ABSTRACT

Title of Dissertation:	CHARACTERIZATION OF SURVIVAL ASSOCIATED GENE INTERACTIONS AND LYMPHOCYTE HETEROGENEITY IN CANCER
	Assaf Magen, Doctor of Philosophy, 2019
Dissertation directed by:	Professor Sridhar Hannenhalli Department of Computer Science, UMD Professor Eytan Ruppin National Cancer Institute, NIH

Cancer is the second leading cause of death globally. Tumors form intricate ecosystems in which malignant and immune cells interact to shape disease progression. Yet, the molecular underpinnings of tumorigenesis and immunological responses to tumors are poorly understood, limiting their manipulation to elicit favorable clinical outcomes. This thesis lays conceptual frameworks for investigating the molecular interactions taking place in tumors as well as the diversity of the immune response to cancer.

In the molecular level of individual cancer cells, the phenotypic effect of perturbing a gene's activity depends on the activity level of other genes, reflecting the notion that phenotypes are emergent properties of a network of functionally interacting genes. In the context of cancer, contemporary investigations have primarily focused on just one type of functional genetic interaction (GI) – synthetic lethality (SL). However, there may be additional

types of GIs whose systematic identification would enrich the molecular and functional characterization of cancer. This thesis describes a novel data-driven approach called EnGIne, that applied to large-scale cancer data identifies 71,946 GIs spanning 12 distinct types, only a small minority of which are SLs. The detected GIs explain cancer driver genes' tissue-specificity and differences in patients' response to drugs, and stratify breast cancer tumors into refined subtypes. These results expand the scope of cancer GIs and lay a conceptual and computational basis for future studies of additional types of GIs and their translational applications.

Furthermore, tumor growth is continuously shaped by the immune response. However, T cells typically adopt a dysfunctional phenotype may be reversed using immunotherapy strategies. Most current tumor immunotherapies leverage cytotoxic $CD8^+$ T cells to elicit an effective anti-tumor response. Despite evidence for clinical potential of $CD4^+$ tumor-infiltrating lymphocytes (TILs), their functional diversity has limited our ability to harness their anti-tumor activity. To address this issue, we have used single-cell mRNA sequencing (scRNAseq) to analyze the response of $CD4^+$ T cells specific for a defined recombinant tumor antigen, both in the tumor microenvironment and draining lymph nodes (dLN). New computational approaches to characterize subpopulations identified TIL transcriptomic patterns strikingly distinct from those elicited by responses to infection, and dominated by diversity among T-bet-expressing T helper type 1 (Th1)-like cells. In contrast, the dLN response includes Follicular helper (Tfh)-like cells but lacks Th1 cells. We identify an interferon-driven signature in Th1-like TILs, and show that it is found in human liver cancer and melanoma, in which it is negatively associated with response to checkpoint therapy. Our study unveils unsuspected differences between tumor and virus $CD4^+$ T cell responses, and provides a proof-of-concept methodology to characterize tumor- control $CD4^+$ T cell effector programs. Targeting these programs should help improve immunotherapy strategies.

A non-technical overview

This dissertation discusses two central problems to cancer development: the interconnected processes taking place within tumors along with the diversity of the immune response to cancer.

Genetic interactions: The formation of malignant tumors is determined by complex interactions between many genes, therefore we cannot use an individual gene to eradicate cancer without assessing the state of its network of interactions. We designed new computational methodologies to characterize multiple types of interactions across thousands of genes which provides an improved characterization of tumorigenesis. These findings should help monitoring disease progression and tailoring the appropriate treatment to specific individuals based on their tumor-specific features.

T cell responses to tumors: The complexity of immune responses to cancer is not well characterized. Hence, we have limited capabilities in controlling the immune system to eradicate cancer. We use new computational approaches and biological technologies to characterize the functional diversity of immune responses to tumors. Therefore, we open possibilities for designing new ways to manipulate anti-tumor immune responses for clinical benefit.

CHARACTERIZATION OF SURVIVAL ASSOCIATED GENE INTERACTIONS AND LYMPHOCYTE HETEROGENEITY IN CANCER

by

Assaf Magen

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2019

Advisory Committee:
Professor Sridhar Hannenhalli, Chair & Advisor
Professor Eytan Ruppin, Co-Advisor
Professor Rémy Bosselut, Co-Advisor
Professor Ramani Duraiswami, Dean's Representative
Professor Steve Mount

Acknowledgements

I would like to express my deepest appreciation to my committee members for their unwavering support during my journey as a PhD trainee. I am deeply indebted to Dr. Ruppin for his profound belief in my work and invaluable advice, especially in the early stages of my training. I am extremely grateful for Dr. Hannanhalli's relentless support in nurturing my scientific curiosity and insightful contributions to my work. I would also like to extend my deepest gratitude to Dr. Bosselut for his invaluable guidance and patience throughout my transition into the scientific environment of experimental biology and immunology.

I also had the great pleasure of working with fantastic colleagues and collaborators. Special thanks to Dr. Schäffer for this unwavering support, guidance and insightful suggestions in our study of genetic interactions. I very much appreciate the key experimental contributions of Dr. Nie to our joint study of immune responses to tumors. I also wish to thank Dr. Ciucci for his valuable advice and unparalleled support in my immunological studies. Additionally, I would like to extend my sincere thanks to the members of the Ruppin, Hannenhalli and Bosselut labs for providing constructive criticism and a vibrant research environment.

Finally, I wish to thank the departments of Computer Science and Computational Biology (University of Maryland – College Park), the GPP Program (NCI-UMD Partnership for Integrative Cancer Research), and the Office of Science and Technology Resources (National Cancer Institute, Center for Cancer Research, National Institutes of Health) for their support.

Table of Contents

List of Figures

List of Abbreviations

GI	Functional genetic interaction
SL	Synthetic lethality
EnGIne	Encyclopedia of clinically significant GIs in cancer
TILs	Tumor-infiltrating lymphocytes
scRNAseq	Single-cell mRNA sequencing
dLN	Draining lymph nodes
Th1	T helper type 1
Tfh	Follicular helper
TSG	Tumor suppressor genes
CSC	Cancer stem cells
SDL	Synthetic Dosage Lethal
SR	Synthetic Rescues
MHC	Major histocompatibility complex
TME	Tumor microenvironment
FDR	False Discovery Rate
PIN	Protein interaction network
ROC–AUC	Receiver operating characteristic area under the curve
NMF	Non-Negative Matrix Factorization
HR	Hazard ratio
LCMV	Lymphocytic choriomeningitis virus
GP	Glycoprotein
MC38	Mouse colon adenocarcinoma cell line
tSNE	t-Distributed Stochastic Neighbor Embedding
UMI	Unique molecular identifiers
FC	Fold-change
NK	Natural Killer
TIL_{HLC}	Human liver cancer TILs
TIL_{Mel}	Human melanoma TILs

Chapter 1: Introduction

Malignant tumors develop through a complex process shaped by accumulation of genetic alterations conferring selective growth advantage and activating complex mechanisms to evade being recognized and eliminated by the immune system. This chapter provides an overview of the basic elements governing tumorigenesis and immune responses to tumors.

Determinants of cancer development

Unlike single cell organisms such as bacteria, animals are composed of an ecosystem of cells which reproduce and coordinate their behavior via multiple communication mechanisms. Oncogenesis, or cancer development, is characterized by defying the set of rules governing normal coordinated development and activity in multicellular organisms to provide an uncontrolled proliferative capacity, thus posing a threat to the host organism.

Genomic alterations drive oncogenesis

During normal process of cell division, the process in which a cell gives rise to two daughter cells, genetic information is copied and stored. Random errors that alter the genetic sequence (mutations) may occur during this process and may result in altered cellular functions (Vogelstein et al., 2013; Weinberg, 1983). In rare cases, such alterations are advantageous and provide the cell, providing higher proliferation, migration or survival capacity than other cells while also suppressing the regulatory processes meant to check and control such aberrations.

A wide variety of mechanisms, termed the ‘hallmarks of cancer’ (Hanahan and Weinberg, 2000, 2011; Lazebnik, 2010), are involved in this transformation, including

evasion of normal control of cell division (mitosis) and programmed cell death (apoptosis), the ability to establish new blood vessels, suppression and evasion of the immune system and the capacity to colonize distant tissues. The next sections will focus on the cell survival capacity and immune response interaction in depth as they are most relevant to the topics discussed in this dissertation.

Over time, mutated cells divide and give rise to a malignant tumor. Malignant tumors, unlike benign lesions, pose a threat to healthy cellular ecosystem by means of migrating to distant organs (metastasis) (Chambers et al., 2000; Fidler, 2003; Nguyen et al., 2009) and an unlimited proliferation at these new sites. Importantly, cancers undergo an evolutionary process where each cancer cell may accumulate its own set of selectively advantageous mutations (driver mutations), rendering it and its decedents a phenotypically distinct clone (Greaves and Maley, 2012). Thus, tumors evolve into a heterogeneous collection of clones with different molecular makeups. Each of which may respond differently to therapeutic interventions, making it harder to develop a single therapeutic regimen to eradicate the disease.

The number of mutations varies considerably (average of 33 to 66 mutations) across a variety of tissue origins, corresponding to distinct cancer types (Vogelstein et al., 2013). Skin (melanoma) and lung cancers outrank other tissues by accumulating hundreds of mutations owing to their direct exposure to mutagens such as ultraviolet light and cigarette smoke, respectively. The majority of mutations reflect single-base substitution (such as A>T or C>G) while the minority of mutations account for insertions or deletions. Additional types of genetic alterations include changes in the number of chromosomes (aneuploidy) as well as deletions and translocations of whole segments of genomes, thereby affecting many genes at a time.

Cancers evade normal cell-cycle control and apoptosis

One of the central mechanisms through which cancers obtain an unlimited proliferative capacity is carried out by avoiding the normal life cycle imposed on normal cells. Particularly, the proper regulation of cell division and cell death is heavily disrupted. Cell division processes are orchestrated in part by extracellular growth and anti-growth signals. Cancer cells are insensitive to anti-growth signals and lack dependence on growth signals. Additionally, cancer cells are typically resistant to induced apoptosis (programmed cell death) by multiple mechanisms. Tumor Protein P53 (encoded by *TP53*) has a central role in inducing cell cycle arrest and apoptosis upon cellular stress such as DNA damage (Harris, 1996; Junttila and Evan, 2009). Mutations of deleterious consequence on the function of *TP53*, i.e., those which prevent the proper function of the protein, confer tumors with the ability to evade apoptosis.

Cancers evade recognition by the immune system

In addition to avoiding normal growth and death signals, cancers evolve to bypass anti-tumor responses generated by the host's adaptive immune system. The adaptive arm of the immune system specializes in generating cells specific to a variety of foreign genetic sequences (antigens), which are naturally presented on the surface of all nucleated cells. The mutated genomes of cancer cells present with new and unfamiliar peptides, termed neo-antigens (Linnemann et al., 2015; Segal et al., 2008; Sjoblom et al., 2006). Upon immune recognition, immune cells may destroy cancer cells and eradicate the tumor. However, cancers activate sophisticated mechanisms to prevent proper immune surveillance or hijack the immune response for their benefit by blocking inflammation signals and driving tolerance (Schreiber et al., 2011).

Cancer drivers as diagnostic markers and therapeutic agents

To understand the genetic drivers of oncogenesis, studies have identified genes contributing to increased risk to develop cancer. These efforts have guided the development of strategies to monitor disease progression and designing drugs to change the course of disease.

Identification of cancer drivers

Cancer-critical genes were classified into two main categories: oncogenes and tumor-suppressors. While oncogenes are genes with the potential to promote oncogenesis when overactivated by genetic alterations which disrupt the normal regulation of that gene, tumor suppressor genes (TSG) inhibit tumorigenesis by slowing proliferation, correcting genetic alterations or inducing apoptosis (Weinberg, 1991). Once the functions of TSGs are lost by genomic alterations affecting both copies of the gene (alleles), the appropriate regulation of cellular processes is disrupted, thus leading to tumor development.

Owing to the different mechanism by which oncogenes and tumor suppressors affect tumorigenesis, the methods used to identify such genes differ. The *Ras* oncogene family was identified by introducing DNA fragments extracted from human tumors to cells in a culture dish (*in vitro*) (Harvey, 1964). Some of these cells outgrow others and show a transformed phenotype. In contrast, tumor suppressors, such as the retinoblastoma protein (*Rb*), were identified by deleting both copies of a specific band on chromosome 13 which is visibly abnormal in retinoblastoma patients (Weinberg, 1995). Recent advances in sequencing technologies allowed the identification of genomic loci with recurring mutations and associations to tumor growth, suggesting new candidate oncogenes and tumor suppressors. To date, almost 600 genes were implicated in a variety of cancer types (Futreal et al., 2004).

Clinical applications of cancer drivers

Cancer driver genes provide means to characterize the molecular makeup of tumors, and most importantly, identify potential interventions to change the course of disease favorably. Targeting specific cancer genes which were altered in a patient's tumor provides an opportunity to induce selective killing of cancer cells without interfering with healthy tissue processes. For instance, overactivation of the human epidermal growth factor receptor 2 oncogene (HER2, coded by *ERBB2*) in breast cancer patients is responsible for regulating cell growth, differentiation and migration, leading to aggressive tumor growth and poor clinical prognosis (Citri and Yarden, 2006; Mendelsohn and Baselga, 2006). Therefore, HER2 is being used in clinical settings as diagnostic marker allowing to stratify patients and guide the appropriate therapeutic regimen (Stuhlmiller et al., 2015a). Additionally, the antibody *Trastuzumab* is being used to block the receptor and inhibit its function, thereby slowing down disease progression of $HER2^{+}$ patients. Thus, cancer genes are being used in clinical settings to either monitor or constrain the progression of oncogenesis.

Cancers develop resistance to therapeutic agents

Although drug targeting of oncogenes such as HER2 have shown promise in therapeutic settings and shrink tumor size considerably, patients typically develop resistance to therapy which allows tumors to relapse (Holohan et al., 2013). Resistance may originate intrinsically from mechanisms present before treatment or through mechanisms acquired during therapy. Intrinsic mechanisms involve the presence of a multiple tumor clones carrying a distinct mutational profile. While some are sensitive to the drug administration, others may be resistant, allowing them to continue proliferating (Greaves and Maley, 2012). Cancer stem cells (CSC) were suggested to be a tumor subpopulation with stem cell-like properties, providing the continuous ability for self-renewal and the capacity to repopulate tumors after drug administration (Al-Hajj et al., 2004; Butti et al., 2019). They are quiescent which

renders them resistant to chemotherapies which target cells undergoing rapid proliferation. CSC are thought to be an underrepresented subpopulation of tumor cells, suggesting that their genomic profiling by traditional techniques may be limited and thereby divert clinical efforts to target the majority of non-stem-like tumor cells.

An alternative theory suggests that the continuously changing mutational landscape of the tumor (Negrini et al., 2010) may allow tumor cells to compensate over the inhibition of a single oncogene by modulating the expression levels of other genes (Bertotti et al., 2015; Rathert et al., 2015), thereby indicating that resistance may be an acquired property of tumors. Although initial responses may be effective in reducing tumor size, resistance to therapy may render the treatment ineffective. For instance, targeting of HER2-positive breast cancers with *Trastuzumab* is typically bypassed by activating mutations in PIK3CA or activation of redundant signaling pathways including IGF1 and HER3. These observations exemplify the inherent limitation of independent studies of cancer genes without consideration of their vast network of functionally interacting genes, which may have an important role in determining therapeutic outcomes. In contrast, studies of cancer genes in the context of their interacting genes should provide the basis for rational drug targeting based on the molecular susceptibilities of individual patients.

Genetic interactions enable developing improved therapeutic strategies

Cellular functions are mediated by functionally interacting networks of genes. Functional genetic interactions (GIs), whereby the phenotypic effects of a gene's activity are modified by the activity of another gene, are thus a key to understanding complex diseases, including cancer, which involves an interplay among a myriad of genes (Ashworth et al., 2011; Jerby-Arnon et al., 2014; Kelley and Ideker; Lu et al., 2013; Wong et al., 2004; Zhong and Sternberg, 2006). This section explores the recent efforts in discovering genetic interactions and the challenges implicated in developing new methods for identifications of new types of genetic interactions.

Synthetic Lethality

In cancer genomics, three types of GIs have been studied so far showing major roles in disease progression and patient survival and suggesting novel therapeutic avenues. The vast majority of GI studies to date have focused on synthetic lethal (SL) gene pairs, describing the relationship between two genes whose individual inactivation results in a viable phenotype while their combined inactivation is lethal to the cell (Ashworth et al., 2011; Miyamoto et al., 2015; Sajesh et al., 2013; Stuhlmiller et al., 2015b). For instance, DNA repair is maintained by two distinct pathways, one that relies on PARP protein and the other which requires BRCA proteins (BRCA1 and BRCA2). While both pathways are active in healthy cells, some cancer cells lose the BRCA-dependent pathway and increase their dependence on the PARP-dependent pathway (**Figure 1.1**). Hence, disruption of PARP yields irreparable genomic instability and selective cancer cell death as only those cells experience dual inactivation of DNA repair pathways. This mechanism is being utilized in clinical settings by administrating PARP inhibitors (such as *Olaparib*) to BRCA-deficient patients and activating the lethal SL

interaction in cancer cells. Thus, inhibiting SL partners of a gene that is inactivated in a given tumor allows for selective killing of tumor cells (Ashworth et al., 2011; Jerby-Arnon et al., 2014; Kroll et al., 1996).

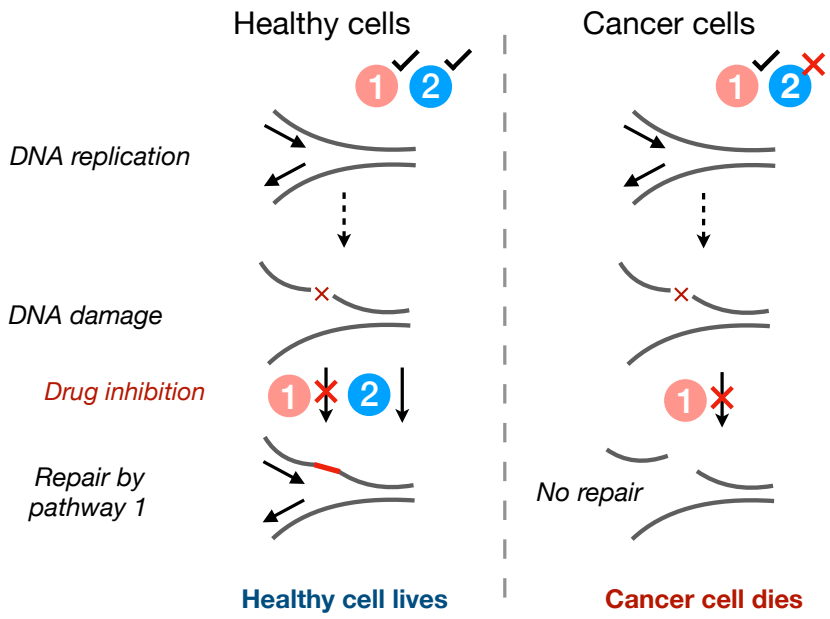


Figure 1.1: Utilizing DNA repair mechanisms for selective anti-tumor therapies. Both DNA repair pathway 1 (PARP-dependent) and 2 (BRCA-dependent) are active in healthy cells while pathway 2 is inactivated in some cancer cells due to a genetic alteration. Inhibiting pathway 1 across all cells activates a lethal SL interaction selectively in cancer cells.

Additional genetic interaction types

Another related class of GIs are Synthetic Dosage Lethal (SDL) interactions, where the underactivity of one gene together with the over-activity of another gene is lethal but not either event alone (Megchelenbrink et al., 2015; Stuhlmiller et al., 2015b; Szappanos et al., 2011). SDLs are promising for oncogenes, many of which are difficult to target directly, by targeting their SDL partners (Luo et al., 2009; Rathert et al., 2015). A third class of GIs are Synthetic Rescues (SR), where a change in the activity of one gene is lethal to the cell but an

alteration in its SR partner 'rescues' cell viability. SRs may play a key role in tumor relapse and emergence of resistance to therapy (Hartwell et al., 1997; McLornan et al., 2014). Indeed, previous investigations have shown that the overall numbers of functionally active SLs and SDLs in a given tumor sample are highly predictive of patient survival (Megchelenbrink et al., 2015).

Unbiased characterization of the landscape of genetic interactions

These three interaction types however represent just the 'tip of the GI iceberg', as there are many additional types of GI that can be defined at a conceptual level, and whose systematic exploration may have important functional ramifications for cancer therapy. To address this gap, we have developed a novel data-driven computational pipeline, called "EnGIne" (Encyclopedia of clinically significant GIs in cancer).

EnGIne identifies clinically relevant genetic interactions by associating molecular data with patients' survival. To enable the unbiased study of genetic interactions, EnGIne addresses multiple key limitations of large-scale survival analysis. Patient survival is a complex phenotype which depends on a wide range of genomic and environmental factors. Multiple clinical and demographic confounding factors contribute to variation in the molecular makeup of tumors, in patient survival and in the response to therapy. EnGIne models and controls for such confounders to identify molecular patterns which generalize across patient cohorts. Additionally, genome wide discovery of interactions between any pair of genes (overall 200 million pairs corresponding to 20,000 genes) is hampered by computational bottlenecks which EnGIne bypasses by incorporating a set of increasingly stringent filters. Finally, EnGIne's findings are shown to generalize to independent datasets, which demonstrates that this methodology is not over fitting dataset-specific features and can learn meaningful insights into the context-specific effects of genes involved in oncogenesis.

Immune responses to tumors

Multicellular organisms' ability to defend themselves from harmful invaders (pathogens) relies on the immune system which contributes to the elimination and control of a wide range of pathogens, including viruses, bacteria and fungi. The two arms of the immune response, the innate and the adaptive responses, act in concert to prevent pathogens from harming the host by complementary mechanisms (Gajewski et al., 2013; Zuniga et al., 2015).

The innate and adaptive immune system

Unlike the adaptive immune response, the innate response is non-specific to a particular invader, which allows for a fast reaction to pathogens in the scale of hours rather than days. It allows to recognize molecular patterns characteristic of invaders, destroy infected cells and activate immune cells of the adaptive immune response via dendritic cells. In contrast, the adaptive immune response provides the ability to tailor the immune response to the specific properties of new invaders. By creating an almost unlimited diversity of T and B cells recognizing different foreign molecules (antigens), the adaptive immune response is able to initiate a targeted response against any known or unknown pathogen.

T and B cell responses to foreign peptides

The adaptive immune response is divided into antibody responses and T cell mediated responses. The antibody responses are mediated by B cells which produce proteins (antibodies) that either block pathogens from infecting host cells or mark pathogens for destructions by the innate immune system (Pieper et al., 2013). T cell mediated responses provide a diverse range of other functions, both in providing help for other immune cells to mount effective responses as well as by direct killing of infected host cells (Halle et al., 2017; Krueger et al., 2017). T cells recognize foreign antigens which are presented on major

histocompatibility complex (MHC) found on the surface of host cells (Davis and Bjorkman, 1988). This mechanism allows the adaptive immune system to monitor the contents of host cells and discover those infected with proteins derived from foreign invaders. Although both T and B cells originate from hematopoietic stem cells in the bone marrow, T cells mature in the thymus. After maturation, naive T and B cells migrate to lymph nodes where they meet antigens and get activated upon meeting the antigen they are specific to (i.e., their cognate antigen), thus becoming effector cells (**Figure 1.2**). Effector cells will migrate to the inflammation site in order to eliminate or control the invaders.

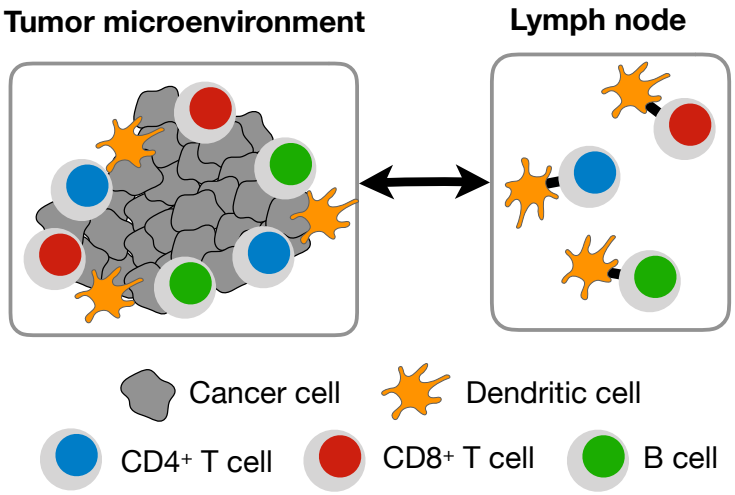


Figure 1.2: Lymphocyte are activated in the lymph nodes and migrate to the tumor microenvironment to control tumor growth. Dendritic cells carry antigens from the tumor site into the lymph nodes. After being activated by dendritic cells, $CD4^+$ and $CD8^+$ T cells, and B cells migrate to the tumor microenvironment to eliminate the tumor.

<u>$CD4^+$ T cell balance between inflammation and immunosuppression</u>

T cells are presented with antigens through two classes of MHC proteins which serve different functions. Cytotoxic $CD8^+$ T cells capable of killing infected cells are presented with antigens on MHC class I expressed on all nucleated cells. In contrast, $CD4^+$ T helper cells are activated by recognizing peptides presented on dendritic cells by MHC class II and

provide helper functions to control the magnitude of the immune response (Bluestone et al., 2009; Sakaguchi et al., 2008).

Upon activation, specific subsets of $CD4^+$ T cells induce inflammation by secreting pro-inflammatory molecules (such as IFN-γ) (Qin and Blankenstein, 2000). Conversely, regulatory $CD4^+$ T cells suppress the immune response (immunosuppression) to prevent the immune response from excessive inflammation which may be destructive for healthy tissues (Sakaguchi et al., 2008). Thus, $CD4^+$ T cells have an essential role in modulating the immune response and controlling the delicate balance between inflammation to immunosuppression.

$CD4^+$ and $CD8^+$ T cells protect against cancer

The mutations acquired by cancer cells result in the generation of molecules that may appear foreign to T cells (neo-antigens), which may in turn activate an anti-tumor response. T cells recognizing tumor neo-antigens were shown to provide effective means to fight cancer (**Figure 1.3**). However, immune responses are typically dampened due to the induction of immunosuppressive microenvironment driven by the activation of inhibitory receptors such as Programmed Cell Death 1 (*Pdcd1* which codes for PD-1). Consequently, novel therapies aim to unleash effective anti-tumor immune responses by blocking inhibitory receptors (Borst et al., 2018; Gajewski et al., 2013; Ribas and Wolchok, 2018; Rosenberg and Restifo, 2015; Wei et al., 2017).

$CD4^+$ T cell diversity limit clinical progress

Cytotoxic $CD8^+$ T lymphocytes are being exploited in clinical settings due to their ability to recognize tumor neo-antigens and kill cancer cells (Ott et al., 2017; Rosenberg and Restifo, 2015). However, effective anti-tumor immunity relies on a complex interplay between diverse lymphocyte subsets that remain poorly characterized. $CD4^+$ T helper cells, which are essential for effective immune responses and control the balance between inflammation and

immunosuppression (Bluestone et al., 2009; Borst et al., 2018; Sakaguchi et al., 2008; Zhu et al., 2010), have recently emerged as potential therapeutic targets (Aarntzen et al., 2013; Borst et al., 2018; Hunder et al., 2008; Malandro et al., 2016; Mumberg et al., 1999; Ott et al., 2017; Tran et al., 2014; Wei et al., 2017).

$CD4^+$ helper cells contribute to the priming of $CD8^+$ T cells and to B cell functions in lymphoid organs (Ahrends et al., 2017; Borst et al., 2018; Crotty, 2015). $CD4^+$ T helper type-1 (Th1) cells secrete the cytokine IFN-γ and affect tumor growth by targeting the tumor microenvironment (TME), antigen presentation through MHC class I and II, and other immune cells (Alspach et al., 2018; Beatty and Paterson, 2001; Bos and Sherman, 2010; Kammertoens et al., 2017; Qin and Blankenstein, 2000; Tian et al., 2017). Conversely, Th2 cells can promote tumor progression and regulatory T cells (Treg) mediate immune tolerance, suppressing the function of other immune cells and thus preventing ongoing anti-tumor immunity (Chao and Savage, 2018; DeNardo et al., 2009; Tanaka and Sakaguchi, 2017).

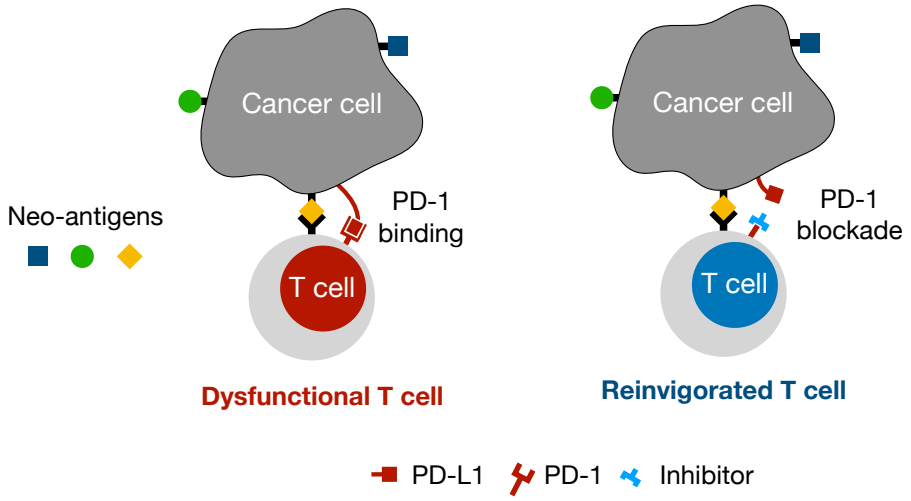


Figure 1.3: Antigen-specific T cell responses to tumor neo-antigens. $CD4^+$ and $CD8^+$ T cells perform various functions upon encountering their cognate antigen on tumor cells. Binding of PD-L1 to PD-1 renders T cells dysfunctional which prevents successful anti-tumor responses. PD-1 blockade reinvigorates T cell function to elicit an effective immune response.

Challenges in characterizing T cell diversity

Despite the anti-tumor potential of $CD4^+$ T cells, disentangling their functional diversity has been the limiting factor for pre-clinical and clinical progress. Investigations of $CD4^+$ T cell diversity leveraged the basic notion where distinct cell types harbor distinct molecular compositions. Thus, identifying subsets of cells with unique molecular makeup has the potential to identify novel cell types.

While several studies have assessed the diversity of Treg cells or their specificity to tumor antigens (tumor-reactivity) (Ahmadzadeh et al., 2019; Chao and Savage, 2018 ; De Simone et al., 2016; Malchow et al., 2013; Plitas et al., 2016; Zhang et al., 2018; Zheng et al., 2017a), the functional diversity of conventional (non-Treg) tumor-infiltrating lymphocytes (TILs) has remained poorly understood. Studies measuring average gene expression across a bulk of cells (bulk RNA sequencing) have limited power at identifying new, and especially rare functional cell states. Conventional single-cell approaches (e.g. flow or mass cytometry) overcome this obstacle by providing protein-abundance measurements at the single cell level, but are necessarily restricted to a limited number of hypothesis-based targets they can analyze, thus limit the unbiased identification of novel subpopulations.

Recent studies, whether of human or in experimental tumors, have leveraged a genome-wide RNA sequencing technique at the single cell level (scRNAseq) to perform unbiased clustering analysis of thousands of cells to identify novel subpopulations. However, these studies did not distinguish tumor antigen-specific from bystander $CD4^+$ T cells, even though bystanders may form the vast majority of conventional (non-Treg) T cells in the TME (Ahmadzadeh et al., 2019; Azizi et al., 2018; Duhen et al., 2018; Sade-Feldman et al., 2018; Simoni et al., 2018; Zhang et al., 2018; Zheng et al., 2017a), in particular in draining lymphoid organs, where immune responses are typically initiated.

To address these challenges, we applied the resolution of scRNAseq to a tractable experimental system assessing tumor-specific responses both in the tumor and in lymphoid organs. While scRNAseq provides an unprecedented opportunity to investigate cellular diversity, the platform suffers from low signal to noise ratio. We designed computational analyses to characterize cell subpopulations in high resolution using a data-driven approach to determine the appropriate resolution while controlling for false discovery of subpopulations. To compare the transcriptomes of the identified subpopulations across distinct tissue-origins and experimental designs, we develop a strategy to mitigate context specific effects while retaining important signal originating from cell lineage-defining biological factors. These strategies enabled us to find novel TIL subpopulations and correspond them to multiple preclinical and clinical datasets.

Our analyses dissect the complexity of the $CD4^+$ T cell response to tumor antigens, both in the tumor itself and in draining lymphoid organs, and identify broad transcriptomic divergences between anti-tumor and anti-viral responses. Emphasizing the power of this approach, new transcriptomic patterns identified in the present study are also found in $CD4^+$ T cells infiltrating human tumors and correlate with response to checkpoint therapy in human melanoma.

