, Brian J et al. (2013). "De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis." In: *Nature protocols* 8.8, pp. 1494–1512.

Habib, Naomi et al. (2016). "Div-Seq: Single-nucleus RNA-Seq reveals dynamics of rare adult newborn neurons." In: *Science* 353.6302, pp. 925–928.

Habib, Naomi et al. (2017). "DroNc-Seq: Deciphering cell types in human archived brain tissues by massively-parallel single nucleus RNA-seq." In: *bioRxiv*, p. 115196.

Haghverdi, Laleh, Florian Buettner, and Fabian J Theis (2015). "Diffusion maps for high-dimensional single-cell analysis of differentiation data." In: *Bioinformatics* 31.18, pp. 2989–2998.

Haghverdi, Laleh et al. (2016). "Diffusion pseudotime robustly reconstructs lineage branching." In: *Nature methods* 13.10, pp. 845–848.

Halko, Nathan, Per-Gunnar Martinsson, and Joel A Tropp (2009). "Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions." In: arXiv: 0909.4061 [math.NA].

Hamid, Omid et al. (2013). "Safety and Tumor Responses with Lambrolizumab (Anti–PD-1) in Melanoma." In: *N. Engl. J. Med.* 369.2, pp. 134–144.

Hartigan, J A and P M Hartigan (1985). "The Dip Test of Unimodality." In: *Ann. Stat.* 13.1, pp. 70–84.

Hartwell, Leland H et al. (1999). "From molecular to modular cell biology." In: *Nature* 402, pp. C47–C52.

Ho, Ping-Chih et al. (2015). "Phosphoenolpyruvate Is a Metabolic Checkpoint of Anti-tumor T Cell Responses." en. In: *Cell* 162.6, pp. 1217–1228.

Hodi, F Stephen et al. (2010). "Improved survival with ipilimumab in patients with metastatic melanoma." en. In: *N. Engl. J. Med.* 363.8, pp. 711–723.

Ilicic, Tomislav et al. (2016). “Classification of low quality cells from single-cell RNA-seq data.” In: *Genome biology* 17.1, p. 29.

Islam, Saiful et al. (2011). “Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq.” en. In: *Genome Res.* 21.7, pp. 1160–1167.

Ito, Tomoki et al. (2006). “OX40 ligand shuts down IL-10-producing regulatory T cells.” en. In: *Proc. Natl. Acad. Sci. U. S. A.* 103.35, pp. 13138–13143.

Jahrsdörfer, Bernd et al. (2010). “Granzyme B produced by human plasmacytoid dendritic cells suppresses T-cell expansion.” In: *Blood* 115.6, pp. 1156–1165.

Jaitin, Diego Adhemar et al. (2014). “Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types.” en. In: *Science* 343.6172, pp. 776–779.

Jebara, T., Kondor, R., & Howard, A. (2004). “Probability product kernels.” In: *J. Mach. Learn. Res.* 5.Jul, pp. 819–844.

Jeffrey, Kate L et al. (2006). “Positive regulation of immune cell function and inflammatory responses by phosphatase PAC-1.” In: *Nature immunology* 7.3, pp. 274–283.

Jiménez-Sánchez, Alejandro et al. (2017). “Heterogeneous Tumor-Immune Microenvironments among Differentially Growing Metastases in an Ovarian Cancer Patient.” In: *Cell* 170.5, pp. 927–938.

Joffre, Olivier et al. (2009). “Inflammatory signals in dendritic cell activation and the induction of adaptive immunity.” en. In: *Immunol. Rev.* 227.1, pp. 234–247.

Johnson, David S et al. (2007). “Genome-wide mapping of in vivo protein-DNA interactions.” In: *Science* 316.5830, pp. 1497–1502.

Josefowicz, Steven Z, Li-Fan Lu, and Alexander Y Rudensky (2012). “Regulatory T cells: mechanisms of differentiation and function.” In: *Annual review of immunology* 30, pp. 531–564.

Keir, Mary E et al. (2008). “PD-1 and Its Ligands in Tolerance and Immunity.” In: *Annu. Rev. Immunol.* 26.1, pp. 677–704.

Kim, Daehwan, Ben Langmead, and Steven L Salzberg (2015). “HISAT: a fast spliced aligner with low memory requirements.” In: *Nature methods* 12.4, pp. 357–360.

Kim, Daehwan et al. (2013). “TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions.” en. In: *Genome Biol.* 14.4, R36.

Klein, Allon M et al. (2015). “Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells.” en. In: *Cell* 161.5, pp. 1187–1201.

Kundaje, Anshul et al. (2015). “Integrative analysis of 111 reference human epigenomes.” In: *Nature* 518.7539, p. 317.

Lane, Keara et al. (2017). “Measuring Signaling and RNA-Seq in the Same Cell Links Gene Expression to Dynamic Patterns of NF-κB Activation.” In: *Cell Systems* 4.4, pp. 458–469.

Langmead, Ben and Steven L Salzberg (2012). “Fast gapped-read alignment with Bowtie 2.” In: *Nature methods* 9.4, pp. 357–359.

Lavin, Yonit et al. (2017). “Innate immune landscape in early lung adenocarcinoma by paired single-cell analyses.” In: *Cell* 169.4, pp. 750–765.

Lefterova, Martina I and Mitchell A Lazar (2009). “New developments in adipogenesis.” In: *Trends Endocrinol. Metab.* 20.3, pp. 107–114.

Levine, Jacob H et al. (2015). “Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis.” In: *Cell* 162.1, pp. 184–197.

Li, Bo and Colin N Dewey (2011). “RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome.” In: *BMC Bioinformatics* 12.1, p. 323.

Li, Heng et al. (2009). “The sequence alignment/map format and SAMtools.” In: *Bioinformatics* 25.16, pp. 2078–2079.

Lippitz, Bodo E (2013). “Cytokine patterns in patients with cancer: a systematic review.” In: *Lancet Oncol.* 14.6, e218–e228.

Lönnberg, Tapio et al. (2017). “Single-cell RNA-seq and computational analysis using temporal mixture modelling resolves Th1/Tfh fate bifurcation in malaria.” en. In: *Sci Immunol* 2.9.

Lun, Aaron TL, John C Marioni, and Karsten Bach (2016). “Pooling across cells to normalize single-cell RNA sequencing data with many zero counts.” In: *Genome biology* 17.1, p. 75.

Maaten, L van der and G Hinton (2008). “Visualizing Data using t-SNE.” In: *J Mach Learn Res* 9, pp. 2579–2605.

Macosko, Evan Z et al. (2015). “Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets.” en. In: *Cell* 161.5, pp. 1202–1214.

Makino, Yuichi et al. (2003). “Hypoxia-inducible factor regulates survival of antigen receptor-driven T cells.” en. In: *J. Immunol.* 171.12, pp. 6534–6540.

Mamedov, T G et al. (2008). “A fundamental study of the PCR amplification of GC-rich DNA templates.” en. In: *Comput. Biol. Chem.* 32.6, pp. 452–457.

Manley, Leigh J, Duanduan Ma, and Stuart S Levine (2016). "Monitoring Error Rates In Illumina Sequencing." en. In: *J. Biomol. Tech.* 27.4, pp. 125–128.

Mantovani, Alberto and Massimo Locati (2013). "Tumor-associated macrophages as a paradigm of macrophage plasticity, diversity, and polarization." In: *Arteriosclerosis, thrombosis, and vascular biology* 33.7, pp. 1478–1483.

Mantovani, Alberto et al. (2008). "Cancer-related inflammation." en. In: *Nature* 454.7203, pp. 436–444.

Marrack, Philippa et al. (2000). "Genomic-scale analysis of gene expression in resting and activated T cells." In: *Current opinion in immunology* 12.2, pp. 206–209.

McDermott, D F et al. (2014). "Immune correlates and long term follow up of a phase Ia study of MPDL3280A, an engineered PD-L1 antibody, in patients with metastatic renal cell carcinoma …." In: *niu.edu.*

McKenna, Aaron et al. (2010). "The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data." In: *Genome research* 20.9, pp. 1297–1303.

McLendon, Roger et al. (2008). "Comprehensive genomic characterization defines human glioblastoma genes and core pathways." In: *Nature* 455.7216, pp. 1061–1068.

Mellman, Ira, George Coukos, and Glenn Dranoff (2011). "Cancer immunotherapy comes of age." In: *Nature* 480.7378, pp. 480–489.

Metzker, Michael L (2010). "Sequencing technologies—the next generation." In: *Nature reviews genetics* 11.1, pp. 31–46.

Michielsen, Adriana J et al. (2012). "Inhibition of dendritic cell maturation by the tumor microenvironment correlates with the survival of colorectal cancer patients following bevacizumab treatment." en. In: *Mol. Cancer Ther.* 11.8, pp. 1829–1837.

Mills, Charles D et al. (2000). "M-1/M-2 macrophages and the Th1/Th2 paradigm." In: *The Journal of Immunology* 164.12, pp. 6166–6173.

Moignard, Victoria et al. (2015). "Decoding the regulatory network of early blood development from single-cell gene expression measurements." In: *Nature biotechnology* 33.3, pp. 269–276.

Monsurrò, Vladia et al. (2004). "Quiescent phenotype of tumor-specific CD8+ T cells following immunization." In: *Blood* 104.7, pp. 1970–1978.

Moreno-Sánchez, Rafael et al. (2009). "The bioenergetics of cancer: Is glycolysis the main ATP supplier in all tumor cells?" In: *Biofactors* 35.2, pp. 209–225.

Mues, C et al. (2009). “Regulation of glucose-6-phosphatase gene expression by insulin and metformin.” en. In: *Horm. Metab. Res.* 41.10, pp. 730–735.

Muller, Alexander J and Peggy A Scherle (2006). “Targeting the mechanisms of tumoral immune tolerance with small-molecule inhibitors.” en. In: *Nat. Rev. Cancer* 6.8, pp. 613–625.

Murray, Peter J et al. (2014). “Macrophage activation and polarization: nomenclature and experimental guidelines.” In: *Immunity* 41.1, pp. 14–20.

Nestorowa, Sonia et al. (2016). “A single-cell resolution map of mouse hematopoietic stem and progenitor cell differentiation.” en. In: *Blood* 128.8, e20–31.

Network, Cancer Genome Atlas Research et al. (2011). “Integrated genomic analyses of ovarian carcinoma.” In: *Nature* 474.7353, pp. 609–615.

Newell, Evan W et al. (2012). “Cytometry by time-of-flight shows combinatorial cytokine expression and virus-specific cell niches within a continuum of CD8+ T cell phenotypes.” en. In: *Immunity* 36.1, pp. 142–152.

Nishimura, H et al. (1999). “Development of lupus-like autoimmune diseases by disruption of the PD-1 gene encoding an ITIM motif-carrying immunoreceptor.” en. In: *Immunity* 11.2, pp. 141–151.

