LEVERAGING SINGLE-CELL GENOMICS TO UNCOVER CLINICAL AND PRECLINICAL RESPONSES TO CANCER IMMUNOTHERAPY

by
E. Davis-Marcisak

A dissertation submitted to The Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy

Baltimore, Maryland
June 2022

Abstract

Immune checkpoint inhibitors (ICIs) provide durable clinical responses in about 20% of cancer patients, but have been largely ineffective for non-immunogenic cancers that lack intratumoral T cells. Most tumors have somatic mutations that encode for mutant proteins that are tumor-specific and not expressed on normal cells (termed neoantigens). Cancers, such as melanoma, with the highest mutational burdens are more likely to respond to single agent ICIs. However, most cancers, including pancreatic ductal adenocarcinoma (PDAC), have lower mutational loads, resulting in fewer T cells infiltrating the tumor. Studies have previously demonstrated that an allogeneic GM-CSF-based vaccine enhances T cell infiltration into human pancreatic cancer. Recent work with Panc02 cells, which express around 60 neoantigens similar to human PDAC, showed that PancVAX, a neoantigen-targeted vaccine, when paired with immune modulators cleared tumors in Panc02-bearing mice. This data suggests that cancer vaccines targeting tumor neoantigens induce neoepitope-specific T cells, which can be further activated by ICIs, leading to tumor rejection. Currently, the impact of ICIs and neoantigen-targeted vaccines on immune cell expression states and the underlying mechanism of therapeutic response remains poorly defined. Comprehensive characterization of responding immune cells, particularly T cells, will be critical in understanding mechanisms of response and providing a rationale for combinatorial therapies. In this work, we develop innovative computational methods and analysis pipelines to analyze the tumor-immune microenvironment at single-cell resolution. We establish an algorithm to quantify differential heterogeneity in single-cell RNA-seq

data, demonstrate the use of non-negative matrix factorization and transfer learning algorithms to identify previously unknown and conserved ICI responses between species, and develop a novel algorithm to physicochemically compare single-cell T cell receptor sequences. We leverage these methods in various contexts to yield new insight into the biological mechanisms underlying positive immunotherapeutic responses in diverse tumor types, including PDAC.

Thesis Readers

Dr. Elana Fertig (Primary Advisor)
Associate Professor
Department of Oncology
Biomedical Engineering and Applied Mathematics and Statistics
Johns Hopkins University School of Medicine

Dr. Neeha Zaidi
Assistant Professor
Department of Oncology
Johns Hopkins University School of Medicine

For Joan.

Acknowledgements

The journey to this body of work has been a true adventure, filled with unexpected twists and turns, constant learning, and precious moments of discovery. This endeavor would not have been possible without the help of many people along the way.

First, I would like to thank my team of mentors Dr. Elana Fertig, Dr. Elizabeth Jaffee, and Dr. Neeha Zaidi for their unwavering support and generosity. Thank you for sharing your profound knowledge with me while also giving me the freedom to learn and grow. I would also like to thank my other thesis committee members—Dr. Loyal Goff and Dr. Kathleen Gabrielson, for their time and insights. I would like to express my deepest gratitude to Dr. Garry Cutting, who introduced me to translational research during my early years as a scientist. Your dedication to improving patients' lives will forever stay with me. I am especially grateful to Dr. Luis Martinez, who inspired and encouraged me to pursue research at a very important time in my life. I would also like to thank my colleagues in the Fertig lab and Jaffee lab for their friendship and guidance throughout this journey.

To my family, words cannot express my love and gratitude. I owe you everything. To my parents, your love has been a constant source of strength throughout my life. Thank you for believing in me and supporting my decisions, even when it meant moving across the country. To my husband, your love, patience, and support have been my foundation. Thank you for always giving me balance in the most unsteady of times. And to my dogs, who add so much joy to my life and always remind me to go outside. I can not thank you all enough.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Abstract

Single-cell technologies are emerging as powerful tools for cancer research. These technologies characterize the molecular state of each cell within a tumor, enabling new exploration of tumor heterogeneity, microenvironment cell-type composition, and cell state transitions that affect therapeutic response, particularly in the context of immunotherapy. Analyzing clinical samples has great promise for precision medicine but is technically challenging. Successfully identifying predictors of response requires well-coordinated, multi-disciplinary teams to ensure adequate sample processing for high-quality data generation and computational analysis for data interpretation. Here, we review current approaches to sample processing and computational analysis regarding their application to translational cancer immunotherapy research.

Introduction

Single-cell analysis has become a widespread tool used in cancer research to characterize the cellular and molecular composition of tumors[1–3]. Technologies to profile single cells are currently able to measure tumor heterogeneity across molecular levels, including DNA[4], RNA[5], protein[6], and epigenetics[7]. Whereas bulk technologies are limited to an averaged signal often representing the molecular states of the most

abundant cell populations, single-cell approaches resolve the cellular composition of the tumor microenvironment (TME). This characterization holds particular promise for the field of tumor immunology, as comprehensive profiling can determine the cell types and pathways involved in anti-tumor responses and immune evasion. In addition, recent spatial transcriptomics and proteomics approaches preserve tissue architecture, enabling the analysis of cell-to-cell interactions and cellular neighborhoods reflective of the interactions in immune responses[8]. Samples derived from immunotherapy clinical trials can benefit from using single-cell-based technologies to capture the nuances of therapeutic immune cell responses in cancer. The development of immune checkpoint inhibitors (ICIs) enhanced cancer therapy by providing clinical benefits to a portion of previously incurable cancers; however, most patients do not respond to ICIs[9]. Understanding the complex immune cell composition and molecular pathways associated with cell state transitions during these therapies can potentially identify mechanistic predictors of response and elucidate new druggable targets to overcome immunotherapy resistance[10].

Current single-cell technologies span a wide array of rapidly advancing methodologies, with the most common examples for tumor immunotherapy including single-cell RNA sequencing (scRNA-seq) for transcriptional profiling[5], mass cytometry (CyTOF) for proteomics profiling [6], and spatial molecular profiling[1, 11, 12] (Fig. 1-1). Each of these technologies provides a high-dimensional molecular profile for individual cells, which can be computationally sorted into distinct cell populations. These technologies profile more than the canonical cell-type markers that are commonly measured in multi-parameter flow cytometry experiments, for example. The high-dimensional nature of these approaches can enable more refined annotation of cell types, inference of cellular state transitions, and association of molecular pathways. These characterizations require complementary computational techniques to determine the pathways that drive the behavior of each distinct cell type and infer the intra- and intercellular

interactions associated with transitions in cell states. The inference of these pathways mirrors the current clinical research in tumor immunology, where precision medicine strategies are being developed to use combination therapeutics to rewire the TME to enable immunotherapy sensitization[13].

Single-cell and spatial technologies for immune profiling

Single-cell and spatial approaches can be used to examine tumors in great detail, characterizing cell-type composition and tumor heterogeneity by gene or protein expression[14, 15]. These approaches have already been implemented to profile the TME of multiple cancer types, including leukemia, melanoma, and breast, lung, and gastrointestinal cancers, among others (Table 1). Here, we summarize benchmarked technologies currently employed for high-dimensional characterization of tumors in the context of immunotherapy research (Table 2).

Single-cell proteomics

Fluorescence-based flow cytometry is currently the gold standard method for cell-type identification. It remains the most commonly used single-cell method for cell-type annotation and sorting in immunology[16]. Although this is a reproducible approach, fluorescence flow cytometry is limited by the number of features that can be simultaneously analyzed (up to 30 markers) due to the inherent limitations related to channel spillover and equipment throughput. Thus, a high-parameter study often requires complex compensation strategies or splitting panels into subpanels with redundancy of key markers to obtain high-dimensional single-cell proteomic characterization. Sampling strategies designed to increase the dimensionality of fluorescence-based characterization ultimately require larger numbers of cells, limiting application for patient biopsies, which have a limited number of cells[16].

Table 1. High-dimensional transcriptomics and proteomics technologies and application in human cancer studies

				Tissue				Cell type profiled			
	Technology	Reference	Cancer type	Tumor	Adj. normal	Lymph node	Blood	Immune	Tumor	Stromal	Therapy
Single-cell proteomics	CyTOF	Gadalla et al., (2019)	ovarian, melanoma, breast	X			X	X			–
		Subrahmanyam et al., (2018)	melanoma				X	X			anti-CTLA4, anti-PD-1
		Krieg et al., (2018)	melanoma				X	X			ANTI-PD-1
		Wu et al. (2020a)	pancreatic				X	X			GVAX, anti-CTLA4
Single-cell transcriptomics	scRNA-seq	Nagaoka et al. (2020)	gastric	X				X			anti-PD-1, anti-IL-17
		Kieffer et al. (2020)	breast	X						X	–
		Schelker et al. (2017)	ovarian	X				X	X	X	–
		Peng et al., (2019)	pancreatic	X				X	X	X	–
		Patel et al., (2014)	glioblastoma	X					X		–
		Tirosh et al., (2016)	melanoma	X				X	X	X	–
		Ma et al., (2019)	hepatocellular, cholangiocarcinoma	X				X	X	X	anti-CTLA4, anti-PD-1
		Bernard et al., (2019)	pancreatic	X				X	X	X	–
		Schlesinger et al. (2020)	pancreatic	X				X	X	X	–
		Sun et al. (2021)	hepatocellular	X				X	X	X	–
		Davidson et al., (2020)	melanoma		X	X		X	X	X	–
		Savas et al., (2018)	breast	X				X	X	X	–
		Zheng et al. (2017a)	hepatocellular	X	X		X	X	X	X	–
		Guo et al., (2018)	NSCLC	X				X	X	X	–
		Sade-Feldman et al., (2018)	melanoma	X				X	X	X	anti-CTLA4, anti-PD-1
Spatial proteomics	MIBI	Angelo et al., (2014)	breast	X				X			–
		Keren et al., 2018	breast	X				X			–
	IMC	Giesen et al., (2014)	breast	X				X			–
		Jackson et al., (2020)	breast	X				X			–
		Xiang et al., (2020)	NSCLC	X				X			–
		Ho et al., (2020)	hepatocellular carcinoma	X				X			cabozantinib, anti-PD-1
	CODEX	Schürch et al., 2020	colorectal	X			X	X			–
	MxIF	Yan et al., (2019)	melanoma	X				X	X		–
Spatial transcriptomics	spatial transcriptomics	Moncada et al., (2020)	pancreatic	X				X	X	X	–
Multi-omics	CITE-seq, scRNA-seq	Cadot et al., (2020)	chronic lymphocytic leukemia				X	X			ibrutinib
	CyTOF, scRNA-seq	Gubin et al., (2018) *	sarcoma	X				X			anti-CTLA4, anti-PD-1
	scRNA-seq, TCR, CyCIF	Yost et al., (2019)	basal cell	X				X	X	X	anti-PD-1
		Azizi et al., (2018)	breast cancer	X	X	X	X	X			–
		Jerby-Arnon et al., (2018)	melanoma	X				X			anti-PD-1
		Wu et al. (2020b)	NSCLC, endometrial, colorectal, renal	X	X		X	X			–
	scRNA-seq, spatial transcriptomics, MIBI	Ji et al., (2020)	squamous cell	X	X			X	X	X	–
	scRNA-seq, IMC	Aoki et al., (2020)	Hodgkin lymphoma	X	X			X			–

CODEX, codetection by indexing; CyCIF, cyclic immunofluorescence; MIBI, multiplexed ion beam imaging; MxIF, multiplexed immunofluorescence; NSCLC, non-small cell lung cancer.

Table 1-1. High-dimensional transcriptomics and proteomics technologies and application in human cancer studies.

As an alternative to fluorescence-based flow cytometry, CyTOF detects metal intensities from antibodies conjugated with isotopically enriched heavy-metal reporter ions. This design enables CyTOF to profile up to 50 markers simultaneously[17]. Based on the mass range of the reporter ions used when conjugating the antibody panel, CyTOF methods can theoretically be developed to detect >100 markers in the same cell to enable high-dimensional molecular profiling. Another advantage of the reliance on heavy-metal conjugated antibodies over fluorescence-based technologies

is the fact that they are rarely present in biological samples, eliminating analytical challenges resulting from false signals from intrinsic cellular background [6, 17]. As antibody-based technologies, both CyTOF and fluorescence-based cytometry can evaluate protein isoforms (e.g., CD45RO) and post-translational modifications (e.g., phosphorylation)[17]. CyTOF profiling relies on antibody panels that are ultimately limited by the number of isotopically enriched metals that can be reliably conjugated and is highly dependent on antibody quality. This reliance on pre-selected antibody panels restricts analysis to anticipated cell types, which limits the discovery of new cell types and molecular changes due to immunotherapy treatment[18]. Still, the metal reporters in CyTOF are robust to freezing and thawing and to a variety of fixation protocols, making this technology versatile in application and storage needs compared with other single-cell methods [19, 20].

The multi-parameter profiling of CyTOF makes it a powerful technique to understand variations in immune cell composition before and after immunotherapy (Fig. 1-1A). This technology has been used to model changes in the distribution of cell-type abundances in pre-clinical models[21] and peripheral blood mononuclear cells (PBMCs) of immunotherapy-treated tumors[22–24]. Notably, application of a panel of 40 markers for analysis of PBMCs from melanoma patients before anti-CTLA-4 or anti-PD-1 therapy revealed that PBMCs from anti-CTLA-4 responders were enriched for naive and effector T cells compared with non-responders. Among anti-PD-1 responders, central memory and effector memory T cells were more frequent, suggesting that different cell-type compositions are potential predictors of response to distinct immunotherapies[23].

Single-cell transcriptomics

Single-cell sequencing approaches perform genome-wide profiling of individual cells. As a result, they are not limited by pre-determined markers and can be applied to globally

characterize transcriptional profiles (scRNA-seq)[5], mutational burden (single-cell DNA sequencing)[25], and chromatin states (single-cell ATAC sequencing)[7]. Of these technologies, gene expression profiling with scRNA-seq is the most commonly used to identify cell types in the TME. In contrast to previous studies with bulk RNA-seq data, scRNA-seq profiling does not require experimental protocols to sort cells before sequencing[26]. The comprehensive, whole-transcriptome profiling of cell types with scRNA-seq allows for inference of cell state transitions, differential gene expression, and functional oncogenic and immunologic pathway analysis[27, 28] (Fig. 1-1B). Such analyses can be performed from scRNA-seq data directly using computational approaches.

Different technologies have been developed for scRNA-seq, and the choice of which platform to apply depends on the biological questions that need to be addressed. SMART-seq allows for single-cell analysis of full-transcripts of hundreds of cells that are sorted by fluorescence-activated cell sorting into microtiter plates for library preparations[29]. Massively parallel RNA single-cell sequencing (MARS-seq) also requires cell sorting, and sequencing is restricted to the 3 end of the transcript. MARS-seq introduced transcript tagging with cell-specific barcodes and a unique molecular identifier (UMI) that allows sequencing counts to be assigned to the respective gene[30]. Fluidigm C1 became an attractive option, as its microfluidic platform automated cell capture and increased the number of cells profiled from a few hundred to nearly a thousand[31]. This microfluidic platform allows full transcript sequencing or 3 sequencing. The development of droplet-based methods such as inDrop [32], Drop-seq[33], and the widely used 10X Genomics platform[34] increased the scalability of single-cell profiling. Droplet-based approaches allow thousands of single cells to be sequenced from an individual sample. In these methods, cells are captured and encapsulated in gel emulsion beads, inside which barcoding and UMI tagging occur. A limitation of these platforms is that sequencing will capture only the 3 or 5 ends of

the transcripts. Still, the barcoding strategies of UMI-based approaches enable the adaptation to single-cell multi-omics profiling across numerous molecular scales[35]. Concurrent profiling of protein and RNA with cellular indexing of transcriptomes and epitopes by sequencing (CITE-seq)[36], as well as T and B cell receptor (TCR/BCR) sequencing and RNA[37, 38], is particularly applicable to tumor immunology.

All scRNA-seq technologies can be used to characterize the cellular composition of tumors. Both the study cohort and the underlying biological question should determine which platform to select for analysis. Methods covering full transcripts (SMART-seq and Fluidigm C1) are ideal for identifying rare gene variants and splicing isoforms at a trade-off of profiling a relatively small number of cells. Therefore, these full-transcript technologies are ideally suited to high-resolution characterization of rare cell populations. UMI-based methods have higher cellular resolution, but lower molecular resolution, and are subject to signal dropout that can result in failed detection of genes. Still, the high-dimensional cellular profiling makes these UMI-based technologies more suitable than their full-transcript counterparts to annotate the diverse cell types in the TME and measure gene expression changes between treatment conditions[39].

Numerous scRNA-seq studies have examined tumor heterogeneity and identified new cell types or functional subtypes that are a result of tumor progression [15, 40–49]. Since tumor heterogeneity is crucial to understanding tumor evolution and anti-tumor immune responses, scRNA-seq has been extensively applied to tumor-infiltrating leukocytes in order to identify the immunosuppressive and effector cell types that populate different tumors and to associate cell types with specific transcriptional signatures to understand immune modulation[40, 43, 46, 48, 50–52]. scRNA-seq analysis has also been used to uncover the mechanisms driving resistance to immunotherapy[21, 53, 54]. For example, Gubin et al. performed both CyTOF and scRNA-seq profiling of tumors from a pre-clinical sarcoma model to assess the cellular composition and functional changes induced by different ICIs (anti-PD-1, anti-CTLA4, and the combi-

Table 2. Single-cell and spatial technologies sample requirements and data analysis opportunities

Technology		Single-cell transcriptomics			Spatial transcriptomics	
	Single-cell proteomics	scRNA-seq	snRNA-seq	Spatial proteomics	Slide-seq, 10X Genomics	SeqFISH, SeqFISH+, MERFISH
Sample	dissociated cells	viable dissociated cells	viable nuclei (fresh or frozen tumor)	frozen, FFPE	frozen	frozen, FFPE
Tissue type	tumor and normal tissue, blood	tumor and normal tissue, blood	tumor and normal tissue, blood	tumor and normal tissue	tumor and normal tissue	tumor and normal tissue
Genome coverage	pre-selected markers (~40 markers)	genome-wide	genome-wide	pre-selected markers (~40 markers)	genome-wide	pre-selected markers (up to 10,000 genes)
Data analysis						
Cell composition	Yes	Yes	Yes	Yes	Yes	Yes
Cell states	Yes	Yes	Yes	Yes	Yes	Yes
Differential expression	Yes	Yes	Yes	Yes	Yes	Yes
Pathway analysis	No	Yes	Yes	No	Yes	Yes
Cell-fate trajectories	No	Yes	Yes	No	No	No
Molecular interaction	No	Yes	Yes	No	Yes	Yes
Cellular interactions	No	Yes	Yes	No	No	No
scTCR/ scBCR	No	Yes	No	No	No	No

scTCR/scBCR, single-cell T cell receptor/single-cell B cell receptor; snRNA-seq, single-nucleus RNA sequencing.

Table 1-2. Single-cell and spatial technologies sample requirements and data analysis opportunities.

nation)[21]. The combined use of scRNA-seq and CyTOF allowed cross-validation of cell-type abundances associated with ICI response in different data modalities. The measurement of additional molecular parameters through scRNA-seq enabled de novo discovery of cell state transitions conserved between mouse and human tumors in data reanalysis by Davis-Marcisak et al., which included a subset of activated natural killer (NK) cells associated with anti-CTLA4 response[55].

Spatial analysis platforms

Single-cell approaches like CyTOF and scRNA-seq that are widely applied to characterize tumors rely on the profiling of dissociated tumor specimens, which results in the loss of the spatial organization of cells within a sample. New technologies that maintain the spatial organization of cells are essential to infer cell-to-cell interactions

within the TME. Thus, in recent years spatial proteomics and transcriptomics analyses are emerging as powerful tools to characterize the spatial distribution of cell types within a tumor. These technologies allow for direct measurement of spatial colocalization of cells, which is often associated with intercellular interactions. The emerging high-molecular coverage of these technologies enables further inference of the cellular and molecular pathways as well as cell state transitions associated with interactions between cells[56].

Chromogenic immunohistochemistry (IHC) has been the gold standard approach for clinical spatial proteomics profiling[57]. However, it has limited multiplexing capacity (four markers), which represents a challenge for research into the comprehensive cellular composition of the TME. The development of fluorescent IHC increased the number of proteins that could be interrogated at the same time (eight markers), but, similar to fluorescence-based flow cytometry, the overlap between wavelengths limits the isolation of large numbers of proteins[58, 59]. Sequential IHC techniques were developed to profile up to 12 proteins simultaneously and then the samples can be stripped to allow for restaining, which increases the molecular resolution [60], but the number of markers is still limited by the quality of the tissue after multiple cycles of antibody stripping, which ultimately limits the resolution of these technologies. Recent advances have led to the development of protein multiplex technologies that allow the mapping of roughly 50 markers in the same section. Image mass cytometry (IMC)[11], multiplexed ion beam imaging[61], and cyclic imaging detection (codetection by indexing [CODEX], cyclic immunofluorescence [CyCIF], and multiplexed immunofluorescence [MxIF])[62–64] are approaches that can measure protein levels of up to 50 markers at the same time and provide the spatial distribution of the signal as well as information on which cells are in contact with one another (cell neighbors) (Fig. 1-1C).

High-dimensional spatial proteomics technologies have been applied to characterize cellular interactions in melanoma[65], breast[66, 67], colorectal[8], cutaneous squamous

cell carcinoma[56], Hodgkin lymphoma[68], liver[69], and lung tumors[70]. In the context of immunotherapy treatment, Ho et al. leveraged IMC to identify cellular neighborhoods containing B cells, helper T cells, and CD68+CD163 myeloid cells, suggestive of an immune response in an immunotherapy-responsive liver tumor[69]. In the case of immunotherapy-treated melanoma, Jerby-Arnon et al.[53] used tissue CyCIF (t-CyCIF) to demonstrate that tumor cells can express markers that decrease T cell infiltration, creating immune cold cell neighborhoods that are detectable prior to immunotherapy initiation.

Similar to the comparison between CyTOF and scRNA-seq, spatial transcriptional (ST) profiling provides higher molecular resolution than spatial proteomics technologies and is reviewed in detail by Maniatis et al.[71]. Approaches such as Slide-seq and the 10X Genomics Visium platform enable whole-transcriptome characterization within spots on a slide that provides near-single-cell resolution in fresh-frozen samples (Fig. 1-1D). These technologies use specially designed slides spotted with DNA-barcoded beads (Slide-seq)[72] or oligo-dT/UMI tags (10X Genomics)[30] that will capture the tissue RNA on the slide. The barcoded spots are around 50–100 m in size, allowing 2–10 cells to be captured in each spot, and the sequencing counts will refer to the population of cells mapped to the slide spots. Computational deconvolution methods to estimate the molecular profile of single cells from each spot are currently an active area of research. Even though the technology lacks single-cell resolution, it is still possible to identify cellular neighborhoods and the cell types frequently interacting within such niches directly from the expression profiles of the spots[73] (Fig. 1-1D).

The development of high-dimensional RNA in situ hybridization technologies led to single-cell-resolution ST analysis with near-genome-wide capabilities. Although these in situ approaches do not involve transcript sequencing, their ability to detect thousands of transcripts in tissues allows their classification as ST platforms[1]. Lubeck et al.[74] developed sequential fluorescence in situ hybridization (seqFISH), which uses

sequential hybridization and fluorescent signal detection for single-cell in situ RNA measurement of a few hundred pre-selected genes. The improved seqFISH+[75] allows for profiling up to 10,000 genes, nearing the resolution of the whole transcriptome, but still requires prior selection of the genes. Another in situ technique that provides accurate spatial single-cell resolution with genome-wide coverage is MERFISH[76]. MERFISH requires multiple steps of hybridization and imaging, resulting in extensive experimental labor, depending on the number of genes to be profiled[76]. Emerging multi-omics technologies, such as DBit-seq, allow for concurrent proteomics and transcriptomics spatial molecular profiling, merging the strengths of both spatial transcriptomics and spatial proteomics for cellular characterization[77].

Tumor sample processing for single-cell profiling in clinical research

Although single-cell profiling has spread rapidly in tumor immunology, the intensive sample processing required limits application to clinical specimens. Notably, the majority of non-spatial single-cell technologies, such as scRNA-seq and CyTOF, require viably dissociated cells for profiling[78] (Fig. 1-1A and B). The most commonly used methods for sample dissociation apply enzymatic-based digestion and heated incubation. The sample storage prior to dissociation, type of enzyme, and time of incubation all affect the single-cell profiling and must be optimized carefully for each tumor type[78]. Digested samples must consist of single cells upon microscopic examination, and accurate characterization of the molecular states can be achieved only for live cells, with viability greater than 70%. Nonetheless, dead cells can be filtered as part of pre-processing after analysis, allowing for lower cellular viability in the case of assays with high-throughput cellular characterization such as CyTOF. The requirement of viable cells for dissociation poses a further barrier for the analysis of samples that are most typically preserved non-viably, such as biopsies. In a clinical

environment, maintaining cell viability requires rapid sample acquisition from the surgical or clinical team, pathological assessment, transportation to the lab, tumor dissociation, sample resuspension, and sequencing library preparation. Thus, a highly coordinated routine is required to obtain, process, and preserve samples rapidly enough to maintain cellular viability—a challenging process for staff-limited groups and for multi-site clinical trials. An additional challenge posed by the need for this immediate profiling is that all single-cell technologies are subject to technical artifacts that arise from processing samples at different times, in distinct profiling batches, or by different technicians. In clinical research, biospecimens necessarily arise at the time of treatment, making it impossible to control for technical artifacts in experimental design in cohort studies or time-course profiling during treatment. These batch effects can be overcome by optimizing preservation protocols so that samples can be processed simultaneously or by including a control sample in each batch that can be used to correct for technical artifacts computationally. Alternative strategies such as flash-freezing for nucleus isolation and single-nucleus RNA sequencing have been shown to compare to scRNA-seq and are emerging as alternatives for single-cell analysis of cryopreserved samples that can overcome some of these limitations[79, 80].

Spatial molecular profiling relies on slide-based technologies that retain the cellular architecture, without requiring tumor dissociation. Requirements for sample preservation and preparation in spatial proteomic assays depend on the technology. Spatial proteomics can be performed for both frozen and formalin-fixed paraffin-embedded (FFPE) samples (Fig. 1-1C). Most current spatial transcriptomics approaches rely on frozen samples, with approaches to use FFPE samples under development[81] (Fig. 1-1D). The ability to profile FFPE-preserved samples enables clinical research on samples processed for long-term storage.

Computational pipelines for single-cell and spatial analysis

The high-dimensional nature of single-cell data makes computational pipelines a critical component for obtaining cellular and molecular interpretation. Analysis methods are advancing in step with new technologies, providing a wide range of pipelines to choose from. These diverse analysis methods fall under several main classifications that together enable biological interpretation (Fig. 1-2). First, the single-cell data from all platforms must be pre-processed from raw outputs into estimates of the molecular expression for each cell while removing poor-quality cells. Subsequently, the data are clustered and visualized with marker genes for annotation of cell types in distinct clusters. Next, differential expression analysis can estimate changes in molecular markers among and within cell types between treatment groups. For single-cell transcriptomics, the increased number of molecular markers allows for in-depth analysis of cell state transitions and intracellular gene-regulatory networks (GRNs). Single-cell network inference algorithms also include intercellular interactions, relying on indirect inference based on ligand-receptor pairs[82–84]. Finally, spatial molecular analysis algorithms utilize additional spatial statistics and neighborhood analysis for cellular colocalization that can provide more direct evidence of intercellular interactions[85–87].

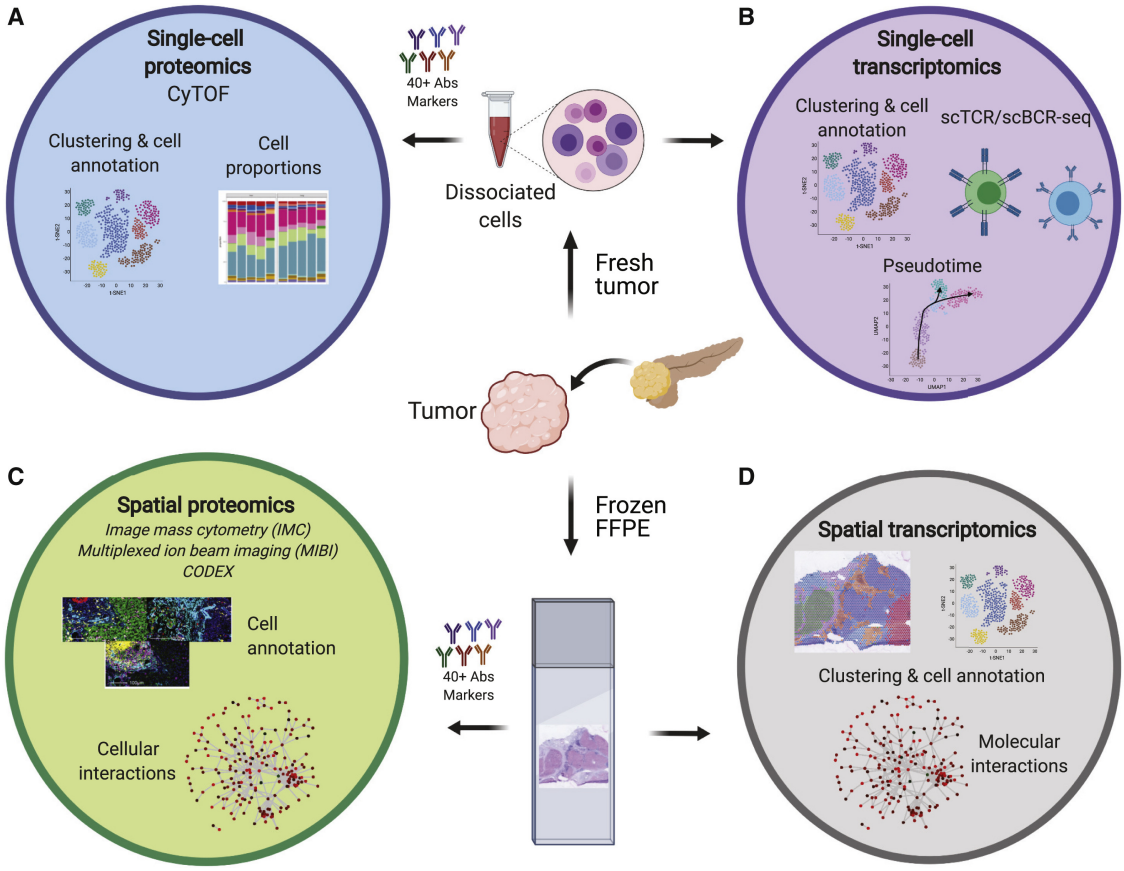

Figure 1-1. High-dimensional transcriptomics and proteomics approaches for cancer profiling. Several high-dimensional approaches are currently available to understand cancers' cellular composition and intercellular interactions. **(A)** Single-cell proteomics (CyTOF) provides cell composition and cell-state information. **(B)** Single-cell transcriptomics allows the same type of analysis, but its genome-wide coverage can also deliver cell trajectory predictions and T and B cell repertoires. To correlate cell composition and states to cellular interactions, spatial technologies are more informative than single-cell suspension analysis. scTCR/scBCR-seq, single-cell T cell receptor/single-cell B cell receptor sequencing. **(C)** With spatial proteomics and its single-cell resolution, it is possible to identify individual cell types and determine specific cell-to-cell interactions. **(D)** Although it lacks single-cell resolution, spatial transcriptomics can predict cell interactions based on the molecular expression of receptors and ligands between different cell neighbors and discover driving oncogenic pathways among the different cell niches because it is not restricted to previously selected markers. The selection of which approach to apply will depend on what samples are available, how they are preserved, and what biological questions need to be answered.

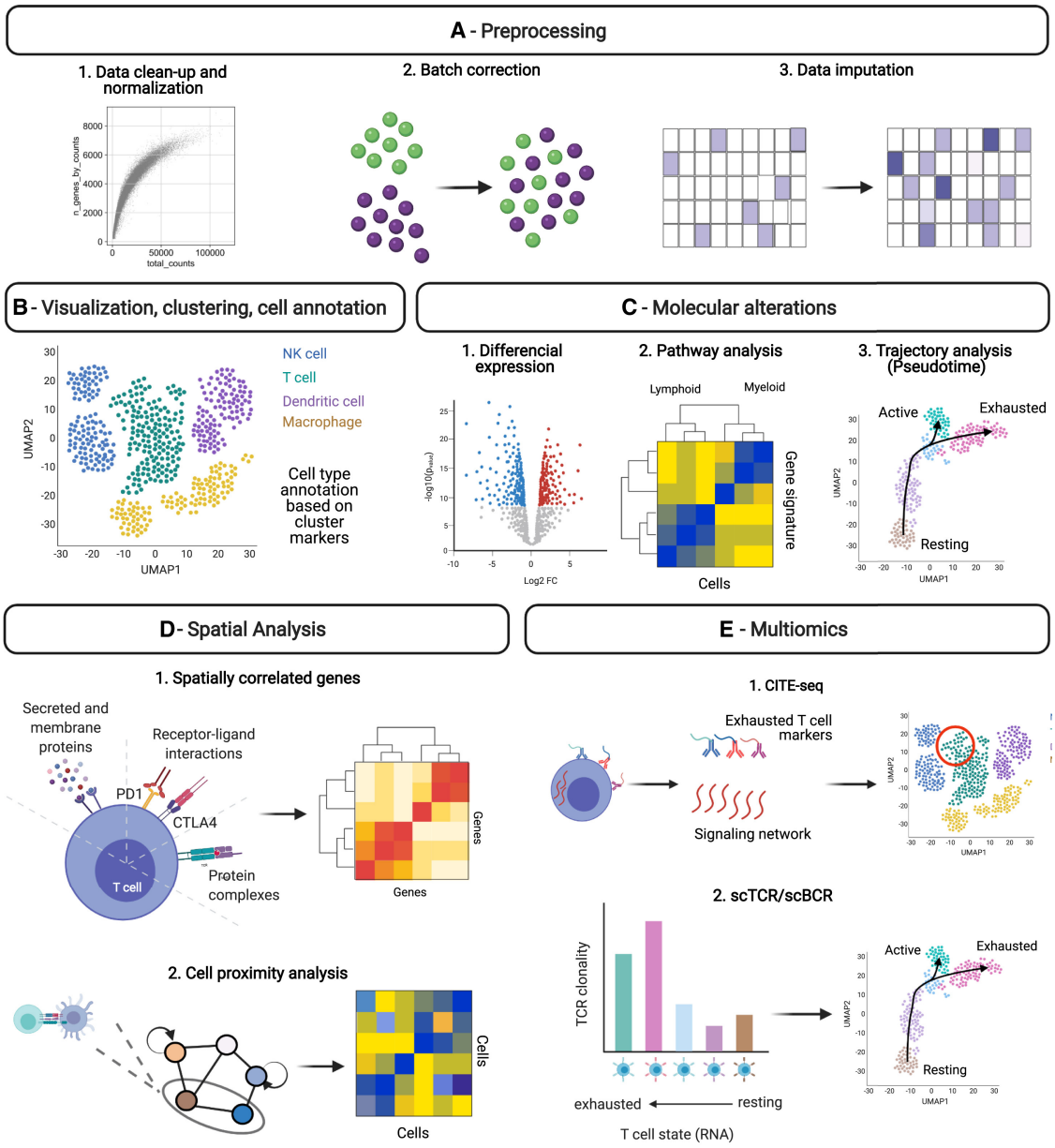

Figure 1-2. Computational workflow and methods for single-cell and spatial analysis. Several open-source benchmarked computational tools are available for high-dimensional dataset analysis. Independent of the tools of choice, analytical steps are required in order to obtain reproducible results and identify markers to predict response and targets for new therapeutics. **(A)** Single-cell and spatial data analysis will start with raw data pre-processing for (1) data clean-up to remove poor-quality cells and normalization to correct for low or high numbers of reads associated with experimental artifacts; (2) batch correction to remove unwanted variation among samples due to experimental discrepancies; and (3) data imputation to correct for the real data dropouts (zeros in the data). **(B)** Subsequently, dimensionality reduction will allow data visualization and cell-type annotation using clusterization tools that assign annotations based on specific markers expressed by each cluster. From there, the data are ready for downstream analysis depending on the methodology applied and biological questions.

Figure 1-2. (C) (1) Molecular alterations can be identified using differential expression analysis. In the case of transcriptomics data, it is also possible (2) to perform pathway analysis to identify drivers of cancer progression and responses to therapies and (3) to predict cell-fate trajectories to understand tumor and TME modulation across time. **(D)** From proteomics and transcriptomics data, it is possible to take a snapshot of the (1) molecular (e.g., protein marker expression, cytokine gene expression, receptor-ligand expression) and (2) cellular interactions (e.g., cell proximity analysis) that potentially drive the different features associated with cancer progression and response to therapies. **(E)** Finally, multi-omics approaches allowing (1) protein and gene expression analysis from the same samples (CITE-seq) or (2) T and B cell repertoire analysis in combination with transcriptional profiling add an additional layer of information that increases accuracy for cell-type annotation and for investigation of their role in cancer evolution and therapeutic responses.

While commercial software for single-cell analysis exists, the majority of analysis approaches are implemented in free, open-source software. To ensure broad adoption, this software is often built upon bioinformatics ecosystems such as the R/Bioconductor project that provide community standards and peer-review (Amezquita et al., 2020) or community-curated pipelines in R (i.e., Seurat, Monocle, and Giotto)[85, 88–95] and Python (i.e., Scanpy)[96]. Implementing these analysis pipelines can require extensive computer resources and computer programming experience. To make these pipelines more generally accessible, platforms such as Galaxy[97] and GenePattern Notebook[98] provide these methods in interactive, user-friendly interfaces for single-cell analysis with direct access to cloud computing. Further improvements in creating user-friendly databases, such as the developing CellxGene platform[99], remain an active area of development for the single-cell community.

Pre-processing and batch correction

The first step of single-cell and spatial data analysis is pre-processing the raw data output from each technology into measurements of protein or transcript abundances for each cell or spatial spot in the respective sample (Fig. 1-2A). All downstream analyses rely on these data summaries, making pre-processing critical to the accuracy of the resulting findings. The pre-processing approaches depend on machine-specific data outputs and biases, requiring techniques that are tailored to each technology.

Single-cell proteomics technologies typically output FCS files, following the standards of lower-throughput flow cytometry experiments. Whereas these FCS files are the final output of CyTOF, in spatial proteomics (i.e., IMC, CODEX) the data are obtained as images. Subsequently, a segmentation step is used to determine cellular boundaries prior to protein quantification and export into the FCS file format. The downstream analyses of FCS files require additional primary analysis steps to obtain protein abundances for each cell: bead-based normalization to standardize the

intensities for each signal, de-barcoding to isolate the cells for each experiment in a single batch if multiplexed, and, in some cases, compensation to account for spillover of signal between channels[100]. Altogether, this pipeline provides an estimate of normalized antibody intensities for each cell that can be carried forward to subsequent analysis of cell types.

Most single-cell transcriptomics technologies are sequencing based and provide FASTQ files containing short reads. The pre-processing of FASTQ files involves alignment to the human transcriptome and quantification of reads for each transcript or UMI, depending on the technology. Some software also performs quality control while pre-processing the raw data[101]. These alignment and pre-processing steps return a matrix containing barcodes specific to each cell captured and the counts for detected transcripts. Whereas single-cell proteomics relies on control beads to normalize the data, single-cell transcriptomics leverages the higher-dimensional nature of the data to derive a distribution for data normalization to correct for the overall differences in read depth for each cell[102]. In normalizing single-cell data, it is important to note that a value of zero read counts means either that a gene is not expressed in a cell or that it is randomly not detected among the sequencing short reads. Imputation methods were developed to estimate missing expression values and are well suited to gene-level visualization. However, these methods can introduce false positives into the data, potentially introducing statistical biases if they are used for downstream analysis[103]. Another step in pre-processing scRNA-seq is ensuring that barcodes refer to a unique cell and not to more than one cell that was captured in the same droplet (doublet) or to empty droplets (no cell). Those barcodes must be detected and filtered prior to analysis[86]. Likewise, dead cells quantified by quantifying the fraction of mitochondrial transcript counts relative to the total transcript counts must also be filtered for accurate analysis[104].

Spatial technologies based on imaging require an initial cell segmentation step

to isolate the cell boundaries in which protein or RNA abundances are estimated. Segmentation tools are under rapid development for both spatial proteomics and spatial transcriptomics, building upon frameworks developed for microscopy[105]. After segmentation, many of the same pre-processing and analysis methods for single-cell technologies can then be applied directly to estimates of molecular abundances. Sequencing-based, spot-level spatial transcriptomics technologies rely on alignment and quantification of the short reads associated with the barcode for each spot. These transcriptional profiles yield a semi-bulk estimate for all the cells captured in an individual spot. Spot-based deconvolution processes are required to estimate the transcriptional profile at single-cell resolution[106].

Normalization procedures are being developed to correct for discrepancies in molecular abundances or signal variation to ensure that cells obtained from a single processing batch (CyTOF) or library (scRNA-seq) in single-cell assays are comparable. Nonetheless, different experimental covariates can introduce unwanted bias, or batch effects, into the estimated molecular profiles from each data modality. Batch effects can arise due to differences in incubation periods during dissociation, handling personnel, reagent lots, or timing of sample processing. Batch effects are pervasive in high-throughput datasets and have been long recognized in previous generations of bulk technologies[107]. Technical noise is amplified in the case of single-cell technologies, requiring even greater attention to study design and batch correction methods to remove these technical artifacts[108]. The choice of normalization and batch correction methods can have a more substantial impact on the experimental results than the choice of downstream method for differential expression, making it a critical step in single-cell analysis pipelines[109]. Experimental designs that ensure each batch shares cells from the same biological condition allow remaining batch effects to be corrected computationally. Several batch correction tools have been designed to remove technical artifacts in the low-dimensional embeddings used to visualize single-cell data. Others

correct the data themselves by either leveraging the correlation structures between genes to preserve only biological variation or explicitly incorporating the batch as a covariate in the model. Batch-aware analysis pipelines should utilize batch correction techniques to standardize the data visualization and then model batch as a covariate for downstream differential expression analyses[110].