Chapter 2: Discovery of multi-type genetic interactions in cancer

Functional genetic interactions (GIs) underlie the complexity of cellular phenotypes and disease. Synthetic lethality provided opportunity to develop selective killing of cancer cells (Kaelin, 2005), but represented only a small subset of the spectrum of genetic interactions. This chapter describes our novel methodology to characterize the diversity of genetic interactions using the EnGIne pipeline. EnGIne substantively expands the current knowledge of genetic interactions in cancer, laying a strong conceptual and computational foundation for future studies of additional GI types.

Pancancer identification of putative genetic interactions

We applied EnGine to analyze 5,157 TCGA (Weinstein et al., 2013) samples of 18 different cancer types, identifying clinically significant GIs of 12 distinct types. Using drug response data from TCGA (Weinstein et al., 2013) and molecular drug target information, we show that the detected GIs are associated with response to therapy by specific drugs. Their activation patterns can account for the tissue-specificity of known driver genes and stratify breast cancer into clinically relevant subtypes.

Overview of the Encyclopedia of Genetic Interactions (EnGIne) Pipeline

The overall EnGine pipeline is summarized in **Figure 2.1** and the technical details are provided in the Methods section. Given a large set of tumor transcriptomes (**Figure 2.1A**), we first partition the expression level of each gene into low, medium and high, following our previous approach to identify SL interactions (Lee et al., 2018). Thus, for a pair of genes, there are $9 = 3 \times 3$ combinations, or bins, of possible co-activity states for the two genes

(**Figure 2.1B**). For a given ordered pair of genes, each tumor sample maps to exactly one of the 9 bins. Our goal is to identify GI pairs of the form $(x, y, b, \pm\alpha)$ such that for the specific gene pair (x, y), the tumors in which the joint activity of (x, y) maps to bin b have a significant fitness advantage (+) or disadvantage (-) with effect size α, relative to all other tumors whose activity of (x, y) maps to a different bin. The effect size α is estimated by measuring the difference in the survival curves between those patients where the activity of (x, y) is in bin b in their tumors and those where it is not, as depicted in **Figure 2.1C**; note that for most gene pairs, there may not be any bin exhibiting a significant fitness differential. A significant GI pair $(x, y, b, \pm\alpha)$ is termed *functionally active* in a particular tumor if the co-activity states of (x, y) in that tumor fall in bin b. We hypothesized that the patients whose tumor has a larger number of functionally active interactions with negative tumor fitness effects will have better prognosis and conversely, the patients whose tumor has a larger number of functionally active interactions with positive tumor fitness effects will have poorer prognosis.

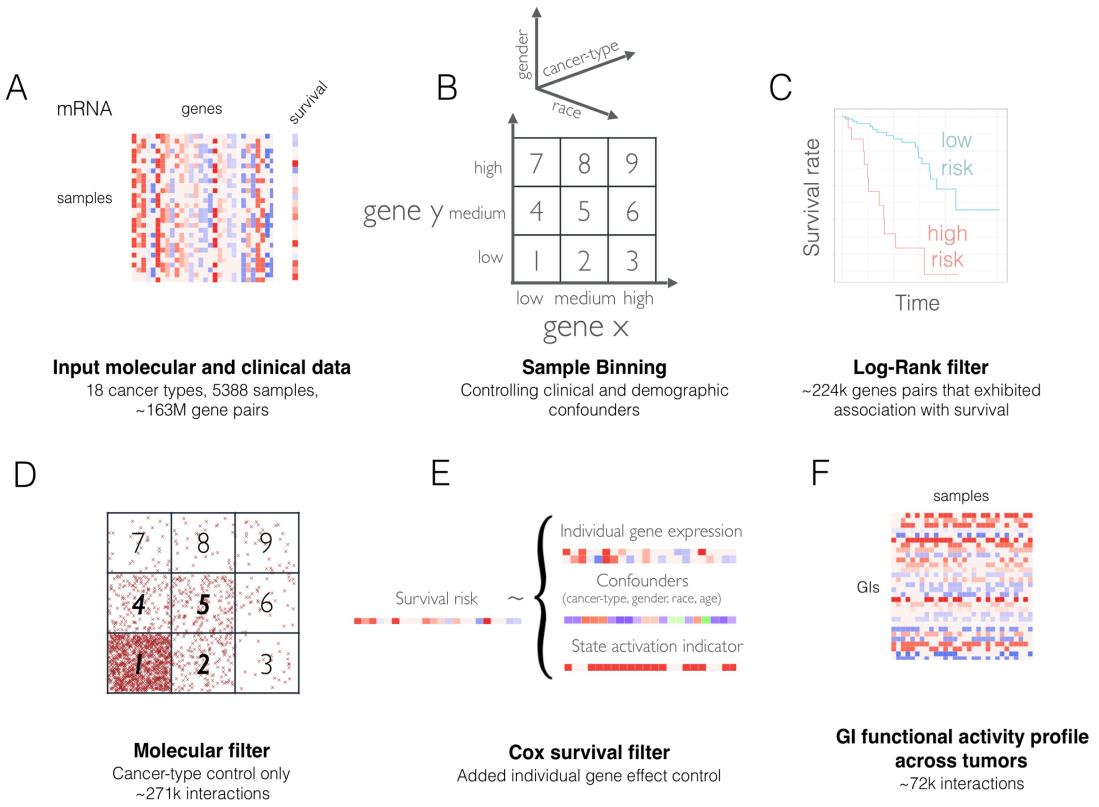


Figure 2.1: Overview of the EnGIne pipeline.

Given a large set of tumor transcriptomes **(A)**, we first partition the expression level of each gene into low, medium, and high activity state, resulting in 9 joint activity state bins for any two genes **(B)**. Each combination of a gene pair and bin *b* induces a bipartition of the set of tumor samples based on whether the co-activity levels of the gene pair in a specific tumor is in bin *b*. The first step of EnGIne screens for the gene pairs that show distinct survival trends in the two sets of tumors in any of the bins, based on log-rank test **(C)**. Next, for a gene pair and a bin identified in **(C)**, we test whether the putative gene interaction in bin *b* has a differential effect on tumor fitness, by testing for depletion or enrichment of samples in the bin *b* relative to expectation based on individual genes **(D)**. Finally, for each retained gene pair, in each of the 9 bins separately, we fit a Cox proportional hazards model to assess whether being in a particular bin is associated with a distinct (positive or negative) pattern of patient survival, followed by correction for multiple hypotheses testing **(E)**. The output of EnGIne is **(i)** a list of GIs of each of the 12 types studied, and **(ii)** GI profile in each of the individual tumor samples, defined as activity state of each GI in the tumor sample **(F)**.

Step-wise filtering of multi-type genetic interaction candidates

We analyzed 5,157 TCGA (Weinstein et al., 2013) samples for 18 cancer types. First, as an initial screening, we performed a Log-Rank survival test (depicted in **Figure 2.1C**) for each gene pair in each of the 9 bins. To make this computationally feasible and to limit the burden of multiple testing correction in the subsequent steps, we used an extremely stringent cutoff for the log rank test leading to the retention of about 1/1,000 gene pairs surveyed, resulting in 223,946 gene pairs that exhibit a significant association with survival in one of the 9 bins. Second, if a potential GI in bin *b* has a differential effect on tumor fitness, we expect the number of tumors that map to bin *b* to be relatively enriched (for a '+' interaction positively affecting tumor survival), or depleted (for a '-' interaction negatively affecting tumor survival). Thus, we applied an additional filter (**Figure 2.1D**) to retain the GIs exhibiting a consistent patient survival and tumor fitness enrichment or depletion statistic, yielding 179,444 gene pairs. Third, for each retained gene pair, in each of the 9 bins, we implemented a Cox proportional hazards model, specifically controlling for age, cancer-type, gender, and race, to assess whether a tumor being in a particular bin is associated with patient survival, either positively or negatively (**Figure 2.1E**). Finally, we applied an empirical False Discovery Rate (FDR) correction based on the significance of the Cox interaction term of the 179,444 gene pairs relative to those obtained for randomly shuffled gene pairing as the null control.

At a False Discovery Rate (FDR) < 1%, this resulted in 71,946 predicted GIs across the 9 bins, of the form $(x, y, b, \pm\alpha)$, which form the final set of TCGA (Weinstein et al., 2013) inferred GIs (**Figure 2.1E**). Considering the symmetry among bins (bin 2 ~ bin 4 corresponding to low-medium expression interaction; bin 3 ~ bin 7 corresponding to low-high expression interaction; bin 6 ~ bin 8 corresponding to medium-high expression interaction), there are 6 unique types of interaction bins, and considering the two directions of

the effect size yields a total of 12 basic types of GIs. We ascertained the robustness of the pipeline to changes in the quantile boundaries for the 3×3 bins and to changes in the log-rank and FDR thresholds (**Appendix A Extended Results 1**).

The landscape of multi-type GIs

EnGIne identified 71,946 clinically significant GIs of 12 different types, ~0.02% of all the possible candidate gene pairs and GI types tested. To gain insight into their biological significance, we analyzed their correspondence to known interactions and known cancer genes, and whether the observed effects generalize to additional datasets.

PIN-supported GIs are enriched with cancer genes

Considering the expectation that neighboring genes in the protein interaction network (PIN) are more likely to be involved in a GI (Schaefer et al., 2012), to obtain a smaller but more biologically grounded PIN-supported GI network, we retained only the gene pairs that are separated by two or fewer edges in the PIN. This PIN-supported GI network was composed of 1704 GIs involving 1786 genes (**Table A.1**) that included 133 known cancer genes (Cosmic dataset) (Futreal et al., 2004) associated with various cancer types (enrichment p-value = 2.5×10^{-22}) and 50 breast cancer specific (Intogen dataset) (Gonzalez-Perez et al., 2013) driver genes (enrichment p-value = 7.7×10^{-13}).

Unbiased GI identification reveals unexpected interaction type diversity

The distribution of the detected 1704 PIN-supported GIs across the 12 GI types reveals that previously characterized interactions may represent only a small fraction of the overall interaction landscape (**Figure 2.2A**). SL interactions are surprisingly one of the least abundant

types of identified GIs, and so are SDLs (1% of all GIs). Remarkably, the positive "anti-symmetric" type of SLs, in which the joint low activity of the two interacting genes is associated with a higher tumor fitness, is 3 times more abundant than SLs.

The interaction between the Cosmic Cancer Census genes *GNAQ* and *JAK2* is one example of such a positive interaction in bin 1 (**Figure A.1**). *GNAQ*, encoding Gq, and *JAK2* are both downstream targets in a signaling pathway with several functions pertinent to cancer, including endothelial cell maintenance and vascular remodeling (Kawai et al., 2017). Interestingly, the two most abundant types of pan-cancer GIs correspond to bin 2 and bin 6) where one of the genes has medium level of activity and only the extreme activity of its partner gene reveals a phenotypic effect. For most GI bins, we see a higher proportion of GIs exerting a positive effect on tumor fitness, consistent with the hypothesis that the GIs uncovered during the evolution of cancer are under positive selection. The above distribution trends are quite similar for the full 71,946 GI network (**Figure A.2A**). Additionally, we ascertained that the inferred GIs are not dominated by correlated gene expression patterns (**Appendix A Extended Results 2**).

Context specific effects of cancer drivers

Cancer genes that encode transcription factors, such as *MYC* and *KLF4*, have proven difficult to target directly (Lambert et al., 2018; Li et al., 2018). One important application of EnGIne is to identify candidate interaction partners of the difficult-to-target cancer genes for indirect interventions. To assess this capability, we identified the GI partners of several cancer genes using target-specific FDR. **Figure 2.2B** shows survival patterns for different activity state combinations of breast tumor suppressor *ERCC2*, a transcription-coupled DNA excision repair gene (Benhamou and Sarasin, 2002; Bernard-Gallon et al., 2008), and a breast cancer oncogene *KLF4*, a zinc finger transcription factor (Akaogi et al., 2009). It reveals two interesting trends: As expected, the over-activation of the oncogene and under-activation of the tumor suppressor

(bin 3) results in poorer patient survival than expected from the individual gene effects (bins, 1, 2, 6, and 9). However, surprisingly, the survival curve reveals a reversal of the effect of the tumor suppressor *ERCC2* inactivity on survival when the oncogene *KLF4* has medium activity (bin 2), whose individual activity is associated with better survival; the (*ERCC2, KLF4*) interaction exemplifies the relevance of medium expression bins in this study.

This and several other examples of GIs involving a cancer driver gene (**Supplementary File 1**) demonstrate that the context-specific effects of driver genes may show very different trends than their previously established effects as individual genes. **Figure A.3** shows the extended GI-network (71,946 GIs prior to PIN filtering) involving the Cosmic and Intogen driver genes. In addition, and consistent with *MYC*'s role as an oncogene, the GIs occurring when *MYC* has low activity mostly have negative effect on tumor fitness. However, when *MYC* is activated, we find that the low expression of *PUF60,* one of the known regulators of *MYC* (Matsushita et al., 2014; Rahmutulla et al., 2013), is associated with higher tumor fitness (type +3 interaction, HR = 1.37, p-value = 5.0E-04) (**Figure 2.2D**). In contrast, we find that high expression of *MYC* does not significantly contribute to poorer prognosis when *PUF60* is expressed at medium or high levels (p-value = 0.9). Thus, this result underscores the importance of molecular context in developing anti-MYC treatments.

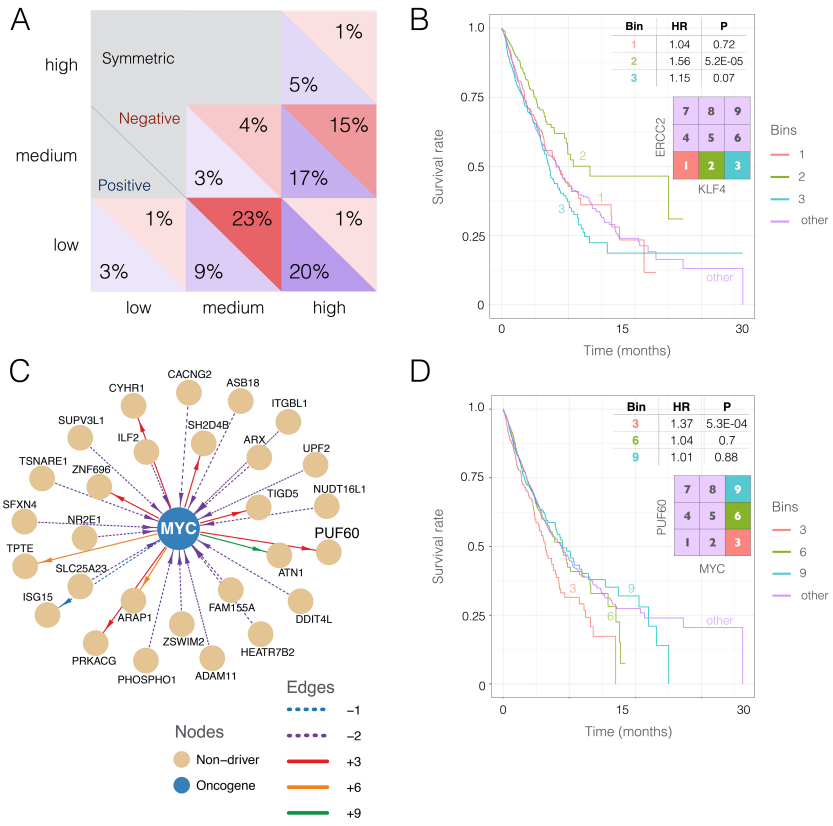


Figure 2.2: Broad distribution and characteristics of the detected GIs and context-specific effect of cancer driver genes on survival.

(A) Distribution of the 1704 significant PIN-supported GIs across 9 joint activity bins. The fractions of GIs in each bin are shown for GIs with positive (blue) and negative (red) effect on tumor fitness. Only the data in the lower triangle of the matrix are shown as the GIs are symmetric relative to the genes in a pair. **(B)** The Kaplan-Meier (KM) survival curve of GI involving *ERCC2*, a transcription-coupled DNA excision repair gene, known to be a breast cancer tumor suppressor, and *KLF4*, a zinc finger transcription factor known to be oncogenic in breast cancer, reveals increasingly poor survival by over-activation of the oncogene and under-activation of the tumor suppressor (bin 3). Strikingly, the effect of *ERCC2* inactivity on survival is reversed when *KLF4* has medium activity level (bin 2). **(C)** The predicted GIs involving the oncogene *MYC*. **(D)** KM survival curve of GI involving *MYC* and its regulator *PUF60*. High expression of *MYC* is associated with poor prognosis specifically at low activation of *PUF60*.

Validations of EnGIne-identified GIs

We validated EnGIne by comparing its SL predictions to previously reported SLs identified via large *in vitro* screens (Bommi-Reddy et al., 2008; Lord et al., 2010; Luo et al., 2009; Steckel et al., 2012; Turner et al., 2008). Each of the three filtering steps of EnGIne (**Figure 2.1C-E**), could discriminate the experimentally determined SLs from the non-SLs, with ROC-AUCs of 0.63 (p-value = 0.0005), 0.62 (p-value = 0.001), and 0.59 (p-value = 0.012), respectively. These results are significant, albeit of modest accuracy (reflecting the widely known discrepancy between in-vitro and in-vivo data (Williams et al., 2000)), support the contribution of each of the individual steps in EnGIne.

In addition, we find the PIN-supported GIs to be predictive of patient survival both in cross-validation setting in TCGA (Weinstein et al., 2013) as well as in an independent breast cancer METABRIC (Curtis et al., 2012a) dataset (Curtis et al., 2012b) (**Figure A.4A, Appendix A Extended Results 3**). The prediction accuracy quantified via the concordance index (CI) show that GI based prediction compares favorably with the gene-wise approach. A bigger improvement is observed in the independent METABRIC (Curtis et al., 2012a) dataset (concordance $\approx$ 0.64), testifying that the GI-based approach is generalizable, while the individual gene-based approach fails to generalize (concordance $\approx$ 0.51). **Figure A.4B** depicts the survival prediction accuracy of each GI type. Interactions involving both genes in their wild type mid-activity levels (i.e. bin 5) have negligible predictive power on survival, testifying that more extreme levels of expression of at least one of the two genes tend to be involved in functional GIs affecting survival. In addition, we have excluded the possibility that additional confounders may impact our results (**Appendix A Extended Results 4**).

GIs activation is associated with drug response

To establish the prognostic value of the GIs inferred by EnGIne, we have associated the GI activation scores to patient drug response. By estimating the activation of gene interactions involving target genes of each drug, we demonstrate the GI significance for developing precision drug administration strategies.

Assessment of GIs activation across responders and non-responders

To avoid circularity, we applied EnGIne to identify GIs based only on TCGA (Weinstein et al., 2013) samples that do not have drug response information, and tested the predicted GIs' ability to discriminate responders from non-responders in the 'unseen' TCGA (Weinstein et al., 2013) samples where the drug response information is available. Notably, because the considered drugs are inhibitory, it suffices to focus on GI bins 1, 2, and 3, where one of the genes (the drug's target) has low activity. For a given drug and cancer type having data on responders and non-responders, we analyzed the GIs involving each of the drug targets (identified via target-specific FDR).

We then tested whether the frequencies of GI activation in responders and the non-responders are significantly different using a Fisher exact test (Fisher, 1922). For positive GIs, we expect a lower GI activation frequency among responders and the opposite for negative GIs (e.g., as in the case of SL-type GIs). However, owing to very small and unbalanced numbers of responders and non-responders (5 to 35 samples per response group per drug), the Fisher test is underpowered, and we therefore tested whether the overall distribution of the obtained ratio of GI activation frequency in responders and non-responders are lower than those obtained using randomly shuffled drug-response labels using paired Wilcoxon tests (Wilcoxon, 1945), performed separately for each drug-cancer type pair.

Differential drug target GI activation between responders and non-responders

We considered the 12 drug-cancer type pairs that have RECIST (Response Evaluation Criteria In Solid Tumors) (Eisenhauer et al., 2009) drug response following treatment for at least 10 patients (at least 5 responders and 5 non-responders) in TCGA (Weinstein et al., 2013). Each of the 6 basic GI types was tested for the 12 drug-cancer type pair. Overall, in 18 of the 72 tests (5 fold enrichment for $P \leq 0.05$) of drug-cancer type combinations, GIs of a particular type exhibit statistically significant differential activation frequencies between responders and non-responders consistent with the expected effects of the GIs (**Figure 2.3A**). Reassuringly, several of those significant drug-target GI's are in bin 1, which contains the SLs, consistent with previous reports showing the role of SLs in mediating drug response (Jerby-Arnon et al., 2014; Lee et al., 2018). Among the drugs, Gemcitabine, Lomustine, and Paclitaxel exhibit differential GI activation for most GI types (aggregate p-values ranging from 4×10^{-16} to 2×10^{-11}).

We also explored the most differentially activated individual GIs. Imposed an empirical FDR threshold of 0.01 on the Fisher test p-value yielded 521 GIs for the 12 drug-cancer type combinations (**Table A.2**). Individual genes comprising the 521 GIs are closer to each other in the PPI network relatively to shuffled pairs (Wilcoxon p-value < 0.001, Methods) and have a significantly increased number of direct PPI interactions between them (Fisher $P < 0.02$, Methods).

Context-specific effects of Paclitaxel-mediated BCL2 inactivation

As an illustrative test case, we explored the GIs associated with the response to Paclitaxel, in TCGA (Weinstein et al., 2013) Head and Neck Squamous Cell Carcinoma (HNSC) cohort. Paclitaxel inhibits the proteins encoded by *BCL2*, *TUBB1*, and *MAP* based on DrugBank (Law et al., 2014). We identified a GI involving the inactivation of *BCL2* (known to suppress apoptosis, indirectly inhibited through phosphorylation (Ruvolo et al., 2001)) and the over-activation of *ITPR1* (Inositol 1,4,5-trisphosphate receptor type 1, also known as IP3 receptor

type 1), negatively affecting tumor fitness (GI type -3). Interestingly, our analysis shows that this GI is functionally active among the responders at a significantly higher ratio than among non-responders (odds-ratio ≈ 11.1).

The interaction between ITPR1 and BCL2 is well characterized (Chen et al., 2004; Oakes et al., 2005; Rong et al., 2009); one of these studies suggests that BCL2 also interacts with the two other human paralogs ITPR2 and ITPR3, but these interactions are not represented in the PIN used in this study and were therefore not detected. BCL2 exerts its oncogenic effect by inhibiting ITPR3-mediated channel opening and Ca^{2+} release from the endoplasmic reticulum, and thus preventing cancer cell apoptosis. Our analysis strongly suggests that BCL2 inhibition by Paclitaxel is especially effective when the ITPR1 expression is abundant, enabling effective Ca^{2+} release. Additional Paclitaxel targets *TUBB1* and *MAP2* are also linked with ITPR1/BCL2 through GIs with literature evidence for experimentally validated or putative interactions (by STRING DB (Szklarczyk et al., 2015)) (**Figure 2.3B**), suggesting promising avenues for additional studies.

GIs explain cancer driver genes' tissue-specificity

Many of the known cancer driver genes affect tumor initiation and development in a tissue-specific manner, despite the cancer gene being expressed in other tissues as well. Next, we explored whether the GIs can explain the tissue-specificity of cancer genes.

Assessment of GI activation across tissue-specific driver genes

We identified 15 oncogenes and 20 tumor suppressors whose effects are likely to be restricted to specific cancer types, based on preferentially high mutation rates in those cancer types,

including breast, bladder, and gastric cancer (**Table A.3**). For each cancer driver, we assigned a risk score to each patient by aggregating functionally active GIs involving the driver gene defined using target-specific FDR; for oncogenes, only the bins with high oncogene activity and for tumor suppressors, only the bins with low tumor suppressor activity were considered. We hypothesized that for a cancer gene, the risk score will be greater in tissues where the cancer gene is implicated relative to other tissues. Indeed, for 15 out of 35 (~43%; 5 oncogenes and 10 tumor suppressors) driver genes, the observations are consistent with our hypothesis (Wilcoxon rank-sum test, FDR < 0.1, **Table A.3**).

HLF lung- and breast-specificity is captured in GI network activity

HLF, a bZIP transcription factor, has been linked to lung and breast cancer based on its significantly greater missense mutation frequency in those cancer types (Gonzalez-Perez et al., 2013). We observed a significant difference (FDR < 1.07×10^{-12}) in GI activation risk score for breast and lung cancer relative to the other tissues. Specifically, we found that positive GIs are preferentially activated in these two tissues while negative GIs are preferentially activated in the other tissues, consistent with the increased tumor fitness in these two foreground tissues (**Figure 2.3C,D**). Overall, these results suggest that cancer type-specific effects of many driver genes may be explained by their tissue-specific GI network activity.

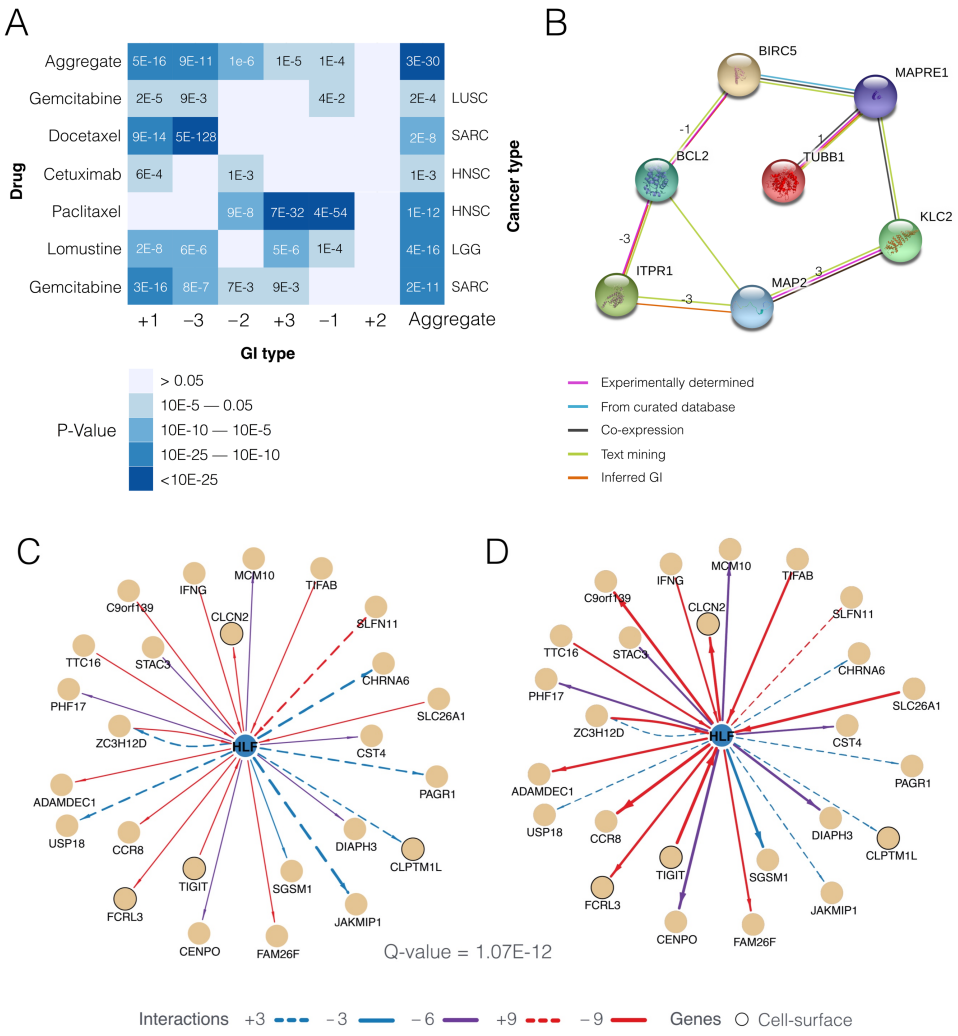


Figure 2.3: Differential GI activation between drug response groups and tissues. **(A-B) Drug response analysis. (A)** For each drug (left row labels) and each cancer type (right row label) combination, and for each GI type (columns), the heat plot shows the significance of differential activation of GIs in responders and non-responders consistent with expectation. The last column shows the significance when all GI types are aggregated. **(B)** The network shows the inferred functional interactions (based on STRING (Szklarczyk et al., 2015)) among the genes interacting with *Paclitaxel* targets, as well as inferred GI types. **(C-D) Tissues specificity analysis.** For the HLF-specific GI network, the figure shows the activity states of GIs in breast and lung cancers **(C - foreground tissues)** and in other cancer types **(D - background tissues)**. The edge weight (thickness) represents the fraction of samples in which the GI was functionally active. Several GIs are differentially active in the two sets of cancer types. The figure also depicts cell surface proteins among the HLF's GI partners. The GI network-based sample-specific risk score is significantly higher (q-value < 1.07×10^{-12}) in breast and lung cancer relative to other cancer types, potentially mediated by a selective activation of positive GIs in the foreground tissues and negative GIs in the background tissues.

GIs have potential prognostic implications in breast cancer

We investigated whether functionally active pan-cancer GIs in a tumor may provide an alternative methodology to tumor stratification into sub-types. We focus on breast cancer because it has a large number of samples in TCGA (Weinstein et al., 2013) and because a second independent dataset, METABRIC (Curtis et al., 2012a), is publicly available.

Stratifying breast cancer tumors based on their GI profiles

To obtain a GI-based representation of each breast cancer sample, we quantified the functional activity (a binary indicator) of each PIN-supported GI detected in TCGA (Weinstein et al., 2013), rather than generating BRCA-specific network, thus avoiding potential circularity of inference and prediction within samples sharing similar characteristics. Based on this 1704-dimensional binary vector representation of each tumor sample we clustered the 1981 breast cancer samples in the independent METABRIC (Curtis et al., 2012a) dataset using a conventional Non-Negative Matrix Factorization (NMF). Optimal clustering was achieved (maximum value of the Dunn index, Methods) for 10 clusters (**Figure A.7A**). Upon closer inspection of the distributions of known breast cancer subtypes in these clusters (**Figure A.7B**) we merged two of the clusters, thus yielding 9 clusters for further analyses.

Kaplan-Meier curves (**Figure 2.4A**) and statistical analysis show that the 9 clusters have distinct survival characteristics with an overall mean hazard ratio (HR) difference of 1.94 (p-value below the lowest reportable threshold and shown as 0). The distinct survival characteristics are consistent with analysis performed using the full 71,946 GI network (**Figure A.7C**). **Figure A.8A** shows the survival characteristics obtained for the previously published clustering of the METABRIC (Curtis et al., 2012a) samples. As evident, both approaches obtain similar survival separation levels, but exhibit differences in their histopathological

composition. Currently, breast cancer has 5 well-established clinically distinct subtypes based on the tumors' histopathological attributes. **Figure 2.4B** shows the fractions of each known subtypes among the 9 GI-based clusters. Several clusters are highly associated with specific subtypes such as Basal [triple-negative] (cluster 5), Luminal A (clusters 3,4) etc. Others show association with several subtypes, e.g., Luminal A and B both have high fractions in cluster 8. Interestingly, the basal subtype, which are largely triple-negative and have poor prognosis, correspond to a distinct cluster in our analysis (cluster 5), consistent with their distinct clinical status. In the original METABRIC (Curtis et al., 2012a) publication, 50% of the samples were left unassigned to any of their 10 clusters, while our GI-based clustering covers all samples.

GI-based stratification provides improved predictive value

We find the GI based approach to provide improved survival predictive value over the classical histopathological ones (**Figure 2.4C,D**). There are two situations in which the GI-based clustering leads to a different classification of patients for survival analysis: (1) cases where known histopathological breast cancer subtypes are split across multiple GI-based clusters (e.g. Luminal B across clusters 1, 2 and 8), and conversely (2), cases where one GI-based cluster harbors multiple known histopathological subtypes (e.g. cluster 2 contains Her-2 and Luminal B subtypes). In the former case, we find that the 1989 Luminal B tumors that are split across different GI-based clusters exhibit statistically significant ($P < 7.14\times10^{-6}$) distinct survival trends (**Figure 2.4C**), supporting the GI approach in separating the Luminal B tumors. Likewise, in the latter case, we find that the survival trends of Her-2 and Luminal B histopathological subtype samples that are assigned to the same GI-based cluster 8 do not show a significant difference ($P < 0.203$) in their survival trends (**Figure 2.4D**), suggesting that the GI-based stratification may in some instances provide better survival prognosis relative to histopathology-based stratification.

We systematically identified 6 additional instances of the above two scenarios where (1) a known tumor histopathological subtype was split across multiple GI-based clusters or (2) multiple known histopathological subtypes were assigned to the same GI-based cluster (and each cluster has at least 30 samples). In each instance of the first kind we tested for statistically significant differences in survival and in each instance of the second kind we tested for lack thereof. As shown in **Figure A.9**, in 5 out of 6 instances we found that the GI-based clusters provided a more accurate survival prognosis. In contrast, we identified 4 cases of the second kind in the original METABRIC (Curtis et al., 2012a) clusters and found that none of their survival trends were significantly different (**Figure A.10**). Thus, these results demonstrate that the GI approach performs better than clustering based on histopathological subtypes or METABRIC (Curtis et al., 2012a) clustering based on gene expression profiles, in terms of survival prognosis.