Nishimura, H et al. (2001). “Autoimmune dilated cardiomyopathy in PD-1 receptor-deficient mice.” en. In: *Science* 291.5502, pp. 319–322.

Novershtern, Noa et al. (2011). “Densely interconnected transcriptional circuits control cell states in human hematopoiesis.” In: *Cell* 144.2, pp. 296–309.

Ohta, Akio et al. (2006). “A2A adenosine receptor protects tumors from antitumor T cells.” en. In: *Proc. Natl. Acad. Sci. U. S. A.* 103.35, pp. 13132–13137.

Patro, Rob et al. (2017). “Salmon provides fast and bias-aware quantification of transcript expression.” en. In: *Nat. Methods* 14.4, pp. 417–419.

Paul, Franziska et al. (2015). “Transcriptional Heterogeneity and Lineage Commitment in Myeloid Progenitors.” en. In: *Cell* 163.7, pp. 1663–1677.

Perera, Ranjan J et al. (2006). “Identification of novel PPARγ target genes in primary human adipocytes.” In: *Gene* 369, pp. 90–99.

Perfetto, Stephen P, Pratip K Chattopadhyay, and Mario Roederer (2004). “Seventeen-colour flow cytometry: unravelling the immune system.” en. In: *Nat. Rev. Immunol.* 4.8, pp. 648–655.

Petukhov, Viktor et al. (2017). "Accurate estimation of molecular counts in droplet-based single-cell RNA-seq experiments." In: *bioRxiv*, p. 171496.

Platanias, Leonidas C (2005). "Mechanisms of type-I- and type-II-interferon-mediated signalling." en. In: *Nat. Rev. Immunol.* 5.5, pp. 375–386.

Prabhakaran, Sandhya et al. (2016). "Dirichlet process mixture model for correcting technical variation in single-cell gene expression data." In: *International Conference on Machine Learning*, pp. 1070–1079.

Pyonteck, Stephanie M et al. (2013). "CSF-1R inhibition alters macrophage polarization and blocks glioma progression." In: *Nature medicine* 19.10, pp. 1264–1272.

Qureshi, Omar S et al. (2011). "Trans-endocytosis of CD80 and CD86: a molecular basis for the cell-extrinsic function of CTLA-4." en. In: *Science* 332.6029, pp. 600–603.

Ramsköld, Daniel et al. (2012). "Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells." en. In: *Nat. Biotechnol.* 30.8, pp. 777–782.

Regev, Aviv et al. (2017). "The human cell atlas." In: *Elife* 6.

Riley, Todd R et al. (2014). "SELEX-seq: a method for characterizing the complete repertoire of binding site preferences for transcription factor complexes." In: *Hox Genes: Methods and Protocols*, pp. 255–278.

Robert, Caroline et al. (2011). "Ipilimumab plus dacarbazine for previously untreated metastatic melanoma." en. In: *N. Engl. J. Med.* 364.26, pp. 2517–2526.

Robinson, Mark D and Alicia Oshlack (2010). "A scaling normalization method for differential expression analysis of RNA-seq data." In: *Genome biology* 11.3, R25.

Rokhlin, Vladimir, Arthur Szlam, and Mark Tygert (2009). "A randomized algorithm for principal component analysis." In: *SIAM Journal on Matrix Analysis and Applications* 31.3, pp. 1100–1124.

Rosenberg, Steven A, James C Yang, and Nicholas P Restifo (2004). "Cancer immunotherapy: moving beyond current vaccines." en. In: *Nat. Med.* 10.9, pp. 909–915.

Sainz, Bruno et al. (2014). "ISG15 is a critical microenvironmental factor for pancreatic cancer stem cells." In: *Cancer research* 74.24, pp. 7309–7320.

Schietinger, Andrea et al. (2012). "Rescued Tolerant CD8 T Cells Are Preprogrammed to Reestablish the Tolerant State." en. In: *Science* 335.6069, pp. 723–727.

Schlitzer, Andreas et al. (2015). “Identification of cDC1- and cDC2-committed DC progenitors reveals early lineage priming at the common DC progenitor stage in the bone marrow.” en. In: *Nat. Immunol.* 16.7, pp. 718–728.

Segal, Eran et al. (2004). “A module map showing conditional activity of expression modules in cancer.” In: *Nature genetics* 36.10, pp. 1090–1098.

Segal, Neil H et al. (2008). “Epitope landscape in breast and colorectal cancer.” en. In: *Cancer Res.* 68.3, pp. 889–892.

Şenbabaoğlu, Yasin et al. (2016). “Tumor immune microenvironment characterization in clear cell renal cell carcinoma identifies prognostic and immunotherapeutically relevant messenger RNA signatures.” In: *Genome biology* 17.1, p. 231.

Setty, Manu et al. (2016). “Wishbone identifies bifurcating developmental trajectories from single-cell data.” In: *Nature biotechnology* 34.6, pp. 637–645.

Shah, Sheel et al. (2016). “In situ transcription profiling of single cells reveals spatial organization of cells in the mouse hippocampus.” In: *Neuron* 92.2, pp. 342–357.

Shalek, Alex K et al. (2013). “Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells.” en. In: *Nature* 498.7453, pp. 236–240.

Shalek, Alex K et al. (2014). “Single-cell RNA-seq reveals dynamic paracrine control of cellular variation.” en. In: *Nature* 510.7505, pp. 363–369.

Sharma, Padmanee et al. (2017). “Primary, adaptive, and acquired resistance to cancer immunotherapy.” In: *Cell* 168.4, pp. 707–723.

Shekhar, Karthik et al. (2016). “Comprehensive classification of retinal bipolar neurons by single-cell transcriptomics.” In: *Cell* 166.5, pp. 1308–1323.

Shiroguchi, Katsuyuki et al. (2012). “Digital RNA sequencing minimizes sequence-dependent bias and amplification noise with optimized single-molecule barcodes.” en. In: *Proc. Natl. Acad. Sci. U. S. A.* 109.4, pp. 1347–1352.

Sica, Antonio and Alberto Mantovani (2012). “Macrophage plasticity and polarization: in vivo veritas.” en. In: *J. Clin. Invest.* 122.3, p. 787.

Singer, Meromit et al. (2016). “A distinct gene module for dysfunction uncoupled from activation in tumor-infiltrating T cells.” In: *Cell* 166.6, pp. 1500–1511.

Smith-Garvin, Jennifer E, Gary A Koretzky, and Martha S Jordan (2009). “T Cell Activation.” In: *Annu. Rev. Immunol.* 27.1, pp. 591–619.

Steinman, R M et al. (2000). “The induction of tolerance by dendritic cells that have captured apoptotic cells.” en. In: *J. Exp. Med.* 191.3, pp. 411–416.

Stockman, J A (2011). “Vaccination against HPV-16 Oncoproteins for Vulvar Intraepithelial Neoplasia.” In: *Yearbook of Pediatrics* 2011, pp. 20–22.

Subramanian, Aravind et al. (2005). “Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.” In: *Proceedings of the National Academy of Sciences* 102.43, pp. 15545–15550.

Swiecki, Melissa and Marco Colonna (2015). “The multifaceted biology of plasmacytoid dendritic cells.” In: *Nature Reviews Immunology* 15.8, pp. 471–485.

Tang, Fuchou et al. (2009). “mRNA-Seq whole-transcriptome analysis of a single cell.” In: *Nature methods* 6.5, pp. 377–382.

Tao, Terence and Van Vu (2005). “On random ± 1 matrices: Singularity and determinant.” In: *Random Struct. Algorithms* 28.1, pp. 1–23.

The HDF Group (1997-2018). *Hierarchical Data Format, version 5.* http://www.hdfgroup.org/HDF5/.

Tirosh, Itay et al. (2016). “Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq.” In: *Science* 352.6282, pp. 189–196. ISSN: 0036-8075. DOI: 10.1126/science.aad0501. eprint: http://science.sciencemag.org/content/352/6282/189.full.pdf. URL: http://science.sciencemag.org/content/352/6282/189.

Tivol, E A et al. (1995). “Loss of CTLA-4 leads to massive lymphoproliferation and fatal multiorgan tissue destruction, revealing a critical negative regulatory role of CTLA-4.” en. In: *Immunity* 3.5, pp. 541–547.

Tusher, Virginia Goss, Robert Tibshirani, and Gilbert Chu (2001). “Significance analysis of microarrays applied to the ionizing radiation response.” In: *Proceedings of the National Academy of Sciences* 98.9, pp. 5116–5121.

Ugel, Stefano et al. (2015). “Tumor-induced myeloid deviation: when myeloid-derived suppressor cells meet tumor-associated macrophages.” en. In: *J. Clin. Invest.* 125.9, pp. 3365–3376.

Valle, Sergio, Weihua Li, and S Joe Qin (1999). “Selection of the Number of Principal Components: The Variance of the Reconstruction Error Criterion with a Comparison to Other Methods†.” In: *Ind. Eng. Chem. Res.* 38.11, pp. 4389–4401.

Vallejos, C A et al. (2017). “Normalizing single-cell RNA sequencing data: challenges and opportunities.” In: *Nature Methods* 14, pp. 565–571.

Van't Veer, Laura J et al. (2002). "Gene expression profiling predicts clinical outcome of breast cancer." In: *nature* 415.6871, pp. 530–536.

Villani, Alexandra-Chloé et al. (2017). "Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors." In: *Science* 356.6335, eaah4573.

Wherry, John E (2011). "T cell exhaustion." en. In: *Nat. Immunol.* 12.6, ni.2035.

Wherry, John E and Makoto Kurachi (2015). "Molecular and cellular insights into T cell exhaustion." en. In: *Nat. Rev. Immunol.* 15.8, nri3862.

Whitfield, Michael L et al. (2002). "Identification of Genes Periodically Expressed in the Human Cell Cycle and Their Expression in Tumors." en. In: *Mol. Biol. Cell* 13.6, pp. 1977–2000.

Wilhelm, Brian T and Josette-Renée Landry (2009). "RNA-Seq-quantitative measurement of expression through massively parallel RNA-sequencing." en. In: *Methods* 48.3, pp. 249–257.

Wing, K et al. (2008). "CTLA-4 Control over Foxp3 Regulatory T Cell Function." In: *Science* 322.5899, pp. 271–275.

Wolchok, Jedd D et al. (2013). "Nivolumab plus ipilimumab in advanced melanoma." en. In: *N. Engl. J. Med.* 369.2, pp. 122–133.

Zheng, Chunhong et al. (2017a). "Landscape of Infiltrating T Cells in Liver Cancer Revealed by Single-Cell Sequencing." In: *Cell* 169.7, pp. 1342–1356.

Zheng, Grace X Y et al. (2017b). "Massively parallel digital transcriptional profiling of single cells." en. In: *Nat. Commun.* 8, p. 14049.