Visualization of data through low-dimensional embeddings

Single-cell and spatial data pre-processing, filtering, and normalization methods yield high-dimensional matrices representing an abundance of molecular species (proteins or transcripts as rows) by cells (columns). In the case of spatial datasets, the tissue position is added as an additional layer to these complex matrices and is used for visualization purposes. The high dimensionality of the matrices limits the direct application of standard data visualization techniques that often rely on two- or three-dimensional plots. Typically, the number of markers measured with these technologies is higher than the number of distinct biological processes (e.g., cell types, cell state transitions, etc.) captured. These features introduce correlations between the molecular species measured, which can be captured through a smaller number of features than the total number of markers in the data, enabling the use of dimensionality reduction techniques for visualization and analysis[111–113].

The most commonly used dimension reduction techniques for single-cell data, and also for spatial datasets, are t-distributed stochastic neighbor embedding (t-SNE) and uniform manifold approximation and embedding (UMAP) (Fig. 1-2B). Dimensional reduction techniques are the first step in analysis to enable visual inspection and data interpretation. Briefly, these methods transform high-dimensional data into a lower-dimension embedding for visualization. In the resulting plot, each point represents a single cell that is plotted using a computational method that ensures that the distance between cells along the coordinate axes corresponds to the distance they would have

from one another if computed for the entire molecular profile. In computing these distances, UMAP balances the global structure of more distant points, whereas the balance between preserving the distance of nearby points and more distant clusters is a tunable parameter in t-SNE[114, 115].

Both t-SNE and UMAP are well suited to visualizing clusters for distinct cell types within the data. Other manifold learning approaches, such as PHATE, have been designed with additional constraints to ensure that the embeddings not only model clusters but also preserve continuous transitions between cell states[116]. Overall, these embeddings provide a visualization tool to explore the variation and structure of the data, but require further analysis methods to infer biological insights from the representations. Most translational analyses select a single embedding, typically UMAP, that best distinguishes cell types and cell states in the TME. This embedding is used to anchor the visualization of results from subsequent analysis, coloring cells based on cell-type annotations or expression values for genes or proteins that significantly change due to treatment. In spatial single-cell analysis, after applying the same dimensionality techniques, these low-dimensional visualizations are often paralleled by visualization of the selected features directly on the tissue image.

Annotation of cell types in the tumor microenvironment

Accurate cell-type identification in scRNA-seq provides the first step to inferring changes in cell proportions between samples and from perturbations such as therapies. Gating strategies used to identify cell types in flow cytometry can also be applied to the proteins or genes in single-cell assays. However, gating approaches fail to realize the potential of the high-throughput profiling to comprehensively identify cell types present in the data, characterize cellular heterogeneity, and discover new cell types. Mirroring the mathematical assumptions of t-SNE and UMAP, cells of the same cell type can be expected to have similar gene expression profiles[114, 115]. Thus,

clustering algorithms are effective tools for cell-type identification (Fig. 1-2B). The large scale of single-cell data can require specialized implementations of clustering algorithms to ensure that these algorithms can run quickly, without requiring extensive computing resources. Many clustering algorithms employed for single-cell analysis leverage tools from social network analysis to identify groups of cells with similar molecular profiles and to mitigate noise from rare cells in the analysis[87, 117, 118]. Genes or proteins that are uniquely expressed in each cluster serve as marker genes that can be used to annotate the cell types associated with those clusters. These cell-type definitions and labels will depend on the number of clusters used for analysis. Determining this optimal number of clusters remains an open question. Indeed, the hierarchical nature of cell types (e.g., subclassification of lymphoid cells into B cells, T cells, and NK cells, and subsequent subclassification of CD8+ and CD4+ T cells) suggests that different dimensions will capture different granularity of cell-type delineation, reflected in emerging methods for ensemble-based clustering[119, 120]. Therefore, standard practice for cell-type assignment currently relies on an iterative process of clustering cells at multiple dimensions and assessing the expression of marker genes for known immunological and stromal populations in the resident tissue type. To avoid the manual nature of this approach, several methods have emerged to leverage reference cell databases to infer identities of individual cells or clusters[121]. By using curated signatures, cell-type annotation becomes robust and reproducible across studies. However, these signature-based methods will not identify cell types that were not previously included in the signatures, and the non-annotated cluster of cells will have to be manually verified and annotated.

As reference atlases of cell types emerge for tumors through projects such as the Human Tumor Atlas Network[122], the first waves of annotation will rely heavily on prior biological knowledge for classifying cell types. However, as new relationships between cell types are discovered, new tools will be necessary to help characterize novel

biology, and approaches to distinguish stable cell types from cell state transitions (e.g., between activated and exhausted T cells) remain an open area of research for single-cell analysis[28]. The common lineages of tumor cells with their normal counterparts can make them difficult to identify through marker genes or clustering analysis alone. To distinguish cancer cells from normal cells in scRNA-seq, copy number variation (CNV) analysis is a robust approach that detects large chromosomal variations (gains and losses of large DNA segments) by examining the gene expression distribution along chromosomes. These methods use RNA expression levels to infer DNA copy number at a given genomic region, which can separate cells with extensive CNV alterations, such as cancer cells, from diploid cells[15, 123–125]. To perform CNV inference, it is important to use methods designed to scale with the size of data being used. Early approaches for CNV inference[15, 123] were developed using first-generation scRNA-seq technologies (Fluidigm C1, SMART-seq)[29, 31], which have lower cell throughput than the more recent high-throughput technologies (inDrop, Drop-seq, 10X Genomics platform) [32–34]. The development of computational tools with improved speed and accuracy for large-scale datasets with sparse molecular coverage remains a critical area of research, with new approaches, such as CopyKAT, starting to emerge that are compatible with widely used high-throughput platforms[124].

Analysis of cell-type-dependent molecular changes

After cell-type identification is performed, functional changes from perturbations such as treatment can be determined through differential expression analysis comparing treatment conditions within each cell type (Fig. 1-2C). Briefly, these analysis methods compare the distribution of expression values for each protein or gene between treatment groups for the subset of cells annotated as a given cell type. The optimal statistical test for this differential expression analysis remains an open question, although approaches based upon negative binomial tests are emerging as providing the

best model of the distribution of the molecular abundances of both scRNA-seq[102, 126] and CyTOF data[100, 127]. Standard pathway analysis tools can then be applied to determine the molecular pathways that were altered based upon the results of these differential expression analyses[128, 129].

In patients with the same cancer type, tumor heterogeneity can contribute to dramatic differences in therapeutic outcomes. Thus, characterization of cellular heterogeneity within the TME is necessary to gain a deeper understanding of tumor progression and treatment. Metrics to assess differences in heterogeneity between sample groups detect molecular variability across overall transcriptional profiles[40] or at the pathway level[130, 131] within individual cell types. Immune cell populations within tumors can also be highly heterogeneous, making it valuable to use these methods to determine the heterogeneity among these cells as well.

Whereas differential expression analysis can infer molecular changes from cells of a pre-specified cell type, changes in molecular pathways and state transitions may occur for multiple cell types simultaneously, resulting in incomplete identification through these analysis approaches. In contrast, non-negative matrix factorization (NMF) approaches seek potentially overlapping, but low-dimensional patterns that contribute additively to the sources of variation in the data. As a result, they capture patterns that may co-occur, better modeling hierarchies in cellular lineages. Each of the patterns learned from matrix factorization analysis can represent a distinct biological process, which can be interpreted biologically through the gene weights of the corresponding features[112, 132, 133]. For example, NMF was used to identify NK cell activation in anti-CTLA4 response in our reanalysis of the scRNA-seq data from Gubin et al.[21, 55]. The gene signatures from these NMF approaches are often robust across multiple datasets, allowing for transfer learning approaches to identify the gene signatures associated with these inferred cell states in new datasets[112]. This approach has been leveraged for cross-species analysis relating pre-clinical and clinical models[55], and

indeed transfer learning is at the core of many supervised signature-based cellular annotations leveraging single-cell atlases. New non-linear approaches can also learn molecular changes from perturbations independent of cell-type annotations[134].

Trajectory inference and pseudotemporal ordering for cell state transitions

The heterogeneous nature of single-cell data allows us to observe not only the diverse cell types in the sample but also a range of molecular states within each cell type. While scRNA-seq data represent a single snapshot of the overall sample, they can still report on individual cells that correspond to a broad range of molecular states[135]. Trajectory inference methods computationally order the individual cells along a biological process according to their molecular states (Fig. 1-2B)[92, 136, 137]. Many trajectory inference methods also assign a "pseudotime" value to each cell that represents its relative position along the trajectory. This process allows us to observe gene expression dynamics and identify cell states on a continuum along biological processes more directly than the inferences of cell state transitions from NMF methods. Numerous trajectory inference methods have been developed in recent years, and they differ on the basis of their underlying algorithms, required prior information, and the expected topology (e.g., cyclic, linear, bifurcating) of the output trajectories. Although some recent methods[92] also infer the topology of the trajectory, most methods order cells along an assumed topology[136, 137]. Thus, the accuracy of the inferred trajectories is dependent on the choice of appropriate analysis method for the dataset and its associated biological process[138]. Since cancer datasets contain a heterogeneous mix of cell types, trajectory inference methods cannot be directly applied to the data. Instead, the common approach is to isolate certain cell types[50] and perform trajectory inference with respect to only these cell types.

The determination of cellular state in these analyses relies on successful trajectory

inference. The key challenge for successful trajectory inference is its dependence on the embedding from techniques, such as UMAP. As a result, they follow cell state transitions in cells only if the shape of that embedding matches the topology of the trajectory. Other dimension reduction approaches that explicitly model cell state transitions could be better suited to this inference. For example, RNA velocity[139] uses the spliced and unspliced mRNAs to calculate a high-dimensional vector representing the time derivative of the gene expression state of each cell in the dataset. RNA velocity has recently been generalized to model gene-specific kinetics[140] and cellular transport mechanisms for spatial transcriptomics data[76]. Estimates from RNA velocity can be used for more detailed visualization of the kinetic state of each cell using directional arrows in the low-dimensional embeddings. The length and direction of these arrows correspond to the high-dimensional RNA velocity vector of the cell. scMomentum[141] incorporates RNA velocity estimates computed by scVelo for predicting cell-type-specific directed GRNs. For every cell-type-specific network, an energy landscape is generated, where a cell's energy represents its differentiation potential. Extending the concept of RNA velocity, the first- and second-order kinetics of protein translation in single-cell multi-omics datasets can be estimated using protein velocity and acceleration[142]. Whereas the unspliced mRNA level of a cell is said to represent its future spliced mRNA levels, the current protein expression in a cell can represent the past spliced mRNA levels. The combination of protein and RNA velocity can be visualized as a curve calculated from the three points corresponding to past, present, and future values of the spliced mRNA, which represent the kinetics of the cell state. Overall, the trajectory inference or velocity analyses are relevant for identifying cell state dynamics and predicting cell fates from the analyses of a single "snapshot" in time, with potential to estimate the evolution of tumor and immune cells during cancer immunotherapy.

Inferring intra- and intercellular interaction networks from single-cell and spatial technologies

GRN inference is a key step to understanding the interactions between genes within and between cells, allowing for inference of the biological processes underlying molecular regulation (Fig. 1-2E). Numerous GRN inference methods have been developed in single-cell data with the goal of learning the structure of gene networks from data directly. Many approaches have been adapted from techniques that were originally developed for bulk transcriptomics analysis, which quantify network structure based upon the correlation between pairs of genes[143] or use machine learning methods to determine which genes can modify the expression profiles of one another[144, 145]. Newer methods have extended these approaches to specifically model the heterogeneity of single-cell data[146], including explicit extensions for time-course data[147]. Notably, the temporal ordering of cells by trajectory inference methods enables further inference of GRNs that can use the relative timing of gene expression changes to infer which gene controls the expression of another based on which is expressed first[148–150].

Whereas data-driven methods for GRN analysis infer intracellular signaling networks, regulatory processes may also occur between cells as through paracrine signaling or direct cell-to-cell interactions. Consider the case of interactions between dendritic cells (DCs) and T cells as an example of cell types that interact during the immune response. DCs are antigen-presenting cells that stimulate the clonal expansion and cytotoxic function of T cells[151]. To estimate these interactions from scRNA-seq data, a number of approaches attempt to infer intercellular interactions by identifying coexpressed ligand-receptor pairs[83, 84, 152] between cell types. The incomplete ability of transcriptional data to model receptor activation and the noisy nature of single-cell data pose limitations to the inference of intercellular signaling networks from single-cell data alone. Spatial molecular technologies provide a promising source of information to enhance these estimates by modeling intercellular interactions more

directly through observations of cellular colocalization. To that end, recently developed methods[153, 154] use spatial transcriptomic data to identify spatially informed GRNs and intercellular-signaling genes. Other interpretations of networks include spatially proximal or interacting cell types to recognize pairs of cells that have a higher likelihood of colocalization and spatially informed identification of coexpressed ligand-receptor pairs[85]. Thus, it is possible to infer the interactions between DCs and T cells from single-cell data through ligand-receptor-network methods, while the direct visualization of cellular colocalization from spatial datasets can confirm such interactions.

Single-cell multi-omics

One of the advantages of UMI-based scRNA-seq is the ability to attach additional oligonucleotide barcodes to cells to allow for concurrent measures of multiple molecular modalities in the same cell with single-cell multi-omics technologies[3]. A notable multi-omics technology for studying the TME is CITE-seq[36] (Fig. 1-2D). CITE-seq simultaneously obtains antibody-based proteomics and transcriptional profiling, combining the benefits of a priori identification of cell types using proteomics with the unsupervised analysis of scRNA-seq. This technology has been applied to monitor the temporal changes in PBMC composition during chronic lymphocytic leukemia therapy with the targeted agent ibrutinib, demonstrating clonal heterogeneity among leukemic cells and therapeutic perturbations in cancer and immune cells[155].

The behavior of certain immune cells can also be traced from multi-omics by using genetic identifiers. T cells and B cells undergo germline DNA recombination that results in a broad repertoire of TCRs and BCRs. Multi-omics technologies enable simultaneous transcriptional profiling of T cells and B cells and their respective receptors (Fig. 1-2D). TCR sequences can be acquired directly from platforms such as 10X Genomics TCR/BCR and paired transcriptome sequencing, or they can be inferred

from raw sequencing reads of scRNA-seq data by computational algorithms such as TraCeR[156], BraCeR[157], and VDJPuzzle[158, 159]. The availability of combined scRNA-seq and single-cell TCR-seq data in these approaches enables the association of cellular states with clonal expansion. In these analyses, the specific TCR and BCR profiling generates information about antigen-specific anti-tumor responses[40, 51, 54].

Paired scRNA and TCR-seq is gaining traction and being applied to a variety of cancers. A recent pan-cancer study of ICI-treated patients used this combined analysis to identify the clonal expansion of effector T cells in patients that respond to anti-PDL1 therapy[160]. In addition, expanded clonotypes were detectable across tumor tissue, normal adjacent tissue, and peripheral blood, suggesting a potential minimally invasive biomarker of immunotherapy response[160]. Clonotype information can also complement trajectory inference analysis of intratumoral T cells and B cells to track the dynamic relationships of these lymphocytes as they mount anti-tumor responses and respond to therapy. For example, a recent study using paired TCR- and RNA-seq to profile CD8+ chimeric antigen receptor (CAR)-T cells from the blood of patients undergoing CD19 CAR-T immunotherapy found distinct clonal transcriptional dynamics and expansion after adoptive transfer[161]. In basal cell carcinoma, combined scRNA-seq of CD8+ T cells treated with anti-PD-1 found an increased presence of activated and exhausted populations, as well as a hybrid population expressing activation and exhaustion markers[51], an expected effect of anti-PD-1 therapy[54, 162]. TCR analysis indicated that the largest clones presented exhaustion gene signatures. Also using TCR clonality, the authors were able to track those clones in pre- and post-treatment samples and observed that anti-PD-1 therapy did not convert exhausted T cells to a non-exhausted state. There was no expansion of the exhausted T cell clones, but new clonotypes, absent from the pre-treatment samples, were detected, suggesting that anti-PD-1 therapy attracts new T cells to the tumor with the potential to identify a new panel of antigens[51]. These findings

provide an immense contribution to understanding responses to ICIs and indicate that these therapeutic agents enhance the ability of tumors to attract additional T cells as opposed to reactivating exhausted T cells already present in the tumor[51].

While the ability to sequence both TCR chains provides an advantage over bulk single-chain methods, determining the antigen specificity of captured T cells remains a critical area of research. Advances into new multi-omics technologies are also actively being developed, with emerging methods for various combinations of proteomics, transcriptomics, spatial, and immune receptor profiling. These advances include technologies for resolving the multi-scale pathways in the TME, including intracellular phospho-proteomic states[163], intranuclear sequencing of transcription factors[164], chromatin[165], CRISPR-based screens[166], barcoding for lineage tracing of single cells[167, 168], and concurrent spatial profiling of RNA and proteins[77].

Chapter 2

Differential variation analysis enables detection of tumor heterogeneity using single-cell RNA-sequencing data

Abstract

Tumor heterogeneity provides a complex challenge to cancer treatment and is a critical component of therapeutic response, disease recurrence, and patient survival. Single-cell RNA-sequencing (scRNA-seq) technologies have revealed the prevalence of intra- and inter-tumor heterogeneity. Computational techniques are essential to quantify the differences in variation of these profiles between distinct cell types, tumor subtypes, and patients to fully characterize intra- and inter-tumor molecular heterogeneity. In this study, we adapted our algorithm for pathway dysregulation, Expression Variation Analysis (EVA), to perform multivariate statistical analyses of differential variation of expression in gene sets for scRNA-seq. EVA has high sensitivity and specificity to detect pathways with true differential heterogeneity in simulated data. EVA was applied to several public domain scRNA-seq tumor datasets to quantify the landscape of tumor heterogeneity in several key applications in cancer genomics such as immunogenicity, metastasis, and cancer subtypes. Immune pathway

heterogeneity of hematopoietic cell populations in breast tumors corresponded to the amount of diversity present in the T-cell repertoire of each individual. Cells from head and neck squamous cell carcinoma (HNSCC) primary tumors had significantly more heterogeneity across pathways than cells from metastases, consistent with a model of clonal outgrowth. Moreover, there were dramatic differences in pathway dysregulation across HNSCC basal primary tumors. Within the basal primary tumors there was increased immune dysregulation in individuals with a high proportion of fibroblasts present in the tumor microenvironment. These results demonstrate the broad utility of EVA to quantify inter- and intra-tumor heterogeneity from scRNA-seq data without reliance on low dimensional visualization.

Introduction

Tumor heterogeneity poses significant challenges in the clinical diagnosis and treatment of cancer. Variation can occur among tumors of the same histological subtype, giving rise to variability in therapeutic responses among patients. Cellular heterogeneity can also occur within tumors, allowing cancer to evolve over the course of disease progression, resulting in drug resistance, treatment failure, and disease recurrence[169–171]. An important source of tumor heterogeneity is the molecular variation among subclones and even individual cells within a tumor. This variation drives tumor progression through dysregulation of key cancer pathways and contributes to the evolutionary fitness of tumors[171, 172]. Differential variability analysis of bulk transcriptional data from microarrays and RNA-sequencing have also demonstrated that tumors with worse prognosis have a corresponding increase in transcriptional variation[173–176]. Single-cell RNA-sequencing (scRNA-seq) technologies provide an unprecedented ability to measure gene expression from individual cells, enabling in-depth exploration of tumor heterogeneity[177, 178].

Accurate characterization of inter-sample variation from scRNA-seq data of tumors

is critical to quantify tumor heterogeneity. Molecular heterogeneity of scRNA-seq data is often analyzed visually, using computational methods for dimensionality reduction that enable qualitative interpretations based upon the dissimilarity in transcriptional profiles between cells[111, 115, 179–185]. These techniques provide visual representations of the cellular composition within high-dimensional data. However, stochasticity, overplotting, and nonlinearity can challenge biological interpretation from visual analysis of scRNA-seq data. Moreover, the embeddings produced by these algorithms often rely on pre-selection of highly variable genes for clustering that may bias analyses based on variation. Highly variable genes are often identified based upon the coefficient of variation[186, 187] or dispersion[52, 93] of each gene across a cell population. Further gene set analysis of these statistics can be applied to quantify pathways or biological processes that contribute to cell-to-cell differences within a group of cells. Multivariate methods for analyzing transcriptional heterogeneity provide alternatives to quantify transcriptional heterogeneity from scRNA-seq data, such as PAGODA which quantifies overdispersion of annotated gene sets[131]. Similarly, phenotypic volume was introduced to quantify the variation between cells in a single sample[40]. These methods are all tailored to identify highly variable gene sets or samples across a population of cells from a single phenotype. Differences in variation within cells from a diseased population relative to variation within cells from a normal population may drive critical phenotypes, such as carcinogenesis or metastasis. Additional analysis techniques are essential to capture relevant pathway level heterogeneity that drives the observed deviations between groups of cells from distinct phenotypes.

In this paper, we extend our algorithm to quantify relative pathway dysregulation between experimental conditions from bulk transcriptional data[188] called Expression Variation Analysis (EVA) to scRNA-seq. Briefly, EVA provides a robust statistical test to compare the heterogeneity of transcriptional profiles of genes in a gene set between groups of cells from two phenotypes. We benchmark EVA using simulated

data and compare its performance to other methods to demonstrate the accuracy, robustness, and interpretability of the algorithm. With the recent outpouring of large scale scRNA-seq studies in cancer, publicly available datasets provide a breadth of transcriptional data to explore the role of heterogeneity in a variety of contexts. We utilize datasets from head and neck[47] and breast[40] cancers, which contain thousands of cells comprising dozens of cell types from different tissues, subtypes, and individuals. These datasets were selected to benchmark the performance of our algorithm to characterize cases with known differences in heterogeneity, such as between tumor and normal cells. Pathways found to be statistically significant from EVA are called differentially variable or heterogeneous between cells from distinct sample groups. These analyses enable novel characterization of the role of tumor heterogeneity in complex processes in cancer. For example, these analyses enable quantification of pervasive, differentially variable pathways between primary tumors and metastases consistent with the hypothesis of clonal outgrowth[189]. They also enable us for the first time to define the relationship between variation in immune pathways and TCR clonality. Finally, they quantify inter-tumor heterogeneity between primary tumors of a single subtype and identify immune dysregulation related to the degree of fibroblasts present in the tumor microenvironment (TME). Together, these results suggest that EVA provides an important tool to quantify inter-cellular heterogeneity directly from scRNA-seq data to yield novel biological insights.

Methods

EVA analysis

We use EVA from the R/Bioconductor package GSReg[188] version 1.17.0 to quantify pathway dysregulation in sets of cells from one group relative to the set of cells in another. Kendall-tau dissimilarities are computed with the function in the GSReg package and other dissimilarity measures using the R package philentropy version

0.2.0. Imputed scRNA-seq data are input to this algorithm, with imputation method described for each dataset below. P-values obtained from EVA analysis are FDR adjusted with the Benjamini-Hochberg correction and FDR adjusted p-values below 0.05 are called statistically significant. Additional details of our methods, including all code and datasets used to generate our results are available online.

Simulated datasets

We generate several simulated datasets to benchmark the performance of EVA, with varying degrees of complexity to balance controlled testing of the algorithm with the complex properties of scRNA-seq data. To examine the sensitivity of distance metrics to missing data, we simulate count data with different amounts of missing data using the squamous cell carcinoma bulk transcriptional dataset with a binary phenotype from the R/Bioconductor package GSBenchmark version 0.112.0. We randomly replace expression values with specified percentages of zeros to generate multiple datasets with varying degrees of missingness. We also generate a dataset with no signal by duplicating the transcriptional profiles for one phenotype. Again, we randomize zeros to determine the effect on the false positive rate in data without signal. We perform 100 iterations of all randomizations and test the performance against 35 distance measures.

While random zeros can be used to examine the general effect of missing data on dissimilarity, this does not accurately capture the nature of zeros in scRNA-seq data. To explore this, we simulate scRNA-seq data generated using the R/Bioconductor package Splatter version 1.0.3[190]. We first generate a simulated dataset with no signal to assess the dependence of EVA to missing data from scRNA-seq data. Count data was simulated for a single group of 100 cells and 10,000 genes using default parameters. A second group was simulated under the same conditions, with the parameter for dropout = TRUE. Merging these outputs resulted in a single dataset

with a population of cells equally distributed between two groups with identical transcriptomes and a varying number of zeros in one group. After imputation of this dataset, we randomize pathway expression profiles for each cell in one group to simulate data with differential heterogeneity in specified pathways.

For comparison of EVA to pre-existing methods we use PAGODA[131] from the R/Bioconductor package scde version 3.8. To generate simulated scRNA-seq data consisting of two groups with a high degree of differential variability, we modified Splatter so that the mean gene expression could follow a different distribution for each gene pathway and we added an additional variance parameter that allowed the variance to be specified for each cell group and pathway. We simulate count data for two groups containing 500 cells and 340 genes representing 10 synthetic pathways. We use default Splatter parameters with no added pathway variance for one group and set the added variance to 1 for each pathway in the second group to simulate differential variation between groups. We impute the simulated datasets described above for EVA analyses with the R package Rmagic version 1.3.0[191].

Public domain scRNA-seq datasets

We use 45,000 immune cells from 8 primary breast carcinomas with matched normal breast tissue, blood, and lymph nodes along with 27,000 T-cells with paired single-cell RNA and single-cell TCR sequencing previously described in Azizi et al.[40]. The scRNA-seq dataset from Azizi et al.[40] was previously imputed from their study using BISCUIT[192]. These datasets are available under GEO: GSE114727, GSE114724, and GSE114725.

We also use scRNA-seq datasets of 6,000 cells from 18 head and neck squamous cell carcinoma (HNSCC) patients containing 5 sets of matched primary tumors and lymph node metastases as previously described in Puram et al.[47] In our study, we impute the scRNA-seq data from Puram et al.[47] with MAGIC version 0.1.0

(Python) prior to analysis[191]. HNSCC subtypes present in the data were called using The Cancer Genome Atlas (TCGA) classification profiles on bulk data from primary cancer cells[193]. For inter-subtype comparisons batch effect correction was performed using the function ComBat from R/Bioconductor package sva version 3.26.0[194], considering each patient as a batch to isolate differences between cells from distinct HNSCC subtypes. This dataset is available under GEO: GSE103322.

TCR repertoire analysis

TCR repertoire clonality, richness, and Morisita-Horn similarity index between samples were computed on the TCR sequencing data from Azizi et al.[40] using the tcrSeqR[195] R package version 1.0.6 available from https://github.com/ahopki14/tcrSeqR.

Differential expression and gene set enrichment analysis

Differential expression analyses were performed across all expressed genes using the Monocle R/Bioconductor package version 2.6.1[95]. In all tests, the number of genes detected in each cell was included in both the full and reduced models as a nuisance parameter. Gene set enrichment was performed on differentially expressed genes with FDR adjusted p-values below 0.05 using the wilcoxGST function from the limma package version 3.32.10[196]. The alternative hypotheses of "up" and "down" were used to determine if genes were generally upregulated or downregulated, respectively.

Pathways and gene sets used in EVA and enrichment analyses

EVA and gene set enrichment analyses are performed for distinct sets of pathways appropriate for each analysis. Molecular signaling pathways are determined from the Hallmark gene sets in MSigDB version 6.1[197], meta-signatures defined from NMF analysis of the scRNA-seq data in Puram et al.[47], and the Myeloid Innate Immunity Panel pathways from NanoString (NanoString Technologies).

Code Availability

All code for the EVA analyses is available from https://github.com/edavis71/scEVA.

Results

The EVA algorithm provides a multivariate statistical framework to quantify differences in transcriptional heterogeneity between sets of cells from two phenotypes

EVA is a statistical algorithm designed to compare the relative heterogeneity of expression profiles within pathways between two phenotypes. It does this by computing the expected dissimilarity of expression profiles between any pair of samples from one phenotype relative to the expected dissimilarity of expression profiles between any pair of samples from another. When applied to the set of genes in a pathway, the expected dissimilarity between any pair of samples from one phenotype provides a measure of pathway dispersion or dysregulation. The difference of empirical estimates of the dispersion from a phenotype to another phenotype is called EVA statistics. EVA statistics test the null hypothesis that pathway dysregulation is equal between the phenotypes. Previously, we derived a computationally efficient approximation of p-values for these EVA statistics from U-theory statistics[188, 198]. Briefly, let xi denote the expression profile for sample i for the set of genes annotated to a specific pathway and d(xi,xj) the dissimilarity between the profiles for sample i and j for any dissimilarity metric d. EVA tests the null hypothesis that the E[d(xi,xj)] = E[d(xk,xl)], where E[] denotes the expectation and i and j index a pair of i.i.d. samples from one phenotype while k and l index a pair of i.i.d. samples from another. U-theory statistics provide an asymptotic approximation for the standard deviation of the dissimilarity measures for each phenotype as described in Asfari et al.[188, 199], resulting in an analytic framework to test the null hypothesis. The resulting EVA algorithm provides a robust, non-competitive gene set measure to quantify the relative

inter-phenotype heterogeneity of pathway usage. In our previous applications, we based our comparisons on the Kendall-tau dissimilarity measure in bulk transcriptional data. This measure was selected both because its rank-based nature reduces sensitivity to data preprocessing and models discordance between the expression of genes in a profile, which is indicative of pathway dysregulation. Bulk data lacks the resolution to quantify cellular heterogeneity because it is inherently an aggregate. EVA is poised to perform variation analysis based upon the measures of cellular heterogeneity in scRNA-seq data. If we treat each individual cell as a sample, we can adapt EVA to compare transcriptional heterogeneity scRNA-seq data between specified sets of cells (Fig. 2-1A and B).

Given that Kendall-tau dissimilarity is rank-based, it is robust to normalization and read depth. However, the abundance of zero counts from scRNA-seq data would lead to an increase of ties in the ranking. Moreover, dropout events in scRNA-seq data occur when an mRNA transcript is not captured by the library preparation reaction prior to sequencing and this generally happens more frequently in genes expressed at low levels. This, combined with the general bursting nature of the transcription machinery, leads to "false" zero counts, indistinguishable from biological zeros of truly unexpressed transcripts, and inappropriate rank assignments in the Kendall-tau dissimilarity.

Simulated data studies reveal varying sensitivities of distance metrics to missing data

The EVA algorithm defaults to comparisons based upon the Kendall-tau dissimilarity metric. Because this metric quantifies the number of gene pairs which switch ranks between two conditions, it directly quantifies how tightly a set of genes in a pathway are regulated[173, 188]. Previous work in simulated data studies for bulk RNA-seq data have shown that this algorithm performs optimally at detecting differences between

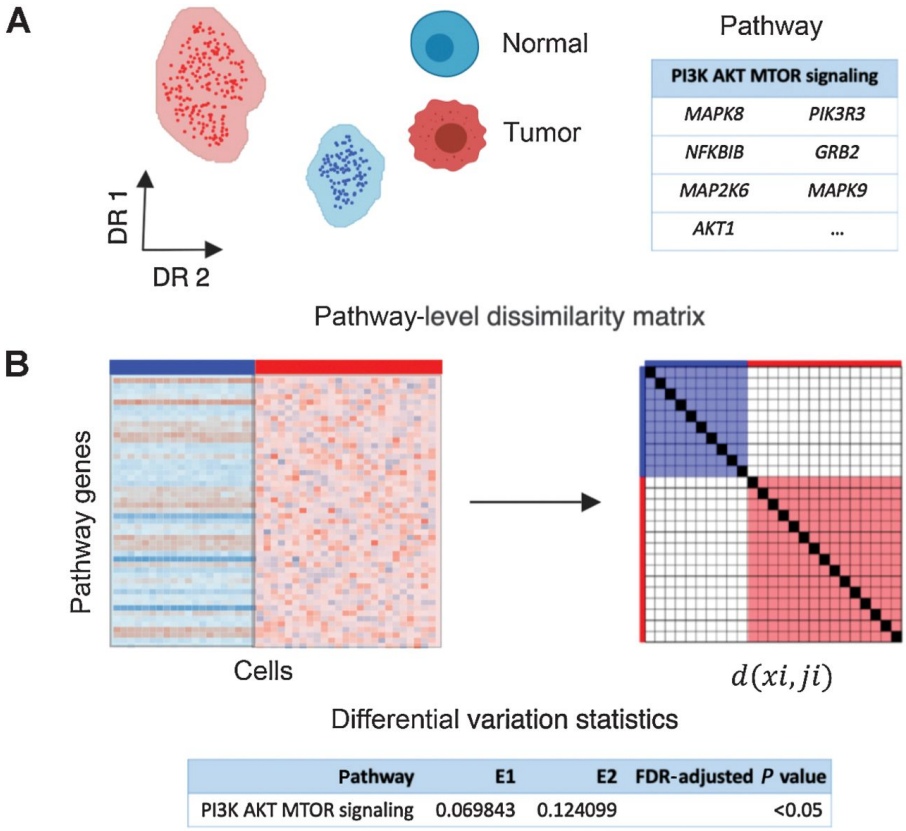

Figure 2-1. Overview of EVA algorithm to compare pathway-level transcriptional heterogeneity between groups of cells from two phenotypes. (A) EVA inputs a single-cell gene expression matrix for cells from two phenotypes, such as tumor and normal cells, and a list of genes annotated to a single pathway. **(B)** EVA extracts the expression profiles for pathway specific genes. It then computes the dissimilarity between the expression profiles for each pair of cells from the same phenotype using a user specified dissimilarity metric. Finally, EVA computes the expected dissimilarity between pairs of cells of each phenotype and U-theory statistics are applied to test the null hypothesis that the expected dissimilarity between pairs of cells from one phenotype is equal to the expected dissimilarity between pairs of cells in the other. The expected dissimilarity between pairs of cells from one phenotype is called the EVA statistic, which quantifies the inter-cellular heterogeneity for a given pathway. The U-theory statistics provide a robust estimate to quantify p-values that compare this relative heterogeneity between phenotypes.

phenotypes whose expression profiles vary in rank between within samples from a single phenotype, but does not detect mean shifts in expression profiles between the phenotypes[198]. In the case of single-cell data, it is critical to quantify the sensitivity of the algorithm to the dissimilarity metric used for analysis, missing data from dropout, and relative to other algorithms for pathway variation in scRNA-seq data. We have devised a series of simulated data experiments described in the Methods to evaluate the performance of EVA for each of these contexts. Briefly, we use the simulated data to examine the sensitivity of EVA to various distance metrics, the dependence of EVA to missing data from scRNA-seq, and to compare EVA to other available techniques.

First, the U-theory statistics to compare the expected dissimilarity between groups of cells from distinct phenotypes in EVA are general and can be applied to any dissimilarity measure. In order to compare the results of EVA using various dissimilarity measures, we use a dataset in GSBenchmark containing expression profiles of 22 matched samples from HNSCC tumor and normal tissues[200]. It was previously observed that pathways between tumor and normal samples are significantly dysregulated in this dataset[188]. We apply EVA using 35 dissimilarity metrics from Philentropy[201] and found that the number of significant pathways between tumor and normal vary widely across metrics (Fig. S2-1A). Several metrics including cosine and Ruzicka found no significant differentially variable pathways between normal and tumor samples. Kendall-tau detected the highest number of significant pathways, followed by Euclidean which is a commonly used distance to compare transcriptomes between single cells in visualization methods such as tSNE[115, 179–184].

We next examined the sensitivity of different dissimilarity measures to variable sparsity by randomly replacing transcription values with specified percentages of zeros. For each metric, the significant pathways calculated on the previously described data with no sparsity are used as our true positives to benchmark the performance. The

number of significant pathways varies greatly depending on the amount of missing data, with an overall loss of signal when the amount of zeros is the highest (Fig. S2-1B-D). While cosine initially found no significant differentially variable pathways, this metric detected the most false positives when count data was replaced with 80% zeros. Of note, Kendall-tau had the most consistency of the results without dropout in these simulations.

Altogether, these simulations demonstrate that each dissimilarity metric has varying degrees of sensitivity to missing data. We select Kendall-tau for the remainder of the analyses in this paper based on the observed accuracy in the two simulated datasets without additional normalization. We note the rank-based nature of the Kendall-tau dissimilarity renders the EVA statistics performed on Kendall-tau dissimilarity independent of common normalization procedures, such as log transformation.

EVA captures differential variation in imputed scRNAseq simulations

The previous simulated datasets were designed with random zeros to test the performance of EVA to missing data. Yet, dropout in scRNA-seq may not be missing at random. To determine the effect of dropout and imputation on EVA's robustness to detect pathway variability, we conducted an additional simulation study using synthetic scRNA-seq datasets generated using the Splatter pipeline[190]. We first examined the performance of EVA on a dataset with no signal and a bias in zeros. We used Splatter to generate a simulated dataset from two identical groups, one containing only biological zeros, and one where dropout was also present (Fig. 2-2A). Due to the abundance of zeros in the group with dropout and the sensitivity of Kendall-tau to missing data, EVA failed to recognize that the groups were otherwise identical and detected differential heterogeneity across 62% (31 out of 50) MSigDB Hallmark gene set pathway comparisons. We then imputed the missing values in the simulated

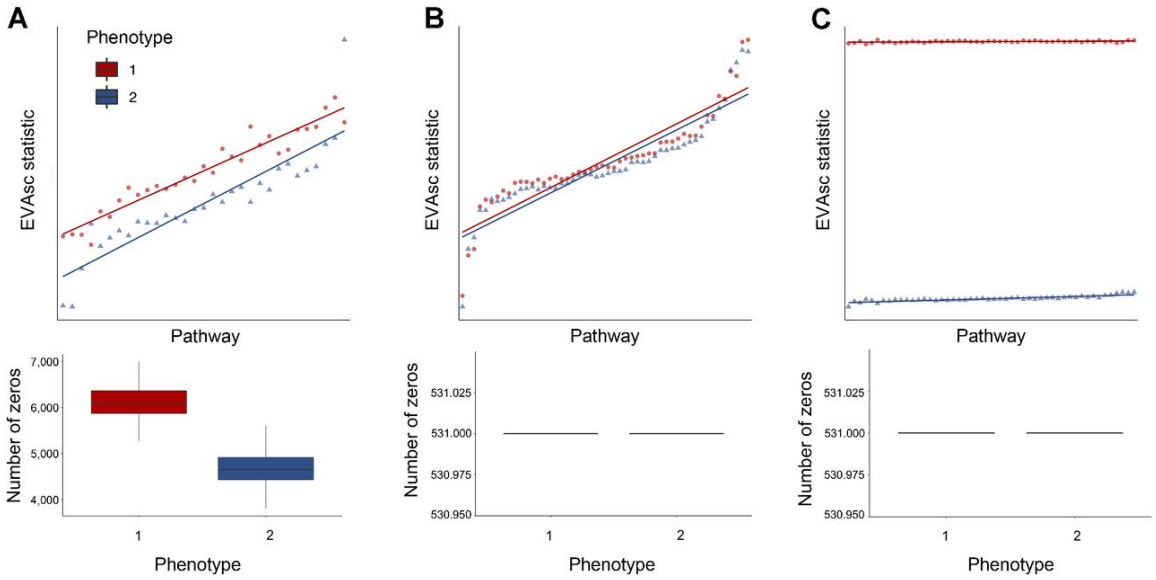

Figure 2-2. Performance of EVA with Kendall-tau dissimilarity on simulated data. (A) We apply EVA to a simulated dataset containing 50 pathways with no differential variation between cells from two phenotypes, but differential bias in their respective dropout rates. EVA statistics using a Kendall-tau dissimilarity have differential heterogeneity consistent with the simulated dropout rates. **(B)** After MAGIC imputation of the data from A, EVA finds no significant differentially variable pathways and EVA statistics overlap for the two groups. **(C)** We generate an additional simulated dataset by adding randomized signal to one group from the imputed data. The EVA statistics for significant pathways reflects the true heterogeneity in the simulated dataset.

dataset using MAGIC[191]. EVA analysis of this imputed data had no pathways with statistically significant differential heterogeneity between the two groups (Fig. 2-2B).

We next examined the performance of EVA to detect known differential variation in imputed scRNA-seq data. Using the previously described imputed dataset, pathway expression profiles for each cell in one group were randomized to simulate heterogeneity. EVA detected dramatic differences in variation between the two groups across all randomized hallmark pathways. 100% (50 out of 50) of the comparisons were statistically significant (Fig. 2-2C). These simulations demonstrate that EVA is able to assess the degree of pathway dysregulation between conditions in imputed scRNA-seq data.

EVA detects differential variation not observed by other methods

To benchmark EVA against previously published methods that are commonly used for variation analysis, namely coefficient of variation[186, 187] and PAGODA[131], we compared the ability of these methods to recognize highly variable pathways in simulated data. To generate simulated scRNA-seq data containing heterogeneity between groups, we modified the Splatter pipeline to include an additional variance parameter. This allowed for increased variance in the simulated gene expression between cells, specifically with the option to specify the amount of variance between groups. We then generated 10 pathway expression profiles for 500 normal cells and 500 tumor cells with different values for this variance. Each of the pathway-level comparisons were statistically significant when analyzed with EVA. In comparison, gene set enrichment performed on the coefficient of variation statistics identified no significantly variable pathways and no significantly overdispersed pathways were identified by PAGODA with adjusted z-scores greater than 1.96 (Supplemental table 1). While previously existing methods quantify the overall variation across transcriptional profiles of all cells, EVA is unique in determining differential variation between cells from two distinct phenotypes.