GI-based clusters are characterized by distinct mutational profiles and GI types

To explore potential mutational basis of the GI-based clusters in another way, we assessed whether the samples in GI-based clusters harbor distinct mutations patterns. We identified 196 genes (**Table A.4**) with significantly greater mutation frequency in one or more of the clusters, relative to their overall mutation frequency in breast cancer. **Figure 2.4E** shows the mutational frequency profiles of these genes across the 9 clusters. Overall, the differentially mutated genes across the GI based clusters include 10 cancer drivers: *CDK12, CDKN1B, DNAJB1, ERBB2, EXT2, FCGR2B, FNBP1, HOXC13, PDGFRB,* and *SEC24D*. A more detailed discussion of the potential biological significance of some of these mutations is provided in the Supplementary Results. We note that no mutation data was used in the GIs inference via EnGIne.

Additionally, we quantified the fraction of each of the 12 types of functionally active GIs among the samples in each cluster. **Figure 2.4F** shows the active GI profiles of each cluster,

and reveals two broad subgroups, one including clusters 4, 3, 8, and 1 and another including clusters 2, 5, 6, and 7. Interestingly, the two subgroups clearly segregate in terms of their survival, testifying that the classification into GI types captures a simplified yet robust characterization of the clinical prognosis. The first broad subgroup of tumors (clusters 4, 3, 8 and 1) are characterized by high fractions of type +2 and +6 GIs, both of which involve a medium expression and low expression bin. Therefore, this analysis demonstrates the relevance of considering medium expression states in molecular stratification. **Figure A.11** compares the GI-profiles of clusters revealed in the TCGA (Weinstein et al., 2013) and the METABRIC (Curtis et al., 2012a) breast cancer data and shows a high degree of consistency. A global comparison of the GI profiles of the 9 clusters in the two datasets shows a Spearman correlation of 0.67 (p-value = 2.4×10^{-14}) between the GI types composition of these clusters, implying that GI-profiles are a robust characteristic of breast cancer tumors across different tumor collections. Thus, the GI-based clustering demonstrates a proof of principle for improved stratification of breast cancer tumors into classes with distinct survival prognosis and mutational profiles.

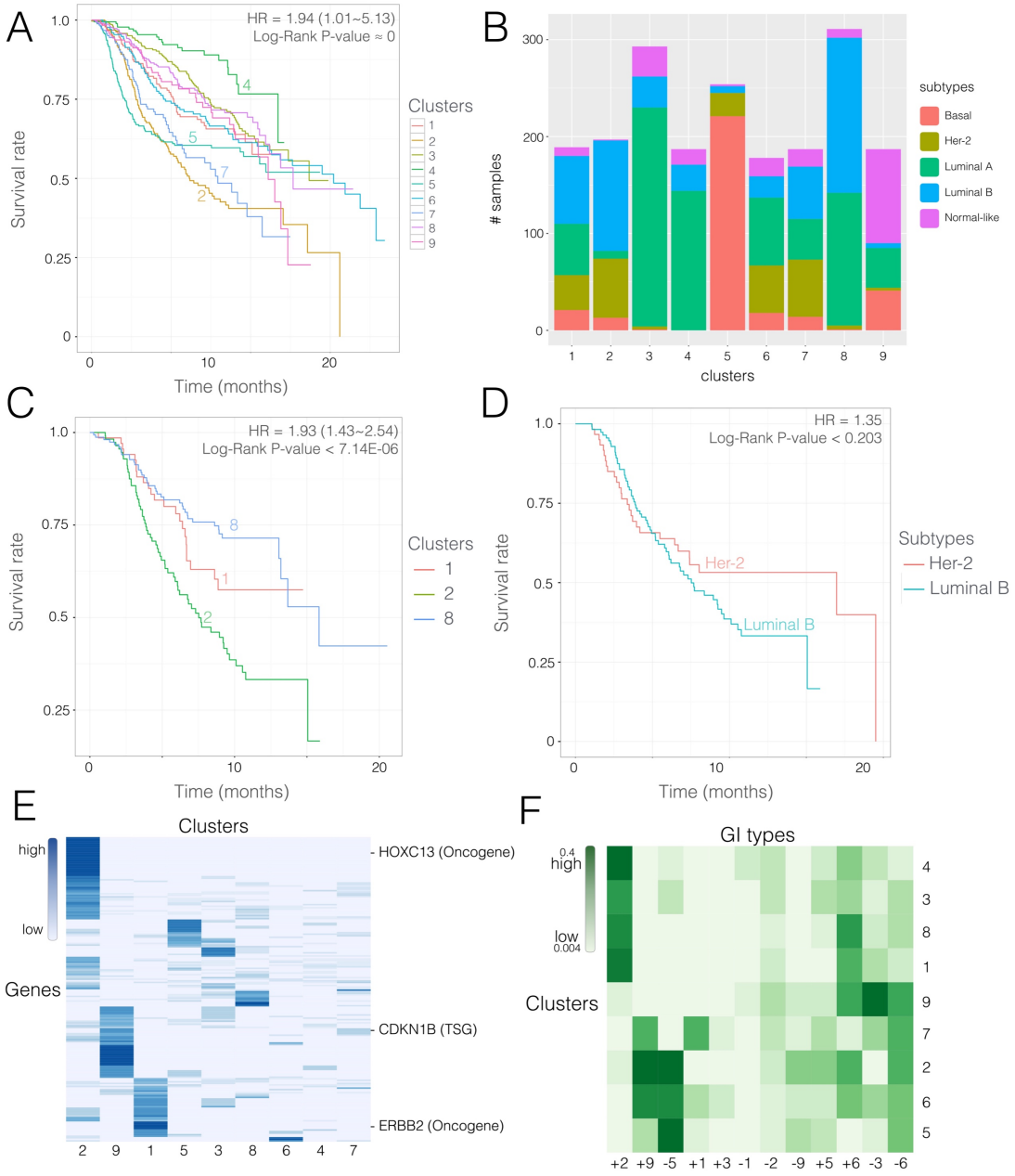


Figure 2.4: Breast cancer patient stratification.
(A) Mean survival curves of the individuals in the 9 inferred GI-based breast cancer subtypes.
(B) Cluster subtype composition based on PAM50 breast cancer sub typing (Bernard et al., 2009).
(C-D) Survival trends of tumors of known histopathological cancer subtypes within and across GI-based clusters. **(C)** Luminal B samples that are split across GI-based clusters 1, 2 and 8 show significant survival differences. **(D)** Her2 and Luminal-B type tumors that are included within the GI-based cluster 2 exhibit similar survival trends.

(E) Mutational profile of GI-based breast cancer subtypes. Mutation profiles of 196 genes (rows) across the 9 GI-based clusters (columns). For each gene and each cluster, the figure depicts the fraction of samples in the cluster in which the gene is mutated.
(F) GI types composition of the GI-based breast cancer subtypes in the METABRIC dataset. In clustering the samples based on GI profile, each GI is probabilistically assigned to a single cluster, based on which, the composition of GIs assigned to each cluster is obtained. The x-axes represent the 12 GI types (6 activity bins and 2 directional effects on survival), and the y-axes represent the clusters. The colors represent the fraction of cluster-assigned GIs of each GI type.

Chapter 3: Single-cell resolution profiling of tumor-reactive $CD4^+$ T-cells

Experimental system to identify tumor antigen-specific $CD4^+$ T cells

To track tumor antigen-specific $CD4^+$ T cells in-vivo, we set up a tractable experimental system where tumor cells present antigens derived from a well characterized glycoprotein, allowing to isolate $CD4^+$ T cells exhibiting reactivity to that glycoprotein after exposure to the tumor microenvironment. Using computational approaches to characterize subpopulations with distinct transcriptional profiles, we identify novel $CD4^+$ T cell subsets.

Tracking tumor-specific $CD4^+$ T cells

We retrovirally expressed the lymphocytic choriomeningitis virus (LCMV) glycoprotein (GP) in colon adenocarcinoma MC38 cells, using a vector expressing mouse Thy1.1 as a reporter (**Figure B.1A**). Subcutaneous injection of the resulting MC38-GP cells produced tumors allowing analysis of immune responses by day 15 after injection. We tracked GP-specific $CD4^+$ T cells through their binding of tetramerized I-A^b MHC-II molecules associated with the GP-derived GP66 peptide (Crawford et al., 2014; Schulz et al., 1989). Such $CD4^+$ cells were found in the tumor and draining lymph node (dLN) of MC38-GP tumor-bearing mice, but neither in non draining LN (nLN) from MC38-GP mice, nor in mice carrying control MC38 tumors (**Figure B.1B**).

To study the $CD4^+$ T cell response to tumor antigens, we aimed to produce genome-wide single cell mRNA expression profiles (scRNAseq) in $CD4^+$ tumor-infiltrating lymphocytes (TILs) and $CD4^+$ dLN cells. We sorted GP66-specific T cells from dLNs, as

these were the only subset of dLN $CD4^+$ T cells for which tumor specificity could be ascertained. Among TILs, we noted that ~87% of GP66-specific $CD4^+$ T cells expressed Programmed Cell Death 1 (PD-1, encoded by *Pdcd1,* **Figure B.1C**), a marker of persistent antigenic stimulation (Agata et al., 1996). Thus, to obtain a broad representation of antigen-specific TILs, not limited to GP-specific cells, we used PD-1 expression as a surrogate for tumor antigen specificity and purified tumor $CD4^+$ $CD44^+$ $PD\text{-}1^+$ T cells ($PD\text{-}1^{hi}$ TIL) for scRNAseq. We verified critical conclusions of the scRNAseq analyses by flow cytometry, comparing GP66-specific and $PD\text{-}1^{hi}$ TILs.

Tumor-responsive $CD4^+$ T cells are highly diverse

We captured GP66-specific dLN (dLN) and $PD\text{-}1^{hi}$ TIL (TILs hereafter) $CD4^+$ cells using the 10x Chromium scRNAseq technology (Zheng et al., 2017b); additionally, we captured GP66-specific $CD4^+$ splenocytes from LCMV (Armstrong strain)-infected mice (Matloubian et al., 1994) as a technical and biological reference (**Figure B.1D**, called 'LCMV cells' here). We excluded cells of low sequencing quality (low number of detected genes), potential doublets, and B cell contaminants, leaving 566 dLN, 730 TILs, and 2163 LCMV $CD4^+$ cells for further analyses (**Table B.1**).

We defined groups of cells sharing similar transcriptomic profiles using Phenograph clustering (Levine et al., 2015). Consistent with previous studies (Ciucci et al., 2019), LCMV-responding cells segregated into transcriptomic patterns characteristic of follicular helper T cells (Tfh, providing help to B cells) and type-1 T helper cells (Th1, secreting the cytokine IFN-γ), among other subsets (**Figure B.2A**). Tfh cells expressed *Tcf7* (encoding the transcription factor Tcf1), *Cxcr5,* and *Bcl6* genes, whereas Th1 cells expressed *Tbx21* (encoding the transcription factor T-bet), *Ifng* (IFN-γ), and *Cxcr6*. Low resolution clustering of TILs and dLN cells identified 5 main groups (**Figure B.2B**). Groups I and II displayed

features of Th1 cells, although group II differed by higher expression of *Cxcr3*, a chemokine receptor implicated in T cell trafficking (Xie et al., 2003) and lower expression of *Ifng*. Group III was characterized by expression of genes typical of regulatory T (Treg) cells, including *Foxp3* and *Il2ra*, encoding CD25, the IL-2 receptor α subunit. Group V had attributes of Tfh cells, including expression of *Bcl6* and *Cxcr5*, while group IV had intermediate levels of *Bcl6* and *Cxcr5* but higher levels of *Ccr7*, which preferentially marks memory cell precursors at the early phase of the immune response (Ciucci et al., 2019; Fritsch et al., 2005; Marshall et al., 2011; Pepper and Jenkins, 2011).

Increasing the resolution of subpopulation identification

To further characterize $CD4^+$ T cell populations, we developed a user-independent, data-driven approach to increase clustering resolution while controlling for false discovery of clusters. Applying such high-resolution clustering separately to TILs and dLN cells, we identified 15 clusters (TIL clusters t1-t7 and dLN clusters n1-n8), refining the original five main groups (**Figure 3.1A**). Revealing unexpected diversity among Th1-like TILs, group I and II resolved into 5 subpopulations, including a distinct cluster (t5) expressing higher levels of *Il7r* (encoding the IL-7 receptor α chain) and lower levels of *Tbx21* and *Ifng*. Only cluster group III (Tregs) included both TIL and a small subset of dLN cells and expressed intermediate levels of *Tbx21*. Groups IV and V, the bulk of dLN cells, resolved into 5 and 2 clusters, respectively. In contrast to high *Tbx21* expression across most TIL subpopulations, and consistent with flow cytometric analysis, dLN cells did not exhibit Th1 attributes nor high *Tbx21* expression (**Figure 3.1A, B.2C and B.2D**).

To support these observations, we analyzed the pooled TILs and dLN cells by t-Distributed Stochastic Neighbor Embedding (t-SNE), a dimensionality reduction approach that positions cells on a two-dimensional grid based on transcriptomic similarity (Laurens van

der Maaten, 2008). Although performed on the pooled populations, t-SNE recapitulated the minimal overlap between TIL and dLN transcriptomic patterns (**Figure 3.1B, left**), irrespective of parameter selection controlling the balance between local and global similarities (**Figure B.2E**). The populations remained separated based on their TIL or dLN origin even after controlling for potential confounders (including TCR engagement on dLN cells as a result of GP66-tetramer-based purification (**Appendix B Extended Data and Figure B.5**), number of unique molecular identifiers (UMIs), and expression of ribosomal and mitochondrial coding genes, **Figure B.2F**). Furthermore, the five cluster groups (**3.1A and B.1B**) almost completely segregated from each other when projected on the t-SNE plot (**Figure 3.1B, right**). Overlay of gene expression confirmed co-localization of cells expressing high levels of cluster-characteristic genes (**Figure 3.1C**).

Reproducibility assessment of subpopulations

To verify the reproducibility of these observations, we analyzed a biological replicate consisting of 1123 TILs and 675 dLN GP66-specific cells captured from a separate set of tumors (**Figure B.2G and Table B.1**). Because batch-specific effects can potentially confound co-clustering from distinct experiments, we separately clustered cells from each replicate and compared these clusters to assess reproducibility. Within each experiment, we generated cluster-specific fold-change (FC) vectors recording expression of each gene in a cluster relative to all other clusters, thus bypassing potential capture-specific biases. We then evaluated pair-wise correlations between FC vectors of clusters across the two replicate experiments to identify clusters with significant transcriptome similarity (reproducible clusters). We found significant inter-experiment matches between most clusters (**Figure 3.1D**), supporting the reproducibility of the underlying transcriptomic patterns. We filtered out the irreproducible clusters, retaining overall 40 out of total 47 clusters across TILs, dLN

and LCMV cells. Thus, scRNAseq analysis of tumor-specific $CD4^+$ T cells identifies an unsuspected diversity of transcriptomic programs in the tumor microenvironment and dLN.

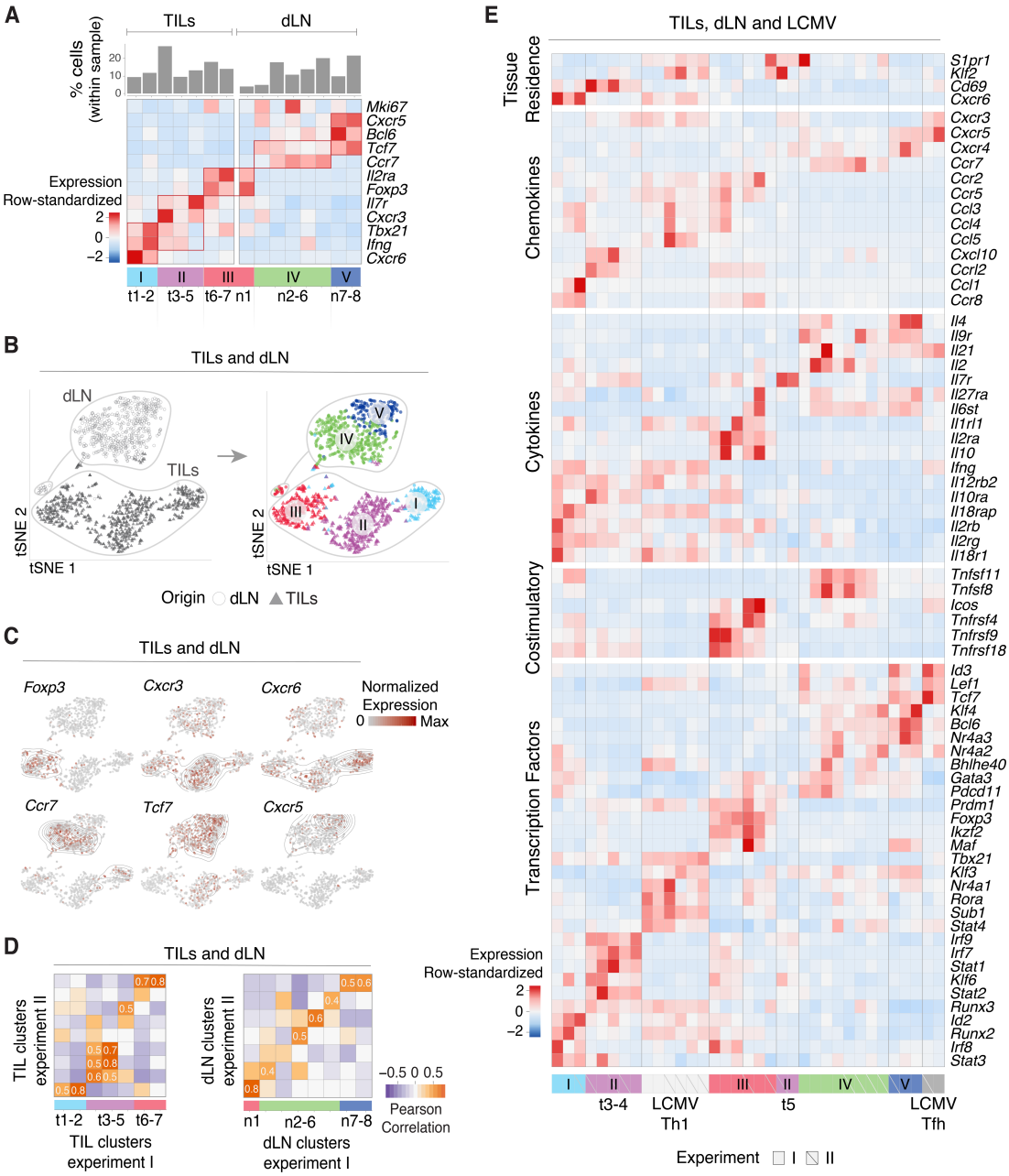


Figure 3.1: Characterization of CD4$^+$ TIL, dLN and LCMV transcriptomes by scRNAseq.

(A-D) TILs and dLN cells from WT mice at day 14 post MC38-GP injection analyzed by scRNAseq. **(A)** Heatmap shows row-standardized expression of selected genes across TIL and dLN clusters. Bar plot indicates the number of cells in each cluster relative to the total TIL or dLN cell number. **(B)** tSNE display of TILs and dLN cells, grey-shaded by tissue origin (**left**) or color-coded by main group (**right**, as defined in A). **(C)** tSNE (TIL and dLN cell positioning as shown in B) display of normalized expression levels of selected

genes. **(D)** Heatmap shows Pearson correlation between clusters' FC vectors (as defined in text) across the two replicate experiments for TILs (**left**) and dLN (**right**).

(E) TILs, dLN and LCMV cells from replicate experiments I and II analyzed by scRNAseq. Heatmap shows row-standardized expression of selected genes across clusters. Group II (purple) t5 separated into a distinct component from t3-4 (as defined in text).

Correlation analyses mitigate tissue-context-specific factors

Meaningful comparison of phenotypes across scRNAseq datasets requires a methodology for synthetic integration of datasets. We demonstrate that correlation analysis provides a sensitive strategy to identify hidden correspondences between datasets.

Tissue-context-specific factors drive conventional clustering

Comparison of TILs, dLN, and LCMV cells showed little overlap, including between TILs and dLN cells (**Figure B.2H, left**). Thus, we considered that the impact of tissue of origin on the transcriptome was the primary driver of clustering and masked potential commonalities in effector programs. Indeed, most TIL subpopulations had attributes of tissue residency, including low *S1pr1* and *Klf2* expression, and high expression of *Cd69* and *Cxcr6*, contrasting with LCMV and most tumor dLN clusters (**Figure 3.1E)** (Bai et al., 2007; Carlson et al., 2006; Kumar et al., 2017). Only group III Tregs, and separately cells undergoing cell cycle clustered together regardless of origin (**Figure B.2H, right**). These observations prompted us to search for potential underlying similarities among these disparate transcriptomic patterns. We found that data integration approaches designed to uncover similarities across experimental conditions could not overcome the strong separation resulting from biological context (**Figure B.3A**), and had limited ability to reveal functionally relevant differences (e.g. between *Foxp3*$^{+}$ and *Foxp3*$^{-}$ TILs, **Figure B.3B**) (Butler et al., 2018).

Correlation analysis identifies correspondence among highly diverse transcriptomic patterns

We considered the correlation analysis used above for cluster reproducibility assessment. This analysis distributed the set of 40 reproducible clusters into 6 ‘meta-clusters’ (with manual curation attaching meta-cluster 1^{b} to 1^{a}) (**Figure 3.2A and Table B.2**). Four meta-

clusters (1, 3, 5 and 6) comprised cells of more than one tissue context (**Figure 3.2A, right**); meta-clusters 3 and 5 covered previously noted similarities among Treg and cycling cells, respectively, and meta-cluster 1 included cells with Tfh transcriptomic patterns (**Figure 3.2A, right**). Thus, the increased sensitivity of the correlation analysis establishes relatedness among the highly diverse transcriptomic patterns identified by conventional clustering.

Characterizing transcriptomic similarities

We further characterized the meta-clusters by identifying their defining overexpressed genes. Genes driving the Treg meta-cluster (meta-cluster 3 encompassing group III) included *Foxp3*, *Il2ra and Ikzf2* (**Figure 3.1E**) and costimulatory molecules *Icos* and *Tnfrsf4*, encoding Ox40 (**Figure 3.2B left, Figure 3.1E**), consistent with flow cytometric analysis (**Figure 3.2D**). In contrast, *Gzmb* (encoding the cytotoxic molecule Granzyme B) and *Lag3* were overexpressed in TIL Tregs relative to dLN Tregs (and relative to other TIL subsets) (**Figure 3.2B right, 3.2C, 3.2E**). Thus, the similarity analysis both confirmed the shared Treg circuitry across TILs and dLN and identified TIL-specific *Gzmb* cytotoxic gene expression in TIL Tregs.

Unexpectedly, and contrasting with Foxp3-expressing Treg clusters, the correlation analysis failed to detect similarities between the three groups of T-bet-expressing cells, which were distributed into meta-clusters 2 (TILs group II t3-4), 4 (LCMV cells) and 6 (TILs group I t1-2) (**Figure 3.2A**). These clusters differed in their relative expression of multiple genes encoding for transcription factors, costimulatory molecules, and cytokines or chemokines or their receptors (**Figure 3.1E**). While all of them expressed T-bet, expression levels substantially differed among clusters. Other transcriptional regulators, including *Stat4*, *Sub1* (a transcriptional co-activator) and *Prdm1* showed cluster-specific expression. Compared to LCMV-responsive Th1 cells, T-bet-expressing TIL clusters also showed higher expression of *Il12rb*, *Il7r* and *Il10ra*, and distinct patterns of chemokine and chemokine receptor

expression. Relative to the other T-bet-expressing cells, TILs group II t3-4 differed by lower expression of *Bhlhe40*, a transcription factor controlling inflammatory Th1 fate determination (Sun et al., 2001; Yu et al., 2018) and the upregulation of multiple type I IFN-induced genes (*Ifit*), including their downstream transcription factors *Irf7* and *Irf9* (**Figure 3.2F top, 3.2G, B.3C**), suggesting a specific impact of interferon type 1 signals in the tumor microenvironment. Thus, we designated group II t3-4 as interferon stimulated cells (Isc hereafter) and group I t1-2 as Th1.

Aside from not expressing these interferon type 1 signals, Th1 TILs differed from other T-bet-expressing cells by expressing Killer Cell Lectin (Klr) genes (**Figure 3.2F bottom, 3.2G, B.3C**), characteristic of terminally differentiated effector cells (Joshi and Kaech, 2008; Jung et al., 2010). Of note, Th1 TILs did not express the Natural Killer (NK) T cell-specific transcription factor PLZF, indicating they were not NK T cells (**Figure B.3D**). Importantly, a recent study of human colon cancer identified a $CD4^+$ TIL subset with elevated *Bhlhe40* expression (Zhang et al., 2018). This subset was clonally expanded and enriched in tumors with micro-satellite instability, suggesting specificity for tumor antigens. Th1 TILs identified in our study exhibited features characteristic of the human colon TIL subset, including expression of *Bhlhe40* and *Lag3* (**Figure 3.2G and B.3C)**, but differed in downregulation of *Gzmb*, *Irf7* and several other molecules (overall 40 upregulated and 10 downregulated genes out of total 216 human colon Th1 genes, GSEA (Subramanian et al., 2005) p = 0.001, **Table B.3**). This suggested that the impact of *Bhlhe40* expression on the TIL transcriptomes is largely consistent but also context-specific.

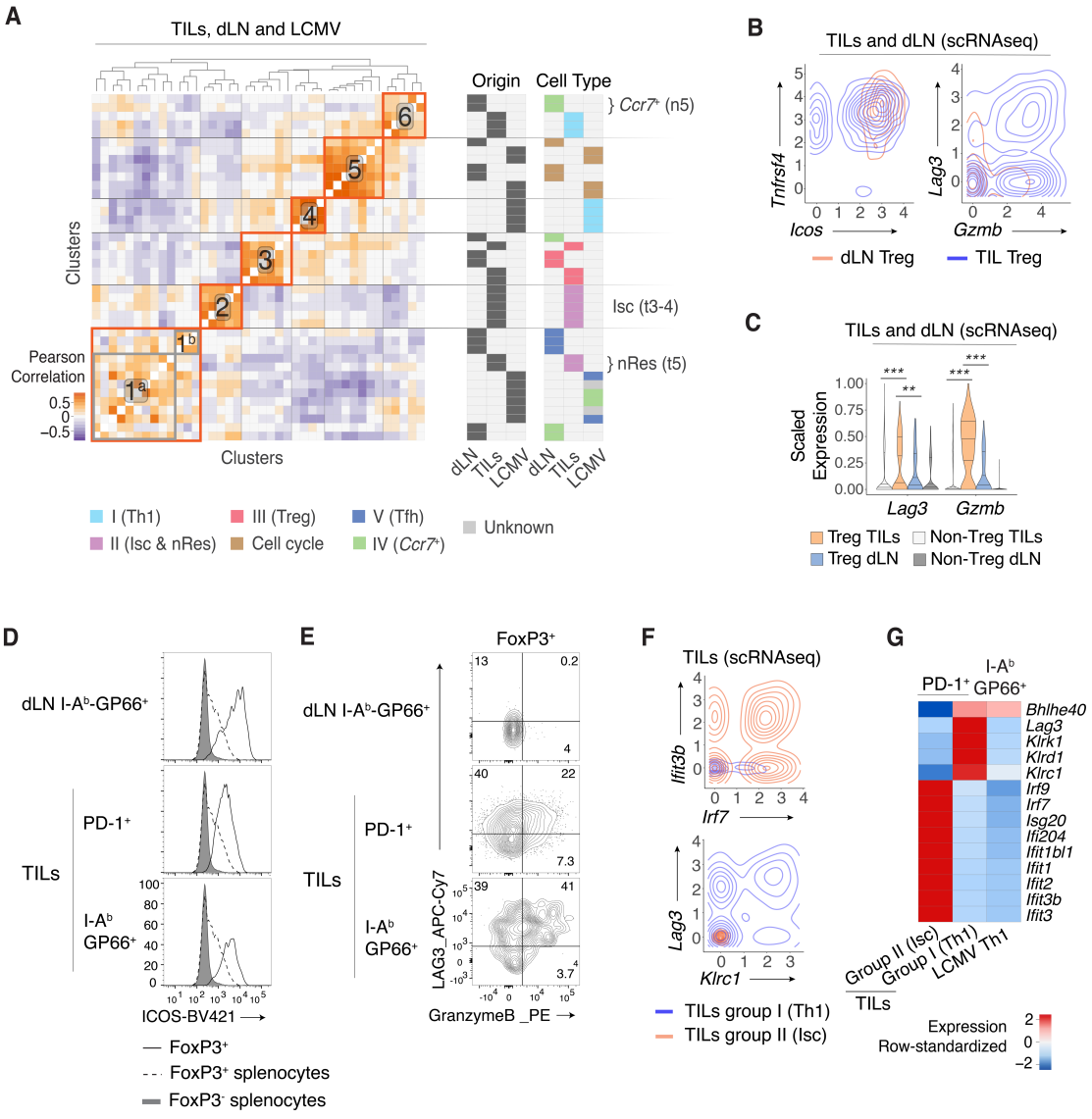


Figure 3.2: Correlation analysis identifies cluster effector fate relatedness and divergence.

(A) Heatmap defines meta-clusters based on Pearson correlation between TIL, dLN and LCMV cluster FC vectors (as defined in text) (**left**). Indicator tables show tissue origin and cell type color-code per cluster (**right**).

(B-E) Comparison of dLN Tregs and TIL Tregs (respectively clusters t6-7 and n1 as shown in **Figure 3.1A**). **(B)** Contour plots of dLN Treg (orange) or TIL Treg (blue) cell distribution according to scRNAseq-detected normalized expression of *Icos* vs. *Tnfrsf4* (**left**) and *Gzmb* vs. *Lag3* (**right**). **(C)** Violin plot of *Lag3* and *Gzmb* scRNAseq expression in Treg vs. non-Treg TIL and dLN populations (Unpaired T test, ** $p < 0.01$, *** $p < 0.001$); bands indicate quartiles (25th, 50th and 75th quantile). **(D)** Overlaid protein expression of ICOS in FoxP3$^+$ TILs, dLN cells and splenocytes from tumor-free mice control. **(E)** Protein expression contours of Granzyme B vs. LAG3 in FoxP3$^+$ TILs and FoxP3$^+$ dLN cells.

(F-G) Comparison of TIL Th1 and Isc (respectively clusters t1-2 and t3-4 as shown in **Figure 3.1A**) to LCMV Th1 (as shown in **Figure 3.1E and S2A**) **(F)** Contour plots of TILs group I Th1 (orange) and group II Isc (blue) cell distribution according to scRNAseq-detected normalized expression of *Irf7* vs. *Ifit3b* (**top**) and *Klrc1* vs. *Lag3* (**bottom**). **(G)** Heatmap shows row-standardized expression of differentially expressed genes across TILs group II Isc, TILs group I Th1 and LCMV Th1.

Similarity analysis identify potential link between dLN and TIL Th1

Meta-cluster 6 unexpectedly associated Th1 TILs and a dLN *Ccr7*$^{+}$ cluster (Group IV cluster n5) (**Figure 3.2A**), suggesting a potential link between TILs and dLN. The association was driven by *Bhlhe40*, the TNF superfamily members 8 (*Tnfsf8* encoding CD30L) and 11 (*Tnfsf11* encoding RANKL), DNA-Binding protein inhibitor *Id2* (**Figure 3.3A and 3.1E**). The potential connection between *Ccr7*$^{+}$ dLN cells and *Ifng*$^{+}$ TIL Th1 was specific to *Ccr7*$^{+}$ cluster n5, which segregated from n6 and other dLN subsets (Tfh and Treg) based on higher expression of *Ifng* (but not *Tbet*) and of *Cd200* (**Figure 3.3B**). Flow cytometry identified a corresponding CD200^{hi} subset among CXCR5^{lo}CCR7^{+} (group IV) but not CXCR5^{+}CCR7^{-} Tfh cells (**Figure 3.3C, B.3E and B.3F**). Multiple dLN *Ccr7*$^{+}$ clusters displayed central memory precursor-like (Tcmp) features defined in LCMV infection (Ciucci et al., 2019), including the expression of transcription factors *Tcf7*, *Klf3* and *Rora*, but lacked expression tissue residence mark *Cd69 (***Figure 3.1E**) (overall 18 upregulated and 12 downregulated genes out of total 69 Tcmp genes (Ciucci et al., 2019), **Table B.3**). This indicated modest correspondence with previous assessments of T cell memory.

Transcriptomic divergences between tumor- and viral-responsive CD4^{+} T cells

Meta-cluster 1 comprised LCMV Tfh clusters and dLN group V Tfh clusters (**Figure 3.2A**). We verified that the abundance of dLN Tfh cells was similar in mice carrying MC38-GP and MC38 tumors (**Figure B.3G**), indicating that this response is not a consequence of GP expression. Flow cytometric analysis confirmed key Tfh attributes in dLN and LCMV cells

(**Figure 3.3D**), although dLN Tfh cells differed from LCMV-responsive Tfh cells by lower expression of *Icos* and the upregulation of the transcription factor *Maf* (**Figure 3.3E, 3.1E and B.3H**). Unexpectedly, meta-cluster 1 associated the dLN and LCMV Tfh clusters with a subpopulation of group II TILs (cluster t5) (**Figures 3.2A and 3.1A**), based in part on intermediate expression of *Tcf7* (1.6 fold relative to other TIL subpopulations) (**Figure 3.3F and 3.1E**), a transcription factor preventing terminal differentiation of effector $CD8^+$ T cells, including TILs (Brummelman et al., 2018; Im et al., 2016; Kurtulus et al., 2019; Siddiqui et al., 2019; Wu et al., 2016; Zhou et al., 2010). Flow cytometric analysis confirmed the abundance of GP66-specific $IL7R^+$ TILs (**Figure 3.3G**). In addition, the $Tcf7^{int}$ t5 cluster showed expression of the transcription factor *Klf2* and its downstream target Sphingosine-1-phosphate receptor 1 (*S1pr1*) (Bai et al., 2007; Carlson et al., 2006), indicating retention of a cell trafficking transcriptional program (**Figure 3.3F and 3.1E**) and contrasting with the interferon-driven Isc TILs. Thus, we designated cluster t5 of group II TILs as putative non-resident cells (nRes hereafter).