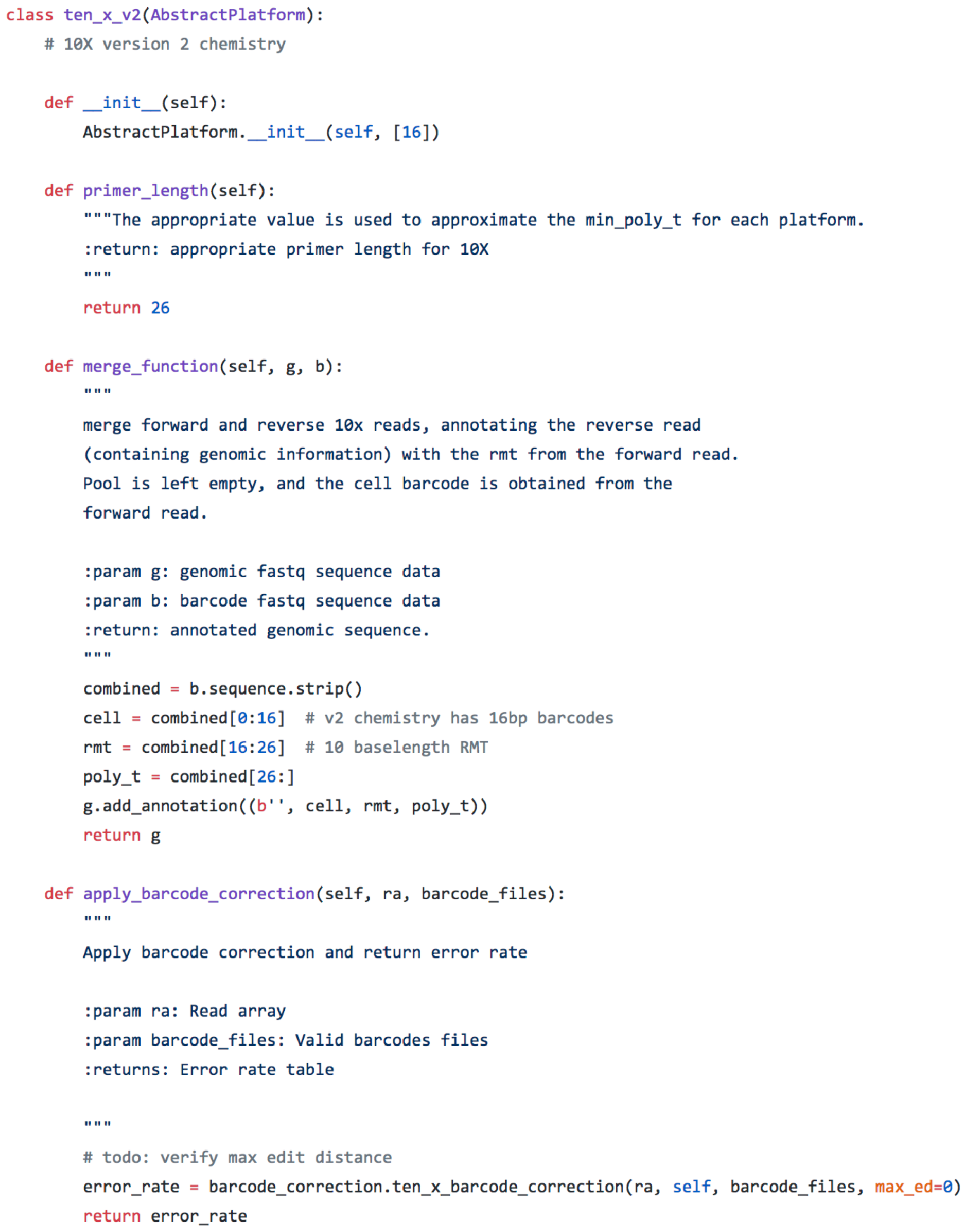
Zheng, Wei, Lisa M Chung, and Hongyu Zhao (2011). "Bias detection and correction in RNA-Sequencing data." en. In: *BMC Bioinformatics* 12, p. 290.

Ziegenhain, Christoph et al. (2017). "Comparative Analysis of Single-Cell RNA Sequencing Methods." In: *Molecular Cell* 65.4, 631–643.e4. ISSN: 1097-2765. DOI: 10.1016/j.molcel.2017.01.023. URL: https://doi.org/10.1016/j.molcel.2017.01.023.

Zilionis, Rapolas et al. (2017). "Single-cell barcoding and sequencing using droplet microfluidics." In: *nature protocols* 12.1, pp. 44–73.

Zitvogel, Laurence and Guido Kroemer (2009). "Anticancer immunochemotherapy using adjuvants with direct cytotoxic effects." en. In: *J. Clin. Invest.* 119.8, pp. 2127–2130.

Supplementary Figures



```python
class ten_x_v2(AbstractPlatform):
    # 10X version 2 chemistry

    def __init__(self):
        AbstractPlatform.__init__(self, [16])

    def primer_length(self):
        """The appropriate value is used to approximate the min_poly_t for each platform.
        :return: appropriate primer length for 10X
        """
        return 26

    def merge_function(self, g, b):
        """
        merge forward and reverse 10x reads, annotating the reverse read
        (containing genomic information) with the rmt from the forward read.
        Pool is left empty, and the cell barcode is obtained from the
        forward read.

        :param g: genomic fastq sequence data
        :param b: barcode fastq sequence data
        :return: annotated genomic sequence.
        """
        combined = b.sequence.strip()
        cell = combined[0:16]  # v2 chemistry has 16bp barcodes
        rmt = combined[16:26]  # 10 baselength RMT
        poly_t = combined[26:]
        g.add_annotation((b'', cell, rmt, poly_t))
        return g

    def apply_barcode_correction(self, ra, barcode_files):
        """
        Apply barcode correction and return error rate

        :param ra: Read array
        :param barcode_files: Valid barcodes files
        :returns: Error rate table

        """
        # todo: verify max edit distance
        error_rate = barcode_correction.ten_x_barcode_correction(ra, self, barcode_files, max_ed=0)
        return error_rate
```

Figure 1: Code for the 10x platform class. Each platform class defines the length of the primer sequence, a merge function to combine the barcodes and the genomic fastq, and optionally, custom barcode correction and multi-alignment resolution methods. 10x uses the default multi-alignment method, and therefore that does not require implementation in this class.

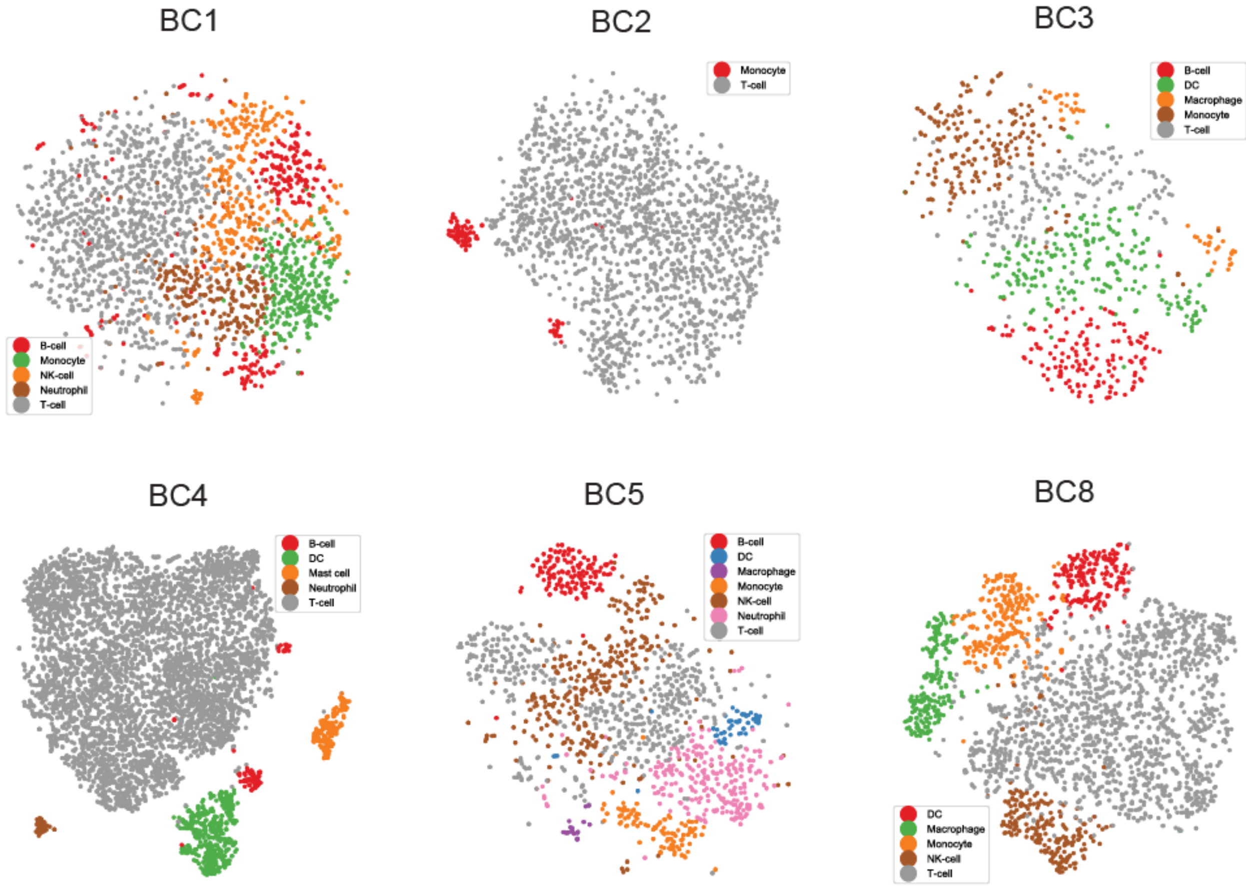


Figure 2: t-SNE projection of complete immune systems from six breast cancer tumors. scRNA-seq data for each tumor is processed with pipeline in Figure S1B and library size-normalized; each dot represents a single- cell colored by PhenoGraph clustering, and clusters are labeled by inferred cell types. Two additional tumors are presented in 3.1.

