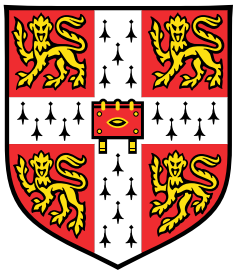
Machine Learning Methods for Cancer Immunology



Leon Chlon

CRUK Cambridge Institute
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Pembroke College November 2017

To Abd'ullah Salem Al Neyadi, your smile may be gone for now, but the happiness you perfused our hearts with radiates into the cosmos and fuels my ambition to make the world a better place. My little brother, this entire volume of work is dedicated to you, your smile lives on in my heart forever, and you are the inspiration behind the next chapter in my life. May Allah watch over your soul and return you to us soon.

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others except in the case of Chapter 3, where some results were derived in collaboration with Dr. Shivan Sivakumar and Dr. Ines de Santiago, and led to the following publication Sivakumar *et al* [1]. The results outlined in section 3.2 outline work done by Dr. Shivan Sivakumar and Dr. Ines de Santiago exclusively. Their work has been included to provide a contextual basis for my own results, as outlined in the remaining sections of chapter 3.

The work in outlined in chapter 5 led to the preprint Chlon *et al.* [2] currently hosted on the bioRxiv repository and awaiting publication - https://doi.org/10.1101/144832. I am the copyright holder of this preprint. Furthermore, I am the only author on this paper alongside my supervisor Dr. Florian Markowetz. All results, figures, text and supporting information are exclusively my own work. Sections 5.3 - 5.5 contain text and figures reproduced from my own preprint Chlon *et al.* [2].

This dissertation contains fewer than 65,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 150 figures.

Leon Chlon
November 2017

This dissertation contains material from the following publications

- Sivakumar, S.*, De Santiago, I.*, **Chlon, L.*** and Markowetz, F.
  Master regulators of oncogenic KRAS response in Pancreatic Cancer: An integrative Network Biology Analysis. 2017.
  PLoS Medicine 14 (1).

- **Chlon, L.**, Markowetz, F.
  Causal Modeling Dissects Tumour–Microenvironment Interactions In Breast Cancer. 2017.
  bioRxiv, 144832.

Other manuscripts published during my PhD or currently in preparation:

- **Chlon, L.***, Louhimo, R.*, Yuan, Y., Hautaniemi, S. and Markowetz F.
  Spatial Patterns of Lymphocyte-Cancer Interactions Predict Patient Survival in Breast Cancer.
  PLoS Computational Biology. (Under Second Review)

- Ali, R., **Chlon, L.**, Pharoah, P., Markowetz, F. and Caldas, C.
  Patterns of Immune Infiltration in Breast Cancer: A Gene-Expression-Based Retrospective Study. 2016.
  PLoS Medicine 13 (12).

Co-first author - *

Acknowledgements

I am very fortunate to have worked with a supervisor as supportive, proactive and intellectually nurturing as Dr. Florian Markowetz. His uncanny ability to combine profound subject knowledge with a much needed touch of comedy made my time in the lab wholly rewarding. My prospective career in academia has benefited greatly by learning from someone with such an unparalleled dedication to his students and science.

I am eternally grateful to Dr. Geoff Macintyre, the paragon of intellectual coolness, for providing much needed 5 minute domain knowledge primers, delivered amicably behind an Aussie accent whilst I was inundated with biological jargon. I owe much to Dr. Federico Giorgi for introducing me to Judea Pearl (in the literary sense) and being a continuing source of bioinformatics wisdom. I thank Dr. Shivan Sivakumar and Dr. Ines de Santiago for being amazing collaborators on the pancreatic cancer subtyping project in addition to great friends. I also thank my fellow PhD survivor Edith Ross for providing an unquestioning line of support during my PhD, both academically and as a master oarswoman. I am indebted to my fellow physicist Dr. Turid Torheim for her discussions on imaging methods and Norse mythology. I was extremely fortunate to learn from one of the most brilliant statisticians in the field, Dr. Oscar Rueda, you are the man.

None of this work would have been possible were it not for the unshakable support of my remarkable father, my superhero mother who never once doubted my ability to succeed, my uncle Yousef, aunty Leila, uncle Hussein and my angelic and patient Teta, a pinnacle of virtue. I am eternally grateful to my uncle Dr. Moustapha Awada for being the scientific role model I have always looked up to. I am indebted to Maggie Chlon, the most intelligent thinker I know, for supplying me with sustained physics discussions to quell my nostalgia. Similarly, Diana Chlon has always been at hand to provide comic relief in days where p exceeded 0.05. My baby brother Jacob, you are the little ball of life and happiness that drove me to succeed.

Most importantly, I owe this entire thesis to a most endearing and incredible woman, whose unfaltering support and sustained encouragement filled each day with unparalleled happiness. Thank you for being a continuing source of inspiration to me. This dissertation is as much yours as it is mine, my dearest Farah.

Abstract

Tumours are highly heterogeneous collections of tissues characterised by a repertoire of heavily mutated and rapidly proliferating cells. Evading immune destruction is a fundamental hallmark of cancer, and elucidating the contextual basis of tumour-infiltrating leukocytes is pivotal for improving immunotherapy initiatives. However, progress in this domain is hindered by an incomplete characterisation of the regulatory mechanisms involved in cancer immunity. Addressing this challenge, this thesis is formulated around a fundamental line of inquiry: how do we quantitatively describe the immune system with respect to tumour heterogeneity?

Describing the molecular interactions between cancer cells and the immune system is a fundamental goal of cancer immunology. The first part of this thesis describes a three-stage association study to address this challenge in pancreatic ductal adenocarcinoma (PDAC). Firstly, network-based approaches are used to characterise PDAC on the basis of transcription factor regulators of an oncogenic *KRAS* signature. Next, gene expression tools are used to resolve the leukocyte subset mixing proportions, stromal contamination, immune checkpoint expression and immune pathway dysregulation from the data. Finally, partial correlations are used to characterise immune features in terms of *KRAS* master regulator activity. The results are compared across two independent cohorts for consistency.

Moving beyond associations, the second part of the dissertation introduces a causal modelling approach to infer directed interactions between signaling pathway activity and immune agency. This is achieved by anchoring the analysis on somatic genomic changes. In particular, copy number profiles, transcriptomic data, image data and a protein-protein interaction network are integrated using graphical modelling approaches to infer directed relationships. Generated models are compared between independent cohorts and orthogonal datasets to evaluate consistency. Finally, proposed mechanisms are cross-referenced against literature examples to test for legitimacy.

In summary, this dissertation provides methodological contributions, at the levels of associative and causal inference, for inferring the contextual basis for tumour-specific immune agency.

Table of contents

Chapter 1

Introduction

This thesis is concerned with studying the interaction between various processes of the human immune system and cancer cells. To facilitate this, two main statistical approaches are proposed that characterise interactions from multiomics datasets. The first approach studies immune-cancer interactions on an association level, identifying features that correlate with immune agency. The second approach moves beyond associations, using probabilistic modeling techniques to identify causal drivers for immune traits. This chapter gives a succinct introduction to tumour heterogeneity, an overview of immunology and how it pertains to cancer and describes several experimental techniques for immune profiling. The end of this chapter contains a summary of the entire thesis.

1.1 The Heterogeneous Architecture of Cancer

Tumour heterogeneity accounts for a majority of the variation in pathology exhibited between cancer samples. Heterogeneity is observed at the genetic level (different combinations of driver and passenger mutations), and at the cellular level (variations in tumour-specific phenotypes). The dysregulation of transcription, translation and signalling mechanisms in cells lead to a characteristic hallmark of cancer termed "immune evasion". Immune system disruption in the context of tumour heterogeneity has yet to be fully characterised, but is becoming better understood in the wake of large-scale genome sequencing projects. Inference of these disruption mechanisms is the central topic of this thesis.

1.1.1 History of Tumour Heterogeneity

Historically, tumours were considered to be homogeneous diseases comprising a uniform population of identical malignant cells with ambiguity regarding their interaction with sur-

rounding tissues. Early categorisation on the basis of histopathology alone failed to explain notable morphological differences in disease presentation or the variability in prognostic parameters such as metastatic potential or survival [3]. Advancements in genetic profiling and other experimental tools have enabled the categorisation of tumour cells into lineages of genotypically and phenotypically distinct populations [4, 5]. This enabled the stratification of patients presenting with similar cancers into subtypes that demonstrated amenability to a variety of novel targeted therapeutics, thus fostering the paradigm of *personalised medicine* [6].

Microarray technology, next generation sequencing and the rapid development of computational methods provide compelling evidence for a genomic basis for tumour heterogeneity. This complements earlier findings by pathologists, who discovered stark intra-tumoural variability in spatial features pertaining to cell shape and tissue morphology in microscopy experiments [7, 8]. Throughout the past decade, high throughput sequencing pipelines enabled the rapid processing of large patient cohorts into large datasets of DNA and mRNA measurements. By mining this data, researchers discovered striking molecular differences between patients regarded as having the same disease pathology. [9, 10]. Tumour heterogeneity ties these observations into a robust theory of disease progression, and is believed to account for the majority of disease phenotypic variability between cancer patients.

1.1.2 Heterogeneity Between Patients

Cancer is widely viewed as an evolutionary disease, whereby accumulated mutations progressively push normal cells into a highly proliferative and malignant state [4], capable of invading surrounding tissues and evading immune destruction. This stems from the accumulation of mutations that disrupts the normal regulation of the cell proliferation machinery. The human genome is approximately 3 billion base pairs long, with over 10 trillion cells that constitute the average human body. In the absence of carcinogens, the normal mutation rate per generation (children and their parents) is approximately 10^{-8} per base pair. Qualitatively, this implies that an offspring's genome will contain a minimum of 30 novel mutations in the period between conception and gamete development [11]. Strikingly, tumour cells sequenced at the exome level demonstrate a burden anywhere between 1 to 1000 somatic aberrations, potentially conferring genome-wide dysfunction [12]. Cancer phylogenies are complex, and no two tumours sampled within a population will share an identical genome or collection of cancer cells as illustrated in Fig. 1.1. This principle is known as *inter-tumoral heterogeneity* and is closely linked to the variability in appearance, behaviour and development of tumours

between patients.

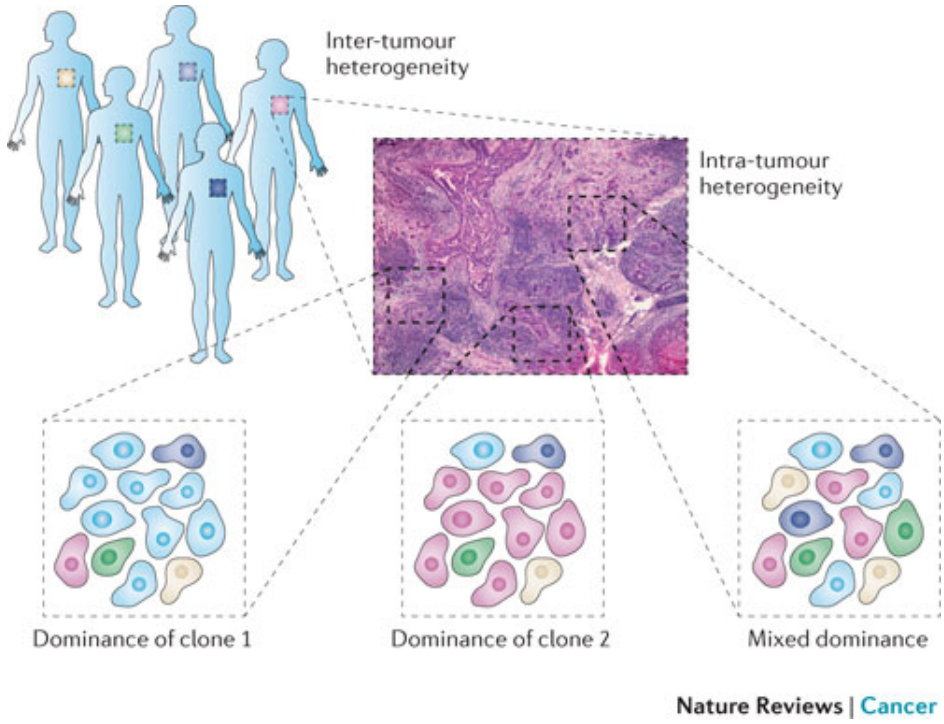


Fig. 1.1 Heterogeneity at the population and tumour level. Here, *clones* refer to subpopulations of genetically distinct cancer cells within the same tumour site. This figure has been reproduced from [8].

Transcription Dysregulation All hierarchies of cellular processing are subject to some kind of regulation, from gene transcription to protein translation to cell signalling. Since these layers are interconnected, dysregulation at the transcriptional level can potentially perturb chains of events along the hierarchy, lending to a molecular biology analogue of the butterfly effect. Transcription factors, for example, regulate the abundance of gene mRNAs, and they are themselves tightly regulated. Transcription factor dysregulation can manifest through aberrant feedback loops, damage to DNA binding sites or aberrant production of intermediary co-factor proteins. As such, transcription factor dysregulation can disrupt normal cell functionality, break homeostasis, and manipulate surrounding tissues through disruption to extracellular signal propagation mechanisms.

Cell Signalling Disorder Normal cells are equipped with a repertoire of molecular "programs" pertaining to cell cycling, migration or local homeostatic maintenance that are

executed during tissue development or repair. Functional studies reveal that such programs are underpinned by complex molecular signalling networks that operate on a cell and tissue-specific manner [13]. Oncogenic mutations disrupt the normal operating protocols of signalling networks, dysregulating mitototic function and reorganising local tissues into a pro-tumourigenic landscape [14]. The normally reticent cross-talk between normal and stromal cells is dominated by a crescendo of reprogrammed signaling, abetting an anti-apoptotic and immunosuppressed environment. [15]. Breaking components associated with antigen presentation, such as defective or poorly expressed members of the MHC class I pathway or transporter associated with antigen processing protein (TAP), is one way tumour cells avoid immune destruction [16] (Fig. 1.2). This results in suppression of tumour-associated antigen presentation, subsequently leading to reduced immune recognition. Furthermore, some tumours overexpress immunosuppresive subfamilies of cytokine and chemokine signalling molecules enhancing immune evasion and promoting proliferation [17].

Mechanisms of immune evasion will be studied in greater detail in section 1.2.2. To conclude this paragraph, we state an obvious observation: since inter-tumoural heterogeneity guarantees distinct genetic profiles between patient tumours, cellular processes will show varying levels of dysregulation, leading to non-uniform immune profiles across cohorts.

1.1.3 Tumour Antigens

All nucleated cells present fragments of proteins known as peptides to the immune system in a process known as *peptide processing and presentation*. An antigen is any peptide capable of evoking an immune reaction and potentially, but not necessarily, lead production of specific antibodies. MHC class I present peptides to CD8+ T cells, while their class II counterparts present to CD4+ T cells. Although similar in functionality, class I deal with intracellular peptides whereas class II deal with exogenous peptides. Fig. 1.2 illustrates the MHC class I pathway, through which degraded peptides are processed and presented on the cell membrane. Firstly, intracellular proteins within the cell cytoplasm are degraded into pep-

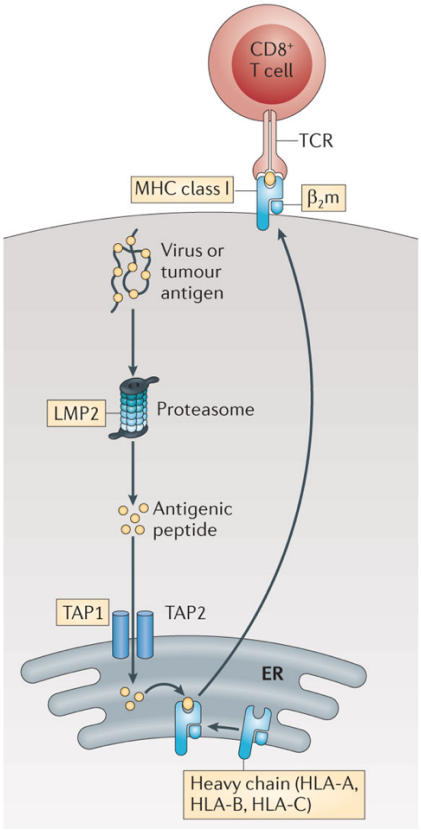


Fig. 1.2 The MHC class I peptide presentation pathway. This figure has been adapted from [18]

tides that translocate to the *endoplasmic reticulum*. They
subsequently bind to the cleavage of MHC I molecules,
which can accommodate peptides of length 8-9 amino acids at most. Finally, the MHC-peptide structure translocates to the cell surface for presentation to CD8+ T cells. Since degradation occurs towards end of the functional life of a protein, the rate of presentation is theoretically proportional to this half life. This is true with the exception of defective ribosomal products (the product of aberrant transcription/translation [19]), which constitute 30% - 70% of all intracellular proteins and which degrade almost immediately after synthesis [20, 21]. Remarkably, a single cell will often express between 10,000-500,000 MHC-peptide molecules on the cell surface [22] for immune inspection. There, cross presentation between the cells and T cells helps decide whether a cell has a suspected pathology and ultimately, if it should be destroyed.

T cells are designed to discriminate between peptide sequences produced by the body (self-antigens) and those which are foreign (nonself) through a concept called *tolerance*. Therefore, a T cell receptor (TCR) has a higher binding affinity towards MHC I molecules presenting nonself antigens than self antigens. Tumour antigens are a class of tumour rejection molecules characterised by two classes of peptides. The first class are self-peptides with incomplete T cell tolerance, which can arise dysregulated patterns of tissue-specific expression. The second class is characterised by peptides not normally present in the genome, and are referred to as *neoantigens*. Neoantigen formation results from the degradation of novel proteins coded for by tumour-altered DNA. This novelty is sufficient to escape central T-cell tolerance and provoke an immune response [23]. Notably, the neoantigen profiles of patients are considered unique, since mutational profiles are not shared at high enough frequencies between patients to create meaningful clusters. Therefore, tumour neoantigens heterogeneity is one way of explaining downstream heterogeneity in immune agency.

A principle question posed by cancer immunology asks: do all neo-antigens elicit an immune response? T cell reactivity studied in the context of human melanoma found that only a small handful of mutations lead to the formation of robust neoantigens [25, 26, 27]. Qualitatively, adequate T-cell reactivity was achieved when melanomas exhibited a somatic mutation burden rate of at least 10 mutations per coding DNA megabase. Extrapolating this to other human malignancies, researchers found T cell reactivity potentially correlates with mutational burden [24]. A seminal paper by Alexandrov *et al* found large variances in mutational burden profiles between different types of cancers [28], and based on reactivity

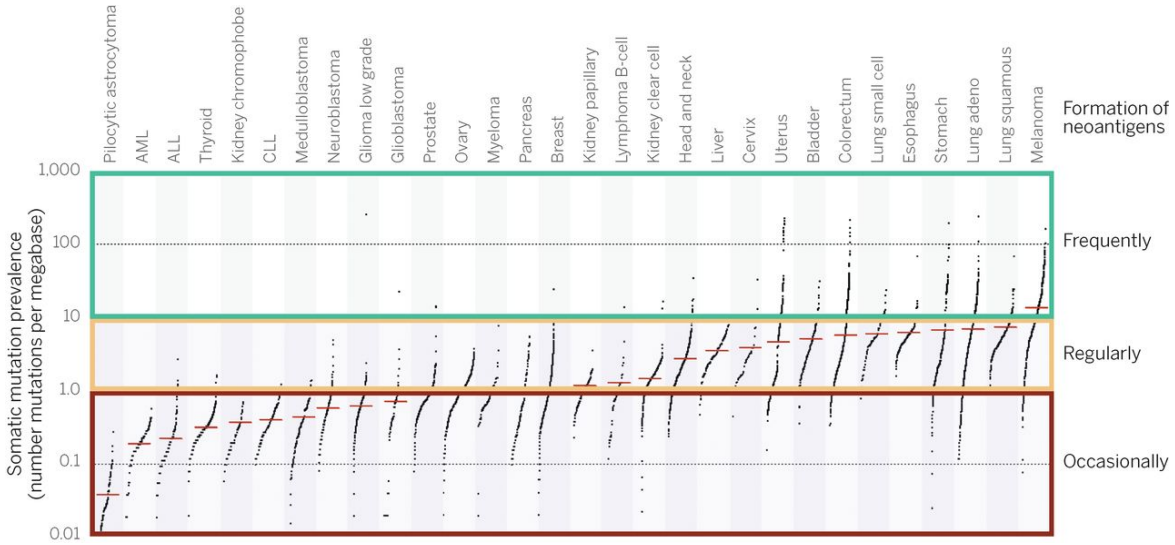


Fig. 1.3 Somatic mutation frequencies across a range of human cancer malignancies, labeled on the right by the likelihood of neoantigen formation. These likelihoods are based on observations from human melanoma studies and adapted from [24].

rates in human melanoma, researchers expect adequate T cell recognition of neo-antigens for cancers with 10 somatic mutations per megabase or higher (Fig. 1.3).

1.2 Innate and Adaptive Immunity in Cancer

1.2.1 The Human Immune System

Thus far, we have introduced the concept of tumour heterogeneity and its impact on the appearance, behaviour and development of the tumour, with a scope on tumour immunity. We will now introduce the general roles of human leukocyte subpopulations, the concept of a *tumour microenvironment* with respect to immune cells and mechanisms by which tumours evade immune destruction.

The immune system is comprised by a complex collective of cells and humoural factors that can be subdivided into "innate" and "adaptive" subpopulations as illustrated in Fig. 1.4. These subpopulations play complementary roles in the surveillance, detection and elimination of pathogens.

Innate Immunity The innate immune system is comprised of a large number of bone marrow derived cell families including monocytes, macrophages, dendritic cells and natural

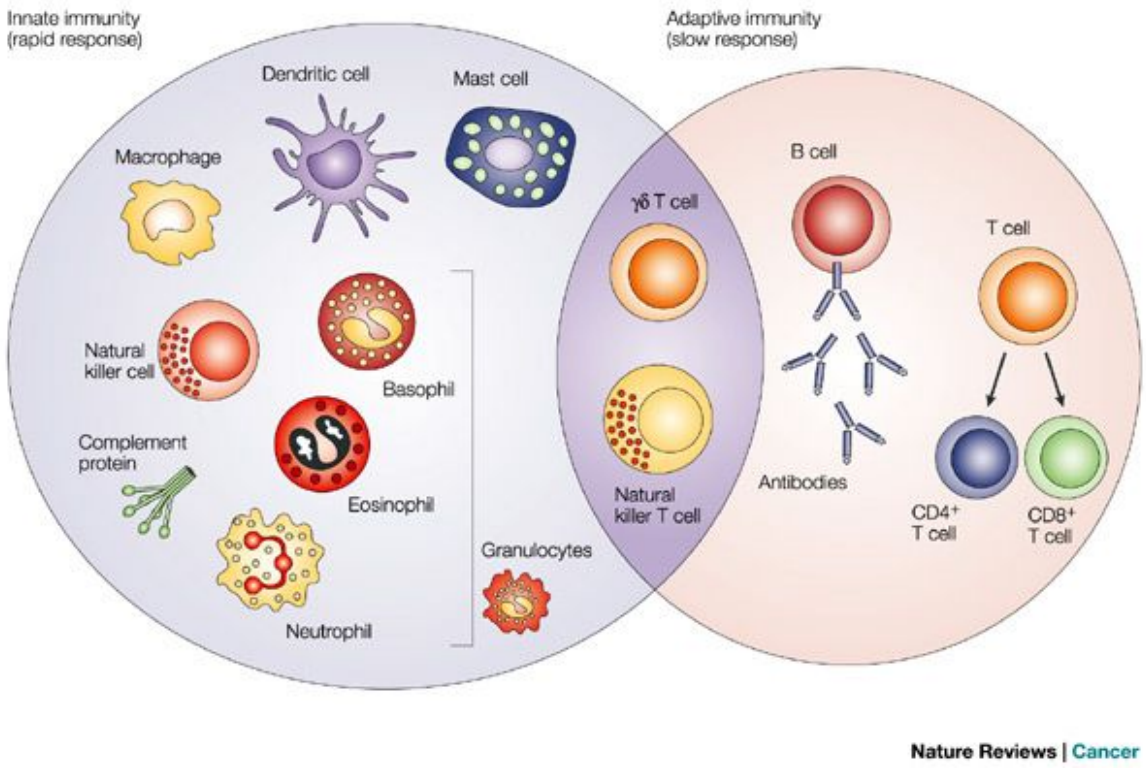


Fig. 1.4 The innate immune system is the first to respond during pathogenesis, and it consists of a range of immune agents including macrophages, dendritic cells and natural killer cells. The adaptive immune reaction is a delayed, but highly specialised response targeting specific antigens. It consists primarily of cell families descended from the T cell and B cell lineages. This figure has been reproduced from [29].

killer cells that form the basis of a pathogen-directed immune response. Cytokine, interleukin and chemokine molecules are examples of humoural factors involved in pathways that link together different components of the immune system.

Adaptive Immunity The central property characterising an adaptive immune system response is the production of a special repertoire of proteins, expressed as antigen-specific antibodies in B cells and T cell receptors (TCR). Non-self antigens, such as tumour neoantigens, have a strong affinity towards matching antibodies and TCRs, resulting in elevated co-stimulatory signalling and the increased proliferation of antigen-specific lymphocytes.

Mechanisms of Antigen Recognition Dendritic cells present antigens via the MHC class I pathway through three well characterised channels. Firstly, virally infected dendritic cells can endogenously process viral proteins into constituent peptide fragments and present them

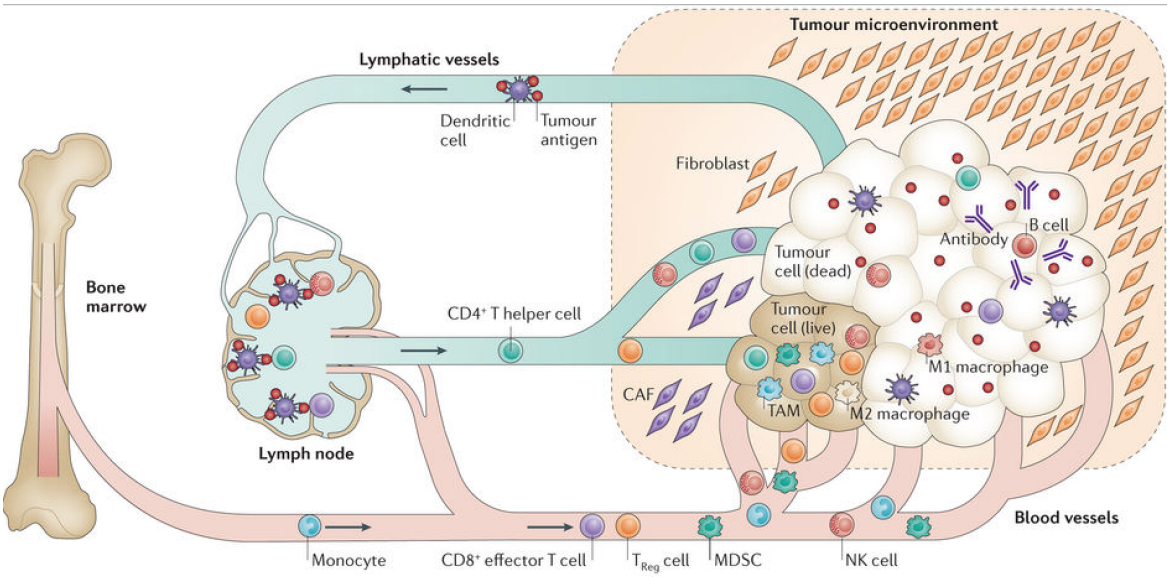


Fig. 1.5 A general schematic for anti-tumoural immunity, illustrating the interaction between the adaptive and innate immune system, and other leukocytes such as immunosuppressive immune cells and humoural factors. This figure has been reproduced from [30].

on their surfaces. Secondly, dendritic cells can cross-present antigens acquired through engulfing fragments of dead cells. Finally, it has been proposed that dendritic cells have the potential to acquire third-party MHC-peptide structures from dead cells through endosomal fusion or trogocytosis, and subsequently present them to CD8+ T cells in a process known as *cross-dressing* [31]. As illustrated in Fig. 1.5, dendritic cells migrate towards proximal lymph nodes via local lymphatic channels where they activate naive T cells and B cells to initiate the adaptive immune response.