EVA detects greater variation in tumor than normal in scRNA-seq data from breast cancer samples

We next evaluated the ability of EVA to compare heterogeneity between normal and tumor samples in scRNA-seq data from breast tumors for distinct immune cell types[40]. Azizi et al.[40] reported an increase in the variance of tumor cell-intrinsic gene expression compared to normal breast tissue. Genes with the largest differential variance were enriched in signaling pathways important to the TME. To demonstrate that EVA enables robust statistical comparison of this heterogeneity in pathways, we

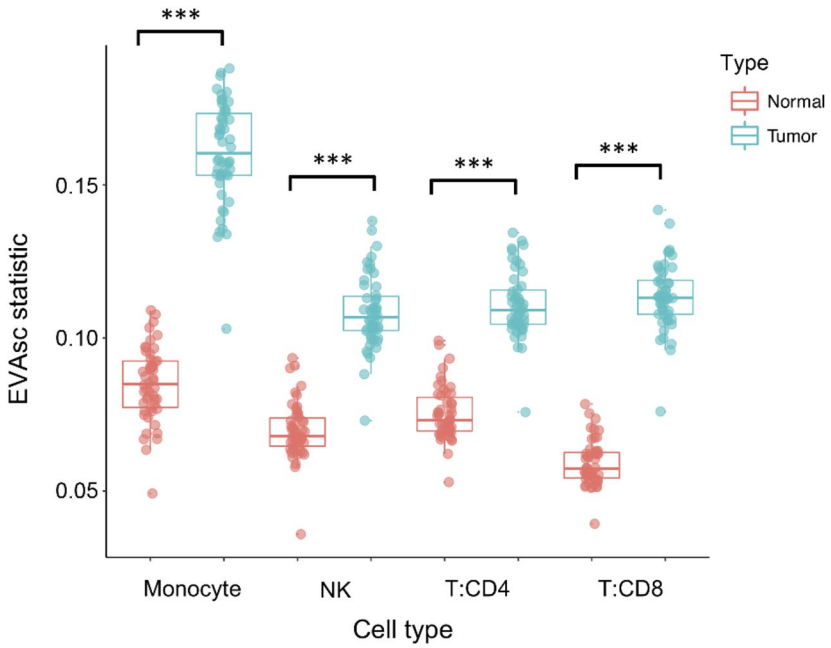

Figure 2-3. All pathways are significantly dysregulated in immune cell types from breast tumors relative to normal breast tissue. Boxplot of EVA statistics of inter-cellular heterogeneity for all 50 hallmark pathways in major immune cell types from both tumor (blue) and normal (red) breast tissue.

compared tumor to normal immune cells across multiple cell types, which included T-cells, myeloid, and NK cells. EVA analysis detected greater variation in breast tumor than normal breast tissue across each immune cell type tested. All 50 pathways tested were statistically significant in each comparison (FDR adjusted p-value < 0.05) (Fig. 2-3). This suggests that increased pathway heterogeneity within tumor-associated immune cell types may be driven by distinct TMEs present within a single tumor.

EVA finds increased immune pathway heterogeneity in tumors with high T-cell clonality

With the rapid increase of interest in the field of immunotherapy, T-cell receptor (TCR) sequencing is becoming a valuable tool for assessing immune response. Accordingly, we used T-cells from breast cancer data[40] to explore the relationship between the TCR repertoire and heterogeneity in immune signaling pathways using 27,000 T-cells with paired single-cell RNA and V(D)J sequencing from three breast cancer tumors.

For each individual tumor, we computed Shannon entropy for TCR clonality and richness as a measure of TCR diversity based on the single-cell TCR sequencing data (Fig. 2-4A). A Morisita-Horn similarity matrix was generated to compare the similarity of TCR repertoires across tumor replicates (Fig. 2-4B). We then applied EVA to the scRNA-seq data using the Myeloid Innate Immunity Panel pathways from NanoString (NanoString Technologies) to compare each T-cell subtype between individuals. Hierarchical clustering of the EVA statistics revealed a gradient of pathway dysregulation directly correlated with the degree of TCR clonality (Fig. 2-4C). We further applied GSEA to differentially expressed genes to compare the overlap between the enrichment of upregulated and downregulated immune pathways and the immune pathway dysregulation found with EVA (Supplemental table 4). The majority of the significantly dysregulated pathways from EVA overlapped with pathways that were enriched for upregulation in higher clonality compared to lower clonality individuals, with seven additional pathway comparisons that are significantly downregulated. We note that clonal expansion of T-cells is generally associated with a mounting immune response after antigen recognition. Our EVA results suggest that increased clonality of the TCR repertoire leads to increased heterogeneity in immune pathway expression as well as upregulated immune pathway expression.

EVA finds increased variation in primary tumors relative to metastases and subtype-specific pathway dysregulation

After demonstrating the ability of EVA to detect heterogeneity between tumors, we sought to characterize intra-tumor heterogeneity within primary tumors and associated metastases. Further, we aimed to identify differences in pathway heterogeneity between cancer subtypes, within subtypes, and within the TME. In order to make these comparisons, we applied EVA to scRNA-seq data for 18 HNSCC patients, including five matched primary tumors and lymph node metastases[47].

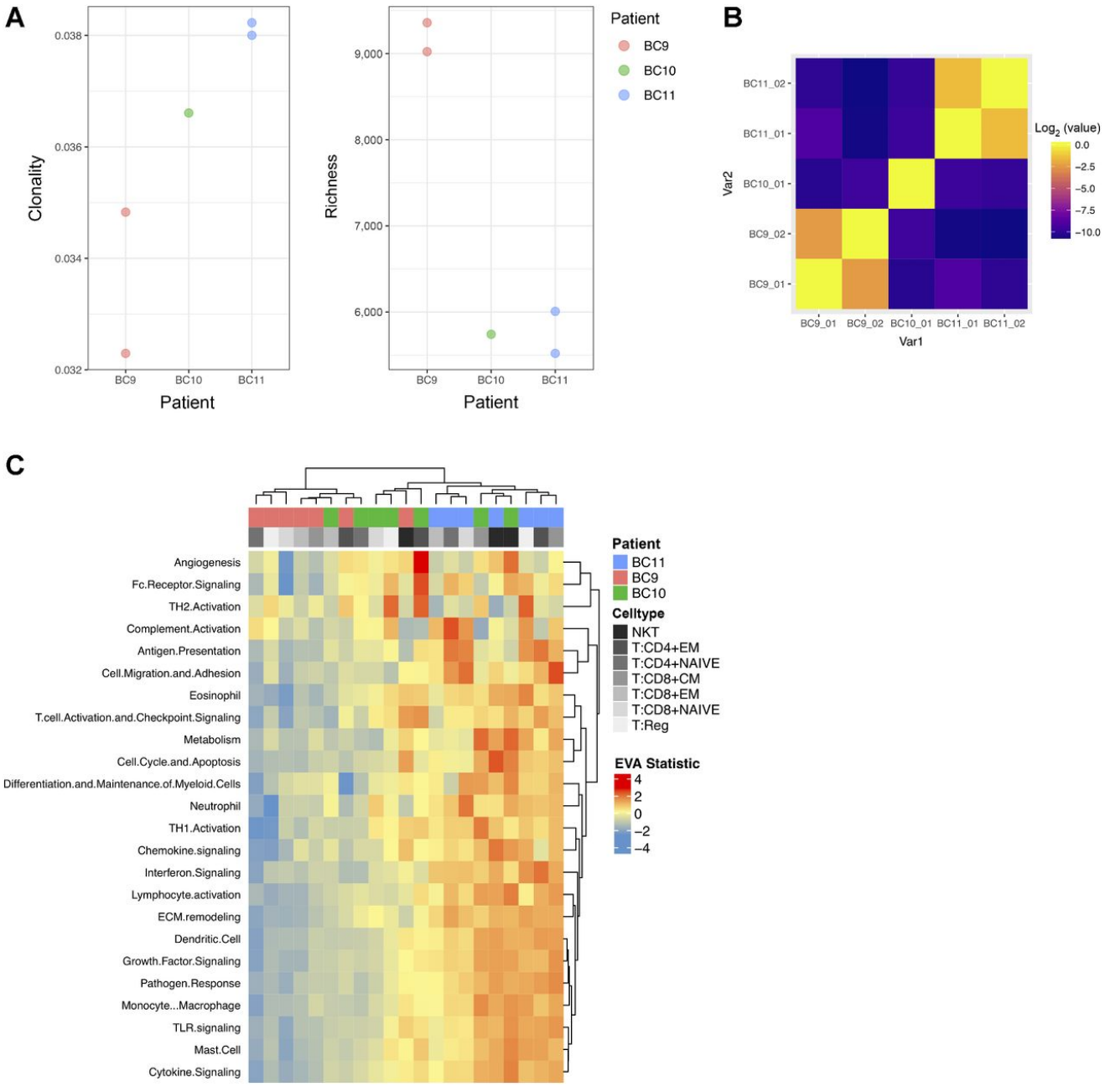

Figure 2-4. TCR clonality is associated with immune pathway dysregulation in breast tumors. (A) TCR clonality and richness for individual breast tumors with matched scRNA-seq and TCR-seq data. **(B)** Heatmap of Morisita-Horn similarity index to quantify agreement of CDR3 clonotypes from duplicate TCR-seq data for the same breast tumor and between individual breast tumors. **(C)** Hierarchical heatmap of EVA statistics of inter-cellular heterogeneity for immune pathways in each breast tumor T cell subtype.

We first applied EVA to matched primary and metastatic cancer cells within five individual HNSCC patients to examine intra-tumor pathway dysregulation. 98% (55 out of 56) of the pathways are statistically significant for patient HN25, with 100% (56 out of 56) statistically significant for patient HN26 (FDR adjusted p-value < 0.05) (Supplemental table 5). In both cases, all significant hits have greater variation in the primary tumor than the metastasis (Fig. 2-5A). For the remaining three patients, no significant pathway dysregulation was observed. Puram et al.[47] previously observed that the expression profiles of lymph node metastases overlapped with the corresponding primary tumors. While this indicates that there appears to be no mean differences between the paired samples, our method is able to capture significant differential variation between these phenotypes which was previously unrecognized.

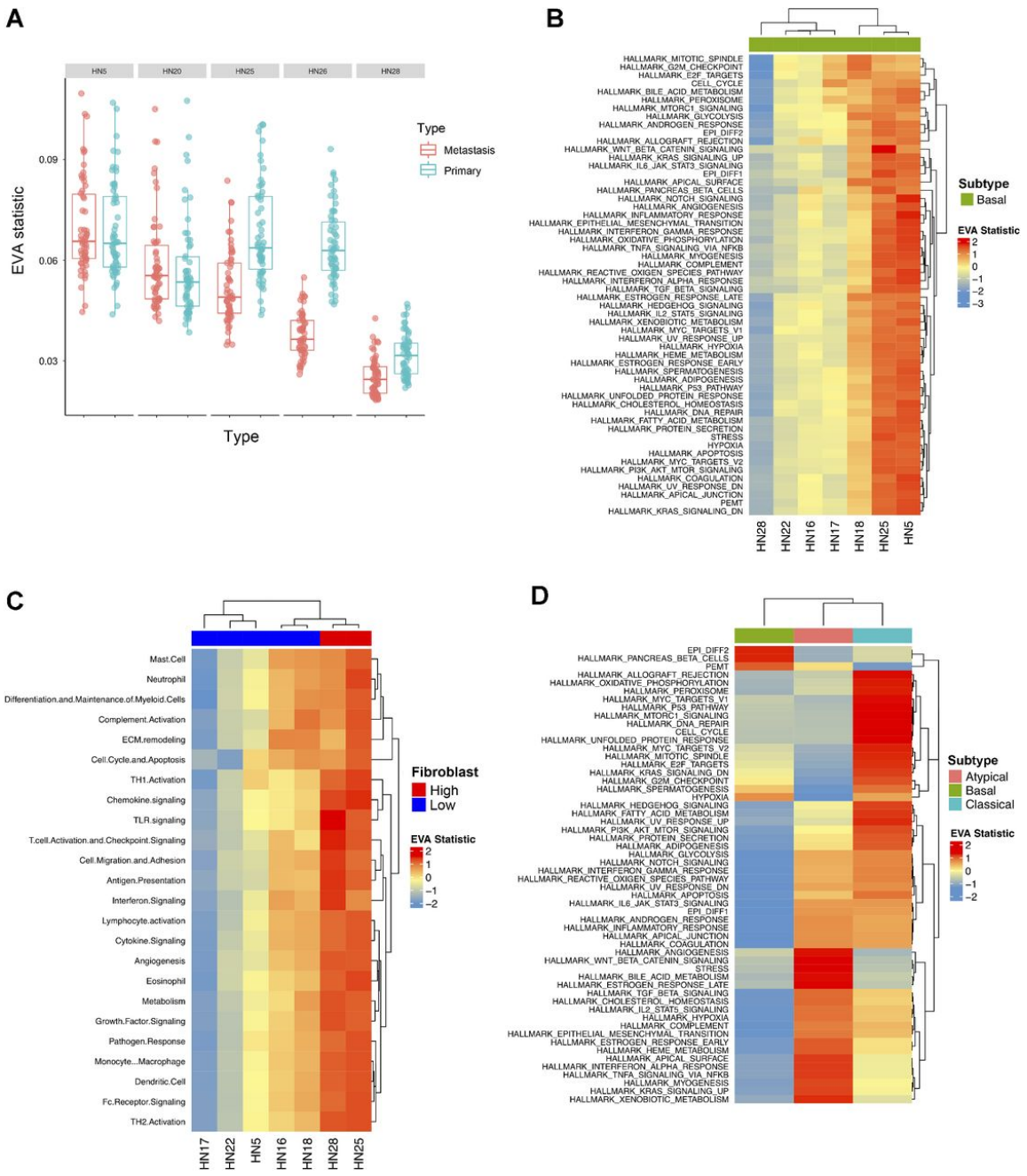

Figure 2-5. Inter- and intra-tumor heterogeneity distinguish HNSCC subtypes and metastases. (A) Boxplot of EVA statistics in primary and metastatic HNSCC cancer cells for each patient demonstrate higher inter-cellular heterogeneity in primary cancer cells than metastatic cells for two patients. **(B)** A heatmap of EVA statistics reveals that inter-cellular heterogeneity varies between primary cancer cells of the basal tumor type for all hallmark pathways, although no differences in mean expression were observed previously with tSNE[47]. **(C)** EVA analysis observes significant increases in inter-cellular variation of immune pathways for fibroblasts that are associated with the total fibroblast content in each basal HNSCC tumor.

Figure 2-5. Previous observations of TCGA subtypes noted that tumors with high fibroblast content (red) were classified as mesenchymal and low fibroblast content (blue) as basal, suggestive of fibroblast mediated differences between immune pathway activity in these subtypes. **(D)** Heatmap of EVA statistics of inter-cellular heterogeneity in hallmark pathways for cancer cells from patients in distinct HNSCC subtypes.

To determine the degree of inter-tumor heterogeneity between patients within a single subtype we compared primary cancer cells between seven individuals with basal primary tumors. 78% (923 out of 1176) of the comparisons are statistically significant when all pairwise combinations of patients were considered (FDR adjusted p-value < 0.05) (Supplemental table 6). EVA analysis revealed dramatic differences in pathway dysregulation across patients (Fig. 2-5B). Additionally, we explored heterogeneity within cells of the primary TME across individuals with basal primary tumors. Previously, Puram et al.[47] observed that the proportion of cell types within the TME vary for each patient. Notably, they found that the differences in the basal and mesenchymal subtypes of HPV-negative head and neck cancer can be attributed to a larger proportion of fibroblasts in the TME. Thus, we stratified these basal samples into a binary classification of high (>40%) or low-fibroblast (<40%). To determine the transcriptional status of immune-pathways within patient-specific fibroblast populations we applied EVA using the Myeloid Innate Immunity Panel pathways from NanoString (NanoString Technologies). Hierarchical clustering of the EVA statistics demonstrated increased immune dysregulation in individuals with a high proportion of fibroblasts present in the TME (Fig. 2-5C). 69% (348 out of 504) of the comparisons are statistically significant when all pairwise combinations of patients were considered (FDR adjusted p-value < 0.05) (Supplemental table 7). We note that the fibroblast composition in each basal tumor is independent of the pathway dysregulation observed across cancer cells from distinct patients.

We next applied EVA to primary cancer cells of HNSCC subtypes to examine the differences between inter-tumor heterogeneity. Subtypes were previously called by TCGA classification and ComBat[202] was performed to remove the impact of patient identity on transcriptional profiles. This batch correction enables EVA to compare cells from several patients to isolate only subtype-specific differences[194]. We include all MSigDB Hallmark gene set pathways and six meta-signatures derived

from non-negative matrix factorization programs that represent common expression programs variable within multiple tumor forms[47] in our comparisons. 46% (77 out of 168) of the comparisons are statistically significant when all pairwise combinations of subtypes were considered (FDR adjusted p-value < 0.05) (Supplemental Table 7). Hierarchical clustering of the EVA statistics demonstrated patterns of subtype-specific pathway dysregulation (Fig. 2-5D).

Discussion

We develop EVA to quantify heterogeneity in pathway level gene expression from imputed scRNA-seq data to quantify differential variability between conditions. We demonstrate the suitability of EVA for identifying differential variability of pathway gene expression by applying it to simulated and real scRNA-seq data. Simulated data generated with Splatter[190] was used to demonstrate the ability of EVA to detect known variability between conditions. Validation was performed by comparing immune cell types between normal breast tissue and breast tumors from Azizi et al.[40]. As expected, EVA detected increased variability in the tumor cells for all cell type comparisons relative to normal cells (Fig. 2-3).

We then applied EVA to perform novel analyses of differential heterogeneity on two publicly available cancer scRNA-seq datasets. We used paired single-cell RNA and single-cell TCR sequencing data[40] to compare inter-patient T-cell subtype heterogeneity in relation to TCR clonality. TCR repertoire analysis showed differences in the level of TCR clonality for each individual (Fig. 2-4A). EVA analysis revealed significant differences in immune pathway heterogeneity between individuals, consistent with the degree of TCR clonality: increased TCR clonality, increased heterogeneity (Fig. 2-4C). We then performed differential expression analysis between individuals to explore the direction of gene set enrichment. There was a large amount of overlap in differentially variable and differentially upregulated pathways, indicating increased

heterogeneity as well as increased gene expression in higher clonality individuals.

Ikeda et al.[203] examined the relationship between intra-tumor expression levels of immune-related genes and TCR repertoire in endometrial cancer. They found increased mRNA expression levels in cases with high T-cell clonality, which was associated with a better prognosis. These results were obtained using total RNA and quantitative real-time PCR in relatively few genes and are consistent with our findings at a comprehensive single-cell RNA level. Recent data has also shown that increased clonal expansion of T-cells and low baseline clonality are associated with longer survival after being treated with anti-CTLA4 inhibitors in pancreatic ductal adenocarcinoma[195]. Thus, characterizing the immune microenvironment by expression of immune pathways, immune pathway heterogeneity, and the clonality of infiltrated T-cell receptors may be an important biomarker for clinical response to immunotherapy. With the advent of paired single-cell RNA and TCR profiling methods, studying the transcriptional effect of TCR repertoire changes across cancer cells may provide further insight into the mechanisms of immunotherapy.

Further, an HNSCC scRNA-seq dataset from Puram et al.[47] was used to examine differences in heterogeneity between HNSCC tumor subtypes. Previously, bulk studies have classified HNSCC tumors into four distinct molecular subtypes based on their expression profiles[193]: atypical, basal, classical, and mesenchymal. EVA analysis revealed unique patterns of pathway dysregulation in each of the subtypes detected by TCGA classification (Fig. 2-5D). Overall, immune pathways are enriched in the atypical subtype. It has been reported that mesenchymal and atypical subtypes have the highest degree of immune infiltration, making them attractive targets for immunotherapy[204]. Our results suggest a key immune component specific to the atypical subtype.

Previous analyses of the HNSCC scRNA-seq data found that the cancer cells in the mesenchymal and basal subtypes have similar expression profiles when stromal

contribution was removed[47] and refined the classification of mesenchymal to basal subtype. We speculated that the cellular compositions of the TME within individual basal tumors could contribute to the molecular heterogeneity. Importantly, fibroblasts have opposing roles in the TME and showed a wide-range of inter-tumor proportional variability. Normal fibroblasts exert anti-tumorigenic effects to suppress tumor growth but can be reprogrammed to a cancer-associated phenotype supportive of tumor evolution. EVA analysis comparing fibroblast populations between individuals with basal primary tumors demonstrated that TMEs with a large proportion of fibroblasts have a high degree of immune pathway dysregulation. This indicated immune pathway heterogeneity within the fibroblast expression states, likely due to the immunomodulatory role of cancer-associated fibroblasts within the TME[205].

Beyond immunology, the intra-patient comparison with EVA enables evaluation of the role of intra-tumor heterogeneity in metastasis. Specifically, we compared cancer cells from primary tumors to metastases from individual patients in HNSCC single-cell data. This analysis revealed a clear pattern: either uniform dysregulation or no significant differences between the primary tumor and metastasis. For the two patients that had differential variability, the heterogeneity within the primary tumor was significantly higher than the metastatic cancer cells (Fig. 2-5A). This observation agrees with Nowell's theory of clonal evolution, which states that cancer originates from a single cell, accumulates genetic alterations, and during the process of metastasis there is an enrichment for the most aggressive clones[189]. This theory would indicate that clonal metastases are more homogeneous, as very few cells gain invasive and metastatic potential. Such intra-tumor discrepancies that may evolve as the disease progresses between the primary tumor and disseminated metastasis can result in incorrect biomarkers being used to make clinical decisions and lead to therapeutic failure[169]. The differences in molecular heterogeneity may also give rise to different therapeutic responses in primary tumors than metastases. We note that

the analyses performed in this study used current landmark cohorts of breast and head and neck tumors, which were limited in sample size. Future work with EVA analysis on larger sample cohorts is essential to establish the role of heterogeneity in complex dynamic processes such as cancer progression and therapeutic response.

Together, the results of these analyses show that EVA is a robust algorithm for detecting inter- and intra-tumor heterogeneity in scRNA-seq data. EVA is applicable to imputed scRNA-seq datasets, which we demonstrate using MAGIC and BISCUIT imputed data. In the applications to some of the cancer datasets in this study, such as tumor versus normal and primary tumor versus metastasis, we observe widespread changes across a majority of pathways between phenotypes. We attribute these changes to the pervasive transcriptional reprogramming in cancer. While the pathways examined in this study are in no way exhaustive, this is suggestive of global disruption of gene expression and makes for broad interpretations. Comparisons within immune cells and primary cancer subtypes show phenotype specific patterns of dysregulation, allowing more specific interpretation of the molecular mechanisms in tumor heterogeneity. We note that EVA can be widely applied beyond cancer, for example to evaluate the role of transcriptional variation on cell fate specification in development[206]. In this context, heterogeneity is more constrained than in cancer and EVA finds different patterns of inter-cellular heterogeneity for distinct pathways, with some pathways increasing over developmental time and others decreasing. Because EVA statistics compare cells from two phenotypes, this time course analysis was performed by applying EVA to compare cells from pairs of consecutive developmental time points. Extensions to EVA to quantify dysregulation relative to continuous phenotypes[207] or adaptation of alternative kernel based methods to scRNA-seq data[208] will be essential to evaluate as extensions to more complex statistical comparisons in future work.

In addition, EVA is broadly applicable to any dissimilarity metric and is not

limited to Kendall-tau (Supplemental Fig. 1). This flexibility in the algorithm allows users to specify appropriate distance metrics for datasets and enables the direct comparison of performance across various metrics. We have demonstrated that different dissimilarity metrics have different sensitivities to missing data and note that these metrics may also have different sensitivities to the preprocessing used for the scRNA-seq datasets. The rank-based Kendall-tau dissimilarity metric used for the majority of this study is independent of many sample-specific normalization procedures, such as log transformation or quantile normalization. Other dissimilarity measures may be sensitive to these transformations, and this effect must be evaluated before applying EVA to compare dissimilarity based upon these metrics. Emerging variance stabilization methods to account for the pervasive heteroscedastic mean variance relationship of scRNA-seq data may impact the results obtained with this algorithm and are essential to evaluate in future studies. Thus, EVA is a robust multivariate statistical method to quantify differential variation of pathway gene expression and provides the ability to explore transcriptional variation in numerous disease and normal contexts at a single-cell resolution. Future work to improve the EVA algorithm will involve integrating mathematical models to compute comparisons on scRNA-seq data without the need for imputation.

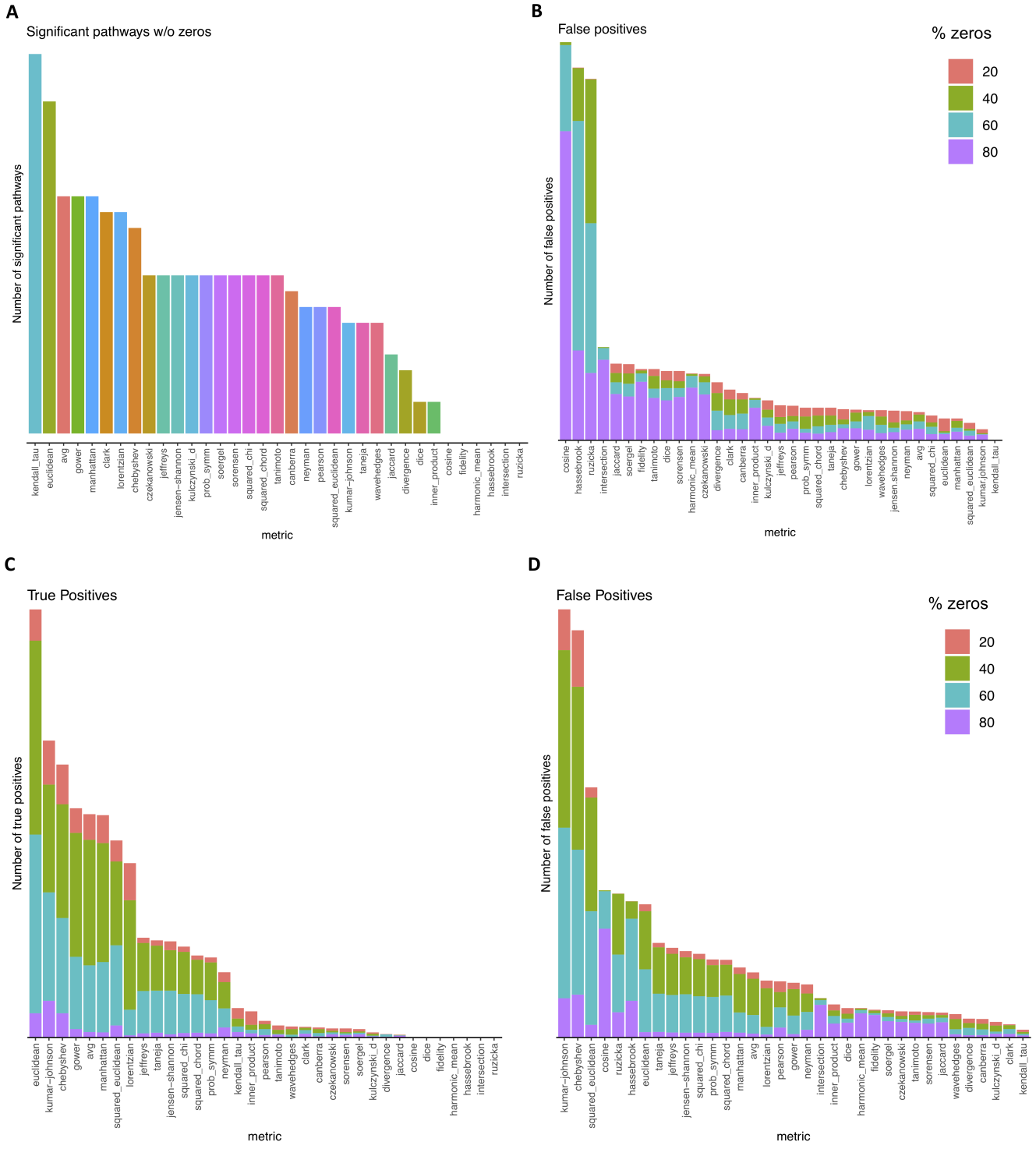

Figure S2-1. Sensitivity of dissimilarity metrics to variable sparsity. (A) Barplot of the number of significant hallmark gene set pathways across 35 metrics from performing EVA on a benchmark bulk transcriptional dataset with normal and cancer samples.

Figure S2-1. (B) Number of significant hallmark gene set pathways across 35 metrics from performing 100 EVA permutations on a mirrored bulk transcriptional data set so that there is no signal. Each bar is colored by the percentage of random zeros added to the dataset. As the number of random zeros increases, the number of significant pathways (all false positives) tends to increase. **(C)** Number of true significant hallmark gene set pathways across 35 metrics from performing 100 EVA permutations on the bulk transcriptional data from (A). True positives are defined as any pathway that was significant for a metric when no zeros are present. Each bar is colored by the percentage of random zeros added to the dataset. For most metrics, as the number of random zeros increases, the number of significant pathways increases around 40% and 60% zeros, and the signal drops at 80%. **(D)** Number of false positive significant hallmark gene set pathways across 35 metrics from performing 100 EVA permutations on the bulk transcriptional data from (A). False positives are defined as any pathway that was not significant for a metric when no zeros are present. Each bar is colored by the percentage of random zeros added to the dataset. For most metrics, as the number of random zeros increases, the number of significant pathways peaks around 40% and 60% zeros, with some metrics peaking at 80%. This indicates variable sensitivity to zeros in a dataset with signal.

Chapter 3

Transfer learning between preclinical models and human tumors identifies a conserved NK cell activation signature in anti-CTLA-4 responsive tumors

Abstract

Background

Tumor response to therapy is affected by both the cell types and the cell states present in the tumor microenvironment. This is true for many cancer treatments, including immune checkpoint inhibitors (ICIs). While it is well-established that ICIs promote T cell activation, their broader impact on other intratumoral immune cells is unclear; this information is needed to identify new mechanisms of action and improve ICI efficacy. Many preclinical studies have begun using single-cell analysis to delineate therapeutic responses in individual immune cell types within tumors. One major limitation to this approach is that therapeutic mechanisms identified in preclinical models have failed to fully translate to human disease, restraining efforts to improve ICI efficacy in translational research.

Method

We previously developed a computational transfer learning approach called projectR to identify shared biology between independent high-throughput single-cell RNA-sequencing (scRNA-seq) datasets. In the present study, we test this algorithm's ability to identify conserved and clinically relevant transcriptional changes in complex tumor scRNA-seq data and expand its application to the comparison of scRNA-seq datasets with additional data types such as bulk RNA-seq and mass cytometry.

Results

We found a conserved signature of NK cell activation in anti-CTLA-4 responsive mouse and human tumors. In human metastatic melanoma, we found that the NK cell activation signature associates with longer overall survival and is predictive of anti-CTLA-4 (ipilimumab) response. Additional molecular approaches to confirm the computational findings demonstrated that human NK cells express CTLA-4 and bind anti-CTLA-4 antibodies independent of the antibody binding receptor (FcR) and that similar to T cells, CTLA-4 expression by NK cells is modified by cytokine-mediated and target cell-mediated NK cell activation.

Conclusions

These data demonstrate a novel application of our transfer learning approach, which was able to identify cell state transitions conserved in preclinical models and human tumors. This approach can be adapted to explore many questions in cancer therapeutics, enhance translational research, and enable better understanding and treatment of disease.

Background

Single-cell RNA-sequencing (scRNA-seq) data provide an unprecedented opportunity to unravel the cellular complexity and diversity of immune cell populations in the tumor microenvironment[209]. When used in the context of immunotherapy, scRNA-seq data of tumors can provide a more comprehensive understanding of the molecular and cellular pathways that drive therapeutic response and resistance. While studies often use preclinical mouse models as a convenient and useful tool for studying therapeutic response mechanisms, they are limited in their ability to infer biology relevant to therapeutic responses in humans. To improve the clinical efficacy of immunotherapies such as immune checkpoint inhibitors (ICIs), we need a deeper understanding of the fundamental mechanisms that underlie the anti-tumor activity of ICIs in humans.

Many aspects of the immune system are conserved between mice and humans, but there are significant species-specific differences[210]. These differences may contribute to the frequent failure of therapies that are effective in mouse models from showing similar efficacy in humans[211]. Discrepancies between ICI mechanisms observed in mice and humans may be further complicated by species-specific differences that mask detection of conserved alterations in responding immune cells. A deeper understanding of human and mouse immune responses to immunotherapy could generate new insights into properties that define therapeutic sensitivity. Emerging scRNA-seq studies that have begun to characterize changes in gene expression after ICI treatment[21, 54, 212] are ideally suited to begin learning these mechanisms. However, computational tools that identify conserved cell state transitions across species are needed to compensate for species-specific immune system differences in transcriptional data. As scRNA-seq becomes increasingly popular in immuno-oncology, such tools will be essential to validate preclinical findings in terms of both robustness and clinical relevance.

To enable cross-species data integration, we previously developed a computational

framework that uses matrix factorization (CoGAPS) and transfer learning (projectR) to integrate transcriptional datasets from different species[112]. This approach has led to the identification of both species-specific and conserved biological processes in the developing retina of mice and humans[213, 214], but it has not yet been applied to cancer therapeutics. To determine if transfer learning can identify conserved and clinically relevant transcriptional alterations within the tumor microenvironment induced by therapy, we applied it to learned cellular patterns from scRNA-seq data of intratumoral immune cells in ICI-treated preclinical models and human patients.

We focused our investigation on the impact of anti-CTLA-4 antibodies because of the numerous cellular mechanisms of action of anti-CTLA-4 antibodies (ipilimumab) found to underlie its efficacy[215, 216]. By blocking the inhibitory T cell receptor CTLA-4, anti-CTLA-4 antibodies enhance T cell effector activity, causing tumor regression[217, 218]. Studies in mice suggest that anti-CTLA-4 efficacy is also dependent on the depletion of CTLA-4 expressing regulatory T cells[219, 220]. However, Sharma et al.[221] found that anti-CTLA-4 treatment does not deplete Tregs in several human cancer types, suggesting there may be a discrepancy in anti-CTLA-4 response between mouse and human tumors. Therefore, attempts to understand the mechanism of action of anti-CTLA-4 antibodies could be improved by computational approaches that can identify biology shared by mice and humans and point to additional cell types beyond T cells that may mediate anti-CTLA-4 therapeutic efficacy.

Altogether, this study provides an application of transfer learning to enable preclinical to clinical evaluation of cellular pathways associated with anti-CTLA-4 treatment. Using scRNA-seq data from Gubin et al.[21], we show that CoGAPS is able to detect robust transcriptional signatures associated with anti-CTLA-4 treatment (Fig. 3-1). The signature most associated with anti-CTLA-4-treated tumors reflected NK cell activation. We use projectR to confirm the association of this signature with positive clinical outcomes in datasets generated from distinct modalities that

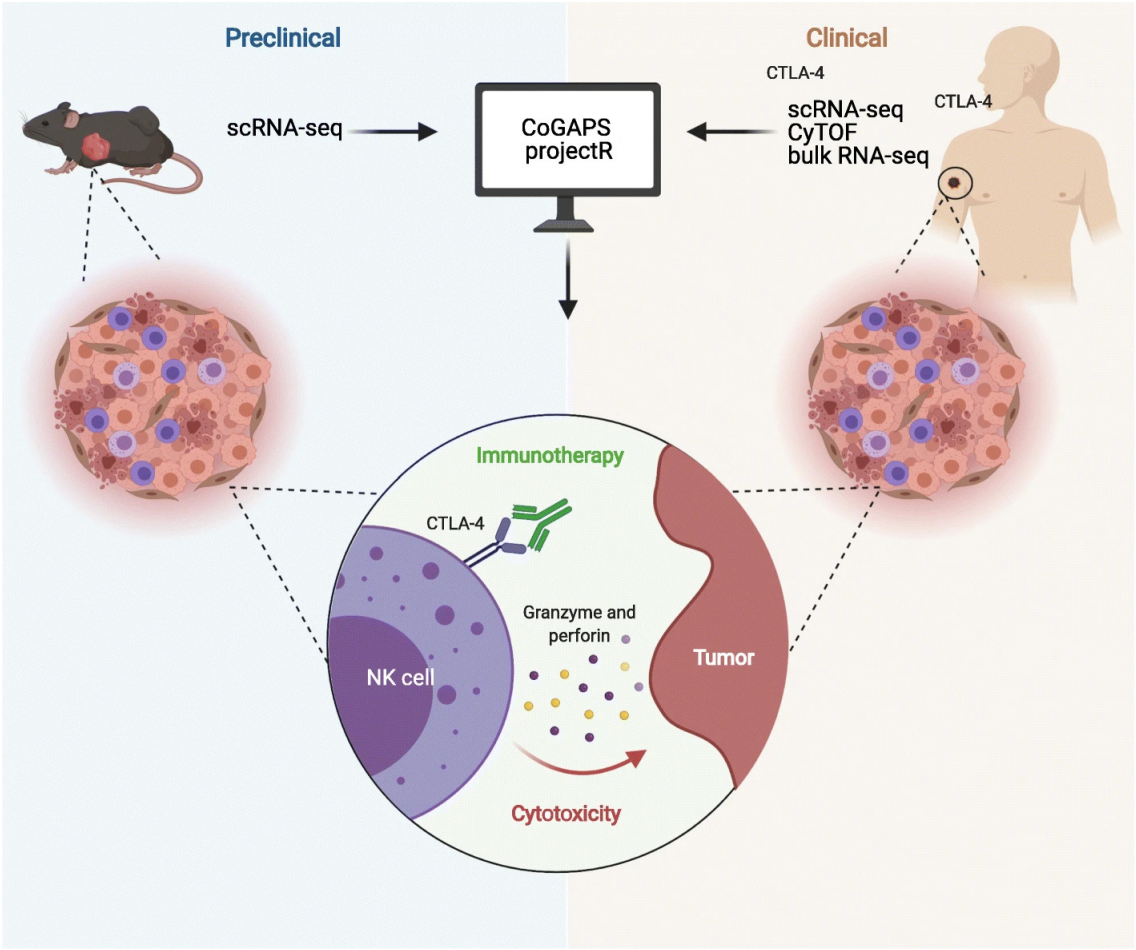

Figure 3-1. Graphical summary. Visual summary of the computational workflow, data types (scRNA-seq, CyTOF, or bulk RNA-seq), and sources (preclinical or clinical) used to identify conserved responses to immunotherapy. In response to anti-CTLA-4 therapy, we detect natural killer cell activation in mice and human tumors and demonstrate that human natural killer cells express CTLA-4 and bind anti-CTLA-4 at the cell surface.

include bulk RNA-seq, mass cytometry, and scRNA-seq. This analysis identifies NK cell activation in anti-CTLA-4-treated human tumors that had not been described previously. We confirm our computational findings with complementary molecular techniques to begin to elucidate how NK cells activate in response to anti-CTLA-4 treatment. These analyses yield novel insights into the role of NK cells in anti-CTLA-4 efficacy and represent a general strategy for the study of shared tumor biology across datasets derived from different tumor types, treatment groups, sequencing platforms, and species.

Methods

Data collection

In this study, we used three public scRNA-seq datasets that were downloaded from NCBI's Gene Expression Omnibus (GEO). For CoGAPS analysis on preclinical immunotherapy samples, we used the dataset from Gubin et al. (accession number GSE119352)[21]. This dataset contains 15,000 flow-sorted CD45+ intratumoral cells from mouse sarcomas that were collected during treatment with either control monoclonal antibody, anti-CTLA-4, anti-PD-1, or combination anti-CTLA-4 and anti-PD-1 acquired with the 10x Genomics Chromium platform, using v1 chemistry. Associations between CoGAPS signatures and immunotherapy treatment were confirmed by transfer learning using paired mass cytometry data from Gubin et al.[21], which was downloaded from the FLOW Repository (FR-FCM-ZYPM) and processed using the R package cytofkit version 0.99.0.

For transfer learning to human samples, we used two scRNA-seq datasets of intratumoral immune cells from metastatic melanoma patients. To first test the relationship between our preclinical CoGAPS patterns and clinical outcome, we used the dataset from Sade-Feldman et al. (accession number GSE120575)[54]. This dataset contains 16,000 flow-sorted CD45+ intratumoral cells obtained from 48 human melanoma tumor biopsies from 32 patients at baseline or after treatment with either anti-CTLA-4, anti-PD-1, or combination anti-CTLA-4 and anti-PD-1. This data was acquired with Smart-seq2[222].

Next, to confirm the observed relationship between our preclinical NK activation signature and response to anti-CTLA-4, we used the scRNA-seq dataset from de Andrade et al. (accession number GSE139249)[212]. This dataset contains 40,000 flow-sorted NK cells from matched blood and tumor samples obtained from 5 patients with melanoma metastases. Two patients had an initial response to treatment with

anti-CTLA-4 or anti-PD-1 with oncolytic virus. Two patients failed to respond to combination anti-CTLA-4 and anti-PD-1 or anti-PD-1. One patient was not treated with immunotherapy. This data was acquired with the 10x Genomics Chromium platform, using v2 chemistry.

In addition, bulk RNA-seq was downloaded from The Cancer Genome Atlas[223]. Normalized gene expression for 33 tumor types were obtained using the R/Bioconductor package TCGAbiolinks version 2.14.1[224]. CIBERSORT scores for this data were obtained from Thorsson et al.[225].

These datasets were used for pattern discovery and transfer learning as described below.

Dimensionality reduction and cell type identification

Cell type inference analyses were performed for the Gubin et al.[21] dataset with the standard Monocle3 workflow using package version 0.2.0[89, 92, 95]. Dimensionality reduction and visualization for scRNA-seq data were performed using Uniform Manifold Approximation and Projection (UMAP)[179]. Briefly, the first 15 principal components were used as input into the reduce_dimension function. Canonical cell type marker genes as described in Gubin et al.[21] were used to annotate cells.

Mouse pattern discovery and gene set analysis using CoGAPS

CoGAPS analysis was performed using the R/Bioconductor package CoGAPS version 3.5.8[226] to analyze the mouse sarcoma dataset from Gubin et al.[21]. Genes with a standard deviation of zero were removed prior to analysis. The input for CoGAPS is a data matrix of single-cell data with genomic features by cells, a number of sets, and number of patterns to learn (nPatterns) on each of the sets of cells. Because single-cell data is large, CoGAPS is performed for random subsets of cells in the complete scRNA-seq data as determined by the number of sets used as an input parameter to

the software. CoGAPS factorizes the input matrix into two related matrices containing the gene weights (the amplitude (A) matrix) and sample weights (the pattern (P) matrix) for each data subset, and then identifies a set of consensus patterns across the data subsets and re-learns the amplitude (A) matrix on the entire dataset. Because consensus patterns are learned across multiple sets, the final number of patterns may not match the input parameter of nPatterns. The log2 transformed count matrix of remaining genes across all samples was used as input to the CoGAPS function. Default parameters were used, except nIterations = 50,000, sparseOptimization = True, and nSets = 12. The input parameters for nPatterns were determined empirically, by testing over a range of dimensions. When the nPatterns input was set to 3, we obtained results that identified immune cell lineage. We reasoned that additional patterns could further identify biological processes in the data related to treatment. We initially tested 50 patterns; however, many of the patterns highlighted few cells, indicating an over-dimensionalization of the data. When nPatterns was set to 25, CoGAPS identified 21 consensus patterns, which separated immune cell types and cell states.

Genes highly associated with each pattern were identified by calculating the PatternMarker statistic[227], which takes the gene weights assigned by CoGAPS and returns those most associated with a particular pattern or set of patterns. The CalcCoGAPSStat function was used to identify pathways significantly enriched in each pattern for the MSigDB hallmark gene sets[197] and PanCancer Immune Profiling panel from NanoString Technologies. This function links each CoGAPS pattern to the activity of input gene sets using a z-score based statistic[228]. p-values obtained from pathway analysis were FDR adjusted with the Benjamini-Hochberg correction and FDR adjusted p-values below 0.05 were called statistically significant.