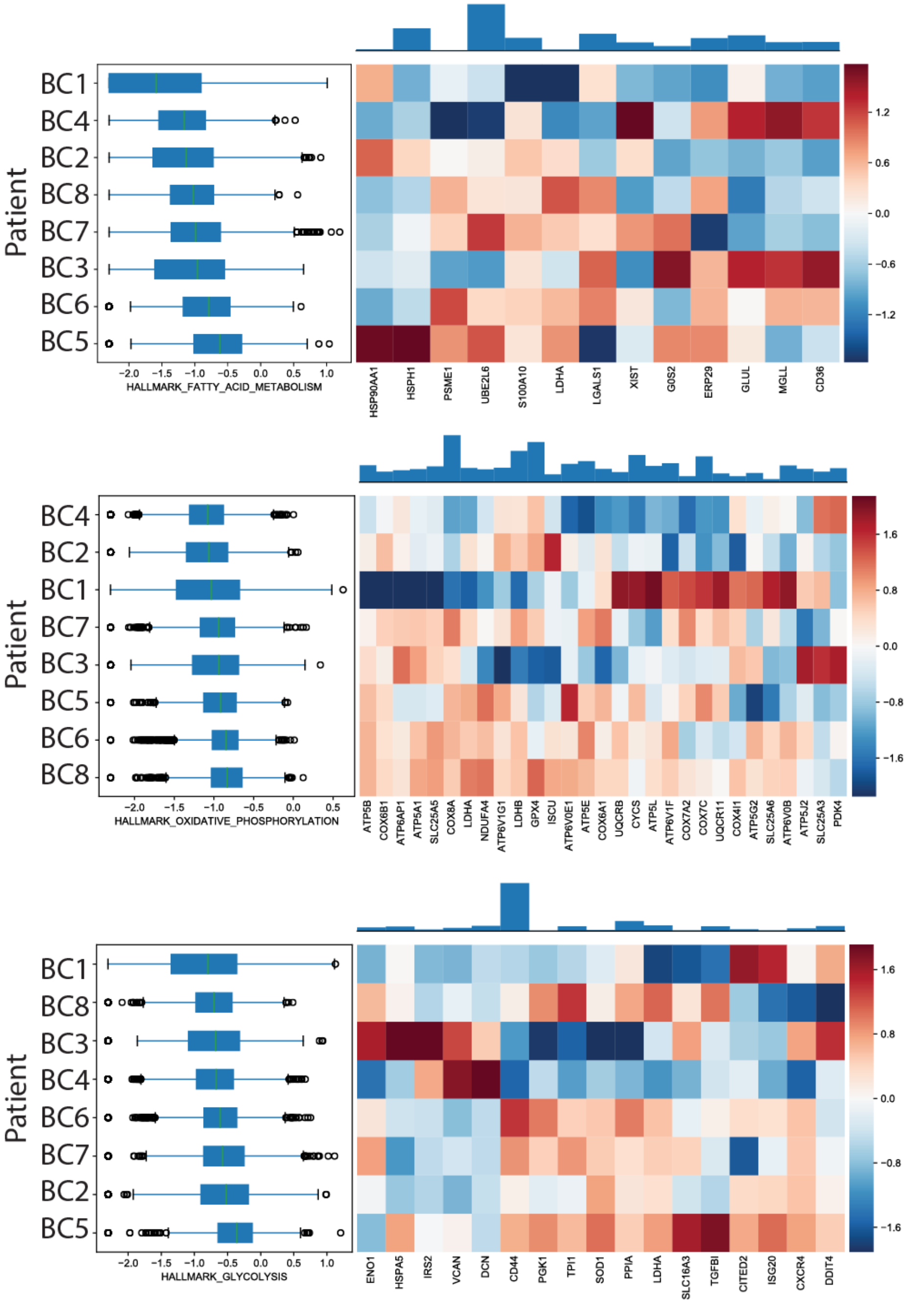


Figure 3: Expression of metabolic signatures: fatty acid metabolism (top), phosphorylation (middle), and glycolysis (bottom), summarized as boxplots (left) showing expression of each respective signature (defined as the mean normalized expression of genes) across immune cells from each patient; and heatmap (right) displaying z-scored mean expression of genes in each signature; (top) barplot showing total expression of each gene indicated in the heatmap across all patients. See Figure 3.13 for one additional signature

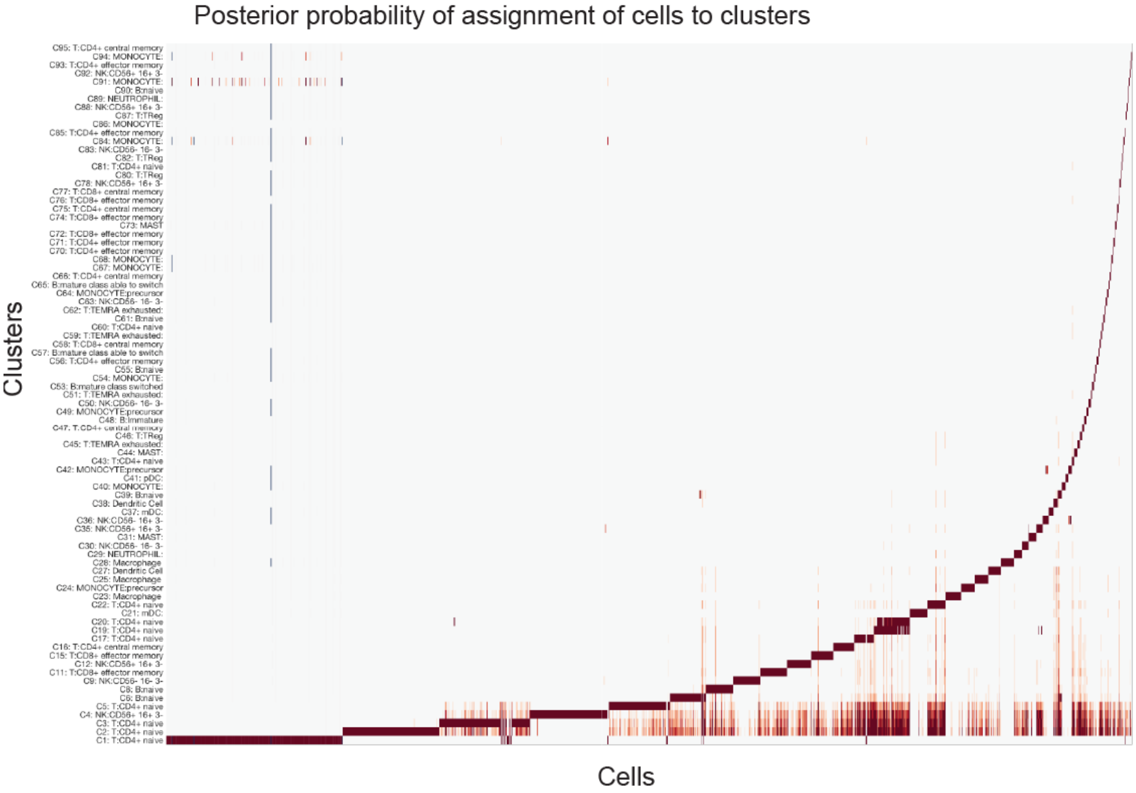


Figure 4: Posterior probability of assignment of cells to clusters in the Biscuit model in the full immune cell atlas of combined tissues and patients presented in Figure 3.10; note broad distributions in assignment of naive T cells (bottom) as compared to other cell types.

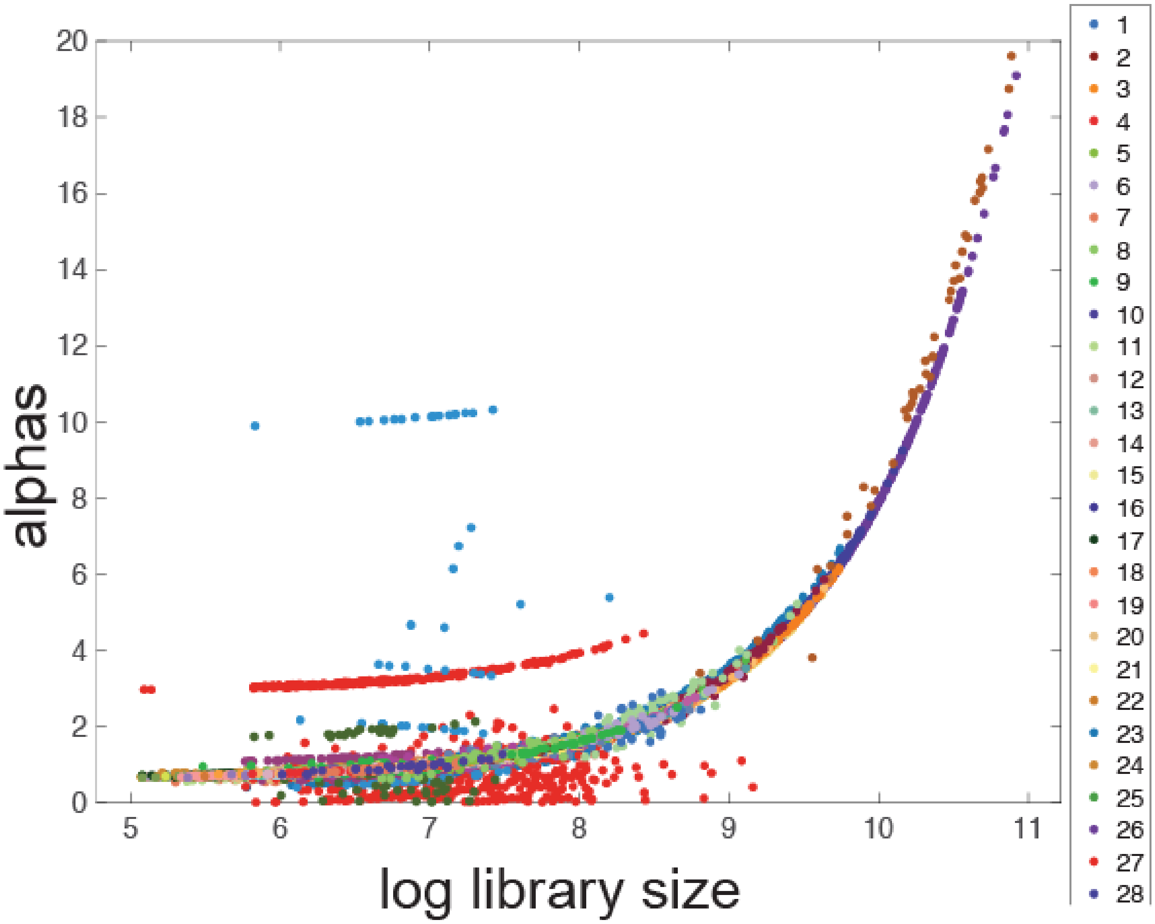


Figure 5: Distribution of Biscuit alpha parameters per cell vs log of library size, with cells colored by clusters; Biscuit alpha parameters correct for differences in library size across and within clusters.