Activation and Dynamics of the Adaptive Response T cells and B cells specialise with with billions of surface antigen receptor combinations in order to maximise the probability of non-self antigen recognition. Upon activation, T cells navigate towards the tumour where they facilitate cytotoxic mediated cancer cell destruction by engaging the host MHC/antigen complex via their T cell receptor. B cells secrete their antigen receptors as antibodies (Fig. 1.5). B-cell antibody production against a specific antigen promotes cancer cell apoptosis through two well characterised mechanisms [32]. First, antibodies can engage the complement system, triggering a cascade of reactions within local tissues that upregulate phagocytosis of cancer cells via phagocytes such as dendritic cells. Secondly, antibodies can bind directly against the antigen, blocking extrinsic proliferative growth signals from stimulating the cell and ultimately, cell death [32].

Immune Cell Regulation If an immune reaction to a pathogen is left unchecked, it can turn rampant and potentially kill the host. A notable example of this is in the case of a cytokine storm, whereby the feedback loop between immune cells and cytokines becomes dysregulated, resulting in an exaggerated immune reaction and damage to local tissues [33]. To prevent this, a system of immune regulatory mechanisms exists that mitigates exaggerated immune responses. Cancer hijacks these systems as part of its immune escape mechanism (see section 1.2.3) and uses several subpopulations of leukocytes to create an immunosuppressed environment [17].

T-regulatory cells (T-regs) are a classic example of regulatory cells that are often overrepresented in tumours relative to normal tissues [34], and are recruited via the overexpression of tumour-derived chemokines [35]. One mechanism by which T-regs suppress effector cell activity is the secretion of inhibitory cytokines such as IL-10 and transforming growth factor beta 1 (TGFβ) [36] that downregulate the killing function (cytolysis) of effector cells. Furthermore, T-regs have been shown to secrete vascular endothelial growth factor (VEGF) [37], a protein promoting both an immunosuppressive tumour microenvironment and angiogenesis. T-reg infiltration is a predictive of poor prognosis in a variety of cancer pathologies and subtypes [38].

Macrophages are examples of specialised myeloid progenitor cells that have the propensity to polarise into two subtypes of contrasting tumour-related functionality [39]. Whereas M1 macrophages enhance anti-tumoural immunity, M2 macrophages have been shown to induce an immunosuppressive environment. They achieve this through IL-10 secretion and facilitate tumour progression through VEGF, chemokine ligand 2 (CCL2) & TGFβ secretion [40, 41, 42]. M2 macrophages are a component of the wound healing response and they often work with non-immune cells to upregulate alternative mechanisms for immune evasion (explained further in section 1.2.2).

Myeloid derived suppressor cells (MDSCs) represent a family of immature myeloid cells that have been prevented from fully specialising into mature granulocytes, macrophages or dendritic cells [43]. Full differentiation is often impeded in the event of pathological conditions such as cancer or sepsis. MDSCs have been shown to suppress T-cell proliferation through several complex mechanisms involving the expression of arginase (*ARG*1) and nitric oxide synthase (iNOS) [43].

1.2.2 Immunity in the Tumour Microenvironment

Thus far, we have introduced the concept of tumour heterogeneity, different components of the immune system, and a brief overview over how they interact. A holistic understanding of cancer immunity depends on incorporating additional information relating to other biological agents in the surrounding tissues. Tumours comprise more than just cancer and immune cells; non-malignant agents such as fibroblasts, lymphatic & blood vessels, pericytes, adipocytes are potentially amenable to recruitment and corruption through tumour-derived signaling pathways. A complex intercellular signaling network consisting of cytokines, chemokines, growth factors and other humoural factors links these agents together to form the *tumour microenvironment* (TME) [44]. In section 1.2.1, examples were given of several cells that suppress the capacity of the immune system to destroy the tumour. Cancer cells recruit immunosuppressive leukocytes into the TME through direct or indirect signalling mechanisms between different TME compartments and the immune system. These mechanisms are becoming progressively better characterised; examples of which are outlined in the following paragraphs.

Immunosuppressive Leukocytes Cancer cells may recruit T-regs into the TME either via the direct secretion of CCL-22 [45], inducing M2 Macrophages to express CCL-22 [46], or promoting a hypoxic microenvironment that results in the expression of CCL-28, a T-reg recruiting cytokine [37]. Naive macrophage polarisation into the immunosuppressive M2 state can be factilitated through exposure to IL-4, IL-10 or IL-13 [47]. Cancer cells and TME infiltrating immune cells have been shown to over-express these factors [48] and promote the overrepresentation of M2 macrophages in the TME.

Immune Inhibitory Checkpoints Another mechanism by which cancer cells influence T effector cell functionality is through the overexpression of molecular checkpoint molecules that are normally involved in downregulating exaggerated T cell killing. For example, epithelial cells express a cell membrane-bound protein known as programmed death ligand 1 ($PDL1$), designed to limit T cell activity during peripheral tissue inflammation, and thus lowering the likelihood of an autoimmune reaction. By overexpressing this molecule, cancer cells significantly increase the probability of engaging the corresponding T cell programmed death receptor ($PD1$) [49]. Furthermore, persistent and prolonged exposure to antigens or inflammation can lead to the progressive loss of effector functionality in cytotoxic T cells, pushing them into an *exhausted* state [50]. Exhausted T cells are characterised by the overexpression of $PD1$ and other inhibitory receptors such as $TIM3$ and $LAG3$ [51].

Embryonic Developmental Pathways Cancer cells have been shown to hijack embryonic developmental pathways such as Hedgehog, Wnt and Notch in order to promote tumour progression [52]. Perturbation to the activity of these pathways has been linked to differential regulation of immune activity within the tumour microenvironment. For example, Hegehog signalling has been implicated in impaired T cell proliferation and activation in both basal cell carcinoma and pancreatic ductal adenocarinoma (PDAC) [1, 53]. Upregulated Notch pathway signaling has been linked to enhanced T cell responses in PDAC [1] and mouse models of cancer [54].

Stromal Cells Studies focusing on the immune cell-stroma crosstalk have found that each component plays an active role in shaping the the other. In one notable example, recruited M2 macrophages drive stromal cell activation through the expression of fibroblast growth factors FGF-7 and FGF-9. In turn, fibroblasts ablation *in vivo* is associated with decreased M2 macrophage infiltration, suggesting a potential recruitment mechanism [55]. Furthermore, cancer associated fibroblasts (CAFs) have been shown to promote Treg infiltration through the secretion of *CCL5* [56]. CAFs have been implicated in inhibiting the intratumoural infiltration of T cells through the secretion of *CXCL*12 which binds to *CXCR*4+ cells [57]

In summary, cancer cells promote a TME that facilitates tumour progression through overexpression of checkpoint elements, infiltrating T cell corruption, signaling pathway perturbation and stromal cell manipulation. The acquisition of these immuno-evasive measures over time are pivotal to tumourigenesis. A fundamental theory of cancer immunity known as the immunoediting hypothesis tries to fit these observations together in the context of tumour heterogeneity, as discussed in the following section.

1.2.3 The Immunoediting Hypothesis

The immunoediting hypothesis stems from the earlier "cancer immunosurveillance" proposed mechanism for the immunity-driven detection and elimination of cancer cells [59]. This hypothesis described a system by which components of the immune system act in sentinel-like fashion to detect and eliminate neoplastic lesions. 50 years later, cancer immunosurviellance now forms part of a much larger hypothesis consisting of three sequential phases: elimination, equilibrium and escape [60] (illustrated in Fig. 1.6). Under this new framework, the immune system not only plays a role in cancer cell destruction, but prunes genetically heterogeneous subpopulations of cancer cells and ultimately shapes the genotype and resulting phenotype of the tumour.

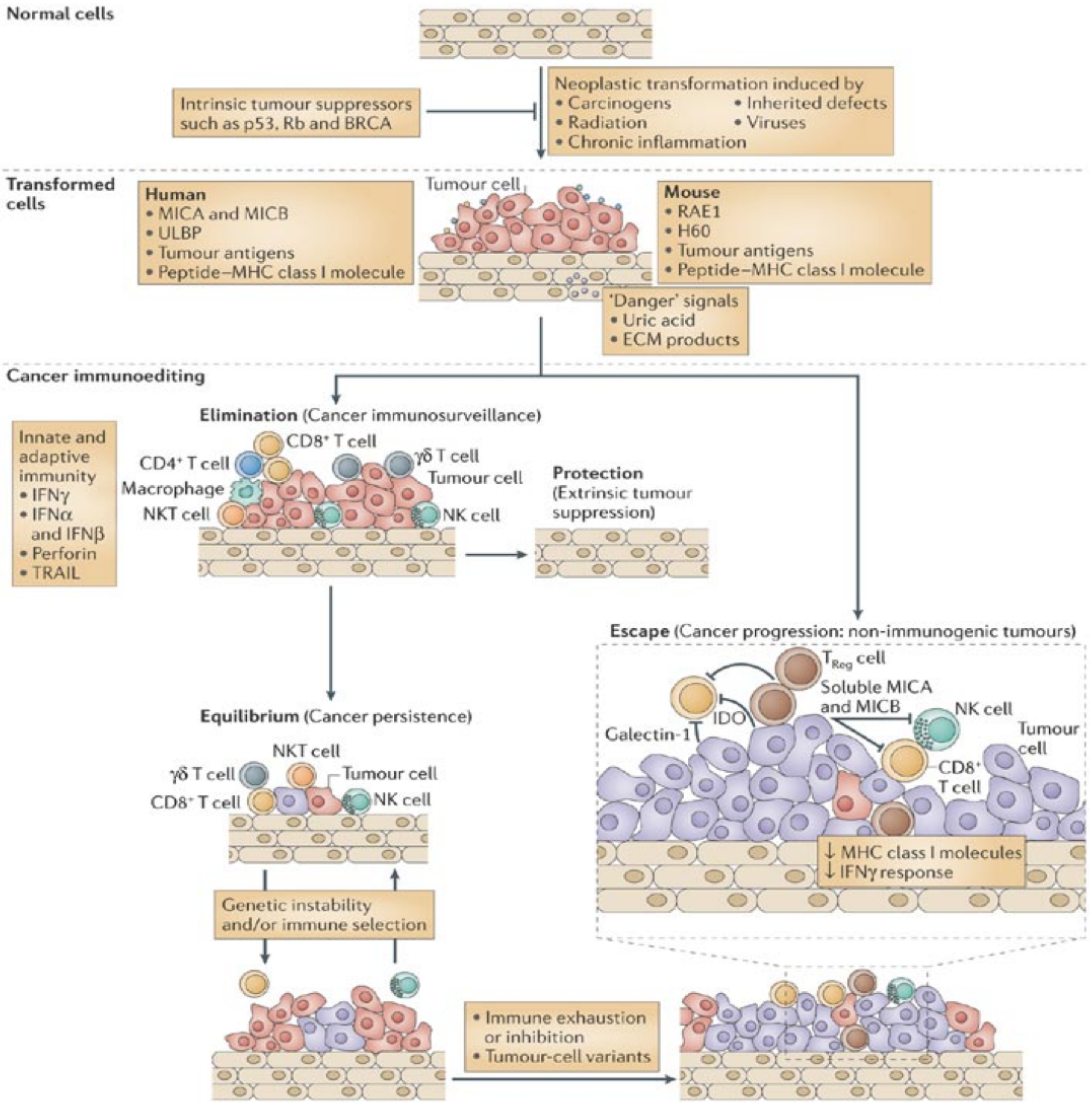


Fig. 1.6 A diagram of the immunoediting process illustrating the immunosurveillance, equilibrium and escape phases. This figure has been adapted from [58].

Elimination The elimination phase of immunoediting is synonymous with cancer immunosurveillance, whereby the innate and adaptive immune system congruently eliminate cancer cells prior to their progression into a clinically observable tumour. This phase has been well characterised, with many studies reporting the higher prevalence of spontaneous and carcinogen-induced tumours in immunodeficient knockdown mice over controls [61, 62]. It has been shown that NK cells play a crucial role in the elimination of these early lesions, responding to chemokines and other signalling molecules secreted by stressed cells [63].

These findings highlight the innate immune system as a key regulator of neoplastic anomalies in the body, although further studies have shown that CD8+ and CD4+ T cells can also be recruited via the cross-talk between NK cells and dendritic cells [64].

Equilibrium Cancer cells can progress into a state of equilibrium with the immune system, with progression and regression pathways upregulated concurrently, pushing the tumour into a state of functional dormancy. Studies have shown that tumours demonstrate prolonged persistence in this state, regularised by a balance between proliferation cytokines such as IL-23/IL-10 and elimination factors such as IL-12/IFN-γ [65]. Immunogenic tumour cells are eliminated at a faster rate than their more evasive counterparts giving rise to natural selection. During this phase, tumour cell variants can evolve with characteristics that enable them to circumvent immune regularisation, resulting in the failure of the immune system to control tumour growth [63]. This final phase of immunoediting is known as "escape" and is described below.

Escape Tumours are defined by their characteristic ability to demonstrate unimpeded growth. When the immune system fails to control the growth of emerging cancer cell variants, a tumour can be classified as clinically observable. The immune system is said to have "edited" the emerging disease in such a way that only variants with acquired mutations conferring immune evasion are observed at the clinical stage. These mutations compromise immunosurveillance by breaking the antigen processing and presentation pathway, exhausting immune efforts by overexpressing immune checkpoint markers such as PD-L1 and galectin-9 or recruiting a repertoire of immunoregulatory cells such as M2 macrophages or MDSCs. The skewed balance of cytokines such as such as IL-10, TGF-β and VEGF in the microenvironment favour tumour progression [63, 60, 66].

1.3 Linking Heterogeneity With Immunity

Intertumoural heterogeneity can give rise to variants that generate a variety of tumour microenvironments, eliciting differential immune responses whose composition and functionality confer varying clinical implications. The characterisation of the immune landscape between subtypes of the same cancer class is a key goal in cancer immunology and can help elucidate which groups of patients stand to benefit the most from treatment options such as immunotherapy. Immune system decomposition, spatial properties and relative abundance are crucial features in relating immune response to clinical parameters and dysregualted cancer signaling. Molecular profiling techniques enable the elucidation of cellular mecha-

nisms responsible for immunity at the transcription/translation level. At the sample level, imaging techniques provide a spatial context for the immune system, guiding the search for useful morphological features. A combination of these approaches can provide profound insight into the regulatory mechanisms of anti-tumoural immunity [67], arguably one of the most important tasks in cancer immunology. Additionally, experiment-derived immune features can be used to build scores that enable robust prediction of outcome and other clinical parameters [68, 69].

1.3.1 Layered Approaches to Immune Profiling

This section provides a brief experimental protocols overview pertaining to immune feature extraction and regulatory mechanism inference from tumour specimens, both at the cellular and sample levels. Each mentioned data modality has been used in the preparation of projects throughout this thesis. Computational approaches to immune profiling are detailed extensively in chapter 2.

Flow Cytometry Flow cytometry provides a cell counting protocol enabling the robust enumeration and sorting of leukocyte families. Heterogeneous cell populations are extracted from fresh tissue samples and labelled using a fluorescent tag, which are typically modified antibodies that bind to a distinguishing cell-specific protein marker. The cells are suspended in media and passed through a machine that measures fluorescence intensity. Subpopulations are isolated by passing single cells through one by one and cataloging them based on the fluorescent antibody tag. This enables features such as purity, volume and abundance of single cells to be extracted from the heterogenous admixture. This method does not offer any spatial context with respect to cells in the microenvironment and thus provides a limited overview of immune cell representation in the tumour. There is an upper bound on the number of antibodies handled at any given time, and hence, comprehensive immune profiling attempts are laborious and expensive procedures.

Gene Expression Profiling Although flow cytometry can sort and enumerate subpopulations of cells from a sample admixture, it can only measure signals from a handful of fluorescent tags. This places a strict upper bound on the number of subpopulations that can be extracted from a sample at any given time. Gene expression assays on the other hand can measure thousands of mRNA transcripts concurrently from bulk tumours or laser microdissected compartments of the tumour. Historically, this was achieved using microarray technology, but is rapidly being overtaken by novel mRNA sequencing pipelines that demonstrate superiority in detecting low abundance transcripts, identifying *de novo* isoforms

and detecting rare variants [70]. Recently, computational methods have been proposed that infer immune system properties such as composition and functional state from expression data; these are overviewed extensively in chapter 2. Molecular assays cannot provide any spatial context for measurements however, and the convoluted nature of expression profiling makes it extremely challenging to localise translational/transcriptional behaviour to single subpopulations of cells.

Haematoxylin & Eosin Stained Slides Staining tissue samples with Haematoxylin & Eosin (H&E) is a widely used protocol for analysing biopsies for clinical markers of suspected neoplastic lesions. Haematoxylin stains only basophilic cell components such as nucleic acids, making the nucleus appear blue under a microscope. Conversely, Eosin is acidic and reacts with the acidophilic components of the cell such as amino groups of proteins in the cell, staining the cytoplasm pink. Pathological analysis of cancer in H&E stained images has long revealed features of tumour architecture and cell morphology characterising disease state [67]. For example, lymphocytes and neutrophils demonstrate distinctive cell morphologies that a trained eye can distinguish from the surrounding tissues, enabling them to be quantified and their relationship with the surrounding microenvironment described. The added dimensionality of geometric features is invaluable to describing how the geometric context of lymphocytes vary between different cancers. H&E stained images contain a limited number of features and thus, lymphocyte subpopulations cannot typically be distinguished from the images alone, unlike immunohistochemistry techniques.

Immunohistochemistry Staining The principle methodology underlying immunohistochemistry (IHC) staining is similar to that of flow cytometry, with the exception that IHC tests the representation of a single specific cellular component per slide. IHC staining can help us visualise the distribution and localisation of specific leukocytes in the tissue architecture, enabling a wider variety of features to be extracted. However, in order to generate a comprehensive overview of immune composition, it is necessary to stain for a large panel of leukocyte markers which can be time consuming and expensive. Furthermore, the feasibility of this is determined by the sample volume since this places an upper bound on the number of slides that can be generated. Less pathological expertise is needed to make quantitative measures of specific cell family abundance however, as cellular components of interest typically stain brown and surrounding tissue stains blue. Quantitative profiling can be achieved using simple scripts that quantify colour intensity [71].

Proteome Composition Mass spectrometry (MS) quantifies the abundance of a particular peptide in a given sample. More abundant peptides give rise to higher peaks in the spectrum, with the intensity specified by a continuous distribution. Algorithms can use a peptide spectrum to quantify the relative abundance of a protein. Samples from protein mixtures can be analysed in independent MS batch runs to give quantify the abundance of a panel of proteins [72]. However, MS requires a complex preparation of samples and is insensitive to low-abundance proteins such as signalling molecules relative to other immunoassays. [73]

Reverse-phase protein arrays (RPPA) enable the high-throughput measurement of multiple samples under the same experimental conditions. A RPPA is basically a miniature array of antibodies that enable highly sensitive measurements of a single protein. It has proven to be extremely useful in quantifying protein abundance from small tumour samples such as a biopsy. This makes the RPPA an extremely useful clinical tool, especially given the minor amount of prior preparation needed for analysis. However, this method is limited in that it requires a specific antibody for every protein, and the limited array size places an upper bound on the number of tests that can be performed at a given time. Therefore, although more sensitive measurements can be made of less abundant proteins, less overall proteins can be measured relative to MS [73].

Somatic Variant Calling *Somatic* alterations in DNA are mutations introduced after conception, and thus are present in all cells aside from germ cells. Somatic variants in cancer range from point mutations termed single-nucleotide variants (SNVs), to insertion/deletion (indels) of nucleotide sequences to chromosomal copy number alterations (CNAs) and rearrangements known as structural variants (SVs). High-throughput next generation sequencing (NGS) techniques such as whole genome sequencing (WGS) have been used to reveal cancer-specific somatic mutations by measuring and cross-referencing the tumour genome with a matched normal genome from adjacent tissue or blood. Whole exome sequencing (WES) is a cheaper alternative to WGS but cannot identify non-coding mutations or SVs. Together, these technologies have enabled the stratification of patients on the basis of tumour genome heterogeneity [74, 75], and have even been integrated with gene expression analysis for more powerful classification schemes [76].

1.3.2 Key Challenges in Cancer Immunology

Understanding potential associations between the immune microenvironment and the outcome of treatment strategies such as surgery, chemotherapy, radiotherapy, hormone therapy or

adjuvant therapy has become a prominent goal in cancer immunology. Tumour heterogeneity often means that many observations are context-dependent and thus, regulators of an immune phenotype will not necessarily be consistent across tumour types and subtypes. Categorising the immune microenvironment between subpopulations is essential to determining whether immunotherapies or potential biomarkers will lend any benefit to said subgroup. Furthermore, understanding the complex interactions between the tumour microenvironment and the immune system is essential to developing novel immunotherapy strategies.

Association studies between different data modalities are used comprehensively to probe interactions between the tumour microenvironment and the immune system. For example, Rooney *et al* integrate expression, sequencing and copy number data into a seminal analysis of cytolytic activity and its dysregulation in the TME [69]. Snyder *et al* use whole genome sequencing to link tumour genetic landscapes to anti-CTLA therapy [77]. Mlecnik *et al* integrate genomic, transcriptomic and imaging data to study the stratification of immune responses by microsatellite instability in colorectal cancer [78]. The success of these approaches has fostered the field of "multiomics", referring to the rapidly developing field of biological inquiry that integrates several "omes" (proteomes, genomes, transcriptomes, imageomes) into a single analysis [79, 80]. Developing multiomics approaches to learn representations between the TME and infiltrating immune system is the first problem this thesis addresses.

Association studies can determine regions of the genome that are most likely to explain the variance in a phenotype, but cannot provide any insight regarding the direction of the association. Since a fundamental goal of cancer immunology is to find mechanisms through which immune traits take on the values they have, this constraint limits our confidence in assigning responsibility to a genomic abnormality. As such, a mechanism describing the emergence of a phenotype given a mutation is not complete if causality cannot be proved. Probabilistic models integrating multiomics datasets have demonstrated considerable success in predicting molecular pathways regulating obesity and diabetes phenotypes [81, 82] . Reconstructing molecular pathways linking genomic abnormalities to downstream cancer immune features is the second problem this thesis aims to address.

1.4 Study Aims and Dissertation Summary

Although great advances have been made in the field of cancer immunology, much work remains to understand the immune microenvironment to the extent where personalised

immunotherapies can be deployed for all patients. To contribute to this effort, my thesis has been designed around a central line of inquiry: How do we qualitatively describe the role of the immune system with respect to tumour heterogeneity? From this stems two primary questions that this thesis aims to answer:

1. What associations exist between tumour cells and their surrounding microenvironment? Can we use these associations to find patients potentially amenable to existing immunotherapy strategies?

2. Having elucidated associations, can we find a statistical method to assign directionality? In other words, does A cause B or can we find another variable that better explains this association?

To address 1), this thesis draws upon computational approaches to derive immune features and draw associations between different compartments of the TME. We address 2) by moving away from association study based schemes to explore causal relationships and reconstruct regulatory mechanisms between the TME and the immune system. The results and achievements of my thesis have been summarised in the paragraphs below.

Computational Approaches to Cancer Immunology Chapter 2 gives an overview of recent computational methods for the quantitative characterisation of immune composition and function from multiomics data. Computational immunology is a fast growing area of research aiming to provide methods for analysing the growing prevalence of multimodal datasets. Gene expression data from bulk tumours contains contributions from a variety of distinct cell populations in the TME, enabling us to ask: how much of leukocyte *A* do I have in my sample and what is its relation with the TME? Sequencing data enable tumour neoantigen characterisation and MHC typing using predictive algorithms. Imaging data permits the design of high-throughput computational pathology pipelines that generate high resolution spatial features of lymphocytes with respect to surrounding tissue structures. The use of these methods in integrative multiomics pipelines form the basis of the work outlined in the following chapters.

The Role of Oncogenic KRAS in Pancreatic Cancer Immunity In chapter 3, I present my work on elucidating key associations between tumour cells and the immune system in the context of pancreatic adenocarcinoma (PDAC). Work done by colleagues led to the identification of dysregulated TF clusters modulating three embryonic developmental pathways contributing to PDAC. Termed Hedgehog/Wnt, Notch and Cell Cycle, these pathways demonstrate varying degrees of dysregulation amongst PDAC samples, enabling

the identification of three main patient subgroups. Building on my colleague's PDAC characterisation, I used gene expression tools to qualitatively describe TME features including leukocyte subset mixing proportions, stromal contamination, immune inhibitory checkpoint overexpression and immune pathway dysregulation. Partial correlation analysis revealed a stromal basis for Hedgehog/Wnt and Notch signaling, suggesting a probable mechanism for PDAC progression. Hedgehog/Wnt samples demonstrate immunosuppressed TMEs, with a characteristic M2 macrophage signature and no evidence of significant T cell recruitment. On the other hand, Notch samples demonstrate a characteristic T cell signature, with upregulated T cell related pathways and overexpressed inhibitory checkpoint molecules suggesting a possible immune evasion mechanism through T cell exhaustion. Although the Cell Cycle subgroup demonstrated the highest mutational burden, it was also characteristic of minimal immune infiltration relative to the remaining subgroups.

Probabilistic Models for Regulatory Network Reconstruction Moving beyond associations, chapter 4 serves as a methods introduction for identifying the dependency structure within and between different data modalities. Termed *graphical modelling*, this field is a rapidly growing area of machine learning that can thankfully be condensed into a few simple principles. A conditional independence statement tries to answer the question: Can the association between two given variables be explained by the variance in other subsets of variables? These statements are encoded into the *nodes* of a graphical model, forming the basis of well-established methodologies such as Gaussian graphical models and Bayesian networks. By identifying caveats in existing approaches, we propose a novel framework for dependency structure visualisation using hypothesis-driven priors. The basic idea is to anchor our analysis on a fundamental truth in cancer biology, thus reducing model space complexity and potentially reconstructing ground truth regulatory hierarchies. This approach sets up the foundation for chapter 5, where I address the second question of my thesis.

Causal modeling dissects tumour- microenvironment interactions in breast cancer Elucidating interactions between cancer cells and their microenvironment is a key goal of cancer research with implications for understanding cancer evolution and improving immunotherapy. Previous studies used association-based approaches to infer relationships in transcriptomic data, but could not infer the direction of interaction. Here I present a causal modeling approach that infers directed interactions between signaling pathway activity and immune activity by anchoring the analysis on somatic genomic changes. My approach integrates copy number profiles, transcriptomic data, image data and a protein-protein interaction network to infer directed relationships. I demonstrate that my approach generates models with a

high validation rate in independent cohorts and orthogonal data types. In particular, discover several novel genomic drivers of lymphocytic infiltration are discovered. This framework is very general and can be extended to other cancer types, data types and clinical parameters.

In summary, this dissertation provides methodological contributions, at the levels of associative and causal inference, for inferring the contextual basis for tumour-specific immune agency.

Chapter 2

Computational Approaches to Cancer Immunology

In this section I give an overview of recent computational methods for immune profiling in the tumour microenvironment. This follows from section 1.3.1, where we introduced experimental methods for immune feature acquisition at the cellular and sample levels of a tumour. In section 2.1, I introduce statistical approaches for inferring the representation of distinct cell populations from expression data. For pathway analysis, gene set enrichment analysis is a well-established class of tools for inferring both immune cell enrichment and immune pathway dysregulation (section 2.2). Beyond gene expression alone, sequencing data permits the typing of MHC molecules involved in the peptide presentation process, and also enables the cancer epitope landscape to be characterised (section 2.3). Finally, I overview standard approaches to computational pathology for high-throughput sample-level immune feature extraction in section 2.4.

2.1 Gene Expression Deconvolution

The immune system comprises a range of phenotypically distinct cell families that form an interaction network between themselves and their environment. Each cell family plays a unique role in the overall immune response, by regulating pathways involved in processes ranging from pathogen recognition to wound-healing. Common experimental approaches to studying immune system composition involve immunohistochemistry (IHC) staining or flow cytometry; both of which use antibodies specific to a particular cell family to quantify their abundance in a sample. There are clear advantages to using these approaches, such

as obtaining an absolute quantification of cell family abundance and enhanced specificity. However, these approaches do not scale up well to large sample sizes due to the laborious process of technical preparation (FACS) [83] and pathologist verification and scoring (IHC) [84]. Furthermore the large panels of antibodies and iterative protocols required make immune profiling more than a handful cell families an infeasible process [85]. As such, these protocols are usually used in a validation setting rather than large-scale discovery applications.