Pseudotime analysis

To perform pseudotemporal ordering, the dataset was subset to relevant cell types and treatments based on the desired analysis. Due to the association between pattern 7 and activation state markers, we chose the most active terminus of the trajectory as the end state. Thus, the root node of the trajectory was assigned by identifying the region in the UMAP-dimensional reduction with low CoGAPS pattern 7 weights. Pseudotime values were assigned to cells using the order_cells function from the R package Monocle3 version 0.2.0[89, 92, 95]. Genes with significant expression changes as a function of pseudotime were identified using the graph_test function, using a multiple-testing corrected q-value cutoff of 0.01.

Construction of multivariate Cox proportional hazards models

TCGA normalized gene expression for 33 tumor types was used as input for transfer learning to relate CoGAPS immune signatures to clinical outcomes. Metadata from Liu et al.[229] was used for measures of overall survival and age at diagnosis for TCGA samples. Samples were restricted to those that were labeled as "Primary solid tumor" (n = 9113), and "Metastatic" (n = 394) in the "definition" column of the TCGA metadata, which resulted in 9507 total samples. Association between CoGAPS pattern weights and overall survival was analyzed using multivariate Cox proportional hazards regression models using the survival coxph function from the R package survival version 3.2-11 and $p < 0.05$ was used as threshold for significance.

Correlation analysis

To compare the expression of CTLA-4 and CIBERSORT scores for various immune cell types across immunogenic solid tumors from TCGA, we calculated the Spearman correlation coefficients using the cor.test function in R.

Transfer learning

To examine whether the mouse patterns corresponded to similar immunotherapy responses in human data, we used The R/Bioconductor package projectR version 1.0.0[230] to project the expression matrix from several datasets into the CoGAPS pattern matrix[112]. The CoGAPS result object and the expression matrix from a human dataset is used as input to the projectR function. Homologous genes present in the mouse and human data were retained for projection. Genes without homologs in the human data were removed. ProjectR returns a new pattern matrix, which estimates the role of each pattern in each cell of the human dataset. This comparison of pattern across species usage enabled us to determine how each pattern defines features present in the human dataset (i.e., cell types and immune cell activation).

Pattern performance of predicting anti-CTLA-4 response

The projected pattern weight is a continuous range of values, instead of a binary outcome. Using the individual projected pattern weight for each cell and a binary response outcome to anti-CTLA-4, we performed ROC curve analysis using the ROCR package, version 1.0-7 to determine the true-positive rates versus false-positive rates of pattern 7 weights to classify response. The area under the ROC curve was used as the quality metric to determine the prediction performance.

Cell lines and materials

All human NK cell lines (NK-92, NK-92-CD16v, NKL, YT and KHYG-1) were kindly provided by Dr. Kerry S. Campbell (Fox Chase Cancer Center, Philadelphia, PA). The NK-92-CD16v expressed GFP due to transduction with pBMN-IRES-EGFP containing the FcγRIIIA construct. All NK cell lines were cultured as previously described[231]. Fresh healthy donor NK cells were purchased from AllCells (PB012-P). These NK cells were positively selected from donor peripheral blood using CD56

positivity. Donor NK cell purity was 98–99%. Donor 3 and donor 4 were expanded using engineered antigen presenting cells (K562-4-1BB-mbIL-21) according to the protocol[232]. CTLA-4 overexpressing Jurkat cell line was generated using lentiviral transduction purchased from G&P BIosciences (Product ID LYV-CTLA4, SKU# LTV0710) which contained full length human CTLA-4 gene subcloned into lentiviral expression vector pLTC with an upstream CMV promoter with puromycin selection marker. Jurkat cells were transduced using millipore sigma's spinoculation protocol. In brief, lentiviral particle solution was added to 2 $\times$ 106 Jurkat cells at a final multiplicity of infection of 1, 5, and 10. Cells were centrifuged at 800$\times$g for 30 min at 32 °C then resuspended in complete growth medium for 3 days. After 3 days, cells were resuspended in complete medium containing 5 μg/mL puromycin overnight for selection. Selection was performed twice.

qRT-PCR

RNA was isolated using the PureLink RNA Mini Kit (Ambion). The RNA concentration was measured using NanoDrop 8000 (Thermo Fisher Scientific). cDNA was generated from 20 to 100 ng of RNA using the GoTaq 2-step RT-qPCR System (Promega). qPCR was performed with SYBR Green on a StepOnePlus real-time PCR system (Applied Biosystems). Gene expression was normalized to HPRT and analyzed using 1/ΔCt method with triplicates.

Primers used were the following:

CTLA-4: (F: CATGATGGGGAATGAGTTGACC; R: TCAGTCCTTGGATAGTGAGGTTC)

CD28: (F: CTATTTCCCGGACCTTCTAAGCC; R: GCGGGGAGTCATGTTCATGTA)

CD28H: (F: CCCTGCAAGAAGCCTCAAG; R: CCTTTGTCCACTTAACACGGAG)

HPRT: (F: GATTAGCGATGATGAACCAGGTT; R: CCTCCCATCTCCTTCATGACA)

Western blot

Cells were lysed in boiling buffer with EDTA (Boston BioProducts) supplemented with 1X protease and 1% phosphatase inhibitor prepared following the manufacturer's protocols (Sigma-Aldrich, Cat. No. 11697498001 and P5726). Cleared lysate concentrations were obtained by a DC Protein Assay (BioRad). Lysates 30–50 μg were run on SDS-PAGE gels and transferred to nitrocellulose membranes (GE Healthcare). Western blots were conducted using anti-CTLA-4/CD152 (LS-C193047, LSbio) at concentrations of 1:1000 diluted in 5% milk in PBST. Secondary antibody was anti-rabbit IgG, HRP linked (Cell Signaling) used at 1:1000. Chemiluminescent substrate (Pierce) was used for visualization.

Flow cytometry

All cells were aliquoted into Eppendorf tubes, spun at 5000 rpm for 1 min at 4 °C, washed twice with HBSS (Fisher Scientific Cat. No. SH3058801), and resuspended in 50 μL of FACS buffer (PBS plus 1% BSA) and blocked with 1 μL human Fc block (BD Biosciences, 564219) for 20 min at 4 °C. Labeled antibodies were then added at the manufacturer's recommended concentrations and incubated at 4 °C for 30 min, with vortexing at 15 min. Cells were then washed with FACS buffer twice and resuspended in FACS buffer or fixative (1% PFA in PBS). Flow antibodies included anti-human CD152 (CTLA-4) (BD Bioscience 555853), CD28 (Biolegend 302907), and CD28H (R&D Systems, cat#MAB83162). The CD152 antibody has previously been shown to adequately detect CTLA-4 expression on both human T and B cells[229]. Samples were run in the Georgetown Lombardi Comprehensive Cancer Center Flow Cytometry & Cell Sorting Shared Resource using BD LSRFortessa. Analyses were performed

using FlowJo (v10.4.1).

Immunofluorescence

Ipilimumab was acquired from the Medstar Georgetown University Hospital. Ipilimumab was labeled with Dylight550 fluorophore using the Dylight550 Conjugation Kit (Fast)-Lightning-Link (abcam, ab201800). In short, ipilimumab was diluted from 5 to 2 mg/mL using sterile PBS. Human IgG (Jackson ImmunoResearch, 009-000-003) was diluted from 11 to 2 mg/mL using sterile PBS. One microliter of modifying reagent was added to 10 μL diluted ipilimumab and 10 μL diluted human IgG. Ten μL antibody was then added to the conjugation mix and incubated at room temperature in the dark for approximately 6 h. One microliter of quencher reagent was added to the labeled ipilimumab and the antibody was stored in the dark at 4 °C. NK-92 and PANC-1 cells were collected and washed with cold PBS and brought to a final concentration of 1 $\times$ 106 cells/mL in staining buffer (1% BSA in PBS) in 50 μL. Fifty microliters of labeled ipilimumab or human IgG was added to cells to yield a final concentration of 1 μg/mL antibody. Cells were incubated in the dark at 4 °C for 1 h. After incubation, cells were pelleted and washed three times with cold PBS. Cells were brought to a final concentration of 0.5 $\times$ 106 cells/mL and 100 μL was immobilized on slides using cytospin (Cytospin 2, Shandon) for 5 min at 1000 rpm. Following immobilization cells were fixed with 4% PFA for 10 min at room temperature then washed three times with cold PBS. Coverslips were mounted using VectraSheild mounting media with DAPI and sealed using clear nailpolish and allowed to dry overnight in the dark. Analyses were performed with the Leica SP8 AOBS laser scanning confocal microscope.

Cell surface biotinylation

Cell surface biotinylation of NK92, NKL, YT, and KHYG-1 cells was performed with the Pierce Cell Surface Protein Isolation kit (Thermo Scientific, cat#89881) according

to the manufacturer's protocol. In brief, 4 × 108 cells were pelleted and washed with cold PBS then incubated with EZ-LINK Sulfo-NHS-SS-biotin for 30 min at 4 °C followed by the addition of a quenching solution. Another 1 × 106 cells were collected and saved for total cell western blotting. Cells were lysed with lysis buffer (500 μL) containing the cOmplete protease inhibitor cocktail (Roche, cat# 11697498001). The biotinylated surface proteins were excluded with NeutrAvidin agarose gel (Pierce, 39001). Samples were diluted 50 μg in ultrapure water supplemented with 50 mM DTT. Lysates were subjected to Western blotting with the anti-CTLA-4 antibody described above.

NK cell stimulation

Cell lines or expanded primary NK cells were stimulated with 100 U/mL IL-2 (NCI preclinical repository), 5 ng/mL IL-12 (R&D Systems, cat#219-IL-005), 10 ng/mL IL-15 (NCI preclinical repository), 50 ng/mL IL-18 (Invitrogen, cat#rcyec-hil18), or 500 U/mL IFNg (Sigma-Aldrich, cat# I3265) for 24 h. Cell pellets were collected and processed for rt-qPCR as described above. Cell lines or expanded primary NK cells were stimulated with 3 μg/mL CD28 activating antibody (Biolegend, cat#302933) for 24 h.

Results

CoGAPS identifies known molecular alterations in response to immunotherapy from scRNA-seq data

Whereas human tumors have limited access for high-dimensional profiling, mouse models can be readily used to generate scRNA-seq data to study the tumor immune microenvironment under a variety of treatment conditions. Analysis of these data is then critical to determine biological processes associated with treatment perturbations, with unsupervised learning providing an opportunity for de novo discovery of cell

state transitions related to therapy. To detect latent spaces (also called "patterns") that represent transcriptional signatures across intratumoral immune cells during immunotherapy response, we used our non-negative matrix factorization (NMF) technique, CoGAPS (Fig. 3-2A)[226]. CoGAPS is an established approach to dissect transcriptional signatures that dictate cell type identity (i.e., NK vs. Treg) and cell state (i.e., activated vs. resting), aiding the evaluation of complex molecular alterations within the tumor immune microenvironment[132, 233]. By combining CoGAPS with projectR, a transfer learning approach, we can then quickly query for shared features across independent datasets across species (Fig. 3-2A)[112, 226].

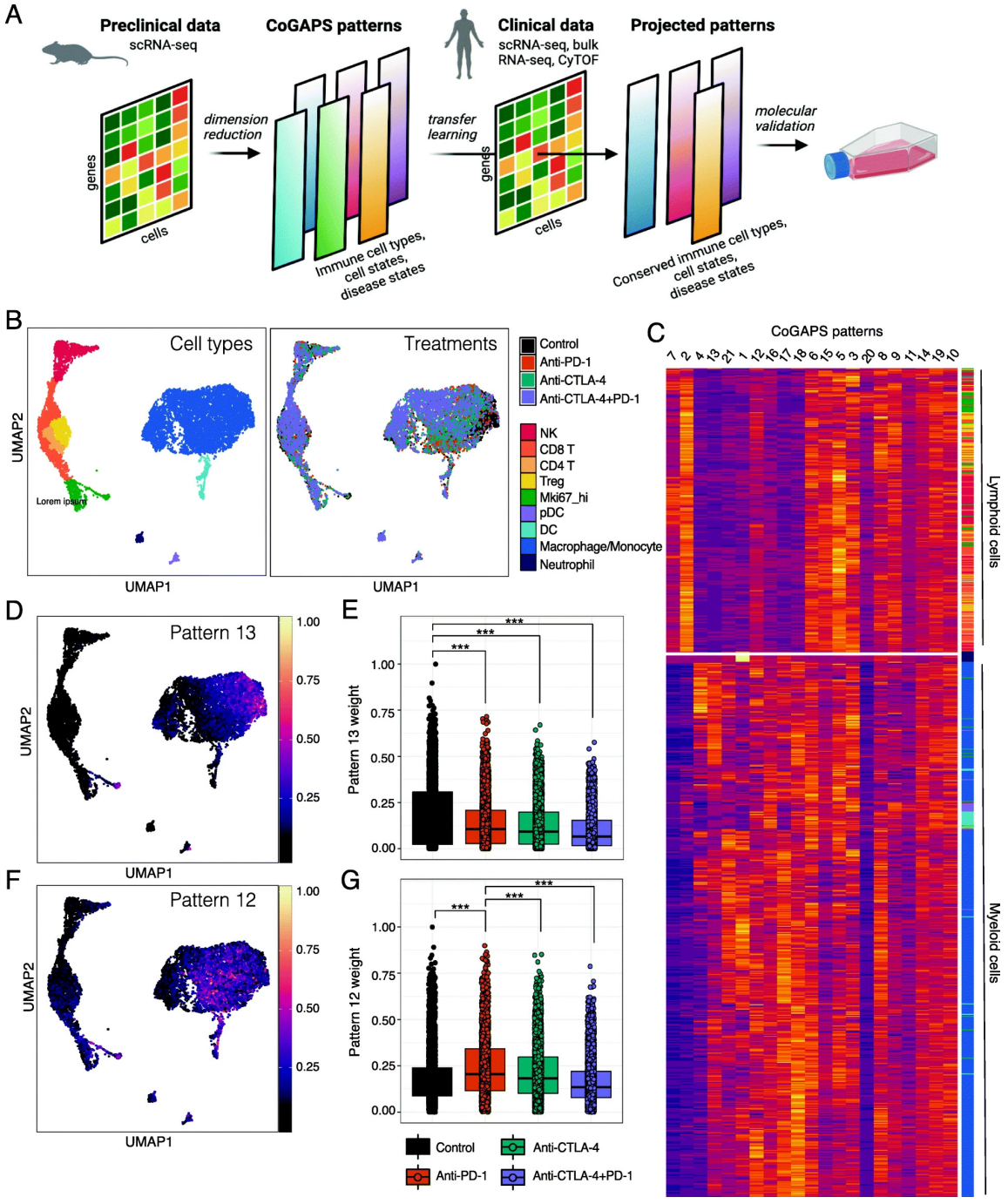

Figure 3-2. CoGAPS identifies gene signatures related to immune cell lineage and treatment response in mouse intratumoral immune cell scRNA-seq data. (A) Overview of the pipeline to relate preclinical and clinical mechanisms of action of therapy using transfer learning. First, CoGAPS, a non-negative matrix factorization algorithm is applied to scRNA-seq data of ICI-treated mouse tumors.

Figure 3-2. Matrix factorization algorithms are unsupervised learning methods that can distinguish low-dimensional gene and cell features (latent spaces) associated with therapeutic responses without prior knowledge of gene regulation or cell type classification. Next, the transfer learning method projectR, is used to project the transcriptional signatures representing the latent spaces (or patterns) identified by CoGAPS into an independent dataset of human tumors treated. Finally, the cell weights representing relative usage of each pattern in the new human dataset can be computationally assessed for relationships to clinical outcomes and as the basis to prioritize candidates for experimental validation. **(B)** UMAP-dimension reduction of droplet-based scRNA-seq of intratumoral immune cells from ICI-treated mouse sarcomas[21]. Samples are colored by annotated cell types (left) and by treatment (right). **(C)** Hierarchical clustered heatmap of 21 CoGAPS patterns demonstrating segregation by immune cell lineage. Rows are individual cells, with row annotations designating cell type. Columns represent different CoGAPS patterns. **(D)** UMAP-dimension reduction colored by CoGAPS pattern 13 weights illustrates a cell type specific signature within the macrophages/monocytes. **(E)** Boxplot of pattern 13 weights in individual macrophage/monocyte cells, faceted by treatment group. Pattern 13 is associated with cells treated with control monoclonal antibody. Significant differences in mean pattern 7 weight between treatment groups are indicated by asterisks where p-values $< 0.05 =$ *, $< 0.01 =$ **, and $< 0.001 =$ ***. **(F)** UMAP-dimension reduction colored by CoGAPS pattern 12 weights illustrates a cell type specific signature within the macrophages/monocytes. G Boxplot of pattern 12 weights in individual macrophage/monocyte cells, faceted by treatment group. Pattern 12 is associated with cells treated with anti-PD-1. Significant differences in mean pattern 7 weight between treatment groups are indicated by asterisks where p-values $< 0.05 =$ *, $< 0.01 =$ **, and $< 0.001 =$ ***

To demonstrate the applicability of our pattern detection and transfer learning approach for cross-species analysis in the context of immunotherapy response, we first applied CoGAPS to identify transcriptional responses induced by ICIs in mouse tumors from a publicly available scRNA-seq dataset including more than 15,000 immune cells isolated from mouse sarcomas[21]. These tumors were treated with a control monoclonal antibody, anti-PD-1, anti-CTLA-4, or combination anti-PD-1 and anti-CTLA-4 antibodies (Fig. 3-2B). A critical challenge in applying matrix factorization algorithms such as CoGAPS to scRNA-seq analysis is selecting an appropriate dimensionality (i.e., number of patterns) to resolve biological features from the data[234]. Consistent with previous studies, running CoGAPS across multiple-dimensionalities revealed that different levels of biological complexity were captured at different dimensionalities[235]. For example, at low dimensionality (3 patterns), CoGAPS separated immune cells into myeloid and lymphoid lineages (Fig. S3-1A). When dimensionality was increased to 21 patterns, the myeloid versus lymphoid lineage distinction was preserved and additional transcriptional signatures reflecting immune cell type and state were captured (Fig. 3-2C).

To identify specific attributes captured by each pattern, we performed gene set analysis using the gene weights for each pattern as input. We used the hallmark gene sets from the Molecular Signatures Database (MSigB)[197] and the PanCancer Immune Profiling gene panel from Nanostring Technologies to assess the enrichment of gene sets controlling well-defined biological processes. We found that several transcriptional signatures identified by CoGAPS were consistent with ICI-mediated changes previously described in the literature. For example, pattern 13 was enriched in macrophages/-monocytes from progressing tumors treated with control monoclonal antibody (Fig. 3-2D and E). In contrast, pattern 12 was prevalent in macrophages/monocytes from tumors treated with anti-PD-1 (Fig. 3-2F and G). Macrophages are commonly divided into two subsets, pro-inflammatory anti-tumor M1 subtype and anti-inflammatory

pro-tumor M2 subtype[236]. Consistent with this, pattern 13, which was enriched in control-treated tumors, reflected M2 macrophage polarization, which promotes tumor growth and metastasis (FDR adjusted p-value = 0.018). In contrast, pattern 12, which was enriched in anti-PD-1 treated tumors, reflected M1 macrophage polarization and interferon responses (FDR adjusted p-value = 0.046). This finding agrees with a recent study, which showed that anti-PD-1 treatment leads to a functional transition within the macrophage compartment towards an immunostimulatory M1 phenotype[237].

CoGAPS analysis identifies a subset of activated NK cells in mouse tumors treated with anti-CTLA-4

In addition to the known transcriptional changes resulting from ICI treatment shown in Fig. 3-2, CoGAPS also identified a transcriptional signature that reflected a subset of activated NK cells—pattern 7 (Fig. 3-3A and B). While tumors from each treatment group contained NK cells with elevated levels of pattern 7, there was a significant enrichment in NK cells from tumors treated with anti-CTLA-4 (Fig. 3-3C). To isolate the genes associated with this pattern, we used the CoGAPS PatternMarker statistic[227]. Instead of being based upon the CoGAPS gene weights, this statistic computes the unique association of genes with a particular pattern to isolate the specific set of genes associated with an inferred biological process to prioritize genes for validation. PatternMarker analysis identified 3195 genes associated with pattern 7. Gene set enrichment analysis on the CoGAPS result object revealed an upregulation of interferon-gamma and IL2-STAT5 gene sets in pattern 7, which are key pathways that govern cytotoxicity and maturation in NK cells (FDR adjusted p-value = 0.013)[238]. In addition, gene weights for pattern 7 were highest for markers of NK cell type and function (NKG7, KLRK1, NCR1, and GZMB) and negative for markers of T cells (CD3D, CD3G, CD3E, CD4, CD8A, and CD8B1) (Fig. S3-1B).

The CoGAPS analysis suggested that pattern 7 identified NK cells undergoing a

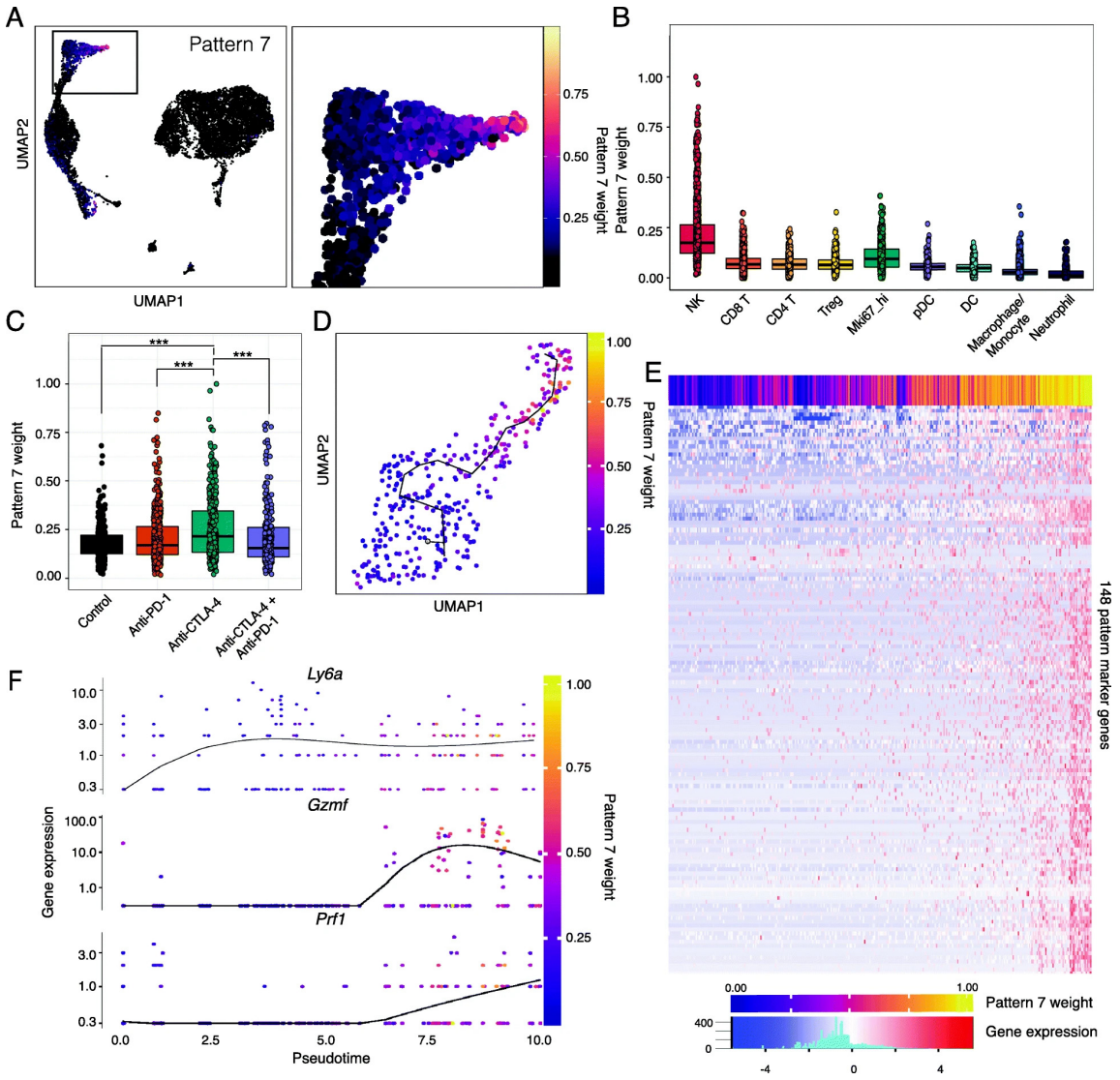

Figure 3-3. CoGAPS and pseudotime analysis reveals a dynamic state change in NK cells during ICI exposure in mouse scRNA-seq data. (A) UMAP-dimension reduction colored by CoGAPS pattern 7 weights across all cells (left) and magnified view (right) showing that pattern 7 marks a population of NK cells delineated in Fig. 3-2A. **(B)** Boxplot of pattern 7 weights across each immune cell type. Cells with high pattern 7 weights are observed only in NK cells. **(C)** Boxplot of pattern 7 weights in individual NK cells faceted by treatment group. Anti-CTLA-4-treated NK cells have increased pattern 7 weights compared to NK cells treated with other immunotherapies. Significant differences in mean pattern 7 weight between treatment groups are indicated by asterisks where p-values $< 0.05 =$ *, $< 0.01 =$ **, and $< 0.001 =$ ***. **(D)** Pseudotemporal trajectory of anti-CTLA-4-treated NK cells colored by CoGAPS pattern 7 weight suggesting that anti-CTLA-4 treatment results in NK cell activation. **(E)** Heatmap of gene expression for 148 pattern markers with variable expression as a function of pseudotime. Columns are individual cells, and column annotation designates pattern 7 weight in each cell. Rows are differentially expressed pattern markers. **(F)** Gene expression of selected NK cell activation genes that are upregulated across pseudotime. Each dot represents a different cell and is colored by CoGAPS pattern 7 weight

cell state change in response to therapy. To further confirm the CoGAPS inference of cell state transitions, we also performed pseudotime analysis on only the NK cells from tumors treated with anti-CTLA-4[95]. While this analysis is not a time course of treatment response, trajectories learned from pseudotime analysis have been shown to enable a quantitative estimation of cellular progression through cell state transitions associated with dynamic biological processes. The pseudotemporal ordering of anti-CTLA-4-treated NK cells showed a sequential progression in cellular trajectory (Fig. 3-3D). This pseudotime trajectory was highly correlated with the pattern 7 weight identified in each cell (0.71 spearman correlation). Notably, the trajectory revealed a single transition state in NK cells as a result of anti-CTLA-4 treatment, with individual cells having transcriptional profiles that reflect various points along the trajectory.

Regression analysis to detect genes significantly associated with changes in pseudotime identified 1968 genes at a q-value threshold of 0.01 in anti-CTLA-4-treated tumors (Table S3). We then looked for genes that were both significantly associated with pseudotime and patternMarkers of the CoGAPS pattern 7 to obtain a subset of 148 genes related to NK cell transitions with anti-CTLA-4 treatment (Fig. 3-3E). This analysis identified 148 genes, including markers of NK cell activation such as perforin, granzymes, and Ly6a[239] (Fig. 3-3F). These data support recent findings that NK cells within mouse tumors can be functionally modulated by ICI treatment[240, 241].

In their original study, Gubin et al.[21] used CyTOF, a mass spectrometry-based flow cytometry method to measure protein expression in parallel with their scRNA-seq. By CyTOF, they found that anti-CTLA-4-induced Granzyme B in a population of KLRG1+ NK cells independently from the scRNA-seq analysis. Still, the relationship between anti-CTLA-4 and NK cell activation in this subpopulation was not evaluated in that study. We hypothesized that immune cells from tumors treated with anti-CTLA-4 in the CyTOF data would have elevated levels of the transcriptional NK cell activation signature we detected in the scRNA-seq data. To test this hypothesis,

we used our transfer learning method, projectR[230], to assess the CyTOF data for the 21 patterns identified by CoGAPS from scRNA-seq. As expected, we found that pattern 7 was highest in immune cells from anti-CTLA-4-treated tumors profiled by CyTOF (Additional file 1: Fig. S3-1C). These findings demonstrate that (1) CoGAPS identified transcriptional changes in response to immunotherapy, which is preserved at the protein and mRNA level and across technological platforms, (2) CoGAPS identified an NK cell activation signature in the scRNA-seq data that was missed by the traditional scRNA-seq analysis methods used in the original study, and (3) ProjectR is capable of identifying gene expression signatures present in both scRNA-seq and CyTOF data.

Preclinical NK cell activation signature is associated with ipilimumab response in metastatic melanoma

To investigate the relevance of the NK cell activation signature (pattern 7) learned in the preclinical mouse model to immunotherapy responses in humans, we used our transfer learning method (projectR), to project two independent scRNA-seq datasets of ICI-treated metastatic melanoma patients[54, 212] into the 21 mouse patterns identified by CoGAPS. We selected melanoma datasets since ICI treatment is widely used in melanoma patients and because previous studies have shown that transcriptional signatures of NK cell infiltration correlate with improved clinical outcomes in melanoma[242]. First, we analyzed a scRNA-seq dataset of 16,000 immune cells isolated from melanoma metastases. Patients in this study were treated with anti-PD-1, anti-CTLA-4, or combination anti-PD-1 and anti-CTLA-4 antibodies, and the biopsies used for scRNA-seq profiling were taken either before or during treatment[54]. Using the projected weights of each signature and treatment outcomes, we evaluated the association of each pattern with therapeutic response in humans. In pre-treatment biopsies, the NK cell activation signature was significantly higher in

anti-CTLA-4 responsive tumors than non-responsive tumors ($p < 1 \times 1015$, Additional file 1: Fig. S3-2A). This is consistent with our initial finding that NK cell activation was enriched in mouse tumors treated with anti-CTLA-4.

Previous scRNA-seq studies that have identified subpopulations of T cells that express transcripts linked to the cytotoxic function of NK cells, such as NKT cells[243, 244]. Consistent with these findings, we observed that cells expressing canonical NK marker genes (NCR1 and FCGR3A) were intermixed with cells expressing T cell marker genes (CD3D) in the lymphocyte cluster (Fig. S3-2B). In addition to showing that pattern 7 is specific for NK cell genes (Fig. S3-1B), to further ensure that T and NKT cells were excluded from analysis and specifically focus on human NK cells, we performed a gene expression gating strategy that required the expression of several transcripts related to NK cell function (NCR1, NKG7, and FCGR3A) and a lack of the T cell transcripts (CD4, CD3D, and CD3G). Gating for NK cells confirmed that the NK cell activation signature was enriched in intratumoral NK cells isolated from anti-CTLA-4 responsive tumors (Fig. 3-4A, $p < 1 \times 108$). Because cells were obtained from tumor biopsies prior to the administration of anti-CTLA-4 treatment, this finding suggests that cytotoxic NK cell infiltration could be predictive of anti-CTLA-4 response. In patients treated with anti-PD-1, there was no significant difference in the NK cell activation signature between responders and non-responders regardless of whether biopsies were taken before (Fig. 3-4A, $p > 0.05$) or during (Fig. 3-4B, $p > 0.05$) treatment. In contrast, the NK cell activation signature was significantly enriched in tumors responsive to combination anti-CTLA-4 and anti-PD-1 taken before (Fig. 3-4A, $p < 0.05$) and during (Fig. 3-4B, $p < 0.01$) treatment. Using receiver operating characteristic curve (ROC) analysis, we found that the NK cell activation signature had a moderate ability to classify anti-CTLA-4 response (Fig. 3-4C, AUC = 0.748), suggesting that the NK activation signature has the potential utility to predict responsiveness to anti-CTLA-4 from pre-treatment tumor biopsies.

These findings indicate that the presence of active NK cells within tumors is important to the clinical usage and success of anti-CTLA-4 therapies.

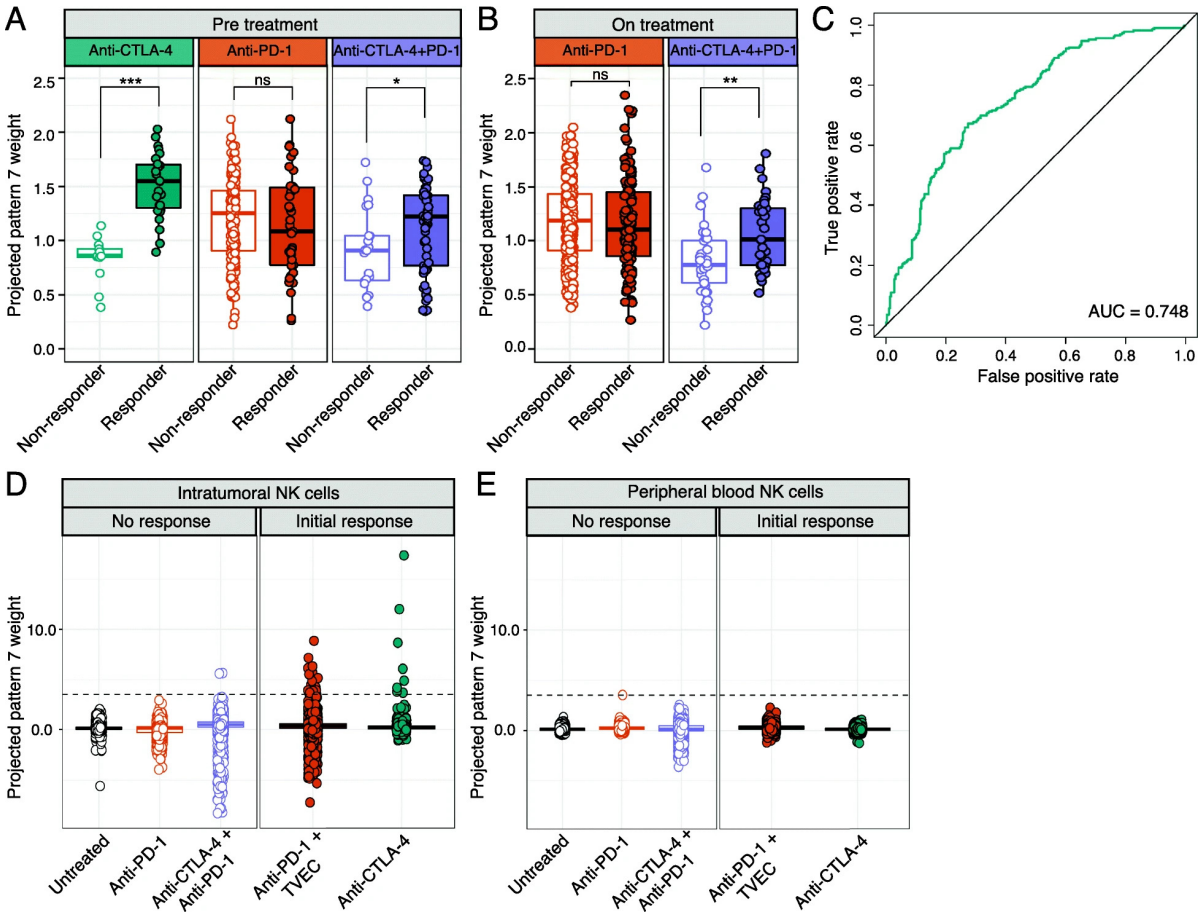

Figure 3-4. ProjectR recovers conserved immunotherapy response in intratumoral NK cells from independent human melanoma scRNA-seq datasets. (A) Box plot of projected pattern 7 weights across intratumoral NK cells from metastatic melanoma patients prior to ICI treatment[54]. Cells are colored by therapy and separated by patient response. Increased pattern 7 is significantly associated with NK cells from patients responsive to anti-CTLA-4 or combined anti-CTLA-4 and anti-PD-1. Significant differences in mean pattern 7 weight between treatment groups are indicated by asterisks where p-values $<$ 0.05 = *, $<$ 0.01 = **, and $<$ 0.001 = ***. **(B)** Box plot of projected pattern 7 weights across intratumoral NK cells from metastatic melanoma patients after treatment with ICI. Cells are colored by therapy and separated by patient response. Increased pattern 7 is associated with NK cells from patients responsive to combination anti-CTLA-4 + anti-PD-1. Significant differences in mean pattern 7 weight between treatment groups are indicated by asterisks where p-values $<$ 0.05 = *, $<$ 0.01 = **, and $<$ 0.001 = ***. **(C)** ROC curve for the performance of pattern 7 weights in predicting response to anti-CTLA-4 prior to the administration of treatment. **(D)** Box plot of projected pattern 7 weights across flow-sorted intratumoral NK cells from metastatic melanoma tumors that were unresponsive ICI (intrinsic resistance) or developed acquired resistance after a period of initial response[212]. The dashed line indicates the average maximum value for pattern 7 across treatment groups. NK cells with elevated pattern 7 weights are seen in patients that had an initial response to ICI, with the highest observed weights from a patient that responded to anti-CTLA-4.

Figure 3-4. (E) Box plot of projected pattern 7 weights across NK cells isolated from peripheral blood of metastatic melanoma patients that had no response to ICI (intrinsic resistance) or developed acquired resistance after a period of initial response. The dashed line indicates the average maximum value for pattern 7 from intratumoral NK cells across treatment groups. Elevated pattern 7 weights are not detected in circulating NK cells, regardless of response.

Although ICI therapy can lead to durable responses in patients with metastatic melanoma, intrinsic and acquired resistance remain major causes of mortalityjenkins2018a. To examine the relationship between NK cell activation and mechanisms of therapeutic resistance, we next projected the transcriptional patterns into a scRNA-seq dataset of NK cells isolated by flow cytometry from matched melanoma metastatic lesions and blood samples of patients that had progressed after immunotherapy[212]. This dataset included two patients that had an initial response to ICI (acquired resistance), two patients that failed to respond to ICI (intrinsic resistance), and one patient that was not given ICI (untreated). We found high levels of the NK cell activation signature in a subset of intratumoral NK cells from the two patients who had an initial response to ICI (Fig. 3-4D). Consistent with our results which indicate that the NK cell activation signature is enriched in anti-CTLA-4 responsive tumors, the highest levels of the NK cell activation signature were found in NK cells from the patient responsive to anti-CTLA-4 (ipilimumab). Elevated NK cell activation signature was also found in the patient responsive to combination treatment with anti-PD-1 and oncolytic virus (pembrolizumab + TVEC). Notably, these observations were specific to intratumoral NK cells, as the NK cell activation signature was detected only at very low levels in NK cells isolated from matched peripheral blood samples (Fig. 3-4E). This result indicates that anti-CTLA-4 treatment leads to NK cell activation specifically within the tumor microenvironment in humans, consistent with observations in mice[241].

Human NK cells express CTLA-4, which is bound by ipilimumab

CTLA-4 is an important regulator of T cells, and there is growing evidence suggesting that CTLA-4 regulates other human immune cell types, including B cells[245, 246], monocytes[247], and dendritic cells[248]. While our computational analysis suggests a

functional role of CTLA-4 in human NK cells, expression of CTLA-4 in human NK cells is controversial in the literature; most studies indicate that human NK cells do not express CTLA-4[241, 249–251]. Our computational association of the intratumoral NK cell activation in response to anti-CTLA-4 treatment suggests that NK cell activity may be modulated directly by CTLA-4 treatment and that CTLA-4 may function as an NK cell immune checkpoint—similar to its role in T cells. To investigate this possibility, we used scRNA-seq data to assess the expression of CTLA-4 transcripts in NK cells and the relationship between CTLA-4 expression and expression of NK cell activation markers. Indeed, we found clusters of intratumoral NK cells from mice and humans that express CTLA-4 and markers of NK cell activation, including GZMB and NKG7 (Fig. 3-5A). Given that CTLA-4 transcripts were detectable in a handful of NK cells, CTLA-4 may be expressed at low to moderate levels and result in poor capture efficiency during scRNA-seq[252]. These technical limitations make the use of in vitro techniques necessary to validate computational findings. Therefore, we turned to molecular biology to further investigate the transcriptional signature of NK cell activation.

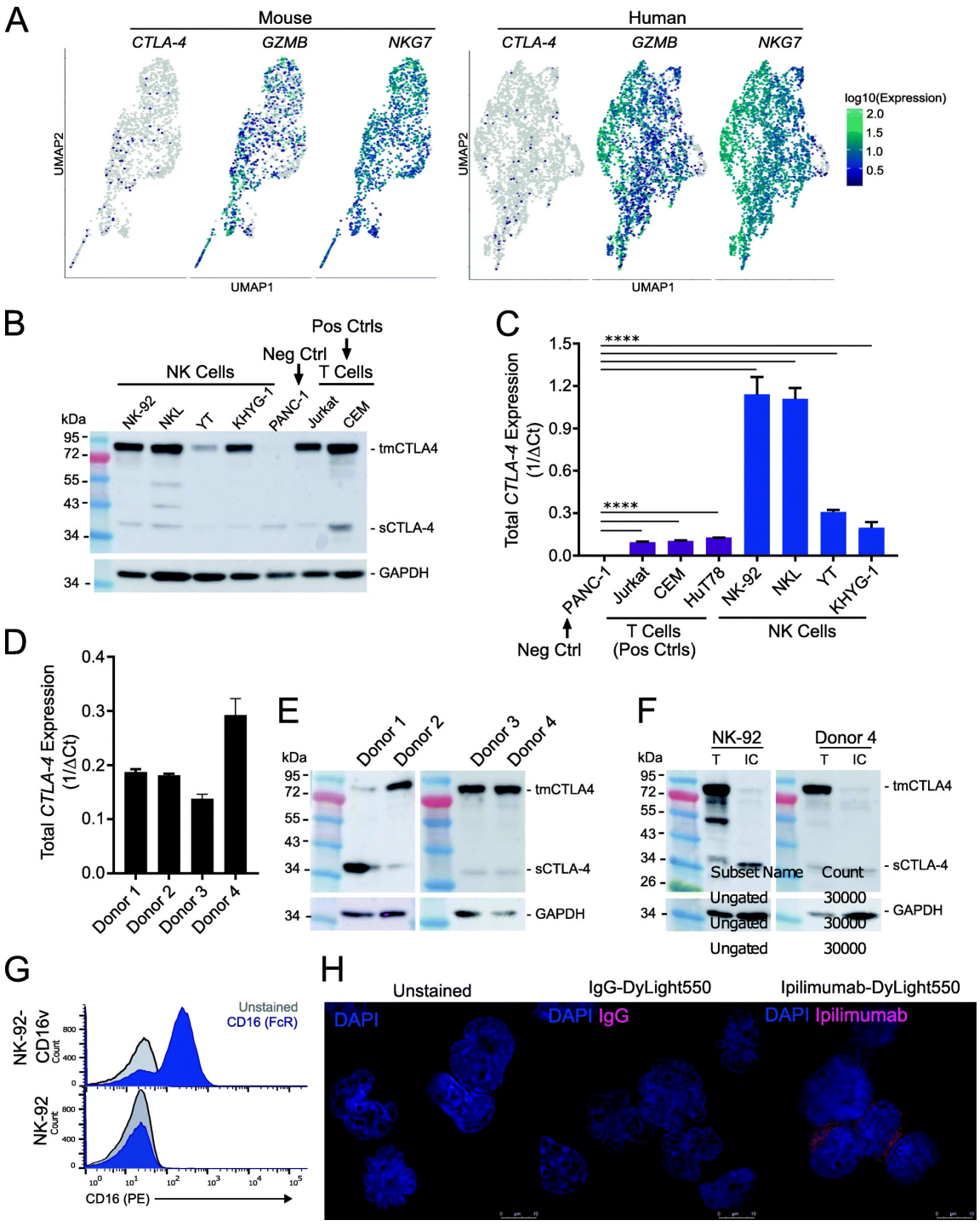

Figure 3-5. CTLA-4 is expressed by both human NK cell lines and healthy human donor-derived NK cells. (A) UMAP-dimension reduction with cells colored by single-cell gene expression for CTLA-4 and representative immune activation genes in mouse (left) and human (right) intratumoral NK cells. The pattern of CTLA-4 expression is consistent with the reduced ability of scRNA-seq to capture low to moderately expressed genes.