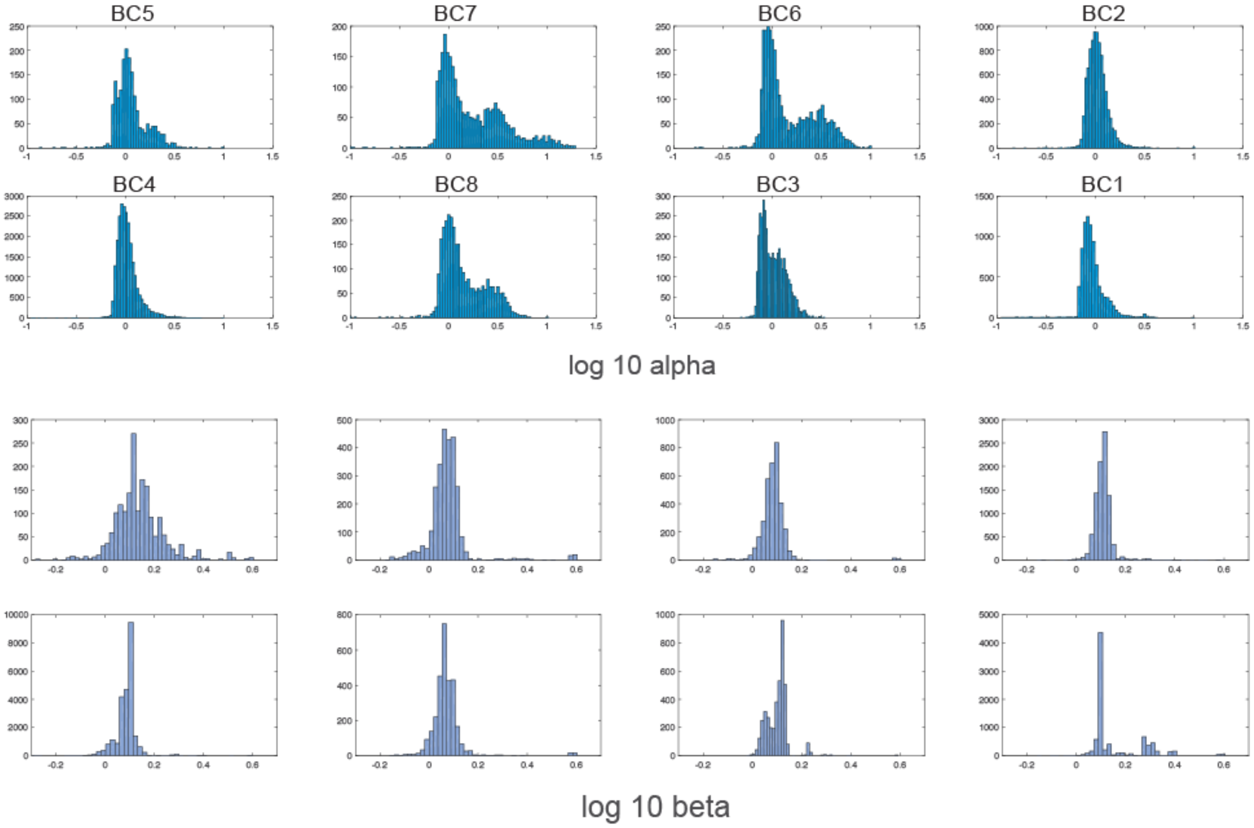


Figure 6: Distribution of inferred cell-specific parameters alpha and beta in Biscuit across cells from each patient. These differences were corrected in normalizing with alpha and beta parameters.

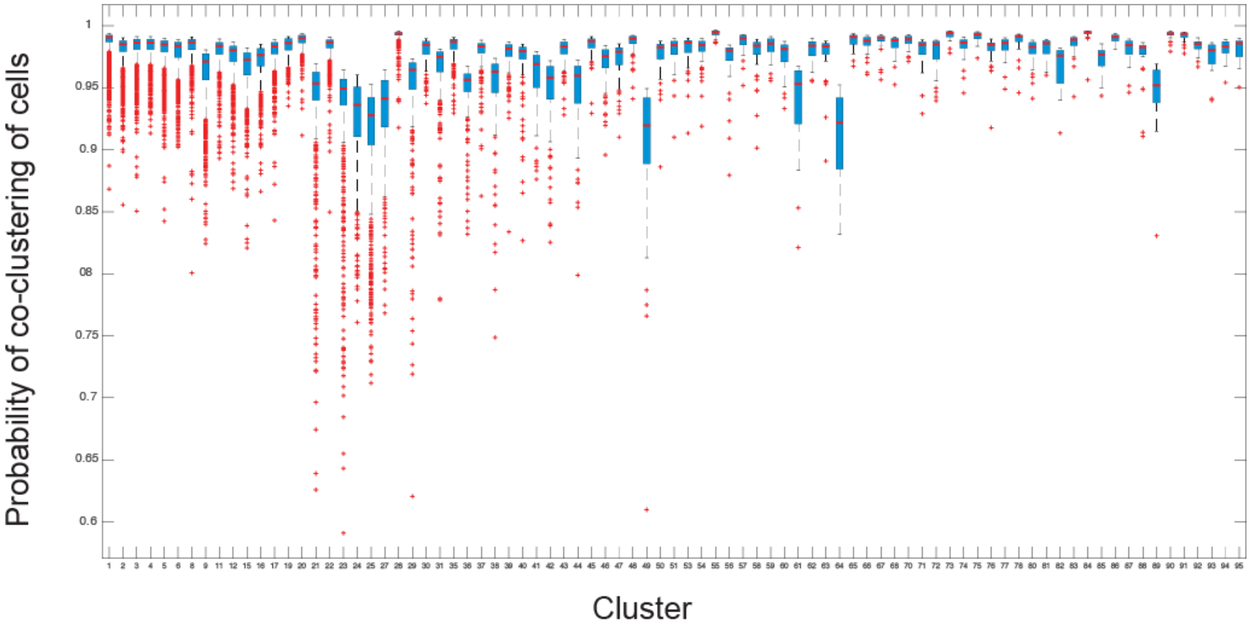


Figure 7: Robustness analysis of clusters performed with 10-fold cross-validation; boxplots summarize the probability of a pair of cells being assigned to the same final cluster across all 10 subsets.

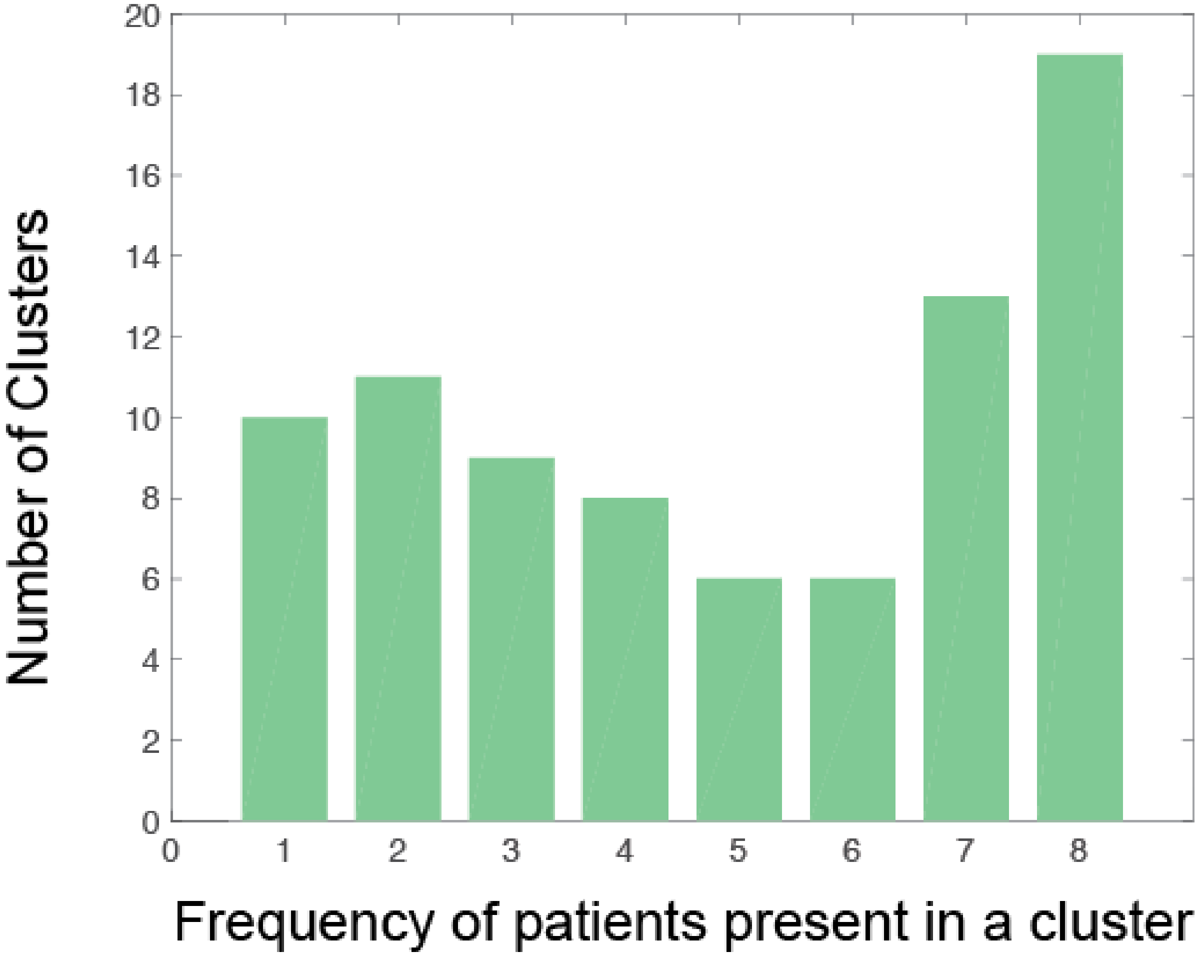


Figure 8: Histogram of frequency of patients contributing to each cluster showing that 19 clusters (out of 95) are present in all 8 patients and 10 clusters are patient-specific.

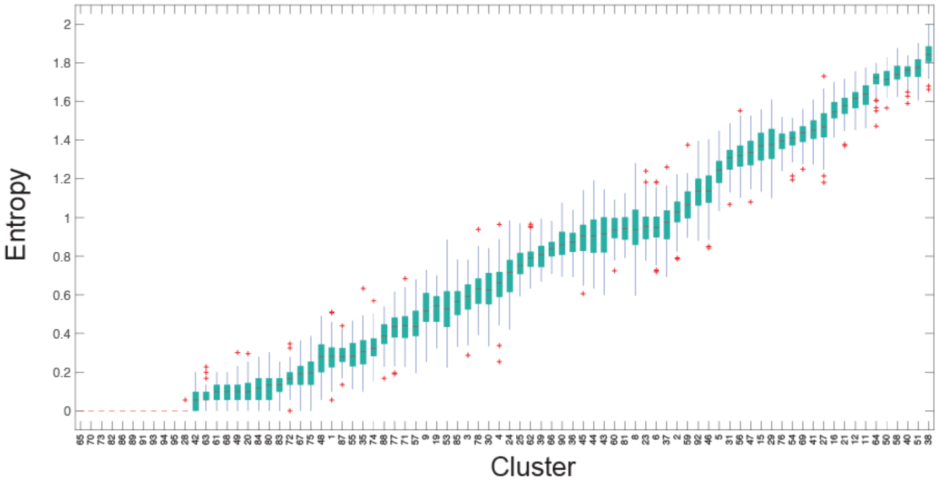


Figure 9: Boxplots showing entropy of distribution of patients in each cluster, computed with bootstrapping to correct for cluster size. Note that cluster labels are given by size (cluster 1 has the most number of cells and cluster 95 has the fewest) and ordering clusters by mean entropy in this plot indicates that entropy does not correlate with size.