Thus, the meta-cluster analysis overcomes divergences resulting from tissue of origin and conventional effector programs to identify subdominant transcriptomic similarities among conventional (non-Treg) cells. To further delineate the relationships between such subsets, we used Reversed Graph Embedding (Trapnell et al., 2014), which has been used to estimate progression through transcriptomic states. This placed the dLN Tfh and TIL Th1 and Isc at the end of an inferred path (**Figure 3.3H**), nRes TILs in the middle of the continuum and $Ccr7^+$ dLN cells between Tfh and nRes. These analyses, combined with the similarities described by meta-clustering, support the notion that the tumor-responsive $CD4^+$ T cell response may be characterized as a transcriptomic continuum; they confirm the transcriptomic distance between Th1 and Isc TILs, even though both subsets express T-bet, the Th1-defining factor.

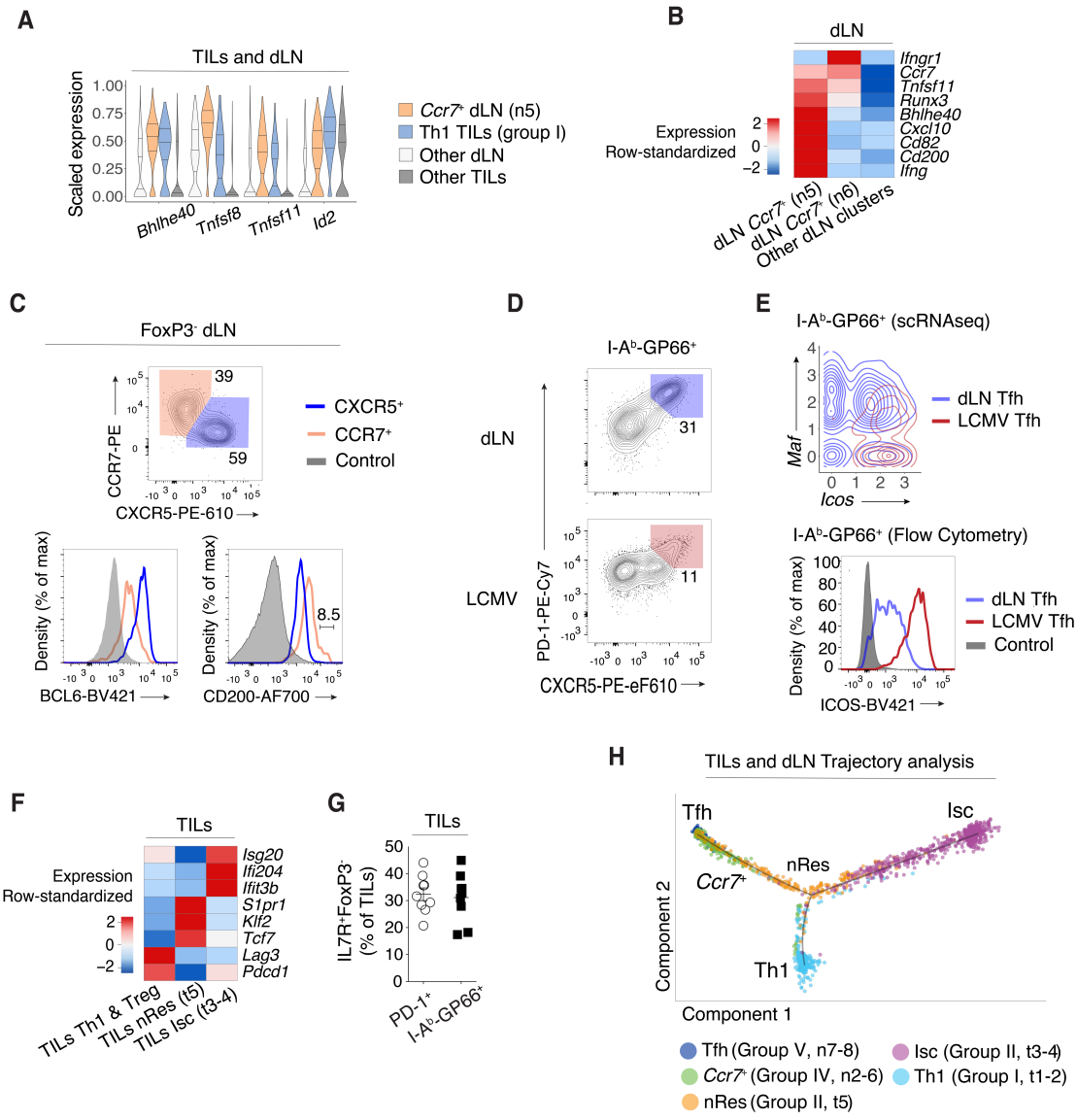


Figure 3.3: Tissue- and subpopulation-specific transcriptomic alterations.

(A) Violin plots of differentially expressed genes across TILs group I Th1, dLN group IV *Ccr7*$^{+}$ (respectively clusters t1-2 and n5 as shown in **Figure 3.1A**) and all other TIL and dLN populations.

(B) Heatmap shows row-standardized expression of differentially expressed genes across dLN *Ccr7*$^{+}$ clusters (group IV n5-6) and other dLN clusters (Treg and Tfh clusters n1 and n7-8, respectively).

(C) Top panel shows protein expression contours of CXCR5 vs. CCR7 in FoxP3^{-} dLN cells. Bottom panel shows overlaid protein expression of BCL6 and CD200 in CCR7^{+} and CXCR5^{+} dLN cells and naive CD4^{+} splenocytes from tumor-free mice control.

(D) Protein expression contours of CXCR5 vs. PD-1 in dLN and LCMV cells.

(E) Contour plot of dLN group V Tfh (red, clusters n7-8) and LCMV Tfh (blue) cell distribution according to scRNAseq-detected normalized expression of *Icos* vs. *Maf* (**top**). Overlaid protein expression of ICOS in dLN and LCMV Tfh cells and naive $CD4^+$ splenocytes from tumor-free mice control (**bottom**).

(F) Heatmap shows row-standardized expression of differentially expressed genes across TILs Isc and nRes clusters (as defined in text, group II t3-4 and t5, respectively) and all other TIL clusters (Th1 and Treg clusters t1-2 and t6-7, respectively).

(G) Fractions of $IL7R^+FoxP3^-$ cells out of total $PD\text{-}1^+$ or $GP66^+$ TILs.

(H) Trajectory analysis of $PD\text{-}1^+$ TILs and $GP66^+$ dLN cells indicating individual cells assignment into a transcriptional continuum trajectory. Group II nRes cluster (t5) is color-coded in orange in contrast to annotations in other figures.

Correspondence to T cell dysfunction and human tumors

We reasoned that expression of a dysfunction-exhaustion program (Thommen and Schumacher, 2018) may account for the limited relatedness between LCMV and TIL Th1 cells, as TILs were sorted on high PD-1 expression for scRNAseq.

TILs subpopulation-specific dysfunction gene programs

To assess the impact of exhaustion on TIL subpopulation, we defined TIL Th1, Isc, nRes and Treg gene signatures as the genes preferentially expressed in each subpopulation relative to all other TILs (**Table B.4**). Consistent with expression of multiple exhaustion marks in TILs (**Figure 3.4A**), we find a significant overlap between multiple viral-response exhaustion gene signatures (MSigDB) (Liberzon et al., 2015) and the Th1 and Treg signatures (**Table B.5**). Separate analysis of a previously reported gene signature characterizing $CD4^+$ T cell dysfunction during chronic infection (Crawford et al., 2014) indicated a significant overlap with the Isc signature, whereas the overlap of that signature with Th1 and Treg signatures was more limited (**Figure B.4A, Table B.6**).

We next examined if the expression of exhaustion genes was homogenous among those TIL subpopulations. Because the overlap with the viral exhaustion signature (Crawford et al., 2014) was too limited for this analysis, we considered a signature encompassing dysfunction marks shared across cancer and chronic viral infection (Chihara et al., 2018). While 55 genes from TIL Th1, Isc, and Treg signatures were also part of this broader dysfunction signature, overlap was heterogeneous, with a pattern of mutually exclusive gene activation among TIL subpopulations (**Figure 3.4B**). This identifies dysfunction programs specific of effector subtypes (**Figure 3.4B, Table B.6**). Of note, we did not detect overlap between any dysfunction-exhaustion signature and nRes cells (**Figure 3.4B, Table B.6**). This is in line

with these cells' residual expression of *Tcf7*, which in $CD8^+$ T cells marks cells with conserved response capabilities (Brummelman et al., 2018; Im et al., 2016; Siddiqui et al., 2019; Wu et al., 2016).

The Isc signature correlates with poor clinical prognosis in human tumors

Last, we examined if MC38-GP TIL transcriptomic patterns were observed in human TILs. We analyzed published *CD4*$^+$ Human liver cancer TILs (TIL_{HLC}) scRNAseq data pooled across six treatment-naive patients (Zheng et al., 2017a). High resolution clustering separated the TIL_{HLC} cells into 11 clusters, which could be combined into groups displaying features of Th1, Isc, Treg TILs and cells undergoing cell cycle (**Figure 3.4C**). While pooled analysis of mouse MC38-GP $CD4^+PD\text{-}1^+$ TILs (TIL) with TIL_{HLC} only identified similarities between cells undergoing cell cycle (**Figure B.4B and B.4C**), cluster correlation analysis indicated significant similarities between TIL and TIL_{HLC} Tregs, cell cycle, and Isc clusters (**Figure 3.4D, top**). We focused on the TIL Isc pattern, which differed the most from previously reported Th1 and Treg transcriptomic profiles. Comparison of overexpression patterns in TIL Isc to the human counterpart indicated a significant overlap which included type I IFN-induced genes and the key transcriptional regulator of type I interferon *IRF7* (**Figure 3.4D bottom and Table B.7**). Thus, the Isc signature newly identified among mouse $CD4^+$ TILs is found in human tumors.

These finding were not unique to liver tumors, as analysis of *CD4*$^+$*CD3*$^+$ human melanoma TILs (Sade-Feldman et al., 2018) across 48 lesions (TIL_{Mel}) identified a cluster enriched in Isc characteristic genes, among other populations (**Figure B.4D**). To investigate the relationships between Isc transcriptomic program and clinical prognosis, we evaluated the association between the expression in TIL_{Mel} of Isc signature genes (defined in MC38-GP TILs) and patient response to checkpoint therapy. Relative to responders, non-responsive tumors had

significantly higher fractions of cells expressing Isc signature genes (49 out of 108 genes, adjusted p-value < 0.05), including the signal transducer *STAT1* and the transcriptional regulators *IRF9* and *IRF7* (**Figure 3.4E and Table B.8**). This indicated negative association between the Isc transcriptomic program and patient response to checkpoint therapy. Thus, the methods used in the present study identify transcriptomic programs shared by multiple tumor types and of potential prognostic significance.

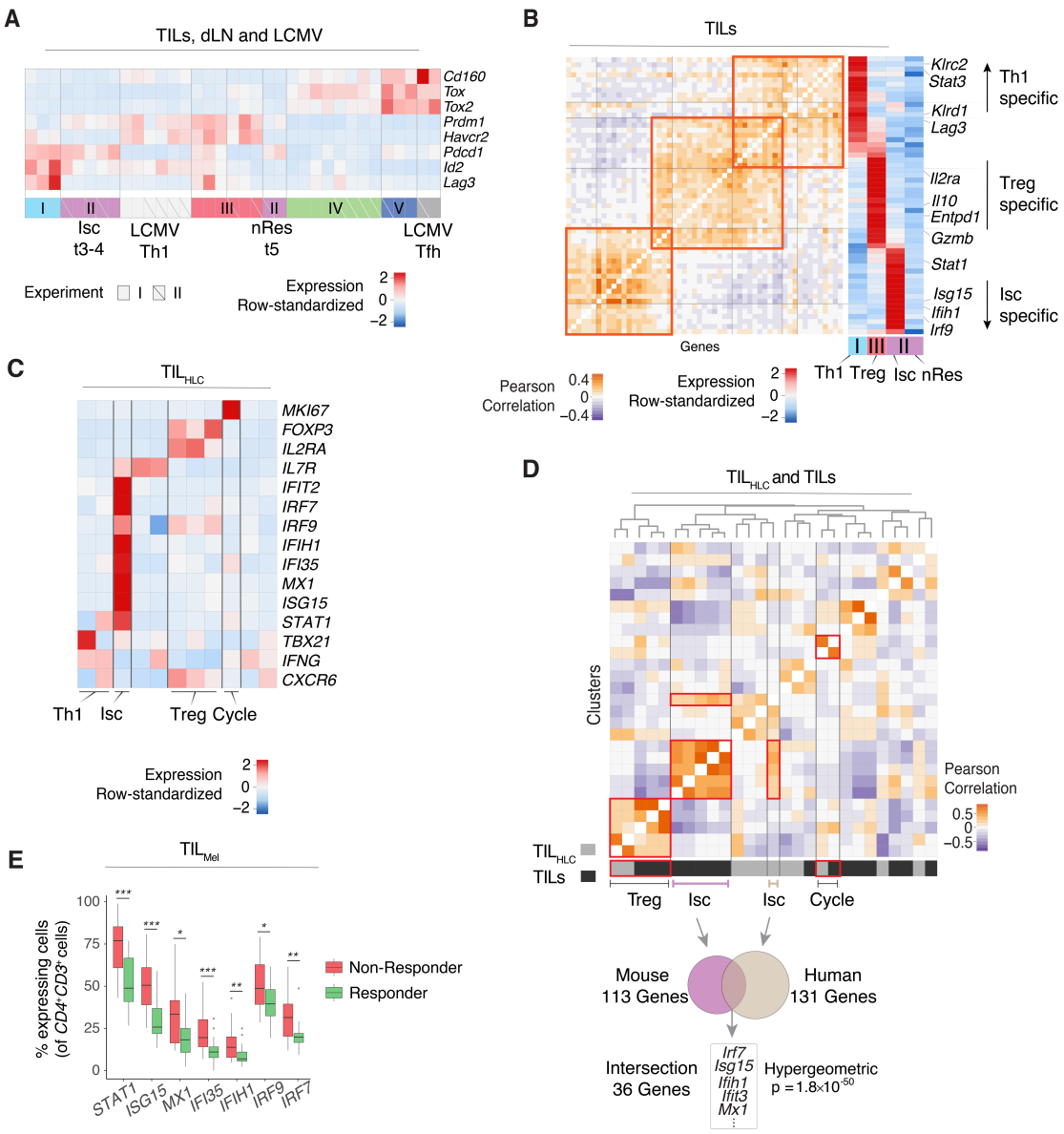


Figure 3.4: Correspondence to human data and dysfunction gene signatures.

(A) Heatmap shows row-standardized expression of selected exhaustion genes across TIL, dLN and LCMV clusters from replicate experiments I and II.

(B) Analysis of IL-27 signature genes overlapping with TIL subpopulation characteristic genes. Heatmap shows Pearson correlation (**left**) and row-standardized expression of overlapping genes across TIL Th1, Treg, Isc and nRes cells (respectively clusters t1-2, t6-7, t3-4 and t5 as shown in **Figure 3.1A**) (**right**).

(C) Analysis of TIL_{HLC} (as defined in text). Heatmap shows row-standardized expression of selected TIL characteristic genes across TIL_{HLC} clusters.

(D) Heatmap defines meta-clusters based on Pearson correlation between TIL_{HLC} and TIL cluster FC vectors (**top**). Overlap of genes characteristic of human liver TIL Isc cluster with mouse TIL Isc gene signature (**bottom**).

(E) Analysis of TIL_{Mel} (as defined in text). Box plots show the percentage of cells expressing selected interferon signaling characteristic genes out of total $CD4^{+}CD3^{+}$ cells across responder and non-responder lesions (Unpaired Wilcoxon test, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$).

Chapter 4: Discussion and Perspective

This dissertation describes new analytical tools providing improved understanding of the clinically-relevant functional interactions in cancer and of the diversity of immune responses to tumors. Using high-throughput molecular data from preclinical cancer models and pancancer clinical studies, the computational frameworks described here touch upon central problems of the fight against cancer and the linkage between genotype to phenotype, including patient survival and response to therapy, and the phenotypic diversity of anti-tumor immune responses.

Genetic interactions

Analyzing molecular and clinical data across thousands of tumors of dozens of types, we comprehensively map the landscape of 12 basic GI types in cancer. Our work extends previous investigations of gene interactions in cancer, which have been almost entirely focused on synthetic lethality (SLs, corresponding to positive effect on survival in bin 1), with a few studies of synthetic dosage lethality (SDL: corresponding to bins 3 and 7), to a total of 12 types of interactions. The identified GIs are predictive of patient survival and drug response, explain tissue-specificity of cancer driver genes, and reveal novel functionally and clinically relevant breast cancer subtypes.

The EnGIne pipeline has bypassed multiple computational bottlenecks by incorporating a Log-Rank analysis allowing to filter gene pairs lacking association with survival regardless of interaction type and applying increasingly stringent methodologies requiring substantial computation time, thereby detecting highly significant multi-type GIs.

Thus, EnGIne provides an unprecedented opportunity to conduct an unbiased whole-genome analysis of multi-type genetic interactions.

The high number of hypotheses being tested poses a significant limitation to data-driven approached for genetic interactions discovery. EnGIne enables to resolve this limitation by using thousands of tumor samples and generating a large-scale background distribution of shuffled GIs for defining the GI significance cutoffs. While this approach is plausible for pairwise genetic interaction identification, the problem exacerbates in studying more complex hypotheses such as exploration of higher-order interactions (e.g. gene triplets). Thus, future studies may consider limiting the hypotheses space by exploring higher-order GIs involving a specific gene of interest, thereby reducing the problem to a pairwise GI identification task. Alternatively, increasing the number of samples by integrating additional data sources into a dataset several folds larger than TCGA may enable such exploration, but computational power will become the bottleneck.

In addition to increasing the statistical power, integrating information across cancer types allowed EnGIne to identify shared transcriptional patterns, rather than tissue-context-specific patterns. While most cancer studies focus on a specific tissue of interest, we hypothesized that although different cancer types are vastly divergent in transcriptomic profiles, some shared mechanisms may be identified and reflected in genetic interactions. Our findings support the paradigm of developing generalized diagnostic tools and therapies for treatments across multiple cancer types. Similarly, we chose to integrate and identify commonalities across various demographic groups rather than studying various them independently. However, previous studies recognized distinct molecular and clinical patterns across genders and races (Dorak and Karpuzoglu, 2012; Yuan et al., 2016). Building tissue-specific GI networks may be of interest for future investigations, provided larger scale datasets. Thus, while integrative analysis enables the discovery of shared molecular patterns, GI

networks may be highly polymorphic across clinical- and demographic-groups. Given the statistical power discussion above, such investigations may benefit from focusing the analysis on specific target genes of interest rather than conducting full genome-wide analysis as described here. Furthermore, contrasting the tissue-specific GI networks' information should identify molecular patterns exclusive to specific cancers and may enhance our understanding of the properties of tumorigenesis unique to each tissue.

Additional complexity inherent to genetic interaction inference originates from modeling interaction effects versus individual gene effects. Genetic studies in yeast (Costanzo et al., 2016; van Leeuwen et al., 2017) considered pair-wise interactions relative to additive phenotypic effect on cell viability *in vitro*. Our study has bypassed assumptions pertaining the nature of the individual gene effect by incorporating their information in a regression model and identifying cases where the interaction term is deemed highly informative for survival prediction in addition to the individual gene information.

In contrast to previous data-driven approaches for GI identification used in-vitro cell line viability measurements across genome-wide knock-out experiments, allowing to assess the concordance between *in-vivo* to *in-vitro* data. However, knock-out *in-vitro* experiments do not capture the GIs involving overactivation states of genes, thus preventing the generalization of previous methodologies to multi-type (including medium and high expression levels) *in-vitro*-supported GI discovery. Therefore, future large-scale *in-vitro* studies may benefit the investigation of genetic interactions provided that the experimental techniques allow for more sensitive control of gene expression levels and capture cell line viability differences across low, medium and high expression levels, similarly to our binning approach. This would allow biologically meaningful integration of *in-vivo* and *in-vitro* data into the EnGIne model as well as provide the ability to perform follow-up experimental validation studies of multi-type GIs.

Multiple factors are worth considering when deciding on the strategy to assign tumors into bins based on gene activity levels. For instance, a previous study of synthetic lethality in glioblastoma (Szczurek et al., 2013) defined the high (low) state as mRNA expression higher than the 80^{th} quantile (lower than the 20^{th} quantile, respectively). We chose to partition gene activity into 3 quantiles of low, medium, and high activity levels within cancer samples. We have shown the robustness of the detected GIs (**Appendix A Extended Results 1**). Besides its simplicity and being non-parametric, our strategy naturally allows us to search for cases where even the normal (or medium) levels of expression of a gene may be associated with fitness or survival effects, depending on the states of other genes (e.g., bins 2 and 6).

The diagnostic implications of the GI network include multiple key strategies for prioritizing drug administration based on patient-specific genomic features. First, the GI network can be used to estimate the impact of a drug target gene inhibition or activation, providing opportunity to administer drugs based on the anticipated functional impact of each drug. Additionally, given a specific drug intervention, resistance mechanisms can be modelled by identifying the positive interactions that could be activated post-treatment. Consequently, one may estimate an optimal combination therapy inducing high negative GI impact on cancer cells and minimizing positive impact associated with resistance.

Given that EnGIne is using bulk RNA-Sequencing rather than single-cell genomic measurements, the pipeline is limited in sensitivity to identify molecular patterns of tumor clones. Since tumors are composed of multiple phenotypically distinct subpopulations, the molecular patterns identified in this study represent the average trends across the highly heterogeneous tumors. Functional studies of GI activation within tumor clones using scRNAseq should provide a substantially advance our understanding of sensitivity or resistance to therapy.

One caveat of genome-wide association studies is asserting which genomic features are causal to change the course of disease. While statistical association between genotypes to phenotypes may be informative for developing diagnostic tools, clinical interventions rely on factors whose modification would cause an alteration in the clinical outcome. The association versus causality problem may exacerbate in large scale association study of pairwise genomic features, such as the analysis described in this thesis. Our correlation analysis indicated that the set of identified GIs is not dominated by correlated features, thus demonstrating that the diversity and abundance of multi-type interactions is likely to be representative of the true diversity. However, integrating additional molecular data types, such as cell lines, mutation, copy number variation (CNV) and protein expression may enable identifying a smaller set of GIs exhibiting consistent trends across a larger feature space and therefore providing additional confidence in their putative causal role. Experimental validations will be required to demonstrate the causal role of these GIs in pre-clinical models.

The set of functionally active GIs provides a complementary molecular characterization of tumors to those obtained by histology and contemporary individual gene-based transcriptomic profiles. A better concordance of the GI-based breast cancer subtypes with survival trends may be partly because GIs were inferred based on their impact on survival. However, interestingly, the detected subtypes are additionally marked by distinct mutational profiles, which were not utilized in inferring the GIs. Overall, these results underscore the importance of molecular context represented by functionally active GIs.

Identifying pairwise gene interaction is only a first step toward capturing the true complexity of cellular networks. Future work can go beyond the 12 basic GI types studied here to investigate more complex GI types that involve different compositions of these basic types; for instance, a given interacting gene pair can confer tumor fitness benefit in multiple co-activity bins and reduce tumor fitness in others. Thus, while the results presented here go

markedly beyond previous definitions and analyses of GIs, they only begin to explore the full scope and clinical potential of GI-based analyses of cancer, awaiting future investigations.

CD4$^+$ T cell diversity by scRNAseq

Using scRNAseq and data-driven computational approaches, our investigation of tumor-responding CD4$^+$ T cells identifies an unsuspected transcriptomic diversity. While recent scRNAseq studies shed light on the Treg component of CD4$^+$ TILs (Ahmadzadeh et al., 2019; Azizi et al., 2018; Zhang et al., 2018; Zheng et al., 2017a), our study assessed the transcriptomic patterns of both regulatory and conventional components, in the tumor itself and in draining lymphoid organs. We identify new transcriptomic patterns and find a heterogeneous distribution of exhaustion gene signatures among TILs subtypes, highlighting the need for extensive analyses of cell-specific effects of treatments targeting exhaustion genes.

Even though most conventional (Foxp3$^-$) tumor-responsive TILs express T-bet, the Th1-defining transcriptional regulator, our study identifies novel and diverse transcriptomic patterns with unexpectedly little similarity to prototypical virus-responsive Th1 cells. This includes a strong TIL-specific IFNγ response across subpopulations. Thus, conventional helper effector definitions, derived from studies of responses to infection, are inaccurate descriptors of responses to tumors. The newly identified Th1-like transcriptome with marks of type I IFN stimulation highlights this conclusion: it was observed among TILs but not LCMV-responding cells, even though LCMV drives a strong type I IFN innate immune response (Cousens et al., 1999).

The role of type I IFN in cancer is not well characterized. Previous studies associated type I IFN with both inflammation and immunosuppression in cancer (Snell et al., 2017). However, our cluster similarity analysis projects the interferon-responsive transcriptomic pattern onto human tumors, overcoming potential sample disparity, and demonstrates its association with poor response to checkpoint therapy. This suggests a negative impact of a strong type I IFN signalling on anti-tumor responses in the context of therapies such as PD-1

blockade. Follow-up studies will provide additional insight into the role of Isc TILs and their emergence in the context of anti-tumor responses. Initially, flow cytometric analysis of Isc genes may enable isolating these cells and facilitate functional validations. Such investigations may demonstrate the effects of depleting or enriching the tumor microenvironment with Isc TILs on tumor growth in pre-clinical models. Additionally, intervention in IFN stimulation mechanisms on $CD4^+$ T cells may demonstrate the ability to control the Isc population abundance in-vivo. Consequently, these findings may be translated into clinical settings as predictive biomarkers allowing to monitor the patient response to checkpoint blockade or developing therapeutic agents to modify the lymphocyte population composition and induce durable anti-tumor responses.

Investigating tumor-specific T cell responses in draining lymphoid organs revealed striking differences with TILs. The absence of Th1 cells from tumor dLN was unexpected and contrasted with infections, including with LCMV or with *Leishmania major*, a typical Th1-driving parasite with kinetics of clinical progression similar to that of experimental tumors, and in which Th1 dLN cells are important contributors to the response (Belkaid et al., 2000). In contrast, the tumor elicited strong, tumor-specific Foxp3-negative Tfh-like responses in dLN. While Tfh differentiation may divert T cells from more efficient (e.g. IFNγ-producing) anti-tumor differentiation, it provides support for the tantalizing possibility that tumor-elicited B cell responses could be exploited against cancer (Carmi et al., 2015). It is also possible that this subset includes a stem cell-like component similar to the $Cxcr5^+$ $CD8^+$ dLN T cells that serve as targets for immunotherapy targeting PD-1 signaling (Im et al., 2016), or cells with similar properties in the tumor micro-environment (Siddiqui et al., 2019).

Studying the migration and differentiation dynamics may validate the similarity and trajectory analysis which was suggestive of a connection between $Ccr7^+$ and nRes ($Tcf7^+S1pr1^+$) transcriptional program in dLN to Th1 and Isc subsets in TME. To this end,

blocking lymphocyte migration or tracking lymphocytes during migration will shed light into the origins of Th1 and Isc in secondary lymphoid organs. An alternative explanation to the disconcordant transcriptomic patterns between dLN and TIL may originate from obtaining only a single snapshot of the population composition at day 15 after MC-38GP injection rather than multiple time points. Since the migration dynamics between the dLN and the TME may occur earlier than 15 days post-injection, a time-course analysis of population composition should indicate whether the TIL-specific subpopulations leave the dLN prior to adopting the Th1 and Isc transcriptomic features or alternatively, they develop these features early and leave shortly after tumor engraftment.

While we have focused on the prognostic utility of the Isc gene signature, a similar strategy can be used to model the multiple associations between subpopulation and clinical responses. By measuring continuous phenotypes such as tumor growth or patient survival, one may integrate the abundance of all the identified subpopulations in a multiple regression model. Integrative analysis of the full diversity of TIL subpopulations may inform us of the joint contribution of cell types and states to the anti-tumor capacity of the immune response. Given these associations, one may validate them in pre-clinical models by modifying the population composition in-vivo and measuring tumor growth. These studies will pave the way for rational combination therapies designed to alter immune composition in patients.

The data-driven approaches introduced in this work will allow researchers to perform unbiased investigations of scRNAseq datasets without imposing arbitrary constraints on the analysis outcomes, such as the clustering resolution. Previous studies have defined the clustering resolution based on presumptions on the population composition or according to saturation of explained variance as a function of number of clusters. However, these approaches are insufficient for unbiased exploratory analysis of poorly characterized cell populations. Our framework allows to measure the confidence one should assign to multiple

high resolution clustering analyses and provide guidance in determining the appropriate resolution.

Furthermore, we provide the means to conduct large scale comparative analysis across experiments, tissue environments and experimental settings, including correspondence between preclinical and clinical genomic measurements (i.e. mouse to human transcriptomes). Thus, the analytical approaches described in this thesis provide new ways to assess the extent of which transcriptomic patterns are shared or distinct across settings and establish the relevance of pre-clinical findings in human tumors. Additionally, we identify reproducible marker genes contributing the divergent phenotypes across experimental settings and establish the putative drivers of shared or unique molecular patterns across datasets.

Future work will extend on the computational contributions by introducing a hierarchical structure to the clustering analyses. Hierarchical modelling should provide improved sensitivity in differential expression analysis and marker discovery. Additionally, hierarchical tree structure may be transformed into a decision tree, allowing the use of hierarchy-defining genes to classify new cells into the correct subpopulation. This idea may resolve the limitation where certain subpopulations defined via conventional clustering are deemed unresolved as they may contain a mixture of cell types.

In conclusion, this study provides a high-resolution characterization of tumor-reactive $CD4^+$ T cell responses in lymphoid organs and the tumor microenvironment. We identify previously unrecognized transcriptomic patterns among tumor-specific T cells and provide an extensive mapping of the $CD4^+$ T cell immune response against cancer. We describe new analytical approaches of broad applicability, including to clinical data, that combine high resolution dissection of transcriptomic patterns and synthetic data integration to identify correspondences between apparently unrelated cell differentiation states.

Appendix A - Genetic Interactions

Extended Results

1. Robustness analysis of genetic interactions

We assessed the robustness of threshold selection throughout the EnGIne pipeline, including:

a) Modifications of bin quantile boundaries: we kept the 3×3 structure and either increased or decreased the corner bins by 10%, corresponding to ~170 added or removed samples per bin

b) Log-rank p-value based most significant quantile: scanning between top 5% – 30%

c) Cox regression FDR threshold: scanning between 0.01 – 0.1

Notably, the relatively small changes in binning thresholds described in (a) result in moving hundreds of samples in or out of the bins (170 ~ 340 samples), thus it is a significant change of the bin size and composition. For computational tractability of exploring a large parameter space, we limited the set of genes to 557 Cosmic Cancer Census (Futreal et al., 2004) genes (Tier 1 and 2) corresponding to 154,846 possible gene pair combinations. As shown in **Table A.7**, for the alternate binning, across all log-rank thresholds, 57%-96% of the gene pairs are recapitulated. Likewise, setting the log-rank quantile threshold to 0.8 (top 20% retained), across various Cox FDR thresholds, 45%-56% of the GI detected using the default setting are captured. For instance, at log-rank threshold = 0.8 and Cox regression FDR = 0.1 yields ~33,000 GIs (regardless of the binning threshold), and these include 53%~55% of the GIs defined using the original binning method. While the EnGIne pipeline offers users to set the above parameters, to demonstrate the utility of EnGIne in the main results we have used a

more stringent Log-rank p-value based quantile and Cox FDR in the genome wide analysis to make the execution time of the pipeline tractable.

2. Correlation analysis of genetic interactions

To assess whether the full GIs network is dominated by correlated gene expression patterns, we compared between correlations found between GI pairs to shuffled GIs. For a pair of GIs A = (x_1,y_1) and B = (x_2,y_2), we defined the Pearson correlation statistic (PCS) between A and B as $Min(\rho(x_1, x_2), \rho(y_1, y_2))$, quantifying whether the two GIs are independently inferred. Then, we calculated the PCS for GI pairs within each GI type and compared them to PCS calculated for the shuffled GI pairs from the same GI type. Given the high number of GIs and the infeasible number of pairs to test, we selected up to 1000 GIs randomly from each GI type and calculated the corresponding shuffled GI list of the same length. The PCS comparisons indicated similar distributions of the actual and shuffled GIs across all GI types (**Figure A.2B, Table A.8**). Although GI type +9 exhibited higher PCS relative to shuffled, the effect size was marginal (**Table A.8**). Thus, we conclude that the full GI network is not dominated by correlated gene expression patterns.