Figure 3-5. **(B)** Western blot demonstrating CTLA-4 expression in human NK cell lines. Representative of two independent experiments. **(C)** Quantitative real-time PCR (qRT-PCR) analysis of total CTLA-4 expression (both isoforms) in a CTLA-4 null line (PANC-1), T cell lines (Jurkat, CEM, HuT78), and NK cell lines (NK92, NKL, YT, KHYG-1). p-value $< 0.001 =$ **** as determined by unpaired, two-tailed t-test. **(D)** qRT-PCR demonstrating CTLA-4 expression in CD56+ selected ex vivo unstimulated NK cells derived from healthy human donors. **(E)** Western blot of CTLA-4 expression in CD56+ selected ex vivo unstimulated NK cells derived from healthy human donors. **(F)** Western blot of total protein (T) and intracellular (IC) protein isolated from human NK cell line NK-92 and unstimulated primary human NK cells using cell surface protein biotinylation for exclusion of surface proteins demonstrating surface expression of CTLA-4 dimers and intracellular expression of CTLA-4 monomers. **(G)** Flow cytometry demonstrating NK-92 does not express antibody receptor CD16. Positive control was the NK-92 line that had been transfected with a CD16 expressing plasmid, NK-92-CD16v. **(H)** Immunofluorescent images of NK-92 cells stained with Dylight550-labeled ipilimumab demonstrating that ipilimumab binds to the NK cell surface. Blue staining indicates DAPI. Shown are representative images of a single field of view taken via confocal microscopy (magnification, 63×; zoom, 3×).

To confirm that human NK cells express CTLA-4, we directly tested four human NK cell lines (NK-92, NKL, YT, and KHYG-1) for CTLA-4 expression at the mRNA and protein level and compared to a negative control CTLA4-null cell line (PANC-1) and positive control T cell lines (Jurkat, CEM, HuT78). While all four cell lines appeared negative for CTLA-4 by flow cytometry (Fig. S3-2A), all NK cell lines revealed robust CTLA-4 expression determined by western blot and qRT-PCR (Fig. 3-5B and C). CTLA-4 is known to be expressed on several tumor-derived human cell lines[253, 254]. To exclude the possibility that this observation was specific to malignant NK cells, we assessed CTLA-4 expression in unstimulated ex vivo CD56+ NK cells isolated from healthy human donor PBMCs. Consistent with the results in NK cell lines, CTLA-4 was undetectable by flow cytometry (Fig. S3-2B). However, western blot and rt-qPCR confirmed that NK cells from each donor constitutively expressed CTLA-4 (Fig. 3-5D and E).

Since the western blots of both the positive control T cell lines and NK cells shows two bands—one representing the 95 kDa dimer that is surface expressed and one representing the 30 kDa monomer that is intracellular—we hypothesized that antibody-specific limitations were precluding successful detection of CTLA-4 on the NK cell surface by flow cytometry. We, therefore, turned to an antibody-independent means of detecting surface expression—surface protein biotinylation—to confirm that NK cells express CTLA-4 on the surface. We biotinylated cell surface proteins and then excluded them from the cell lysate via magnetic separation. Using the NK cell line NK92 and healthy donor NK cells, we determined that CTLA-4 dimers and monomers are present in total cell lysate, but the CTLA-4 dimers are absent from the intracellular protein lysate, confirming that NK cells express CTLA-4 dimers on their surface (Fig. 3-5F).

In T cells, CTLA-4 competes with co-stimulatory receptor CD28 for B7 ligands. When CTLA-4 outcompetes CD28 for B7 binding, it prevents CD28 co-stimulatory

signaling and instead provides inhibitory signaling. Anti-CTLA-4 treatment results in T cell activation by inhibiting the inhibitor, by blocking CTLA-4-B7 interactions and promoting CD28-B7 interactions. To determine if CTLA-4 could be functioning similarly in NK cells, we tested NK cells for CD28 and CD28H expression. Consistent with previous reports, we found that some NK cell lines and donor NK cells expressed CD28 and CD28H[255] by flow cytometry and qRT-PCR (Fig. S3-4). Thus, human NK cells express both CTLA-4 and CD28, supporting a similar role for these receptors in T cells and NK cells.

Ipilimumab binds to CTLA-4 expressed on the NK cell surface independent of CD16

We next wanted to determine if the anti-CTLA-4 antibody, ipilimumab, was capable of binding to CTLA-4 expressed on the NK cell surface. To do so, we fluorescently labeled anti-CTLA-4 (Ipilimumab) to probe for ipilimumab binding to the NK cell surface by immunofluorescence microscopy. One potential complication is a nonspecific binding of ipilimumab to NK cells. Human NK cells express antibody receptors (e.g., Fc receptor CD16) which can bind to the constant region of an antibody regardless of the antibody's specificity[256]. To exclude the possibility of nonspecific ipilimumab-NK cell interactions, we used the human NK cell line NK-92, which lacks generic antibody receptors (i.e., CD16) (Fig. 3-5G). Immunofluorescence imaging demonstrated that fluorescently labeled anti-CTLA-4, but not the IgG control, was capable of binding to NK-92 through recognition of CTLA-4 on the cell surface (Fig. 3-5H). The specificity of the stain was confirmed using the CTLA-4 null line PANC-1 (Fig. S3-2E). We saw abundant surface expression of CTLA-4 by immunofluorescence, confirming the results shown in Fig. 3-5F. To the best of our knowledge, this is the first demonstration that anti-CTLA-4 (ipilimumab) can directly interact with human NK cells via a CD16-independent mechanism.

NK cell activation regulates CTLA-4 expression

In T cells, CTLA-4 expression is modulated in response to T cell activation via CD28 and T cell receptor signaling[257]. To investigate if in vitro NK cell activation would similarly modify CTLA-4 expression in NK cells, we exposed NK cells to a variety of cytokines (IL-2, IL-12, IL-15, IL-18) that activate NK cells and alter NK cell expression of other immune checkpoints (i.e., PD-1)[258, 259] (Fig. 3-6A). Human NK cells, with the exception of NK cell line NK-92, had a drastic reduction in CTLA-4 after 24-h exposure to IL-2. IL-15 also caused a reduction in CTLA-4 expression in all NK cells tested except NKL. Alternatively, IL-12 and IL-18 increased CTLA-4 expression in a subset of NK cell lines, including primary donor NK cells. The variability in CTLA-4 expression in response to cytokine stimulation may be attributed to intrinsic differences in the NK cell lines, which can alter their response to certain stimuli. For instance, the NK92 cell line does not express any of the KIR family of inhibitory receptors; therefore, this cell line is thought to be hyper-sensitive to cell-mediated activation[260].

Target cell recognition is another means to activate NK cells. Since cytokine-activated and target cell-activated NK cells have distinct transcriptional phenotypes[261], we also investigated target cell-mediated NK cell activation on NK cell CTLA-4 expression by exposing NK cells to engineered target cells (K562-4-1BB-mbIL-21 cells) (Fig. 3-6B). Although we saw divergent responses in the primary NK cells from two donors, target cell exposure clearly modulated CTLA-4 expression. These data demonstrate that although responses are variable, human NK cell activation, via cytokine and target cell stimulation, alters NK cell expression of CTLA-4. Combined with the observation that anti-CTLA-4 antibodies bind human NK cells, these results suggest CTLA-4 may be an NK cell checkpoint and drive the computationally identified signature of NK cell activation in anti-CTLA-4 responsive tumors. Taken together, these results confirm the utility of CoGAPS and projectR to identify

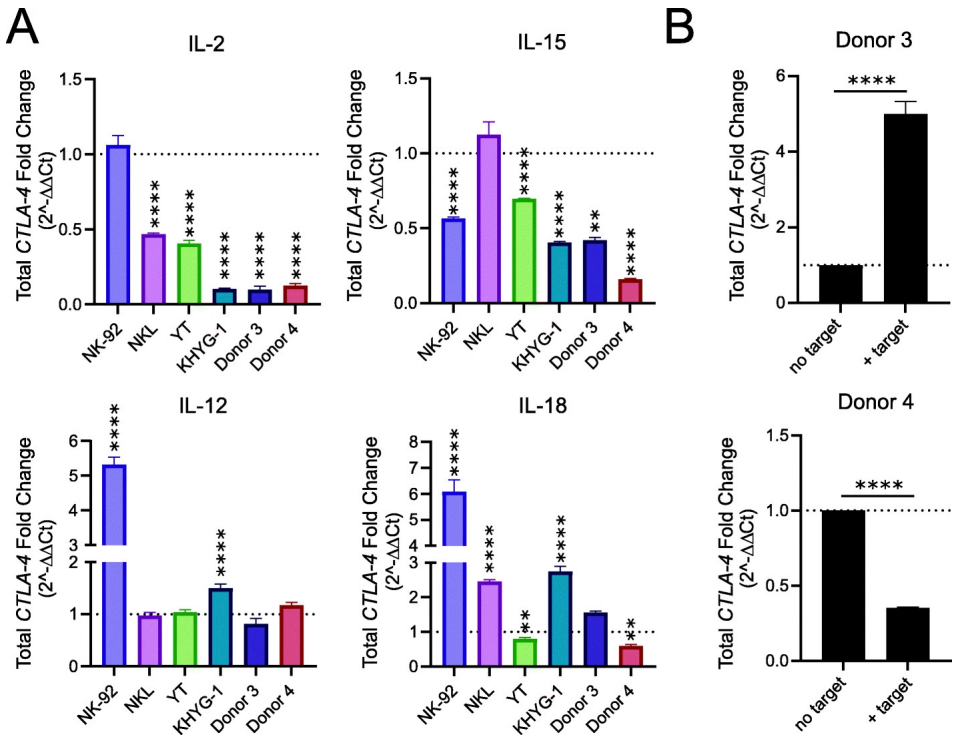

Figure 3-6. NK cell activation regulates CTLA-4 expression. (A) Effect of 24-h stimulation with IL-2, IL-12, IL-15, and IL-18 on NK cell CTLA-4 expression as determined by qRT-PCR (n = 3 for NK cell lines, 2 for donor NK cells; p-values < 0.01 = ** and < 0.0001 = **** when comparing ΔCt for that cell after exposure to cytokine to that cell line unexposed using an unpaired two-tailed t-test). **(B)** Effect of target cell exposure (K562-4-1BB-mbIL-21) on NK cell CTLA-4 expression as determined by qRT-PCR (n = 3, p-value < 0.0001 = **** when comparing Ct using an unpaired two-tailed t-test).

conserved biological processes between preclinical models and human patients that contribute to clinical outcomes.

Preclinical NK cell activation signature is associated with overall survival in metastatic melanoma patients

We hypothesized that the CoGAPS identified NK cell activation signature might be detectable in untreated tumors that naturally elicit an anti-tumor NK cell response, such as melanoma metastases[212]. In addition to the ability to relate biological processes across species, our transfer learning approach can be used to compare across sequencing platforms. Therefore, to investigate if NK cell activation was associated with clinical outcomes in untreated cancer patients, we used projectR to project

bulk RNA-seq data from TCGA of 9507 human tumors representing 32 solid tumor types[223] into the 21 CoGAPS patterns originally identified in scRNA-seq. An association between CoGAPS pattern weight and overall survival was determined using a multivariate Cox proportional hazards model, adjusted for age. In melanoma, pattern 7 weight in metastatic lesions (n = 368) was associated with a longer overall survival (Fig. 3-7A, HR = 0.99, p = 0.017). Pattern 7 weight in primary melanoma lesions (n = 103) was not associated with any statistically significant difference in overall survival (Fig. S3-5). These results show that NK cell activation is significantly associated with overall survival in untreated metastatic melanoma patients. The association between our NK cell activation pattern and clinical outcomes in metastatic lesions is consistent with the role of NK cells in controlling cancer progression and metastasis[262].

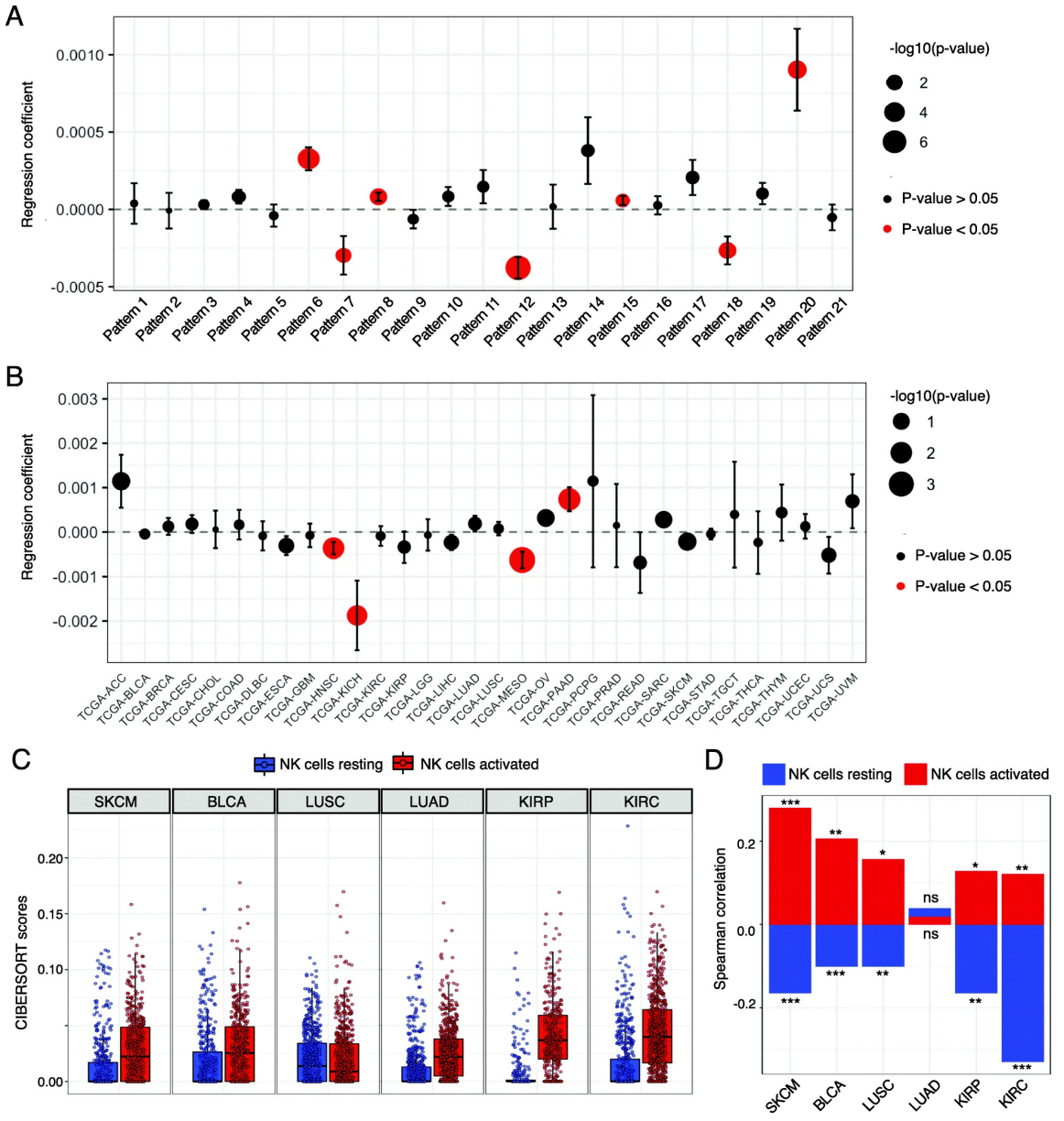

Figure 3-7. Preclinical NK activation signature is associated with overall survival in human melanoma. (A) Coefficients of an age-adjusted multivariate Cox proportional hazards regression model that relates CoGAPS patterns and overall survival in metastatic melanoma lesions from TCGA. Point size scaled to the coefficient's p-value. Red points indicate patterns with significant coefficients. A positive coefficient indicates a worse overall survival and a negative coefficient indicates a better prognosis for the associated variable. **(B)** Coefficients of an age-adjusted multivariate Cox proportional hazards regression model that relates CoGAPS pattern 7 and overall survival across 32 primary tumor types from TCGA. Point size scaled to the coefficient's p-value. Red points indicate patterns with significant coefficients.

Figure 3-7. (C) Boxplot of CIBERSORT scores estimating the abundance of resting and activated NK cells from TCGA RNA-seq data by tumor subtype in TCGA. **(D)** Bar plot of Spearman correlation coefficients between CTLA-4 and CIBERSORT cell type score for immunogenic cancers. CTLA-4 expression is positively correlated with estimation of activated NK cells from TCGA RNA-seq data. Significant correlations for NK scores and CTLA-4 expression are indicated by asterisks where p-values $< 0.05 =$ *, $< 0.01 =$ **, and $< 0.001 =$ ***

When fitting separate Cox proportional hazards models by cancer type across all primary tumor types and adjusting for age, head and neck squamous cell carcinoma (HNSCC), kidney chromophobe (KICH), and mesothelioma (MESO) showed a significant association between pattern 7 weight and overall survival (Fig. 3-7B). Consistent with this, several studies have similarly found an association between infiltrating NK cell abundance or function and overall survival in solid tumor types, including HNSCC[113, 204, 263–266]. Interestingly, pattern 7 weight in primary pancreatic adenocarcinoma (PAAD) was associated with a significantly worse overall survival. Notably, studies of the association between NK cells and disease prognosis in PAAD have had inconsistent findings[267–270]. The association between pattern 7 and worse overall survival in PDAC may be driven by abnormal NK activation or dysregulation of the innate immune system within some lesions. As there is no universal cell type marker to define NK cells and different subsets express standard marker genes differently, studies investigating the relationship between NK cell infiltration and overall survival are limited in their ability to assess the relationship between overall survival and the abundance of functional subpopulations[266]. Bulk RNA-seq similarly suffers from a limited ability to delineate cell types and states from aggregate transcriptional data. In contrast, our results demonstrate we can computationally project transcriptional signatures identified from scRNA-seq data into bulk RNA-seq data to rapidly detect immune cell states shared between distinct species and data modalities. In addition, these results confirm that NK cell activation is associated with overall survival in metastatic melanoma[242].

CTLA-4 expression is positively correlated with the infiltration of active NK cells in immunogenic human tumors

Given that the NK cell activation signature was enriched in anti-CTLA-4-treated mouse tumors, we hypothesized that there may be a correlation between CTLA-4

expression and intratumoral NK cell content. To explore this hypothesis, we used bulk RNA-seq data from TCGA then applied CIBERSORT, a widely used computational approach that infers immune cell content from bulk RNA-seq data[271]. In this analysis, we assessed six immunogenic solid tumor types: skin cutaneous melanoma (SKCM), kidney renal clear cell carcinoma (KIRC), cervical kidney renal papillary cell carcinoma (KIRP), squamous cell carcinoma of the lung (LUSC), lung adenocarcinoma (LUAD), and bladder carcinoma (BLCA). When running CIBERSORT, we used the LM22 signature matrix designed by Newman et al.[271] to estimate the relative fraction of 22 immune cell types within input mixture samples, including an estimation of resting and activated NK cell proportions (Fig. 3-7C). Correlation analysis across the 21 CoGAPS patterns for the genes present in both the CoGAPS amplitude matrix and the LM22 signature matrix (n = 391) found that pattern 7 had the highest correlation (Pearson = 0.497) to the CIBERSORT NK cell activation signature (Table S4), further supporting the association between pattern 7 and NK cell activation. Correlation analysis between CTLA-4 expression and CIBERSORT cell type estimation revealed that the direction of correlation in NK cells was dependent upon the activation state (Fig. 3-7D, Table S5). Across several tumor types, the proportion of activated NK cells was positively correlated with CTLA-4 expression, while the proportion of resting NK cells was negatively correlated. CTLA-4 expression was negatively correlated with estimated proportions of resting NK cells in SKCM ($p < 1 \times 104$), BLCA ($p < 1 \times 103$), LUSC ($p < 1 \times 102$), KIRP ($p < 1 \times 102$), and KIRC ($p < 1 \times 109$). On the other hand, estimated proportions of activated NK cells were positively correlated with CTLA-4 expression in SKCM ($p < 1 \times 106$), BLCA ($p < 1 \times 102$), LUSC ($p < 0.05$), KIRP ($p < 0.05$), and KIRC ($p < 1 \times 102$). As expected, CTLA-4 expression was also positively correlated with the estimated proportions of regulatory T cells (Tregs) in each tumor type (Table S5). This analysis complements our experimental results and further supports a relationship between NK cell activation, CTLA-4 expression,

and clinical outcomes in human tumors.

Discussion

In this application of matrix factorization and transfer learning to cancer immunotherapy, we demonstrate both computationally and experimentally that this approach can elucidate complex immunotherapy responses from scRNA-seq data that are conserved across species. Specifically, we show that our matrix factorization approach (CoGAPS) detected a signature of intratumoral NK cell activation in anti-CTLA-4-treated mice which our transfer learning method (projectR) associated with positive clinical outcomes in metastatic melanoma. We interrogate and validate this NK cell activation signature in several datasets, including proteomics (CyTOF), bulk RNA-seq (TCGA), and additional scRNA-seq. Ultimately, the application of these computational techniques identified novel biology—that human NK cells express CTLA-4, bind anti-CTLA-4 (ipilimumab), and NK cell activation associates with anti-CTLA-4 activity in human tumors.

Both CoGAPS and projectR offer unique advantages to interpreting complex tumor immune cell scRNA-seq data. For instance, traditional clustering methods such as those employed by Gubin et al.[21] group cells according to transcriptional signatures that reflect cell type. However, a single cell's transcriptional profile represents more than just cell type, encompassing additional cellular processes such as activation, exhaustion, and cell signaling, which are not necessarily captured by traditional clustering approaches. Identifying these cellular processes is particularly important when studying immune cells within the tumor microenvironment, where cells may undergo stimulation or dysregulation. In the scRNA-seq data, Gubin et al.[21] did not detect NK cell activation in anti-CTLA-4-treated tumors; however, their subsequent CyTOF analysis revealed prominent upregulation of NK cell granzyme expression specific to anti-CTLA-4 treatment[21]. In contrast, our matrix factorization

method, CoGAPS, was able to identify NK cell activation in response to treatment directly—without the need for clustering, differential expression analyses, or additional technologies— highlighting the advantage of CoGAPS compared to standard analysis methods when studying tumoral immune cells. Using projectR to project the NK cell activation signature into several additional datasets allowed us to ultimately confirm that the transcriptional signature we identified in mice was clinically relevant in humans as well. This is particularly impressive when you factor in the known differences between mouse and human NK cell surface receptors and markers[272]. In this application, we use gene signatures from CoGAPS for projection and transfer learning. Other transfer learning methods have been developed to relate features in a target scRNA-seq dataset to a reference atlas, often relying on non-linear methods for feature identification[273, 274]. In contrast to these other approaches, our projectR software is robust for transfer learning from single-cell data (e.g., PCA, clustering, and other forms of linear matrix factorization) and may capture additional features of cell state transitions based upon all of these methodologies[112, 230]. Future extensions to projectR are needed to enable transfer learning from an ensemble of features across these latent space methods and from emerging non-linear methods for inference of more complex cell state transitions and gene regulatory networks.

The CoGAPS analysis of the scRNA-seq data from an immunotherapy-treated mouse model identified several immune cell states associated with treatment status, including the myeloid compartment. Notably, CoGAPS detected an M2 macrophage signature enriched in untreated mice and an M1 macrophage signature enriched in tumors from anti-PD-1 treated mice (Fig. 3-2D-G). We chose to focus our experimental validation on the NK cell activation signature identified by CoGAPS (pattern 7) for several reasons: (1) pattern 7 was the most clearly associated with a specific cell type and treatment, (2) increased expression of NK cell activation markers had been noted in anti-CTLA-4-treated mice from the original CyTOF analysis[21], (3) there

is growing evidence that CTLA-4 is expressed by non-T cell human immune cell types[245–248], and (4) recent work found that human NK cells express PD-1 and are modulated by anti-PD-1 therapy[275, 276]. Therefore, we hypothesized that CTLA-4 was similarly expressed by human NK cells and activated by anti-CTLA-4 antibodies.

In addition to the experimental validation, our computational analysis with transfer learning demonstrated that the NK cell activation signature is associated with improved overall survival and anti-CTLA-4 response in melanoma patients. This signature was detected in anti-CTLA-4 responsive metastatic melanoma prior to the administration of treatment and correlated with response to therapy. This leads us to hypothesize that the presence of activated NK cells already within tumors improves tumor clearance mediated by anti-CTLA-4. The NK cell activation signature was also elevated in a patient that initially responded to a combination of anti-PD-1 and oncolytic virus therapy. This observation is consistent with previous studies showing that infection of tumors with oncolytic viruses can activate NK cells and stimulate NK-mediated anti-tumor immunity[277]. We note that this observation was specific to intratumoral NK cells and not present in circulating NK cells (Fig. 3-4E), indicating that approaches using peripheral blood to transcriptionally profile NK cell activation with respect to clinical outcomes may be limited. Future transfer learning analyses on large cohort studies of anti-CTLA-4-treated tumors with genomics data could further delineate the role of tumor NK cell activation as a potential predictive biomarker. However, these datasets are currently lacking in the literature, limiting our ability for such computational-driven biomarker analysis in this current study.

While our study is computationally focused, the application of our transfer learning pipeline for cross-species analysis to cancer immunotherapy still suggests that the role of NK cells in anti-CTLA-4 response is preserved between preclinical mouse models and human tumors. Despite growing evidence for the role of checkpoint receptors in NK cell-mediated anti-tumor responses, the expression of CTLA-4 in NK cells has

been disputed in the literature for both mice and humans. Although mouse NK cells have been shown to inducibly express CTLA-4 in response to IL-2[258], a recent study was unable to detect CTLA-4 on the surface of intratumoral mouse NK cells[241]. A study in humans also reported that NK cells from healthy donors do not express CTLA-4[249]. Contrary to these earlier reports, our results demonstrate CTLA-4 is constitutively expressed by circulating healthy human donor NK cells and human NK cell lines. One possible explanation for why previous studies have failed to identify the expression of CTLA-4 by human NK cells is the reliance on flow cytometry in these studies. Flow cytometry can be limited by challenges related to the generation of antibodies and further complicated by the rapid surface expression dynamics of CTLA-4[278]. In support of this explanation, we too fail to detect intracellular or surface CTLA-4 expression when using flow cytometry (Fig. S3-3A and B), even though we are able to unequivocally demonstrate CTLA-4 expression at the RNA and protein level by qRT-PCR and western blot in ex vivo unstimulated healthy donor NK cells (Fig. 3-5B–E), as well as surface expression using immunofluorescence and biotinylation (Fig. 3-5G). Consistent with previous studies[279, 280], we show that human NK cells express CD28 and CD28H (Fig. S3-4), a co-stimulatory receptor that competes with CTLA-4 for the binding of B7 ligands. The expression of B7 on tumor cells also enhances NK recognition and lysis of tumors through CD28-B7 interactions[279–285]. In addition, we show that CTLA-4 expression by human NK cells cultured in vitro is modulated in response to NK cell activation (Fig. 3-6). These findings suggest that CTLA-4 may have similar functions in NK cells and effector T cells[257]. Taken together, these results build upon previous studies that highlight a relationship between NK cells and anti-CTLA-4 response in humans. In melanoma patients treated with anti-CTLA-4, a higher percentage of circulating mature NK cells is correlated with improved overall survival, and NK cells isolated from responsive patients have increased cytolytic activity compared to NK cells isolated from non-

responders[286]. In B16 melanoma models, NK cells and CD8+ T cells synergistically clear tumors in response to anti-CTLA-4 and IL-2 treatment[287]. Furthermore, anti-CTLA-4 has been shown to increase transcriptional markers of NK cell cytotoxic activity in CT26 colon carcinoma tumors[241]. While future mechanistic studies are needed to fully elucidate the specific function(s) of CTLA-4 in NK cell biology, these findings support the computationally driven translational approach employed in this study.

Conclusions

As scRNA-seq datasets of immunotherapy-treated tumors become increasingly prevalent in cancer research, we need appropriate computational tools that can delineate actionable cellular mechanisms of action from these data. This inference can play a critical role in advancing basic science in the preclinical research pipeline, where relating findings to human datasets enables translation for precision immunotherapy strategies. This work describes a framework using latent space discovery through matrix factorization and transfer learning for cross-species data analysis which allows the integration of preclinical and clinical genomics datasets. We provide a powerful method for extrapolating relevant information while avoiding the unique biases of individual technologies (i.e., dropout in scRNA-seq, biased selection of genes in CyTOF, or aggregate transcriptional profiles in bulk RNA-seq). In addition, our approach enables the comparison of different tumor types and treatment conditions. While our study focused on the relation of preclinical models to human tumors, this approach can be readily applied within human tumors to relate mechanisms across tumor subtypes and can be broadly used in other disease contexts as well as drug repurposing. The ability to rapidly identify conserved therapeutic responses between mice and humans will help bridge basic science and clinical research to improve patient outcomes.

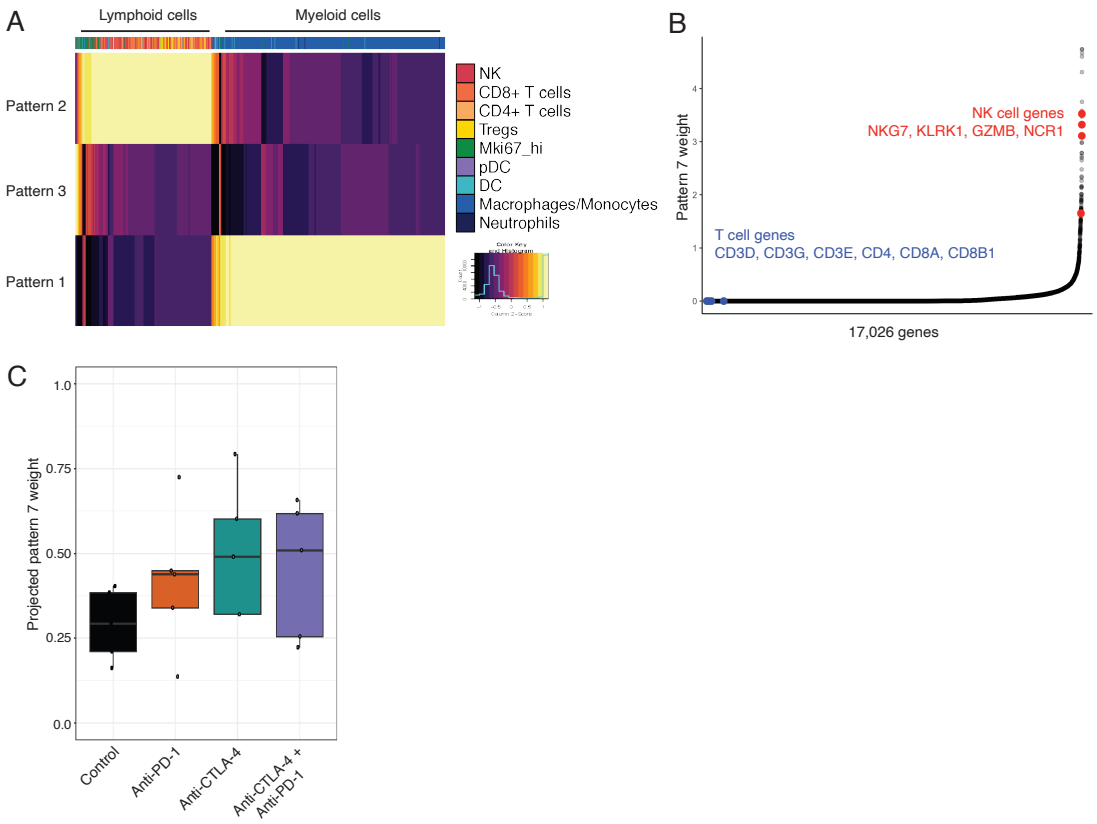

Figure S3-1. CoGAPS patterns identify immune cell lineage and transfer across data modalities. (A) Heatmap of transcriptional signatures (patterns) identified with CoGAPS. When CoGAPS is performed at low dimensionality, here being 3 patterns, the identified transcriptional signatures segregate cells by immune cell lineage. Pattern 3 is relatively flat across all cells, while patterns 1 and 2 define myeloid and lymphoid lineage cells, respectively. **(B)** Scatter plot of pattern 7 gene (amplitude) weights for all 17,026 genes included in the CoGAPS result object, indicating that pattern 7 is specific for NK cell marker genes compared to T cell marker genes. **(C)** Boxplot of the projected NK cell activation signature (pattern 7) weights in tumor infiltrates from mouse tumors analyzed by mass cytometry on day 11 after treatment. Each point represents a replicate sample. For each replicate, the mean protein expression of 37 genes was used as input for projectR. The NK cell activation signature is highest in lymphocyte samples treated with anti-CTLA-4, either alone or in combination with anti-PD-1.

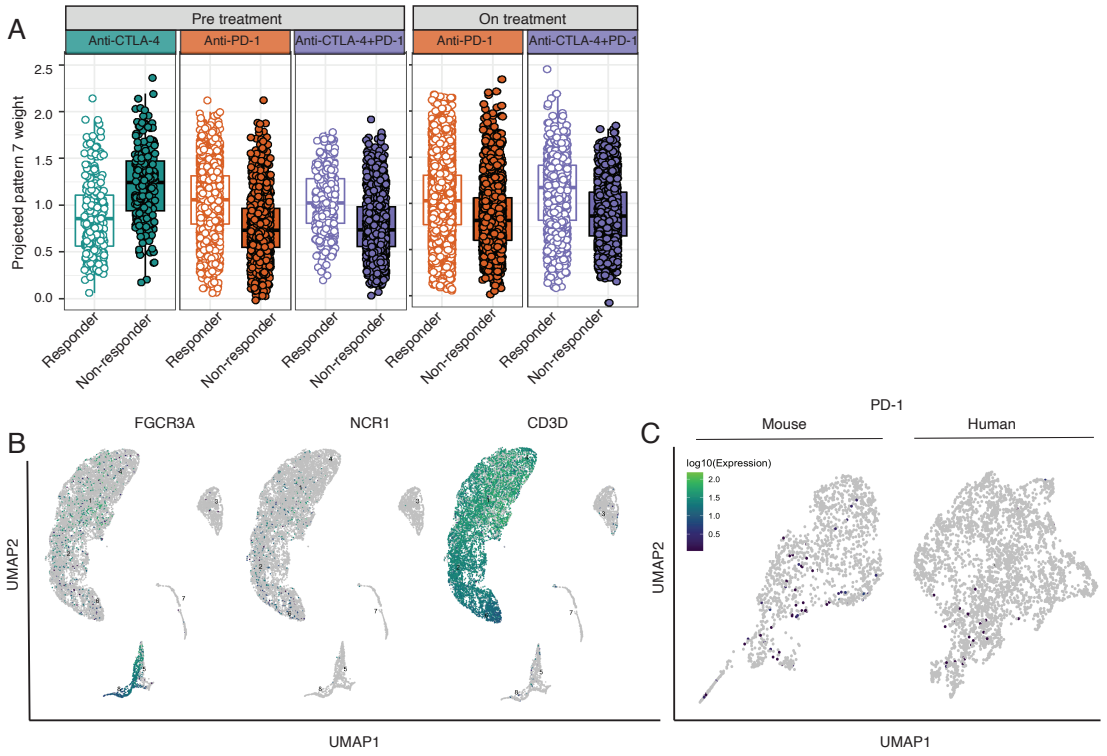

Figure S3-2. NK cell activation signature is associated with anti-CTLA-4 response. **(A)** Box plot of projected pattern 7 weights across intratumoral immune cells from metastatic melanoma patients prior to ICI treatment. Cells are colored by therapy and separated by patient response. Increased pattern 7 is associated with immune cells from patients responsive to antiCTLA-4. **(B)** UMAP dimension reduction with cells colored by single-cell gene expression for representative NK and T cell marker genes. **(C)** UMAP dimension reduction with cells colored by single-cell gene expression for PD-1 in mouse (left) and human (right) intratumoral NK cells. Activated NK cells are known to express PD-1, demonstrating that the observed pattern of PD-1 expression is consistent with the reduced ability of scRNA-seq to capture low to moderate expressed genes.

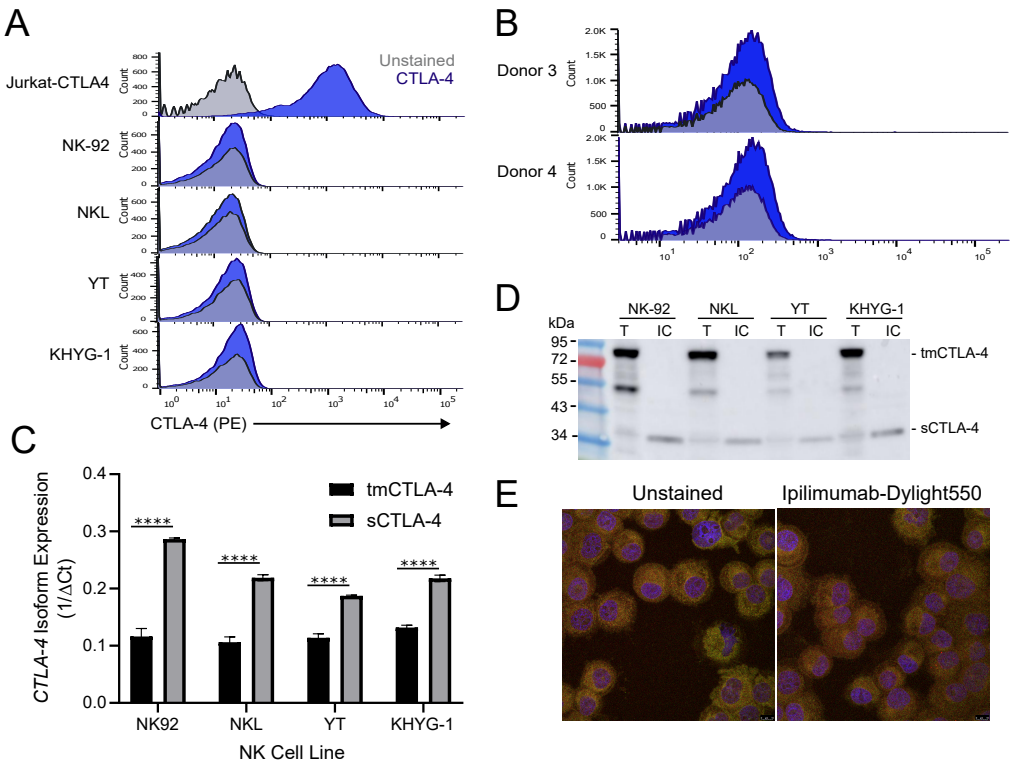

Figure S3-3. Human NK cells express CTLA-4. (A) Flow cytometry for surface expression of CTLA-4 in positive control (Jurkat-CTLA4) and NK cell lines (NK-92, NKL, YT, KHYG1). **(B)** Flow cytometry for surface expression of CTLA-4 on CD56+ selected ex vivo unstimulated NK cells derived from healthy human donors. **(C)** Quantitative qRT-PCR analysis of transmembrane (tmCTLA-4) and soluble (sCTLA-4) isoforms in human NK cell lines, (n=6, p-value < 0.001 = ****). **(D)** Western blot of total protein (T) and intracellular (IC) protein isolated from human NK cell lines NK-92, NKL, YT and KHYG-1 using cell surface protein biotinylation for exclusion of surface proteins demonstrating surface expression of CTLA-4 dimers and intracellular expression of CTLA-4 monomers. Western blot is representative of two independent experiments. **(E)** Immunofluorescent images of PANC-1 cells stained with Dylight550-labelled ipilimumab. Blue staining indicates DAPI. Shown are representative images of a single field of view taken via confocal microscopy (magnification, 63X).