IHC and flow cytometry make measurements on a cellular level whereas molecular assays of gene expression typically looks at bulk populations of cells, with the exception of single-cell profiling. As such, gene expression profiling is said to be done on a sample level rather than a cellular level. For each gene assayed, the expression signal is a combination of contributions from individual cells in the TME, such as those belonging to the cancer-cell autonomous, normal, stromal or immune compartments. A number of algorithmic techniques have been developed to quantify immune cell profiles from molecular assays of gene expression (which are much more abundant datasets than IHC or flow cytometry).

The basic idea is to estimate the mixing proportions $\mathbf{k} = \{k_1, k_2, ..., k_n\}$ of n cell subpopulations where $k \geqslant 0 \; \forall \; k \in \mathbf{k}$, by assuming a linear contribution model. The simplest model can be written by assuming our data admits zero noise and perfect linearity. Let $\mathbf{G} = \{g_1, g_2, ..., g_m\}$ be a set of m unique gene identifiers such that for any $g \in \mathbf{G}$, its expression in a given cell subpopulation i is e_{gi}. The total expression in our sample E_g can therefore be written as the sum of mRNA contributions from each subpopulation:

$$E_g = \sum_{i=1}^{n} e_{gi}. \tag{2.1}$$

E_g can also be written as the dot product between the mixing proportions $\mathbf{k}$ and the average expression of gene g in each cell family: $\mathbf{k} \cdot \{e_{g1}, e_{g2}, ..., e_{gn}\}$. Since this holds for all genes, we can extend this relationship across the entire transcriptome $\mathbf{E}$ by writing the dot product as a matrix multiplication equation:

$$\mathbf{E} = \begin{pmatrix} E_{g_1} \\ E_{g_2} \\ \vdots \\ E_{g_m} \end{pmatrix} = \begin{pmatrix} e_{g_1,1} & e_{g_1,2} & \cdots & e_{g_1,n} \\ e_{g_2,1} & e_{g_2,2} & \cdots & e_{g_2,n} \\ \vdots & \vdots & \ddots & \vdots \\ e_{g_m,1} & e_{g_m,2} & \cdots & e_{g_m,n} \end{pmatrix} \begin{pmatrix} k_1 \\ k_2 \\ \vdots \\ k_n \end{pmatrix} \tag{2.2}$$

where $e_{g_i,j}$ represents the average expression of gene g_i in cell family j, and the mixing proportions $\{k_1, k_2, ..., k_n\}$ are non-zero and sum to unity. From this starting point, we want to address the following questions:

- How do we estimate, or at least place an upper bound on the number of unique cell families in the sample?
- The model outlined by Eqs 2.1 and 2.2 assumes zero noise contamination; a far cry from reality when working with biological data. How do we estimate mixing parameters when noise is introduced to the system? Does our assumption of perfect linearity still hold?
- Can we place an upper bound on m, the number of genes, to minimise volatility in our estimates of mixing proportions?

The following subsections provide an overview of recent computational approaches that aim to address these points.

Gene expression deconvolution represents a class of methods that can be used to estimate unknown parameters in Eq 2.2. There are two types of deconvolution methods typically used for expression data. The first is known as *partial* deconvolution, and attempts to estimate only the unknown mixing proportions vector $\mathbf{k} = \{k_1, k_2, ..., k_n\}$. The second method, *complete* deconvolution estimates both $\mathbf{k}_g$ and the signature matrix. In this section we will evaluate the context in which each approach should be used, and give an overview of recent algorithms.

2.1.1 Complete Deconvolution

Given an expression matrix $\mathbf{E}$, the general aim of complete deconvolution approaches is to measure both $\mathbf{k} = \{k_1, k_2, ..., k_n\}$ and the signature matrix $S = e_{g_i,j}$ simultaneously. Venet *et al* were the first to propose an approach for complete gene expression deconvolution. Their method sequentially searches for $\mathbf{k}$ and S that minimises the norm of the reconstruction error $||\mathbf{E} - S \times \mathbf{k}||^2$ [86]. Applying their approach to colon cancer, they identified four distinct cell populations including those of a hematopoietic and fibroblast lineage. However, the authors state that setting an upper bound on the number of cell populations n is a significant challenge using this approach.

Building on this, Repsilber *et al* , use a least squares non-negative matrix factorisation approach to iteratively compute $\mathbf{k}$ and S such that $||\mathbf{E} - S \times \mathbf{k}||^2 \leqslant a$ where a is a user defined threshold [87]. Unlike Venet *et al* , the authors verify their results experimentally, using

paired flow cytometry and gene expression data from blood samples to compare predicted leukocyte mixing proportions against the ground truth. Other complete deconvolution methods have been proposed that require signature matrix related hyperparameters. For example, Erkkilä *et al* propose a Gaussian mixture model approach with Dirichlet priors over m mixing components to infer $\mathbf{k}$ and S [88]. Kuhn *et al* 's approach requires a list of cell-specific signature genes to help build the signature matrix [89].

2.1.2 Partial Deconvolution

The aim of a partial deconvolution approach is to estimate $\mathbf{k} = \{k_1, k_2, ..., k_n\}$ from a measured transcriptome $\mathbf{E}$ using a pre-defined signature matrix S. This approach reduces the number of unknown parameters in Eq 2.2, making it easier to rearrange and solve. First implemented by Lu *et al* [90], they used feature selection to construct a compact signature matrix of yeast cells at various phases of the cell cycle process and solved for $\mathbf{k}$ using simulated annealing. Abbas *et al* were the first to estimate cell mixtures from blood data using a linear least-squares regression approach in conjunction with a signature matrix [91]. Gong *et al* implemented a quadratic programming approach using expression signatures derived from purified subsets of blood cells [92]. All mentioned approaches have recently been combined under the umbrella of a single R package *CellMix* [93], taking user-provided cell-specific expression signatures as input. Other methods such as perturbation models and robust linear regression approaches have been put forward to deconvolve mixing proportions in blood sample transcriptomes[94, 95].

Thus far, these approaches have mostly focused on deconvolving blood data, which is a far more homogeneous and less noisy tissue than solid tumour tissue. The first application of partial deconvolution in tumour data came from Ghosh, who used mixture modelling methods to estimate cell population mixtures in colorectal cancer [96]. A recent approach by Newman *et al* uses a ν-support vector regression approach to estimate the mixing proportions of 22 leukocyte subsets from bulk tumour transcriptomic data[85]. The authors use feature selection to build a signature matrix containing the transcriptomic profiles of 22 purified leukocyte populations such as T cells, B cells and macrophages. Their algorithm, CIBERSORT, has been shown to outperform quadratic programming, linear least-squares regression and other approaches when reconstructing ground truth immune composition examples in both simulated and real datasets. Since this method motivated a major component of the analysis performed in chapter 3, it is worthwhile describing the underlying mathematics.

To define ν-support vector regression (ν-SVR), we must first introduce the more general concept of Support Vector Machines (SVMs) of which ν-SVR is an instance.

Support Vector Machines Let $\mathbf{X}^n$ be a set of m, n-dimensional measurements of the form $\{x_1, x_2, ..., x_n\}$ belonging to one of two classes $\{y_1, y_2\}$. Let $\mathbf{Y}$ be a vector of length m representing class membership for points in $\mathbf{X}^n$. Finally, we define $\mathbf{X}^n_{y1} \subset \mathbf{X}^n$ to denote points belonging to y_1 and define $\mathbf{X}^n_{y2} \subset \mathbf{X}^n$ to denote points belonging to y_2.

Definition $\mathbf{X}^n_{y1}$ and $\mathbf{X}^n_{y2}$ are *linearly separable* if there exists a set of $n+1$ real numbers $\{w_1, w_2, ..., w_n, k\} \in \mathbf{R}^{n+1}$ such that $\sum_{i=1}^n w_i x_i > k \; \forall x \in \mathbf{X}^n_{y1}$ and $\sum_{i=1}^n w_i x_i < k \; \forall x \in \mathbf{X}^n_{y2}$.

Our assumption that samples in $\mathbf{X}^n$ are linearly separable by category enables us to write a simple overview of SVMs. The basic idea of a SVM is to learn a function $f(\mathbf{X}^n)$ that can separate the two sets of points $\mathbf{X}^n_{y1}$ and $\mathbf{X}^n_{y2}$ in n-dimensional space. In particular we want a function that *maximises* the separation between the two classes, by finding a $(n-1)$-dimensional separating surface with the largest distance between it and the nearest data point on either side. In this case, $f(\mathbf{X}^n)$ is referred to as an *optimal hyperplane* (a simple example is illustrated in Fig. 2.1). Formally,

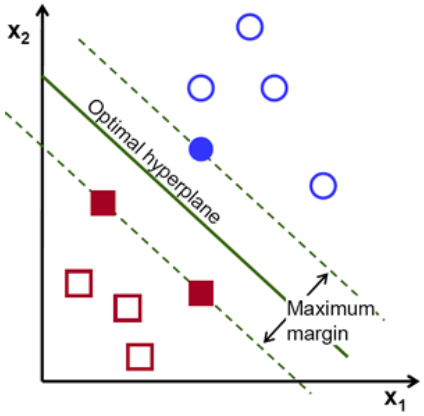


Fig. 2.1 A simple example of optimal hyperplane construction, maximally separating the distance between two linearly separable groups of observations in 2D feature space. This figure has been reproduced from opencv.org.

$$f(\mathbf{X}^n) = \beta_0 + \beta^T x \tag{2.3}$$

where β_0 is a bias term and β is a weight vector. For any point $\mathbf{x} \in \mathbf{X}^n$, the distance D between $\mathbf{x}$ and the hyperplane parameterised by (β_0,β) is given by

$$D = \frac{|\beta_0 + \beta^T \mathbf{x}|}{||\beta||}. \tag{2.4}$$

The closest points to the hyperplane are known as *support vectors* (SV)s: $\mathbf{X}^n_{SV} \subseteq \mathbf{X}^n$. Using these points, we can define the *canonical hyperplane* by choosing β_0 and β such that

$$|\beta_0 + \beta^T \mathbf{x}| = 1 \qquad \forall \mathbf{x} \in \mathbf{X}^n_{SV}. \tag{2.5}$$

Considering only the support vectors, the distance can be written

$$D_{SV} = \frac{|\beta_0 + \beta^T \mathbf{x}|}{||\beta||} = \frac{1}{||\beta||} \qquad \forall \mathbf{x} \in \mathbf{X}^n_{SV}. \tag{2.6}$$

In SVM theory, the *margin* M_{SV} is defined as twice the distance:

$$M_{SV} = \frac{2}{||\beta||} \tag{2.7}$$

which we maximise over with respect to a constraint in order to compute the optimal hyperplane. Since maximising Eq 2.7 is equivalent to minimising $f(\beta) = \frac{||\beta||^2}{2}$, the problem now becomes:

$$\underset{\beta_0,\beta}{\arg\min} f(\beta) \quad \text{subject to} \quad y_i|\beta_0 + \beta^T \mathbf{x}_i| \geqslant 1 \;\; \forall y_i \in \mathbf{Y}, \mathbf{x}_i \in \mathbf{X}^n, \tag{2.8}$$

which can be solved using Lagrangian optimisation.

Support Vector Regression Moving on to SVR, we still minimise $f(\beta)$, but subject to a new restraint dependent on a constant distance parameter ε. Here, SVs are values that lie outside the boundary of the ε-distance region:

$$\underset{\beta_0,\beta}{\arg\min} f(\beta) \quad \text{subject to} \quad |y_i - (\beta_0 + \beta^T \mathbf{x}_i)| \leqslant \varepsilon \;\; \forall y_i \in \mathbf{Y}, \mathbf{x}_i \in \mathbf{X}^n. \tag{2.9}$$

This case of support vector regression is called ε-SVR and is characterised by the ε-insensitive loss function:

$$L_i = \max\{0, |y_i - (\beta_0 + \beta^T \mathbf{x}_i)| - \varepsilon\} \qquad \forall i \in \{1, 2, ..., m\}. \tag{2.10}$$

The basic idea is to fit a "tube" of diameter ε to the data. The strict nature of these constraints may result in no solution for $f(\beta)$ such that Eq 2.9 is satisfied. One way around this is to introduce *slack variables* ξ_m and ξ^*_m to act as upper bounds on regression errors. This guarantees a solution for $f(\beta)$ subject to slightly modified constraints. Eq 2.9 now becomes:

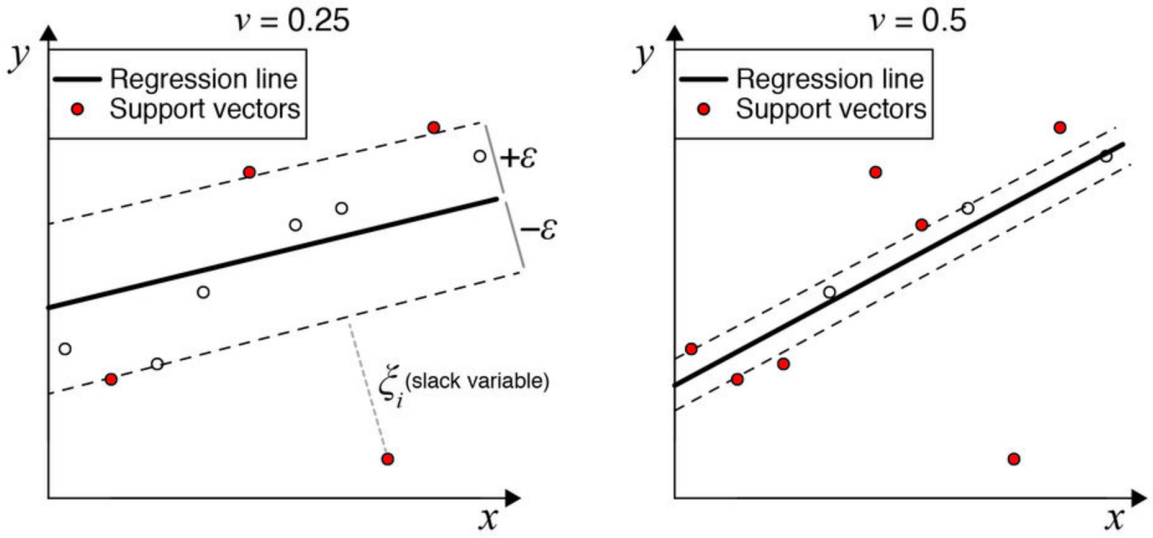


Fig. 2.2 An illustration of ν-support vector regression in 2 dimensions. The ν parameter places a lower bound on the number of support vectors, and simultaneously, an upper bound on the classification error. As such, we see that higher values of ν give rise to narrower ε-tubes. This figure has been reproduced from [85].

$$\underset{\beta_0,\beta}{\arg\min}\ \frac{||\beta||^2}{2} + C\sum_{i=1}^{m}(\xi_i^* + \xi_i) \quad \text{subject to} \tag{2.11}$$

$$y_i - (\beta_0 + \beta^T \mathbf{x}_i) \leqslant \varepsilon + \xi_i, \tag{2.12}$$

$$(\beta_0 + \beta^T \mathbf{x}_i) - y_i \leqslant \varepsilon + \xi_i^*, \tag{2.13}$$

$$\xi_i^*, \xi_i \geqslant 0 \quad \forall i \in \{1,2,..,m\}. \tag{2.14}$$

C is a positive unbound numeric constant that controls the tradeoff between the model complexity characterised by $\frac{||\beta||^2}{2}$ and the ε-insensitive training error $\sum_{i=1}^{m}(\xi_i^* + \xi_i)$. However, this formulation assumes that we know *a priori* the desired accuracy of our model given by $\varepsilon \geqslant 0$. ν-SVR is a special instance of ε-SVR that automatically minimises ε, giving us a new objective function and optimisation problem:

$$\underset{\beta_0,\beta}{\arg\min}\ \frac{||\beta||^2}{2} + C\cdot(\varepsilon\nu + \sum_{i=1}^{m}(\xi_i^* + \xi_i)) \tag{2.15}$$

subject to constraints given by Eqs 2.12, 2.13, 2.14. Consequently, different values of ν alter the diameter of the ε-tube, placing a direct lower bound on the number of SVs and also an upper bound on the loss function.

In the CIBERSORT SVR implementation, SVs are genes from the signature matrix for the leukocyte subsets [85]. Leukocyte mixing proportions can be viewed as the *orientation* of the ε-tube in feature space. Following their estimation, the mixing proportions vector $\mathbf{k}$ are the normalised regression coefficients such that $\beta_i > 0 \; \forall \; i \in \{1,2,..,m\}$ and $\sum_{i=1}^{m} \beta_i = 1$. Two immediate setbacks can be identified in this methodology. Firstly, The ν-SVR formulation means that CIBERSORT only returns mixing proportions and not absolute measures of leukocyte subset abundance. Secondly, CIBERSORT fails to correct for co-linearities in their signature matrix arising from the similarity of leukocyte subsets (for example, activated and resting CD4+ T cells), which may consequently bias the estimation of leukocyte subset mixing proportions. Addressing the first limitation, Becht *et al* introduce the Microenvironment Cell Populations (MCP) counter deconvolution method which computes an abundance score for 8 distinct leukocyte subsets and 2 stromal populations [97]. To address the second issue, Li *et al* propose a deconvolution method that focuses on 6 leukocyte subsets, with subsequent constraints placed on the signature matrix to minimise inter-subset colinearity [98].

Subpopulations of cells often express the same genes, making it difficult to construct signatures specific to a single cell type. Groups of genes with similar expression patterns amongst leukocyte transcriptomic profiles set up a correlation structure. This is particularly problematic in partial deconvolution approaches, since large covariances across a signature matrix can manifest in colinear mixing proportions. One way of dealing with this is by using cell-specific signatures.

2.1.3 Cell-specific Signatures

Ideally, we would like to annotate each cell population by a set of characteristic genes expressed uniquely by that cell population and nowhere else. In addition to minimising artifacts arising from collinearity, this helps with method regularisation, and eliminates redundant features in a signature matrix S such as genes with minimal or constant expression variance across the populations. To achieve this, we formally introduce the notion of *cell-specific markers*.

Definition Given a gene-set **G** and cell population $i \in \{1,2,...,m\}$, a gene $g \in \mathbf{G}$ is said to be a *cell-specific marker* for i if the following conditions hold: $e_{g,i} > 0$ and $e_{g,j} = 0$ $\forall j \in m \backslash \{i\}$.

Identifying cell-specific signatures for each population enables us to rewrite Eq 2.2 with a subsampled signature matrix containing only cell-specific markers for each population

$$\mathbf{E} = \begin{pmatrix} E_{g_1} \\ E_{g_2} \\ \vdots \\ E_{g_m} \end{pmatrix} = \begin{pmatrix} e_{g_{1,1}} & e_{g_{1,2}} & \cdots & 0 \\ 0 & e_{g_{2,2}} & \cdots & e_{g_{2,n}} \\ \vdots & \vdots & \ddots & \vdots \\ e_{g_{m,1}} & 0 & \cdots & e_{g_{m,n}} \end{pmatrix} \begin{pmatrix} k_1 \\ k_2 \\ \vdots \\ k_n \end{pmatrix}, \tag{2.16}$$

which enables the delineation of unique cell population-specific contributions to **E**, and therefore a more accurate measure of **k**. This is evident when we restrict genes in the signature matrix to cell-specific genes of a single cell population i. If $\alpha_i \subset \mathbf{G}$ is the set of cell-specific markers for cell population i, the following equality holds:

$$\sum_{p=1}^{m} \sum_{j \in \alpha_i} g_{j,p} \times \mathbf{k}_p = \mathbf{k}_i \times \sum_{j \in \alpha_i} g_{j,i}. \tag{2.17}$$

The system of equations generated by applying Eq 2.17 to all cell populations is overdetermined and almost always has no solution. Alternative approaches to partial deconvolution, such as taking the geometric or arithmetic mean of marker expression, can be used to infer cell population abundance. Rooney *et al* demonstrate that the geometric mean of *GZMA* and *PRF*1 can be used as a score for cytolytic activity, and use their approach to predict survival across a range of cancer types [69]. However, interpreting the mean expression of larger gene signatures can be challenging if we do not account for the expression of genes outside of the signature. Addressing this concern, Gene Set Enrichment Analysis (GSEA) was proposed, which ranks gene sets relative to the remaining gene *universe* [99].

2.2 Gene Set Enrichment Analysis

GSEA encompasses a class of techniques used to measure the enrichment of a a gene signature in a set of samples. GSEA can be used to measure the enrichment of leukocyte

signatures and pathways. The original formulation proposed by Subramanian *et al* tests the over-representation of a signature between sets of samples under different experimental conditions [99]. Barbie *et al* proposed a method termed Single Sample Gene Set Enrichment Analysis (ssGSEA), which tests gene set overrepresentation on a sample-by-sample basis [100]. ssGSEA forms an integral part of the work done in chapters 3 and 5, thus motivating a detailed overview of the method in this section.

2.2.1 ssGSEA and Population Signatures

Single sample gene set enrichment analysis (ssGSEA) can be used to infer relative representation of cell signatures between samples. Unlike mean approaches mentioned in section 2.1.3, the ssGSEA approach measures over-representation of m signature genes $\mathbf{s}=\{s_1, s_2, s_3, ..., s_m\}$ relative to the remaining sample gene set $\mathbf{G} \setminus \mathbf{s}$. The basic idea is to order genes in $\mathbf{s}$ by their absolute expression in a given sample's transcriptome $\mathbf{E}$. The expression of each signature gene $\{e_{s_1}, e_{s_2}, ..., e_{s_m}\} \subset \mathbf{E}$ is subsequently replaced by a rank. Ordering the list from highest rank m to lowest 1 enables us to construct a weighted empirical cumulative distribution function (ECDF) given by:

$$ECDF_{signature}(i) = \sum_{j=0}^{i} \frac{|e_{s_j}|^{\alpha}}{\sum_{j=0}^{m} |e_{s_j}|^{\alpha}} \tag{2.18}$$

where α is a weighting parameter. The ECDF for the remaining genes $\mathbf{G} \setminus \mathbf{s}$ is given by

$$ECDF_{remaining}(i) = \sum_{j \leqslant i} \frac{1}{m - |\mathbf{G} \setminus \mathbf{s}|}. \tag{2.19}$$

An *enrichment score* is then calculated by computing the integral of the differences between Eqns. 2.20 and 2.19. Barbie *et al* demonstrate the robustness of their method over Kolmogorov-Smirnov statistic-based approaches in lung adenocarcinoma gene expression data.

ssGSEA has been used extensively to delineate cell population representation in complex heterogeneous tissue samples such as cancer. Yoshihara *et al* use ssGSEA in conjunction with stromal and immune cell-specific signatures derived from leukocyte methylation data and laser-capture micro-dissection. Their method, ESTIMATE [101] independently recapitulated the well known immunogenic profile of clear cell renal cell carcinomas and furthermore,

suggested a novel immunogenic component in lung squamous cell carcinoma. Unfortunately, their immune signature is too general and does not provide enough resolution to delineate individual immune profiles from gene expression data.

Abbas *et al* proposed cell-specific signatures using purified cell populations from blood samples, but lacked validation in a solid tissue setting [91]. Work done by Bindea *et al* resulted in the curation of 24 leukocyte-specific signatures from publically available purified cell population data using solid tumour controls [102]. These signatures have recently been used in conjunction with ssGSEA to immunoprofile 19 cancer types [103]

2.2.2 Pathway Analysis

GSEA techniques can also be used to measure the relative activity of a molecular signaling pathway gene expression data. Molecular signaling pathways are typically annotated based on prior biological knowledge of protein interaction, typically inferred through perturbation experiments or co-expression studies [104]. Manually curated pathways often contain an amalgamation of interacting biological processes, such as gene expression regulation, and only a subset of genes tend to contribute to the series of molecular interactions constituting a specific event [105]. Furthermore, GSEA methods can delineate distinct processes from gene sets, by scoring and identifying gene subsets on the basis of contribution to a key event.

The 2005 paper by Subramanian *et al* provided 1,325 pathways as part of their Molecular Signatures Database (MSigDB) [99]. Today, MSigDB provides 17779 gene sets including 4872 immunological signatures corresponding to immune cell types, phenotypes and activity. Typically, these gene sets are curated from microarray experiments measuring the immune system under two different states and selecting the top 200 most differentially expressed genes. Gene sets generated in this way are likely to contain genes that are co-expressed as a result of co-regulation and not phenotype association [99]. This issue is mitigated by GSEA, since the method implicitly delineates the main biological process from the pathway admixture. Gene ontology (GO) terms encompass a range of immunological pathways ranging from antigen processing and presentation to cytotoxic mediated apoptosis. These have been curated and stored in immune-specific repositories such as InnateDB [106] or DC-ATAS [107], or more general repositories such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) [108] or MSigDB.

One shortfall of the GSEA method arises from overlapping pathways, since the same genes can contribute to many different biological processes. Overlapping gene sets intro-

duce an implicit dependency structure amongst GO terms that can bias enrichment tests, even when multiple hypothesis testing correction is applied [109]. To address these issues, several followup methods to GSEA have been proposed. For example, Alexa *et al* model the dependency structure between GO terms as a hierarchical graph, and reduce biases by iteratively pruning genes belonging to significant terms higher in the hierarchy. They also propose a scoring method, whereby genes in a GO term are weighted based on the scores of their neighbours [110]. Jiang *et al* address overlap bias by calculating p-values for each pair of pathways and also the intersection [111].

A further issue arises when attempting to apply multiple hypothesis testing correction to p-values generated from GSEA. A typical GSEA analysis will typically query hundreds, if not thousands of gene sets successively generating large sets of p-values. The overlap between GO terms violates the non-independence criterion of many methods. Correspondingly, several elegant solutions have been proposed that attempt to circumvent the multiple testing correction stage altogether. Bauer *et al* propose embedding each GO term in a Bayesian network, and leveraging probabilistic models to examine all the terms at once, thus avoiding multiple hypothesis testing correction. Lu *et al* use generative models to identify GO terms most likely to have generated the experimental observation [112].

2.3 Understanding Antigen Presentation

The antigen processing and presentation pathway forms an integral component of the immune system. In section 1.1.3, it was mentioned how neoantigen burden can correlate with T cell recognition and activation. Upregulating T cell activation mechanisms is a fundamental goal of cancer immunotherapy. As such, characterising mechanistic breakages in the antigen processing and presentation pathway is a substantial goal in computational immunology. This challenge is typically addressed through two well-established approaches:

1. **Human Leukocyte Antigen (HLA) Typing** asks: does this person's HLA allele confer a selective for tumour cells in the microenvironment?

2. **Cancer Epitope Analysis** asks: what neoantigens are expressed by this tumour? How are they related to the immune system and clinical outcome?

HLA typing does not give rise to immunotherapies itself, but can provide robust risk factors enabling patient stratification. Cancer epitope analysis can be used to stratify patients and also forms the basis of a novel type of immunotherapy. In this section, we will provide a brief overview of approaches 1 and 2 along with their clinical implications.

2.3.1 HLA Typing

Heuman Leukocyte Antigens (HLA) genes code for the Major Histocompatibility Complex (MHC) proteins in humans which play a central role in antigen processing and presentation [113]. Introduced in section 1.13, these complexes transport peptide fragments to the surface of cells. Although there are only 6 HLA genes in the body (3 coding for MHC class I and 3 for class II), HLAs are examples of *polymorphic* genes, with over 17,000 different identified alleles existing within and between individuals [114]. This high level of polymorphism means that with the exception of identical twins, no two people share exactly the same set of HLA alleles [115].