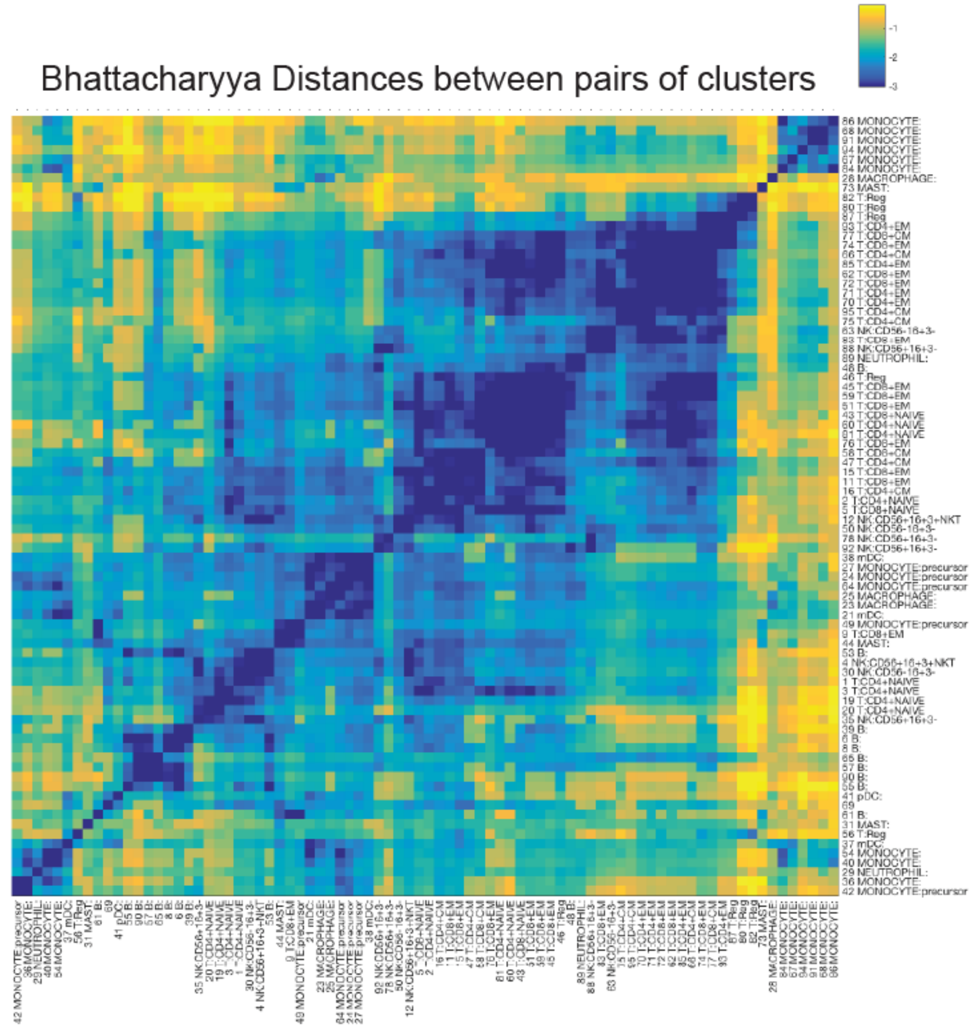


Figure 10: Bhattacharyya pairwise distances between clusters of Figure 3.12 (blue: small distance to yellow: large distance).

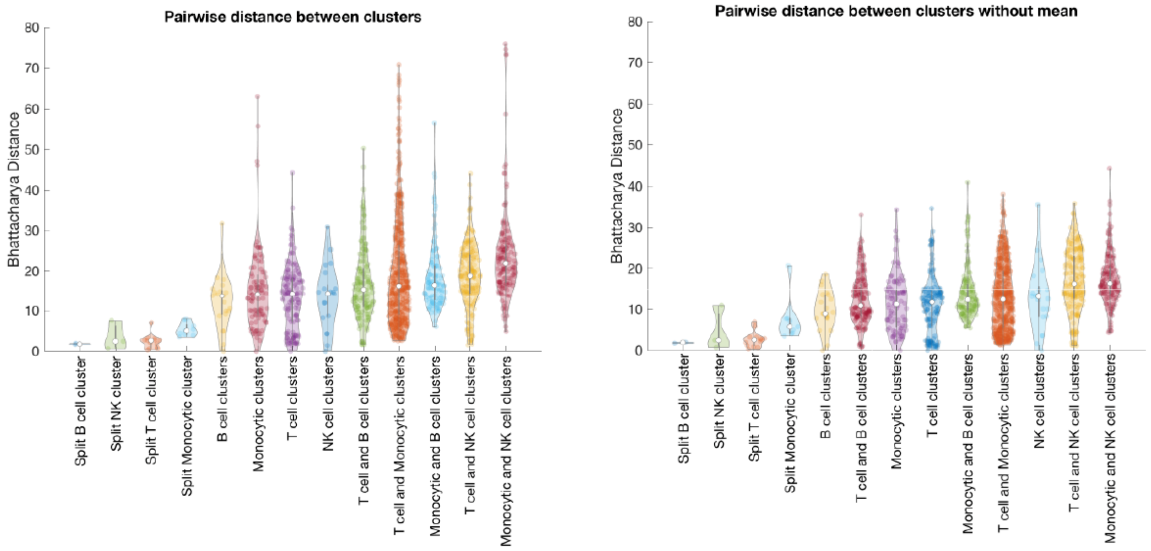
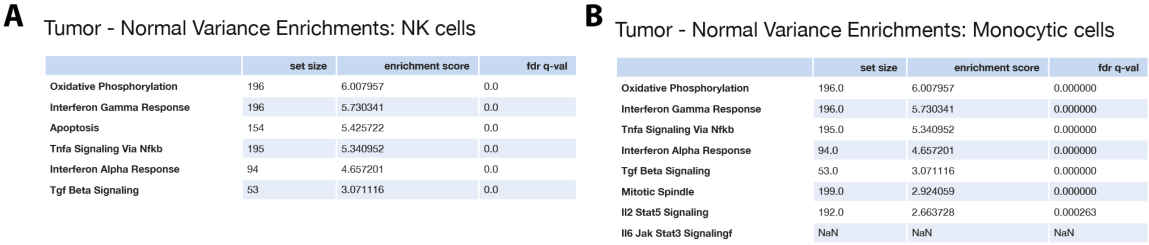


Figure 11: Left: Violin plot of pairwise Bhattacharyya distances between distribution of expression of each gene between all pairs of clusters in the same or different cell types considering mean and covariance of expression, averaged across all genes. Right: same as left, but after removing the effect of cluster mean in computing similarity, thus considering only covariance.

Figure S3



A Tumor - Normal Variance Enrichments: NK cells

	set size	enrichment score	fdr q-val
Oxidative Phosphorylation	196	6.007957	0.0
Interferon Gamma Response	196	5.730341	0.0
Apoptosis	154	5.425722	0.0
Tnfa Signaling Via Nfkb	195	5.340952	0.0
Interferon Alpha Response	94	4.657201	0.0
Tgf Beta Signaling	53	3.071116	0.0

B Tumor - Normal Variance Enrichments: Monocytic cells

	set size	enrichment score	fdr q-val
Oxidative Phosphorylation	196.0	6.007957	0.000000
Interferon Gamma Response	196.0	5.730341	0.000000
Tnfa Signaling Via Nfkb	195.0	5.340952	0.000000
Interferon Alpha Response	94.0	4.657201	0.000000
Tgf Beta Signaling	53.0	3.071116	0.000000
Mitotic Spindle	199.0	2.924059	0.000000
Il2 Stat5 Signaling	192.0	2.663728	0.000263
Il6 Jak Stat3 Signalingf	NaN	NaN	NaN

Figure 12: Hallmark GSEA enrichment results on genes with highest difference in variance in tumor vs normal tissue in (A) NK and (B) monocytic cells. See Figure 4.5 for enrichment in T cells; complete lists of enrichments are presented in Table S5.

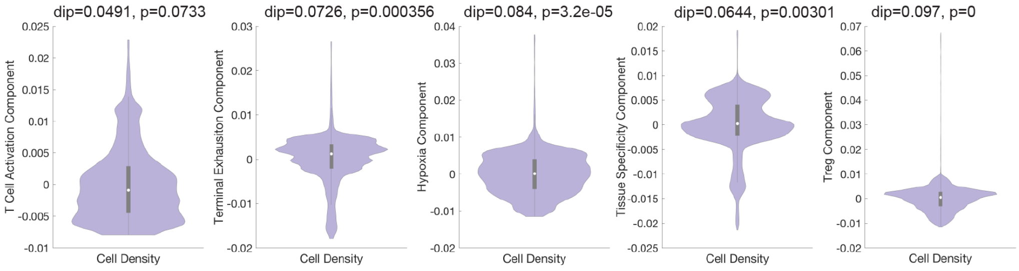


Figure 13: Hartigan's dip test on density of cells projected on diffusion components, showing statistically significant continuity (lack of "dips") in cells along T cell activation component (component 3, third panel from left), whereas other components exhibit more defined states (multimodality).

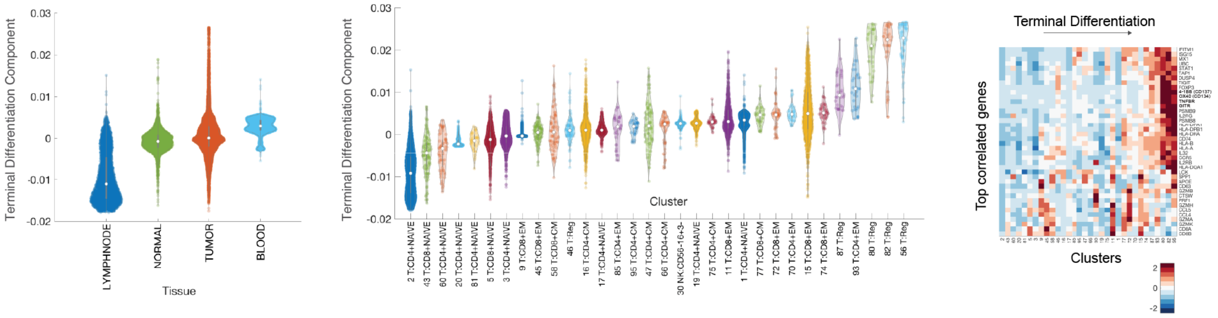


Figure 14: Violin plot of cells projected on terminal differentiation diffusion component: component organized by tissue type (left panels) and cluster (center panel). Also, heatmap showing expression of immune-related genes with the largest positive correlations with component, averaged per cluster and z-score standardized across clusters; columns (clusters) are ordered by mean projection along the component.