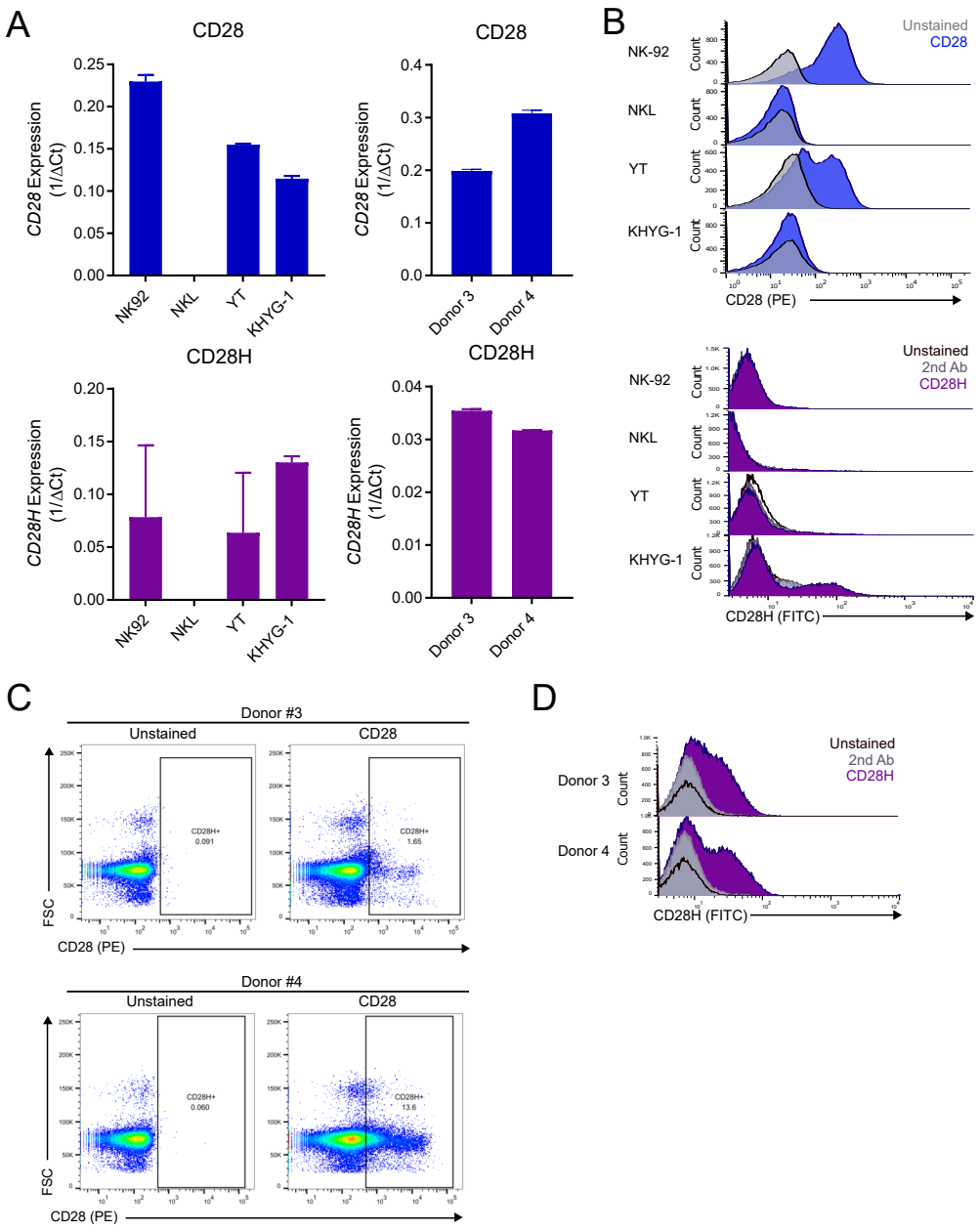

Figure S3-4. CD28 and CD28H expression on human NK cells. (A) qRT-PCR assessment of CD28 and CD28H expression in human NK cell lines and primary donor NK cells, (n=6 for NK cell lines, n= 2 for donor NK cells). **(B)** Flow cytometry assessment of CD28 and CD28H surface expression by human NK cell lines **(C)** Flow cytometry assessment of CD28 surface expression by primary donor NK cells. **(D)** Flow cytometry assessment of CD28H surface expression by primary donor NK cells.

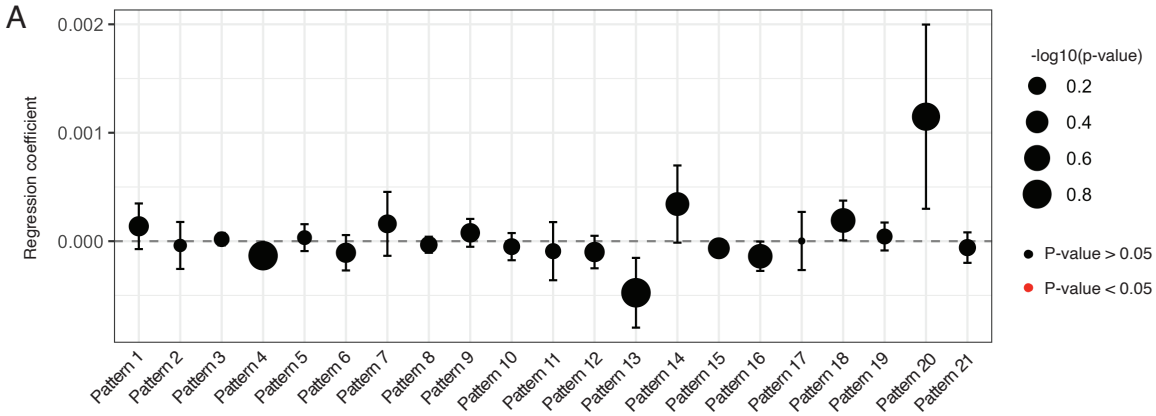

Figure S3-5. Regression coefficients of pattern associations with TCGA tumor survival. (A) Regression coefficients of an age-adjusted multivariate Cox proportional hazards regression model that relates CoGAPS patterns and overall survival in primary melanoma lesions from TCGA. Point size scaled to the coefficient's p-value. Red points indicate patterns with significant coefficients. No regression coefficients for any CoGAPS pattern are significant in primary melanoma.

Chapter 4

Physicochemical features of T cell receptor sequences identify circulating mutant KRAS-specific T cells after peptide vaccination

Abstract

Neoantigen vaccines trigger T cells against the somatic mutations expressed in cancer cells leading to robust anti-tumor immunity. Clinical trials have begun to characterize the phenotype of vaccine-induced neoantigen-specific T cells and their contribution to durable immunological control of tumor growth. Nonetheless, there are currently no established approaches to identify vaccine-induced T cells directly from T cell receptor (TCR) sequencing data, without the need for in vitro expansion to screen for T cell specificity. We have developed Homolig, an algorithm to quantify the degree of physiochemical homology between TCR sequences. Using 30,315 TCRs with known specificity from public databases and published literature, we have shown that our method reliably groups TCRs that share highly conserved physiochemical motifs. We applied this approach to 15,413 TCRs captured by longitudinal single-cell RNA and TCR-sequencing of peripheral blood from a patient vaccinated with 6 mutant KRAS peptides. We identified TCRs enriched in the periphery post-vaccination

with high physiochemical homology to TCRs with known specificity to immunogenic mutant KRAS epitopes. To establish that Homolig identifies antigen-specific TCRs, we used CRISPR-mediated TCR replacement to clone receptors of interest into healthy donor T cells. We confirmed that Homolig identifies TCRs with neoantigen-specific recognition against mutant KRAS peptides. We show for the first time that distinct neoantigen-specific TCRs can be detected directly from unstimulated peripheral blood by physicochemical sequence homology of the TCR alone. This work provides new insights into the nature of TCR recognition in a way that will accelerate the analysis of T cell responses to immunotherapy and the identification of neoantigen-specific T cell populations.

Introduction

Immune checkpoint inhibitors (ICIs) have benefited 20% of patients with previously incurable cancers, including melanoma and non-small cell lung cancer. However, immunologically 'cold' or insensitive tumors such as pancreatic ductal adenocarcinoma (PDAC) fail to respond to single-agent or combination ICIs. This resistance in part arises from the low tumor mutational burden or the low number of neoantigens of such immunologic 'cold' tumors. Neoantigen-specific T cells have been shown to be critical drivers in generating effective anti-tumor responses to ICIs. This has led to testing vaccines targeting neoantigens in human studies that have demonstrated the induction of peripheral de novo high-quality T cells post-vaccination[288, 289]. However, in immunologic cold tumors such as PDAC, multiple clinical and preclinical studies have demonstrated the need to combine neoantigen-targeted vaccines with immunomodulatory agents in order to reduce T cell exhaustion and generate robust and durable anti-tumor immunity with the potential to improve clinical benefit.

Neoantigen-targeted vaccines have the ability to induce de novo high-quality cytotoxic neoantigen-specific T cells in the periphery and tumor site. These vaccines

can either be individualized to each patient's tumor neoantigen profile or can be 'off-the-shelf' targeting known recurrent hotspot neoantigens, such as oncogenic products (e.g KRAS mutations). KRAS mutation is present in over 90% of pancreatic tumors and serves as an attractive target for vaccine therapy. Prior attempts using adoptive T cell therapy targeting mutant KRAS have demonstrated the clinical efficacy of immunologically targeting mutant KRAS. Here, we have performed in-depth analyses of peripheral blood samples from a patient with resected PDAC vaccinated with a mutant KRAS long peptide vaccine in combination with ipilimumab and nivolumab (anti-CTLA-4 and anti-PD-1 antibodies, respectively).

A major rate-limiting step for interpreting responses to immunotherapy is the ability to identify neoantigen-specific T cell receptors (TCRs) from patient specimens. Advances in single-cell technologies have enabled simultaneous analysis of TCR sequences and their corresponding transcriptomes. Recent studies have begun to leverage single-cell transcriptomics to characterize the phenotype of neoantigen-reactive T cells within tumors[290–293]. However, these studies relied on either antigen stimulation approaches to enrich for neoantigen-specific T cells or on the previous identification of neoantigen-specific TCRs. The reliance on single-cell gene expression to identify neoantigen-specific T cells may be limited due to the pervasive nature of dropout, resulting in the inability to completely capture the expression of a gene across cells. Moreover, these studies rely on access to fresh tumor samples. Methods to identify neoantigen-specific T cells directly from single-cell TCR sequences could circumvent these limitations and enable characterization beyond the tumor microenvironment to peripheral blood.

To address this problem, we developed Homolig, an algorithm to assess TCR sequence homology based on physicochemical similarity. We applied Homolig to longitudinal scRNA and TCR-sequencing to identify TCR clones involved in the vaccine-induced immune response. We demonstrate that this approach successfully

distinguishes mutant KRAS-specific TCRs present after vaccination. For the first time, we show the ability to identify neoantigen-specific T cells directly from unstimulated peripheral blood without the influences of ex vivo peptide stimulation and expansion. These results provide a framework for using the peripheral blood T cell repertoire to monitor immunotherapy responses and will accelerate the development of personalized cancer therapies based on the infusion of TCR-engineered T cells.

Methods

KRAS vaccine treatment schedule

The patient underwent surgical resection of pathology confirmed PDAC lesion followed by perioperative chemotherapy/radiation. The tumor specimen underwent molecular testing for KRAS mutations including KRAS G12V, G12C, G12R, G12A, G12D, or G13D. XX weeks after the final round of standard-of-care chemotherapy, the patient received a pool of 6, 21-mer KRAS peptides dissolved at a final amount of 0.3mg/peptide in saline and 0.5mg of PolyICLC (Hiltonol; Oncovir). The patient received the shared mutant KRAS vaccine plus ipilimumab (1mg/kg, IV, every 6 weeks for 2 doses) and nivolumab (3mg/kg, IV, every 3 weeks in the priming phase) followed by nivolumab (480mg, IV, flat dose in boost phate) for one year. Peripheral blood samples were taken every four weeks.

Single-cell RNA (scRNA-seq) and TCR (scTCR) sample preparation and sequencing

PBMCs were thawed on ice, washed, and resuspended in PBS1X containing 1% BSA, and cell counts and viability were made using Trypan Blue staining (ThermoFisher) in the hemocytometer. scRNA library preparations were performed using the 10× Genomics Chromium™ Single Cell system and Chromium™Single Cell 5' Library Gel Bead Kit v2 (10x Genomics). TCRs were enriched using the TCR Amplification

kit (10x Genomics), following the manufacturer's instructions. The initial cell input was 17,000 PBMCs to recover a total of 10,000 cells. Sequencing was performed on the NovaSeq platform (Illumina) using 10x Genomics recommended features and with a depth of 50,000 reads per cell. Sequences were processed using the Cellranger 5.0.1 pipeline (10x Genomics) and mapped to the human reference genome (GrCh38).

Single-cell RNA and TCR-seq analysis

The raw feature-barcode matrix was imported into Scanpy (version 1.8.1)[96] for processing. Putative cell doublets were removed using the Scrublet package[294]. Leiden clustering was performed at a resolution of 2.0 to capture major cell types and subtypes. Differentially expressed genes across the leiden clusters were determined using the *scanpy.tl.rank_genes_groups* function. Clusters were manually annotated based on the RNA expression of known cell type marker genes and confirmed using annotation pipelines like Azimuth[90] and Celltypist[295]. Differentially expressed genes between clusters or between the time points (C1, C2, C6) were determined through Wilcoxon statistical tests (p-value<0.01). TCR repertoire analysis was performed with Scirpy (version 0.9.1)[296]. Productive TCR chain pairing status was determined with the *scirpy.tl.chain_qc()* function. For Homolig analysis, only single pair TCRs were included.

Generation of a physiochemical substitution matrix

The presence of different amino acids in a TCR with similar physicochemical properties may confer similar antigen recognition. To identify TCR sequences with different amino acid sequences but similar antigen-binding capabilities, we first mapped amino acid sequences into a lower-dimensional space that captures functional similarities between two each possible acid pair. We used the 566 numerical indices from AAindex (Version 9.2) [297] to produce a 20 x 566 matrix containing the physicochemical descriptors.

A 20 x 20 euclidean distance matrix was used to represent the physiochemical space between amino acids. To generate a substitution matrix for pairwise alignment scoring, the inverse was taken to provide a similarity measure. This substitution matrix provides a pairwise numerical value based on the underlying physicochemical properties of the component amino acids. Other available blast substitution matrices were downloaded from NCBI (ftp://ftp.ncbi.nih.gov/blast/matrices/) and are provided as alternatives in Homolig.

Pairwise TCR alignment and clustering with Homolig

Homolig takes N TCR sequences as input. TCR data from various platforms typically provides a CDR3 amino acid sequence and a variable gene name. Homolig first parses the variable genes to extract the CDR1, CDR2, and CDR2.5 amino acid sequences using reference FASTA files obtained from IMGT[298, 299] (http://www.imgt.org/vquest/refseqh.htmlreferences).

CDR alignment scores are calculated on the basis of the pairwise scoring matrix described above, which is derived from 566 physicochemical properties of amino acids[297]. For CDRs with different lengths, Homolig performs a sliding window comparison based on the length of the shorter sequence and the highest-scoring alignment is used.

For single-pair TCRs with α and β information generated by single-cell technologies, 8 N-by-N pairwise CDR matrices are generated. The CDR matrices are added together so that the CDR3 and variable gene CDRs contribute equally. After calculation of the N-by-N pairwise TCR matrix, Homolig stores the input file and output matrix in anndata format, which allows for leiden clustering analysis using the scanpy framework[96].

Homolig supports the analysis of variable and hypervariable CDRs from the α and β chain alone or in combination. CDR3 sequences can also be compared without the

inclusion of variable gene CDRs. While only human sequences are analyzed in this work, Homolig supports the analysis of T cell or B cell receptor sequences from any of the 30 species present in IMGT.

Data collection of TCR sequencing data with known specificity

Antigen-specific TCR sequences with paired α and β information were obtained from VDJdb[300] and previous literature[301, 302].

T cell engineering

10 patient TCRs that clustered with known mKRAS-specific TCRs were selected for validation. For construct design, full-length coding TCR sequences based on V/J/CDR3 information were automatically generated using a custom python script. This code is publicly available on the Homolig Github.

TCR knock-ins were generated through nucleofection of Cas12a protein complexed with *TRAC* and *TRBC*-targeted gRNAs as well as a double-stranded DNA homology-directed repair template (HDRT). The HDRT contains homology arms for the *TRAC* locus and encodes the exogenous TCR α and β chains as well as a truncated NGFR tag under the control of an EF1a promoter. Successful knock-in of the construct disrupts the expression of the endogenous α chain from that locus. Murine constant domains were used instead of human constant domains to improve the expression of the exogenous TCR in the 5-10% of cells in which TRBC knockout was incomplete.

Cell lines and cell culture

Engineered T cell lines were grown in RPMI 1640 supplemented with 10% FBS, 1% penicillin-streptomycin, 100 IU/mL IL-2 (Peprotech, cat200-02), and 5 ng/mL IL-7 (Peprotech, cat200-07). PBMCs were cultured in media containing a 1:1 mixture of AimV and RPMI supplemented with 5% human serum, 2mM L-Glutamine, 25mM

HEPES, and 10mM Pen/Strep (Gibco, cat15140-122). Monocyte-derived dendritic cells were cultured in RPMI containing 5% human serum, 2mM L-Glutamine, 1% Pen/Strep and 50mM HEPES. All cells were cultured and maintained at 37 °C and 5% CO2.

Peptides

Mutant KRAS vaccine peptides were synthesized by JPT at GMP-grade with sequences as follows:

G12V- YKLVVVGAVGVGKSALTIQLI,

G12C- YKLVVVGACGVGKSALTIQLI,

G12A- YKLVVVGAAGVGKSALTIQLI,

G12R- YKLVVVGARGVGKSALTIQLI,

G12D- YKLVVVGARGVGKSALTIQLI,

G13D- YKLVVVGAGDVGKSALTIQLI.

The sequence for the control peptide corresponds to an irrelevant fusion gene product as follows:

RKREIFDRYGEEVKEFLAKAKEDF.

Lyophilized peptides were dissolved in DMSO at 4mg/ml and aliquots were stored at -80 °C.

Antibodies for flow cytometry

anti-CD3-FITC (clone HIT3a Biolegend cat 300306, or clone SK7 Biolegend cat 344804), anti-CD8-BV421 (clone RPA-T8, Biolegend cat301036), anti-CD4-BV605 (clone RPA-T4, Biolegend cat300556), anti-mTCRb-APC (clone H57-597, Biolegend cat109211), anti-NGFR-PE/Dazzle (clone ME20.4, Biolegend, cat345119), anti-CD45RO-AF700 (clone UCHL1, cat304218), anti-CD62L-PerCpCy5.5 (clone DREG-

56, Biolegend cat 304823), anti-CD69-PE (clone FN50, Biolegend cat310905), anti-CD137-PECy7 (clone 4B4-1, Biolegend cat309817), anti-IFNg-APC (clone 4S.B3, Biolegend cat502512), anti-IL-2-PE/Dazzle (clone MQ1-17H12, Biolegend cat500343), anti-TNFa-BV650 (clone Mab11, Biolegend cat502938), anti-CD11c-PE (clone 3.9, Biolegend cat301605), anti-HLA-ABC-APC (clone W6/32, Biolegend cat311409), anti-CD86-BV421 (clone BU63, Biolegend cat374209), anti-CD83-PE/Dazzle (clone HB15e, cat305327). Zombie NIR fixable viability stain (Biolegend, cat423105) was used for live/dead cell differentiation.

PBMC peptide restimulation to assess vaccine-induced T cell responses

PBMCs were thawed and rested overnight in complete media at 1e6 cell/well in a 96 well plate. The next day, DMSO control, the control peptide, or KRAS G12V, G12C, G12R, G12A, G12D, G13D 24-mer peptides were added to each well at a final concentration of 2ug/mL. PBMCs were incubated for 48 hours. Six hours prior to collection, eBioscience protein transport inhibitor cocktail (Thermo Scientific, cat00-4980-03) was added at a final dilution of 1X. Cells were collected and first stained with anti-CD3, anti-CD4, anti-CD8, anti-CD45RO, anti-CD62L, anti-CD69, anti-CD137. Cells were permeabilized with BD Perm/Fix kit (BD Biosciences, cat 554714) followed by staining with anti-IFNγ, anti-IL-2, anti-TNF.

Generation of monocyte-derived dendritic cells

CD14+ cells were sorted from patient PBMCs with the EasySep Human CD14 positive selection kit II (Stem cell, cat 17858). Isolated CD14+ cells were seeded at 3e6 cell/well in 6 well dishes and cultured in complete media containing 800IU/mL hGM-CSF (Peprotech, cat300-03) and 200IU hIL-4 (Peprotech, cat200-04) for 6 days. Media was replenished with fresh cytokine containing media every 48 hours. On day 6, moDCs

were matured with 2ng/mL hIL-1β, 1000IU/mL hIL-6, 10ng/ml hTNF, and 1ug/mL hPGE2. Forty-eight hours later, moDCs were collected using 0.9mM EDTA in PBS. moDCs were washed with PBS and pulsed with 4ug/mL control peptide or pooled KRAS G12V, G12C, G12R, G12A, G12D, G13D 24-mer peptides. Flow analysis of moDCs was performed to quantify CD11c and HLA class I expression as well as CD86 and CD83 maturation markers.

CD3- antigen-presenting cell population isolation

CD3+ cells were separated from patient PBMCs with the EasySep Human CD3 positive selection kit II (Stem cell, cat17851). The CD3- population was seeded at 1e6 cell/well in 48 well plates and pulsed overnight with 4ug/mL control peptide or individual KRAS G12V, G12C, G12R, G12A, G12D, or G13D 24-mer peptides in the presence of 250IU hIFNγ (Peprotech, cat300-02) at 37 °C. The next day peptide pulsed CD3- APCs were collected, washed once with PBS, and co-cultured with TCRs of interest. Flow analysis of CD3- cells was performed to quantify the CD11c, CD14, CD19 populations present in addition to HLA class I, HLA class II, and CD86 expression.

TCR coculture

Antigen presenting cell populations were pulsed with pooled or individual KRAS peptides as described previously. After collection and one wash in PBS, 2-5e4 APCs were seeded in 96 well U-bottom plates in T cell media. 1e5 recombinant TCR expressing cells were co-cultured with peptide-pulsed APC populations overnight at 37 °C.

ELISPOT

Multiscreen 96-well filtration plates (Millipore Sigma, cat MSHAS4B10) were coated overnight at 4 °C with 100ul/well of anti-human-IFNγ monoclonal Ab (Clone 1-D1K, Mabtech, cat 3420-3-1000). Wells were washed 3 times with PBS and blocked for 2 hours with RPMI + 10%FBS + P/S at 37 °C. 4e4 mature monocyte-derived dendritic cells pulsed with the control peptide or the KRAS peptide pool were added to the plate in 100ul T cell media. 1e5 TCR knock-in or KO control T cell populations were added to each well in 100ul T cell media and co-cultures were incubated overnight, 37 °C. Cells were removed and the plate was washed 6 times with 0.05% Tween-20 (Millipore Sigma, catP1379) in PBS. Wells were incubated for 2 hours at room temperature with 10ug/mL biotinylated anti-human IFNγ monoclonal antibody (clone 7-B6-1, Mabtech, cat 3420-6-1000) in 0.05%FBS in PBS. Wells were washed as before and wells were incubated with avidin peroxidase complex (Vectastain ELITE ABS kit, Vector Laboratories, catPK-6100) for 1 hour at room temperature. Wells were washed as before. AEC substrate (Vector Laboratories, catSK-4200) was added and wells were developed for 10-15 minutes at room temperature. The reaction was stopped with 6 washes with tap water and plates were allowed to dry for 24 hours before they were counted using an automated ELISPOT reader (ImmunoSpot).

Flow cytometry to assess T cell reactivity

Cells were collected and washed twice with PBS. Zombie NIR viability stain was added to each well and incubated for 15 minutes at room temperature in the dark. Cells were then washed twice in FACS buffer containing 1x HBSS (Gibco, Serum 2%, Na Azide 0.1% and HEPES 0.1%). Extracellular antibody stain was added in FACS buffer to each well and incubated for 20 minutes at 4 °C. Cells were washed twice with FACS buffer. In the case where intracellular cytokine staining was performed, cells were permeabilized with BD Fix/Perm kit according to manufacturer's instructions.

Intracellular antibody stain diluted in 1X perm/wash buffer was added to each well and incubated for 30 minutes at 4 °C. Cells were washed twice with 1X Perm/wash buffer and resuspended in FACS buffer.

All samples were run on a Beckman Coulter Cytoflex cytometer. Data was analyzed in Flowjo version 10.8.1.

TCR sequences from TCGA RNA-sequencing data

MiTCR processed TCR CDR3 sequences across 10,000 tumors comprising 33 cancer types from TCGA were downloaded from GDC upon approval(Thorsson et al. 2018). TCR sequences from tumors with KRAS mutations were identified using cBioPortal. TCR sequences from mutant KRAS tumors were selected for Homolig analysis. This resulted in 24,152 α chain TCR sequences from 624 samples and 27 tumor types and 76,249 β chain TCR sequences from 648 samples and 27 tumor types.

Code availability

The code will be publicly available on GitHub at the time of publication from https://github.com/edavis71/homolig

Data availability

The scRNA-seq data will be available on GEO at the time of publication.

Results

Vaccination induces *de novo* T cell response against mutant KRAS peptides

In this phase I trial we aim to establish the safety and immunogenicity of a pooled mutant KRAS peptide-based vaccine including six of the most common KRAS mutations identified from the TCGA PDAC database: KRAS G12V, G12A, G12R, G12C,

G12D, and G13D. Eligibility for enrollment included molecular confirmation of one of the KRAS mutations included in the vaccine and no evidence of disease on scan after surgical resection and completion of perioperative standard-of-care chemotherapy. Following this enrollment criteria confirmation, the tumor sample from this patient was confirmed to express the G12R mutation. The patient received four doses of the pooled mutant KRAS (mKRAS) peptides delivered in solution with adjuvant polyICLC in combination with immune checkpoint blockade (ICB) as part of the priming phase (Fig. 4-1A). Booster doses were given at weeks 13 and 21 after initial vaccination along with maintenance ICB. Peripheral blood was collected prior to treatment and every four weeks after initial vaccine delivery. Ex vivo peptide restimulation of peripheral blood mononuclear cells (PBMCs) and IFNγ ELISpot confirmed a significant expansion of de novo effector T cells against several mKRAS peptides post-vaccination (Fig. 4-1B). To determine the contribution of the CD4 and CD8 T cell compartment to the observed mKRAS reactivity, we performed flow cytometry analysis of peptide restimulated PBMCs (Fig. 4-1C). We found upregulation of CD69 and CD137 on antigen-experienced CD4 T cells after restimulation with KRAS G12V, G12R, G12A, and G12C peptides (Fig. 4-1D). Interestingly, we observed lower levels of activation in samples restimulated with the KRAS G12D and G13D peptides. The mutation-specific magnitude and kinetics of these responses mirrored those observed in the ELISpot. Robust expression of IFNγ, IL2, and TNF was detected in CD69+ CD4 T cells beginning at cycle 2 (week 4) post-vaccine (Fig. 4-1E). Expansion of polyfunctional cytokine-producing CD4 T cells continued to cycle 6 with increases in double- and triple-cytokine positive T cell populations (Fig. 4-1E). Similarly, an increase in CD69+ CD137+ CD8 T cells was observed G12V, G12R, G12A, and G12C restimulated PBMCs with lower responses to G12D and G13D (Fig. 4-1F). Effector cytokine expression of CD69+ CD8 T cells was also detected, however, was less diverse than the responding CD4 T cell populations (Fig. 4-1G). Analysis of the memory phenotype

markers CD45RO and CD62L at cycle 6 post-treatment demonstrated distinct memory phenotypes between responding CD4 and CD8 T cells (Fig. 4-1H). Activated CD4 T cells expressed Tcm (CD45RO+CD62L+) and Tem (CD45RO+CD62L-) memory markers while the responding CD8 T cell compartment largely clustered into the Teff (CD45RO-CD62L-) and Tem memory phenotypes. These findings demonstrate distinct functional and memory states of vaccine-induced mKRAS-specific T cells and underscore a predominant role of responding CD4 T cells to this vaccine strategy.

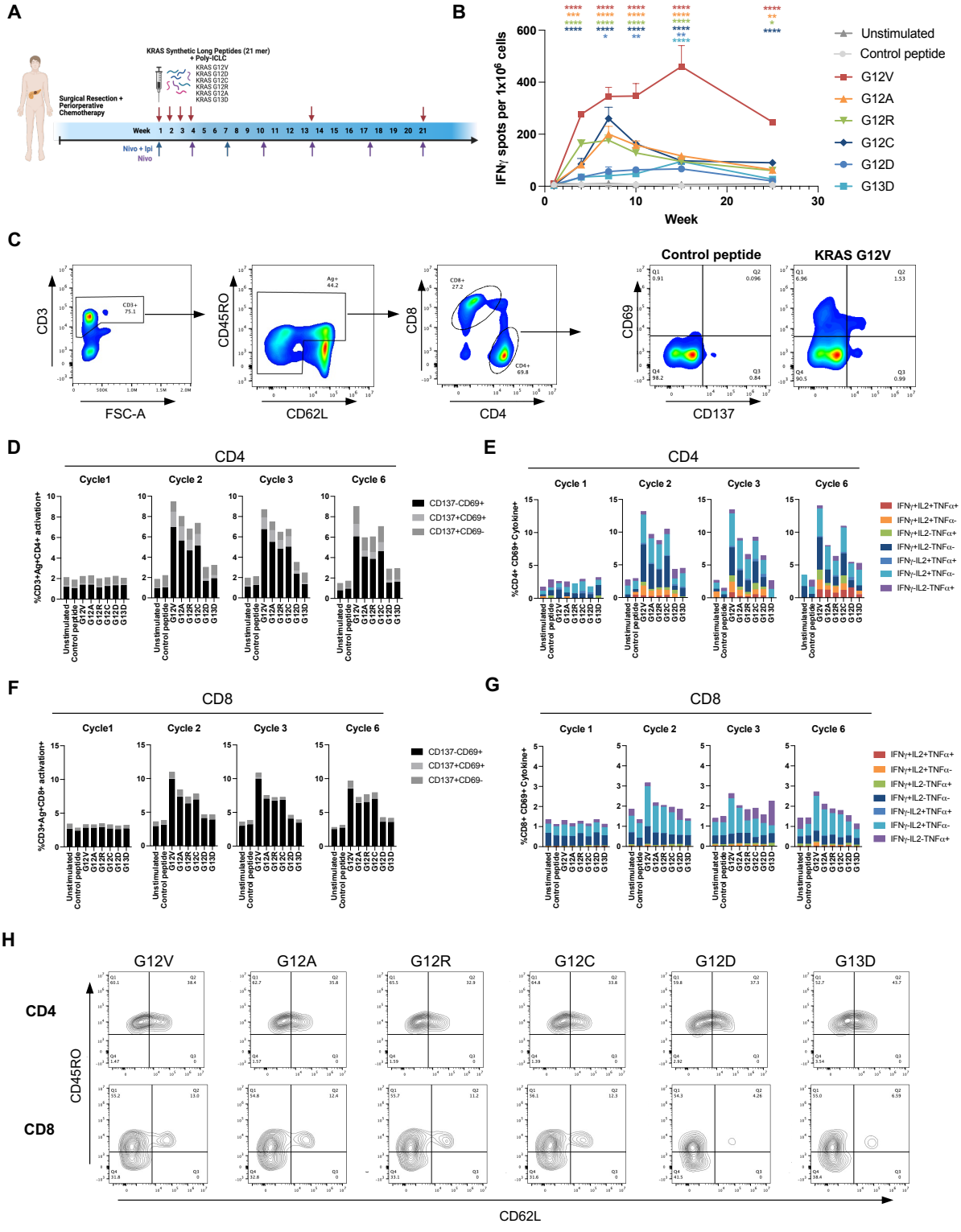
A
KRAS Synthetic Long Peptides (21 mer)
+ Poly-ICLC
KRAS G12V
KRAS G12D
KRAS G12C
KRAS G12R
KRAS G12A
KRAS G13D
Surgical Resection +
Perioperative
Chemotherapy
Week 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21
Nivo + Ipi
Nivo
B
IFNγ spots per 1x10^6 cells
Week
Unstimulated
Control peptide
G12V
G12A
G12R
G12C
G12D
G13D
C
CD3
FSC-A
CD45RO
CD62L
CD8
CD4
Control peptide
KRAS G12V
CD69
CD137
D
CD4
Cycle1
Cycle 2
Cycle 3
Cycle 6
%CD3+Ag+CD4+ activation+
CD137-CD69+
CD137+CD69+
CD137+CD69-
E
CD4
Cycle 1
Cycle 2
Cycle 3
Cycle 6
%CD4+ CD69+ Cytokine+
IFNγ+IL2+TNFα+
IFNγ+IL2+TNFα-
IFNγ+IL2-TNFα+
IFNγ+IL2-TNFα-
IFNγ-IL2+TNFα+
IFNγ-IL2+TNFα-
IFNγ-IL2-TNFα+
F
CD8
Cycle1
Cycle 2
Cycle 3
Cycle 6
%CD3+Ag+CD8+ activation+
CD137-CD69+
CD137+CD69+
CD137+CD69-
G
CD8
Cycle 1
Cycle 2
Cycle 3
Cycle 6
%CD8+ CD69+ Cytokine+
IFNγ+IL2+TNFα+
IFNγ+IL2+TNFα-
IFNγ+IL2-TNFα+
IFNγ+IL2-TNFα-
IFNγ-IL2+TNFα+
IFNγ-IL2+TNFα-
IFNγ-IL2-TNFα+
H
G12V
G12A
G12R
G12C
G12D
G13D
CD4
CD8
CD45RO
CD62L

Figure 4-1

Figure 4-1. mKRAS peptide vaccination induces CD4 and CD8 memory T cell responses. (A) Treatment scheme for KRAS peptide vaccine. Pooled KRAS peptides (0.3mg/peptide) were delivered subcutaneously on days 1,8,15, and 22 (priming) followed by booster doses at weeks 13, 21, 29, 37, and 45. Ipillumimab was administered intravenously at 1mg/kg every 6 weeks for two doses starting on day 1. Nivolumab was administered intravenously at 3mg/kg every 3 weeks in the priming phase, followed by 480mg flat dose during the boost phase. Peripheral blood was collected at weeks 4, 7, 10, 15, and 25. **(B)** IFNγ ELISPOT from PBMCs collected at weeks 4, 7, 10, 15, and 25 post-treatment, restimulated with 2ug/mL KRAS G12V, G12A, G12R, G12C, G12D, or G13D peptide. Unstimulated or control peptide-stimulated PBMCs were used as controls. Two-way ANOVA followed by Dunnett's multiple comparisons test, ns=$p>0.05$, *$p<0.05$, **$p<0.005$, ***$p<0.0005$, ****$p<0.00001$. **(C)** T cell population and activation marker (CD69, CD137) gates from KRAS-peptide restimulated PBMCs. PBMCs were stimulated with 2ug/mL of each KRAS peptide for 48hours. **(D)** Activation and E. cytokine profile of CD4 T cells in peptide-restimulated PBMCs **(F)**. Activation and G. cytokine profile of CD8 T cells in peptide restimulated PBMCs profile of responding CD8 T cells **(H)** CD45RO and CD62L memory phenotype of peptide-specific CD69+ CD4 and CD8 T cells.

Single-cell T cell responses to mKRAS peptide vaccination in unstimulated PBMCs

To assess the cellular immune response to mKRAS peptide vaccination, we collected unstimulated peripheral blood prior to vaccination (C1, baseline) and on weeks 4 (C2) and 25 (C6) after vaccination and performed 5' single-cell RNA-Seq combined with parallel repertoire analysis. Dimension reduction of 63,991 cells by Uniform Manifold Approximation and Projection (UMAP) separated clusters of cells into myeloid and lymphoid compartments (Fig. 4-2A).

Focusing on the T cell compartment resulted in 22,614 cells. UMAP analysis resolved T cell subtypes and functional states across timepoints into 10 clusters. We identified 9 T cell subtypes, including naïve CD4+ T cells (*CCR7, LEF1, SELL*), naïve CD8+ T cells (*CCR7, LEF1, SELL*), central memory T cells (CD4/CD8 TCM, *TNFRSF4*), regulatory T cells (Treg, *FOXP3, IL2RA, and CTLA4*), effector memory (TEM, *CCL5, CST7*), terminal effector memory T cells (TEMRA, *PRF1, NKG7*), MAIT (*KLRB1, SLC4A10, TRAV1-2*), activated T cells (*MKI67* int), and proliferating (*MKI67* high) (Fig. 4-2B).

We next quantified changes in the cell type proportions across vaccine timepoints. While some T cell populations remained consistent (ie. MAIT), others increased significantly in later time points. Activated and cycling T cells were most abundant in cycle 2. Notably, we observed an expansion of memory (CD4/CD8 TCM) T cells in cycle 6 post-vaccination (Fig. 4-2C). This result establishes that the vaccine-induced memory T cells identified by flow cytometry are recapitulated in scRNA-seq data of unstimulated peripheral blood.

Homolig identifies patient TCRs post-vaccine that cluster with known mKRAS-specific TCRs

Previous work has demonstrated that receptors sharing high general sequence similarity based on edit distance are associated with common antigen binding(Dash et al. 2017; Glanville et al. 2017). It is also known that T cell receptors (TCRs) with distinct sequences are capable of recognizing the same self-antigen due to biochemically similar amino acid motifs and influence from both the α and β chain, which existing methods are not poised to identify(Derré et al. 2008). Single-cell TCR-sequencing data provide unprecedented measurements of the clonal identity of T cells by capturing the sequences of both the α and β chains of individual T cells. We hypothesized that comparing single-cell TCRs based on the physicochemical features of amino acids would group biologically similar TCRs into clusters based on shared antigen specificity. We reasoned that including previously identified antigen-specific TCR sequences to compare with TCRs of unknown specificity could be used to rapidly prioritize TCR clusters specific for an antigen of interest, such as mKRAS.

To analyze antigen-specific TCR repertoires at single-cell resolution and obtain a quantitative measure of physiochemical similarity between TCRs, we developed Homolig, a novel algorithm to compute the distance between TCRs (Fig. 4-2D). Each TCR is mapped to the amino acid sequences of the complementarity-determining region (CDR) loops present on the α and β chains that govern peptide:MHC target recognition (CDR1, CDR2, CDR2.5, and CDR3). The distance between two TCRs is computed by performing pairwise sequence alignment and scoring each CDR using a physiochemical substitution matrix derived from physiochemical indices(Kawashima and Kanehisa 2000). Germline encoded CDRs derived from the variable gene (CDR1, CDR2, and CDR2.5) are down-weighted to allow for equal contribution to the CDR3. The resulting pairwise distance matrix is used to visualize the TCR repertoire in two dimensions, where TCRs with a high degree of physiochemical similarity are close to

each other. Unsupervised clustering methods are used to identify groups of similar TCRs and these clusters can be considered physiochemical clonotypes.

We sought to apply our approach to identify and isolate TCRs targeting mKRAS epitopes induced by vaccination. Using paired single-cell α, β TCR sequencing, we compared patient TCRs with a single α and β chain to TCRs with known specificity to pathogenic or mKRAS epitopes[300–302]. 176 out of 22,614 patient TCRs clustered with TCRs that have known reactivity to mKRAS, indicating that 0.77% of peripheral T cells may have putative KRAS-specificity (Fig. 4-2E). To investigate the ability of Homolig to correctly identify mKRAS-specific T cells, we selected 9 TCRs from distinct mKRAS clusters and diverse cell states for functional validation (Fig. 4-2F). TCRs were selected based on relative distance to the known mKRAS TCR within each cluster, established HLA restriction, and presence after vaccination. We found that the majority of putative mKRAS-specific TCRs were observed only once in the dataset or were present in both C2 and C6 post-vaccine.

Isolation of TCRs reactive to mutant KRAS vaccine peptides

To test the reactivity of the Homolig-identified α and β chains against mKRAS antigens, we used CRISPR–Cas12a-based genome editing to introduce a recombinant TCRα and TCRβ chain sequence at the endogenous T cell receptor α and β constant (*TRAC/TRBC*) locus of healthy donor T cells (Fig. 4-3A). All TCR constructs had a successful knock-in rate of approximately 30% as measured by extracellular expression of the murine TCRβ constant region and NGFR tag included in the recombinant TCRα, β (Fig 4-3B). We then performed T cell co-cultures with autologous monocyte-derived dendritic cells (moDCs) pulsed with the pooled mutant KRAS peptides. Several TCRs showed a minor, but not significant, increase in reactivities against peptide pools by IFNγ ELISPOT (Fig 4-3C and D). However, TCR8 showed a significant increase in IFNγ production compared to co-culture with moDCs pulsed with the control peptide

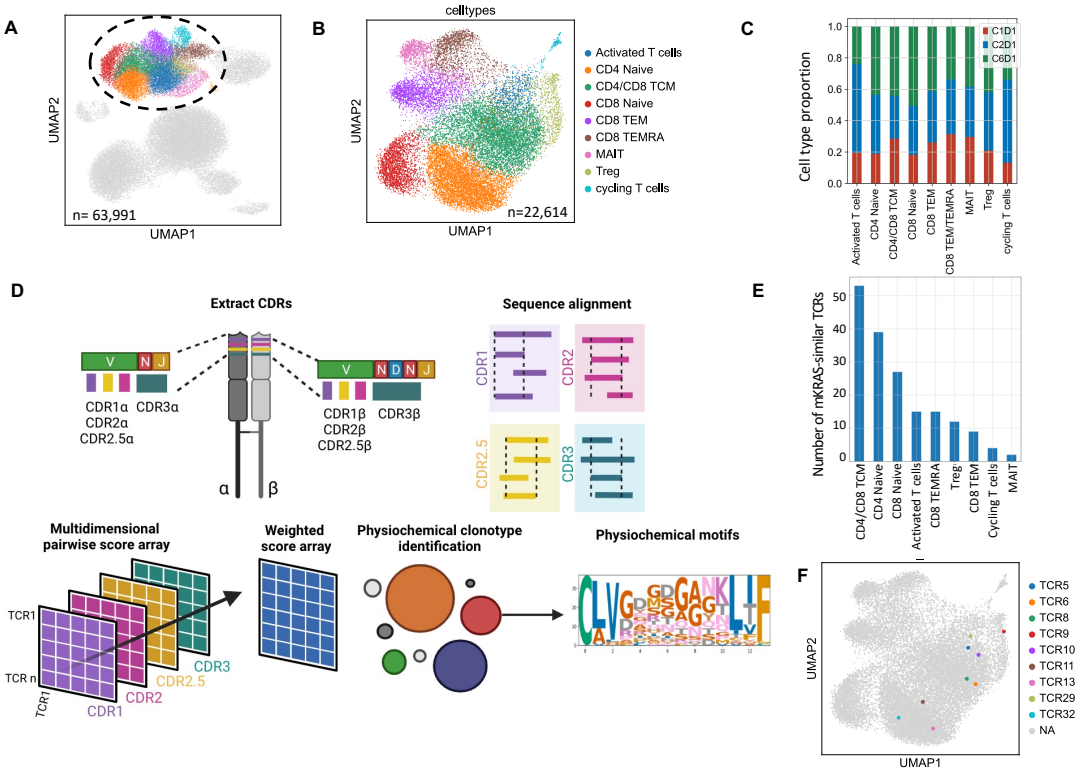

Figure 4-2. **(A)** UMAP of longitudinal scRNA-seq of unstimulated PBMCs. T cell populations are colored by leiden cluster. **(B)** UMAP plot of the T cell types present in peripheral blood samples from 3 vaccine time points. **(C)** Proportions of cell types in each cluster by time point. **(D)** Algorithm overview of Homolig for de novo prediction of mKRAS-specific TCRs from single-cell data. First, germline CDR sequences are extracted using the variable gene. For a set of CDR sequences, a pairwise alignment is performed and scored. CDR1, CDR2, CDR2.5, and CDR3 are scored separately for each TCR pair and then aggregated into a weighted score array to be used in downstream clustering analysis and motif detection. **(E)** Single-cell TCR sequencing reveals T cell clones present after vaccination that cluster with known mKRAS-specific TCRs. 176 patient TCRs have physiochemical similarity to TCRs that recognize mKRAS epitopes. These putative mKRAS TCRs span multiple transcriptional phenotypes but are predominantly memory T cells. **(F)** UMAP plot of T cells selected for validation. 9 prioritized patient TCRs are highlighted.