Specific germline HLA alleles have been associated with a variety of ailments including inflammatory diseases and cancer [116, 117], subsequently promoting the use of HLA-*typing* approaches in understanding cancer progression and anti-tumoural immunity. Additionally, somatic mutations in HLA genes have been linked to upregulation of cytolytic activity in cancer [69]. HLA variant calling is challenging due to the polymorphic nature of the gene and the fact that most reference genomes do not account for this allelic heterogeneity. Therefore, most approaches for resolving HLA alleles make use of a HLA allele database such as one maintained by International ImmunoGenetics (IMGT) [118]. An example of such an approach was proposed by Boegel *et al* , who call HLA alleles by aligning RNA-seq reads to IMGT reference sequences [119]. Shukla *et al* , use an approach combining whole exome sequencing (WES) data and Bayesian modeling techniques to predict HLA alleles. Szolek *et al* propose a linear integer programming approach to predict HLA alleles from RNA-seq, WES and WGS data, demonstrating an overall accuracy of 97% in benchmarking experiments [120].

2.3.2 Cancer Epitope Analysis

The processing and presentation of intercellular protein fragments to the immune system (covered in section 1.4) is a crucial system for maintaining homeostasis. Fragments of presented antigens that elicit T cell recognition are known as *epitopes*. In particular, these are the parts of antigens to which B-cell generated antibodies bind. Higher epitope prevalence is characteristic of increased T cell recognition, and it has been shown that mutational epitopes are positively correlated with the influx of reactive cytotoxic T cell and patient survival [121, 122]. Identifying patient-specific mutational epitopes can abet immunotherapy efforts, through therapeutic antibodies and chimeric antigen receptor (CAR) engineered T cell administration [123]. In addition, patient stratification on the basis of mutational epitopes

can distinguish whom are most likely to benefit from immune checkpoint blockade or other immunotherapies [124, 122].

Peptide binding to a MHC I molecule is implicated as the most significant step in the antigen processing and presentation pathway. As such, the binding strength between a peptide and MHC I, otherwise known as the *binding affinity*, can be used as a proxy for cytotoxic T cell immunogenecity [125]. A range of different algorithms have been proposed to predict the binding affinity of a peptide given a HLA class I allele. Nielsen *et al* were the first to train neural networks to predict binding affinities of peptides. They credit their impressive classification accuracies to an encoding schema that takes into account the chemical similarities of amino acids [126]. Alternatively, Bui *et al* use an matrix method that summarises binding affinities along the peptide chain in terms of individual amino acid contributions, and compute an overall score using a polynomial function [127]. Their methodology, termed Average Relative Binding (ARB), returns rankings of specific sized peptides and MHC class I alleles. More recent MHC I binding affinity predictors are now capable of taking into account a larger range of peptide lengths and HLA alleles [128, 129].

Binding affinity prediction for class II MHC molecules is less straightforward due to poorer characterisation of binding motifs and greater variability in the length of peptides that can bind to them [130]. Therefore, unlike class I methods, sequence alignment must be used to determine the binding motif of the class II molecules [131]. State of the art methods such as NETMHCII take into account not only the peptide-binding cleft of the molecule, but the flanking amino acids too, since they have the can potentially affect peptide-binding affinities [132].

Proteomics datasets generated from mass spectrometry (MS) or reverse phase proteomics array (RPPA) techniques can reveal the extent to which the proteome influences the cancer epitope landscape. Bassani-Sternberg *et al* use MS to demonstrate a correlation between inter-cellular protein abundance and HLA presentation, whilst validating several HLA-bound peptides as epitopes in a colon cancer cell line [133]. The increasingly high throughput nature of these assays increases the attractiveness of proteomics approaches for epitope profiling and sample-wise measures of immunogenecity.

Methods that interrogate the direct interaction between the T cell receptor (TCR) and MHC-peptide structures are currently under development. These measurements are better metrics for immunogenecity than MHC-peptide binding affinities since they implicitly incorporate features from the adaptive immune system. For example, Birnbaum *et al* use

deep-sequencing and yeast epitope curation tools to predict the T cell response of unseen examples [134]. Zhang *et al* use florescence microscopy approaches to measure the rate of dissociation between the MHC-peptide molecule and the TCR [135].

Although much has been achieved in the study of cancer epitopes, progress is impeded by an incomplete understanding of the origins of antigenic peptides bound to MHC molecules [136]. For example, it is hypothesised that a significant fraction of antigenic peptides originate from defective ribosomal products (DRiPs) [137], or peptides that were malformed during the translation stage of mRNAs into proteins.

2.4 Computational Pathology

Thus far, we have discussed techniques such as gene expression analysis, WES, WGS and MS which permit the characterisation of sub-cellular biological processes in bulk or single cell tissues, but provide no morphological context. Pathologist assessments of suspected lesions are routinely carried out through the morphological examination of biopsy slides, where features are tallied to provide clinical scores of stage and grade. Much of computational pathology is concerned with developing models that can make similar conclusions regarding tumour presentation, or proposing novel features that can better stratify patients on the basis of survival [67, 71]. Historically, pathologist assessments were limited to H&E stained images, but with the advent of novel cancer biomarkers, their assessments have widened to include IHC stained images and flow cytometry [138]. In this section, we provide a brief overview of automated computational methods for image analysis, focusing on their applications to cancer immunology.

The overwhelming majority of primary cancer diagnostics and post-resection follow-up studies proceed through H&E stained image examination, making them a widely available resource. Furthermore, almost all slides will have accompanying patient clinical information, providing opportunities for large-scale feature extraction and survival correlation studies [67]. Mining large numbers of features is a laborious task that will often involve a series of pathologists working in parallel. Not only is this manually expensive, but it also introduces noise through the variability in pathologist scoring of images. The purpose of computational pathology is to automate this process and return summary statistics of interesting features. Generally, pipelines consist of three primary steps: data pre-processing, object detection and segmentation, and finally classification. These steps are detailed in the following paragraphs, along with context-specific examples from gold-standard pipelines.

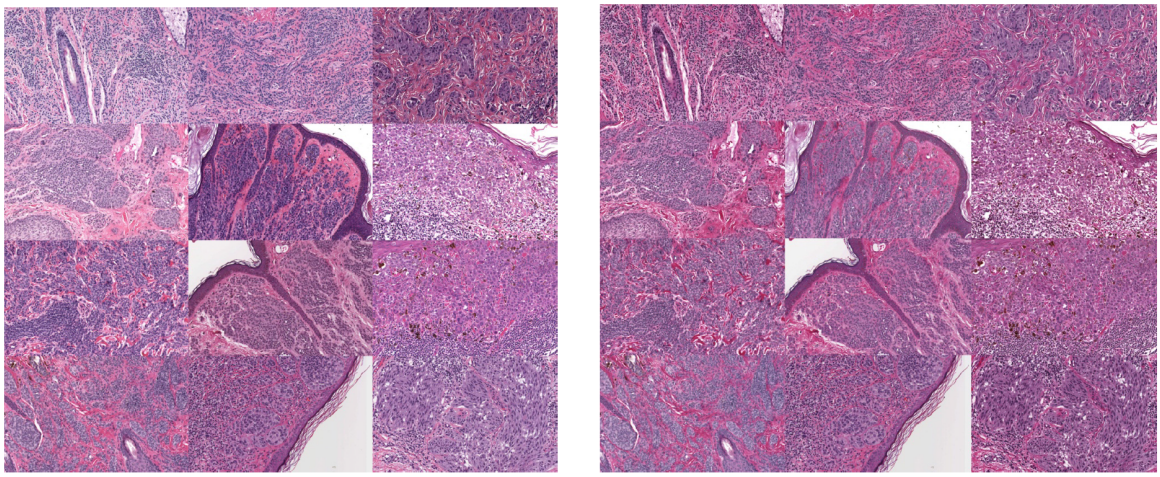


Fig. 2.3 A set of histological images stained under different protocols (left) co-normalised using the Macenko method (right). This image has been reproduced from [139].

Data Pre-Processing A computational pathology pipeline must be robust to analysing histological slides that are different from one another due to age or differences in preparation technique that lead to staining inconsistencies between samples (as is often the case with data that has been aggregated from different processing centres). For optimum accuracy during the image analysis phase, the data must be preprocessed to minimise these inconsistencies. Macenko *et al* . suggested a normalisation method that co-normalises the colour space of images processed under different protocols and conditions using the optical density and singular value decoposition projection [139]. Fig. 2.4 shows an example of a set of images with staining inconsistencies before normalisation and the same set co-normalised using the Macenko method. Since this method is completely unsupervised, it can be readily incorporated into a computational pathology workflow. The embedding, freezing and sectioning process involved in the initial stages of image generation could potentially give rise to other artifacts in the image such as folded tissue, air bubbles, and object clumping due to low spatial resolution. Pre-processing protocols can be designed to eliminate these artifacts before the image processing step or otherwise, they could be classified as artifacts during the image processing step as suggested by the methodology employed by Yuan *et al* [67].

Object Detection and Segmentation Object detection and segmentation are thriving and rapidly developing fields of computer vision that encompass a broad range of disciplines ranging from statistics to computer science to neuroscience. In this section we will briefly overview several methods and discuss the influence of novel deep learning tools on the future

of image analysis in computational pathology. Generally, most object detection and segmentation methods in computational pathology employ a method to distinguish the foreground (cell nuclei) from the background (stroma, extracellular matrix, etc.) and apply the watershed algorithm to segment the cells.

CRImage is a pipeline developed by Yuan *et al* [67] that generates quantitative scores of H&E stained tissues using a classifier trained by an expert pathologist. It achieves cell segmentation through first distinguishing the foreground from the background using Otsu histogram thresholding [140] and then applying a morphological opening operator to break up clusters of objects and distinguish cell-shaped objects from artifacts. A distance transform is then applied, followed by the watershed algorithm in order to separate locally connected cell nuclei. Given that the object detection and segmentation process is unsupervised, this pipeline is robust to many datasets with minimal prior preparation. A significant disadvantage with this method lies with the fact that the combination of thresholding and watershed is known to heavily oversegment images and thus may give rise to misclassified objects in downstream analysis.

The approach implemented by Veta *et al* [141] uses a watershed transform after applying both a radial symmetry transform to mark regions of the image that are close to being disk-like (nuclei are relatively symmetric), and a regional minima markers to avoid oversegmentation of cells. A series of post-processing steps are then applied to the image to clean it up as indicated by the overall schematic in Fig 2.4. A limitation of this method is that a single 1000x1000 pixel image can take 90 seconds or more to segment effectively which increases the computational cost of running it on a series of large slides.

An emerging, extremely powerful classifier known as the *convolutional neural network* (CNN) is revolutionising the field of image analysis with applications to object detection, segmentation and classification. CNNs belong to a class of methods known as Deep Neural Networks (DNNs), and are made up of several convolutional and max-pooling layers that naturally automatically learn the features contained within an image as an internal representation [142]. CNNs first sample local patches in an image and passes them through a filter bank, linking them to units in the next layer which are organised into feature map. Each unit in the same feature map shares the same filter bank due to the fact that most local image features tend to be highly correlated [142], constituting part of a motif. Furthermore, local image statistics are invariant to location, meaning that distinctive motifs can be detected at different locations in the image by units sharing the same weight. The layout of a typical

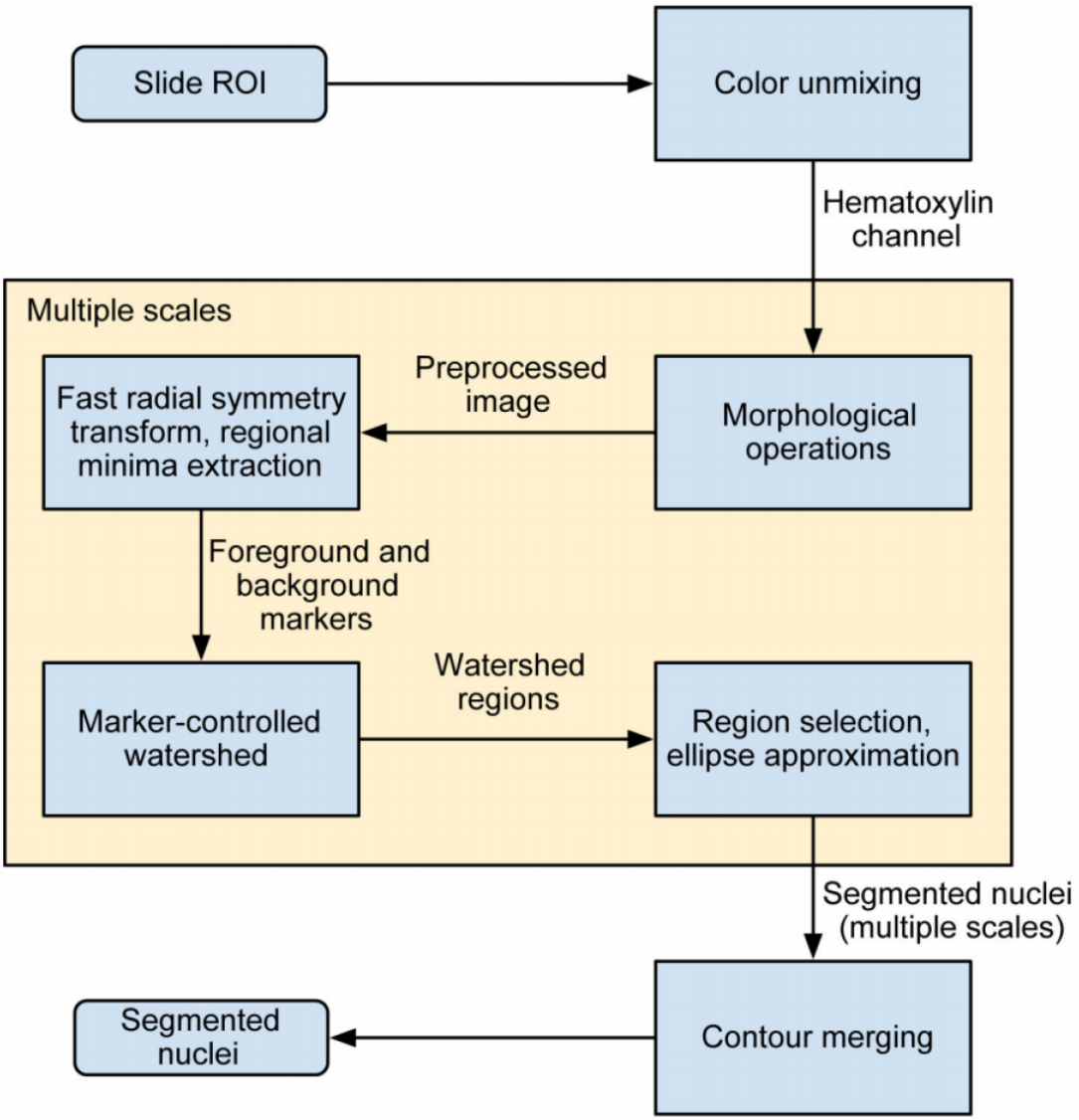


Fig. 2.4 A schematic for the automated cell detection and segmentation package developed by Veta et al. This image has been reproduced from [141].

CNN architecture has been illustrated in Fig 2.5. Proof-of-concept studies have shown the superiority of CNNs in histopathology [143] and it is anticipated that future pipelines will shift towards these machine learning paradigms.

Object Classification After detecting the objects of interest and segmenting the image, the challenge becomes to classify the instances as either stromal, lymphocyte or cancer cells. Typically, an expert pathologist classifies a set of ground truth cells by eye and the

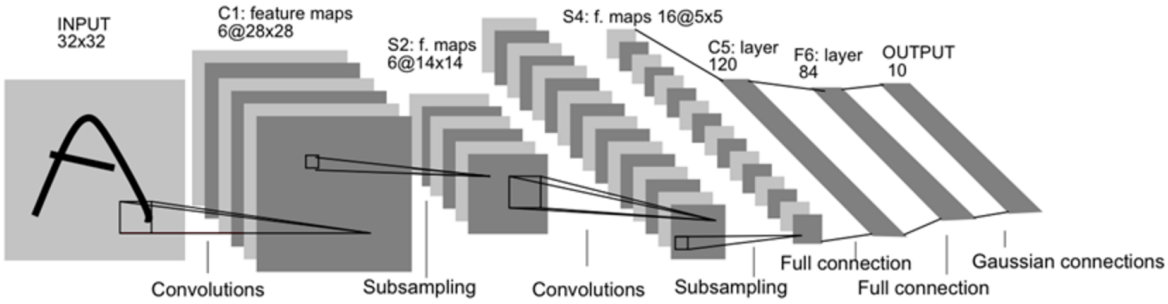


Fig. 2.5 The layout of a typical CNN illustrating the successive layers of feature banks connected to subsampling layers with a fully connected layer for the classification of instances from the feature extraction layers. This image has been reproduced from [144]

morphological features of the cell such as diameter, area and so forth are extracted and used to train a classifier to distinguish a test set [67]. The fact that a selected bag of features is used means that the internal representation of the data piped into the classifier will be constrained, and hence the results would not be as accurate as a system where the features are learned naturally. Deep learning shows superior performance in the classification of cell types as demonstrated by the classification of white blood cells from IHC images by [145], with the one limitation being that the learned features are often too abstract to be informative.

Chapter 3

The Role of Oncogenic KRAS in Pancreatic Cancer Immunity

The last chapter overviewed a range of approaches to TME computational immune profiling at both the sample and cell levels. This chapter evaluates the utility of these methods in elucidating associations between dysregulated tumour cell signalling and the immune landscape of pancreatic ductal adenocarinoma (PDAC). Firstly, I introduce the motivation behind studying immunity in the PDAC setting, and provide a review of subtyping initiatives (section 3.1). Secondly, I introduce work done by my colleagues in identifying embryonic programmes underpinning the oncogenic KRAS transcriptional signature in PDAC, which we subsequently use for patient-based stratification (section 3.2). My main contribution to the project involves mining PDAC gene expression data for immune and stromal features and using them to elucidate key associations between immune agency and the regulators of the KRAS signature. (section 3.3).

3.1 Introduction

Pancreatic cancer has the poorest prognosis of any cancer type, with survival time typically measured in months rather than years after initial diagnosis [147, 148]. There are limited treatment options available for pancreatic cancer and diagnostic efforts to catch the disease at a curable stage are exceedingly difficult. In other cancer types, disease molecular characterisation has directly led to the identification of subgroups amenable to inhibitory checkpoint blockade immunotherapies. Furthermore, identifying biological processes dysregulated by can potentially lead to the identification of new therapeutic targets. However, efforts to identify potential immunotherapies in PDAC are hindered by challenges in molecular

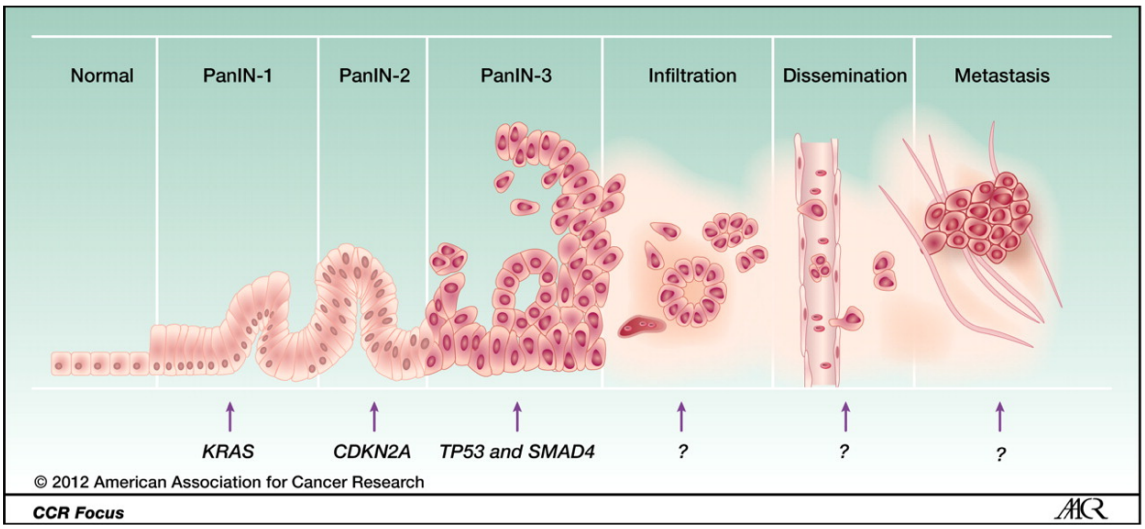


Fig. 3.1 The progressive acquisition of mutations driving normal pancreatic epithelial cells through several neoplastic transformations into a metastatic state. This figure has been reproduced from [146]

characterisation. Motivated by this, the main focus of this chapter is studying the effect of dysregulated PDAC transcriptional processes on immune agency and stromal infiltration.

3.1.1 The role of *KRAS* in PDAC Oncogenesis

PDAC samples display complex and heterogeneous configurations of genomic instability, mostly characterised by frequently occurring point mutations and copy number aberrations [149]. *KRAS* activating mutations are a strikingly common feature amongst PDAC samples, with 90% of tumours typically presenting with a mutation exclusively located on codon 12. [150]. This mutation frequently co-occurs with inactivating mutations in *TP*53 and *SMAD*4, and intragenic mutations in *CDKN*2*A* [151].

Oncogenic *KRAS* is not unique to the pancreas; in fact its prevalence across multiple organs renders it unsuitable as a biomarker for PDAC. In the absence of additional mutations, activated oncogenic KRAS alone forms a histopathologically distinct pre-cancerous lesion called Pancreatic Intraepithelial Neoplasia 1, (PanIn-1). These lesions are typically associated with episodes of chronic inflammation termed pancreatitis. In the absence of additional mutations, cells with sporadic KRAS mutations typically clear away as part of normal cell senescence [152]. Progression into a cancerous lesion is usually the consequence of a function-altering mutation in a tumour suppressor gene, such as p16 or SMAD4/TP53 as

illustrated in Fig. 3.1.

This repertoire of mutations causes perturbations to intra- and inter-signalling mechanisms between cancer cells and the surrounding tissue. Notably, activated oncogenic *KRAS* has been implicated in dysregulating the *PI3K*, *RAF*, and *MEK* signaling pathways, thereby hijacking cell cycle regulation, evading apoptosis and promoting tumour progression [153]. Another example concerns the dysregulation of the $TGF\text{-}\beta$ pathway by *SMAD*4 -inactivating mutations, which are present in at least 50% of all PDAC cases [154]. *SMAD*4 plays a crucial role in the nuclear translocation component of the pathway, with its deletion linked to the formation of more aggressive tumours in activated oncogenic *KRAS* mice [155]. Lastly *CDKN2A* codes for the p16 protein, which plays an integral part of the p16 tumour suppressor pathway. *CDKN2A* inactivating mutations and subsequently, loss of p16 signalling are observed in the majority of PDAC cases [156].

Signalling pathway dysregulation between the tumour epithelial compartment and the remaining TME is not fully understood. In part, this stems from an incomplete characterisation of the transcriptional response to activated oncogenic *KRAS*. Most research into *KRAS*-mediated tissue transformation focuses on mechanism dysregulation in the tumour epithelial compartment and the characteristic stromal component of PDAC, which plays a decisive role in tumour progression [157]. Immunity inference in this domain is made challenging by the fact that PDAC is notoriously difficult to subtype using expression methods - strikingly, PDAC tends to cluster closer to normal pancreatic tissue than to purely neoplastic PDAC cell lines [158], highlighting the extent to which tumour epithelial cells integrate amongst normal and stromal tissues. The next section focuses on this perplexing characteristic in greater detail.

3.1.2 PDAC and the TME

PDAC has a characteristic inflammatory stroma phenotype, with evidence that local fibroblast and stellate cells undergo phenotypic changes during the mesenchymal transition phase of early PanIN lesions [152]. This provides evidence that *KRAS* promotes a tumourigenic microenvironment even at low activity levels. PDAC is vascularised by angiogenic factors expressed by the stellate cells [160]. Furthermore, they produce hyaluronic acid and collagen fibers, helping to maintain the extracellular matrix supporting the tumour bed [161]. Although little is known regarding the exact mechanism of active stroma formation, studies inactivating *KRAS* in early PanIN lesions observe an immediate ablation of activated fibroblasts from the pancreas and a mitigation of pancreatitis [159]. This provides overwhelming evidence

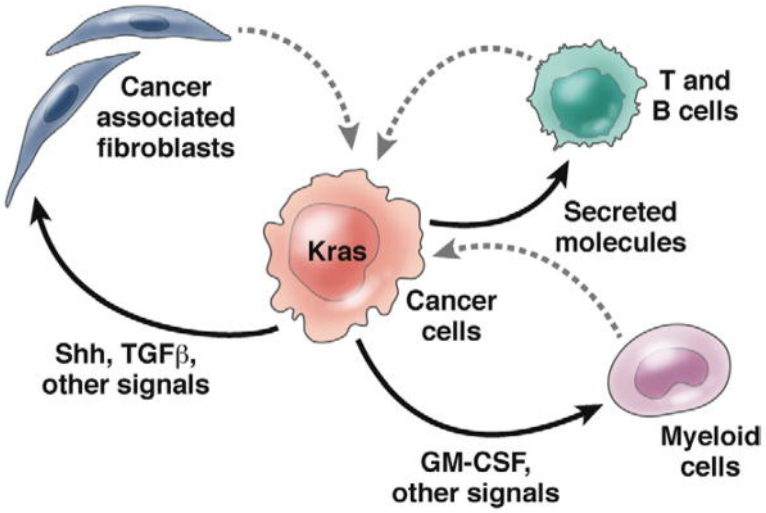


Fig. 3.2 The influence of activated oncogenic *KRAS* on various compartments of the TME including the stromal component and the adaptive and innate immune systems. The origin and nature of *KRAS* derived signalling factors is an area of active research. This figure has be reproduced from [159]

that *KRAS* plays a direct role in maintaining the active stroma, through mechanisms that are actively being characterised (a schematic of potential interactions is illustrated in Fig. 3.2).

One working theory suggests that cells harbouring activated oncogenic *KRAS* express Sonic hedgehog (Shh), which upregulates the Hedgehog signalling pathway in stromal cells and contributes to the inflammatory stroma [162]. Lesina *et al* use mouse models to demonstrate the importance of inflammatory cytokine *IL*-6 in early PanIN progression and development of PDAC [163]. Finally, Charo *et al* suggest that *KRAS* activity is connected with the overexpression of prostaglandin E2 (*PGE*2), which acts on stellate cell receptors to promote pancreatic fibrosis [164].

The influence of the tumour epithelial TME compartment on infiltrating immune cells is not well understood. Early PanIN lesions demonstrate extensive infiltration of immunosuppressive leukocytes including TRegs, MDSCs and mast cells [159] with no consistent mechanism behind their representation in the tumour. PDACs have been stratified on the basis of M2 macrophage abundance [165], but no working theory has yet explained the origins of this stratification. *KRAS* and its consequent transcriptional reprogramming is a potential candidate as a regulator of PDAC immunity. In fact, PanIN immune cell depletion has been demonstrated when *KRAS* is switched off [159], signifying a direct role of the mutation in leukocyte recruitment. The following section focuses on characterising oncogenic *KRAS* in terms of PDAC-specific transcriptomic profiles.

3.1.3 Identifying *KRAS*-specific Subtypes

The goal of this chapter is to study the effect of dysregulated PDAC mechanisms on immune agency and stromal infiltration. To this aim, we seek to characterise PDAC in terms of activated *KRAS*-induced molecular changes. Basically, this enables us to identify mechanisms that are differentially regulated between normal and PDAC epithelial cells. Building a picture of *KRAS* from proteomics data would be ideal, since proteins are ultimately responsible for mechanisms contributing to cell physiology. However, proteomics datasets are laborious to assemble and assays typically generate a handful of features (see section 1.3.1). Transcriptomic studies on the other hand are high-throughput and can generate tens of thousands of features from a single assay. Therefore, most approaches focus on subtype identification from gene expression data [166, 167, 168, 169].

At the time this project was proposed, there existed only two other major gene expression subtyping schemes for PDAC. Collisson *et al* were the first to propose a classification scheme, using non-negative matrix factorisation (NMF) with consensus clustering to identify three clinically distinct subtypes [147]. They demonstrate that their schema can be used to stratify cell lines by *KRAS* mutational burden, but do not illustrate the role of these subtypes in the generation of TME phenotypes. Moffitt *et al* use NMF to perform a deconvolution of pancreatic gene expression data into stromal and epithelial compartments [148]. They proceed to perform consensus clustering on stromal component factor genes, deriving two stromal-centric PDAC subtypes: activated and normal. The activated stroma subtype demonstrates a characteristic macrophage and activated fibroblast signature, indicative of chronic inflammation. However, the authors do not place their findings in the context of any governing biological processes, and fail to explain the role of *KRAS* in initiating the transcriptomic changes giving rise to these subtypes. Finally they do not provide a link between their subtypes and the immune landscape.