3. GI based prediction of patient survival

Recall that a GI is represented by a quadruple comprising of a gene pair (x,y), a symmetric bin b (1, 2, 3, 5, 6, 9; **Figure 2.1**), and its effect on tumor fitness (positive or negative), and a GI is deemed *functionally active* in a specific tumor sample if the joint expression levels of the genes x, y, in the tumor fall in bin b. We turned to assess the extent to which *the aggregate effect* of functionally active GIs in a tumor predicts patient survival (the individual GIs are indeed inferred while considering survival but here, we are interested in the predictive power of their combined effects). The naïve tumor-specific GI survival score is computed as

the difference between the number of functionally active GIs in a tumor with positive effect on survival and the number of those with negative effect on survival.

We compared the GI-based survival prediction accuracies with a comparable gene-level approach (which is based on assessing the expression levels of the genes that are significantly associated with patient survival) as well as a combined gene and GI-based approach. **Figure A.4A** shows the survival prediction accuracy using the widely used C-index (CI) metric, both via cross-validation within TCGA (Weinstein et al., 2013) and on independent METABRIC (Curtis et al., 2012a) breast cancer dataset. As shown, pan-cancer risk prediction based on the predicted GIs compares favorably with the comparable gene-level approach both in cross-validation, and more so on the METABRIC (Curtis et al., 2012a) independent dataset — suggesting good generalizability.

Applying the fully supervised individual gene-level approach (Yuan et al., 2014) (individual gene feature selection based on association with survival, further dimensionality reduction with LASSO Cox and a final Cox model to obtain genes' coefficients for survival risk prediction) yields a comparable accuracy in cross validation (CI ≈ 0.63) but a lower accuracy over the independent dataset (CI ≈ 0.55). Finally, a previous study has reported a CI of 0.71 using a supervised approach specifically on Kidney Renal Clear Cell Carcinoma (KIRC). We ensured that a lower gene-wise accuracy that we observe is not simply due to our filtering and implementation (methods); applying the fully supervised individual gene-level approach on KIRC subset of the filtered dataset yields CI ≈ 0.71, recapitulating the accuracy reported in the original publication (Yuan et al., 2014).

In **Figure A.4B**, we present the performance of the GI-based approach using each of the 6 GI types. As evident, interactions involving both genes in their wild type mid-activity levels have negligible predictive power on survival, testifying that more extreme levels of

gene expression tend to be involved in functional GIs affecting survival. **Figure A.5** shows the performance based on an alternative metric where we dichotomized the extreme (at varying thresholds from 10% to 50%) predicted low- and high-risk groups and quantified the difference in their area under their KM survival curves . The survival prediction performance of the full compendium of 12 (positive and negative) GI types is shown in **Figure A.6**.

4. Assessment of potential confounding factors

We assessed genomic-instability and tumor purity as potential confounding factors and found that the vast majority of the discovered GIs remain highly significant. The genomic instability index measures the relative amplification or deletion of genes in a tumor based on the SCNA (Bilal et al., 2013) and the tumor purity is an estimate of the proportion of cancer cells in the sample (Aran et al., 2015). We recomputed the controlled Cox survival step (**Figure 2.1E**) for all the candidates and shuffled candidates obtained from the previous step of the pipeline, with the addition of the two covariates — genomic instability and tumor purity, to calculate the empirical FDR threshold (0.01 quantile). We obtain 77630 confounder-corrected GIs where (1) 60302 of the 71946 original GIs and (2) 1405 of the 1704 original PPI GIs were detected. Modifying the FDR threshold to 0.1 yields a list of 154450 confounder-corrected GIs which contain almost all of the 72k original GIs with the exception of 48 GIs that are not included (these 48 GIs remain the exception when increasing the FDR quantile threshold to 0.2, reflecting the set of highly confounded GIs, **Table A.6**). 45 of the confounded GIs involve DEFB21, a member of the beta subfamily of defensins (antimicrobial peptides), suggesting that the majority of the confounded effects are limited to one gene as oppose to a wide-spread phenomenon. None of the GIs is discussed in the main text are affected by this additional confounding control.

5. Quality assessment of TCGA Breast cancer data

The TCGA (Weinstein et al., 2013) breast cancer cohort is composed of several subsets of independent studies, with potential batch effects and other confounders. To assess the quality of the TCGA (Weinstein et al., 2013) breast cancer molecular and clinical data, we analyzed the association with survival of 79 known breast cancer genes (Intogen dataset (Gonzalez-Perez et al., 2013)). We find 31 (~40%) of the genes to be significantly associated with survival ($P < 0.05$). Finally, although the TCGA (Weinstein et al., 2013) cohort may be noisy due to the data collection procedures, we find the survival prediction accuracy of the discovered GIs to generalize to the METABRIC (Curtis et al., 2012a) dataset, supporting a reasonable quality of the data.

Extended Figures

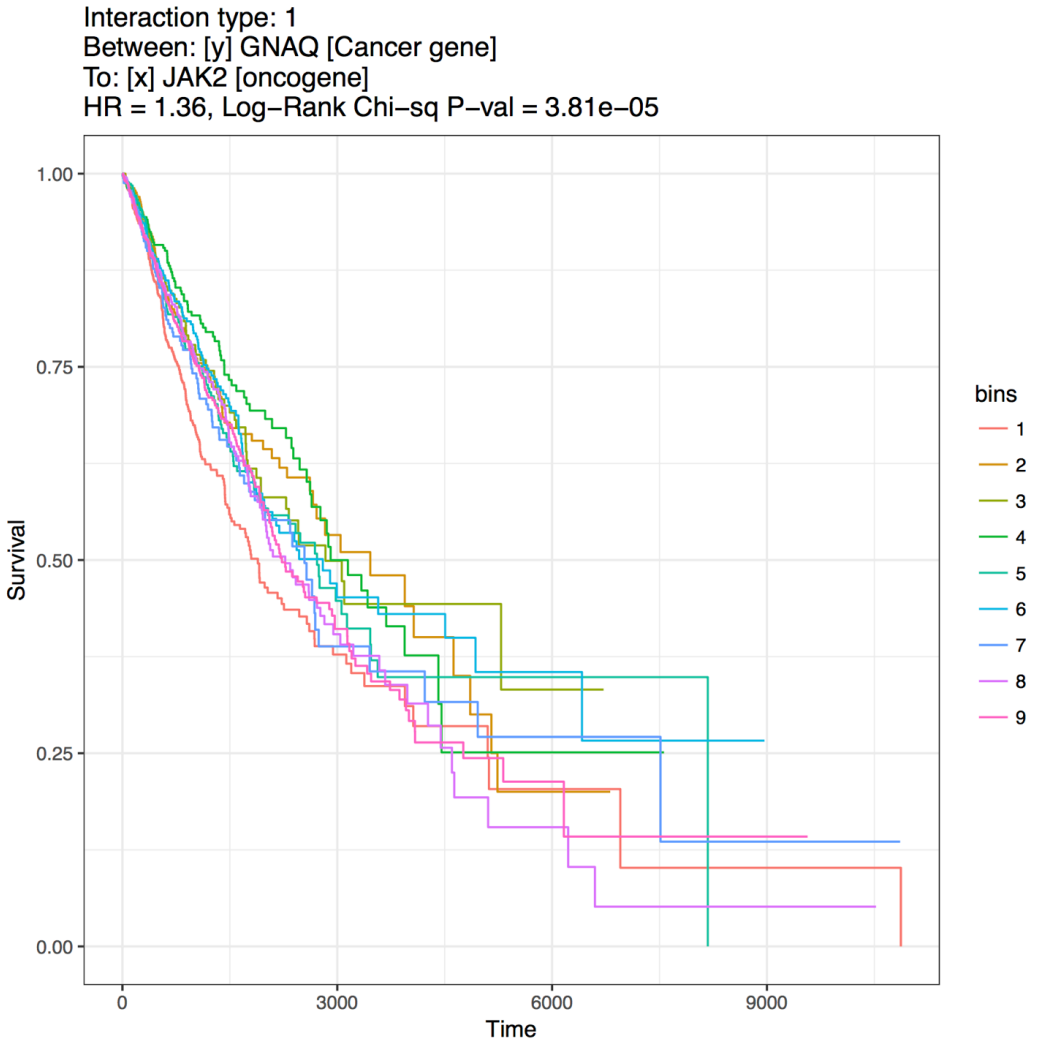


Figure A.1: The Interaction between the cancer genes GNAQ and JAK2 as an example of a positive interaction in bin-1.

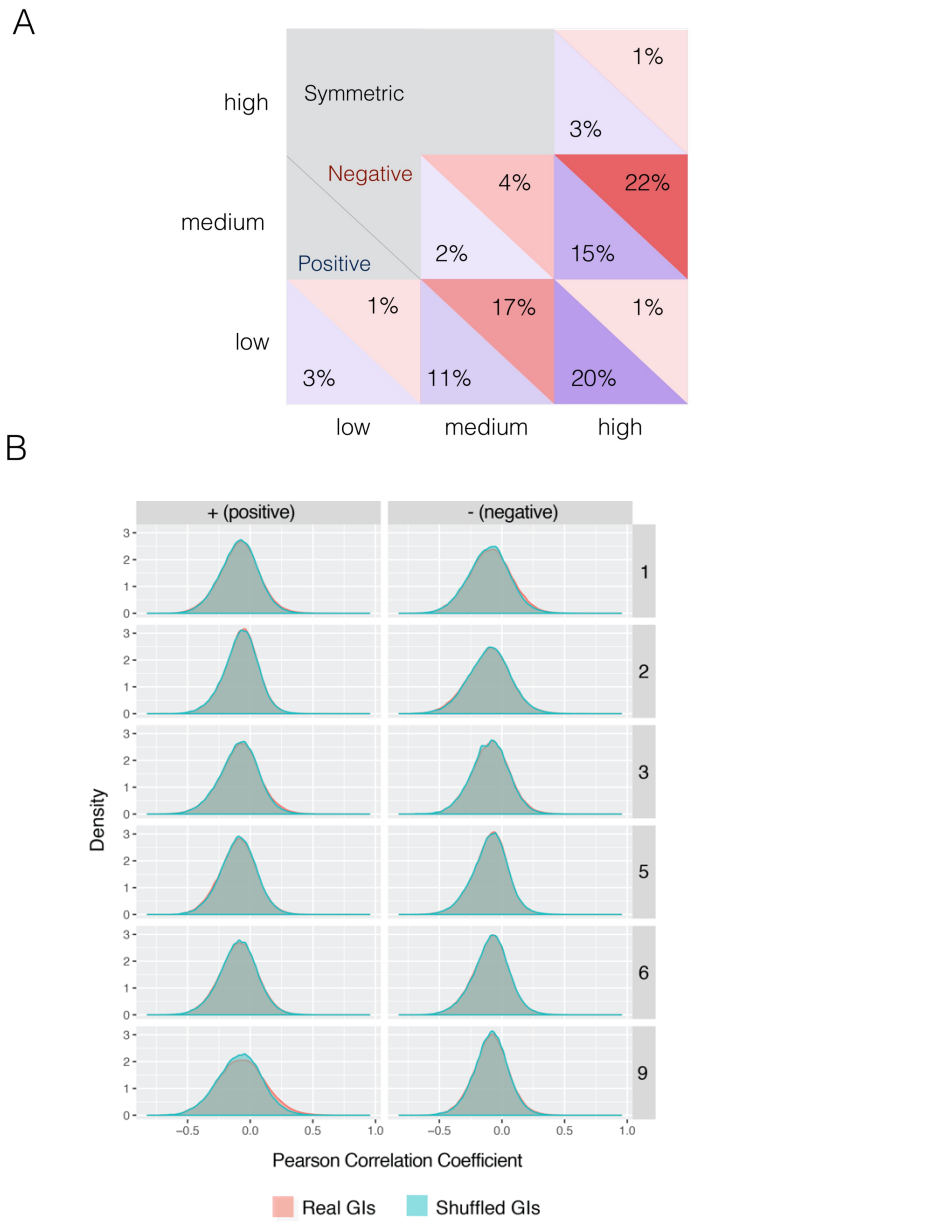


Figure A.2: GI abundance distribution and correlation analysis.
(A) Distribution of the 71,946 significant GIs across 12 joint activity bins. The fractions of GIs in each bin are shown for GIs with positive (blue) and negative (red) effect on tumor fitness. Only the data in the lower triangle of the matrix are shown as the GIs are symmetric relative to the genes in a pair.
(B) Correlation between GI pairs and shuffled GI pairs across the 12 activity bins. Minimum statistic distribution of Pearson correlation between genes defining the Real GIs (non-shuffled, red) and Shuffled GIs (blue).

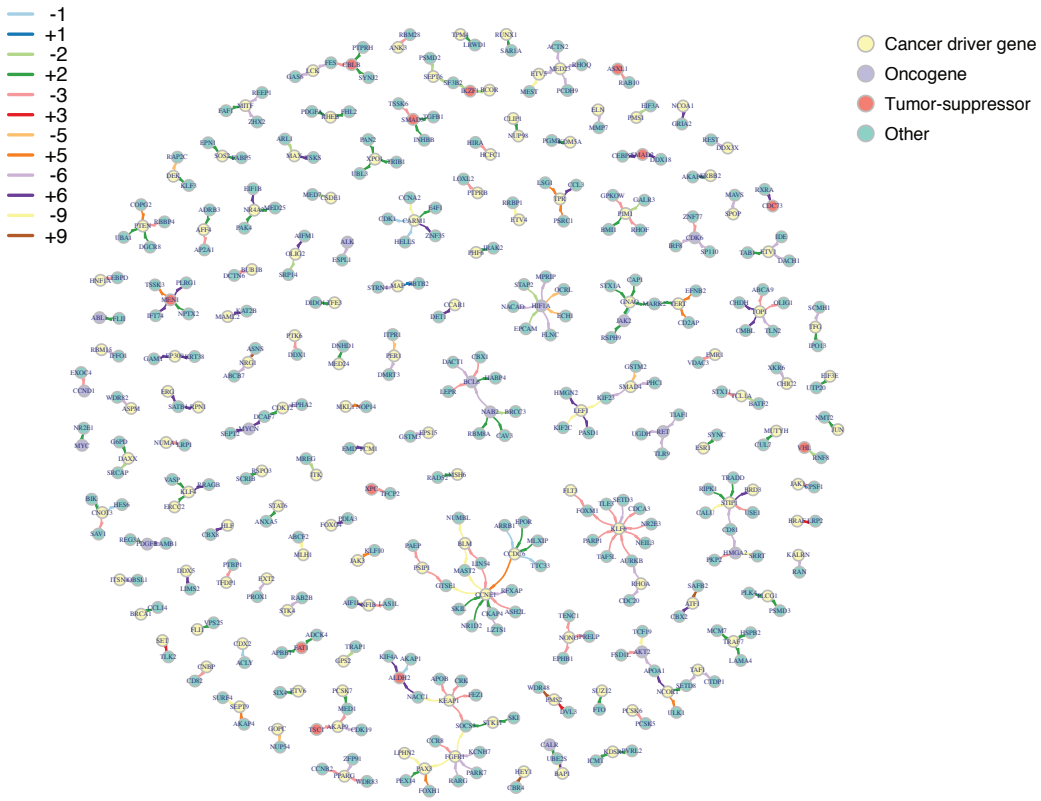


Figure A.3: The GI-network involving the known driver genes. Nodes color coded by oncogenic role (yellow – driver gene, purple – oncogene, red – tumor suppressor, blue – unknown). Edges color coded by GI type.

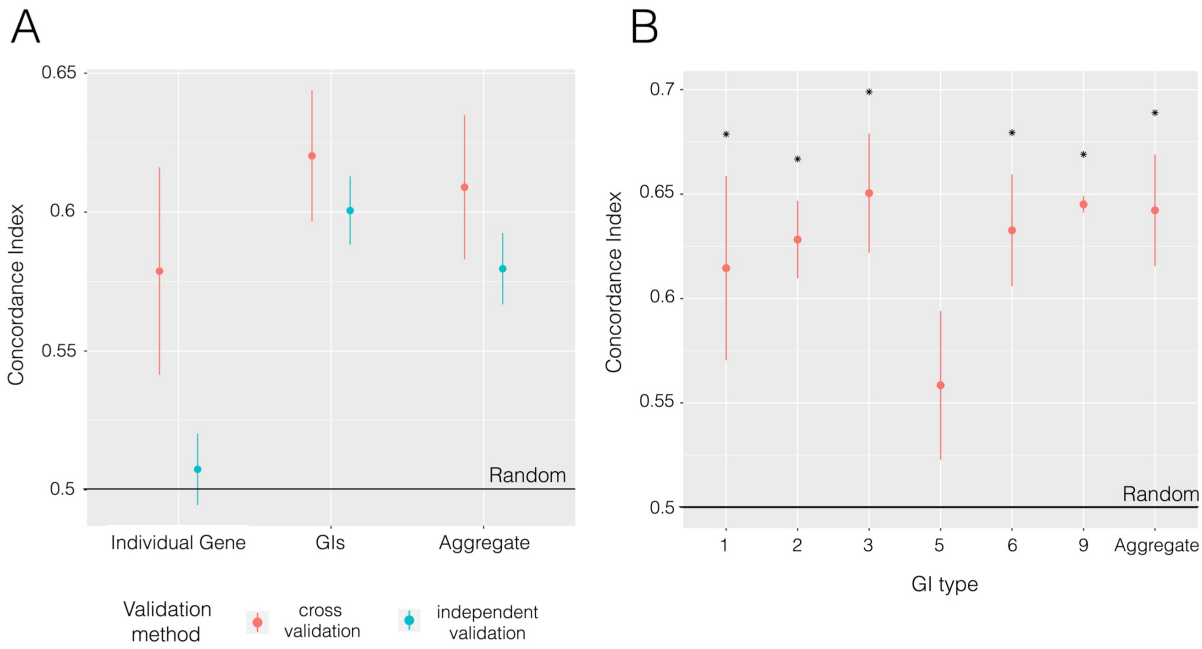


Figure A.4: GI-based approach survival risk prediction performance.
(A-B) Survival prediction accuracy. The prediction accuracy (y-axes) is measured by the concordance of predicted and observed survival (C-index), where 0.5 represents the null expectation. **(A)** C-index comparison between the individual gene-wise approach, the GI-based approach and the approach that aggregates both individual gene signals and the GIs. The accuracy is estimated based on cross-validation within TCGA (Weinstein et al., 2013) (red) as well as in an independent METABRIC (Curtis et al., 2012a) breast cancer dataset (blue). The individual gene approach performs poorly in cross validation and fails in the independent validation while the GIs achieve superior accuracy in both validation approaches. **(B)** Comparison of cross-validation prediction accuracies when the 6 different GIs types are used in isolation. 'All' on x-axis refers to overall GI-based prediction accuracy using all six types (Significant predictions relative to null expectation (p-value < 0.01) are marked by an asterisk).

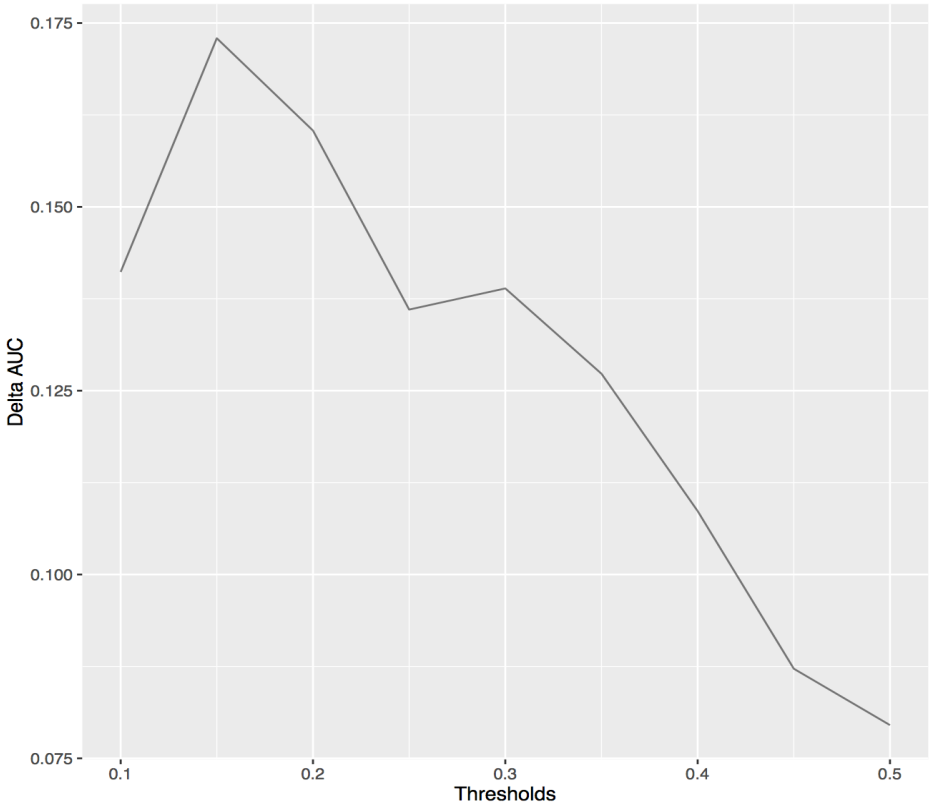


Figure A.5: Survival analysis performance scores based on an alternative metric. We dichotomized the extreme (at varying thresholds from 10% to 50%) predicted low- and high-risk groups and quantified the difference in their area under their Kaplan Meyer (KM) survival curves.

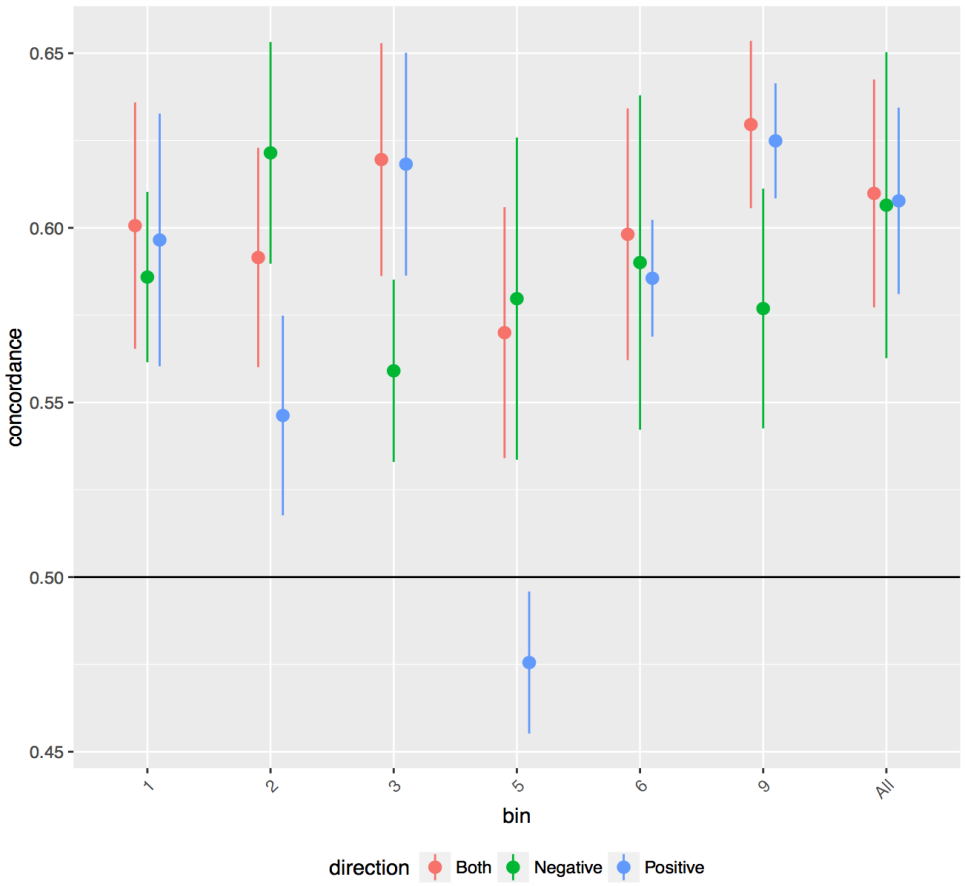


Figure A.6: GI-based approach survival risk prediction performance across all GI types.
GI-based approach survival risk prediction performance restricting to GIs in each of the 6 bins and further segregated them into those with positive and negative effects on survival.

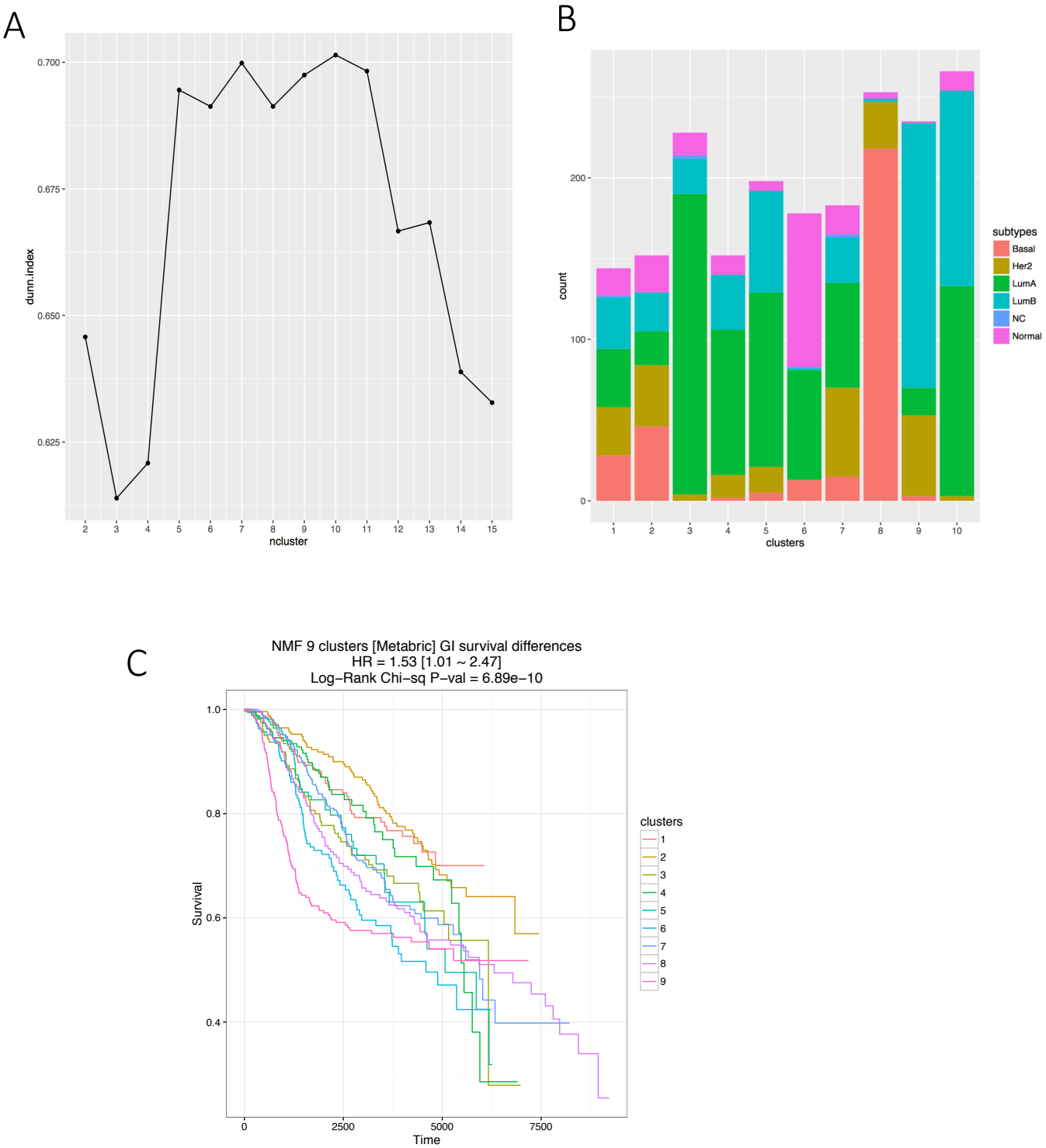
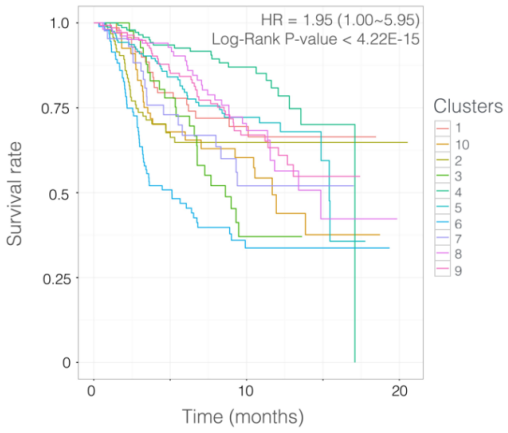
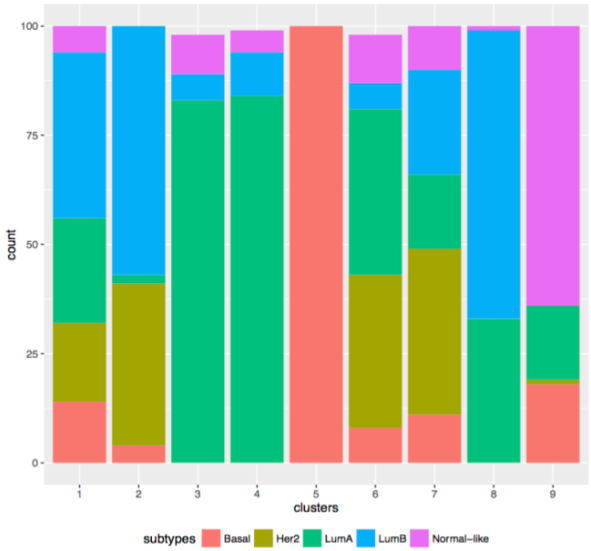


Figure A.7: GI-based Breast cancer patient clustering analysis.
(A) Dunn's index clustering quality score for varying number of clusters. **(B)** Breast cancer subtype distribution in NMF clustering using 10 clusters. **(C)** Survival characteristics of clustering analysis performed using the full GI network.

A


HR = 1.95 (1.00~5.95)
Log-Rank P-value < 4.22E-15
Survival rate
Time (months)
Clusters
1
10
2
3
4
5
6
7
8
9

B


count
clusters
subtypes
Basal
Her2
LumA
LumB
Normal−like

C

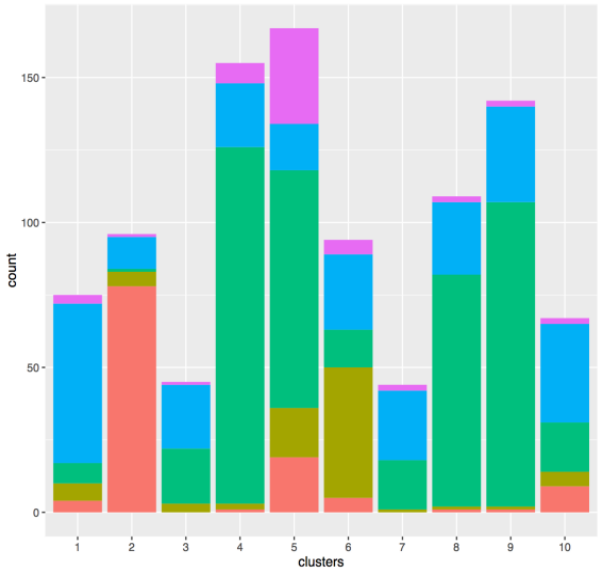

count
clusters

Figure A.8: Cluster clinical subtype composition based on PAM50 breast cancer subtyping.
(A) METABRIC (Curtis et al., 2012a) original publication clustering survival trends. **(B)** Cluster composition based on refined GI clusters, **(C)** Cluster composition provided in the original METABRIC (Curtis et al., 2012a) publication.