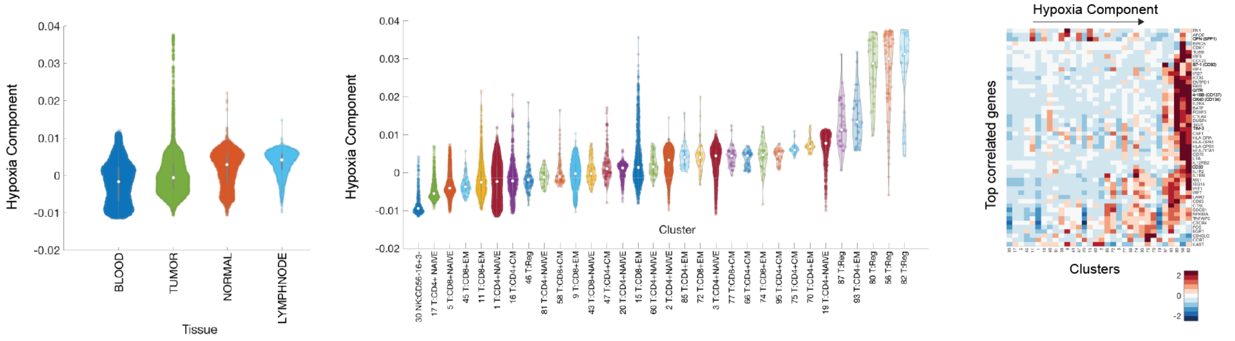


Figure 15: Violin plot of cells projected on hypoxia diffusion component: component organized by tissue type (left panels) and cluster (center panel). Also, heatmap showing expression of immune-related genes with the largest positive correlations with component, averaged per cluster and z-score standardized across clusters; columns (clusters) are ordered by mean projection along the component.

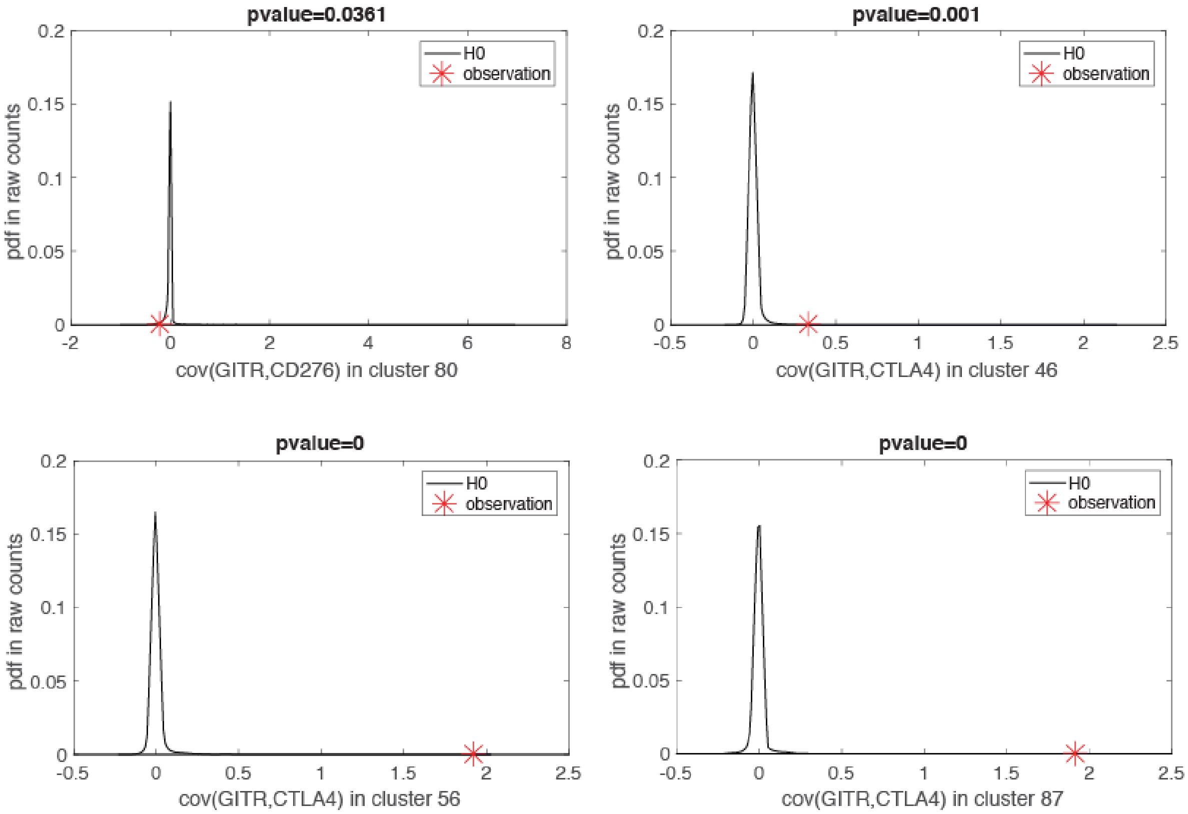


Figure 16: Displaying null distributions and observed covariances between CTLA-4 and GITR in raw, un- normalized data using hypothesis testing, subsampling, and permutation (see STAR methods); shows that the differences in covariance shown in biscuit-normalized data are also present in un-normalized and un-imputed data, and hence are not an artifact of computation.

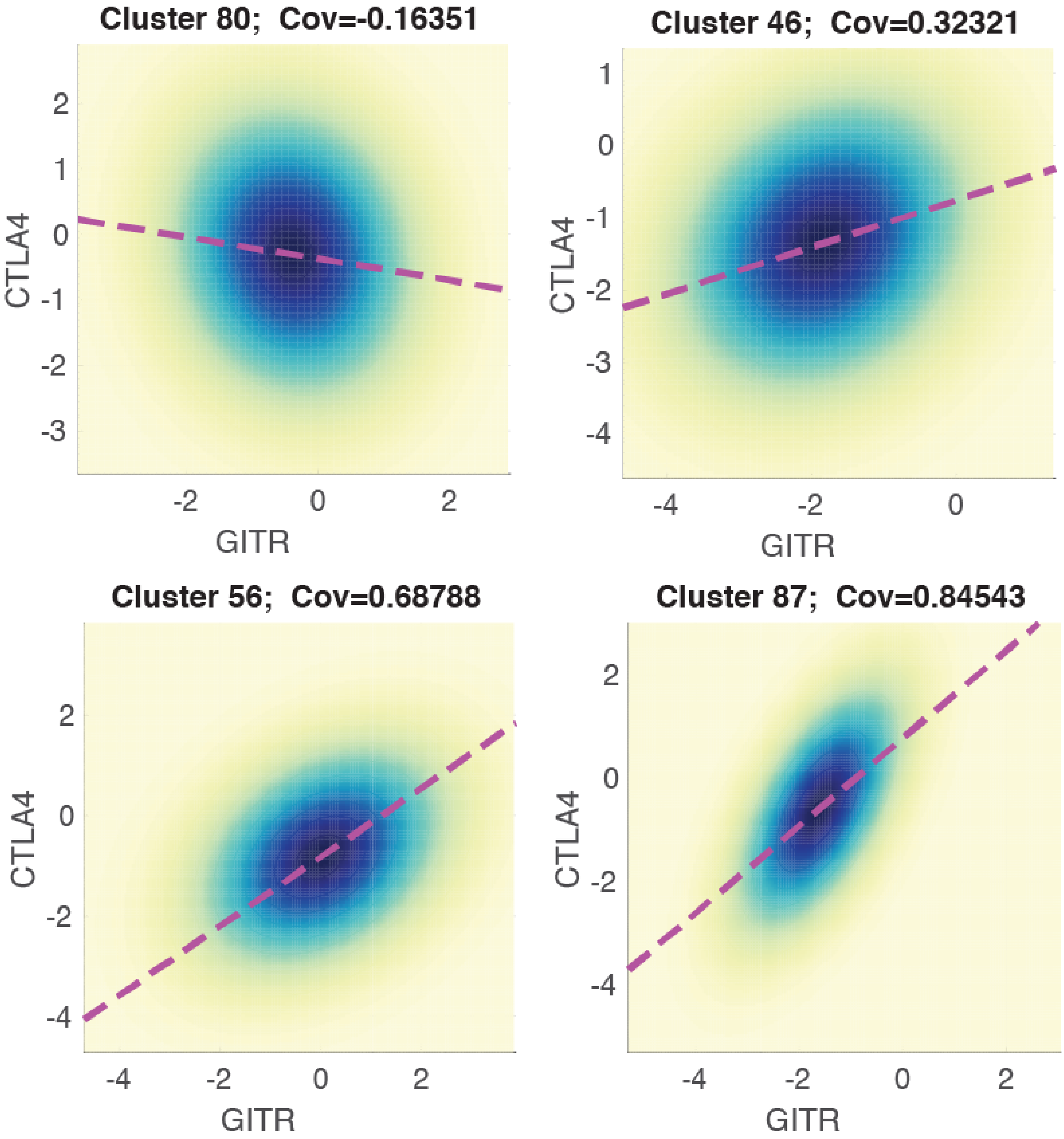


Figure 17: Bivariate plots of expression levels of GITR and CTLA-4 in Treg clusters based on inferred mean and covariance parameters from Biscuit. Dark blue color indicates the highest density of cells and light yellow the lowest density of cells.