In order to identify the cellular processes and transcriptional dysregulation brought on by activated *KRAS*, we make use of a powerful well established regulatory network approach termed *master regulator analysis* (MRA) [170, 171, 172]. MRA aims to identify subsets of TFs whose primary function is to coordinate the regulatory activity of a specific trait, thus conferring a distinct transcriptional signature upon the system. Our approach builds on this, whereby we first construct a *KRAS*-specific signature and then look for a specific repertoire of TFs generating it. We term these TFs the master regulators (MRs) of the oncogenic KRAS signature. Theoretically, identified MRs should elucidate the regulatory programs underlying TME features, since activated *KRAS* is intimately linked with stromal and immune cell

recruitment and phenotype transition [152]. The remainder of this chapter introduces our classification schema for PDAC based on MRA, and how the biological processes governing each subtype link to distinct profiles of stromal recruitment and immune agency.

3.2 MRA Subtyping of PDAC

The work done in this section is split into three stages. Firstly we generate an oncogenic KRAS-specific signature from a murine PDAC cell line. Next, we use MRA to identify regulatory mechanisms that underlie the signature. Finally, we use community detection tools to cluster identified MRs into groups of biologically meaningful processes. We demonstrate how the activity of these processes can be used to stratify patients into clinically distinct clusters. The work done in this section was exclusively carried out by my colleagues Dr. Shivan Sivakumar and Dr. Ines de Santiago, and its purpose is to provide the contextual basis for my own independently generated results, presented in section 3.3.

3.2.1 Transcriptomic Datasets

Our study makes use of existing published resources, including both microarray and RNA-seq derived datasets. All microarray data was generated from Affymetrix GeneChip® Human Genome U133 Plus 2.0 arrays. These included raw gene expression data files for six PDAC studies, which we obtained from Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/gds) with accession numbers: GSE17891, GSE15471, GSE16515, GSE32676, GSE2109, GSE49149 and GSE36924 (the last two accession numbers arise from the same ICGC study) [147, 173, 174, 175, 176, 177, 178]. TCGA RNA-seq data was downloaded from the Cancer Genomics Hub using the following search terms: Assembly: GRCh37/HG19; Disease: Pancreatic adenocarcinoma (PAAD); Sample Type: Primary Solid Tumor; Library Type: RNA-Seq; State: Live; Disease: Pancreatic adenocarcinoma. The ICGC and TCGA expression datasets had matched clinical metadata; access to ICGC clinical metadata was granted directly by Peter Bailey, whereas TCGA clinical data was downloaded via cBioPortal.

3.2.2 Computing a Transcriptional Signature for Actviated Oncogenic *KRAS*

KF508 is an epithelial pancreatic ductal cell line that contains a mutant variant of the KRAS allele, with transcriptional repression mediated using a Lox-Stop-Lox cassette right before the gene locus. Following the introduction of a Cre-expressing adenovirus to the cell, the

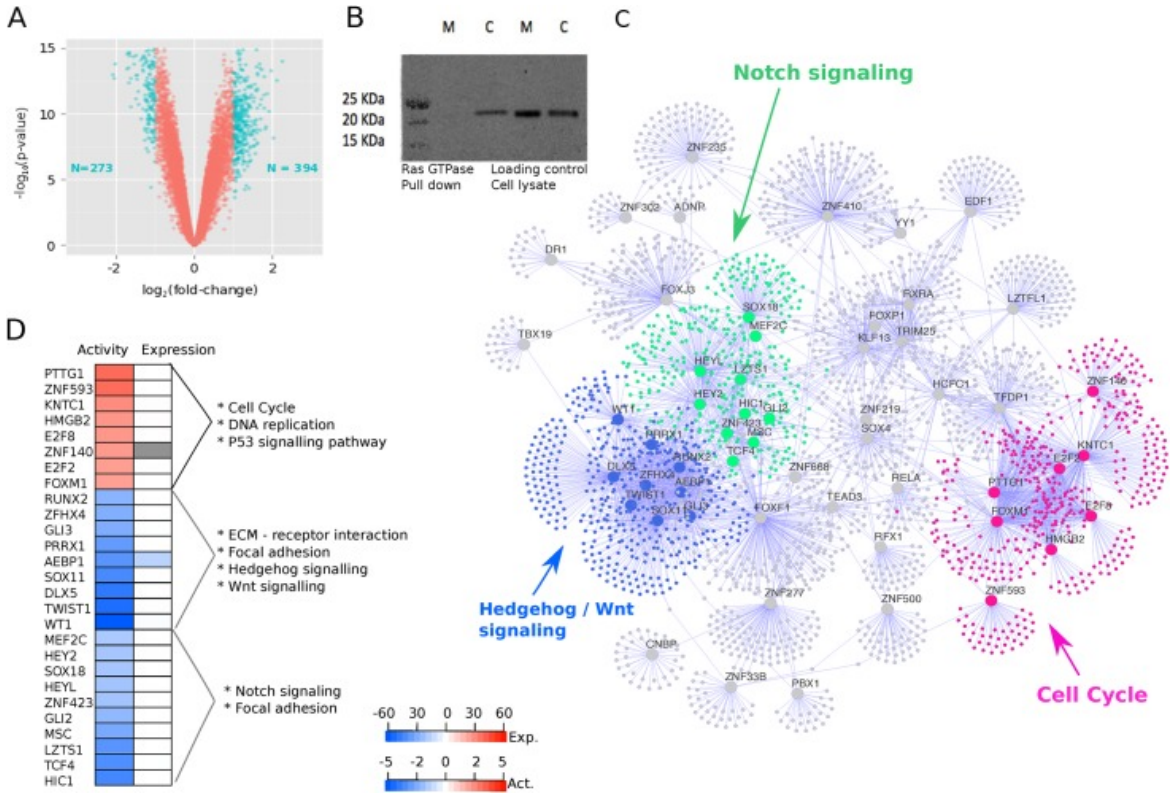


Fig. 3.3 **A.** Volcano plot illustrating differential expression between *KRAS*-on and *KRAS*-off murine PDAC cell lines. Each point corresponds to a microarray probe with non-zero expression. **B.** GTPAse assay showing elevated RAS in *KRAS*-on cells (M - mock; C - cre). **C.** A graph illustrating the 55 identified master regulators of the *KRAS* signature (large nodes) and their regulons (small nodes). Edges here represent regulatory relationships between the MRs and their inferred targets. Coloured subgraphs highlight the three largest communities of MRs identified by the fast greedy community detection algorithm. The three subgroups encompass 27/55 oncogenic *KRAS* MRs and correspond to three biological processes overrepresented for the Hedgehog, Notch and Cell Cycle pathways. **D.** A list of the 27 MRs from the community search, ordered by community membership. The first column represents the activity of the MR as measured by VIPER. The second column illustrates the t-statistic of MR differential expression between the *KRAS*-on and *KRAS*-off murine PDAC cell lines. This figure has been reproduced from [1].

stop sequence is deleted, permitting expression of oncogenic *KRAS*. As a negative control, an equal number of cell samples were exposed to an adenovirus-mock. Six sets of mock- and cre- treated KF508 samples were generated and harvested for total mRNA content. mRNA hybridisation was conducted using Illumina Mousev2 BeadChips using 24 arrays with 46,235 randomly distributed bead types, ultimately interrogating 20,562 genes. 8,472 genes were

interrogated by multiple bead types and 12,014 bead types did not map to any genes.

Differential gene expression analysis between the mock- and cre- groups was performed using the bioconductor package "limma" [179]. 667 differentially expressed probe sets were found in total (Fig 3.3**A**), and the bioconductor package "biomart" was used to map the mouse Ensembl gene identifiers to their equivalent human orthologues (or removed from the list if no orthologue exists). Gene Ontology annotation of the probe sets revealed that the the MAPK and cell growth regulation pathways are overrepresented in the signature, which coincides with our introduction to *KRAS*-mediated oncogenesis (section 3.1.1) .

3.2.3 Characterising PDAC subtypes using MRA

In sections 3.1.2/3.1.3 we discussed how oncogenic *KRAS* can lead to downstream perturbations of signalling pathways. We hypothesise that this mechanism dysregulation arises from differential TF activity presenting between the active and inactive oncogenic *KRAS* states. To identify these MRs, we built a coexpression network for PDAC using six independent pancreatic transcriptomic datasets, for a total of 560 samples. Each dataset was normalised separately and corrected for batch effects. Regulatory networks between TFs and their set of inferred target genes (henceforth known as "regulons"), were built using partial correlations, implemented via the shrinkage estimates of partial correlations method. The significance of each partial correlation was computed using the fdrtool R package [180]. fdrtool enables parameter inference for a variety of null distributions over correlation coefficients, Z-scores, t-statistics, etc. Z-scores are useful statistical instruments since they simultaneously encode information regarding both the significance and directionality of the association. As such, Z-scores were computed for each association pair by hypothesis testing against an estimated null distribution over all associations.

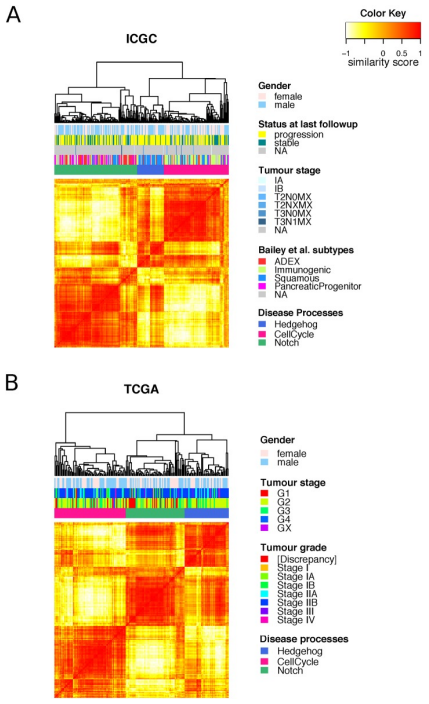


Fig. 3.4 Heatmaps illustrating the unsupervised clustering of patient-wise MR activity into three clear subgroups corresponding to the differential activity of the Hedgehog, Notch and Cell Cycle pathways. Our observations are consistent across the ICGC and TCGA cohorts. This figure has been reproduced from [1].

We compared each network to test the consistency of regulon substructures was conserved across each cohort. The 7 separate networks were collapsed into a single network using Stouffer's method for Z-score integration [181], where the weights assigned to each network are proportional to the size of each study. This was done to ensure that larger cohorts contributed more to the overall integrated network.

The oncogenic KRAS signature was then overlaid onto the integrated network using the msviper function of the VIPER (Virtual Inference of Protein-activity by Enriched Regulon analysis) bioconductor package [182]. Basically, msviper tests for a significant overlap between the genes in the signature and each regulon, with the regulators of overrepresented regulons returned as the MR of the signature. 55 such MRs were identified, with several previously associated with PDAC such as GLI3, AEBP1, and CASP5, and novel MRs proposed such as TCF21, TWIST1, and FOXF2. 8/27 MRs demonstrated upregulated transcriptional activity whereas the remaining 19 MRs demonstrated downregulated activity (Fig. 3.3**D**).

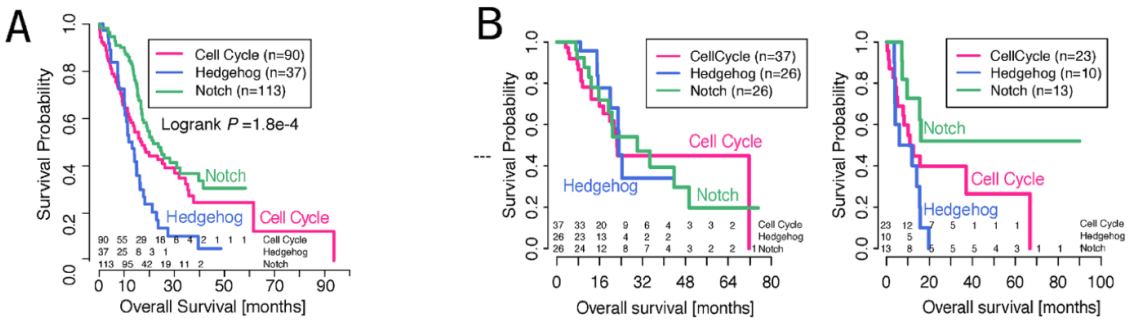


Fig. 3.5 **A.** Kaplan-Meier curves representing survival differences between our schema-stratified ICGC cohort. The logrank test demonstrates a significant survival difference amongst the groups ($p = 1.8e-4$). **B.** Kaplan-Meier curves representing survival differences between our schema-stratified TCGA cohort for patients receiving adjuvant therapy and those not receiving adjuvant therapy. Cox proportional hazards modelling showed a significant stratification (p = 0.02) after correcting for clinical covariates including gender, age, therapy status and stage. This figure has been reproduced from [1].

A community detection approach was used to identify hubs of enriched biological processes in our regulatory network. The three largest communities were comprised of 27/55 identified MRs, and were overrepresented for the Hedgehog/Wnt, Notch and Cell Cycle signalling pathways (Fig. 3.3**C**). The Notch and Hedgehog/Wnt groups clustered more closely together whereas the cell cycle MRs formed a more distant subgroup. When the

activity of the 27 MRs was measured in patient cohorts, three distinct processes could clearly be discerned clustering the sample space. This observation is consistent between both the ICGC and TCGA datasets (Fig. 3.4). Although most PDAC cases have oncogenic *KRAS* as a driver, these results indicate that tumour development forks into three separate paths, dominated by the activity of the Hedgehog/Wnt, Notch and Cell Cycle signalling pathways. These subtypes are clinically distinct, with Notch having the best and Hedgehog the worst survival after correcting for age, gender and tumour stage (Fig. 3.5). The group of patients with dominant cell cycle activity demonstrated the highest mutational burden relative to the other subgroups.

We next focused on identifying immune and stromal features of the TME in the context of Hedgehog/Wnt, Notch and Cell Cycle signalling. Focus was directed towards the two biggest studies in our dataset collection, using the ICGC cohort (n = 242) as a discovery set and the TCGA cohort (n = 178) as an independent validation set.

3.3 The Role of Activated *KRAS* in PDAC Immunity

In the previous chapter, we derived a molecular characterisation for oncogenic *KRAS*-mediated transcriptional dysregulation in terms of Hedgehog/Wnt, Notch and Cell Cycle signalling. We would now like to investigate the association between these signalling pathways and immune/stromal features of the TME. This section is organised in the following way: Firstly, datasets are pre-processed and normalised according to the generating platform. Secondly, we mine the ICGC and TCGA expression datasets for features using computational methodologies introduced in the previous chapter, such as CIBERSORT, ESTIMATE and ssGSEA. We make use partial correlation approaches to link the activity of each signaling pathway to our mined features. Finally, we investigate the assignment of well established PDAC cell lines to our schema.

3.3.1 Data Preprocessing

TCGA Normalisation In RNA-Seq count data, genes with higher average expression across samples also tend to have larger variances than those with lower average expression. This implies that genes with more abundance will be more scattered than genes with lower expression. In statistical jargon, this is referred to as heteroscedasity, and is characterised by an exponential relationship between the rank mean and the standard deviation of gene expression across all genes and samples. There is no straightforward method to correct

this, and normalisation typically proceeds by finding a method that stabilises the variance across the dynamic range of gene expression. The variance stabilising transform has been developed specifically to deal with this issue, and was applied to the TCGA dataset prior to the application of inference tools.

ICGC Normalisation The ICGC GEP is microarray data and thus is normalised differently from RNA-seq data. The output of the Affymetrix array experiment contains close to twenty probes per mRNA target, half of which are used to measure mis-match spots and the non-specific binding of a particular target. The ICGC expression matrix was normalised using the robust multi-array average (RMA) approach, which summarises probe sets using the median polish algorithm and finally employs quantile normalisation in order to make inter-sample comparisons meaningful. Multiple probes mapping to a single gene is a classic issue in the field of bioinformatics since most bioinformatics tools work on the gene level instead of the probe level. These probes typically correspond to different mRNA segments in the gene or splice variants of the same gene. Therefore, the commonly accepted solution of averaging probe signals is not optimal since there is no guarantee that the signals correlate between individual probes. Bourgon *et al* propose a solution that filters probesets by overall variance [183]. Using this, they demonstrate that the discovery rate during differential expression analyses increases by as much as 50%. They confirm that this represents a significant performance increase over filtering by overall mean. Therefore, we selected probes with maximum variance since they are most likely to deliver the most discriminatory power within the sample space. Probes mapping to more than one gene were removed from the expression matrix.

3.3.2 Master Regulator Activity Predicts Immune and Stromal Infiltration

To produce individual stromal and immune content estimations for each sample in our cohorts, ESTIMATE [101] was run using the gene expression matrices as input. The Wilcoxon rank sum test was used to look for any significant stratification of the immune and stromal signatures across our subtypes. The stromal signature demonstrated significant overrepresentation in samples belonging to the Hedgehog/Notch subtypes relative to the Cell Cycle subtype; this result was consistent in both the TCGA and ICGC cohorts (Fig. 3.6**B**). The Hedgehog and Notch subtypes demonstrated similar patterns of stromal infiltration, with Notch being slightly more enriched for the signature in the ICGC cohort (ICGC: p = 1.7e–05, TCGA: p = 7.8e–08; Wilcoxon rank sum test). Interestingly the Hedgehog, Notch and Cell

Cycle subtypes demonstrated significantly varied patterns of immune enrichment, with Notch being the most immunogenic and Cell Cycle being the least. Although Notch and Hedgehog samples demonstrate similar stromal contamination profiles, there is a notable contrast in terms of immune enrichment (ICGC: p = 1.7e–05, TCGA: p = 7.8e–08; Wilcoxon rank sum test).

3.3.3 Notch Activity Associated with Upregulated Adaptive Immunity

We wanted to gain a representative overview of the activity of pathways in the tumour microenvironment that are specific to immune function. To facilitate this, the MSigDB [99]), KEGG [108] and BIOCARTA v5.0 [184] gene set collections were used to build a compilation of immune-specific pathways. This was done by filtering all three aggregated gene sets for immunity-related terms including "T cell", "Immune", "Cytolytic" and so forth. 77 pathways were found that satisfied these filtering requirements, and were subsequently included in the analysis. ssGSEA [100] was implemented via the R package "GSVA" [185] to compute enrichment scores for each immunological pathway across all samples in both the TCGA and ICGC cohorts. To infer the activity of Hedgehog and Notch signalling, gene signatures for each biological process were constructed by aggregating their representative MRs and regulons. These signatures enabled sample-specific Hedgehog and Notch signalling enrichment to be measured using ssGSEA. Notably, sample-wise Hedgehog and Notch signalling were found to be significantly intercorrelated across both the TCGA cohorts (Students t-test $p \leqslant 10^{-16}$ and $p \leqslant 10^{-16}$) respectively. The strong association sets up a latent dependency structure between Hedgehog/Notch signalling and the rest of the microenvironment that must be accounted for. First order partial correlations enable associations to be made between pairs of variables whilst correcting for the variance introduced by a third related variable. Using this principle, correlations between immune pathway enrichment scores and Notch/Hedgehog signalling were evaluated using the R partial correlation toolbox "ppcor" [186] which returns association t-statistics and p-values as output. Bonferroni multiple hypothesis testing correction was applied to p-values in order to correct for type I errors. The significance threshold for each association was set at $p \leqslant 0.05$. Of the 77 immune pathways tested, 29 significantly correlated with Notch activity when conditioned upon Hedgehog signalling (Fig. 3.6**A**). Notably, 7/29 of these pathways pertained exclusively to T-cell mechanisms, including activation and proliferation pathways (Fig. 3.7**B**). The remaining 22 pathways referred to general immune functionality, including cytokine production and extracellular-induced apoptosis. In contrast to Notch signalling, the Hedgehog pathway demonstrated zero significant partial correlations with T cell pathways given Notch signalling. Notch signalling dominated Hedgehog signalling for immune pathway enrichment as shown

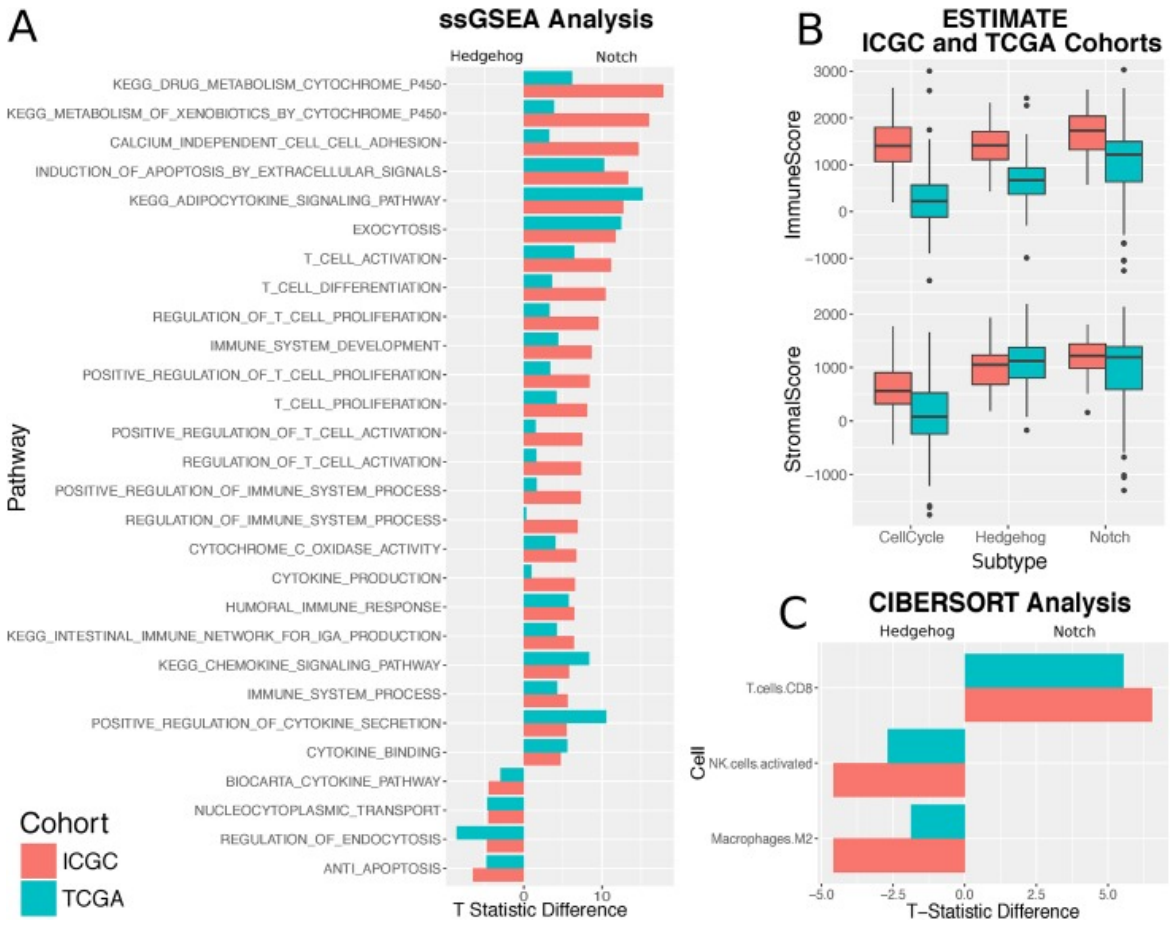


Fig. 3.6 **A.** Associations between ssGSEA scores for each immunological pathway and Notch/Hedgehog signalling were evaluated using hypothesis testing for partial correlations. Each bar represents the difference between the Notch and Hedgehog t-statistics. This metric enables the visualisation of the dominant direction of association between the immune pathway and Hedgehog/Notch signalling. Positive differences imply a stronger association with Notch signalling, whereas negative differences imply a stronger association with Hedgehog signalling. **B.** Boxplots illustrating the stratification of ESTIMATE-derived immune and stromal signature scores by our subtyping schema. **C.** Associations with CIBERSORT-derived leukocyte mixing proportions and the Hedgehog/Notch pathways were evaluated using hypothesis testing for partial correlations. As with **A.**, the t-statistic difference was used to visualise the dominant direction of association. This figure has been reproduced from [1].

in Fig. 3.6**A**.

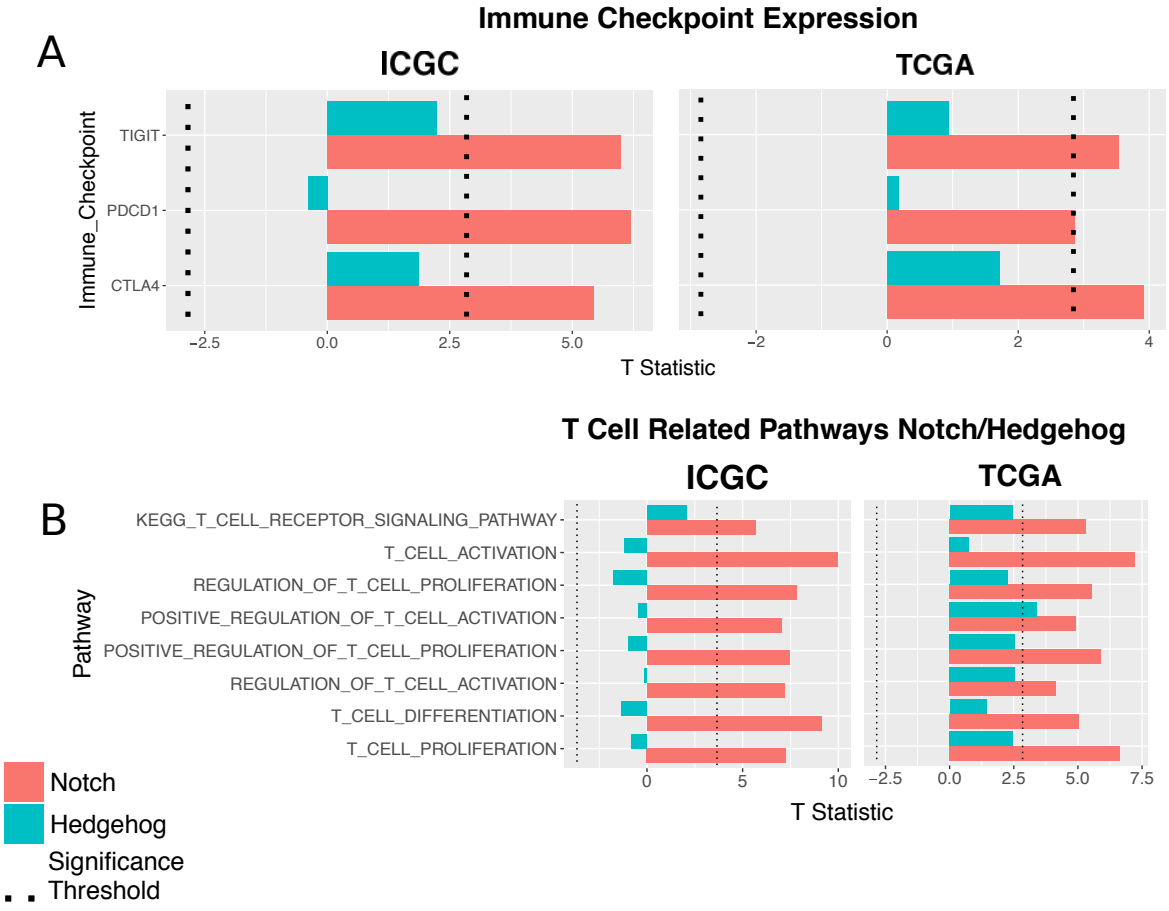


Fig. 3.7 **A.** The association between Hedgehog/Notch signalling and a panel of immune checkpoint markers was quantified using partial correlation hypothesis testing. Each bar represents the t-statistic measurement of association between *TIGIT*, *PDCD*1 (PD-1), and *CTLA*4 expression conditioned on Hedgehog/Notch signalling. Bars crossing the dotted line are significant associations. **B.** The association between Hedgehog/Notch signalling and T-cell related pathways as quantified by partial correlation testing. Each bar represents the t-statistic of partial association and the dotted line represents the significance threshold. This figure has been reproduced from [1].