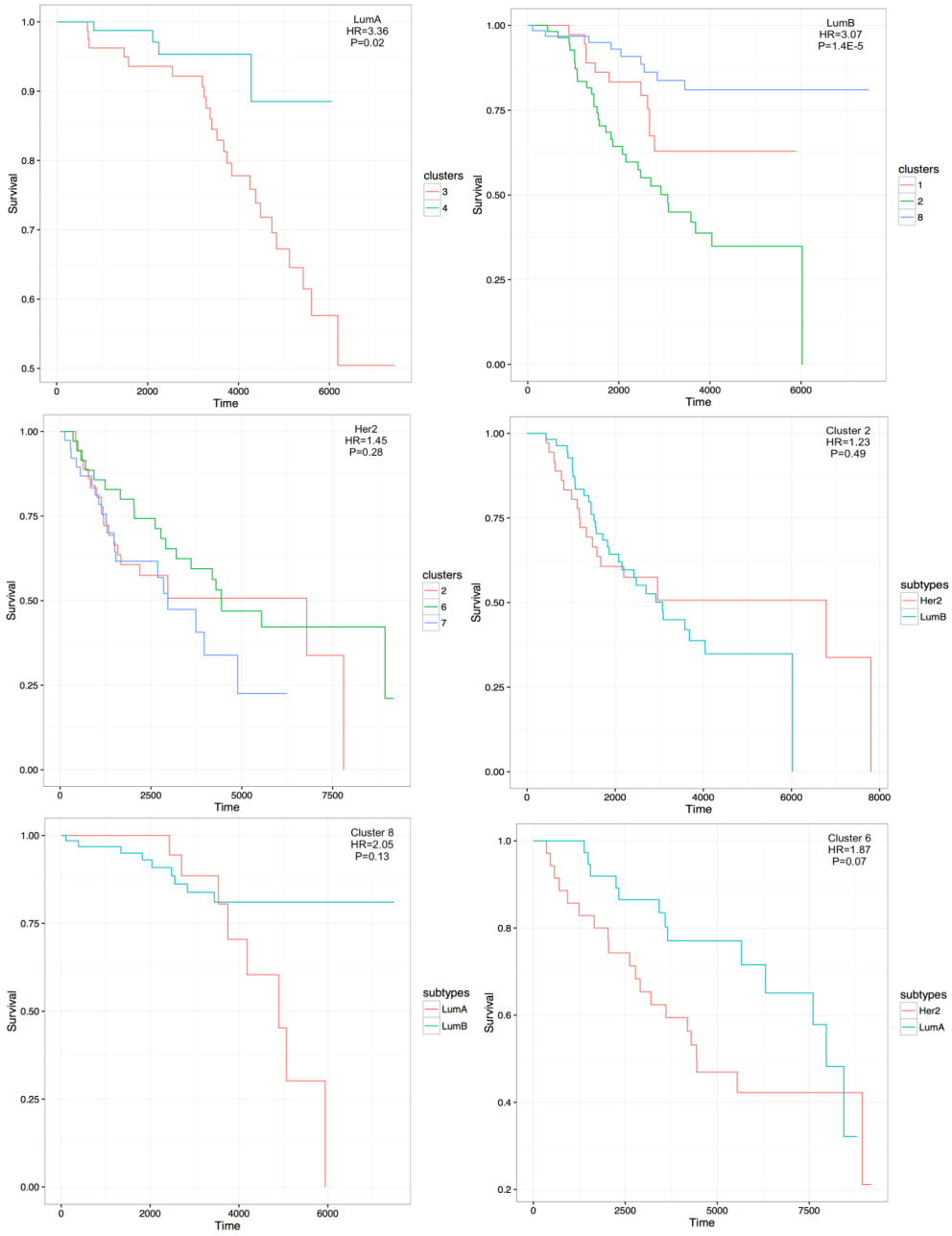


Figure A.9: GI clustering accuracy measures relative to histopathological types.
This Figure shows six tested cases containing at least 30 samples including: three cases of known histopathological breast cancer subtypes are split across multiple SPAGE-based clusters (of which two exhibit statistically significant distinct survival trends) and three cases of SPAGE-based clusters harboring multiple known histopathological subtypes (of which two do not show a significant difference in their survival trends, consistent with the expectation).

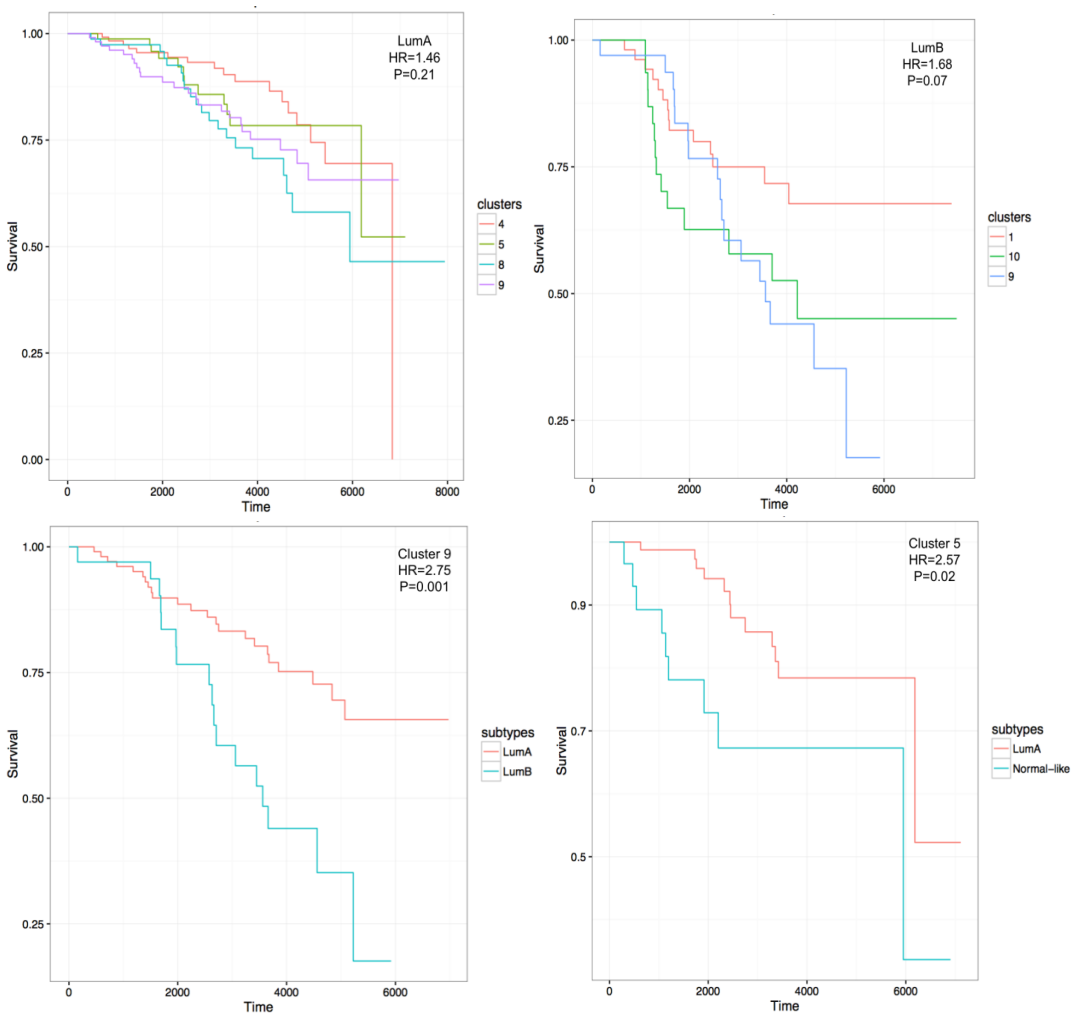


Figure A.10: METABRIC clustering accuracy relative to histopathological types. Among four tested cases containing at least 30 samples, two cases of known histopathological breast cancer subtypes are split across multiple METABRIC clusters (of which neither exhibits statistically significant distinct survival trends) and two cases of METABRIC clusters harboring multiple known histopathological subtypes (of which both show a significant difference in their survival trends, inconsistent with the expectation).

Methods

Data origins and bin construction

We downloaded The Cancer Genome Atlas (TCGA (Weinstein et al., 2013)) (Chang et al., 2013) molecular profiles and clinical covariates via the Broad Firehose (https://gdac.broadinstitute.org/, downloaded on Jan 28, 2016). This covers RSEM-normalized RNAseq data, mutation, and clinical information such as age, sex, race, tumor types, and overall survival of the 8,749 patients (data quality testing is described in **Appendix A Extended Results 5**). Drug response information was downloaded from TCGA (Weinstein et al., 2013) data portal available in the form of RECIST criteria (Eisenhauer et al., 2009) and mapped using DrugBank (Law et al., 2014) database V4.0. For the drug response analysis, to consider only gene inactivation mechanism, we excluded those drugs whose DrugBank mechanism of action label is either potentiator, inducer, positive allosteric modulator, intercalation, stimulator, positive modulator, activator, partial agonist, or agonist.

The PIN was obtained from a previously published resource called HIPPIE (version 2.0, http://cbdm-01.zdv.uni-mainz.de/~mschaefer/hippie/), which aggregates physical protein interaction data from 10 source databases and 11 studies (Schaefer et al., 2012). This network consists of 15,673 human proteins and 203,159 interacting pairs.

We performed gene expression binning specifically for each combination of cancer type, race and gender. Furthermore, we control for various clinical and demographic group specific effects. Because using small groups of samples may result in insufficiently robust models, we filtered rare combinations of clinical and demographic groups, resulting in 5,157 mRNA samples derived from patients spanning 18 cancer types, 3 races and 2 genders. The data were not stratified by stage and grade for three reasons: 1) the grade is missing for most cancer types 2) the stage/grade system varies across tumor types 3) further stratification would

result in further loss of data due to small group sizes. We applied quantile normalization within each sample of the expression data. The METABRIC (Curtis et al., 2012a) breast cancer dataset (Curtis et al., 2012) (as described in reference (Jerby-Arnon et al., 2014)) consists of 1989 microarray samples and was used for independent validation. Similarly, quantile normalization was applied in each sample

Identification of genetic interactions associated with cancer patient survival

As shown in **Figure 2.1**, we divided the range of each gene's expression across tumor samples into 3 equal-sized bins that correspond to the 3 activity states: low, medium and high expression. Given a gene pair, each tumor sample is thus mapped into one of the 9 joint activity states of the two genes. The choice of dividing a gene's activity into three classes, while somewhat arbitrary, was made in consideration of interpretability of functional states and robustness of inference. However, to account for differences in expression distributions across clinical and demographic confounders, we apply a subpopulation specific binning approach. We considered the following categorical confounders: cancer type, race, and gender. We considered the combination of confounder states for which there were at least 100 tumor samples (**Table A.5**). Our binning is thus not confounded by various clinical and demographic variables.

The GI pipeline consists of three steps that successively refine the predictions to arrive at high-confidence set of predicted GIs. As such, the number of all pair-wise combinations of genes is excessively large to apply a comprehensive Cox regression model. For the principal analysis shown in the manuscript, specific parameter thresholds were chosen to make the subsequent analysis tractable, but users of the GI pipeline may choose other thresholds. To inform such decisions, we did a robustness analysis of the parameter settings with a smaller input set of gene pairs (**Appendix A Extended Results 1**).

Step 1: Log-Rank. In the first step, for each of the ~163 million gene pairs, say (x,y), we compute the Log-rank statistics (Harrington and Fleming, 1982) estimating the survival difference between the samples that map to one of the 9 activity bins and the other 8 bins. We implemented the log-rank test in C++ for computational speed. To control for gene-wise effect, we compare the Log-rank statistics of the gene pair (x,y) (in a bin) with those for (x,U) and separately with those for (V,y), where U and V represent all other genes. For a candidate gene pair (x,y), we consider Log-rank(X,Y) to be significant if it is among the top 0.1% relative to all (x,U) and top 0.1% among all (V,y) gene pairs. This threshold of 0.1% (1/1000) can be controlled by the user. We retain a gene pair if it is deemed significant in any of the 9 bins. This procedure retained 223,946 gene pairs of the total of ~163M.

Step 2: Molecular enrichment and depletion. For a gene interaction having positive (respectively, negative) effect on survival, we expect the tumor having that interaction to be under negative (respectively, positive) selection, and therefore we expect the fraction of such tumors (i.e., those mapped to the corresponding activity bin relative to the interacting gene pair to be depleted (respectively, enriched). We only retained the potentially interacting gene pairs for which the fraction of samples in a particular bin, suggested by the log-rank test, were lower (bottom 45 percent) or respectively, greater (top 45 percent among all gene pairs) than expectation, reducing the number of GIs to 271,096 across 179,444 gene pairs. Recall that a gene pair can participate in multiple GIs corresponding to multiple bins and effect on survival. The threshold of 45% can be controlled by the user. Our choice of threshold ascertained that molecular enrichment/depletion is consistent with log-rank test without being overly punitive.

Step 3: Cox proportional hazard test. The Cox proportional hazards model is the most widely accepted approach for modeling survival while accounting for censored data as well as confounding factors. For each gene pair passing the filter at step 2, we modeled its effect on survival in each of the 9 bins using the interaction status λ (active if the sample mapped to the

bin and inactive otherwise), along with the confounders. Specifically, we introduced the expression levels of the two individual genes $\sigma_1,\ \sigma_2$ to model each gene's independent effect on survival, and additionally, clinical and demographic confounders, namely, cancer type, race, gender, and age. The model is stratified based on the discrete confounders, to account for differences in the baseline hazard (risk) characteristics. We did not control for tumor stage and grade as these classifications reflect the very same tumor characteristics our model aims to capture, and such control would prevent us from learning an important element of the disease. Control for genomic-stability and tumor purity as a potential confounder is described in **Appendix A Extended Results 4**.

$$risk \sim \sigma_1 + \sigma_2 + \lambda + age + strata(type, gender, race)$$

Cox modeling provides a p-value representing the significance of the effect of joint gene pair activity on survival. To obtain a null distribution for the p-values, we repeated this process for corresponding list of randomly shuffled pairs (only among the pairs qualifying step 2 above). We retained ~71K gene pairs in the above the most significant 99th quantile of the null p-values distribution as an empirical FDR control.

Optional Step 4: Filtering by protein interactions. To gain additional confidence in the predicted GIs, given the greater tendency (and expectation) for neighbors in the Protein Interaction Network (PIN) to exhibit functional interactions (see Results), we further refined the GI set by retaining the pairs that are found within distance of 2 in the HIPPIE PIN. Overall, we obtain a set of 1704 GIs that exhibit molecular and clinical evidence in cancer as well as evidence from the PIN network.

Survival Risk Assessment

We applied a semi-supervised approach to assign a risk score to each patient according to the functionally active GIs in the sample. Consider a GI involving genes x and y conferring a

positive effect on the tumor fitness in a particular bin B (**Figure 2.1**). If in a sample, genes x and y fall in bin B, then the GI is said to be 'functionally active' in the sample, and a score of +1 is contributed to the overall tumor 'fitness'. Likewise, if the GI has negative effect on the tumor fitness, then a score of -1 is contributed. The overall risk score given a set of GIs is the sum of the individual GI +1 or -1 scores. For each sample, we computed the overall risk score (either in a bin-specific and effect direction-specific fashion, or overall). However, to make our approach comparable to gene-wise approaches (Yuan et al., 2014), we assigned each gene the sum of the contributions by all active GIs involving that gene, with multiplicity for gene pairs involved in multiple GIs. The estimated gene-wise GI score is used as a predictor variable in a Cox model along with the confounding factors discussed above to predict patient's survival.

For cross validation, this model was trained on the same data used for the GIs training and then validated on its cross-validation counterpart. For independent validation, the model was trained on the full TCGA (Weinstein et al., 2013) dataset, and tested on the independent METABRIC (Curtis et al., 2012a) breast cancer data with 1989 samples (Curtis et al., 2012). The prediction accuracy is estimated in terms of the C-index (Harrell et al., 2005). Several previous publications have assessed survival risk prediction accuracy based on dichotomized analysis where samples are separated into distinct low- and high-survival groups and their survival curves then compared (Harrington and Fleming, 1982), which is prone to overestimating prediction accuracy. For comparison, we also performed accuracy estimation following the dichotomized comparison of survival risks between the extreme cases of predicted survival risk groups, for variable thresholds to define the extreme (such as top versus bottom 10% or top versus bottom 20% and so on, **Figure A.5**).

To compare the GI-based survival prediction with the individual gene approach, we implemented an analogous scheme for individual genes where the gene expression values were discretized into 3 expression levels (low, medium and high), and the discretized representation

was used as a predictor variable in a controlled Cox regression model to obtain the significance (p-value) of each gene with respect to survival prediction. The most significant predictors (top 5%) were chosen and precisely used as described for the GIs survival prediction procedure. An analogous procedure was used to estimate the prediction accuracy based on both individual gene effects and GIs.

Identifying gene target(s)-specific GIs

To investigate the GIs involving specific genes of interest (e.g., one or more target genes inhibited by a drug), we used a modified gene set-specific FDR approach. For a set of one or more genes X, we compare the GI significance (Cox regression p-value) of GIs involving any gene in X across the quadruples derived from step 2 (Molecular enrichment/depletion). We defined significant GI interactions as those where the GI significance is more extreme than (lower p-value, higher quantile) the 90^{th} quantile of shuffled GIs involving any member of X.

Applications of genetic interactions

Characterization of differential GI activation between drug-response groups

We retrieved the drug response data as explained in the first subsection of Methods; some patients have response information, and some do not. We inferred the GIs involving each drug's known target gene based on the TCGA (Weinstein et al., 2013) samples that do not have that drug's response information to avoid circularity and filtered based on FDR restricted to the target-specific GIs (using target-specific FDR above). For each of the drug-specific GIs, we compared its activation frequency (whether the GI was functionally active or inactive) among the responders (stable disease, partial response and complete response categories) and non-responders group (clinical progressive disease categories), using one-sided Fisher's exact test (Fisher, 1922), where the alternative hypothesis was that negative (respectively, positive) GIs

are more frequently active among responders (respectively, non-responders). However, given the extremely small and imbalanced sample sizes, and the conservative nature of Fisher's exact test (Berkson, 1978), we tested whether the overall distribution of the obtained odds-ratios are lower than those obtained using randomly shuffled drug-response labels, using one-sided Wilcoxon tests (Wilcoxon, 1945). We thus obtained a p-value for each drug-cancer type pair, segregated by GI type.

Then, for each gene pair in the inferred GIs list and, as a control, in a shuffled list of size 10x as the original GIs list size, we computed the distance between the genes in the PPI network (Schaefer et al., 2012). We then used one-sided Wilcoxon tests (Wilcoxon, 1945) to assess whether GI gene pairs are closer to each other than random expectation. Alternatively, we also compare the number of directly GI gene pairs having direct interaction using one-sided Fisher's exact test (Fisher, 1922).

Characterization of tissue-specific effect of cancer driver genes

A study (Rubio-Perez et al., 2015) of genes' somatic mutation profile across cancer types has identified and characterized the tissue specificity of 459 candidate drivers. For each such candidate, we matched the driver role annotation (oncogene or tumor-suppressor) obtained from the Cosmic Census (Futreal et al., 2004) cancer genes dataset, to obtain a set of 33 tumor suppressors and 25 oncogenes matching the tissue/tumor type annotations. For each of these 58 genes, we calculated the significant GI interactions involving this gene (target-specific FDR). We further excluded genes with 5 or fewer interactions or with 300 or fewer samples where they are expressed, reducing the set of genes of interest to 20 tumor suppressors and 15 oncogenes spanning 10 cancer types (**Table A.3**). For a gene, a sample-specific risk score was calculated based on the functionally active GI partners of the gene (as above for the drug response analysis above), but only considering high activity bin for oncogenes and low-activity

bins for tumor suppressors. For each gene, the cancer types are partitioned into affected types (cancer types affected by the driver) and the other unaffected cancer types. Using a one-sided Wilcoxon rank-sum test, we tested for higher risk score in the samples in the affected cancer type in comparison to those in unaffected types. After correcting for multiple hypotheses testing, 15 out of the 35 (~43%) driver genes were found to have significant tissue-specific GI-based risk score (FDR q-value < 0.1, **Table A.3**).

Breast cancer tumor stratification

We represent a tumor sample as a vector indicating the functional activity status of each predicted GI. This provides a survival-cognizant alternative to the widely-used gene expression profile representation of a sample. We used this representation to partition the METABRIC (Curtis et al., 2012a) (as well as independently for TCGA (Weinstein et al., 2013)) breast cancer patients into clusters using Non-Negative Matrix Factorization (NMF using the brunet algorithm and assigning each sample to the cluster with the highest weight) (Lee and Seung, 2000; Paatero and Tapper, 1994), which has suitable statistical properties and has been shown to be effective in a variety of contexts (Lee and Seung, 1999). Since NMF requires a predetermined number of clusters, we performed the analysis for 2-15 clusters, and assessed their fitness using Dunn's index (Dunn†, 1974), which quantifies compactness within and separation across clusters. The hazard-ratio significance values were computed for each pair of clusters, while the p-values were generated using multi-class log-rank test. For comparison purposes, to match our estimated clusters' sizes to previously published METABRIC (Curtis et al., 2012a) cluster sizes (~900 samples), we constrained the number of samples in each cluster to the 1000 samples that were found to be most highly associated with the cluster. Each cluster's GI profiles were constructed as follows. Our clustering approach – NMF, assigned each GI to a single cluster. For each cluster, and for each of the 12 GI-types (6 bins in **Figure**

2.1 and the two directional effects on survival), we obtain the frequency of GIs of that type, relative to all GI assigned to the clusters.

Mutation frequency analysis was performed on the TCGA (Weinstein et al., 2013) clusters. We defined the gene-wise mutation frequency as the fraction of samples in the cluster in which the gene has a mutation predicted to be deleterious as explained in the next paragraph. Then, we tested whether the mutation frequency distribution of each gene differs significantly across clusters using Chi-square tests. The genes with significant Chi-square statistic (FDR q-value < 0.1) were then used to illustrate the mutational profiles of the clusters. Each mutation's predicted effect on the protein function was obtained from the cBioPortal repository. Out of the 196 differentially mutated genes, 138 genes had matching extended mutation information indicating their SIFT (sorts intolerant from tolerant amino acid substitutions) and PolyPhen (polymorphism phenotyping) predictions. We calculated a gene-wise fraction of mutations predicted to have a significant effect on the protein, separately for SIFT and PolyPhen.

The breast cancer subtypes were derived using the widely accepted PAM50 algorithm (Bernard et al., 2009). The METABRIC (Curtis et al., 2012a) PAM50 subtypes were annotated in the original publication (Curtis et al., 2012), while the TCGA (Weinstein et al., 2013) breast cancer subtypes were calculated using the original published class centroids (Bernard et al., 2009).

Software distribution

The EnGIne software is available on GitHub [https://github.com/asmagen/SPAGEfinder]. The GI network is accessible online via a web portal [https://amagen.shinyapps.io/spage/].

Appendix B – CD4$^+$ T Cell Diversity

Extended Figures

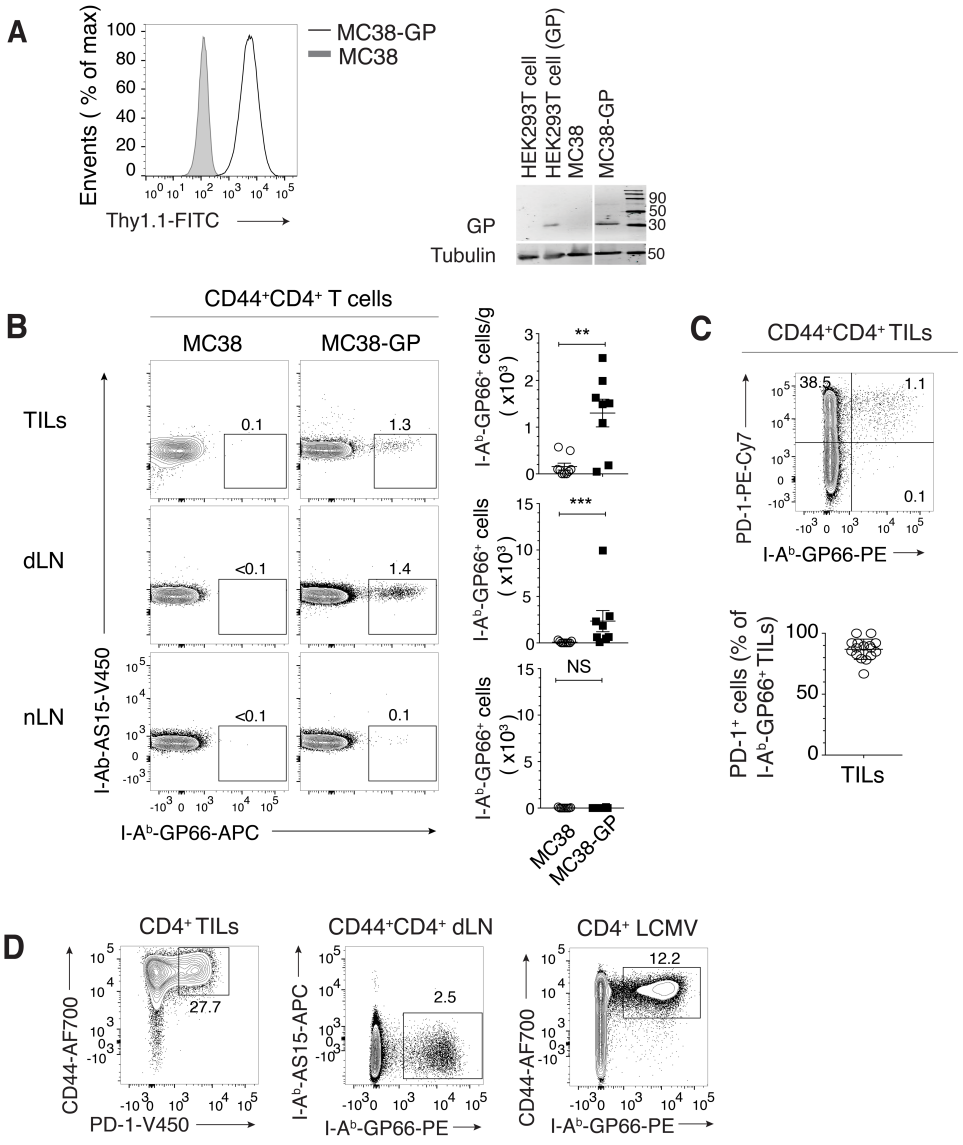


Figure B.1: Characterization of antigen-specific CD4$^+$ T cell responses in MC38 colon adenocarcinoma tumors.

(A) Left panel shows overlaid protein expression of Thy1.1 in MC38 and MC38-GP cells. Right panel shows immunoblot analysis of GP protein expression in HEK293T cells,

HEK293T cells transfected with pMRX-GP-IRSE-Thy1.1 plasmid, MC38 cells or MC38-GP cells.

(B) C57BL/6 mice were subcutaneously injected MC38 or MC38-GP cells and analyzed at day 14 post-injection. Left panel shows protein expression contours of GP66 vs. control (AS15 peptide from *T. gondii*) class II tetramer staining in TILs, dLN and nLN from MC38 and MC38-GP tumor-bearing mice. Right panel shows the number of $GP66^+$ TILs per gram of tumor and total number of $GP66^+$ dLN and nLN cells, separately for MC38 and MC38-GP tumor-bearing mice (Unpaired Mann-Whitney U test, ** $p < 0.01$, *** $p < 0.001$, NS: not significant).

(C) Top panel shows protein expression contours of GP66 tetramer staining vs. PD-1 in TILs. Bottom panel shows the percentage of $PD\text{-}1^+$ cells out of $GP66^+$ TILs.

(D) GP66-specific $CD44^+CD4^+$ splenocytes were isolated from WT animals 7 days post-infection with LCMV Armstrong. Protein expression contour of populations used for scRNAseq captures from MC38-GP tumor-bearing mice (left: TILs PD-1 vs. CD44, middle: dLN GP66 vs. AS15 control) and LCMV Armstrong infected mice (right: GP66 vs. CD44).

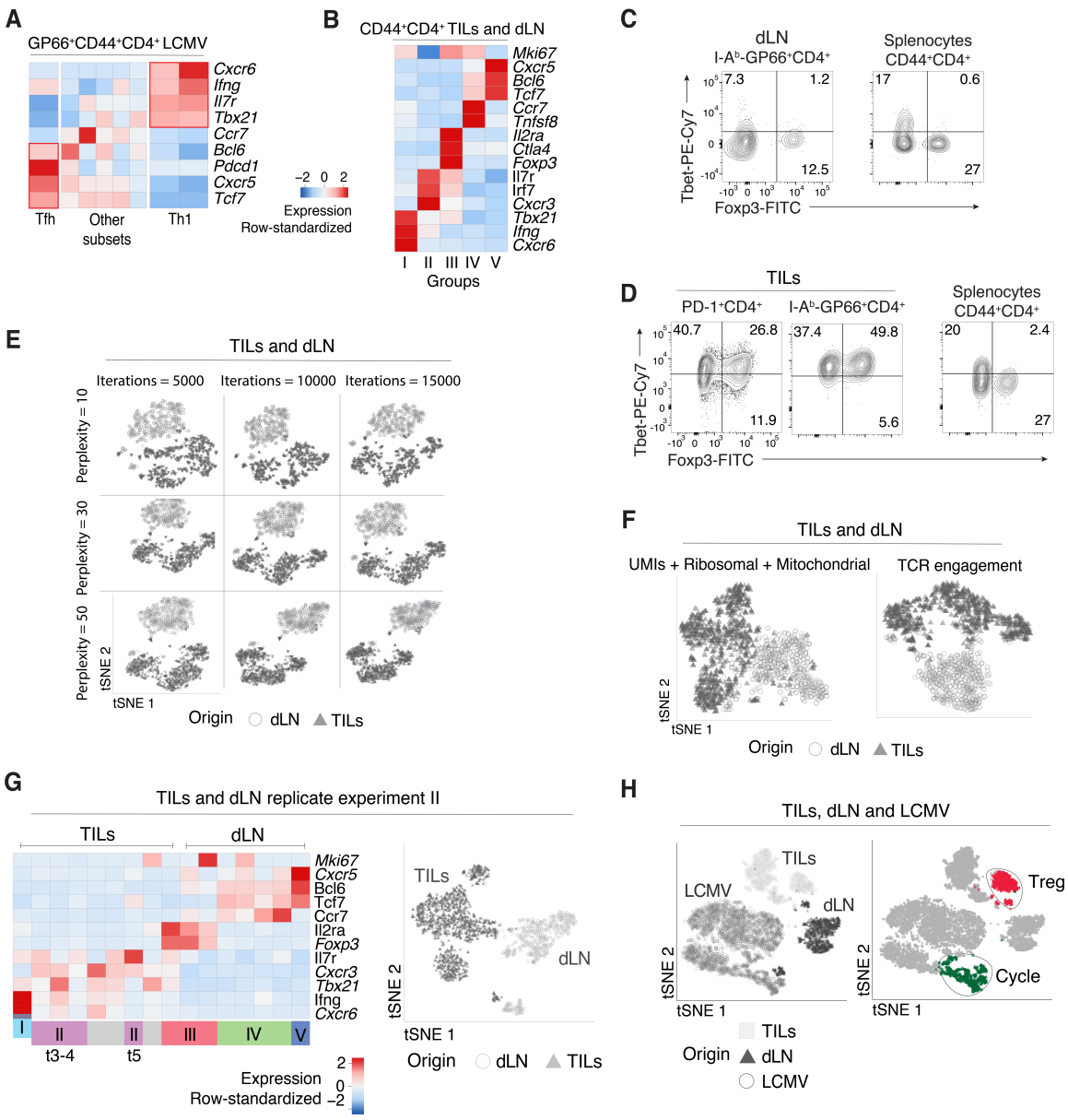


Figure B.2: Characterization of immune responses to LCMV and MC38-GP by scRNAseq.

(A) GP66-specific CD4$^+$ splenocytes from WT animals 7 days post-infection with LCMV Armstrong analyzed by scRNAseq. Heatmap shows row-standardized expression of selected genes across LCMV clusters.

(B-G) TILs and dLN cells from WT mice at day 14 post MC38-GP injection analyzed by scRNAseq. **(B)** Heatmap shows row-standardized expression of selected genes across main TIL and dLN groups (as defined in text). **(C)** Protein expression contours of FoxP3 vs. Tbet in CD44$^+$GP66$^+$ dLN cells (**left**) and in CD44$^+$CD4$^+$ splenocytes from tumor-free mice control (**right**). **(D)** Protein expression contours of FoxP3 vs. Tbet in PD-1$^+$ and GP66$^+$ TILs (**left**) and in CD44$^+$CD4$^+$ splenocytes from tumor-free mice control (**right**). **(E)** tSNE display

of TILs and dLN cells generated using different parameter combination of perplexity and number of iterations, grey-shaded by tissue origin. **(F)** tSNE displays of TILs and dLN cells, grey-shaded by tissue origin, post confounder corrections. **(G)** scRNAseq analysis of TILs and dLN cells from replicate experiment II. Heatmap shows row-standardized expression of selected genes across TIL and dLN clusters (**left**). tSNE display of TILs and dLN cells, grey-shaded by tissue origin (**right**).

(H) TILs, dLN and LCMV cells from replicate experiments I and II analyzed by scRNAseq. tSNE plots show TILs, dLN, and LCMV cells, grey-shaded by origin (**left**) or color-coded by Treg or cell-cycle (Cycle) clustering assignment (grey for all other clusters) (**right**).