(Fig 4-3D). Further co-culture of TCR8 with CD3- antigen-presenting cells isolated from PBMCs and pulsed individually with each mutant KRAS peptide resulted in significant activation and upregulation of CD137 and CD25 after recognition of the KRAS G12V peptide (Fig 4-3E and F). Interestingly, a weaker, but significant response over APCs pulsed with control peptide was observed against the KRAS G12C peptide. This data demonstrates the cross-reactive potential of KRAS-reactive TCRs, which has recently been described[302].

Detection of putative mKRAS-specific TCRs within mutant KRAS tumors

A recent publication identified neoantigen-specific TCRs from peripheral blood that were more frequent in the tumor[303]. To test whether putative mKRAS TCR clusters could be detected within tumors bearing KRAS mutations, we applied Homolig to the TCR repertoire in The Cancer Genome Atlas (TCGA)[304] (Fig. 4-4A and B). We compared tumor-infiltrating TCRα sequences from diverse tumor types to known mKRAS-specific TCRs. α chain TCRs clustered into distinct multi-cluster groups based on TRAV gene usage (Fig. 4-4C). 16 clusters from mutant KRAS tumors grouped with mKRAS-specific TCRs, demonstrating that putative mKRAS-specific TCRs can infiltrate tumors. This analysis also allowed us to assess whether putative mKRAS TCR clusters were associated with specific KRAS mutations. A chi-squared test was performed to compare the proportion of TCRs from each cluster by KRAS mutation. Notably, several clusters that contained a known mKRAS TCR were associated with tumors bearing the same mutation (Fig. 4-4D). For example, in comparison of α chain TCRs, cluster 47 contained a TCR specific to G12D and is significantly associated with G12D tumors (Fig. 4-4D). Given that this cluster is only enriched in G12D tumors, this suggests that these T cells are reactive to KRAS G12D epitopes. Hierarchical clustering of the resulting p-value matrix grouped most G12

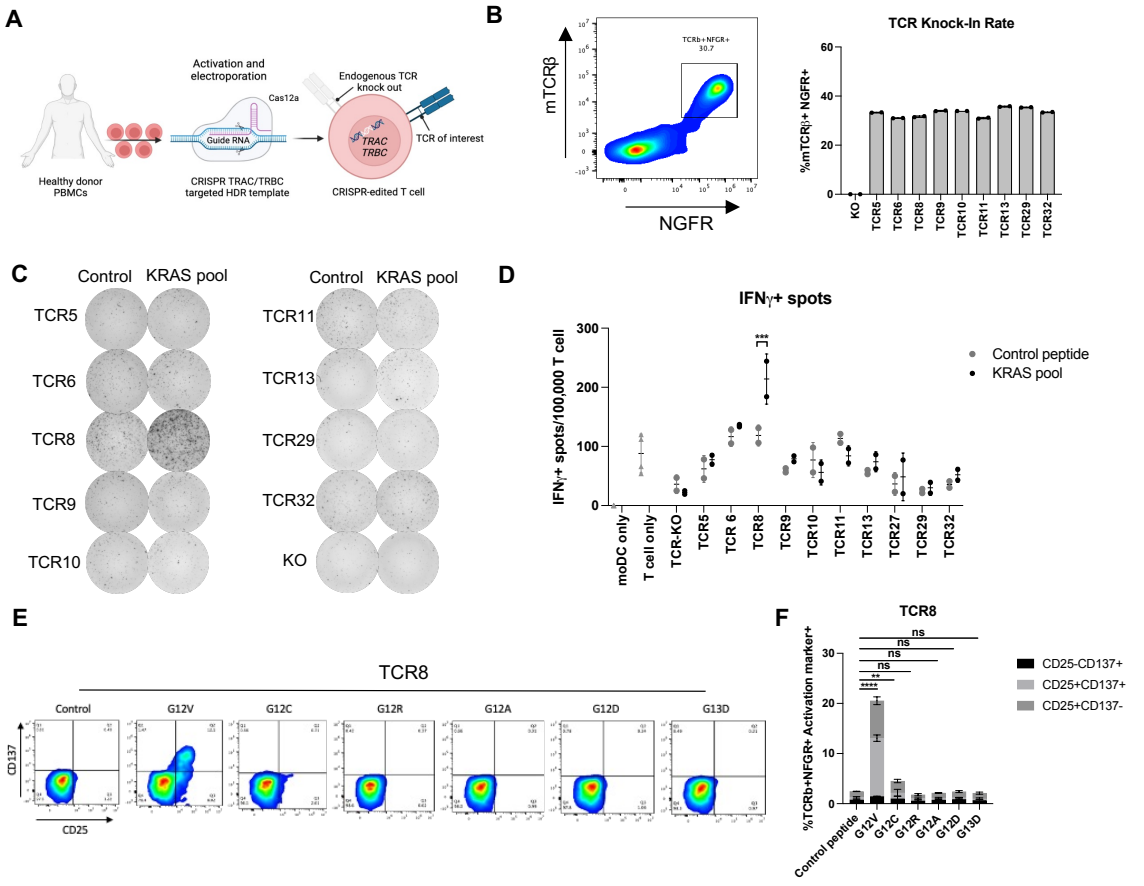

Figure 4-3. Functional validation of predicted mKRAS-specific TCRs. (A) Schematic of CRISPR-based TCR knock-in strategy. **(B)** Quantification of recombinant TCR knock-in into healthy donor T cells by expression of murine TCRβ and NGFR coexpression on healthy donor T cell populations. **(C)** IFNγ production measured by ELISPOT of recombinant TCR expressing cells cocultured with autologous moDCs pulsed with 2ug/mL control peptide or pooled KRAS G12V, G12D, G12C, G12R, G12A, G13D peptides. **(D)** Quantification of IFNγ ELISPOT per recombinant TCR of interest. Two-way ANOVA followed by Sidak's multiple comparisons test, ns=p>0.05, *p<0.05, **p<0.005, ***p<0.0005, ****p<0.00001. **(E)** CD137 and CD25 activation marker expression on TCR8 expressing cells co-cultured with CD3- antigen-presenting cells pulsed with 2ug/mL individual KRAS peptides. **(F)** Quantification of CD25 and CD137 upregulation of TCR8 in response to individual KRAS peptide recognition. Two-way ANOVA followed by Dunnett's multiple comparisons test, ns=p>0.05, *p<0.05, **p<0.005, ***p<0.0005, ****p<0.00001.

mutations together on the far left with lower significance levels. These results indicate that Homolig can be applied broadly to single-chain data of tumor-infiltrating T cells from TCGA and suggest that putative mKRAS TCR clusters which can be detected in the blood after vaccination have the capacity to infiltrate mutant KRAS tumors.

Discussion

Here we have shown that robust CD4 and CD8 memory T cell responses are elicited after vaccination with mutant KRAS peptides in a patient with pancreatic ductal adenocarcinoma (PDAC) and present a method to rapidly identify mutant KRAS-specific T cell receptors (TCRs) directly from single-cell data of peripheral blood.

Understanding the physicochemical principles that drive TCR-antigen recognition is essential to clinical applications. The quality of TCRs recruited during vaccination significantly affects the potency of immune responses (Gallimore et al. 1998). Methods to rapidly identify antigen-specific TCRs based on these features would enable in-depth functional characterization and improve immunotherapy interventions, including the design of vaccine antigens, the choice of optimal TCRs for adoptive T cell therapy, and TCR gene transfer. Here, we employed a novel approach that enabled us to identify and isolate neoantigen-reactive T cells after vaccination directly from unstimulated peripheral blood. Expression of paired α and β TCRs confirmed the predicted antigen reactivity to the G12V vaccine peptide.

A few studies have identified neoantigen-specific T cells in the peripheral blood of individuals with cancer [305–307]. These relied on successive rounds of in vitro stimulation or large-scale screening approaches, laborious techniques that can skew the T cell repertoire, resulting in a potential loss of therapeutically relevant but subdominant clones. Our strategy to identify circulating neoantigen-specific T cells differs in that we use a computational algorithm to predict mKRAS-specific TCRs

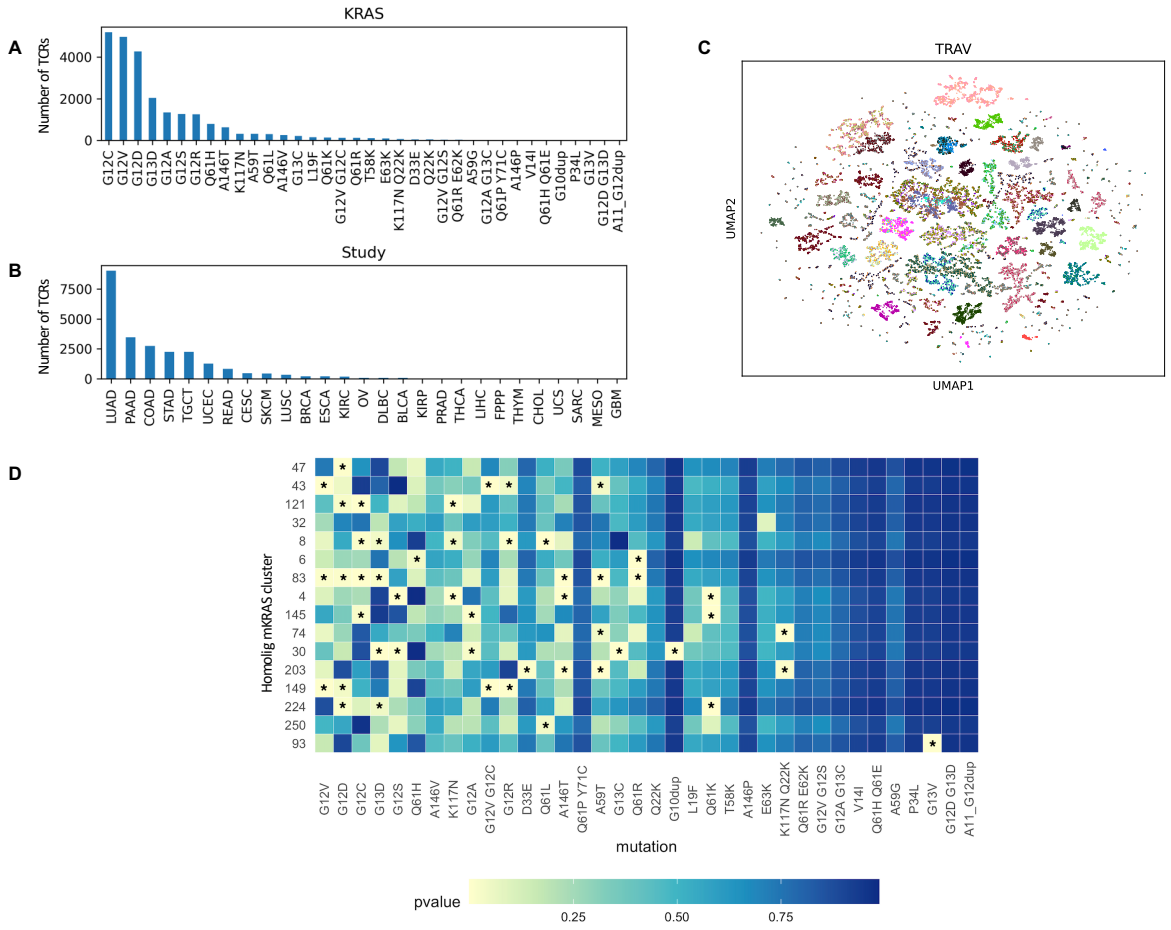

Figure 4-4. Analyses of mutant KRAS tumors from TCGA. (A) Barplot of the number of tumor-infiltrating α chain TCRs by KRAS mutation. **(B)** Barplot of the number of tumor-infiltrating α chain TCRs by TCGA study. **(C)** UMAP plot showing Homolig clustering of tumor-infiltrating α chain TCRs from TCGA colored by TRAV allele. Each dot in the UMAP represents a TCR. **(D)** Chi-squared analysis identifies predicted mKRAS tumor-infiltrating TCR clusters enriched in mutant KRAS tumors based on α chain similarity. P-values < 0.05 are indicated with an asterisk.

solely based on the paired single-cell α, β TCR sequence. Although neoantigen-specific T cells are present at very low frequencies in circulation, which prevents their detection in unstimulated peripheral blood, we demonstrate that Homolig can detect mKRAS-specific TCRs after neoantigen vaccination. In addition to potential applications to improve immunotherapy, this work sets the stage for using the peripheral T cell repertoire to monitor immunotherapy responses, which would greatly benefit patients due to the invasive nature and limited availability of patient tumor specimens.

While our data provide insights into the ability of cancer vaccines to induce T cells specific to oncogenic driver mutations such as those which commonly occur in KRAS, this study focused on therapeutic safety and did not address survival outcomes. Future studies are needed to address the longevity of immunity after vaccination and whether vaccine-induced neoantigen-specific T cells are able to recognize and kill tumor targets in vivo. Ongoing studies related to this project will determine the HLA restriction and cytotoxic potential of predicted mKRAS-specifc TCRs.

Conclusions

Single-cell and spatial molecular profiling technologies and complementary computational analysis pipelines are rapidly advancing as tools for cancer research. The inferences from these technologies rely on the study design, sample processing, and analysis pipelines selected for profiling. Due to the rapid advance of these technologies, many of the computational pipelines that enable interpretation of these data are still being developed. As single-cell data evolve as translational tools, computational methodologies will play a role in driving new discoveries. While powerful, these high-throughput technologies primarily serve as profiling tools to generate new hypotheses about the TME and therapeutic modalities. Therefore, mechanistic bench studies remain an important complement to translate single-cell research into actionable therapeutic targets.

In translational immunotherapy research, the ultimate test of mechanism is that a therapeutic intervention yields the hypothesized immune modulation on the TME within a patients' tumor. While single-cell technologies can be applied to measure these effects, full mechanistic characterization requires time-course profiling that would involve serial sample collection from the same patient, which is unethical and unfeasible to perform. Although monitoring the immune cell repertoire from a patient's peripheral blood is more feasible for time-course studies, single-cell studies comparing the immune cell composition of human tumors and peripheral blood have identified intrinsic differences[40]. Therefore, future studies are needed to provide a more comprehensive comparison between the tumor immune landscape and that of

the periphery to enable the use of single-cell technologies as therapeutic biomarkers. The heterogeneity between patient tumors and the inability to test more than one treatment regimen in a patient further challenge mechanistic single-cell studies in translational research. Single-cell atlas studies pooling clinical trial studies and perturbation studies from pre-clinical models can provide important references to support such human profiling studies. Single-cell profiling of pre-clinical models treated with immunotherapies can point to the cell types and pathways relevant to therapeutic response, while cross-species analysis of human samples treated with the same therapies can reveal which therapeutic responses are conserved. Emerging computational tools to identify shared responses in mice and humans from single-cell datasets can further support model selection for pre-clinical analysis to inform the design of human clinical trials[55, 274] (Fig. 5-1).

Single-cell experiments must be carefully designed to achieve the desired depth of immune characterization and to avoid confounding technical biases with phenotypic covariates. Technology selection should align with the underlying hypothesis of the study. For example, single-cell and spatial transcriptomics is well suited to drive genome-wide discovery across unbiased cell populations, and single-cell proteomics is better suited to studies that aim to profile known molecular targets and cell types. Pathologists can play an important role in selecting regions that capture the appropriate biological region (e.g., tumor-dense regions) and account for tumor heterogeneity. Statistical evaluation of sample size is also an important consideration of study design. Single-cell datasets from large cohorts are important for biomarker discovery and applications of machine learning to predict patient outcomes, particularly to avoid overfitting these models with the large number of molecular features they measure. However, the significant costs of these technologies impose a practical limitation to designing powered cohorts in single-cell atlas studies. Thus, a balance of low-dimensional profiling with proteomics technologies on large cohorts and leveraging

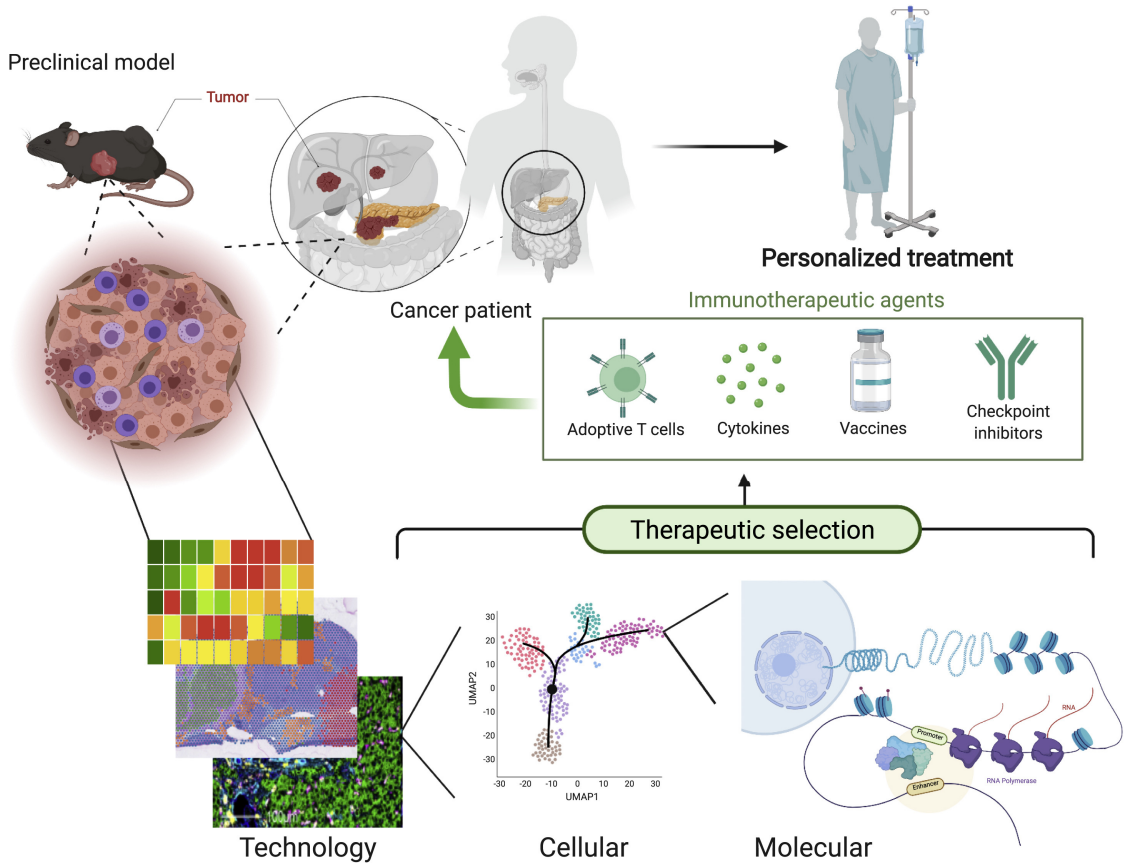

Figure 4-5. Single-cell and spatial technologies have the power to drive discoveries based on the cell types that are commonly affected by immunotherapies in pre-clinical and human tumors. Following identification of the commonalities between models, studies can focus on identifying molecular and cellular markers of response using multi-omics approaches. The combination of different layers of data will drive patient selection for the most adequate therapy, better clinical trial designs, and development of new immunotherapies.

selected samples for higher-dimensional, mechanistic studies is a critical step during experimental design. Computational algorithms for cross-platform data integration can provide an important complement to balance molecular depth with sample size in these mixed designs. Close collaboration between experimental, computational, and statistical scientists can support optimal study design and also prioritize new computational approaches for data analysis tailored to the translational research goals of each study. A multi-disciplinary approach will help the cancer research community to overcome these challenges and use single-cell and spatial platforms to make new discoveries for cancer immunotherapy.

References

1. Marx, V. Method of the Year: spatially resolved transcriptomics. *Nat. Methods* **18,** 9–14 (2021).
2. Sandberg, R. Entering the era of single-cell transcriptomics in biology and medicine. *Nat. Methods* **11,** 22–24 (2014).
3. Zhu, C., Preissl, S. & Ren, B. Single-cell multimodal omics: the power of many. *Nat. Methods* **17,** 11–14 (2020).
4. Navin, N. & Hicks, J. Future medical applications of single-cell sequencing in cancer. *Genome Med* **3,** 31 (2011).
5. Tang, F. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* **6,** 377–382 (2009).
6. Bandura, D. *et al.* Mass cytometry: technique for real time single cell multitarget immunoassay based on inductively coupled plasma time-of-flight mass spectrometry. *Anal. Chem* **81,** 6813–6822 (2009).
7. Buenrostro, J. *et al.* Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523,** 486–490 (2015).
8. Schürch, C. *et al.* Coordinated cellular neighborhoods orchestrate antitumoral immunity at the colorectal cancer invasive front. *Cell* **182,** 1341–1359 (2020).
9. Ribas, A. & Wolchok, J. Cancer immunotherapy using checkpoint blockade. *Science* **359** (2018).
10. Giladi, A. & Amit, I. Single-cell genomics: a stepping stone for future immunology discoveries. *Cell* **172,** 14–21 (2018).
11. Giesen, C. *et al.* Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. *Nat. Methods* **11,** 417–422 (2014).
12. Ståhl, P. *et al.* Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* **353,** 78–82 (2016).
13. Gohil, S., Iorgulescu, J., Braun, D., Keskin, D. & Livak, K. Applying high-dimensional single-cell technologies to the analysis of cancer immunotherapy. *Nat. Rev. Clin. Oncol* (2021).
14. Berglund, E. *et al.* Spatial maps of prostate cancer transcriptomes reveal an unexplored landscape of heterogeneity. *Nat. Commun* **9,** 2419 (2018).
15. Patel, A. *et al.* Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344,** 1396–1401 (2014).

16. Gadalla, R. *et al.* Validation of cytof against flow cytometry for immunological studies and monitoring of human cancer clinical trials. *Front. Oncol* **9,** 415 (2019).

17. Bendall, S. *et al.* Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. *Science* **332,** 687–696 (2011).

18. Hartmann, F. & Bendall, S. Immune monitoring using mass cytometry and related high-dimensional imaging approaches. *Nat. Rev. Rheumatol* **16,** 87–99 (2020).

19. Leipold, M. *et al.* Comparison of CyTOF assays across sites: results of a six-center pilot study. *J. Immunol. Methods* **453,** 37–43 (2018).

20. Sumatoh, H., Teng, K., Cheng, Y. & Newell, E. Optimization of mass cytometry sample cryopreservation after staining. *Cytometry A* **91,** 48–61 (2017).

21. Gubin, M. *et al.* High-Dimensional Analysis Delineates Myeloid and Lymphoid Compartment Remodeling during Successful Immune-Checkpoint Cancer Therapy. *Cell* **175,** 1014–1030 (2018).

22. Krieg, C. *et al.* High-dimensional single-cell analysis predicts response to anti-PD-1 immunotherapy. *Nat. Med* **24,** 144–153 (2018).

23. Subrahmanyam, P. *et al.* Distinct predictive biomarker candidates for response to anti-CTLA-4 and anti-PD-1 immunotherapy in melanoma patients. *J. Immunother. Cancer* **6,** 18 (2018).

24. Wu, A. *et al.* A phase 2 study of allogeneic GM-CSF transfected pancreatic tumor vaccine (GVAX) with ipilimumab as maintenance treatment for metastatic pancreatic cancer. *Clin. Cancer Res* **26,** 5129–5139 (2020).

25. Navin, N. *et al.* Tumour evolution inferred by single-cell sequencing. *Nature* **472,** 90–94 (2011).

26. Avila Cobos, F., Alquicira-Hernandez, J., Powell, J., Mestdagh, P. & Preter, K. Benchmarking of cell type deconvolution pipelines for transcriptomics data. *Nat. Commun* **11,** 5650 (2020).

27. Lim, B., Lin, Y. & Navin, N. Advancing cancer research and medicine with single-cell genomics. *Cancer Cell* **37,** 456–470 (2020).

28. Trapnell, C. Defining cell types and states with single-cell genomics. *Genome Res* **25,** 1491–1498 (2015).

29. Ramsköld, D. *et al.* Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat. Biotechnol* **30,** 777–782 (2012).

30. Islam, S. *et al.* Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat. Methods* **11,** 163–166 (2014).

31. Xin, Y. *et al.* Use of the Fluidigm C1 platform for RNA sequencing of single mouse pancreatic islet cells. *Proc. Natl. Acad. Sci. USA* **113,** 3293–3298 (2016).

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

33. Macosko, E. *et al.* Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161,** 1202–1214 (2015).

34. Zheng, G. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat. Commun* **8,** 14049 (2017).

35. Lee, J., Hyeon, D. & Hwang, D. Single-cell multiomics: technologies and data analysis methods. *Exp. Mol. Med* **52,** 1428–1442 (2020).

36. Stoeckius, M. *et al.* Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods* **14,** 865–868 (2017).

37. Goldstein, L. *et al.* Massively parallel single-cell B-cell receptor sequencing enables rapid discovery of diverse antigen-reactive antibodies. *Commun. Biol* **2,** 304 (2019).

38. Tu, A. *et al.* TCR sequencing paired with massively parallel 3' RNA-seq reveals clonotypic T cell signatures. *Nat. Immunol* **20,** 1692–1699 (2019).

39. See, P., Lum, J., Chen, J. & F. Ginhoux A single-cell sequencing guide for immunologists. *Front Immunol* **9,** 2425 (2018).

40. Azizi, E. *et al.* Single-Cell Map of Diverse Immune Phenotypes in the Breast Tumor Microenvironment. *Cell* **174,** 1293–1308 (2018).

41. Bernard, V. *et al.* Single-Cell Transcriptomics of Pancreatic Cancer Precursors Demonstrates Epithelial and Microenvironmental Heterogeneity as an Early Event in Neoplastic Progression. *Clin. Cancer Res* **25,** 2194–2205 (2019).

42. Davidson, S. *et al.* Single-Cell RNA Sequencing Reveals a Dynamic Stromal Niche That Supports Tumor Growth. *Cell Rep* **31,** 107628 (2020).

43. Guo, X. *et al.* Global characterization of T cells in non-small-cell lung cancer by single-cell sequencing. *Nat. Med* **24,** 978–985 (2018).

44. Li, H. *et al.* Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. *Nat. Genet* **49,** 708–718 (2017).

45. Ma, L. *et al.* Tumor cell biodiversity drives microenvironmental reprogramming in liver cancer. *Cancer Cell* **36,** 418–430 6 (2019).

46. Peng, J. *et al.* Single-cell RNA-seq highlights intra-tumoral heterogeneity and malignant progression in pancreatic ductal adenocarcinoma. *Cell Res* **29,** 725–738 (2019).

47. Puram, S. *et al.* Single-Cell Transcriptomic Analysis of Primary and Metastatic Tumor Ecosystems in Head and Neck Cancer. *Cell* **171,** 1611–24 (2017).

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

49. Zhao, J. *et al.* Single-cell RNA sequencing reveals the heterogeneity of liver-resident immune cells in human. *Cell Discov* **6,** 22 (2020).

50. Savas, P. *et al.* Single-cell profiling of breast cancer T cells reveals a tissue-resident memory subset associated with improved prognosis. *Nat. Med* **24,** 986–993 (2018).

51. Yost, K. *et al.* Clonal replacement of tumor-specific T cells following PD-1 blockade. *Nat. Med* **25,** 1251–1259 (2019).

52. Zheng, C. *et al.* Landscape of Infiltrating T Cells in Liver Cancer Revealed by Single-Cell Sequencing. *Cell* **169,** 1342–1356 16 (2017).

53. Jerby-Arnon, L. *et al.* A cancer cell program promotes T cell exclusion and resistance to checkpoint blockade. *Cell* **175,** 984–997 24 (2018).

54. Sade-Feldman, M. *et al.* Defining T Cell States Associated with Response to Checkpoint Immunotherapy in Melanoma. *Cell* **175,** 998–1013 20 (2018).

55. Davis-Marcisak, E. F. *et al.* Transfer learning between preclinical models and human tumors identifies a conserved NK cell activation signature in anti-CTLA-4 responsive tumors. *Genome Medicine* **13** (2021).

56. Ji, A. *et al.* Multimodal analysis of composition and spatial architecture in human squamous cell carcinoma. *Cell* **182,** 497–514 22 (2020).

57. Ramos-Vara, J. & Miller, M. When tissue antigens and antibodies get along: revisiting the technical aspects of immunohistochemistry–the red, brown, and blue technique. *Vet. Pathol* **51,** 42–87 (2014).

58. Gorris, M. *et al.* Eight-Color Multiplex Immunohistochemistry for Simultaneous Detection of Multiple Immune Checkpoint Molecules within the Tumor Microenvironment. *J. Immunol* **200,** 347–354 (2018).

59. Pulsawatdi, A. *et al.* A robust multiplex immunofluorescence and digital pathology workflow for the characterisation of the tumour immune microenvironment. *Mol. Oncol* **14,** 2384–2402 (2020).

60. Tsujikawa, T. *et al.* Quantitative Multiplex Immunohistochemistry Reveals Myeloid-Inflamed Tumor-Immune Complexity Associated with Poor Prognosis. io. *Cell Rep* **19,** 203–217 (2017).

61. Angelo, M. *et al.* Multiplexed ion beam imaging of human breast tumors. *Nat. Med* **20,** 436–442 (2014).

62. Gerdes, M. *et al.* Highly multiplexed single-cell analysis of formalin-fixed, paraffin-embedded cancer tissue. *Proc. Natl. Acad. Sci. USA* **110,** 11982–11987 (2013).

63. Goltsev, Y. *et al.* Deep Profiling of Mouse Splenic Architecture with CODEX Multiplexed Imaging. *Cell* **174,** 968–981 (2018).

64. Lin, J.-R., Fallahi-Sichani, M. & Sorger, P. Highly multiplexed imaging of single cells using a high-throughput cyclic immunofluorescence method. *Nat. Commun* **6,** 8390 (2015).

65. Yan, Y. *et al.* Understanding heterogeneous tumor microenvironment in metastatic melanoma. *PLoS One* **14** (2019).

66. Jackson, H. *et al.* The single-cell pathology landscape of breast cancer. *Nature* **578,** 615–620 (2020).

67. Keren, L. *et al.* A Structured Tumor-Immune Microenvironment in Triple Negative Breast Cancer Revealed by Multiplexed Ion Beam Imaging. *Cell* **174,** 1373–1387 (2018).

68. Aoki, T. *et al.* Single-Cell Transcriptome Analysis Reveals Disease-Defining T-cell Subsets in the Tumor Microenvironment of Classic Hodgkin Lymphoma. *Cancer Discov* **10,** 406–421 (2020).

69. Ho, W. *et al.* Integrated immunological analysis of a successful conversion of locally advanced hepatocellular carcinoma to resectability with neoadjuvant therapy. *J. Immunother. Cancer* (2020).

70. Xiang, H. *et al.* Cancer-Associated Fibroblasts Promote Immunosuppression by Inducing ROS-Generating Monocytic MDSCs in Lung Squamous Cell Carcinoma. *Cancer Immunol Res* **8,** 436–450 (2020).

71. Maniatis, S., Petrescu, J. & Phatnani, H. Spatially resolved transcriptomics and its applications in cancer. *Curr. Opin. Genet. Dev* **66,** 70–77 (2021).

72. Rodriques, S. *et al.* Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science* **363,** 1463–1467 (2019).

73. Moncada, R. *et al.* Integrating microarray-based spatial transcriptomics and single-cell RNA-seq reveals tissue architecture in pancreatic ductal adenocarcinomas. *Nat. Biotechnol* **38,** 333–342 (2020).

74. Lubeck, E., Coskun, A., Zhiyentayev, T., Ahmad, M. & Cai, L. Single-cell in situ RNA profiling by sequential hybridization. *Nat. Methods* **11,** 360–361 (2014).

75. Eng, C.-H. *et al.* Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH. *Nature* **568,** 235–239 (2019).

76. Xia, C., Fan, J., Emanuel, G., Hao, J. & Zhuang, X. Spatial transcriptome profiling by MERFISH reveals subcellular RNA compartmentalization and cell cycle-dependent gene expression. *Proc. Natl. Acad. Sci. USA* **116,** 19490–19499 (2019).

77. Liu, Y. *et al.* High-spatial-resolution multi-omics sequencing via deterministic barcoding in tissue. *Cell* **183,** 1665–1681 18 (2020).

78. Lafzi, A., Moutinho, C., Picelli, S. & Heyn, H. Tutorial: guidelines for the experimental design of single-cell RNA sequencing studies. *Nat. Protoc* **13,** 2742–2757 (2018).

79. Denisenko, E. *et al.* Systematic assessment of tissue dissociation and storage biases in single-cell and single-nucleus RNA-seq workflows. *Genome Biol* **21,** 130 (2020).

80. Slyper, M. *et al.* A single-cell and single-nucleus RNA-Seq toolbox for fresh and frozen human tumors. *Nat. Med* **26,** 792–802 (2020).

81. Gracia Villacampa, E. *et al.* Genome-wide spatial expression profiling in formalin-fixed tissues. *Cell Genomics* **1,** 100065 (2021).

82. Browaeys, R., Saelens, W. & Saeys, Y. NicheNet: modeling intercellular communication by linking ligands to target genes. *Nat. Methods* **17,** 159–162 (2020).

83. Cherry, C. *et al.* Computational reconstruction of the signalling networks surrounding implanted biomaterials from single-cell transcriptomics. *Nat. Biomed. Eng.* **5,** 1228–1238 (2021).

84. Efremova, M., Vento-Tormo, M., Teichmann, S. & Vento-Tormo, R. CellPhoneDB: inferring cell-cell communication from combined expression of multi-subunit ligand-receptor complexes. *Nat. Protoc* **15,** 1484–1506 (2020).

85. Dries, R., Zhu, Q. & Dong, R. Giotto: a toolbox for integrative analysis and visualization of spatial expression data. *Genome Biol* **22,** 78 (2021).

86. Luecken, M. & Theis, F. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol* **15,** 8746 (2019).

87. Gassen, S. *et al.* FlowSOM: Using self-organizing maps for visualization and interpretation of cytometry data. et. *Cytometry A* **87,** 636–645 (2015).

88. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol* **36,** 411–420 (2018).

89. Cao, J. *et al.* The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566,** 496–502 (2019).

90. Hao, Y., Hao, S. & Andersen-Nissen, E. Integrated analysis of multimodal single-cell data. *Cell* **184,** 3573–3587 (2021).

91. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv* (2018).

92. Qiu, X. *et al.* Reversed graph embedding resolves complex single-cell trajectories. *Nat Methods* **14,** 979–982 (2017).

93. Satija, R., Farrell, J., Gennert, D., Schier, A. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol* **33,** 495–502 (2015).

94. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177,** 1888–1902 (2019).

95. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol* **32,** 381–6 (2014).

96. Wolf, F., Angerer, P. & Theis, F. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19,** 15 (2018).

97. Tekman, M. *et al.* A single-cell RNA-sequencing training and analysis suite using the Galaxy framework. *GigaScience* **9.** giaa102, (2020).

98. Reich, M. *et al.* The genepattern notebook environment. *Cell Syst* **5,** 149–151 (2017).

99. Megill, C. *et al.* Cellxgene: a performant, scalable exploration platform for high dimensional sparse matrices. *bioRxiv* (2021).

100. Nowicka, M. *et al.* CyTOF workflow: differential discovery in high-throughput high-dimensional cytometry datasets. *F1000Res,* 748 (2017).

101. Gao, M. *et al.* Comparison of high-throughput single-cell RNA sequencing data processing pipelines. *Brief. Bioinformatics* **22** (2020).

102. Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol* **20,** 296 (2019).

103. Hou, W., Ji, Z., Ji, H. & Hicks, S. A systematic evaluation of single-cell RNA-sequencing imputation methods. *Genome Biol.* **21,** 218 (2020).

104. Ilicic, T. *et al.* Classification of low quality cells from single-cell RNA-seq data. *Genome Biol* **17,** 29 (2016).

105. Caicedo, J. *et al.* Data-analysis strategies for image-based cell profiling. *Nat. Methods* **14,** 849–863 (2017).

106. Elosua-Bayes, M., Nieto, P., Mereu, E., Gut, I. & Heyn, H. SPOTlight: seeded NMF regression to deconvolute spatial transcriptomics spots with single-cell transcriptomes. *Nucleic Acids Res* **49,** 50 (2021).

107. Leek, J. *et al.* Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet* **11,** 733–739 (2010).

108. Hicks, S., Townes, F., Teng, M. & Irizarry, R. Missing data and technical variability in single-cell RNA-sequencing experiments. *Biostatistics* **19,** 562–578 (2018).

109. Bullard, J., Purdom, E., Hansen, K. & Dudoit, S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* **11,** 94 (2010).

110. Tran, H. *et al.* A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biol* **21,** 12 (2020).

111. Cleary, B., Cong, L., Cheung, A., Lander, E. & Regev, A. Efficient generation of transcriptomic profiles by random composite measurements. *Cell* **171,** 1424–1436 (2017).

112. Stein-O'Brien, G. *et al.* Decomposing Cell Identity for Transfer Learning across Cellular Measurements, Platforms, Tissues, and Species. *Cell Syst* **8,** 395–411 8 (2019).

113. Wagner, A., Regev, A. & Yosef, N. Revealing the vectors of cellular identity with single-cell genomics. *Nat. Biotechnol* **34,** 1145–1160 (2016).

114. Becht, E. *et al.* Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol* **37,** 38–44 (2018).

115. Maaten, L. & Hinton, G. Visualizing Data using t-SNE. *Journal of Machine Learning Research* (2008).

116. Moon, K. *et al.* Visualizing structure and transitions in high-dimensional biological data. *Nat. Biotechnol* **37,** 1482–1492 (2019).

117. Blondel, V., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech,* 10008 (2008).

118. Xu, C. & Su, Z. Identification of cell types from single-cell transcriptomes using a novel clustering method. *Bioinformatics* **31,** 1974–1980 (2015).

119. Mohammadi, S., Davila-Velderrain, J. & Kellis, M. A multiresolution framework to characterize single-cell state landscapes. *Nat. Commun* **11,** 5399 (2020).

120. Way, G., Zietz, M., Rubinetti, V., Himmelstein, D. & Greene, C. Compressing gene expression data using multiple latent space dimensionalities learns complementary biological representations. *Genome Biol* **21,** 109 (2020).

121. Huang, Q., Liu, Y., Du, Y. & Garmire, L. Evaluation of Cell Type Annotation R Packages on Single-cell RNA-seq Data. *Genomics Proteomics Bioinformatics* **19(2)** (2020).

122. Rozenblatt-Rosen, O. *et al.* The Human Tumor Atlas Network: Charting Tumor Transitions across Space and Time at Single-Cell Resolution. *Cell* **181,** 236–249 (2020).

123. Fan, J. *et al.* Linking transcriptional and genetic tumor heterogeneity through allele analysis of single-cell RNA-seq data. *Genome Res* **28,** 1217–1227 (2018).

124. Gao, R. *et al.* Delineating copy number and clonal substructure in human tumors from single-cell transcriptomes. *Nat. Biotechnol* **39** (2021).

125. Tickle, T., Tirosh, I., Georgescu, C., Brown, M. & Haas, B. *inferCNV of the Trinity CTAT Project* 2019.

126. Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S. & Vert, J.-P. A general and flexible method for signal extraction from single-cell RNA-seq data. *Nat Commun* **9** (2018).

127. Crowell, H. *et al.* An R-based reproducible and user-friendly preprocessing pipeline for CyTOF data. *F1000Res* **9** (2020).

128. Irizarry, R., Wang, C., Zhou, Y. & Speed, T. Gene set enrichment analysis made simple. *Stat Methods Med Res* **18,** 565–575 (2009).

129. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* **102,** 15545–15550 (2005).

130. Davis-Marcisak, E. *et al.* Differential Variation Analysis Enables Detection of Tumor Heterogeneity Using Single-Cell RNA-Sequencing Data. *Cancer Res* **79,** 5102–5112 (2019).

131. Fan, J. *et al.* Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. *Nat. Methods* **13,** 241–244 (2016).

132. Stein-O'Brien, G. *et al.* Enter the Matrix: Factorization Uncovers Knowledge from Omics. *Trends Genet* **34,** 790–805 (2018).

133. Zhu, X., Ching, T., Pan, X., Weissman, S. & Garmire, L. Detecting heterogeneity in single-cell RNA-Seq data by non-negative matrix factorization. *PeerJ* **5,** 2888 (2017).

134. Burkhardt, D. *et al.* Quantifying the effect of experimental perturbations at single-cell resolution. *Nat. Biotechnol* (2021).

135. Ji, Z. & Ji, H. TSCAN: Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. *Nucleic Acids Res* **44,** 117 (2016).

136. Setty, M. *et al.* Wishbone identifies bifurcating developmental trajectories from single-cell data. *Nat. Biotechnol* **34,** 637–645 (2016).

137. Shin, J. *et al.* Single-Cell RNA-Seq with Waterfall Reveals Molecular Cascades underlying Adult Neurogenesis. *Cell Stem Cell* **17,** 360–372 (2015).

138. Saelens, W., Cannoodt, R., Todorov, H. & Saeys, Y. A comparison of single-cell trajectory inference methods. *Nat. Biotechnol* **37,** 547–554 (2019).

139. Manno, G. *et al.* RNA velocity of single cells. *Nature* **560,** 494–498 (2018).

140. Bergen, V., Lange, M., Peidli, S., Wolf, F. & Theis, F. Generalizing RNA velocity to transient cell states through dynamical modeling. *Nat. Biotechnol* **38,** 1408–1414 (2020).

141. Soto, L. *et al.* in *scMomentum: Inference of Cell-Type-Specific Regulatory Networks and Energy Landscapes* (BioRxiv, 2020).

142. Gorin, G., Svensson, V. & Pachter, L. Protein velocity and acceleration from single-cell multiomics experiments. *Genome Biol* **21,** 39 (2020).

143. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9,** 559 (2008).

144. Huynh-Thu, V. & Sanguinetti, G. Combining tree-based and dynamical systems for the inference of gene regulatory networks. *Bioinformatics* **31,** 1614–1622 (2015).

145. Huynh-Thu, V., Irrthum, A., Wehenkel, L. & Geurts, P. Inferring regulatory networks from expression data using tree-based methods. *PLoS One* **28** (2010).

146. Chan, T., Stumpf, M. & Babtie, A. Gene Regulatory Network Inference from Single-Cell Data Using Multivariate Information Measures. *Cell Syst* **5,** 251–267 (2017).

147. Papili Gao, N., Ud-Dean, S., Gandrillon, O. & Gunawan, R. SINCERITIES: inferring gene regulatory networks from time-stamped single cell transcriptional expression profiles. *Bioinformatics* **34,** 258–266 (2018).

148. Deshpande, A., Chu, L., Stewart, R. & Gitter, A. Network inference with Granger causality ensembles on single-cell transcriptomics. *Cell Reports* **38,** 110333 (2022).

149. Matsumoto, H. *et al.* SCODE: an efficient regulatory network inference algorithm from single-cell RNA-Seq during differentiation. *Bioinformatics* **33,** 2314–2321 (2017).

150. Specht, A. & Li, J. LEAP: constructing gene co-expression networks for single-cell RNA-sequencing data using pseudotime ordering. *Bioinformatics* **33,** 764–766 (2017).

151. Wculek, S. *et al.* Dendritic cells in cancer immunology and immunotherapy. *Nat. Rev. Immunol* **20,** 7–24 (2020).

152. Kumar, M. *et al.* Analysis of Single-Cell RNA-Seq Identifies Cell-Cell Communication Associated with Tumor Characteristics. *Cell Rep* **25,** 1458–1468 4 (2018).

153. Li, Y. *et al.* Elucidation of Biological Networks across Complex Diseases Using Single-Cell Omics. *Trends Genet* (2020).

154. Yuan, Y. & Bar-Joseph, Z. GCNG: graph convolutional networks for inferring gene interaction from spatial transcriptomics data. *Genome Biol* **21,** 300 (2020).

155. Cadot, S. *et al.* Longitudinal CITE-Seq profiling of chronic lymphocytic leukemia during ibrutinib treatment: evolution of leukemic and immune cells at relapse. *Biomark Res* **8,** 72 (2020).

156. Stubbington, M. *et al.* T cell fate and clonality inference from single-cell transcriptomes. *Nat. Methods* **13,** 329–332 (2016).

157. Lindeman, I. *et al.* BraCeR: B-cell-receptor reconstruction and clonality inference from single-cell RNA-seq. *Nat. Methods* **15,** 563–565 (2018).

158. Eltahla, A. *et al.* Linking the T cell receptor to the single cell transcriptome in antigen-specific human T cells. *Immunol. Cell Biol* **94,** 604–611 (2016).

159. Rizzetto, S. *et al.* B-cell receptor reconstruction from single-cell RNA-seq with VDJPuzzle. *Bioinformatics* **34,** 2846–2847 (2018).

160. Wu, T. *et al.* Peripheral T cell expansion predicts tumour infiltration and clinical response. *Nature* **579,** 274–278 (2020).

161. Sheih, A. *et al.* Clonal kinetics and single-cell transcriptional profiling of CAR-T cells in patients undergoing CD19 CAR-T immunotherapy. *Nat. Commun* **11,** 219 (2020).

162. Wei, S. *et al.* Distinct Cellular Mechanisms Underlie Anti-CTLA-4 and Anti-PD-1 Checkpoint Blockade. sv. *Cell* **170,** 1120–1133 17 (2017).

163. Gerlach, J. *et al.* Combined quantification of intracellular (phospho-)proteins and transcriptomics from fixed single cells. *Sci. Rep* **9,** 1469 (2019).

164. Chung, H. *et al.* Simultaneous single cell measurements of intranuclear proteins and gene expression. *BioRxiv* (2021).

165. Cao, J. *et al.* Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science* **361,** 1380–1385 (2018).

166. Hill, A. *et al.* On the design of CRISPR-based single-cell molecular screens. *Nat. Methods* **15,** 271–274 (2018).

167. Al'Khafaji, A., Deatherage, D. & Brock, A. Control of Lineage-Specific Gene Expression by Functionalized gRNA Barcodes. *ACS Synth. Biol* **7,** 2468–2474 (2018).

168. Kong, W. *et al.* CellTagging: combinatorial indexing to simultaneously map lineage and identity at single-cell resolution. *Nat. Protoc* **15,** 750–772 (2020).

169. Bedard, P., Hansen, A., Ratain, M. & Siu, L. Tumour heterogeneity in the clinic. *Nature* **501,** 355–64 (2013).

170. Pogrebniak, K. & Curtis, C. Harnessing Tumor Evolution to Circumvent Resistance. *Trends Genet* **34,** 639–51 (2018).

171. Gatenby, R., Gillies, R. & Brown, J. Of cancer and cave fish. *Nat Rev Cancer* **11,** 237–8 (2011).

172. Burrell, R., McGranahan, N., Bartek, J. & Swanton, C. The causes and consequences of genetic heterogeneity in cancer evolution. *Nature* **501** (2013).

173. Eddy, J., Hood, L., Price, N. & Geman, D. Identifying tightly regulated and variably expressed networks by Differential Rank Conservation (DIRAC. *PLoS Comput Biol* **6:e1000792** (2010).

174. Bravo, H., Pihur, V., McCall, M., Irizarry, R. & Leek, J. Gene expression anti-profiles as a basis for accurate universal cancer signatures. *BMC Bioinformatics* **13** (2012).

175. Dinalankara, W. & Bravo, H. Gene Expression Signatures Based on Variability can Robustly Predict Tumor Progression and Prognosis. *Cancer Inform* **14:CIN.S23862** (2015).

176. Dinalankara, W. *et al.* Digitizing omics profiles by divergence from a baseline. *Proc Natl Acad Sci U S A* **115,** 4545–52 (2018).

177. Levitin, H., Yuan, J. & Sims, P. Single-Cell Transcriptomic Analysis of Tumor Heterogeneity. Trends Cancer Res. *Elsevier* **4,** 264–8 (2018).

178. Saadatpour, A., Lai, S., Guo, G. & Yuan, G.-C. Single-Cell Analysis in Cancer Genomics. *Trends Genet* **31,** 576–86 (2015).

179. McInnes, L. & UMAP, H. J. Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv* (2018).

180. Moon, K. *et al.* PHATE: A Dimensionality Reduction Method for Visualizing Trajectory Structures in High-Dimensional Biological Data. *BioRxiv* (2017).

181. Angerer, P. *et al.* destiny: diffusion maps for large-scale single-cell data in R. *Bioinformatics* **32,** 1241–3 (2016).

182. Pierson, E. & Yau, C. ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biol* **16** (2015).

183. DeTomaso, D. & Yosef, N. FastProject: a tool for low-dimensional analysis of single-cell RNA-Seq data. *BMC Bioinformatics* **17** (2016).

184. Amir, E.-A. *et al.* viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nat Biotechnol* **31,** 545–52 (2013).

185. DiGiuseppe, J., Cardinali, J., Rezuke, W. & Pe'er, D. PhenoGraph and viSNE facilitate the identification of abnormal T-cell populations in routine clinical flow cytometric data. *Cytometry B Clin Cytom* **94,** 588–601 (2018).

186. Mantsoki, A., Devailly, G. & Joshi, A. Gene expression variability in mammalian embryonic stem cells using single cell RNA-seq data. *Comput Biol Chem* **63,** 52–61 (2016).

187. Brennecke, P. *et al.* Accounting for technical noise in single-cell RNA-seq experiments. *Nat Methods* **10,** 1093–5 (2013).

188. Afsari, B., Geman, D. & Fertig, E. Learning dysregulated pathways in cancers from differential variability analysis. *Cancer Inform* **13,** 61–7 (2014).

189. Greaves, M. & Maley, C. Clonal evolution in cancer. *Nature* **481,** 306–13 (2012).

190. Zappia, L., Phipson, B. & Oshlack, A. Splatter: simulation of single-cell RNA sequencing data. *Genome Biol* **18** (2017).

191. D, D. *et al.* Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. *Cell* **174,** 716–29 (2018).

192. Azizi, E., Prabhakaran, S., Carr, A. & Pe'er, D. Bayesian Inference for Single-cell Clustering and Imputing. *Genomics and Computational Biology* **3** (2017).

193. Walter, V. *et al.* Molecular subtypes in head and neck cancer exhibit distinct patterns of chromosomal gain and loss of canonical cancer genes. *PLoS One* **8:e56823** (2013).

194. Leek, J., Johnson, W., Parker, H., Jaffe, A. & Storey, J. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* **28,** 882–3 (2012).

195. Hopkins, A. *et al.* T cell receptor repertoire features associated with survival in immunotherapy-treated pancreatic ductal adenocarcinoma. *JCI Insight* **3** (2018).

196. Ritchie, M. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* **43:e47** (2015).

197. Liberzon, A. *et al.* The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst* **1,** 417–25 (2015).

198. Afsari, B. *et al.* Splice Expression Variation Analysis (SEVA) for inter-tumor heterogeneity of gene isoform usage in cancer. *Bioinformatics* **34,** 1859–67 (2018).

199. Afsari, B. *et al.* Splice Expression Variation Analysis (SEVA) for inter-tumor heterogeneity of gene isoform usage in cancer. *Bioinformatics* **34,** 1859–67 (2018).

200. Kuriakose, M. *et al.* Selection and validation of differentially expressed genes in head and neck cancer. *Cell Mol Life Sci* **61,** 1372–83 (2004).

201. Drost, H.-G. Philentropy: Information Theory and Distance Quantification with R. *JOSS* **3** (2018).

202. Johnson, W., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8,** 118–27 (2007).

203. Ikeda, Y. *et al.* Clinical significance of T cell clonality and expression levels of immune-related genes in endometrial cancer. *Oncol Rep* **37,** 2603–10 (2017).

204. Mandal, R. *et al.* The head and neck cancer immune landscape and its immunotherapeutic implications. *JCI Insight* **1:e89829** (2016).

205. Alkasalias, T., Moyano-Galceran, L., Arsenian-Henriksson, M. & Lehti, K. Fibroblasts in the Tumor Microenvironment: Shield or Spear? *Int J Mol Sci* **10** (2018).

206. Clark, B. *et al.* Comprehensive analysis of retinal development at single cell resolution identifies NFI factors as essential for mitotic exit and specification of late-born cells. *Neuron* **102** (2019).

207. Afsari, B., Favorov, A., Fertig, E. & Cope, L. REVA: a rank-based multi-dimensional measure of correlation. *BioRxiv* (2018).

208. Zhao, N. *et al.* Testing in Microbiome-Profiling Studies with MiRKAT, the Microbiome Regression-Based Kernel Association Test. *Am J Hum Genet* **96,** 797–807 (2015).

209. Chen, H., Ye, F. & Guo, G. Revolutionizing immunology with single-cell RNA sequencing. *Cell Mol Immunol* **16,** 242–249 (2019).

210. Mestas, J. & Hughes, C. Of mice and not men: differences between mouse and human immunology. *J Immunol* **172,** 2731–2738 (2004).

211. Perel, P. *et al.* Comparison of treatment effects between animal experiments and clinical trials: systematic review. *BMJ* **334** (2007).

212. LF, A. *et al.* Discovery of specialized NK cell populations infiltrating human melanoma metastases. *JCI Insight* **4** (2019).

213. Clark, B. *et al.* Single-Cell RNA-Seq Analysis of Retinal Development Identifies NFI Factors as Regulating Mitotic Exit and Late-Born Cell Specification. *Neuron* **102,** 1111–26 (2019).

214. Lu, Y. *et al.* Single-cell analysis of human retina identifies evolutionarily conserved and species-specific mechanisms controlling development. *Dev Cell* (2020).

215. Hodi, F. *et al.* Improved survival with ipilimumab in patients with metastatic melanoma. *N Engl J Med* **363,** 711–723 (2010).

216. Robert, C. *et al.* Ipilimumab plus dacarbazine for previously untreated metastatic melanoma. *N Engl J Med* **364,** 2517–2526 (2011).

217. Krummel, M. & Allison, J. CTLA-4 engagement inhibits IL-2 accumulation and cell cycle progression upon activation of resting T cells. *J Exp Med* **183,** 2533–2540 (1996).

218. Sutmuller, R. *et al.* Synergism of cytotoxic T lymphocyte-associated antigen 4 blockade and depletion of CD25(+) regulatory T cells in antitumor therapy reveals alternative pathways for suppression of autoreactive cytotoxic T lymphocyte responses. *J Exp Med* **194,** 823–832 (2001).

219. Du, X. *et al.* A reappraisal of CTLA-4 checkpoint blockade in cancer immunotherapy. *Cell Res* **28,** 416–432 (2018).

220. Simpson, T. *et al.* Fcdependent depletion of tumor-infiltrating regulatory T cells co-defines the efficacy of anti-CTLA-4 therapy against melanoma. *J Exp Med* **210,** 1695–1710 (2013).

221. Sharma, A. *et al.* Anti-CTLA-4 Immunotherapy Does Not Deplete FOXP3+ Regulatory T Cells (Tregs) in Human Cancers. *Clin Cancer Res* **25,** 1233–1238 (2019).

222. Picelli, S. *et al.* Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat Methods* **10,** 1096–1098 (2013).

223. JN, C. G. A. R. N. W., EA, C., GB, M., S, K. & BA, O. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet* **45,** 1113–1120 (2013).

224. Colaprico, A. *et al.* TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res* **44:e71** (2016).

225. Thorsson, V. *et al.* The Immune Landscape of Cancer. *Immunity* **48,** 812–30 (2018).

226. Sherman, T., Gao, T. & Fertig, E. CoGAPS 3: Bayesian non-negative matrix factorization for single-cell analysis with asynchronous updates and sparse data structures. *BMC Bioinformatics* **21** (2020).

227. Stein-O'Brien, G. *et al.* PatternMarkers GWCoGAPS for novel data-driven biomarkers via whole transcriptome NMF. *Bioinformatics* **33,** 1892–1894 (2017).

228. Ochs, M. *et al.* Detection of treatmentinduced changes in signaling pathways in gastrointestinal stromal tumors using transcriptomic data. *Cancer Res* **69,** 9125–9132 (2009).

229. Liu, J. *et al.* An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. *Cell* **173,** 400–16 (2018).

230. Sharma, G., Colantuoni, C., Goff, L., Fertig, E. & Stein-O'Brien, G. projectR: An R/Bioconductor package for transfer learning via PCA, NMF, correlation, and clustering. *BioRxiv* (2019).

231. Aldeghaither, D. *et al.* A Mechanism of Resistance to Antibody-Targeted Immune Attack. *Cancer Immunol Res* **7,** 230–243 (2019).

232. Somanchi, S., Senyukov VV, D., CJ, L. & D.A. Expansion, purification, and functional assessment of human peripheral blood NK cells. *J Vis Exp* (2011).

233. Fertig, E. *et al.* CoGAPS matrix factorization algorithm identifies transcriptional changes in AP-2alpha target genes in feedback from therapeutic inhibition of the EGFR network. *Oncotarget* **7,** 73845–73864 (2016).

234. Stein-O'Brien, G. *et al.* Enter the Matrix: Factorization Uncovers Knowledge from Omics. *Trends Genet* **34,** 790–805 (2018).

235. Way, G., Zietz, M., Rubinetti, V., Himmelstein, D. & Greene, C. Sequential compression of gene expression across dimensionalities and methods reveals no single best method or dimensionality. *BioRxiv* (2019).

236. Lu, D. *et al.* Beyond T Cells: Understanding the Role of PD-1/PD-L1 in Tumor-Associated Macrophages. *J Immunol Res* **2019** (2019).

237. Xiong, H. *et al.* Anti- PD-L1 Treatment Results in Functional Remodeling of the Macrophage Compartment. *Cancer Res* **79,** 1493–1506 (2019).

238. Gotthardt, D. & Sexl, V. STATs in NK-Cells: The Good, the Bad, and the Ugly. *Front Immunol* **7** (2016).

239. Lanier, L. Turning on natural killer cells. *J Exp Med* **191,** 1259–1262 (2000).

240. Hsu, J. *et al.* Contribution of NK cells to immunotherapy mediated by PD-1/PD-L1 blockade. *J Clin Invest* **128,** 4654–4668 (2018).

241. Sanseviero, E. *et al.* Anti-CTLA-4 Activates Intratumoral NK Cells and Combined with IL15/IL15R Complexes Enhances Tumor Control. *Cancer Immunol Res* **7,** 1371–1380 (2019).

242. Cursons, J. *et al.* A Gene Signature Predicting Natural Killer Cell Infiltration and Improved Survival in Melanoma Patients. *Cancer Immunol Res* **7,** 1162–1174 (2019).

243. Patil, V. *et al.* Precursors of human CD4+ cytotoxic T lymphocytes identified by single-cell transcriptome analysis. *Sci Immunol* **3** (2018).

244. Smith, S. *et al.* Diversity of peripheral blood human NK cells identified by single-cell RNA sequencing. *Blood Adv* **4,** 1388–1406 (2020).

245. Kuiper, H., Brouwer, M., Linsley, P. & Lier, R. Activated T cells can induce high levels of CTLA-4 expression on B cells. *J Immunol* **155,** 1776–1783 (1995).

246. Pioli, C., Gatta, L., Ubaldi, V. & Doria, G. Inhibition of IgG1 and IgE production by stimulation of the B cell CTLA-4 receptor. *J Immunol* **165,** 5530–5536 (2000).

247. Wang, X.-B. *et al.* Expression of CTLA-4 by human monocytes. *Scand J Immunol* **55,** 53–60 (2002).

248. Laurent, S. *et al.* CTLA-4 is expressed by human monocyte-derived dendritic cells and regulates their functions. *Hum Immunol* **71,** 934–941 (2010).

249. Lang, S., Vujanovic, N., Wollenberg, B. & Whiteside, T. Absence of B7.1-CD28/CTLA-4- mediated co-stimulation in human NK cells. *Eur J Immunol* **28,** 780–786 (1998).

250. Beldi-Ferchiou, A. & Caillat-Zucman, S. Control of NK Cell Activation by Immune Checkpoint Molecules. *Int J Mol Sci* **18** (2017).

251. Lanuza, P. *et al.* Recalling the biological significance of immune checkpoints on NK cells: A chance to overcome LAG3, PD1, and CTLA4 inhibitory pathways by adoptive NK cell transfer? *Front Immunol* **10** (2019).

252. Chen, G., Ning, B. & Shi, T. Single-Cell RNA-Seq Technologies and Related Computational Data Analysis. *Front Genet* **10** (2019).

253. Pistillo, M. *et al.* CTLA-4 is not restricted to the lymphoid cell lineage and can function as a target molecule for apoptosis induction of leukemic cells. *Blood* **101,** 202–209 (2003).

254. Contardi, E. *et al.* CTLA-4 is constitutively expressed on tumor cells and can trigger apoptosis upon ligand interaction. *Int J Cancer* **117,** 538–550 (2005).

255. Zhuang, X. & Long, E. CD28 Homolog Is a Strong Activator of Natural Killer Cells for Lysis of B7H7+ Tumor Cells. *Cancer Immunol Res* **7,** 939–951 (2019).

256. Simmons, D. & Seed, B. The Fc gamma receptor of natural killer cells is a phospholipid-linked membrane protein. *Nature* **333,** 568–570 (1988).

257. Rowshanravan, B., Halliday, N. & Sansom, D. CTLA-4: a moving target in immunotherapy. *Blood* **131,** 58–67 (2018).

258. Stojanovic, A., Fiegler, N., Brunner-Weinzierl, M. & Cerwenka, A. CTLA-4 is expressed by activated mouse NK cells and inhibits NK Cell IFN- production in response to mature dendritic cells. *J Immunol* **192,** 4184–4191 (2014).

259. Quatrini, L. *et al.* Glucocorticoids and the cytokines IL-12, IL-15, and IL-18 present in the tumor microenvironment induce PD-1 expression on human natural killer cells. *J Allergy Clin Immunol* (2020).

260. Tonn, T., Becker, S., Esser, R., Schwabe, D. & Seifried, E. Cellular immunotherapy of malignancies using the clonal natural killer cell line NK-92. *J Hematother Stem Cell Res* **10,** 535–544 (2001).

261. Sabry, M. *et al.* Tumor- and cytokine-primed human natural killer cells exhibit distinct phenotypic and transcriptional signatures. *PLoS One* **14:e0218674** (2019).

262. López-Soto, A., Gonzalez, S., Smyth, M. & Galluzzi, L. Control of Metastasis by NK Cells. *Cancer Cell* **32,** 135–154 (2017).

263. Schantz, S., Savage, H., Racz, T., Taylor, D. & Sacks, P. Natural killer cells and metastases from pharyngeal carcinoma. *Am J Surg* **158,** 361–366 (1989).

264. Schantz, S. & Ordonez, N. Quantitation of natural killer cell function and risk of metastatic poorly differentiated head and neck cancer. *Nat Immun Cell Growth Regul* **10,** 278–288 (1991).

265. Lu, J. *et al.* Detailed analysis of inflammatory cell infiltration and the prognostic impact on nasopharyngeal carcinoma. *Head Neck* **40,** 1245–1253 (2018).

266. Nersesian, S. *et al.* NK cell infiltration is associated with improved overall survival in solid cancers: A systematic review and meta-analysis. *Transl Oncol* **14** (2021).

267. Iannone, F. *et al.* Effect of surgery on pancreatic tumor-dependent lymphocyte asset: modulation of natural killer cell frequency and cytotoxic function. *Pancreas* **44,** 386–393 (2015).

268. Davis, M. *et al.* Effect of pemetrexed on innate immune killer cells and adaptive immune T cells in subjects with adenocarcinoma of the pancreas. *J Immunother* **35,** 629–640 (2012).

269. Karakhanova, S. *et al.* Prognostic and predictive value of immunological parameters for chemoradioimmunotherapy in patients with pancreatic adenocarcinoma. *Br J Cancer* **112,** 1027–1036 (2015).

270. Xu, Y.-F. *et al.* Abnormal distribution of peripheral lymphocyte subsets induced by PDAC modulates overall survival. *Pancreatology* **14,** 295–301 (2014).

271. Newman, A. *et al.* Robust enumeration of cell subsets from tissue expression profiles. *Nat Methods* **12,** 453–457 (2015).

272. Murphy, W., Parham, P. & Miller, J. NK cells–from bench to clinic. *Biol Blood Marrow Transplant* **18:S2–S7** (2012).

273. Mieth, B. *et al.* Using transfer learning from prior reference knowledge to improve the clustering of single-cell RNASeq data. *Sci Rep* **9** (2019).

274. Peng, D. *et al.* Evaluating the transcriptional fidelity of cancer models. *Genome Med* **13,** 73 (2021).

275. Concha-Benavente, F. *et al.* PD-L1 Mediates Dysfunction in Activated PD-1+ NK Cells in Head and Neck Cancer Patients. *Cancer Immunol Res* **6,** 1548–1560 (2018).

276. Mariotti, F. *et al.* PD-1 in human NK cells: evidence of cytoplasmic mRNA and protein expression. *Oncoimmunology* **8** (2019).

277. Alvarez-Breckenridge, C., Yu, J., Kaur, B., Caligiuri, M. & Chiocca, E. Deciphering the Multifaceted Relationship between Oncolytic Viruses and Natural Killer Cells. *Adv Virol* **2012** (2012).

278. Valk, E., Rudd, C. & Schneider, H. CTLA-4 trafficking and surface expression. *Trends Immunol* **29,** 272–279 (2008).

279. Azuma, M., Cayabyab, M., Buck, D., Phillips, J. & Lanier, L. Involvement of CD28 in MHCunrestricted cytotoxicity mediated by a human natural killer leukemia cell line. *J Immunol* **149,** 1115–1123 (1992).

280. Galea-Lauri, J. *et al.* Expression of a variant of CD28 on a subpopulation of human NK cells: implications for B7-mediated stimulation of NK cells. *J Immunol* **163,** 62–70 (1999).

281. Chambers, B., Salcedo, M. & Ljunggren, H. Triggering of natural killer cells by the costimulatory molecule CD80 (B7-1. *Immunity* **5,** 311–317 (1996).

282. Wilson, J. *et al.* NK cell triggering by the human costimulatory molecules CD80 and CD86. *J Immunol* **163,** 4207–4212 (1999).

283. Martín-Fontecha, A., Assarsson, E., Carbone, E., Kärre, K. & Ljunggren, H. Triggering of murine NK cells by CD40 and CD86 (B7-2). *J Immunol* **162,** 5910–5916 (1999).

284. Luque, I., Reyburn, H. & Strominger, J. Expression of the CD80 and CD86 molecules enhances cytotoxicity by human natural killer cells. *Hum Immunol* **61,** 721–728 (2000).

285. Terrazzano, G. *et al.* Differential involvement of CD40, CD80, and major histocompatibility complex class I molecules in cytotoxicity induction and interferon-gamma production by human natural killer effectors. *J Leukoc Biol* **72,** 305–311 (2002).

286. Tallerico, R. *et al.* IL-15, TIM-3 and NK cells subsets predict responsiveness to anti-CTLA-4 treatment in melanoma patients. *Oncoimmunology* **6:e1261242** (2017).

287. Kohlhapp, F. *et al.* NK cells and CD8+ T cells cooperate to improve therapeutic responses in melanoma treated with interleukin-2 (IL-2) and CTLA-4 blockade. *J Immunother Cancer* **3** (2015).

288. Ott, P. *et al.* An immunogenic personal neoantigen vaccine for patients with melanoma. *Nature* (July 13, 2017).

289. Sahin, U. *et al.* Personalized RNA mutanome vaccines mobilize poly-specific therapeutic immunity against cancer. *Nature* (July 13, 2017).

290. Zheng, C. *et al.* Transcriptomic profiles of neoantigen-reactive T cells in human gastrointestinal cancers. *Cancer Cell* **423.** PMID: 35413272. (2022).

291. Oliveira, G. *et al.* Phenotype, specificity and avidity of antitumour CD8+ T cells in melanoma. *Nature* **Aug;596(7870):119-125** (2021).

292. Caushi, J. *et al.* Transcriptional programs of neoantigen-specific TIL in anti-PD-1-treated lung cancers. *Nature* **Aug;596(7870):126-132** (2021).

293. Lowery, F. *et al.* Molecular signatures of antitumor neoantigen-reactive T cells from metastatic human cancers. *Science* (2022).

294. Wolock, S., Lopez, R. & Klein, A. Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. *Cell Syst* **291** (2019).

295. C, D. C. *et al.* Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science* (May 13, 2022).

296. Sturm, G. *et al.* Scirpy: a Scanpy extension for analyzing single-cell T-cell receptor-sequencing data. *Bioinformatics* (Sept. 15, 2020).

297. Kawashima, S. & Kanehisa, M. AAindex: amino acid index database. *Nucleic Acids Res* **1;28(1):374** (Jan. 2000).

298. Giudicelli, V., Duroux, P. & Ginestoux, C. IMGT/LIGM-DB, the IMGT comprehensive database of immunoglobulin and T cell receptor nucleotide sequences. *Nucleic Acids Res* **34** (2006).

299. Lefranc, M. *et al.* IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. *Dev Comp Immunol* **Jan;27(1):55-77** (2003).

300. Shugay, M. *et al.* VDJdb: a curated database of T-cell receptor sequences with known antigen specificity. *Nucleic Acids Res* (Jan. 4, 2018).

301. Levin, N. *et al.* in *Identification and Validation of T-cell Receptors Targeting RAS Hotspot Mutations in Human Cancers for Use in Cell-based Immunotherapy. Clin Cancer Res* (Sept. 15, 2021).

302. Bear, A. *et al.* Biochemical and functional characterization of mutant KRAS epitopes validates this oncoprotein for immunological targeting. *Nat Commun* (July 16, 2021).

303. Gros, A. *et al.* Recognition of human gastrointestinal cancer neoantigens by circulating PD-1+ lymphocytes. *J Clin Invest* (Nov. 1, 2019).

304. Li, B. *et al.* Landscape of tumor-infiltrating T cell repertoire of human cancers. *Nat Genet* (July 7, 2016).

305. Cafri, G. *et al.* Memory T cells targeting oncogenic mutations detected in peripheral blood of epithelial cancer patients. *Nat Commun* **10,** 449 (2019).

306. Leko, V. *et al.* Identification of neoantigen-reactive T lymphocytes in the peripheral blood of a patient with glioblastoma. *J Immunother Cancer* (July 9, 2021).

307. Martin, S. *et al.* A library-based screening method identifies neoantigen-reactive T cells in peripheral blood prior to relapse of ovarian cancer. *Oncoimmunology* (Sept. 21, 2017).

June, 2022

Emily Davis-Marcisak

Education

PhD in Human Genetics *August 2017 - June 2022*
Johns Hopkins School of Medicine
McKusick-Nathans Institute of Genetic Medicine
Baltimore, MD, USA

Master of Science in Biotechnology *August 2014 - May 2016*
Johns Hopkins University
Baltimore, MD, USA

Bachelor of Science in Biochemistry *August 2011 - May 2013*
University of Florida
Gainesville, FL, USA

Associate of Arts *August 2008 - July 2011*
Eastern Florida State College
Melbourne, FL, USA

Positions and Employment

Research Technologist *May 2015 - May 2017*
Johns Hopkins University, McKusick-Nathans Institute of Genetic Medicine

Worked on the development and functional characterization of immortalized human bronchial epithelial cell lines originally derived from a patient with cystic fibrosis (CFBEs). Analysis aims to cover 500+ rare variants, including variants with conflicting genotype-phenotype correlation in relation to what is observed clinically. I created 90 such lines.

Development of cell lines involved: plasmid design, site directed mutagenesis, DNA sequencing, stable and transient transfection, integration PCR, qRT-PCR, western blotting, and Ussing chamber short circuit current analysis for base function and assessment of pharmacological treatment.

Production Associate I *March 2014 - April 2015*
Thermo Fisher Scientific / Dharmacon, GE Healthcare

Responsibilities included completing the primary tasks in areas of Synthesis, Quality Control, Post Process, and Packaging for the manufacturing of all gene editing products. This involved daily operation and trouble shooting of liquid handling robots and continually seeking ways to improve the production process. This also required maintaining proper Housekeeping per Standard Operating Procedures (SOP) and 5S standards within the department and compliance with EHS regulations and policies.

Professional Research Assistant *June 2013 - January 2014*
University of Colorado, Biochemistry Department

Sought to find novel small molecule TLR4 inhibitors as a potential treatment for sepsis via high throughput screening of a drug library. Hit compounds were confirmed and analogues of these molecules were synthesized and studied for in vitro activity and mechanism of action.

Compounds were tested for viability using a variety of in vitro cell-based assays such as: nitric oxide, ELISA, SEAP, and toxicity assays. Also performed experiments such as: flow cytometry, PCR, western blot, FRET, and localization microscopy.

Undergraduate Research Assistant *March 2012 - May 2013*
Univeristy of Florida, Katritzky Lab of Heterocyclic Organic Chemistry

Performed benzotriazole based natural product synthesis as well as solution-phase peptide synthesis, NMR, HPLC, and mass spectrometry in the development of a novel heterocyclic peptide with potential biological activity, Rolloamide B.

Laboratory Assistant *August 2010 - July 2011*
Eastern Florida State College

Duties consisted of: Preparing weekly labs for general biology and chemistry courses; regularly taking inventory and stocking lab benches; calibrating and cleaning all equipment; preparing necessary solutions.

Funding

T cell mechanisms of immunotherapy response in pancreatic ductal adenocarcinoma
National Cancer Institute (Bethesda) 2021-02-01 to 2024-01-31
Grant number: F31CA250135

Teaching Experience

Teaching Assistant, Introduction to Single Cell Analysis *May 2020 - June 2020*
Johns Hopkins School of Medicine, Sidney Kimmel Comprehensive Cancer Center

Teaching Assistant, Advanced Topics in Human Genetics *January 2020 - May 2020*
Johns Hopkins School of Medicine

Teaching Assistant, Pathology for Graduate Students *August 2019 - October 2019*
Johns Hopkins School of Medicine

Teaching Assistant, Biochemistry *August 2013 - December 2013*
University of Colorado Boulder

Tutor, Chemistry *August 2010 - July 2011*
Eastern Florida State College

Awards and Honors

NCI Graduate Student Recruitment Program *May 2022*

JHU SOM Graduate Student Association Travel Award *March 2022*

Invited Speaker, SACB *November 2020*

JXTX Foundation Scholar, Galaxy Project *November 2020*

Invited Speaker, AACR *June 2020*

Women in Cancer Research Scholar, AACR *June 2020*

Florida Bright Futures Academic Scholars Award *2010 - 2013*

Phi Theta Kappa Vice President of Service *2010 - 2011*

Eastern Florida State College Graduation with Honors *2008 - 2011*

Sam Walton Community Scholarship Winner *2010*

Professional memberships and organizations

American Society of Human Genetics *November 2018 -*

Women in Cancer Research *August 2018 -*

American Association for Cancer Research *August 2018 -*

Conferences Attended

Single Cell Genomics Gordon Research Conference *May 2022*
Les Diablerets, Switzerland
Invited speaker

Biological Data Science Conference *November 2020*
Virtual
JXTX Foundation Scholar

Systems Approaches to Cancer Biology *November 2020*
Virtual
Invited speaker

American Association for Cancer Research Annual Meeting *June 2020*
Virtual
Women in Cancer Research Scholar and invited speaker

American Association for Cancer Research Annual Meeting *March 2019*
Atlanta, GA

5th Annual Metastatic Breast Cancer Conference *November 2018*
Baltimore, MD

30th Anniversary AACR Special Conference Convergence: Artificial Intelligence, Big Data, and Prediction in Cancer *October 2018*
Newport, RI

Publications and Presentations

Publications (in chronological order)

1) Sidiropoulos D, Rafie C, Jang J, Castanon S, Baugh A, Garcia E, Christmas B, Narumi V, **Davis-Marcisak E**, et al. Entinostat decreases immune suppression to promote an anti-tumor response within a HER2+ breast tumor microenvironment. *Cancer Immunology Research*, 10.1158/2326-6066.CIR-21-0170.
2) Fitzgerald A, **Davis-Marcisak E**, Fertig E, Weiner L. DPP inhibition alters the CXCR3 axis and enhances NK and CD8+ T cell infiltration to improve anti-PD1 efficacy in murine models of pancreatic ductal adenocarcinoma. *Journal for ImmunoTherapy of Cancer*, **9**(11)::e002837.
3) **Davis-Marcisak E**, Fitzgerald A, Kessler M, Danilova L, et al. Transfer learning between preclinical models and human tumors identifies a conserved NK cell activation signature in anti-CTLA-4 responsive tumors. (2021) *Genome Medicine*, **13**(1):129.
4) **Davis-Marcisak E**, Deshpande A, Stein-O'Brien G, Ho WJ, et al. From bench to bedside: single-cell analysis for cancer immunotherapy. (2021) *Cancer Cell*, **39**(8):1062-1080.
5) Zhang S, Gong C, Ruiz-Martinez A, Wang H, **Davis-Marcisak E**, et al. (2021) Integrating single cell sequencing with a spatial quantitative systems pharmacology model spQSP for personalized prediction of triple-negative breast cancer immunotherapy response. *ImmunoInformatics*.
6) Joynt AT, Evans TA, Pellicore MJ, **Davis-Marcisak E**, et al. (2020) Evaluation of both exonic and intronic variants for effects on RNA splicing allows for accurate assessment of the effectiveness of precision therapies. *PLoS Genetics*, **16**(10):e1009100.
7) Kagohara LT, Zamuner F, **Davis-Marcisak E**, Sharma G, Considine M, Allen J, et al. (2020) Integrated single cell and bulk gene expression and ATAC-seq reveals heterogeneity and early changes in pathways associated with resistance to cetuximab in HNSCC sensitive cell lines. *Br. J. Cancer*, **123**:101–113.
8) **Davis-Marcisak E**, Sherman T, Orugunta P, Stein-O'Brien G, Puram SV, Roussos Torres E, et al. (2019) Differential variation analysis enables detection of tumor heterogeneity using single-cell RNA-sequencing data. *Cancer Research*, **79**(19):5102-5112.

9) Brian C, Stein-O'Brien G, Shiau F, Cannon G, **Davis-Marcisak E**, et al. (2019) Single-Cell RNA-Seq Analysis of Retinal Development Identifies NFI Factors as Regulating Mitotic Exit and Late-Born Cell Specification. *Neuron*, **102**(6):1111-1126.
10) McCague AF, Raraigh KS, Pellicore MJ, **Davis-Marcisak E**, et al. (2019) Correlating cystic fibrosis transmembrane conductance regulator function with clinical features to inform precision treatment of cystic fibrosis. *Am J Respir Crit Care Med*, **199**(9):1116-1126.
11) Sharma N, Evans TA, Pellicore MJ, **Davis E**, Aksit MA, McCague AF, et al. (2018) Capitalizing on the heterogeneous effects of CFTR nonsense and frameshift variants to inform therapeutic strategy for cystic fibrosis. *PLoS Genetics*, **14**(11):e1007723.
12) Han ST, Rab A, Pellicore MJ, **Davis E**, McCague AF, Evans TA, et al. (2018) Residual Function of Cystic Fibrosis Mutants Predicts Response to Small Molecule CFTR Modulators. *JCI Insight*, **3**(14):e121159.
13) Raraigh KS, Han ST, **Davis E**, Evans TA, Pellicore MJ, McCague AF, et al. (2018) Functional assays are essential for interpretation of missense variants associated with variable expressivity. *American Journal of Human Genetics*, **102**(6):1062-1077.
14) Lee M, Roos P, Sharma N, Atalar M, Evans TA, Pellicore MJ, **Davis E**, et al. (2017) Systematic computational identification of variants that activate exonic and intronic cryptic splice sites. *American Journal of Human Genetics*, **100**(5):751-765.
15) Sharma N, LaRusch J1, Sosnay PR, Gottschalk LB, Lopez AP, Pellicore MJ, Evans T, **Davis E**, et al. (2016) A sequence upstream of canonical PDZ binding motif within CFTR COOH terminus enhances NHERF1 interaction. *American Journal of Physiology-Lung Cellular and Molecular Physiology*, **311**(6):L1170-L1182.
16) Csakai A, Smith C, **Davis E**, Martinko A, Coulup S, Yin H. (2014) Saccharin Derivatives as Inhibitors of Interferon-Mediated Inflammation. *Journal of Medicinal Chemistry*, **57**(12):5348-55.
17) El Khatib M, Elagawany M, Caliskan E, **Davis E**, Faidallah HM, El-feky SA, Katritzky AR. (2013) Total Synthesis of Cyclic Heptapeptide Rolloamide B. *Chemical Communications*, **49**(26):2631-3.

Posters and Published Abstracts (in chronological order)

- **Davis-Marcisak E**, Fitzgerald A, Zaidi N, Jaffee E, et al. (2020) Transfer learning identifies common cellular determinants of immune checkpoint inhibitor response between preclinical tumor models and patients. *Cancer Research*, **80**:16 Supplement 3413.
- **Davis-Marcisak E**, Orugunta P, Stein-O'Brien G, Puram SV, Roussos Torres E, et al. (2019) Expression variation analysis for tumor heterogeneity in single-cell RNA-sequencing data. *Cancer Research*, **79**:13 Supplement 4697.
- Roussos Torres E, Rafie C, **Davis E**, Kagohara L, Fertig E, Jaffee E. (2019) Sequencing the tumor microenvironment and myeloid derived suppressor cells to understand response to immunotherapy in primary HER2 positive breast cancer. *Cancer Research*, **79**:13 Supplement 5027.
- N Sharma, TA Evans, A Joynt, **Davis E**, M Pellicore, M Atalar, et al. (2018) Theratyping of CFTR splicing variants. *Pediatric Pulmonology*, **53**(S2):204-204.
- Raraigh KS, Han ST, Rab A , Matthew P, Evans TA, **Davis E**, et al. (2018) Theratyping Rare and common CFTR missense variants based on residual function an phenotype. *Pediatric Pulmonology*, **53**(S2):262-263.
- KS Raraigh, S Han, **Davis E**, T Evans, M Pellicore, A Mccague, et al. (2018) WS17. 3 Functional characterization and CFTR2 disease liability assignment of 48 missense variants. *Journal of Cystic Fibrosis*, **17**:S31-S32.
- ST Han, A McCague, M Atalar, M Pellicore, **Davis E**, TA Evans, et al. (2017) Functional assessment of ultra rare CFTR missense variants in human airway cells informs disease liability and drug therapy. *Pediatric Pulmonology*, **52**(S47):S266-S266.
- N Sharma, TA Evans, M Pellicore, **Davis E**, ST Han, A McCague, et al. (2017) CFTR targeted therapy for a subset of splice-site and nonsense variants that allow protein production. *Pediatric Pulmonology*, **52**(S47):S262-S262.
- TA Evans, N Sharma, **Davis E**, A Joynt, GR Cutting, JL Taylor-Cousar. (2017) Identification of a loss of function CFTR variant in orangutans. *Pediatric Pulmonology*, **52**(S47):S269-S269.

- ST Han, M Pellicore, T Evans, **Davis E**, KS Raraigh, GR Cutting. (2016) Amino acid substitutions of pore-lining residues may comprise a distinct theratype of CF-causing CFTR variants. *Pediatric Pulmonology*, **51**(S45):246-246.
- Atalar, M, Vecchio-Pagán B, **Davis E**, Akhtar Y, Sharma N, Blackman SM, Cutting GR. (2016) Analysis of the SLC26A9 locus as a modifier of CF. *Pediatric Pulmonology*, **51**(S45):243-243.
- ST Han, M Pellicore, T Evans, **Davis E**, KS Raraigh, GR Cutting. (2016) Interpretation and stratification of CFTR splice-site variants to identify potential therapeutic targets for small molecule correctors and potentiators. *Pediatric Pulmonology*, **51**(S45):245-246.
- N Sharma, M Pellicore, T Evans, **Davis E**, A McCague, ST Han, et al. (2016) NMD inhibiton improves the outcome of corrector and potentiator NMD treatments for the nonsense mutations expressing complex glycosylated truncated protein. *Pediatric Pulmonology*, **51**(S45):293-293.

Programming skills

- Computer Languages: R; Python; Bash; Shell
- Software Tools: LaTeX; Vim; Illustrator
- Version Control: Git; SVN