3.3.4 Leukocyte Composition Related to MR Processes

In section 2.1, we overviewed several deconvoluton methods for delineating leukocyte mixing proportions from bulk tumour gene expression data. CIBERSORT [85] is a partial deconvolution approach that provides a signature matrix containing the transcriptomic for 22 distinct purified subsets of leukocytes. CIBERSORT accepts a gene expression matrix as input, and returns the a estimated leukocyte mixing proportions vector, a p-value and a correlation measuring the success of each deconvolution for each sample. For each cohort,

CIBERSORT was run with 1000 permutations and deconvolutions with p-value of 0.05 or less were considered significant.

The correlation score output from CIBERSORT can be used as a metric to indicate the degree to which CIBERSORT leukocytes are overrepresented in a sample. Comparing this to the immune score output of ESTIMATE, we observed strong correlations for both TCGA and ICGC (Pearson's test $R = 0.55$ and $R = 0.84$ respectively). Samples classified as "Cell Cycle" illustrated a negligible median ESTIMATE immune score; and therefore the Cell Cycle signalling pathway was not interrogated for further associations. T-statistic associations between Hedgehog/Notch signalling and leukocyte mixing proportions were evaluated using first order partial correlations. p-values determined by the t-statistic were corrected for multiple hypothesis testing type I errors using the Bonferroni formulation in the ICGC cohort only. A p-value threshold of $p \leqslant 0.05$ defined the significance threshold in both cohorts. We illustrate the dominant phenotype by computing the t-statistic difference between the Hedgehog and Notch associations (Fig. 3.6**C**).

These tests revealed a significant overrepresentaiton of CD8+ T cells in samples with elevated Notch signalling. In contrast, Hedgehog signalling was significantly associated with a dominant M2 Macrophage and Natural Killer cell signature. These results are conserved well across the ICGC and TCGA cohorts.

Finally, a panel of immune inhibitory checkpoints, including CTLA4, PD-1, PD-L1, TIM3 and TIGIT, were tested for association with both the Hedgehog and Notch pathways through the use of first-order partial correlation analysis. Results between the ICGC and TCGA cohorts correlated well, with CTLA4, PD-1, and TIGIT expression demonstrating significant positive correlations with Notch activity, and no significant association with Hedgehog activity (Fig. 3.7**A**).

3.3.5 Cell Line Classification

A microarray dataset containing the expression profiles of 44 pancreatic carcinoma cell lines was acquired from the Broad Institute's cell line resource. Gene expression centroids for each subtype were computed from the ICGC cohort using the R package "pamr". The ICGC cohort was chosen over TCGA since it is also a microarray dataset. The centroids were used to classify each cell line into our schema using the shrunken centroids method [187] and strikingly, all 44 cell lines classified into the "Cell Cycle" group. Cell lines are hypothetically homogeneous cell populations of a single cellular lineage, leading us

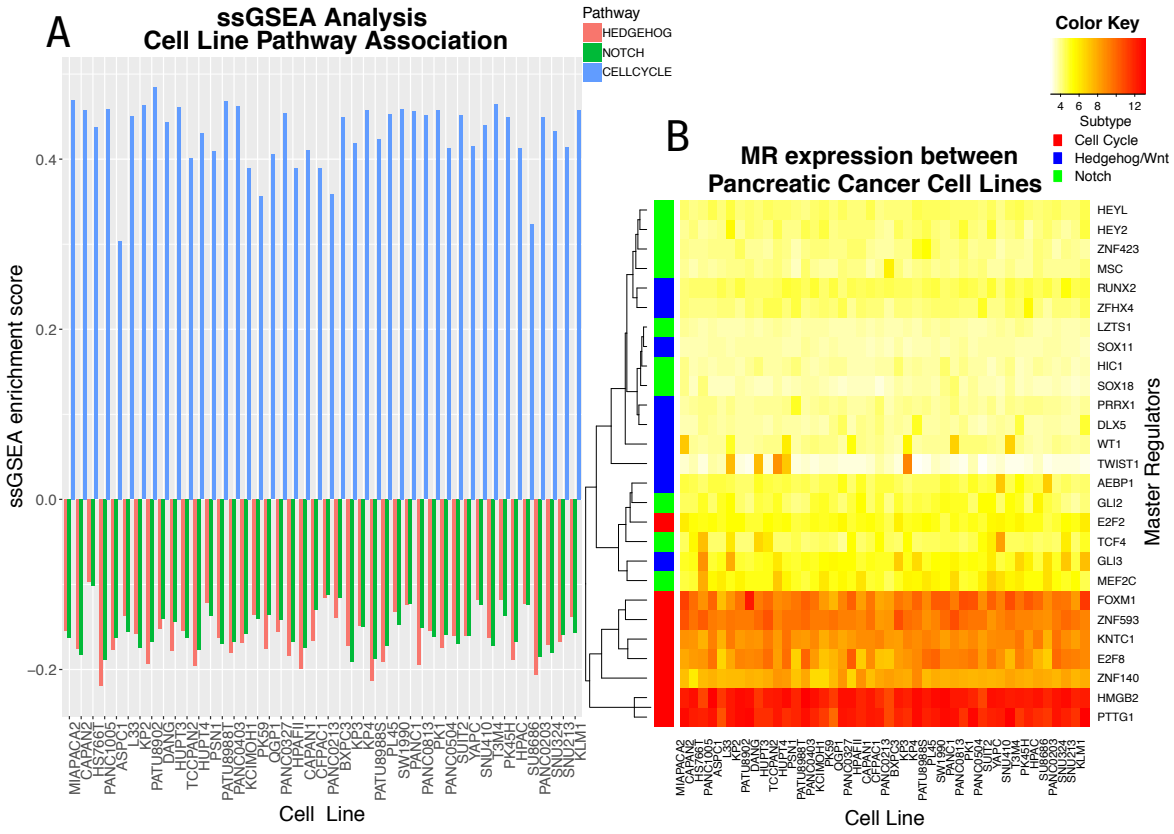


Fig. 3.8 **A.** A transcriptomic dataset containing the expression profiles of 44 pancreatic carcinoma cell lines was acquired from the Broad Institute's cell line resource. ssGSEA analysis revealed that the majority of cell lines demonstrated a dominant pattern of Cell Cycle signalling with downregulated Notch and Hedgehog signalling. **B.** Clustering the cell lines on the basis of master regulator expression clearly defines a distinct cluster of overexpressed MRs corresponding to the Cell Cycle pathway. In contrast, the master regulators of Hedgehog and Notch display underrepresented expression relative to the Cell Cycle MRs. This figure has been reproduced from [1].

to hypothesise that Hedgehog and Notch signalling are part determined by cancer cell interactions with other components in the microenvironment such as the stroma. As an orthogonal mode of validation, Hedgehog, Notch and Cell Cycle activity were computed for each cell line using ssGSEA, with Cell Cycle signalling activity dominating the other pathways (Fig. 3.8**A**).

3.4 Discussion

PDAC is a devastating disease that confers a poor prognosis relative to other cancer histologies. Attempts to describe PDAC in terms of mutational drivers commonly identify activating mutations in *KRAS* as a consistent disease progenitor. Molecular subtyping efforts do exist, but they attempt to place *KRAS* in the context of clustered subtypes rather than directly derive clusters from a *KRAS* transcriptional signature. Hence, they are often unable to find mechanisms linking the progenitor mutation to TME features and clinically observable phenotypes.

Our approach attempts to find the master regulators (MR)s of the *KRAS*-induced transcriptional response. To achieve this aim, we use MR analysis to isolate 55 TFs forming the regulatory apparatus for a *KRAS*-specific transcriptomic signature derived from the KF508 murine cell line. Community detection approaches identified 27/55 MRs as the central components of three embryonic signalling pathways: Hedgehog, Notch and Cell Cycle. The activity of these MRs enabled the stratification of patients into three clinically distinct subtype with distinct mutational burden profiles. Each subtype is characterised by three routes of PDAC development facilitating differential transcriptomic responses. Computational immunology tools were used to mine each transcriptional response for features of adaptive and innate immune agency. We find that each route of PDAC development is characterised by a unique immune and stromal profile, possibly contributing to the prognostic discrepancy between subtype.

3.4.1 Notch/Hedgehog Signalling and Stromal Recruitment

Patients classified as either "Hedgehog" or "Notch" by our schema were represented by high levels of stromal infiltration. Sonic hedgehog (Shh) factors expressed by epithelial cells are known to have a paracrine like-function on the surrounding tissue, upregulating the effect of Hedgehog signalling in the surrounding stroma [188]. Furthermore, Hedgehog signalling in stromal tissue is implicated in the maintenance of a reactive stroma and thus, chronic pancreatitis could be an explanatory factor for the poor prognosis of the "Hedgehog" subtype [189]. The expression of Hedgehog factors takes place in both the cancer-cell autonomous and stromal compartments of the TME, making it difficult to resolve how much signalling is taking place where. Notch signalling is mostly isolated to the cancer-cell autonomous compartment with little known about its activity in the stroma [157]. The overrepresentation of stromal content in Notch-dominated samples most likely arises due to concurrent Hedgehog signalling; the two pathways are strongly inter-correlated. Furthermore,

the epithelial growth factor receptor *EGFR* pathway is upstream from Notch signalling, and has previously been implicated as a regulator of pancreatic fibrosis [190]. As such, the exact role of Notch in stromal signalling warrants further inquiry.

3.4.2 Hedgehog Upregulates Immunosuppression

Our study found significant overlaps between Hedgehog signalling and downregulated adaptive leukocyte functionality. The Hedgehog process was especially enriched for M2 macrophages, a well established mediator of immunosuppression. We mentioned earlier that M2 macrophages are used to stratify patients on the basis of survival through unknown recruitment mechanisms. Since the activated stroma employs several well characterised systems for M2 macrophage polarisation (section 1.2.2), our findings recapitulate an important feature of the PDAC microenvironment in the context of mechanisms driven by oncogenic *KRAS*. Interestingly, our study found that Hedgehog was overrepresented for Natural Killer cells but underrepresented for cytotoxic T cells. This suggests that immunosuppression takes place at the adaptive immune level. Although Hedgehog patients have the worse prognosis, they could stand to benefit from newer targeted immunotherapies.

3.4.3 Notch Signalling Promotes Immune-Induced Tumour Cytotoxicity

We uncovered a significant over-representation of adaptive immune function in the Notch process, including pathways annotating T cell activation and proliferation. Our observation is consistent with work done by Palaga *et al* , who demonstrate that Notch regulation is required for *IFN-*γ production and the proliferation of CD8+/CD4+ T cells [191]. Furthermore, Maekawa *et al* suggest that the cytotoxic action of CD8 T cells can be upregulated by Notch ligands binding to the Notch1 and Notch2 transmembrane receptors [192]. Indeed, several studies have demonstrated that some tumours promote a Notch-suppressive TME to escape immunosurveillance [135, 193]. This may account for our observation of greater CD8+ T cell representation in tumours with greater Notch signalling. Furthermore, Mathieu *et al* show that Notch signalling upregulates the binding of the NICD-RBPj molecule to the *PDCD*1 promoter, causing it to become overexpressed on the surface of activating CD8+ T cells [194]. This is directly consistent with our observations of a correlation between PD-1 and Notch signalling, suggesting that patients in the "Notch" category may be potentially amenable to inhibitory checkpoint therapies.

Chapter 4

Probabilistic Models for Regulatory Network Reconstruction

In this chapter, I introduce the motivation behind causal inference and how it can be used to deepen our understanding of cancer-immunity interactions (section 4.1). I introduce a range of graphical modeling methods, which differ with respect to how extensively they explain the correlation between two events given a third event (section 4.2). Bayesian networks are dealt with in section 4.3, and how they pertain to genomic signaling events. Section 4.4 introduces methodologies for learning these networks from real data, and the benchmarking of well established approaches. We end the chapter by proposing a new method that aims to overcome limitations encountered whilst benchmarking other approaches, thus forming the basis for chapter 5.

4.1 Introduction

4.1.1 Association Studies in Cancer Immunology

Investigating cancer-immune interactions is challenging given the heterogeneous and evolving nature of the cell populations. The two compartments form complex multicellular ecosystems: cancer cell development widely viewed as an evolutionary process [4] whereas the immune system is comprised of adaptive and innate immune agents that demonstrate different levels of phenotypic plasticity and memory [195]. High throughput technologies such as NGS generate vast quantities of data that have been successfully mined to provide insight into cancer by developing mechanistic theories that support clinical decision making [76, 1]. These techniques have been extended to cancer immunogenomics, where NGS data such as whole exome sequencing (WES), whole genome sequencing (WGS) and RNA-seq

are probed for links to the immune system [69, 85, 117].

For example, Rooney *et al* integrate RNA-seq, WES and SNP array data into a working theory of tumour-specific cytolitic activity [69]. Martins *et al* combine RNA-seq and IHC data to validate PTEN loss as a driver for ovarian cancer [196]. Orru *et al* gauge the genomic basis for variations observed in 95 immune cell types across over 1600 patients, using a combination of WGS and flow cytometry [197]. Finally, Yuan *et al* use lymphocyte expression signatures alongside H&E features to build an integrative classifier for survival [67]. These studies propose that using a combination of data modalities for inference may be more advantageous than focusing on singular data types.

Association-based methods have been used to characterise TME interactions within bulk [102, 198] and micro-dissected tumour transcriptomes [199]. For example, Ali *et al* [198] showed that patterns of immune infiltration varied between molecular subtypes of breast cancer in bulk tumour transcriptomes; and Oh *et al* [199] derived stromal-epithelial co-expression networks from micro-dissected tumour data to investigate crosstalk within the tumour microenvironment. Although these studies can determine regions of the genome most likely to explain phenotypic variance, they cannot provide any insight regarding the direction of the association. This problem has prompted a need for causal inference using more sophisticated statistical methods.

4.1.2 The Need for Statistical Causal Inference

A fundamental goal of cancer immunogenetics is to find mechanisms in the tumour that give rise to observable immune phenotypes. Association studies alone are non-mechanistic and lack the context required to temporally order two correlated events. As such, a mechanism describing the emergence of a phenotype given a mutation is not complete if causality cannot be proved. Causal inference provide a means of predicting the perturbation of a variable given a change in a related causal variable [200]. As described in the following paragraphs, causal inference can be extremely resourceful and provides an excellent means of reconstructing mechanisms from existing datasets.

Mechanism Discovery and Prediction A well-defined biological mechanism is desirable for two reasons. Firstly, tracing the origins of a phenotype through layers of molecular interactions can potentially reveal targets amenable to therapy. Secondly, it will enable the prediction of perturbations to a phenotype under different experimental settings, helping

guide future hypotheses.

Association analyses amongst variables involved in an experiment aims to infer parameters of a "true distribution" from which samples were drawn. In addition to gauging the association between variables, these parameters help make predictions regarding past and future events, and probabilities can be updated when further experimental evidence is incorporated into our beliefs. So as long as the experimental conditions that generated the data are invariant, the parameters and thus, inference should not be perturbed. This directly arises from there being nothing in probability distribution functions to enable the prediction of the change of one distribution parameter with respect to another [200]. Causal inference generalises this analysis, enabling us to infer probabilities under both static and changing conditions, making it an excellent approach for mechanism discovery.

In a typical NGS experiment, the number of variables typically ranges in the tens of thousands, with a limited means of reducing the search space of alternative causal hypotheses. Even with the high throughput capacity of NGS, the number of experiments required draw causal conclusions between thousands of genes becomes extremely large and impractical. Non-automated causal discovery protocols do not demonstrate adequate robustness given the limited feasibility of experimental deployment, whereas the rate of development of computational processing power and storage space can support search algorithms that operate over large causal hypotheses domains. Statistical approaches such as graphical modeling can fit many variables into the context of an overall causal structure, describing the influence of variables over one another. These statistical implementations narrow down many causal hypotheses into only a handful.

Data Recycling

The availability of extensive multi-omics resources from projects such at The Cancer Genome Atlas (TCGA) [201], the International Cancer Genome Consortium (ICGC) [202] and other datasets published alongside smaller studies (Gene Expression Omnibus for example [203]) have been extensively mined for associations but not causal interactions. A causal model can integrate multiple data modalities and provide sufficient resolution for reconstructing a mechanism of interest. These types of analysis rarely require supplementary experimentation other than end-stage validation to functionally characterise the causal drivers of a phenotype [81]. As such, causal inference methodologies stands to benefit from the ever increasing

sample sizes of data collecting initiatives.

The remainder of this chapter focuses on providing a statistical methods overview for causal inference. These approaches focus on visualising the dependency structure between observations in the form of probabilistic graphical models. The fundamental premise of causal inference stems from the notion of *conditional independence*, which we formally introduce in the next following section.

4.2 Conditional Independence Models

A graph is defined by $G = (V, E)$ ordered pairs of V vertices and E edges. In probabilistic models, each vertex $v \in V$ corresponds to a random variable X_v and we let the set $\mathbf{X} = (X_1, X_2 ..... X_n)$ represent a set of n random variables. Let $\mathbf{Y}$ denote M draws from $\mathbf{X}$, such that $y \in \mathbf{Y} = (x_1, x_2, ..., x_n)$. $\mathbf{Y}$ can be represented as a $n \times M$ matrix with rows representing each random variable and columns representing each sample. E represents the dependency between the random variables $X \in \mathbf{X}$ represented by V and thus forms the graph topology. A special case of G has directed edges and permits no cycles, referred to as a directed acyclic graph (DAG), and plays a prominent role in causal inference.

The biological interpretation of each $v \in V$ depends on the type of data being analysed. Microarray data is used to construct transcriptonal gene regulatory networks, protein data used to build protein-protein interaction networks and so forth. Vertices needn't be restricted to singular modalities: v_1 may represent RNA-seq measurements and v_2 represent image features for example. Multimodal conditional independence models built and validated using multi-omics datasets form the basis of work done in chapter 5.

4.2.1 Skeletal Association Graphs

Biological mechanisms typically proceed through chains of biochemical interactions between molecules. Consequently, much of functional genomics focuses on attempting to cluster together similar groups of random variables representing molecules such as mRNA transcript abundance [204, 205]. This stems from a fundamental heuristic that genes which are coexpressed are more likely to be coregulated: i.e. their expression is controlled by the same regulatory programs. Coexpression networks can be built to represent these relationships, with the vertex space V representing genes and E representing pairwise correlations between

genes. Typically, the quality of this correlation is judged by generating a p-value, with a threshold set to determine if an edge is kept or discarded. Association networks can also be constructed from multimodal datasets, illustrating the relationship between alternative families of molecules, such as mRNAs and proteins.

Initialising Skeletal Association Graphs To proceed with an association analysis, we first formally define the notion of a similarity measure between variables. A similarity measure is any relationship between observed variables that can be represented in terms of a joint distribution. Correlations are the simplest similarity measures, with 0 correlation signifying statistical independence between two sets of observations. Correlations networks are accurate even when considering the issue of the large n small M problem (the number of genes in an expression assay will almost always exceed the number of samples in a dataset). In $\mathbf{R}^2$ space, two vectors $\{y_i, y_j\} \in \mathbf{Y}$ are said to be linearly dependent if α can be determined such that $y_i = \alpha \times y_j$ holds. Depending on the value of α, the correlation between y_i and y_j is either exactly -1 or 1. If α cannot be determined, $\{y_i, y_j\}$ are said to be linearly independent, and will take on $\mathbf{R}^2 \in (-1, 1)$. Variables will not always be linearly related, especially if they represent different data modalities. Discovering non-linear relationships between $\{y_i, y_j\}$ requires moving away from correlations to pair-wise mutual information or non-parametric estimation techniques like kernel functions [206, 207].

The second step of graph construction involves controlling for the multiplicity of simultaneous association inferences. Rejecting the test statistic at the error level of 0.05 means that 5% of all hypotheses will on average be rejected incorrectly. Considering that the expected number of incorrect rejections (formally known as Type I errors) grows linearly with the number of pairwise tests forming a complete graph, ($n(n-1)/2$ in the context of n vertices), this quickly becomes problematic. The p-value is defined as the probability under the null hypothesis (i.e. the statement "no relationship exists between A and B) of sampling extrema values from a statistical model. A practical method consists of building a $\mathbf{R}^2$ *null distribution* through comprehensive column and row-wise permutations of the data matrix, and comparing the edge distribution of the real data graph to that of the null distribution. Type I errors can be addressed using methods ranging from the more conservative Bonferroni correction to the more passive Benjamini-Hochberg procedure [208].

Skeletal Association Graphs are not Causal Let X, Y and Z be random variables that demonstrate an underlying non-zero correlation structure. In the context of a biological system, we can let X, Y and Z pertain to the expression profiles of three genes. The depen-

dency relations between these variables cannot be distinguished through associations alone. A functional genomics solution exists to solve this dilemma: perturbation experiments. By perturbing the "resting" state of a cell, the resultant values of X, Y and Z can be used to figure out a dependency structure between the variables. However, this approach becomes laborious when the number of genes in the correlation structure becomes large, as is typical in NGS experiments. Thankfully, a statistical approach can be used in place of perturbation experiments. Such methods not only search for significant correlations, but those that cannot be explained by the variance in other variables. If the association between X and Y can be explained by the variance in Z, X and Y are said to be *conditionally independent* given Z.

Definition If X, Y and Z are three disjoint subsets of variables under a joint distribution $P(X,Y,Z)$, then *X and Y are conditionally independent given Z* (denoted $X \perp Y|Z$) if and only if $P(X,Y|Z) = P(X|Z) \cdot P(Y|Z) \; \forall \; X = x, Y = y$ and $Z = z$ assignments for which $P(Z = z) > 0$.

For interpretation, $X \perp Y|Z$ can be stated as "After observing Z, subsequently observing Y gives me no new knowledge to understand X". In the context of associations, $X \perp Y|\mathbf{Z}$ implies that the correlation between X and Y can be explained by the variance in Z. This is equivalent to writing $P(X|Y,Z) = P(X|Z)$, a generalisation for the independence relation $P(X|Y) = \frac{P(X,Y)}{P(Y)} = \frac{P(X) \cdot P(Y)}{P(Y)} = P(X)$. The definition also holds when conditioning on a vector of variables $\mathbf{Z}$.

In the context of graphs, conditional independence statements concerning variables encoded by V constitute edges in E. Depending on the size of the conditioning set, special cases of graphs can be built. For example, skeletal association graphs can be seen as a special case with $\mathbf{Z} = \emptyset$. All statistical models we deal with in this chapter are built upon the notion of conditional independence. We will see in the upcoming sections that biological assumptions can be used to minimise the number of variables $|\mathbf{Z}|$ conditioned upon the correlation between X and Y and thus simplify graph structures.

4.2.2 Gaussian Graphical Models

Let $(X_1, X_2, ..., X_i, ..., X_j, ..., X_n) \in \mathbf{X}$ be a n-dimensional vector of random variables drawn from the Gaussian N(μ,Σ), where the covariance matrix Σ is positive definite and therefore invertible. The distribution has the precision matrix $\mathrm{P} = \Sigma^{-1} = (p_{ij})_{ij}$.

Definition The *partial correlation* between $(X_i, X_j) \in \mathbf{X}$ given $\mathbf{Y} = \mathbf{X} \backslash \{X_i, X_j\}$ is defined by

$$\rho_{i,j \cdot \mathbf{Y}} = -\frac{p_{ij}}{\sqrt{p_{ii} p_{jj}}}. \tag{4.1}$$

It follows that between any two vertices $i, j \in V$ where $i \neq j$,

$$X_i \perp X_j | \mathbf{Y} \Leftrightarrow \rho_{i,j \cdot \mathbf{Y}} = 0. \tag{4.2}$$

Partial correlations enable us to construct a class of undirected graphs called *Gaussian Graphical Models*, which encode statements of conditional independence between pairs of nodes given all other nodes in the network [209].

Definition Let V be a vertex set such that each $i \in V$ represents a random variable given by $X_i \in \mathbf{X}$. A *Gaussian Graphical Model* (GGM) is a graph where the edge set E is defined such that $\{i, j\} \in V$ are adjacent iff the partial correlation between them given the remaining vertex space $Y = V \setminus \{i, j\}$ is non-zero (i.e., $\rho_{i,j \cdot \mathbf{Y}} = 0$).

In practice, GGMs can be built from most datasets with invertible covariance matrices Σ. Partial correlation coefficients can be computed using entries in the precision matrix $P = \Sigma^{-1}$. Statistical tests are then employed to gauge whether $\rho_{i,j \cdot \mathbf{Y}}$ is significantly different than zero. The GGM over the vertex space is populated by edges where this condition is satisfied.

Advantages over Association Graphs Association graphs are built from entries in the data covariance matrix Σ and provide a useful visualisation of relationship structures between variables. GGMs are built instead from the precision matrix Σ^{-1}, and provide a visualisation of the correlation between variables after correcting for the variance of all other variables in the data. Correlation coefficients and partial correlation coefficients both demonstrate strong and weak criteria for dependency testing between variables as shown in Table 1. Correlation coefficients tend to be non-zero for most pairs of variables and thus are strong criterions of independence, but concurrently, weak for establishing meaningful dependence statements. Conversely, partial correlation coefficients tend to be zero after conditioning upon all other variables, and are thus strong measures of dependence but a weak criterion of independence.

GGMs are encoded by strong criteria for dependence and thus will be more sparse than association graphs. The vanishing associations prove useful in filtering correlations between variables whose values stem from an underlying regulator. In the context of biology,

	Dependence Criterion	Independence Criterion
Correlation Coefficient	Weak	Strong
Partial Correlation Coefficient	Strong	Weak

Table 4.1 Comparison between Correlations and Partial Correlations

these may be genes whose expression profiles are coregulated by a transcriptional program. Furthermore, GGMs can isolate variables of interest, such as interactions between a TF and a regulated gene that may be weakly correlated in an association graph, but strongly related in terms of partial correlations with respect to neighbouring genes.

GGM construction is not trivial GGMs are easy to build in theory, but the process becomes less trivial when we work with real data. Most omics datasets are high dimensional, with the number of variables almost always tending to exceed the number of samples ($|\mathbf{X}| \gg N$). A covariance matrix built from data with relatively few observations tend to be singular and therefore non-invertible. This is highly problematic, as partial correlations depend on a precision matrix that cannot be computed from a singular covariance matrix. One way of addressing the issue of $|\mathbf{X}| \gg N$ is simply to make $|\mathbf{X}|$ smaller using variable selection. Toh and Horimoto were the first to do this, implementing a clustering analysis as a preprocessing step of GGM construction, thereby reducing the effective size of $|\mathbf{X}|$ [210]. Alternatively, Schafer *et al.* recommend using the Moore-Penrose pseudoinverse in place of regular matrix inversion for singular matrices [211].

GGMs are examples of fully conditional graphs. They ask "Does the correlation between two variables hold if I condition on every other variable?". The next section examines a powerful but simpler class of graphs with a conditioning space size of 1.

4.2.3 Triplet Graph Models

Triplet graph models are special cases of GGMs, where associations are conditioned upon only a single other variable rather than the remaining vertex set. Formally,

Definition Let $\{i, j, k\}$ be three unique points in a vertex set V corresponding to random variables $\{X_i, X_j, X_k\} \in \mathbf{X}$. A *Triplet Graph* is a graph where the edge set E is defined such that i and j are adjacent iff $\rho_{i,j \cdot k} \neq 0 \; \forall \; k \in V \setminus \{i, j\}$. These edges encode conditional

independence statements of the form

$$X_i \not\perp X_j | X_k \Leftrightarrow \rho_{i,j \cdot k} \neq 0 \qquad \forall k \in V \setminus \{i, j\} \tag{4.3}$$

into the graph. For gene coexpression, triplet graphs greatly reduce the dimensionality problem of GGMs since we only test triplets of genes at any given time. This property makes triplet graphs ideal for any multiomics dataset, where the set of variables $\mathbf{X}$ can include transcriptomic, genomic and proteomic data or any combination thereof. Furthermore, it can be shown that a triplet graph coincides with a GGM over the same vertex space provided the GGM contains no cycles.