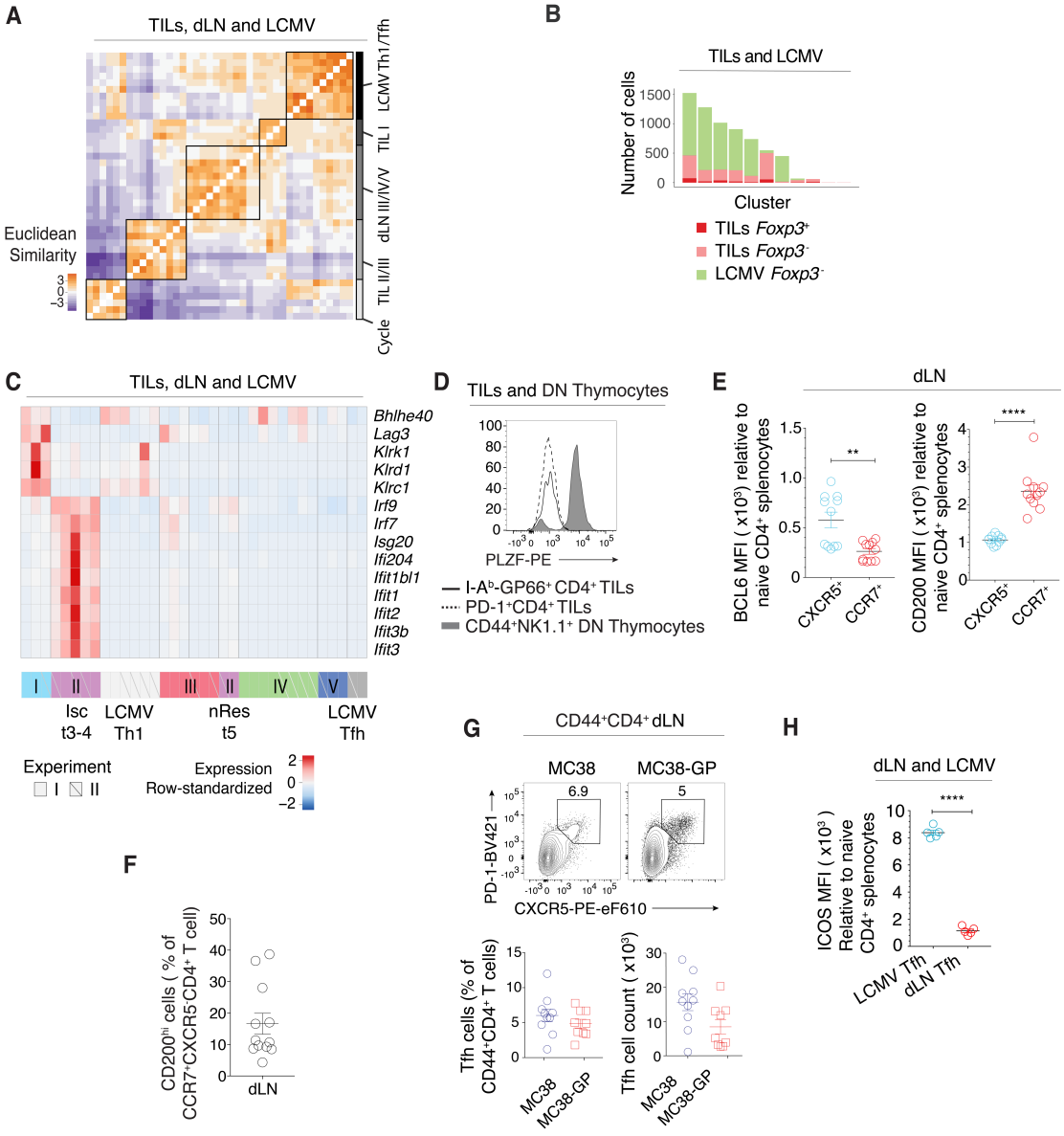


Figure B.3: Assessment of tissue-context-specific effects on clustering analyses and TILs-dLN heterogeneity.

(A-C) TILs, dLN and LCMV cells from replicate experiments I and II analyzed by scRNAseq. **(A)** Heatmap shows Euclidean similarity between cluster-specific average expression vectors (as defined in text) (**left**) annotated with cluster origin and cluster group or type (**right**). **(B)** Bar plot shows relative cluster composition of FoxP3^{+} or FoxP3^{-} TILs and FoxP3^{-} LCMV (no FoxP3^{+} cells found in GP66^{+} LCMV) after applying a data integration approach (Butler et al., 2018). **(C)** Heatmap shows row-standardized expression of TIL Isc and Th1 characteristic genes across TIL, dLN and LCMV clusters.

(D) Overlaid protein expression of PLZF in GP66^{+} and PD-1^{+} TILs and CD44^{+}NK1.1^{+} DN (double negative CD4^{-}CD8^{-}) thymocytes from tumor-free mice control.

(E) Mean fluorescence intensity (MFI) of BCL6 and CD200 in $CXCR5^+$ or $CCR7^+GP66^+$ dLN cells relative to naive $CD4^+$ splenocytes from tumor-free mice control (Unpaired t-test, ** $p < 0.005$, **** $p < 0.0001$).

(F) Percentage of $CD200^{hi}$ cells out of $CCR7^+CXCR5^+$ dLN cells.

(G) Top panel shows protein expression contours of CXCR5 vs. PD-1 in $CD44^+CD4^+$ dLN cells from MC38 and MC38-GP tumor-bearing mice. Bottom panel shows percentage of Tfh cells out of total $CD44^+CD4^+$ T cells in dLN (**left**) and total number of Tfh cells (**right**).

(H) Mean fluorescence intensity (MFI) levels of ICOS in LCMV Tfh and dLN Tfh relative to naive $CD4^+$ splenocytes from tumor-free mice control (Unpaired t-test, $p < 10^{-5}$).

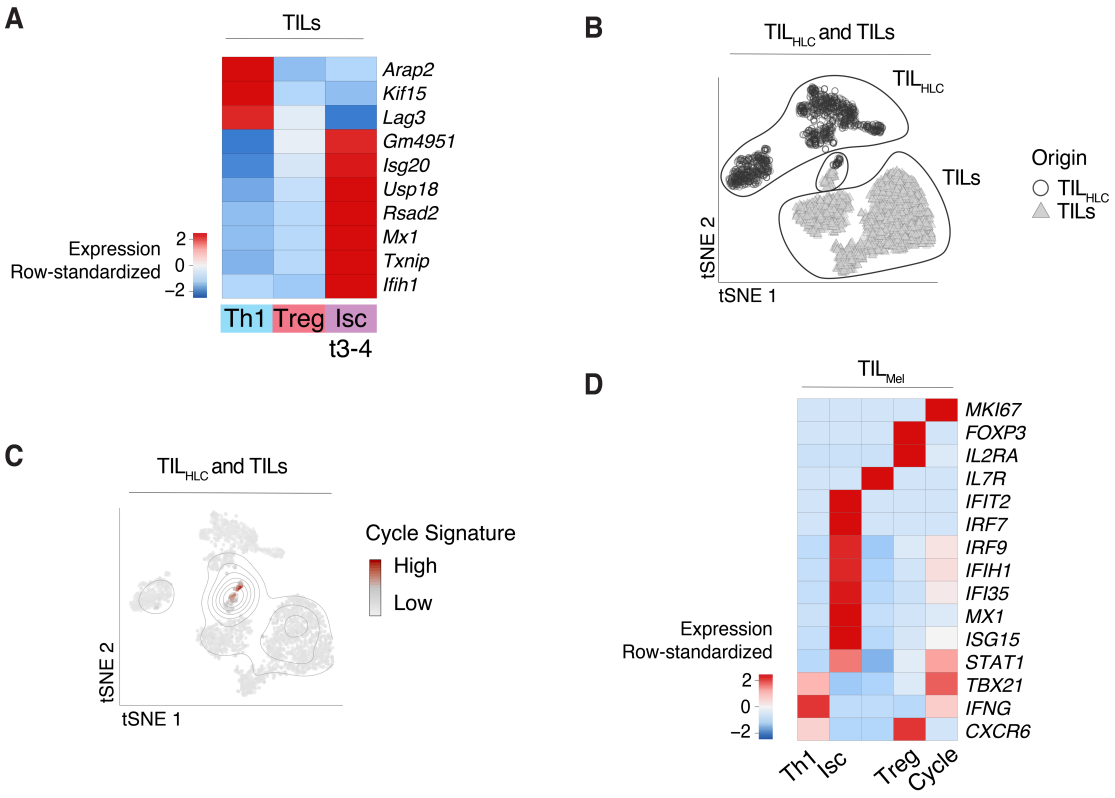


Figure B.4: Correspondence to human data and dysfunction gene signatures.

(A) Heatmap shows row-standardized expression of selected exhaustion genes across TIL Th1, Treg and Isc clusters (respectively clusters t1-2, t6-7 and t3-4 as shown in **Figure 3.1A**).

(B-C) Analysis of TIL_{HLC} and TILs (as defined in text). **(B)** tSNE plots show cells grey-shaded by origin. **(C)** tSNE plots show cells color-coded by cell cycle signature activation level.

(D) Analysis of TIL_{Mel} (as defined in text). Heatmap shows row-standardized expression of selected TIL characteristic genes across TIL_{Mel} clusters.

Extended data

GP66-tetramer binding results in potential cross-linking of and signaling by the TCR of GP66-specific T cells. To model the transcriptomic effect of TCR engagement as a result of GP66-tetramer-based purification, we sought to compare LCMV-specific $CD4^+$ T cells obtained either after GP66-tetramer purification or without tetramer-based purification. To enrich in such cells without tetramer staining, we noted that ~94% of GP66-specific $CD4^+$ splenocytes from LCMV-infected mice express little or no IL7R [IL-7 receptor α chain] (**Figure B.5A**). Thus, we considered that most $CD44^{hi}CD4^+Il7R^+$ splenocytes were not LCMV-specific, and sorted $CD44^+IL7R^-$ (LCMV $IL7R^-$) T cells for scRNAseq; in addition to antigen-specific $CD44^+$ GP66-tetramer purified (LCMV $GP66^+$) T cells (**Figure B.5B**). Pooled clustering of the two samples revealed 2 (out of 6) clusters heavily dominated by stained cells (**Figure B.5C, top**), suggesting staining bias limited to those clusters. As expected from GP66 tetramer engagement with the TCR, GP66-specific clusters were characterized by genes involved in T cell receptor signaling and NFKB signaling (**Table B.9**), while clusters containing cells from both samples displayed features of Tfh and Th1 cells (**Figure B.5C, bottom**). We designated the GP66-characteristic genes as the TCR engagement GP66 signature (**Table B.10**) and regressed the activation scores of the signature from the expression matrix using a linear regression model fitted to each gene.

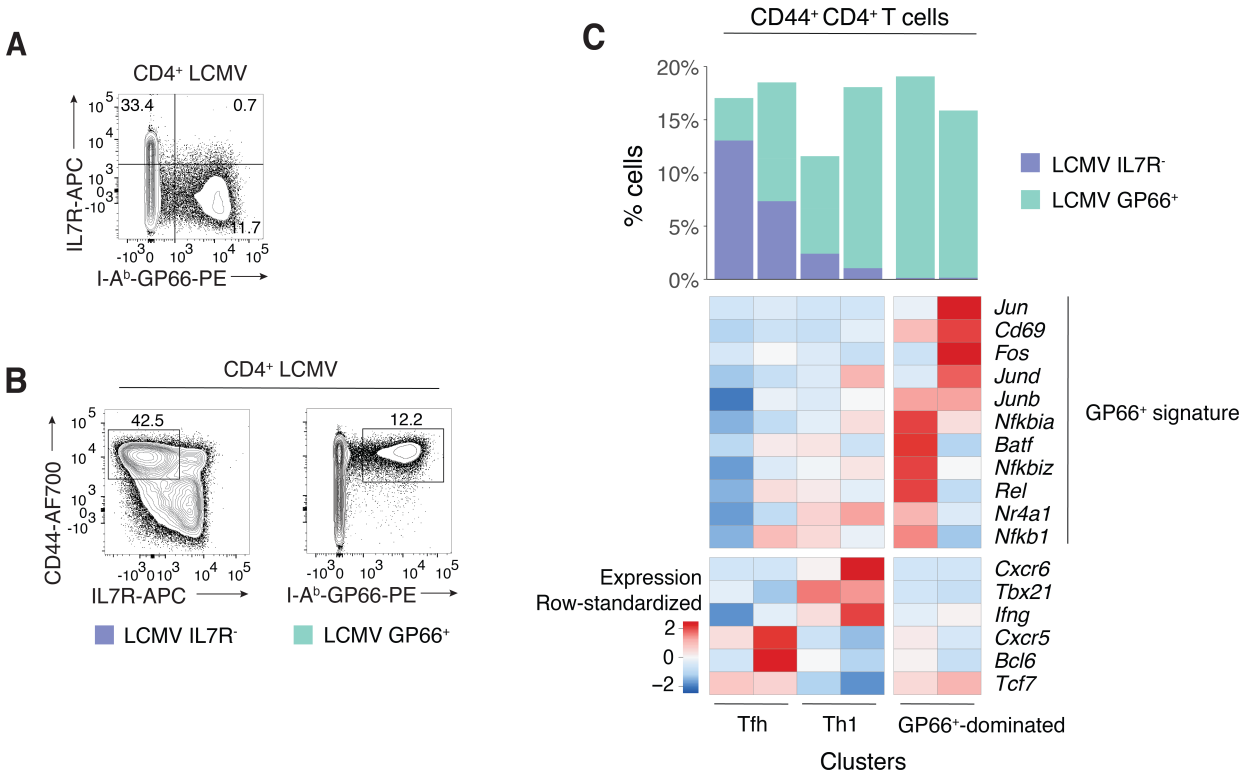


Figure B.5: Transcriptomic effects of TCR engagement as a result of GP66-tetramer-based purification.

(A-B) Analysis of $CD4^+$ splenocytes from C57BL/6 animals 7 days post-infection with LCMV Armstrong. **(A)** Protein expression contour of GP66 tetramer staining vs. IL7R in $CD4^+$ LCMV cells. **(B)** Protein expression contours of IL7R vs. CD44 (for LCMV $IL7R^-$ sample, **left**) and GP66 vs. CD44 (for LCMV $GP66^+$ sample, **right**).

(C) LCMV $IL7R^-$ and LCMV $GP66^+$ cells analyzed by scRNAseq. Heatmap shows row-standardized expression of selected genes across pooled LCMV $IL7R^-$ and LCMV $GP66^+$ clusters (**bottom**). Bar plot indicates the number of LCMV $IL7R^-$ and LCMV $GP66^+$ cells in each cluster relative to the total number of cells (**top**).

Methods

scRNAseq data processing

De-multiplexing, alignment to the mm10 transcriptome and unique molecular identifier (UMI) calculation were performed using the 10X Genomics Cellranger toolkit (v2.0.1, http://software.10xgenomics.com/single-cell/overview/welcome). Pre-processing, dimensionality reduction and clustering analyses procedures were applied to each dataset (that is, specific tissue origin in each experiment) independently to account for dataset-specific technical variation such as sequencing depth and biological variation in population composition, as follows. We filtered out low quality cells with fewer than 500 detected genes (those with at least one mapped read in the cell). Potential doublets were defined as cells with number of detected genes or number of UMIs above the 98th quantile (top 2% owing to up to 2% estimated doublets rate in the 10X Chromium system). Potentially senescent cells (more than 10% of the reads in the cell mapped to 13 mitochondrial genes) were also excluded. Library size (LS_j, number of UMIs in cell j) normalization and natural log transformation were applied to each cell library, i.e., $norm_j^i = \ln\left(\frac{raw_j^i}{LS_j} + 1\right)$ to quantify the expression of gene i in cell j, where raw_j^i is the number of reads for gene i in cell j.

Highly variable genes were defined as genes with greater than one standard deviation of the dispersion from the average expression of each gene. However, to account for heteroscedasticity, variable genes were identified separately in bins defined based on average expression. PCA analysis was performed on the normalized expression of the set of dataset-specific highly variable genes. We selected the top PCs based on gene permutation test (Buja and Eyuboglu, 1992). 'Barnes-hut' approximate version of t-SNE (van der Maaten, 2014) (perplexity set to 30, 10k iterations) was applied on the top PCs to obtain a 2D projection of the data for visualization.

Gene signature activation was quantified relative to a technically similar background gene set as described in (Haber et al., 2017). Briefly, we identify the top 10 most similar (nearest neighbours) genes in terms of average expression and variance, then define the signature activation as the average expression of the signature genes minus the average expression of the background genes. GP66 tetramer staining signature definition is described in **Appendix B Extended Data**. Additionally, we defined lists of ribosomal, mitochondrial, and cell cycle genes (Kowalczyk et al., 2015) for confounder controls (**Table B.10**).

High-resolution clustering analysis

Phenograph clustering (Levine et al., 2015) using the top PCs (see dimensionality reduction) was performed independently on each dataset to allow full control of the clustering resolution based on dataset-specific coverage and heterogeneity features. The clustering resolution (number of clusters) is controlled by the K nearest neighbour (KNN) parameter. We designed a simulation analysis to estimate the optimal clustering resolution, i.e., at what resolution the clustering is superior in quality to clustering driven by technical biases inherent to scRNAseq, as follows. Here we define the clustering quality as the clustering modularity reported by Phenograph, which indicates intra-cluster compactness and inter-cluster separation. The simulations consist of repeating the clustering analysis on 100 shuffled expression matrices to estimate the 'null' distribution of the clustering quality, where the gene expression measurements are permuted within each cell to retain the cell-specific coverage biases. We repeated this process for varying value of the KNN parameter k to compare the clustering modularity of the original O_k to the shuffled S_k data. The final resolution was defined as the maximal resolution where $\frac{O_k}{S_k} \geq 2$. Pooled clustering analysis (joint rather than separated by dataset) and visualization was performed using PCA on the aggregate list of highly variable genes defined on each dataset. Clustering was done with and without controlling for

confounding factors (number of UMIs, number of detected genes and gene signatures activation of ribosomal, mitochondrial, cell cycle and GP66 staining signature).

After obtaining the initial clusters and identifying the overexpressed genes in each cluster, we apply two filters: (1) we exclude small clusters of B cells ($Cd79^{+}$ populations) from each dataset. (2) We identify PCs driven by B cell marker genes and remove the individual cells whose expression profile has high scores for those PCs (outliers). We then repeat the entire processing and clustering to prevent detecting highly variable genes and PCs driven by contaminations, which may in turn reduce the signal of other small populations of interest.

Population matching analysis

Differential expression was performed using Limma (version 3.32.10). We initially performed differential expression analysis between each cluster against the pool of all other clusters within a given dataset. Identified clusters were labelled as a known T cell subtype if the majority of the known subtype-defining genes were differentially over-expressed in that cluster. We then matched populations across experiments to assess the reproducibility of the populations and to uncover similarities across datasets that are masked due to overall tissue-context-specific differences. To reduce the effects of tissue-context-specific effects on the similarity calculation, we used the fold change (FC) measure of each gene $FC_g^c = \frac{\langle foreground_g \rangle}{\langle background_g \rangle}$ (average of gene g in cluster c (foreground) relative to all other clusters (background) of the same dataset). Then we measured the Pearson correlation between the FC vectors of all pairs of clusters across datasets. We compare this approach with an alternative approach that uses Euclidean distances between the average expression vectors, defined as average expression of all genes in a cluster and a recent data integration approach

(Butler et al., 2018) following tutorial specifications [https://satijalab.org/seurat/immune_alignment.html; version 2.0.1].

Robust cluster calling and population comparisons

For each dataset, we defined 'robust clusters' as those that had highly similar match in the biological replicate. High similarity is defined as Pearson correlation coefficient greater than ~ 1.28 standard deviations from the mean for each dataset, corresponding to nominal p-value of 0.1. Hierarchical clustering was performed on the identified robust clusters using the inter-cluster similarity matrix, where the similarity was defined as above using the Pearson correlation between the FC vectors. Using the vector of average expression vectors did not achieve similar result; specifically, using hierarchical clustering of the Euclidean distances between the clusters average expression vector retained the grouping of clusters based on origin tissue (**Figure B.3A**).

We then analysed differential expression patterns for clusters belonging to each meta-cluster, excluding cell cycle clusters. For a given pair of clusters of interest, A and B in datasets X and Y respectively, we performed three differential expression analyses: (1) differential expression in A relative to other clusters in X, (2) differential expression in B relative to other clusters in Y, and (3) differential expression in A relative to B. In addition to average expression differences, we quantified the detection rate of gene X as proportion of cells where 1 or more reads was mapped to X and prioritized differentially expressed genes exhibiting also differential detection across conditions. This analysis was performed for the two replicates separately and the results interpreted jointly; a gene was deemed as over-expressed in cluster A in tissue X if it is over-expressed relative to other clusters in X as well as relative to B, in both replicates. Selected genes' normalized scRNAseq expression measurements were visualized as contours, where zero (0) values were assigned random value drawn from a normal distribution centered around 0.

Trajectory analysis (Reversed Graph Embedding) of TIL populations (group I and II, excluding group III Tregs) was performed using Monocle (version 2.9.0, parameters max_components = 2, method = DDRTree).

Correspondence to external gene signatures and human data

For each TIL subpopulation (group I Th1, group II Isc, group II nRes and group III Treg) we selected overexpressed genes exhibiting differential detection (as defined above) relative to all other TILs across both experiments (**Table B.4**).

Gene set enrichment analysis of immunologic gene signatures was performed using mSigDB (Liberzon et al., 2015) [C7: immunologic signatures database with clusterProfiler package (version 3.4.3). Other gene signatures discussed throughout were downloaded from the original publications' supplementary materials. Correspondence to Tcmp signature was performed by differential expression of dLN *Ccr7*$^{+}$ clusters n5-6 relative to other dLN and TIL (n1, n7-8, t1-7) rather than dLN subpopulations alone to satisfy the background conditions used in the original publication. IL-27 co-inhibitory gene signature heterogeneity was characterized by gene differential expression analysis across Th1, Isc, and Treg TIL, indicating which genes are preferentially expressed in one subpopulation versus the others.

Human liver cancer TIL scRNAseq counts were downloaded from GEO [GSE98638]. Non-*CD4*$^{+}$ T cells were filtered based on the classification in the original publication (Zheng et al., 2017a). Human gene symbols were translated to Mouse gene symbols using package biomaRt (version 2.37.8). Pre-processing, clustering and population matching analysis were applied as described above. Human melanoma TILs data scRNAseq counts were downloaded from GEO [GSE120575]. We selected *CD4*$^{+}$ T cells as cells with at least one mapped read to *CD4* and [*CD3D* or *CD3E* or *CD3G*], following the authors definition (Sade-Feldman et al.,

2018). 108 out of 136 Isc signature genes were mapped to human gene symbols. The detection rate of each Isc signature gene (as defined above) in each lesion were used to assess differential detection across responders and non-responders. We used two-sided Wilcoxon test to quantify the significance of differential activation.

Software distribution

The computational pipeline is available on [https://github.com/asmagen/robustSingleCell]. The pipeline requires access to Slurm high-performance computing core for efficient simulation analyses.

Experimental procedures

Wet experimental work was performed by Dr. Jia Nie and provided as part of this dissertation for consistency and completeness.

Mice. C57BL/6 mice were purchased from the National Cancer Institute Animal Production Facility and were housed in specific pathogen-free facilities. Animal procedures were approved by the NCI Animal Care and Use Committee.

Cell lines and constructs. MC38 murine colon cancer cell lines (Corbett et al., 1975) were obtained from Jack Greiner's lab and cultured in DMEM that contained 10% heat-inactivated FCS, 0.1 mM nonessential amino acids, 1 mM sodium pyruvate, 0.292mg/ml L-glutamine, 100 pg/ml streptomycin, 100 U/mL penicillin, 10mM Hepes. MC38-GP cells were generated as follows: LCMV-*gp* gene was amplified from pHCMV-LCMV-Arm53b (addgene#15796) and inserted into pMRX-IRSE-Thy1.1 by BamH1 and Not1. Then pMRX-Thy1.1 contained LCMV-*gp* gene was transfected into Plat E cell to package retrovirus. MC38 cell line was transduced by above retrovirus collection and followed by single cell

sorting in 96-well plate after 48hs. The monoclonal cell lines were identified by flow cytometry and western blot.

LCMV infection model and Tumor model. 2 x 10^5 pfu of LCMV Armstrong (Matloubian et al., 1994) were injected intra-peritoneal in 6-12 weeks old C57BL/6 mice. Mice were analysed 7 days post infection. MC38 and MC38-GP tumor cells (0.5×10^6) were s.c. injected in the flank of C57BL/6 mice.

Antibodies. Antibodies for the following specificities were purchased either from Affymetrix Becton-Dickinson Pharmingen or ThermoFisher Ebiosciences: CD4 (RM4.4 or GK1.5), CD8β (H35-17.2), CD45.2 (104), CD45 (30-F11), TCRβ (H57-597), CD5 (53-7.3), B220 (RA3-6B2), Siglec F (E50-2440), NK1.1 (PK136), CD11b (M1/70), CD11c (N418), CD44 (356 IM7), IL7R (A7R34), CCR7 (4B12), CXCR5 (SPRCL5), Bcl6 (K112-91), Lag3 (C9B7W), Cxcr6(SA051D1), CD25(PC61.5), CD278(7E,17G9), PD-1 (J43), Foxp3(FJK-16s), Granzyme B(FGB12), Tbet (4B10), CD200(OX-90). Streptavidin, MHC tetramers loaded with the *Toxoplasma gondii* AS15 (Grover et al., 2012) and LCMV GP66 peptides (AVEIHRPVPGTAPPS and DIYKGVYQFKSV, respectively) were obtained from the NIH Tetramer Core Facility.

Cell preparation and flow cytometry. Lymph node and spleen were prepared and stained as previously described (Wang et al., 2008). For TIL preparation, tumors were dissected 14 to 18 days post-injection, washed in HBSS, cut into small pieces, and subjected to enzymatic digestion with 0.25mg/ml liberase (Roche) and 0.5mg/ml DNAase I (SIGMA) for 30 minutes at 37 degrees. The resulting material were passed through 70um filters and pelleted by centrifugation at 1500rpm. Cell pellets were resuspended in 44% Percoll (GE Healthcare) on an underlay of 67% Percoll and centrifuged for 20min at 1600 rpm without brake. TILs were isolated from the 44%/67% Percoll interface. Following isolation, cells were blocked with anti-FcγRIII/FcγRII (unconjugated, 2.4G2) and subsequently stained for

flow cytometry. Staining for AS15:I-A^b tetramer, GP66:I-A^b tetramer and CXCR5 was performed at 37 degrees for 1 hour prior to staining for other cell surface markers. For intracellular staining, cell surface staining were preformed first, following fixation using the Foxp3-staining kit (eBioscience). Flow cytometry data was acquired on LSR Fortessa cytometers (BD Biosciences) and analysed with FlowJo (TreeStar) software. Dead cells and doublets were excluded by LiveDead staining (Invitrogen) and forward scatter height by width gating. Purification of lymphocytes by cell sorting was performed on a FACS Aria or FACS Fusion (BD Biosciences).

Single cell RNAseq. 3000-13000 T cells sorted from LCMV infected or tumor-bearing mice were loaded on the Chromium platform (10X Genomics) and libraries were constructed with a Single Cell 3′ Reagent Kit V2 according to the manufacturer instructions. Libraries were sequenced on multiple runs of Illumina NextSeq using paired-end 26x98bp or 26x57bp to reach a sequencing saturation greater than 70% resulting in at least 49000 reads/cell.

Bibliography

- Aarntzen, E.H., De Vries, I.J., Lesterhuis, W.J., Schuurhuis, D., Jacobs, J.F., Bol, K., Schreibelt, G., Mus, R., De Wilt, J.H., Haanen, J.B.*, et al.* (2013). Targeting CD4(+) T-helper cells improves the induction of antitumor responses in dendritic cell-based vaccination. Cancer Res *73*, 19-29.
- Agata, Y., Kawasaki, A., Nishimura, H., Ishida, Y., Tsubata, T., Yagita, H., and Honjo, T. (1996). Expression of the PD-1 antigen on the surface of stimulated mouse T and B lymphocytes. Int Immunol *8*, 765-772.
- Ahmadzadeh, M., Pasetto, A., Jia, L., Deniger, D.C., Stevanovic, S., Robbins, P.F., and Rosenberg, S.A. (2019). Tumor-infiltrating human CD4(+) regulatory T cells display a distinct TCR repertoire and exhibit tumor and neoantigen reactivity. Sci Immunol *4*.
- Ahrends, T., Spanjaard, A., Pilzecker, B., Babala, N., Bovens, A., Xiao, Y., Jacobs, H., and Borst, J. (2017). CD4(+) T Cell Help Confers a Cytotoxic T Cell Effector Program Including Coinhibitory Receptor Downregulation and Increased Tissue Invasiveness. Immunity *47*, 848-861 e845.
- Akaogi, K., Nakajima, Y., Ito, I., Kawasaki, S., Oie, S.h., Murayama, A., Kimura, K., and Yanagisawa, J. (2009). KLF4 suppresses estrogen-dependent breast cancer growth by inhibiting the transcriptional activity of ERalpha. Oncogene *28*, 2894--2902.
- Al-Hajj, M., Becker, M.W., Wichal, M., Weissman, I., and Clarke, M.F. (2004). Therapeutic implications of cancer stem cells. Curr Opin Genet Dev *14*, 43-47.
- Alspach, E., Lussier, D.M., and Schreiber, R.D. (2018). Interferon gamma and Its Important Roles in Promoting and Inhibiting Spontaneous and Therapeutic Cancer Immunity. Cold Spring Harb Perspect Biol.
- Aran, D., Sirota, M., and Butte, A.J. (2015). Systematic pan-cancer analysis of tumour purity. Nat Commun *6*, 8971.
- Ashworth, A., Lord, C.J., and Reis-Filho, J.S. (2011). Genetic interactions in cancer progression and treatment. Cell *145*, 30--38.
- Azizi, E., Carr, A.J., Plitas, G., Cornish, A.E., Konopacki, C., Prabhakaran, S., Nainys, J., Wu, K., Kiseliovas, V., Setty, M.*, et al.* (2018). Single-Cell Map of Diverse Immune Phenotypes in the Breast Tumor Microenvironment. Cell *174*, 1293-1308 e1236.
- Bai, A., Hu, H., Yeung, M., and Chen, J. (2007). Kruppel-like factor 2 controls T cell trafficking by activating L-selectin (CD62L) and sphingosine-1-phosphate receptor 1 transcription. J Immunol *178*, 7632-7639.

- Beatty, G., and Paterson, Y. (2001). IFN-gamma-dependent inhibition of tumor angiogenesis by tumor-infiltrating CD4+ T cells requires tumor responsiveness to IFN-gamma. J Immunol *166*, 2276-2282.
- Belkaid, Y., Mendez, S., Lira, R., Kadambi, N., Milon, G., and Sacks, D. (2000). A natural model of Leishmania major infection reveals a prolonged "silent" phase of parasite amplification in the skin before the onset of lesion formation and immunity. J Immunol *165*, 969-977.
- Benhamou, S., and Sarasin, A. (2002). ERCC2/XPD gene polymorphisms and cancer risk. Mutagenesis *17*, 463--469.
- Bernard-Gallon, D., Bosviel, R., Delort, L., Fontana, L., Chamoux, A., Rabiau, N., Kwiatkowski, F., Chalabi, N., Satih, S., and Bignon, Y.-J. (2008). DNA repair gene ERCC2 polymorphisms and associations with breast and ovarian cancer risk. Molecular cancer *7*, 36.
- Bertotti, A., Papp, E., Jones, S., Adleff, V., Anagnostou, V., Lupo, B., Sausen, M., Phallen, J., Hruban, C.A., Tokheim, C.*, et al.* (2015). The genomic landscape of response to EGFR blockade in colorectal cancer. Nature *526*, 263--267.
- Bilal, E., Dutkowski, J., Guinney, J., Jang, I.S., Logsdon, B.A., Pandey, G., Sauerwine, B.A., and Shimoni, Y.a. (2013). Improving Breast Cancer Survival Analysis through Competition-Based Multidimensional Modeling. PLoS Computational Biology *9*.
- Bluestone, J.A., Mackay, C.R., O'Shea, J.J., and Stockinger, B. (2009). The functional plasticity of T cell subsets. Nat Rev Immunol *9*, 811-816.
- Bommi-Reddy, A., Almeciga, I., Sawyer, J., Geisen, C., Li, W., Harlow, E., Kaelin, W.G., and Grueneberg, D.A. (2008). Kinase requirements in human cells: III. Altered kinase requirements in VHL-/- cancer cells detected in a pilot synthetic lethal screen. P Natl Acad Sci USA *105*, 16484--16489.
- Borst, J., Ahrends, T., Babala, N., Melief, C.J.M., and Kastenmuller, W. (2018). CD4(+) T cell help in cancer immunology and immunotherapy. Nat Rev Immunol *18*, 635-647.
- Bos, R., and Sherman, L.A. (2010). CD4(+) T-Cell Help in the Tumor Milieu Is Required for Recruitment and Cytolytic Function of CD8(+) T Lymphocytes. Cancer Research *70*, 8368-8377.
- Brummelman, J., Mazza, E.M.C., Alvisi, G., Colombo, F.S., Grilli, A., Mikulak, J., Mavilio, D., Alloisio, M., Ferrari, F., Lopci, E.*, et al.* (2018). High-dimensional single cell analysis identifies stem-like cytotoxic CD8(+) T cells infiltrating human tumors. J Exp Med *215*, 2520-2535.
- Buja, A., and Eyuboglu, N. (1992). Remarks on Parallel Analysis. Multivariate Behav Res *27*, 509-540.
- Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol *36*, 411-420.
- Butti, R., Gunasekaran, V.P., Kumar, T.V.S., Banerjee, P., and Kundu, G.C. (2019). Breast cancer stem cells: Biology and therapeutic implications. Int J Biochem Cell Biol *107*, 38-52.