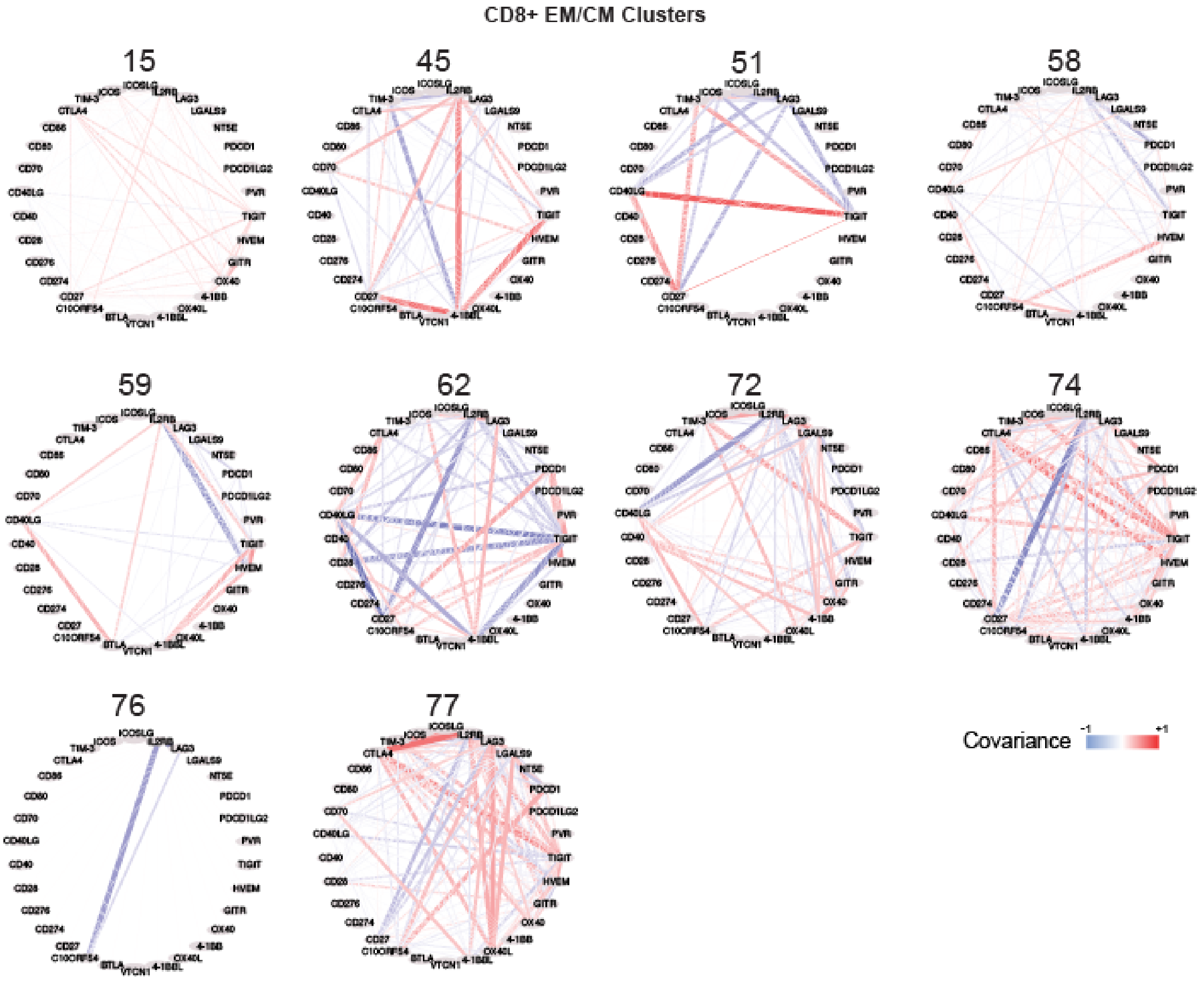


Figure 18: Network graphs showing covariance between checkpoint receptors in activated T cell clusters. Edge width denotes absolute magnitude (strength) of covariance and color denotes sign of covariance (red positive and blue negative). Note diversity across clusters.

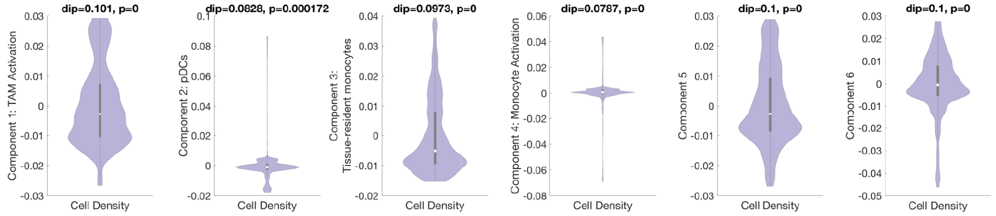


Figure 19: Hartigan's dip test on density of cells projected on diffusion components indicating no diffusion components across myeloid cells show statistically significant continuity, implying myeloid cells reside in defined (multimodal) states along major components explaining variation.

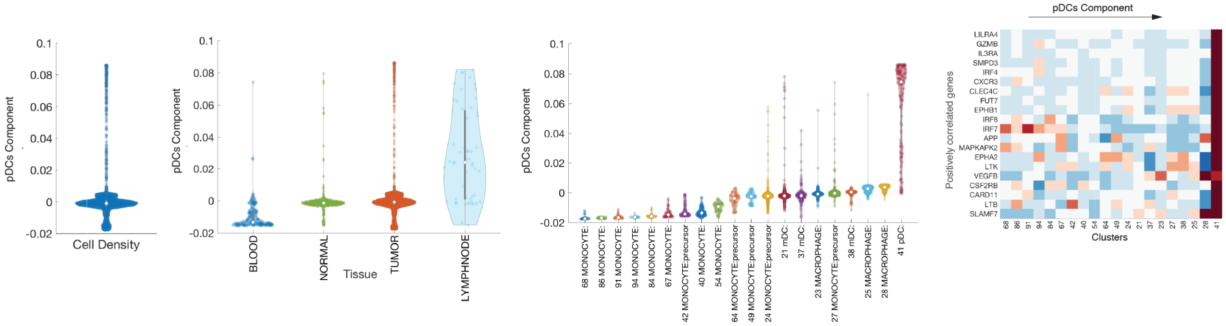


Figure 20: Violin plot showing the density of cells projected along pDC component and organized by tissue type and cluster.

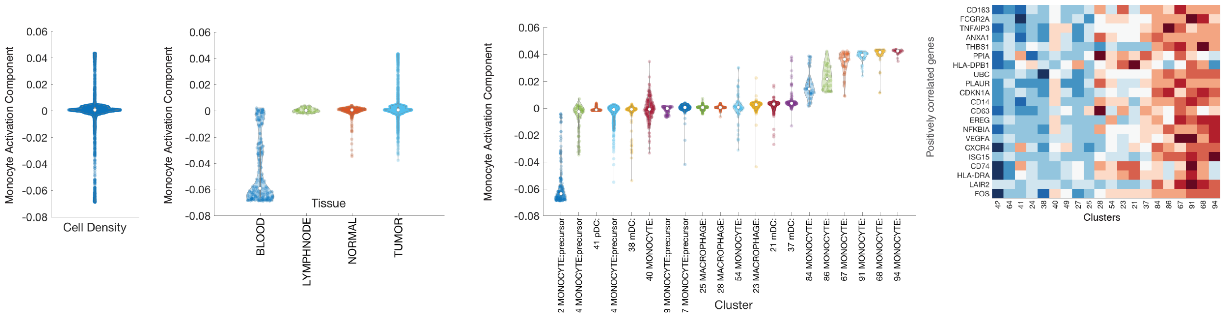


Figure 21: Violin plot showing the density of cells projected along monocyte activation component and organized by tissue type and cluster.

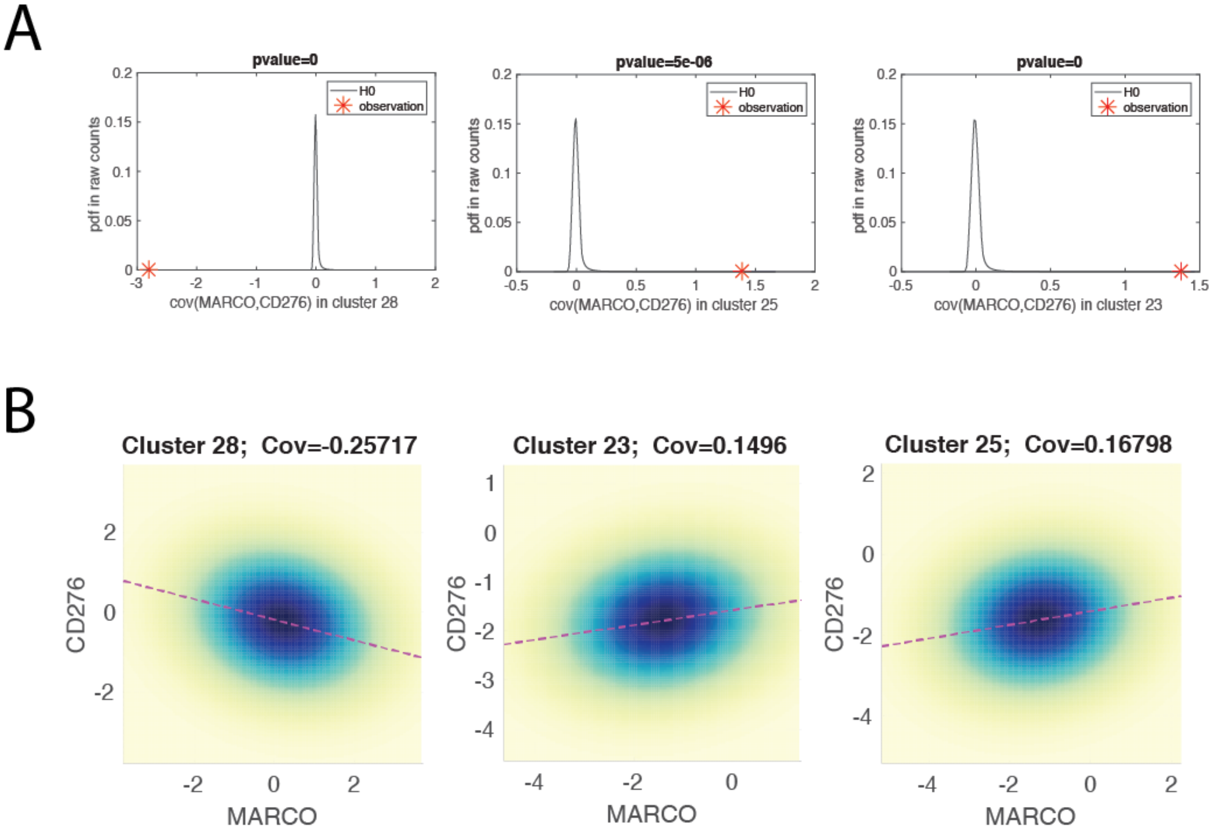


Figure 22: Displaying null distributions and observed covariances between MACRO and CD276 in raw, unnormalized data using hypothesis testing, subsampling, and permutation (see STAR methods), showing that the differences in covariance in normalized data as shown in Figure 7B are also present in un-normalized and un-imputed data, and hence is not an artifact of computation. Bivariate plots of expression levels of MARCO and CD276 in Treg clusters based on inferred mean and covariance parameters from Biscuit. Dark blue color indicates the highest density of cells and light yellow the lowest density of cells.