4.3 Bayesian Networks

Thus far, we built association graphs by asking: "what does the correlation structure of $\mathbf{X}$ look like?". We then considered fully conditional graphs by asking: "what does the correlation structure between X_i and X_j look like when we condition upon all of $\mathbf{X} \backslash \{X_i, X_j\}$?" Finally, we built triplet graphs by asking: "what is the correlation structure between X_i and X_j look like when we condition upon each other variable X_k?".

We can summarise these graphs into a query concerning general conditional independence by asking: "what does the correlation structure between X_i and X_j look like hen we condition upon all *subsets* of the remaining nodes $\mathbf{X}_S \subseteq \mathbf{X} \backslash \{X_i, X_j\}$?". The edges in this graph are undirected, although the dependency between vertices in GGMs and triplet graphs already imply some form of directionality in terms of conditional independence. We can formalise this directionality in the form of a *Bayesian Network*. According to Jensen, a graph $G = (V, E)$ must satisfy four criteria to be a Bayesian Network [212].

1. E must consist of directed edges between variables.
2. The set of random variables $\mathbf{X}$ assigned to each vertex in V must consist of a set of mutually exclusive, finite values.
3. The graph must be acyclic, and hence by property 1) a DAG.
4. A vertex $v \in V$ is represented by the conditional probability $P(X_v | \mathbf{X}_{pa(v)})$ where $pa(v) \subseteq V \setminus v$ represents v's parents in the graph.

The joint probability distribution over all vertices is expressed as the product of all node conditional probabilities:

$$p(X_V) = \prod_{v \in V} p(X_v | \mathbf{X}_{pa(v)}). \tag{4.4}$$

The union set of v's parents, its children and its children's parents is known as a *Markov Blanket*.

Definition Given a a vetex $v \in V$ in a Bayesian network, the collective union of its parents, children and children's parents is known as a Markov blanket. It follows that if δV is the Markov blanket of v,

$$p(X_v | \delta V, \mathbf{X} \setminus \delta V) = p(X_v | \delta V), \tag{4.5}$$

and thus v is conditionally independent of any subset of nodes in the network given its Markov blanket.

This property, along with the DAG structure of Bayesian networks naturally implies an ordering of vertices. Now that an introduction to Bayesian networks has been given, the process by which to build them from data can be outlined. This will be broken down into several subsections that answer the following questions:

1. How do we compute $p(X_v | \mathbf{X}_{pa(v)})$?
2. What is conditional independence in the context of DAGs?
3. How do we select the best Bayesian network structure?

4.3.1 Node Probability Distributions

The definition of a Bayesian network includes a probability distribution attached to each vertex $v \in V$, conditional on its parents $pa(v)$: $p(X_v | \mathbf{X}_{pa(v)})$. X_v is a random variable that can be draw from many different discrete and continuous distributions. In practice, multinomial distributions are used for discrete X_v and Gaussian distributions for continuous X_v. For Bayesian networks modeling multi-omics data, a combination of discrete and continuous distributions will often be used. We briefly introduce each of these distributions:

- *Discrete Case*: Let X_v represent a random variable assigned to $v \in V$ and let $\mathbf{X}_{pa(v)}$ represent the set of v's discrete parent random variables. The probability distribution

attached to node v can be described by a multinomial distribution function:

$$X_v|X_{pa(v)} \sim Multin(1, \theta_{v|\mathbf{X}_{pa(v)}}) \tag{4.6}$$

where $\theta_v = \{\theta_{v|\mathbf{x}_{pa(v)}}\}$ parameterises the function.

- *Continuous Case*: Let X_v represent a random variable assigned to $v \in V$ and let $X_{pa(v)}$ represent the set of v's continuous parent random variables. The probability distribution attached to node v follows the Gaussian distribution function:

$$x_v|x_{pa(v)} \sim N(\mu_v, \sigma_v^2) \tag{4.7}$$

where μ_v is given by the linear contributions from each parent node in the following model:

$$\mu_v = \beta_v^0 + \sum_{i \in pa(v)} \beta_v^i x_i \tag{4.8}$$

and β_v^i are regression coefficients. The parameterisation of Eq. 4.7 can therefore be expressed $\theta_v = \{\beta_v, \sigma_v^2\}$.

4.3.2 Conditional Independence in DAGs

In the case of Bayesian networks, each edge refers to a *conditional independence statement*. We have seen how to compute independence statements by reading the topology of undirected graphs, but how does it work in the case of directed edges? A criterion called directed *d-separation* was formulated to address this issue. Basically, d-separation is a measure of independence between a two sets of disjoint observations $\mathbf{X}$ and $\mathbf{Y}$ given a third disjoint set of observations $\mathbf{Z}$. The idea here is to associate the notion of dependence with that of path connectedness. Let $\mathbf{X}$ and $\mathbf{Y}$ be a set of nodes in a DAG $G = (V, E)$ separated by a sequence of consecutive edges known as a *path*.

Definition Let $\mathbf{Z}$ correspond to the set of nodes that lie along all paths between $\mathbf{X}$ and $\mathbf{Y}$. A path p between two nodes is said to be d-separated by $\mathbf{Z}$ if and only if:

1. p contains a chain $x \to j \to y$ or a fork $x \leftarrow j \to y$ such that the middle node j is in $\mathbf{Z}$
2. p contains a "collider" $x \to j \leftarrow y$ such that the middle node j is not in $\mathbf{Z}$ and no descendant of j (in G) is in $\mathbf{Z}$.

Under this definition $\mathbf{Z}$ is said to d-separate a node $x \in \mathbf{X}$ from $y \in \mathbf{Y}$ in G (denoted $(x \perp y|Z)_G$ if and only if it d-separates every path from x to y.

Bayesian networks over the same vertex space may be defined by a different edge space, but still contain the same statements of conditional independence. These networks are said to *Markov Equivalent*.

Definition Let $G_1 = (V, E_1)$ and $G_2 = (V, E_2)$ be two DAGs defined over the same vertex space V. G_1 and G_2 are said to be Markov equivalent if for every three subsets of vertices $X, Y, Z \subseteq V$, $(x \perp y|z)_{G_1} \leftrightarrow (x \perp y|z)_{G_2}$ $\forall\, x \in X, y \in Y, z \in Z$.

Markov equivalent networks are statistically indistinguishable from one another, placing a direct theoretical limit on structure inference from even data with many observations. Fortunately, Markov equivalent networks share the same underlying skeletal structure, meaning we can distinguish them from other less relevant structures [213].

4.4 Structure Learning

Learning dependency relations E over a set of nodes V in a Bayesian network involves computing the conditional independence state of node pairs given all other subsets of nodes in V. This is more challenging than in the case of our conditional independence graph examples where we conditioned on either a single other variable or all other variables. The mathematics behind Bayesian network structure learning is tractable in settings where $|V|$ is small, but becomes more challenging as we scale up the number of variables. A straightforward way of bypassing this issue is through scoring-based approaches. This section will briefly describe different scoring techniques that can be used for structure selection.

4.4.1 Maximum Likelihood Estimation

Bayesian networks are basically DAGs with conditional independence distributions attached to each node, and the joint distribution over all nodes given by equation 4.4. Therefore, we can set up a one-to-one correspondence between the set of DAG structures and the set of joint distributions over the node space. This enables us to do model selection by choosing the DAG that best fits the data $\mathbf{X}$.

The basic idea behind maximum likelihood estimation (MLE) is to search for parameters θ that maximise the likelihood $p(\mathbf{X}|M,\theta)$ where M is the joint distribution over the nodes corresponding to a particular DAG. This technique is useful in cases where parameters are not known *a priori*. Formally, MLE computes the following likelihood L for each DAG

$$L = \arg\max_{\theta} p(\mathbf{X}|M,\theta) \tag{4.9}$$

Model selection proceeds by selecting the network with the greatest likelihood. However, models with many parameters can lead to overfitting, and consequently, larger likelihoods may be observed for more complex DAGs than simpler ones. It is possible to correct for overfitting using scoring regularisation that factors in the number of model parameters. A widely used approach to penalising MLE scores by model complexity are the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC).

Akaike Information Criterion (AIC) The AIC provides a criterion for penalised model selection from the maximum likelihood estimates of several different models given the same data. It basically regularises the MLE by adding a term to the negative log likelihood that is directly proportional to the number of model parameters. A good way to think about the AIC is that it seeks the model with the greatest fit to the data with the smallest number of parameters. For any given model M, its AIC is given by

$$AIC_M = -2\log_e(\arg\max_{\theta} p(\mathbf{X}|M,\theta)) + 2P \tag{4.10}$$

where P is the number of free parameters. Generally, the model with the smallest AIC is chosen. Unforunately, the AIC does not perform null hypothesis testing and thus it cannot make statements regarding the absolute quality of a model. Consequently, all models may fit the data poorly, and the AIC will only permit "least-worst" ranking. Lastly, the AIC typically selects for high dimensional *approximating* models that do not tend to the true model that generated the data. This also means that the AIC has a tendency to find more overfitting models.

Bayesian Information Criterion (BIC) Like the AIC, the BIC also provides a criterion for model selection which is based in part on model likelihood estimation. The equations themselves differ very little: the multiplicand of the number of free parameters P in BIC

scales logarithmically with the number of samples N.

$$BIC_M = -2\log_e(\arg\max_{\theta}(\mathbf{X}|M,\theta)) + \log_e(N)P, \tag{4.11}$$

where $N \gg P$. When $N \geqslant e^2$, the BIC confers a stricter penalty on complex models than the AIC. Furthermore, the BIC attempts to find the true model amongst a set of models and is consistent, i.e. as $N \to \inf$, the BIC selects the correct model with probability 1. This is achieved by placing an uninformative uniform prior over the number of free parameters and estimating the model with the highest posterior probability.

When it comes to choosing between the BIC or AIC as an information criterion for model selection, it is useful to bear in mind the individual assumptions both methods make. The AIC is a constant plus the relative measure of the distance between the latent true likelihood function of the data and the proposed fitted likelihood function of the model. The BIC is an estimate of the posterior probability of the model being true under a Bayesian framework. In both cases, the model with the smallest criterion is the one closest to the true unknown likelihood function. Despite their theoretical differences, the main discriminant between equations 4.10 and 4.9 lies in the size of the penalty assigned to more complex models, which is greater in the BIC than the AIC for large N. Therefore, the AIC has a higher probability of selecting an overfitting model, and the BIC is more prone to underfitting. So how do we select the best information criterion for a given set of models? As a rule of thumb, the AIC should be used in circumstances where a false negative result would be more misleading than a false positive. Correspondingly, the BIC should be used in circumstances where a false positive result would be considered more misleading than a false negative.

4.4.2 Bayesian Model Scoring

In Bayesian network scoring, we search for a model topology M that best describes the data D by computing the posterior probability

$$p(M|D) = \frac{p(D|M)\cdot p(M)}{p(D)} \tag{4.12}$$

over all the model space. In Bayes's formalism, $p(M)$ is intended to describe our *prior* beliefs over the model space. With no reason to believe otherwise, $p(M)$ can be represented by a uniform distribution over all models. If we have any *prior* knowledge on what we expect the dependency structure in the data to look like, $p(M)$ may have a different distribution. $p(D)$ averages the data likelihood over all possible models and is used to normalise $p(M|D)$,

(i.e. force its integral between $(-\infty,\infty)$ to equal unity). $p(D|M)$ represents the marginal probability of observing data D given model M. It is computed by marginalising out all parameter configurations Θ in the full model:

$$p(D|M) = \int_{\theta\in\Theta} p(D|M,\theta)p(\theta|M)d\Theta. \tag{4.13}$$

Unlike the MLE approach, we are not optimising over the model parameters θ, but instead we integrating them out of the equation. For this reason the θ term appears in the MLE framework but not in Eq. 4.12. As a consequence, the Bayesian setup implicitly prevents overfitting by averaging over many parameter configurations.

Network Structure Priors Structure priors abet model simplification by integrating prior biological knowledge from orthogonal datasets. Refining and building upon existing networks is often more advantageous than starting from scratch, since it enables computational overheads to be reduced and provides statistical confidence in new inference. There is a straightforward approach to building DAGs from structural priors. Given a prior network $\zeta = (V,E)$, we can define a search space $\mu(\zeta) \subseteq \zeta$. DAGs $M \in \mu$ are equally likely to be observed in this new subgraph. The prior probability can therefore be described by a uniform distribution over the search space

$$p(M) = \begin{cases} \frac{1}{|\mu(\zeta)|}, & D \in \mu(\zeta) \\ 0, & D \notin \mu(\zeta). \end{cases} \tag{4.14}$$

We include M in the model space if $p(M) \neq 0$. The prior network therefore acts as a filter over the model space, drawing focus to those supported by biological evidence. Examples of structural priors include protein-protein interaction networks and co-expression matrices derived from other studies. $p(M)$ can be used both as an informative prior in Eq. 4.12, and can be used to inform model selection in the MLE framework.

Prior conjugacy To be able to analytically compute the integral in Eq. 4.13, the model prior $p(\theta|M)$ and the marginal likelihood $p(D|M,\theta)$ must display *conjugancy*. If the posterior distribution $p(M|D)$ is from the same family as the prior distribution $p(\theta|M)$, $p(\theta|M)$ is said to be a conjugate prior for the marginal likelihood function $p(D|M,\theta)$. Without conjugancy, the posterior distribution cannot be computed analytically and we require heuristic approximations [214]. Fortunately, conjugate priors already exist for the node probability distributions introduced in section 4.3.1, and we briefly outline each of these here.

- *Multinomial Case*: Likelihoods described by multinomial distribution exhibit conjugancy with the Dirichlet prior. Let θ_v be a vector of multinomial parameters. Then

$$\theta_v \sim Dirichlet(\alpha_v) \tag{4.15}$$

  where α_v is the Dirichlet scale parameter.

- *Gaussian Case*: Likelihoods described by a multivariate Gaussian function exhibit conjugancy with a Normal-Wishart distribution prior. If $\mu|\mu_0,\lambda,\Delta \sim \mathrm{N}(\mu|\mu_0,\lambda\Delta^{-1})$), with mean $\mu|\mu_0$ and covariance matrix $\lambda\Delta^{-1}$ where $\Delta|\mathbf{W},\nu \sim$ Wishart($\Delta|\mathbf{W},\nu$), then (μ,Δ) has the Normal-Wishart distribution:

$$(\mu,\Delta) \sim NW(\mu_0,\lambda,\mathbf{W},\nu) \tag{4.16}$$

Likelihood Equivalence DAGs may exhibit structural variations but generate the same likelihood scores given any dataset given the same parameterisation. Given two graphs G and G' where $p(G) > 0$ and $p(G') > 0$, G and G' are said to be *likelihood equivalent* iff $p(G|\theta) = p(G'|\theta)$ [215].

An example to illustrate this can be shown simply using a three variable domain $\{X,Y,Z\}$ with parameterisation $\Theta = \{\theta_X,\theta_Y,\theta_Z\}$. If we let $G = X \rightarrow Y \rightarrow Z$ and G' = $X \leftarrow Y \leftarrow Z$, it can be seen that $\theta_{X|Y,Z} = \theta_{X|Y}\theta_{Z|Y}$ is the conditional independence statement for both G and G'. These networks are called *hypothesis equivalent* and therefore imply likelihood equivalence. In causal Bayesian networks, hypothesis equivalence is avoided by asserting that each non-root node is caused directly by its parents and no other node in the network. Heckermann asserts that the assumption of likelihood equivalence is valid regardless, and can be used to learn network structures from transcriptomic data [215].

Enforcing this assumption requires that we place restrictions on our choice of prior parameters. In the case of discrete data, our marginal likelihood takes the form of a multinomial and our prior is a Dirichlet distribution, parameterised by a set $\{\alpha_{i_\delta,\mathbf{i}_{pa(\delta)}}\}$ where the node δ in a DAG is in the state i_δ and whose parents are in the states $\mathbf{i}_{pa(\delta)}$. Ensuring likelihood equivalence is achieved by placing the following restrictions over the Dirichlet parameters

$$\{\alpha_{i_\delta,\mathbf{i}_{pa(\delta)}}\} = \alpha P(I_\delta = i_\delta|\mathbf{I}_{pa(\delta)} = \mathbf{i}_{pa(\delta)}) \tag{4.17}$$

where $\alpha \geqslant 0$ captures the strength of our prior beliefs, termed the "concentration parameter". α is a hyperparameter and is independent of network node values, and whose value strongly determines the regularisation of the network structure.

Network Regularisation Regularisation is a technique used in machine learning to address model overfitting and solve ill-posed problems (specifically finding unique solutions in the context of network inference). Earlier we overviewed solutions to the overfitting problem in GGMs, specifically where the number of variables exceeded the number of samples ($|\mathbf{X}| \gg N$). In fact, regularisation is always needed in cases where $|\mathbf{X}| \gg N$, although different approaches are used in Bayesian networks.

Steck and Jaakola showed that the value of α in Eq. 4.20 consequently influences network regularisation [205]. α is partitioned on the basis of the number of parent node configurations, and thus $\{\alpha_{i_\delta, \mathbf{i}_{pa(\delta)}}\}$ tends to be small for complex models. This implies that large α networks are weakly regularised, since more complex networks have a higher probability of being selected. Another approach involves informatively selecting the distribution around each network node. For example Bulashevska *et. al.* select distributions that build parent-child dependency structures around noise-based logic gates [216]. An obvious disadvantage arising from this is the risk of loss of conjugacy between the marginal likelihood and the model prior. Under these circumstances, the marginal likelihood can no longer be evaluated analytically and heuristic approaches such as Gibbs sampling or variational inference need to be applied.

4.4.3 Model Selection

The framework for inferring optimal models can be generalised into a three component strategy. The basic idea is to first define a *search space* over all models and transverse it using a *search strategy*, assigning "optimality" scores along the way using an appropriate *scoring metric*. We extensively discussed Bayesian and MLE scoring metrics in the previous sections.

Search Space For Bayesian networks, the search space is the set of all networks defined over a vertex set V, each encoding a different dependency structure in the form of conditional independence statements. The number of DAGs permissible on a set of n edges can be written analytically

$$a_n = \sum_{i=1}^{n} (-1)^{i-1} \binom{n}{i} 2^{i(n-i)} a_{n-i} \tag{4.18}$$

where $a_0 = 1$ [217]. This number becomes large extremely quickly, and exhaustively transversing the entire set can become computationally intractable. Since many of these networks will be Markov equivalent, they will encode the same conditional independence statements and thus their scoring metrics will be identical. This implies that the set M of all DAGs over V vertices is strictly larger than the set of DAGs encoding unique conditional independence statements. Chickering *et.al.* employ a methodology that defines a search space only over DAGs that uniquely describe a joint distribution [218].

Castelo and Kocka demonstrate that by moving from DAG space into the space of canonical essential graph representations, the overall number of representations can be reduced by a factor of 4, including Markov equivalent DAGs, [219]. They showed that there are on average 4 times as many equivalent DAGs as there are essential graphs. The sparsity of essential graphs is related to the number of Markov equivalent DAGs, with the number decreasing rapidly as the essential graphs become less dense.

The search space can also be narrowed by focusing on the set of DAG subgraphs rather than entire network structures. Friedman and Koller argue that ordering of nodes form a more regular search space with a smoother posterior distribution [220].

Search Strategy The aim of an efficient search strategy is to return the highest scoring network from a search space using the least computational overhead. The following list of algorithms have been developed to perform optimised selection over DAG search spaces.

1. **Hillclimbing Greedy Search** This relatively simple algorithm initialises a k-connected graph (k being a hyperparameter) graph $G = (V, E)$ where V is the set of DAGs. We call $N_v \subset V$ the *neighbourhood* of $v \in V$ if $\forall n_v \in N_v$, v and n_v are *neighbours* (connected by an edge). The algorithm selects a node at random, then computes the Bayesian or MLE scores for each DAG in the node's neighbourhood. The one with the highest score is selected. The algorithm iterates until no DAG in the neighbourhood has a higher score higher score than the current DAG. This search can become computationally expensive at in large DAG search spaces regardless of the choice of k.

2. **Sparse Candidate Algorithm** This algorithm iteratively constrains the search space to belong to a small subset of candidates by searching for highly dependent sets of nodes.

3. **Ideal Parent Algorithm** For each variable V in the network, this algorithm generates an *ideal parent profile* representing a hypothetical parent variable P_v that is most likely to predict the values in V [221]. Network selection proceeds by searching for DAGs that contain the most number of ideal parents given a set of variables.

Model Confidence Even the most optimal learned probabilistic models will harbour some degree of uncertainty. The reasoning lies in the name: *probabilistic* model. Representing complex multivariate datasets in terms of combinations of conditional distributions and expecting zero residuals is an unreasonable venture. Instead of selecting one *true* model, the plan is to scan the posterior distribution of model likelihoods and make an informed selection. In the context of Bayesian scoring, the ideal method of directly sampling the posterior distribution is unfeasible given the intractability of $p(D)$ in Eq. 4.15.

Markov Chain Monte Carlo (MCMC) can be used to sample posterior distributions $p(M|D)$ that are analytically intractable. The idea is to propose network structures by sampling parameters from $p(M|D)$ under the Metropolis-Hastings acceptance criterion [222]. This sampling is iterated until an approximation to the posterior distribution can be built. An important result in MCMC is that the sampled distribution is guaranteed to converge to the posterior distribution as the number of samples tends to infinity. This method is used widely in the Bayesian framework of learning dynamic DAG structures and node ordering.

Other methods exist to deal with both MLE and Bayesian score evaluated networks. Bootstrapping is a commonly used approach to evaluate model robustness. The idea is to sample variables with replacement from a dataset and assemble a collection of "bootstrap datasets". Network inference is then performed on each bootstrap dataset and the relative frequency of network features is used as a metric for model confidence. Replacement sampling can become problematic by giving rise to identical datapoints in some bootstrap datasets. Consequently, this can increase the collinearity between variables that would otherwise demonstrate weak or no correlation. In datasets with fewer samples, this collinearity effect becomes more dramatic.

Steck and Jakkola correct for this using the commonly used *leave-k-out* approach [205]. Instead of sampling with replacement, all except k samples are chosen at random, effectively

generating a sample-permuted bootstrap dataset. Repeating this process multiple times provides a reliable measure of model uncertainty.

4.4.4 Method Benchmarking

Baysian networks are ultimately visualisation tools for conditional dependence structures in multivariate data. Recapitulating ground truth biological examples is needed to evaluate model robustness and reliability. One way of achieving this is testing how well they reconstruct cellular networks of various complexity. Most benchmarking studies focus on reconstructing networks with a small number of nodes as a proof of concept. Zak *et al.* use a system of differential equations to simulate a set of conditional independence statements and encode a 10-gene transcriptional network [223]. Their benchmarking attempt failed to generate the same network given the data and a mixture of linear and log-linear approaches. However, Husmeier showed the same simulated 10-node network could be better recapitulated using dynamic Bayesian networks (DBNs) by fine-tuning parameters such as the training set size. In fact DBNs were found to display improved accuracy over other methods on the same simulated 19-node network [224]. Smith *et al.* demonstrated how Bayesian networks were highly efficient in recovering the functional network structure of data simulated at multiple biological organisation hierarchies such as gene expression, neural anatomy and behaviour [225].

These examples demonstrate inconsistent success when it comes to recovering network structures from simulated data. We can explain this volatility in three ways. Firstly, model uncertainty can play a decisive role in network reconstruction, a factor which is strongly controlled by the number of data samples. More data also lead to more accurate measurements of model uncertainty (i.e. through leave-k-out bootstrapping). Secondly, these examples focused primarily on simulated observational data whereas most networks in biology are characterised by node perturbations. Data generated from real or simulated perturbation experiments have been shown to enhance network reconstruction relative to observational data alone [226, 227]. Lastly, we saw that simulated data encoding the hierarchical dependency structures of biological processes were reconstructed more efficiently than data simulated from standard distributions. In chapter 5, We propose a methodology that aims to minimises reconstruction errors through hypothesis-driven network simplification - modelling perturbative events in hierarchical biological processes. Cancer is system where perturbations propagate through a chain of biological processes, rooted at a single defining event. In theory, this will enable us to control for uncertainty and improve network reconstruction accuracy.

Chapter 5

Causal Modeling Dissects Tumour–Microenvironment Interactions In Breast Cancer

In the last chapter, I overview statistical approaches to causal inference including graphical modeling methods. Building on these frameworks, I propose an hypothesis-driven approach to learn regulatory mechanisms driving lymphocyte infiltration (section 5.1). I use methodologies introduced in chapter 2 to mine gene expression and image data for immune features (section 5.2). In particular, this analysis incorporates large independent cohorts of genomic, transcriptomic and imageomic data, that are independently pre-processed. (section 5.3). By applying my methods to breast cancer data, I discover novel regulatory hierarchies and mutational regulators of lymphocyte infiltration in expression data, and validate these findings in image data (section 5.4). Finally, I discuss my results and evaluate limitations of my methods in section 5.5.

The work in this project led to the preprint Chlon *et al.* [2] currently hosted on the bioRxiv repository and awaiting publication - https://doi.org/10.1101/144832. I am the copyright holder of this preprint. Furthermore, I am the only author on this paper alongside my supervisor Dr. Florian Markowetz. All results, figures, text and supporting information are exclusively my own work. Sections 5.3 - 5.5 contain text and figures reproduced from my own preprint Chlon *et al.* [2].

5.1 Hypothesis-driven Network Reconstruction

In the previous chapter, the concept of Bayesian networks and statistical tools were introduced that can be used to learn graphical model topologies from real data. The purpose of this chapter is to use these methods to investigate regulatory mechanisms underpinning immune responses in breast cancer. This project aims to look for scenarios where a perturbation in the cancer genome can propagate through the cancer cell and manifest downstream at the immune level. For example, one mechanism might start with the amplification of *PIAS*3, leading to the downregulation of $TFEB$ activity followed by decreased lymphocyte recruitment. This would distinguish it from another scenario where $TFEB$ activity and lymphocyte recruitment could be independent events given *PIAS*3 amplification. To do this an approach was designed with the following questions in mind:

1. What breast cancer mutations are associated with an immune phenotype?
2. Are there any transcription factors (TFs) whose activity correlates with both the mutation *and* the immune phenotype?
3. Is the dysregulated TF activity causal for the immune phenotype, or is the immune phenotype causal for the dysregulated TF activity?
4. Considering questions 1, 2 and 3, can a regulatory structure be found that links together the mutation, TF activity dysregulation and the immune phenotype?

The proposed method orders DAG nodes with respect to sequences of biological processes. The aim is to find an event that we can *anchor* the rest of the analysis to, as this will simplify the resultant network. The anchoring node has no parents, and therefore its Markov blanket limited to its children and its children's descendants. The network anchor can be thought of as the beginning of a sequence of events, and must be empirically substantiated in the context of biology. Cancer is commonly referred to as a *disease of the genome*, and much of what we know about carcinogenesis and progression can be qualitatively described by genomic alterations [228]. It has become the consensus that cancer development begins in the genome through of loss of tumour suppressor function, or gain of oncogenic function [229, 230].

Point genetic mutations cannot be anything other than anchor nodes, since knowing the mutational state of one gene does not enable you to make predictions about another. Genes can sometimes be replicated or deleted concurrently in a chromosomal event known as a copy number variation (CNV) [231]. This sets up a gene-level copy number correlation

structure in an amplified/deleted chromosomal region. Conditional independence can easily demonstrated by conditioning such associations on the chromosomal event itself. As such, CNVs can be used alongside point mutations as anchor nodes.

5.1.1 Model Definitions

The biological process of transcription is required to code genes into proteins. In transcription regulation, a *trans-acting* element is a gene whose translation codes for a protein/RNA directly involved in the transcription of another gene [232]. *Cis-acting* elements are non-coding DNA regions that regulate the transcription of an adjacent or closely proximal gene, typically by acting as transcription factor binding sites [233]. As such, genomic alterations to trans/cis-acting DNA sites can perturb gene expression, examples being loss/gain-of-function (LOF/GOF) mutations or amplification/deletion of proteins/binding sites [234]. As such, a hierarchical perturbation event starting at the genomic level and affecting the transcriptomic level is a biologically sound hypothesis. We can say that the anchor mutation M is causal for a gene expression fluctuation δg

$$M \rightarrow \delta g \tag{5.1}$$

for an arbitrary gene g in the set of all genes $g \in G$. Cancer cells are not the only players in the tumour microenvironment, and δg could represent expression changes brought about by varying levels of immune cell infiltration and functionality.

$$\delta I_p \rightarrow \delta g \tag{5.2}$$

where δI_p is a metric quantifying the change in representation or activity of a leukocyte population p. Eq 5.1 can lead to dysregulated cancer-immune signalling or antigen presentation, giving rise to the reciprocal of Eq 5.2:

$$\delta g \rightarrow \delta I_p. \tag{5.3}$$

The immune system *responds* to cancer, there is no known biological function by which immune cells can mutate DNA and cause non-hematopoietic cancer. The immune system either responds to the signalling of stressed cells or recognises neo-antigens on the cancer cell surface [60]. Therefore, δI_p can be anchored on M

$$M \rightarrow \delta I_p. \tag{5.4}$$

Stemming from this formulation are three fundamental hypotheses which are summarised below.