- Carlson, C.M., Endrizzi, B.T., Wu, J., Ding, X., Weinreich, M.A., Walsh, E.R., Wani, M.A., Lingrel, J.B., Hogquist, K.A., and Jameson, S.C. (2006). Kruppel-like factor 2 regulates thymocyte and T-cell migration. Nature *442*, 299-302.
- Carmi, Y., Spitzer, M.H., Linde, I.L., Burt, B.M., Prestwood, T.R., Perlman, N., Davidson, M.G., Kenkel, J.A., Segal, E., Pusapati, G.V.*, et al.* (2015). Allogeneic IgG combined with dendritic cell stimuli induce antitumour T-cell immunity. Nature *521*, 99-U254.
- Chambers, A.F., Naumov, G.N., Vantyghem, S.A., and Tuck, A.B. (2000). Molecular biology of breast cancer metastasis - Clinical implications of experimental studies on metastatic inefficiency. Breast Cancer Res *2*, 400-407.
- Chao, J.L., and Savage, P.A. (2018). Unlocking the Complexities of Tumor-Associated Regulatory T Cells. J Immunol *200*, 415-421.
- Chen, R., Valencia, I., Zhong, F., McColl, K.S., Roderick, H.L., Bootman, M.D., Berridge, M.J., Conway, S.J., Holmes, A.B., Mignery, G.A.*, et al.* (2004). Bcl-2 functionally interacts with inositol 1,4,5-trisphosphate receptors to regulate calcium release from the ER in response to inositol 1,4,5-trisphosphate. Journal of Cell Biology *166*, 193--203.
- Chihara, N., Madi, A., Kondo, T., Zhang, H., Acharya, N., Singer, M., Nyman, J., Marjanovic, N.D., Kowalczyk, M.S., Wang, C.*, et al.* (2018). Induction and transcriptional regulation of the co-inhibitory gene module in T cells. Nature *558*, 454-459.
- Citri, A., and Yarden, Y. (2006). EGF-ERBB signalling: towards the systems level. Nat Rev Mol Cell Bio *7*, 505-516.
- Ciucci, T., Vacchio, M.S., Gao, Y., Tomassoni Ardori, F., Candia, J., Mehta, M., Zhao, Y., Tran, B., Pepper, M., Tessarollo, L.*, et al.* (2019). The Emergence and Functional Fitness of Memory CD4(+) T Cells Require the Transcription Factor Thpok. Immunity *50*, 91-105 e104.
- Corbett, T.H., Griswold, D.P., Jr., Roberts, B.J., Peckham, J.C., and Schabel, F.M., Jr. (1975). Tumor induction relationships in development of transplantable cancers of the colon in mice for chemotherapy assays, with a note on carcinogen structure. Cancer Res *35*, 2434-2439.
- Costanzo, M., VanderSluis, B., Koch, E.N., Baryshnikova, A., Pons, C., Tan, G., Wang, W., Usaj, M., Hanchard, J., Lee, S.D.*, et al.* (2016). A global genetic interaction network maps a wiring diagram of cellular function. Science *353*, aaf1420--aaf1420.
- Cousens, L.P., Peterson, R., Hsu, S., Dorner, A., Altman, J.D., Ahmed, R., and Biron, C.A. (1999). Two Roads Diverged: Interferon α/β– and Interleukin 12–mediated Pathways in Promoting T Cell Interferon γ Responses during Viral Infection. The Journal of Experimental Medicine *189*, 1315-1328.
- Crawford, A., Angelosanto, J.M., Kao, C., Doering, T.A., Odorizzi, P.M., Barnett, B.E., and Wherry, E.J. (2014). Molecular and transcriptional basis of CD4(+) T cell dysfunction during chronic infection. Immunity *40*, 289-302.

- Crotty, S. (2015). A brief history of T cell help to B cells. Nat Rev Immunol *15*, 185-189.
- Curtis, C., Shah, S.P., Chin, S.-F., Turashvili, G., Rueda, O.M., Dunning, M.J., Speed, D., Lynch, A.G., Samarajiwa, S., Yuan, Y.*, et al.* (2012a). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups.
- Curtis, C., Shah, S.P., Chin, S.-F., Turashvili, G., Rueda, O.M., Dunning, M.J., Speed, D., Lynch, A.G., Samarajiwa, S., Yuan, Y.*, et al.* (2012b). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature *486*, 346--352.
- Davis, M.M., and Bjorkman, P.J. (1988). T-cell antigen receptor genes and T-cell recognition. Nature *334*, 395-402.
- De Simone, M., Arrigoni, A., Rossetti, G., Gruarin, P., Ranzani, V., Politano, C., Bonnal, R.J.P., Provasi, E., Sarnicola, M.L., Panzeri, I.*, et al.* (2016). Transcriptional Landscape of Human Tissue Lymphocytes Unveils Uniqueness of Tumor-Infiltrating T Regulatory Cells. Immunity *45*, 1135-1147.
- DeNardo, D.G., Barreto, J.B., Andreu, P., Vasquez, L., Tawfik, D., Kolhatkar, N., and Coussens, L.M. (2009). CD4(+) T cells regulate pulmonary metastasis of mammary carcinomas by enhancing protumor properties of macrophages. Cancer Cell *16*, 91-102.
- Dorak, M.T., and Karpuzoglu, E. (2012). Gender differences in cancer susceptibility: an inadequately addressed issue. Front Genet *3*, 268.
- Duhen, T., Duhen, R., Montler, R., Moses, J., Moudgil, T., de Miranda, N.F., Goodall, C.P., Blair, T.C., Fox, B.A., McDermott, J.E.*, et al.* (2018). Co-expression of CD39 and CD103 identifies tumor-reactive CD8 T cells in human solid tumors. Nat Commun *9*.
- Eisenhauer, E.A., Therasse, P., Bogaerts, J., Schwartz, L.H., Sargent, D., Ford, R., Dancey, J., Arbuck, S., Gwyther, S., Mooney, M.*, et al.* (2009). New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1). European Journal of Cancer *45*, 228--247.
- Fidler, I.J. (2003). Timeline - The pathogenesis of cancer metastasis: the 'seed and soil' hypothesis revisited. Nature Reviews Cancer *3*, 453-458.
- Fisher, R.A. (1922). On the Interpretation of χ 2 from Contingency Tables, and the Calculation of P. Journal of the Royal Statistical Society *85*, 87.
- Fritsch, R.D., Shen, X., Sims, G.P., Hathcock, K.S., Hodes, R.J., and Lipsky, P.E. (2005). Stepwise differentiation of CD4 memory T cells defined by expression of CCR7 and CD27. J Immunol *175*, 6489-6497.
- Futreal, P.A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N., and Stratton, M.R. (2004). A census of human cancer genes. Nature reviews Cancer *4*, 177--183.
- Gajewski, T.F., Schreiber, H., and Fu, Y.X. (2013). Innate and adaptive immune cells in the tumor microenvironment. Nat Immunol *14*, 1014-1022.

- Gonzalez-Perez, A., Perez-Llamas, C., Deu-Pons, J., Tamborero, D., Schroeder, M.P., Jene-Sanz, A., Santos, A., and Lopez-Bigas, N. (2013). IntOGen-mutations identifies cancer drivers across tumor types. Nature Methods *10*, 1081--1082.
- Greaves, M., and Maley, C.C. (2012). Clonal evolution in cancer. Nature *481*, 306-313.
- Grover, H.S., Blanchard, N., Gonzalez, F., Chan, S.A., Robey, E.A., and Shastri, N. (2012). The Toxoplasma gondii Peptide AS15 Elicits CD4 T Cells That Can Control Parasite Burden. Infect Immun *80*, 3279-3288.
- Haber, A.L., Biton, M., Rogel, N., Herbst, R.H., Shekhar, K., Smillie, C., Burgin, G., Delorey, T.M., Howitt, M.R., Katz, Y.*, et al.* (2017). A single-cell survey of the small intestinal epithelium. Nature *551*, 333-+.
- Halle, S., Halle, O., and Forster, R. (2017). Mechanisms and Dynamics of T Cell-Mediated Cytotoxicity In Vivo. Trends Immunol *38*, 432-443.
- Hanahan, D., and Weinberg, R.A. (2000). The hallmarks of cancer. Cell *100*, 57-70.
- Hanahan, D., and Weinberg, R.A. (2011). Hallmarks of Cancer: The Next Generation. Cell *144*, 646-674.
- Harris, C.C. (1996). Structure and function of the p53 tumor suppressor gene: Clues for rational cancer therapeutic strategies. Jnci-J Natl Cancer I *88*, 1442-1455.
- Hartwell, L.H., Szankasi, P., Roberts, C.J., Murray, A.W., and Friend, S.H. (1997). Integrating genetic approaches into the discovery of anticancer drugs. Science *278*, 1064--1068.
- Harvey, J.J. (1964). An Unidentified Virus Which Causes the Rapid Production of Tumours in Mice. Nature *204*, 1104-1105.
- Holohan, C., Van Schaeybroeck, S., Longley, D.B., and Johnston, P.G. (2013). Cancer drug resistance: an evolving paradigm. Nature Reviews Cancer *13*, 714-726.
- Hunder, N.N., Wallen, H., Cao, J., Hendricks, D.W., Reilly, J.Z., Rodmyre, R., Jungbluth, A., Gnjatic, S., Thompson, J.A., and Yee, C. (2008). Treatment of metastatic melanoma with autologous CD4+ T cells against NY-ESO-1. N Engl J Med *358*, 2698-2703.
- Im, S.J., Hashimoto, M., Gerner, M.Y., Lee, J., Kissick, H.T., Burger, M.C., Shan, Q., Hale, J.S., Lee, J., Nasti, T.H.*, et al.* (2016). Defining CD8+ T cells that provide the proliferative burst after PD-1 therapy. Nature *537*, 417-421.
- Jerby-Arnon, L., Pfetzer, N., Waldman, Y.Y., McGarry, L., James, D., Shanks, E., Seashore-Ludlow, B., Weinstock, A., Geiger, T., Clemons, P.A.*, et al.* (2014). Predicting cancer-specific vulnerability via data-driven detection of synthetic lethality. Cell *158*, 1199--1209.
- Joshi, N.S., and Kaech, S.M. (2008). Effector CD8 T cell development: A balancing act between memory cell potential and terminal differentiation. Journal of Immunology *180*, 1309-1315.
- Jung, Y.W., Rutishauser, R.L., Joshi, N.S., Haberman, A.M., and Kaech, S.M. (2010). Differential Localization of Effector and Memory CD8 T Cell Subsets

in Lymphoid Organs during Acute Viral Infection. Journal of Immunology *185*, 5315-5325.
- Junttila, M.R., and Evan, G.I. (2009). p53-a Jack of all trades but master of none. Nature Reviews Cancer *9*, 821-829.
- Kaelin, W.G. (2005). The concept of synthetic lethality in the context of anticancer therapy. Nature Reviews Cancer *5*, 689--698.
- Kammertoens, T., Friese, C., Arina, A., Idel, C., Briesemeister, D., Rothe, M., Ivanov, A., Szymborska, A., Patone, G., Kunz, S.*, et al.* (2017). Tumour ischaemia by interferon-gamma resembles physiological blood vessel regression. Nature *545*, 98-+.
- Kawai, T., Forrester, S.J., O'Brien, S., Baggett, A., Rizzo, V., and Eguchi, S. (2017). AT1 receptor signaling pathways in the cardiovascular system, pp. 4--13.
- Kelley, R., and Ideker, T. (2005). Systematic interpretation of genetic interactions using protein networks. Nature Biotechnology *23*, 561--566.
- Kowalczyk, M.S., Tirosh, I., Heckl, D., Rao, T.N., Dixit, A., Haas, B.J., Schneider, R.K., Wagers, A.J., Ebert, B.L., and Regev, A. (2015). Single-cell RNA-seq reveals changes in cell cycle and differentiation programs upon aging of hematopoietic stem cells. Genome Res *25*, 1860-1872.
- Kroll, E.S., Hyland, K.M., Hieter, P., and Li, J.J. (1996). Establishing genetic interactions by a synthetic dosage lethality phenotype. Genetics *143*, 95--102.
- Krueger, A., Zietara, N., and Lyszkiewicz, M. (2017). T Cell Development by the Numbers. Trends Immunol *38*, 128-139.
- Kumar, B.V., Ma, W., Miron, M., Granot, T., Guyer, R.S., Carpenter, D.J., Senda, T., Sun, X., Ho, S.H., Lerner, H.*, et al.* (2017). Human Tissue-Resident Memory T Cells Are Defined by Core Transcriptional and Functional Signatures in Lymphoid and Mucosal Sites. Cell Rep *20*, 2921-2934.
- Kurtulus, S., Madi, A., Escobar, G., Klapholz, M., Nyman, J., Christian, E., Pawlak, M., Dionne, D., Xia, J., Rozenblatt-Rosen, O.*, et al.* (2019). Checkpoint Blockade Immunotherapy Induces Dynamic Changes in PD-1(-)CD8(+) Tumor-Infiltrating T Cells. Immunity *50*, 181-194 e186.
- Lambert, M., Jambon, S., Depauw, S., and David-Cordonnier, M.H. (2018). Targeting transcription factors for cancer treatment.
- Laurens van der Maaten, G.H. (2008). Visualizing Data using t-SNE. Journal of Machine Learning Research, 2579-2605.
- Law, V., Knox, C., Djoumbou, Y., Jewison, T., Guo, A.C., Liu, Y., MacIejewski, A., Arndt, D., Wilson, M., Neveu, V.*, et al.* (2014). DrugBank 4.0: Shedding new light on drug metabolism. Nucleic Acids Research *42*.
- Lazebnik, Y. (2010). What are the hallmarks of cancer? Nature Reviews Cancer *10*, 232-233.
- Lee, J.S., Das, A., Jerby-Arnon, L., Arafeh, R., Auslander, N., Davidson, M., McGarry, L., James, D., Amzallag, A., Park, S.G.*, et al.* (2018). Harnessing synthetic lethality to predict the response to cancer treatment. Nat Commun *9*.

- Levine, J.H., Simonds, E.F., Bendall, S.C., Davis, K.L., Amir el, A.D., Tadmor, M.D., Litvin, O., Fienberg, H.G., Jager, A., Zunder, E.R.*, et al.* (2015). Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. Cell *162*, 184-197.
- Li, H., Fang, Y., Niu, C., Cao, H., Mi, T., Zhu, H., Yuan, J., and Zhu, J. (2018). Inhibition of cIAP1 as a strategy for targeting c-MYC–driven oncogenic activity. Proceedings of the National Academy of Sciences *115*, E9317--E9324.
- Liberzon, A., Birger, C., Thorvaldsdottir, H., Ghandi, M., Mesirov, J.P., and Tamayo, P. (2015). The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst *1*, 417-425.
- Linnemann, C., van Buuren, M.M., Bies, L., Verdegaal, E.M., Schotte, R., Calis, J.J., Behjati, S., Velds, A., Hilkmann, H., Atmioui, D.E.*, et al.* (2015). High-throughput epitope discovery reveals frequent recognition of neo-antigens by CD4+ T cells in human melanoma. Nature medicine *21*, 81-85.
- Lord, C.J., McDonald, S., Swift, S., Turner, N.C., and Ashworth, A. (2010). A high-throughput RNA interference screen for DNA repair determinants of PARP inhibitor sensitivity. DNA Repair *7*, 2010--2019.
- Lu, X., Kensche, P.R., Huynen, M.a., and Notebaart, R.a. (2013). Genome evolution predicts genetic interactions in protein complexes and reveals cancer drug targets. Nat Commun *4*, 2124.
- Luo, J., Emanuele, M.J., Li, D., Creighton, C.J., Schlabach, M.R., Westbrook, T.F., Wong, K.K., and Elledge, S.J. (2009). A Genome-wide RNAi Screen Identifies Multiple Synthetic Lethal Interactions with the Ras Oncogene. Cell *137*, 835--848.
- Malandro, N., Budhu, S., Kuhn, N.F., Liu, C., Murphy, J.T., Cortez, C., Zhong, H., Yang, X., Rizzuto, G., Altan-Bonnet, G.*, et al.* (2016). Clonal Abundance of Tumor-Specific CD4(+) T Cells Potentiates Efficacy and Alters Susceptibility to Exhaustion. Immunity *44*, 179-193.
- Malchow, S., Leventhal, D.S., Nishi, S., Fischer, B.I., Shen, L., Paner, G.P., Amit, A.S., Kang, C., Geddes, J.E., Allison, J.P.*, et al.* (2013). Aire-dependent thymic development of tumor-associated regulatory T cells. Science *339*, 1219-1224.
- Marshall, H.D., Chandele, A., Jung, Y.W., Meng, H., Poholek, A.C., Parish, I.A., Rutishauser, R., Cui, W., Kleinstein, S.H., Craft, J.*, et al.* (2011). Differential expression of Ly6C and T-bet distinguish effector and memory Th1 CD4(+) cell properties during viral infection. Immunity *35*, 633-646.
- Matloubian, M., Concepcion, R.J., and Ahmed, R. (1994). Cd4(+) T-Cells Are Required to Sustain Cd8(+) Cytotoxic T-Cell Responses during Chronic Viral-Infection. J Virol *68*, 8056-8063.
- McLornan, D.P., List, A., and Mufti, G.J. (2014). Applying Synthetic Lethality for the Selective Targeting of Cancer. New England Journal of Medicine *371*, 1725--1735.
- Megchelenbrink, W., Katzir, R., Lu, X., Ruppin, E., and Notebaart, R.a. (2015). Synthetic dosage lethality in the human metabolic network is highly

predictive of tumor growth and cancer patient survival. P Natl Acad Sci USA *112*, 12217--12222.
- Mendelsohn, J., and Baselga, J. (2006). Epidermal growth factor receptor targeting in cancer. Semin Oncol *33*, 369-385.
- Miyamoto, D.T., Zheng, Y., Wittner, B.S., Lee, R.J., Zhu, H., Broderick, K.T., Desai, R., Fox, D.B., Brannigan, B.W., Trautwein, J.*, et al.* (2015). RNA-Seq of single prostate CTCs implicates noncanonical Wnt signaling in antiandrogen resistance. Science (New York, NY) *349*, 1351--1356.
- Mumberg, D., Monach, P.A., Wanderling, S., Philip, M., Toledano, A.Y., Schreiber, R.D., and Schreiber, H. (1999). CD4(+) T cells eliminate MHC class II-negative cancer cells in vivo by indirect effects of IFN-gamma. Proc Natl Acad Sci U S A *96*, 8633-8638.
- Negrini, S., Gorgoulis, V.G., and Halazonetis, T.D. (2010). Genomic instability - an evolving hallmark of cancer. Nat Rev Mol Cell Bio *11*, 220-228.
- Nguyen, D.X., Bos, P.D., and Massague, J. (2009). Metastasis: from dissemination to organ-specific colonization. Nature Reviews Cancer *9*, 274-U265.
- Oakes, S.A., Scorrano, L., Opferman, J.T., Bassik, M.C., Nishino, M., Pozzan, T., and Korsmeyer, S.J. (2005). Proapoptotic BAX and BAK regulate the type 1 inositol trisphosphate receptor and calcium leak from the endoplasmic reticulum. Proceedings of the National Academy of Sciences *102*, 105--110.
- Ott, P.A., Hu, Z., Keskin, D.B., Shukla, S.A., Sun, J., Bozym, D.J., Zhang, W., Luoma, A., Giobbie-Hurder, A., Peter, L.*, et al.* (2017). An immunogenic personal neoantigen vaccine for patients with melanoma. Nature *547*, 217-221.
- Pepper, M., and Jenkins, M.K. (2011). Origins of CD4(+) effector and central memory T cells. Nat Immunol *12*, 467-471.
- Pieper, K., Grimbacher, B., and Eibel, H. (2013). B-cell biology and development. J Allergy Clin Immunol *131*, 959-971.
- Plitas, G., Konopacki, C., Wu, K., Bos, P.D., Morrow, M., Putintseva, E.V., Chudakov, D.M., and Rudensky, A.Y. (2016). Regulatory T Cells Exhibit Distinct Features in Human Breast Cancer. Immunity *45*, 1122-1134.
- Qin, Z., and Blankenstein, T. (2000). CD4+ T cell--mediated tumor rejection involves inhibition of angiogenesis that is dependent on IFN gamma receptor expression by nonhematopoietic cells. Immunity *12*, 677-686.
- Rathert, P., Roth, M., Neumann, T., Muerdter, F., Roe, J.-S.J.S., Muhar, M., Deswal, S., Cerny-Reiterer, S., Peter, B., Jude, J.*, et al.* (2015). Transcriptional plasticity promotes primary and acquired resistance to BET inhibition. Nature *525*, 543--547.
- Ribas, A., and Wolchok, J.D. (2018). Cancer immunotherapy using checkpoint blockade. Science *359*, 1350-1355.
- Rong, Y.-P., Bultynck, G., Aromolaran, A.S., Zhong, F., and Parys, J.B.a. (2009). The BH4 domain of Bcl-2 inhibits ER calcium release and apoptosis

by binding the regulatory and coupling domain of the IP3 receptor. Proceedings of the National Academy of Sciences *106*, 14397--14402.
- Rosenberg, S.A., and Restifo, N.P. (2015). Adoptive cell transfer as personalized immunotherapy for human cancer. Science *348*, 62--68.
- Ruvolo, P.P., Deng, X., and May, W.S. (2001). Phosphorylation of Bcl2 and regulation of apoptosis. Leukemia : official journal of the Leukemia Society of America, Leukemia Research Fund, UK *15*, 515--522.
- Sade-Feldman, M., Yizhak, K., Bjorgaard, S.L., Ray, J.P., de Boer, C.G., Jenkins, R.W., Lieb, D.J., Chen, J.H., Frederick, D.T., Barzily-Rokni, M.*, et al.* (2018). Defining T Cell States Associated with Response to Checkpoint Immunotherapy in Melanoma. Cell *175*, 998-1013 e1020.
- Sajesh, B.V., Guppy, B.J., and McManus, K.J. (2013). Synthetic genetic targeting of genome instability in cancer, pp. 739--761.
- Sakaguchi, S., Yamaguchi, T., Nomura, T., and Ono, M. (2008). Regulatory T cells and immune tolerance. Cell *133*, 775-787.
- Schaefer, M.H., Fontaine, J.F., Vinayagam, A., Porras, P., Wanker, E.E., and Andrade-Navarro, M.A. (2012). Hippie: Integrating protein interaction networks with experiment based quality scores. PLoS ONE *7*.
- Schreiber, R.D., Old, L.J., and Smyth, M.J. (2011). Cancer immunoediting: integrating immunity's roles in cancer suppression and promotion. Science *331*, 1565-1570.
- Schulz, M., Aichele, P., Vollenweider, M., Bobe, F.W., Cardinaux, F., Hengartner, H., and Zinkernagel, R.M. (1989). Major Histocompatibility Complex - Dependent T-Cell Epitopes of Lymphocytic Choriomeningitis Virus Nucleoprotein and Their Protective Capacity against Viral Disease. Eur J Immunol *19*, 1657-1667.
- Segal, N.H., Parsons, D.W., Peggs, K.S., Velculescu, V., Kinzler, K.W., Vogelstein, B., and Allison, J.P. (2008). Epitope landscape in breast and colorectal cancer. Cancer Res *68*, 889-892.
- Siddiqui, I., Schaeuble, K., Chennupati, V., Fuertes Marraco, S.A., Calderon-Copete, S., Pais Ferreira, D., Carmona, S.J., Scarpellino, L., Gfeller, D., Pradervand, S.*, et al.* (2019). Intratumoral Tcf1(+)PD-1(+)CD8(+) T Cells with Stem-like Properties Promote Tumor Control in Response to Vaccination and Checkpoint Blockade Immunotherapy. Immunity *50*, 195-211 e110.
- Simoni, Y., Becht, E., Fehlings, M., Loh, C.Y., Koo, S.L., Teng, K.W.W., Yeong, J.P.S., Nahar, R., Zhang, T., Kared, H.*, et al.* (2018). Bystander CD8(+) T cells are abundant and phenotypically distinct in human tumour infiltrates. Nature *557*, 575-+.
- Sjoblom, T., Jones, S., Wood, L.D., Parsons, D.W., Lin, J., Barber, T.D., Mandelker, D., Leary, R.J., Ptak, J., Silliman, N.*, et al.* (2006). The consensus coding sequences of human breast and colorectal cancers. Science *314*, 268-274.
- Snell, L.M., McGaha, T.L., and Brooks, D.G. (2017). Type I Interferon in Chronic Virus Infection and Cancer. Trends Immunol *38*, 542-557.

- Steckel, M., Molina-Arcas, M., Weigelt, B., Marani, M., Warne, P.H., Kuznetsov, H., Kelly, G., Saunders, B., Howell, M., Downward, J.*, et al.* (2012). Determination of synthetic lethal interactions in KRAS oncogene-dependent cancer cells reveals novel therapeutic targeting strategies. Cell Res *22*, 1227--1245.
- Stuhlmiller, Timothy J., Miller, Samantha M., Zawistowski, Jon S., Nakamura, K., Beltran, Adriana S., Duncan, James S., Angus, Steven P., Collins, Kyla A.L., Granger, Deborah A., Reuther, Rachel A.*, et al.* (2015a). Inhibition of Lapatinib-Induced Kinome Reprogramming in ERBB2-Positive Breast Cancer by Targeting BET Family Bromodomains. Cell Reports *11*, 390--404.
- Stuhlmiller, T.J., Miller, S.M., Zawistowski, J.S., Nakamura, K., Beltran, A.S., Duncan, J.S., Angus, S.P., Collins, K.A.L., Granger, D.A., Reuther, R.A.*, et al.* (2015b). Inhibition of lapatinib-induced kinome reprogramming in ERBB2-positive breast cancer by targeting BET family bromodomains. Cell Reports *11*, 390--404.
- Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S.*, et al.* (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. P Natl Acad Sci USA *102*, 15545-15550.
- Sun, H., Lu, B.F., Li, R.Q., Flavell, R.A., and Taneja, R. (2001). Defective T cell activation and autoimmune disorder in Stra13-deficient mice. Nature Immunology *2*, 1040-1047.
- Szappanos, B., Kovcs, K., Szamecz, B., Honti, F., Costanzo, M., Baryshnikova, A., Gelius-Dietrich, G., Lercher, M.J., Jelasity, M., Myers, C.L.*, et al.* (2011). An integrated approach to characterize genetic interaction networks in yeast metabolism. Nature genetics *43*, 656--662.
- Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K.P.*, et al.* (2015). STRING v10: Protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Research *43*, D447--D452.
- Tanaka, A., and Sakaguchi, S. (2017). Regulatory T cells in cancer immunotherapy. Cell Res *27*, 109-118.
- Thommen, D.S., and Schumacher, T.N. (2018). T Cell Dysfunction in Cancer. Cancer Cell *33*, 547-562.
- Tian, L., Goldstein, A., Wang, H., Lo, H.C., Kim, I.S., Welte, T., Sheng, K.W., Dobrolecki, L.E., Zhang, X.M., Utluri, N.P.*, et al.* (2017). Mutual regulation of tumour vessel normalization and immunostimulatory reprogramming. Nature *544*, 250-+.
- Tran, E., Turcotte, S., Gros, A., Robbins, P.F., Lu, Y.C., Dudley, M.E., Wunderlich, J.R., Somerville, R.P., Hogan, K., Hinrichs, C.S.*, et al.* (2014). Cancer immunotherapy based on mutation-specific CD4+ T cells in a patient with epithelial cancer. Science *344*, 641-645.

- Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N.J., Livak, K.J., Mikkelsen, T.S., and Rinn, J.L. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nature biotechnology *32*, 381--386.
- Turner, N.C., Lord, C.J., Iorns, E., Brough, R., Swift, S., Elliott, R., Rayter, S., Tutt, A.N., and Ashworth, A. (2008). A synthetic lethal siRNA screen identifying genes mediating sensitivity to a PARP inhibitor. The EMBO journal *27*, 1368--1377.
- van der Maaten, L. (2014). Accelerating t-SNE using Tree-Based Algorithms. Journal of Machine Learning Research *15*, 3221-3245.
- van Leeuwen, J., Boone, C., and Andrews, B.J. (2017). Mapping a diversity of genetic interactions in yeast. Curr Opin Syst Biol *6*, 14-21.
- Vogelstein, B., Papadopoulos, N., Velculescu, V.E., Zhou, S.B., Diaz, L.A., and Kinzler, K.W. (2013). Cancer Genome Landscapes. Science *339*, 1546-1558.
- Wang, L., Wildt, K.F., Castro, E., Xiong, Y., Feigenbaum, L., Tessarollo, L., and Bosselut, R. (2008). The zinc finger transcription factor Zbtb7b represses CD8-lineage gene expression in peripheral CD4+ T cells. Immunity *29*, 876-887.
- Wei, S.C., Levine, J.H., Cogdill, A.P., Zhao, Y., Anang, N.A.A.S., Andrews, M.C., Sharma, P., Wang, J., Wargo, J.A., Pe'er, D.*, et al.* (2017). Distinct Cellular Mechanisms Underlie Anti-CTLA-4 and Anti-PD-1 Checkpoint Blockade. Cell *170*, 1120-1133.
- Weinberg, R.A. (1983). Oncogenes and the Molecular-Biology of Cancer. Journal of Cell Biology *97*, 1661-1662.
- Weinberg, R.A. (1991). Tumor suppressor genes. Science *254*, 1138-1146.
- Weinberg, R.A. (1995). The Retinoblastoma Protein and Cell-Cycle Control. Cell *81*, 323-330.
- Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R.M., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., Stuart, J.M., and Network, C.G.A.R. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. Nature Genetics *45*, 1113-1120.
- Wilcoxon, F. (1945). Individual Comparisons by Ranking Methods. Biometrics Bulletin *1*, 80.
- Williams, C.S., Watson, A.J.M., Sheng, H., Helou, R., Shao, J., and DuBois, R.N. (2000). Celecoxib prevents tumor growth in vivo without toxicity to normal gut: Lack of correlation between in vitro and in vivo models. Cancer Research *60*, 6045--6051.
- Wong, S.L., Zhang, L.V., Tong, A.H.Y., Li, Z., Goldberg, D.S., King, O.D., Lesage, G., Vidal, M., Andrews, B., Bussey, H.*, et al.* (2004). Combining biological networks to predict genetic interactions. P Natl Acad Sci USA *101*, 15682--15687.
- Wu, T., Ji, Y., Moseman, E.A., Xu, H.C., Manglani, M., Kirby, M., Anderson, S.M., Handon, R., Kenyon, E., Elkahloun, A.*, et al.* (2016). The TCF1-Bcl6

axis counteracts type I interferon to repress exhaustion and maintain T cell stemness. Sci Immunol *1*.

- Xie, J.H., Nomura, N., Lu, M., Chen, S.L., Koch, G.E., Weng, Y., Rosa, R., Di Salvo, J., Mudgett, J., Peterson, L.B.*, et al.* (2003). Antibody-mediated blockade of the CXCR3 chemokine receptor results in diminished recruitment of T helper 1 cells into sites of inflammation. J Leukoc Biol *73*, 771-780.
- Yu, F., Sharma, S., Jankovic, D., Gurram, R.K., Su, P., Hu, G., Li, R., Rieder, S., Zhao, K., Sun, B.*, et al.* (2018). The transcription factor Bhlhe40 is a switch of inflammatory versus antiinflammatory Th1 cell fate determination. J Exp Med *215*, 1813-1821.
- Yuan, Y., Allen, E.M.V., Omberg, L., Wagle, N., Amin-Mansour, A., Sokolov, A., Byers, L.a., Xu, Y., Hess, K.R., Diao, L.*, et al.* (2014). Assessing the clinical utility of cancer genomic and proteomic data across tumor types. Nature Biotechnology *32*, 644--652.
- Yuan, Y., Liu, L.X., Chen, H., Wang, Y.M., Xu, Y.X., Mao, H.Z., Li, J., Mills, G.B., Shu, Y.Q., Li, L.*, et al.* (2016). Comprehensive Characterization of Molecular Differences in Cancer between Male and Female Patients. Cancer Cell *29*, 711-722.
- Zhang, L., Yu, X., Zheng, L., Zhang, Y., Li, Y., Fang, Q., Gao, R., Kang, B., Zhang, Q., Huang, J.Y.*, et al.* (2018). Lineage tracking reveals dynamic relationships of T cells in colorectal cancer. Nature *564*, 268-272.
- Zheng, C., Zheng, L., Yoo, J.K., Guo, H., Zhang, Y., Guo, X., Kang, B., Hu, R., Huang, J.Y., Zhang, Q.*, et al.* (2017a). Landscape of Infiltrating T Cells in Liver Cancer Revealed by Single-Cell Sequencing. Cell *169*, 1342-1356 e1316.
- Zheng, G.X., Terry, J.M., Belgrader, P., Ryvkin, P., Bent, Z.W., Wilson, R., Ziraldo, S.B., Wheeler, T.D., McDermott, G.P., Zhu, J.*, et al.* (2017b). Massively parallel digital transcriptional profiling of single cells. Nat Commun *8*, 14049.
- Zhong, W., and Sternberg, P.W. (2006). Genome-wide prediction of C. elegans genetic interactions. Science (New York, NY) *311*, 1481--1484.
- Zhou, X.Y., Yu, S.Y., Zhao, D.M., Harty, J.T., Badovinac, V.P., and Xue, H.H. (2010). Differentiation and Persistence of Memory CD8(+) T Cells Depend on T Cell Factor 1. Immunity *33*, 229-240.
- Zhu, J., Yamane, H., and Paul, W.E. (2010). Differentiation of effector CD4 T cell populations (*). Annu Rev Immunol *28*, 445-489.
- Zuniga, E.I., Macal, M., Lewis, G.M., and Harker, J.A. (2015). Innate and Adaptive Immune Regulation During Chronic Viral Infections. Annu Rev Virol *2*, 573-597.