1. The gene expression change δg is *causal* for δI_p, conditioned on M. We illustrate it in the following DAG model

$$M \rightarrow \delta g \rightarrow \delta I_p. \tag{5.5}$$

   Using conditional independence statements, we can factorise the joint distribution over Eq 5.5: $P(M, \delta g, \delta I_p) = P(M)P(\delta g|M)P(\delta I_p|\delta g)$.

2. The gene expression change δg is *reactive* to δI_p conditioned on M. We illustrate it in the following DAG model

$$M \rightarrow \delta I_p \rightarrow \delta g \tag{5.6}$$

   Using conditional independence statements, we can factorise the joint distribution over Eq 5.6: $P(M, \delta g, \delta I_p) = P(M)P(\delta I_p|M)P(\delta g|\delta I_p)$ describes this DAG.

3. The gene expression change δg and δI_p are *independent* given M. We illustrate it in the following DAG model

$$\delta I_p \leftarrow M \rightarrow \delta g_i \tag{5.7}$$

   Using conditional independence statements, we can factorise the joint distribution over Eq 5.7: $P(M, \delta g, \delta I_p) = P(M)P(\delta g|M)P(\delta I_p|\delta g, M)$ describes this DAG.

The probability density functions of 1,2 and 3 follow from standard Markov assumptions. The joint distribution at each node factors into at most, 2 two-dimensional Gaussian functions, placing an upper bound on the number of parameters and greatly reducing overfitting and complex model selection bias. These models are hierarchical in the sense that events start at the genome, propagate up the transcriptome and manifest at the TME level. We can exploit this to write simple likelihood functions based on regression models. To do this, we must first formally introduce the mathematical concept of bivariate normal regression.

5.1.2 Bivariate Normal Regression

Conditional probabilities of random variables allow us to introduce the notion of *linear regression*.

Definition Let X_1 and X_2 be two normally-distributed variables such that their conditional distribution function $f(X_2|X_1)$ is a Gaussian. It follows that the conditional expectation of X_2 on X_1 can be written

$$E(X_2|X_1) = \alpha X_1 + \beta \tag{5.8}$$

where α and β are constants to be found. 5.8 is termed the *linear regression* of X_2 on X_1.

The exact form of 5.8 can be found by explicitly writing out the marginal distribution $f(X_2|X_1)$. The joint distribution of X_1 and X_2 is written

$$f(X_1,X_2) = \frac{1}{2\pi\sigma_{X_1}\sigma_{X_2}\sqrt{1-\rho^2}}\exp[-\frac{(\frac{x_1-\mu_{x_1}}{\sigma_{X_1}})^2 - 2\rho(\frac{x_1-\mu_{x_1}}{\sigma_{X_1}})(\frac{x_2-\mu_{x_2}}{\sigma_{X_2}}) + (\frac{x_2-\mu_{x_2}}{\sigma_{X_2}})^2}{2(1-\rho^2)}]. \tag{5.9}$$

Considering the relation $f(X_2|X_1) = \frac{f(X_1,X_2)}{f(X_1)}$, the marginal distribution can be written explicitly by dividing Eq. 5.9 by the X_1's distribution function, giving us

$$f(X_2|X_1) = \frac{1}{\sqrt{2\pi\sigma_{X_2}^2(1-\rho^2)}}\exp[-\frac{[x_2-\mu_{X_2}-\rho(\frac{\sigma_{X_1}}{\sigma_{X_2}})(x_1-\mu_{X1})]^2}{2\sigma_{X_2}^2(1-\rho^2)}]. \tag{5.10}$$

The expectation value of Eqn. 5.10 gives us the exact form of Eqn. 3.28

$$E(X_2|X_1 = x_1) = \mu_{X_2} + \rho(\frac{\sigma_{X_1}}{\sigma_{X_2}})(x_1-\mu_{X1}). \tag{5.11}$$

which is the exact form of the linear regression of X_2 on X_1.

5.1.3 Likelihood Function Definitions

We are now in a position where we can define the joint probabilities at each model node. To construct the marginal distribution components, we first define regression functions to be

used as expectation values. We can model gene expression changes and immune variation as functions of an underlying mutation

$$g_i = \mu_{g_i} + \alpha_g f(M) + \varepsilon_g \tag{5.12}$$

$$c_i = \mu_c + \alpha_c f(M) + \varepsilon_c \tag{5.13}$$

where g_i and c_i are the respective gene expression and immune trait measurements for sample i, α and β are regression parameters to be learned, and the residual ε is normally distributed with mean 0 and variance σ_g or σ_c respectively.

f is a function defined to map the genotype probability of M onto a suitable domain. In the original parameterisation by Falconer for SNP data [235], f is a signed indicator function sampling values from $\{-1,0,1\}$ depending on the genotype of M. This works well in the context of SNPs since there is no linear dependency structure between acquired mutations and expression; you cannot have more than one point mutation in the same DNA position. By contrast, CNV events replicate entire regions of the genome, setting up a correlation structure with gene expression. To account for this correlation structure, f is defined as a weighted signed function, effectively normalising the CNV signal.

$$f(L_M) = \frac{L_M - \mu_{L_M}}{\sigma_{L_M}} \tag{5.14}$$

is a simple z-score transform with L_M representing the continuous copy number signal of mutation M. A high positive value of $f(M)$ provides confidence that M is amplified, whereas a low negative value provides confidence that M is deleted. These regression models also account for the probability of M being normal since in that case, $f(M) \to 0$.

Using the parameterisation of Eqns 5.12 and 5.13, we can begin outlining the conditional expectation definitions needed to write out the joint probability distribution components of these models.

$$P(g|M) = \frac{1}{\sqrt{2\pi}\sigma_g} \exp\left(-\frac{(g_i - \mu_{g_M})^2}{2\sigma_g^2}\right),$$

$$P(c|M) = \frac{1}{\sqrt{2\pi}\sigma_c} \exp\left(-\frac{(c_i - \mu_{c_M})^2}{2\sigma_c^2}\right),$$

$$P(g|c) = \frac{1}{\sqrt{2\pi\sigma_g(1-\rho^2)}} \exp\left[\frac{\left(g_i - \mu_g - \rho\frac{\sigma_g}{\sigma_c}(c_i - \mu_c)\right)^2}{2\sigma_g^2(1-\rho^2)}\right],$$

$$P(c|g) = \frac{1}{\sqrt{2\pi\sigma_c(1-\rho^2)}} \exp\left[\frac{\left(c_i - \mu_c - \rho\frac{\sigma_c}{\sigma_g}(g_i - \mu_g)\right)^2}{2\sigma_c^2(1-\rho^2)}\right],$$

$$P(c|g,M) = \frac{1}{\sqrt{2\pi\sigma_c(1-\rho^2)}} \exp\left[\frac{\left(c_i - \mu_{c_M} - \rho\frac{\sigma_c}{\sigma_g}(g_i - \mu_g)\right)^2}{2\sigma_c^2(1-\rho^2)}\right],$$

$$P(g|c,M) = \frac{1}{\sqrt{2\pi\sigma_g(1-\rho^2)}} \exp\left[\frac{\left(g_i - \mu_{g_M} - \rho\frac{\sigma_g}{\sigma_c}(c_i - \mu_c)\right)^2}{2\sigma_g^2(1-\rho^2)}\right],$$

where $\sigma_g, \sigma_c, \mu_g, \mu_c$ and ρ are parameters to be learned. μ_{g_M} and μ_{c_M} come directly from the parameterisations of the regression formulae Eqns 5.12 and 5.13, and represent the mean given the mutation genotype at M. For example, in the case of CNVs

$$\mu_{g_M} = \begin{cases} \mu + \alpha f(L_M), & M = amplified \\ \mu - \alpha f(L_M), & M = deleted \\ \mu & M = normal \end{cases} \tag{5.15}$$

is an analogue of the Falconer parameterisation for single point mutations [235]. For each possible genotype, we assume the distribution around the mutation locus M to be normal

with a constant variance. These components multiply to form the probability density function around each DAG, which we will label henceforth as M_1 (Eqn. 5.5), M_2 (Eqn 5.6) and M_3 (Eqn. 5.7). The model likelihood functions are computed as the product of the sum of sample-wise model likelihood probabilities over all genotypes $j \in \{Amplified, Deleted, Normal\}$ for all samples $i \in N$.

$$L(\theta_{M1}; M1) = \prod_{i}^{N} \sum_{j} P(m_j) P(g_i|m_j) P(c_i|g_i), \tag{5.16}$$

$$L(\theta_{M2}; M2) = \prod_{i}^{N} \sum_{j} P(m_j) P(c_i|m_j) P(g_i|c_i), \tag{5.17}$$

$$L(\theta_{M3}; M3) = \prod_{i}^{N} \sum_{j} P(m_j) P(g_i|m_j) P(c_i|g_i, m_j), \tag{5.18}$$

where $\{\theta_{M1}, \theta_{M2}, \theta_{M3}\}$ are parameter vectors taken as the union set of the individual component parameters composing each respective model.

Now that we have established a model framework aimed at addressing benchmarking issues outlined in section 4.4.4, we can begin addressing issues that may cause this approach to fail. These are enumerated below, and solutions discussed in the following subsections.

1. How do we control for correlation structures arising from gene expression co-regulation?
2. How do we filter interesting gene-specific CNV associations from false positives that happen to be co-amplified/deleted on the same chromosomal arm?
3. What is the optimal approach to model selection?

5.1.4 Transcription Factor Network

Genes are encoded to mRNAs through the transcription process. This process is facilitated by proteins called transcription factors (TFs) that selectively bind to different areas of the genome. Basically, TFs bind to regions near a gene, appropriately termed transcription factor sites, and regulate the expression of a gene. One TF species can regulate many different genes, and union of these coregulated genes is known as a *regulon*. Much of the correlation structure in gene expression data can be attributed to groups of genes being coregulated by the same TF. As such, it is easier to establish links between TFs and other variables rather

than individual genes.

Establishing links between TFs and the immune system requires establishing a robust score for TF activity. Scoring TF activity is non-trivial since the word "activity" is vague and differs between studies. The gene expression readout of a TF is not enough to assay its viability as a metric of activity since it can be misleading. For example, the tumour suppressor Rb can mutate and consequently upregulate the TF $E2F1$, leading not to upregulated expression of $E2F1$ but that of its regulon [236]. One well established computational method for scoring TF activity tests the over/under-representation of regulon expression in a sample.

Activity measurement methods TF activity inference methods generally assume that the aggregate expression of a TF regulon can be used as a multiplexed proxy for the TF's activity [182, 236]. As such, TF activity inference can be performed computationally using gene expression data. Rhodes *et al.* propose computing TF activity as the overlap between a TF regulon and enriched cancer transcriptional signatures. Their method effectively quantifies the contribution of a TF to a disease, but the dependency on curated signatures can give misleading results when taken out of context, especially in cancer where intra-sample heterogeneity needs to be accounted for [236]. Their method also tells us nothing about TF activity on the level of individual samples.

Alvarez *et al.* propose a score based on the idea of regulons, using gene expression data to generate sample-wise statistics for TF activity. Their algorithm, virtual inference of protein activity by enriched regulon analysis (VIPER), also operates under the assumption that TF activity can be measured through regulon enrichment [182]. The idea here is to first build a TF regulatory network and then use it to compute scores for TF activity in gene expression samples. The TF regulatory network is constructed from gene expression data, using either Markov network reconstruction algorithms or more simple first order partial correlations approaches. The final network encodes information regarding each TF and its regulon. Regulon enrichment is computed using an analogue of single sample Gene Set Enrichment Analysis (ssGSEA) that corrects for pleiotropic effects. In this context, pleitropic effects refers to non-active TFs falsely appearing activated if their regulons overlap significantly with the regulon of an active TF. Their method corrects for false positives by penalising the proportion of the activity that comes from genes in the overlap.

VIPER has been validated extensively in CHiP-seq experiments that measure TF activity as a function of DNA binding. Furthermore, it has been used to infer TF activity in various

cancer transcriptomic studies, including breast cancer [172]. In summary, VIPER produces the sample-wise coregulation scores we need to overcome the first limitation outlined at the end of section 5.1.3.

5.1.5 Protein Interactome Structural Prior

Since whole chromosomal arms are prone to amplification/deletion, events at the gene level are not necessarily mutually exclusive events. Therefore, the copy number of adjacent genes in a CNV region may be correlated and finding associations between TF activity and gene-level copy number is complicated by false positives. Attributing the variance in a phenotype to the copy number profile of a single gene is very challenging using comparative studies alone. The proposed solution to this builds on the premise that TFs can be up or down regulated by trans-acting modulator proteins [236, 237, 238]. The idea is that false positives can be minimised by limiting inference to genes that code for proteins *already known* to interact with a TF of interest. This places a direct prior over the domain of significant associations between CNV events and TF activity.

In section 4.4.2, we described how network structure priors can be used to narrow the model search space and improve statistical confidence in new inference. Protein interaction networks are excellent candidates as structure priors, since they implicitly encode interactions between TFs and all other intercellular proteins. In the following paragraphs, we briefly describe several well known approaches to building protein interaction networks.

Proteomics Data In section 1.3.1 we overviewed the mass spectrometry (MS) and reverse-phase protein array (RPPA) approaches for quantitatively profiling protein abundance from tissue samples. Probabilistic methods have been tremendously successful in reconstructing protein-protein interaction networks from these data. Breitkreutz *et al.* use a Bayesian mixture modelling approach to estimate protein interaction likelihoods in yeast [239]. In human data, Sowa *et al.* assign deterministic scores to interactions based on both the frequency and uniqueness and co-occurrence of protein pairs to build networks [240]. Unfortunately, validating protein interaction networks generated this way is laborious, requiring extensive *in vitro* knock-downs of network nodes and cross-referencing networks against the resulting perturbations [241].

Protein Interaction Databases Dedicated experimental data repositories annotate protein interaction both at the level of pairwise interactions and in the context of molecular pathways. Experimental data are distributed amongst these repositories, detailing hundreds of

millions of protein interactions across thousands of organisms [242, 243, 244]. A number of text-mining approaches have been proposed to extract protein-protein interactions from these databases including GeneMANIA [245], VisANT [246] and STRING [247]. These algorithms typically use word co-occurrence scoring to test if two proteins appear in the same sentence, paragraph or whole documents more often by chance.

STRING has over 200 million interactions across 5 million proteins and over 1000 organisms, holding both experimentally verified and predicted interactions [247]. Protein-protein interaction networks (PPI) built from these resources are designed to be experimentally verified from the ground up, making them excellent candidates for structural priors. Since STRING's PPI is an ongoing project, it does not yet include every experimentally verified interaction between the entire human proteome. Nonetheless, networks built from experimental data provide more confident statements regarding protein interaction than those built from probabilistic models. This network asserts a uniform prior over the search space, thus narrowing the number of models in the selection phase of DAG learning and reducing the false positive rates. I use STRING's PPI to address the second challenge outlined at the end of section 5.1.3.

5.1.6 Model Search Strategy

To address the final challenge outlined in section 5.1.3, we turn to our primer on model selection (section 4.4.3), where I describe a three component strategy consisting of a search space, a scoring metric and a search strategy.

Search Space Let $P = (V, E)$ be a protein-protein interaction network where V corresponds to protein identifiers and E represents the dependency relation between pairs of vertices. Let $\mathbf{T}$ be a set of transcription factors such that $\mathbf{T} \subseteq V$ and denote $\mathbf{n}_t$ as the neighborhood of any $t \in \mathbf{T}$ in P. Let I denote the lymphocyte infiltration metric, as computed from image data, FACS, transcriptomic data or otherwise. The aim is to see which hypothesis (represented by models $M1$, $M2$ and $M3$) is most likely to describe the dependency structure between the activity at $t \in \mathbf{T}$, copy number aberrations at $\mathbf{n}_t$ and I. Since this approach is hypothesis-driven, the search space is always populated by exactly three models:

$$M1 \qquad CN_{g\in\mathbf{n}_t} \rightarrow Activity_t \rightarrow I \tag{5.19}$$

$$M2 \qquad CN_{g\in\mathbf{n}_t} \rightarrow I \rightarrow Activity_t \tag{5.20}$$

$$M3 \qquad I \leftarrow CN_{g\in\mathbf{n}_t} \rightarrow Activity_t \tag{5.21}$$

where $CN_{g\in\mathbf{n}_t}$ is a copy number event at g in the neighbourhood $\mathbf{n}_t$ and $Activity_t$ is the activity of TF t. For each $g \in \mathbf{n}_t$, search space is created and populated by Eqs. 5.20, 5.21 and 5.22 iff $CN_{g\in\mathbf{n}_t}$, $Activity_t$ and I form a complete association graph after multiple correction testing, (all three vertices must be interconnected). This demonstrates a significant degree of linear dependence amongst the variables, an essential requirement for subsequently establishing conditional independence. Association significance is measured using Student's T test with Benjamini-Hochberg FDR correction.

Scoring method The likelihood functions of $M1, M2$ and $M3$ are given by Eqs. 5.17, 5.18 and 5.19 respectively. The mean and variance of the absolute number of free parameters across the search space are small, and so complex model selection bias is unlikely to be an issue here. Overfitting will not pose as big an issue in these models as with more complex likelihood functions. Although Bayesian scoring would eliminate overfitting altogether, MLE is a more straightforward solution and less computationally expensive than computing posterior distributions across thousands of search spaces. This structural prior implicitly controls for false positive discovery and false negatives still remain to be controlled for. The AIC is advantageous here when penalising Eqs. 5.17, 5.18 and 5.19, as it works well when false negative results are more misleading.

Search method Given a PPI $P = (V,E)$, and a TF $t \in (T)$ where $\mathbf{T} \subseteq V$, let $\mathbf{n}_t$ be the neighborhood of t in P. For each protein $g \in \mathbf{n}_t$, this algorithm tests if a complete association graph can be constructed from the copy number profile $CN_{g\in\mathbf{n}_t}$, the transcription factor activity $Activity_t$ and a metric for immune infiltration I. If so, t and g are fed into the search method, where for each model, the likelihood function is maximised over the data and the parameter space using MLE. MLE returns the negative log likelihood (NLL) of each model and populates the search space with the AIC of each model. The method then uses insertion sort to find the minimum value. Given a search space $\mathbf{S}$ where $|\mathbf{S}| = 3$, it is easy to see that insertion sort will find the smallest AIC given at worst, 9 tries. This algorithm is illustrated

using the following pseudocode.

Data: $\{\mathbf{PPI}, t, I\}$
Result: most likely models, if any.
n = neighbours(t, **PPI**);
for *g in* ***n*** **do**
    D = {CN(*g*), Activity(*t*), *I*};
    g = correlationGraph(**D**);
    if *g.isComplete* **then**
        NLL, nParams = {MLE(M1,**D**), MLE(M2,**D**), MLE(M3,**D**)};
        AIC = -2 * **nParams** + 2 * **NLL**;
        return minimum(**AIC**);
    else
        return NULL;
    end
end

Algorithm 1: Pseudocode describing the search method

5.2 Quantifying Anti-Tumoural T-Cell Response

5.2.1 Motivation

Cytotoxic T-cells are significant players in the anti-tumoural adaptive immune response. CD8+ T cell infiltration, as measured absolutely or relative to the overall tumour mass, is often associated with positive prognosis in a number of cancer types including breast cancer [248]. Novel immunotherapies focus on increasing the efficacy of this response by down-regulating immunosupressive factors, such as checkpoint blockade or signaling molecule therapies [249]. However, progress in this field is hindered by an incomplete understanding of suppressive factors in the tumour microenvironment. We have yet to characterise many of the mechanisms that cancer leverages to escape CD8+ T cell mediated destruction.

Motivated by this, this project aims to quantify T cell infiltration for use in the causal inference framework. The work presented here is the first of its kind to place T cells in a causal context with the cancer cell autonomous compartment of the tumour microenvironment. In section 1.4 several methods were introduced for measuring the abundance of a specific cell type from a sample, including IHC, flow cytometry and gene expression signatures. Gene expression signals are robust metrics of lymphocyte infiltration, but not as accurate

as cell-counting experiments. Large online multimodal datasets will not often have IHC and flow cytometry measurements for T cells; especially not with matched copy number and transcriptomic data per-sample. H&E stained images, however, are readily available with these datasets, since biopsies are routinely produced for diagnostic purposes. These images can be mined for lymphocyte statistics using computational approaches as described in section 2.4.

5.2.2 Gene Expression Signal

T cells stem from a specific hematopoietic lineage of thymocytes and express unique membrane-bound markers shared by no other family of cells in the tumour. As such, computational methods make it possible to elucidate a T cell "signal" from the bulk tumour transcriptome using either a panel of T-cell specific markers or a more general T-cell signature. Markers and signatures differ in that the latter admits non-specific T-cell genes whilst the former contains only T-cell genes. Their utility depends on the approach being used, for example, gene expression deconvolution approaches work better with larger gene sets. However, deconvolution approaches are designed to resolve admixtures of cell populations, whereas this metric aims to estimate the infiltration of a single family of thymoctic lineage cells. Therefore an independently developed marker-based approach termed the "T Cell Score" (TCS) was developed. This subsection will be split into two parts; firstly defining the TCS mathematically and secondly validating it on ground truth examples.

Geometric Average of Markers Inspired by a similar approach for cytolytic activity [69], a relative score for T cell infiltration was defined that is applicable to both RNA-seq and microarray data. Given a panel of m T cell markers $\mathbf{N} = \{N_1, N_2, ..., N_m\}$, the overall T cell representation in sample i is given by the geometric mean of the expression vector $\mathbf{G}^i_N$:

$$TCS_i = \sqrt[m]{G^i_{N_1} \times G^i_{N_2} ... \times G^i_{N_m}} \tag{5.22}$$

The geometric mean is chosen over arithmetic mean since it is less sensitive to outliers.

Choosing Marker Genes Marker gene selection aims to find a subset of genes that can maximally separate a cell population from the remaining transcriptome. T cells are categorised by specific "clusters of differentiation", or immunophenotyping markers, such as the

CD3, CD8, CD4 receptors coded for by the genes *CD3D*, *CD8A*, *CD8B*, *CD4*. These genes are expressed primarily in cells of the thymocyte lineage with little noise contamination from the rest of the tumour.

Validation of TCS Typical approaches to validating gene expression scores involve comparison with a ground truth metric, typically generated using an orthogonal dataset. To facilitate this, a paired gene expression and flow cytometry blood sample dataset was downloaded that examines 9 leukocytes subsets including CD8+ T cells (see section 5.3 Data and Preprocessing). Following this, the TCS was computed for each sample, correlations between FACs evaluated using Pearson's test and significance tested using Student's T test. No positive significant correlations were observed for any leukocyte subset other than CD8+ T cells ($\rho = 0.675$, $P = 0.001$). This validation strongly corroborates the TCS is a robust predictor of CD8+ T cell infiltration in gene expression data.

5.2.3 H&E Images

H&E images illustrate a snapshot of the morphological state of a tumour region prior to biopsy/ resection. They encode tissue architecture features such as cancer cell, immune cell and stromal cell spatial distributions. The admixture is more linearly separable at the morphological level than in typical bulk tumour molecular assays; trained pathologists routinely use H&E features for diagnostic purposes. Unlike molecular assays containing tens of thousands of features (genes), H&E image features are limited to colour, shape and texture descriptors, severely constraining the resolution we can get with cell phenotyping. On the other hand, pathologist assessment of lymphocyte infiltration is more direct and relies on fewer statistical assumptions [67]. Since H&E images are in theory, as abundant as patient gene expression profiles, they are an excellent orthogonal measure for lymphocyte infiltration. Computational approaches to image segmentation and object classification have produced large quantitative H&E feature datasets [71, 67]. METABRIC is a notable breast cancer dataset with matched transcriptomic, copy number profiles and images per sample, making it an excellent candidate for this study [76].

5.3 Data and Preprocessing

This will briefly overview the sources of data and preprocessing protocols used in this study, such that the results of the analysis can be reproduced more easily.

Gene Expression Data Microarray transcriptomic profiles corresponding to 1980 patients from the METABRIC [76] cohort were downloaded from the European Genome-Phenome archive under the accession id: EGAD00010000268. The issue of multiple probes mapping to the same gene was addressed by selecting the probe with the highest variance. RNA-seq count data comprising 1154 BRCA samples was downloaded from the TCGA archive [201] (https://tcga-data.nci.nih.gov/tcga) and processed using a two-step process: applying the variance stabilising transform and quantile normalising the matrix with respect to the METABRIC gene expression distribution. This was done to correct for the large heteroscedasticity between genes and make the expression distributions more comparable.

DNA copy number aberrations METABRIC Affymetrix SNP 6.0 data were downloaded from the same resource as the transcriptomic data. SNP array genomic positions were mapped to gene symbols using the hg18 build. TCGA GISTIC2 [250] gene-level normalised copy number calls for each patient were accessed from GDAC Firehose (http://gdac.broadinstitute.org/).

Protein-protein interactions A network detailing protein-protein interactions was downloaded from the STRINGdb resource (http://string-db.org/) and ENSEMBL identifiers were mapped to HUGO gene symbols using the R biomaRt package.

Transcripton Factor Network Enrichment To infer TF activity, a coexpression network for 788 experimentally verified TFs derived using mutual information was used [172] to calculate the activity of each TF for each sample using the R package viper (virtual inference of protein activity by enriched regulon analysis) [182] using default parameters.

T Cell Score Validation using Flow Cytometry Gene expression profiles for 20 peripheral blood mononuclear cell admixture samples and their corresponding flow cytometry profiles as measured by Newman *et al* [85] were downloaded from (http://cibersort.stanford.edu).

H&E Section Data This project use of the image dataset published by Yuan and colleagues [67], comprising the segmented H&E stained primary tissue sections of 564 patients sampled from the METABRIC cohort. Segmented objects were classified using a SVM trained by an expert pathologist and metrics pertaining to the absolute number of lymphocytes and the lymphocyte density relative to the number of overall objects were measured. The absolute number of lymphocytes were then log-transformed to enable more robust comparison across the patient space. Furthermore, this transformation ensures input to the MLE process exhibits a similar range of values and that the algorithm can be initialised with the same parameters for all cohorts.

5.4 Results

The aim of this project was to establish a framework that formally describes the regulatory structure between somatic genomic events, signaling pathway activity and immune activity in the tumour. This approach is implemented in the statistical environment R [251] and all code to reproduce the results presented here is available as part of an annotated document hosted on the bioRxiv [2] repository - https://doi.org/10.1101/144832 .

Genomic events were measured using the copy number profiles of 19,702 genes, as provided by the METABRIC [76] and TCGA projects [201]. To measure signaling pathway activity, a breast cancer regulatory network [172] for 788 experimentally verified TFs [252] was used as input alongside the TCGA/METABRIC transcriptional profiles to VIPER [182], a method for network based prediction of transcription factor activity.

To measure immune activity, two orthogonal approaches were used: the first approach uses the mean expression of marker genes to define a cytolytic score (CS) [69] and a novel T-cell score (TCS). While the CS trait is a measure of lymphocyte activity, the TCS measures the degree to which they are represented in the tissue. The second approach uses paired H&E images from the METABRIC cohort to measure the absolute number of lymphocytes and their density per tumour [67].

5.4.1 A multi-step causal inference approach to assign directionality to signaling-immune associations

The proposed framework to assign directionality to signaling-immune associations is outlined in section 5.1. Different causal relationships are formalised using three different DAGs (Fig. 5.1**A**). Model 1 (M_1: the causative model) represents a case in which a genomic event changes immune activity by dysregulating signaling activity. Model 2 (M_2: the reactive model) represents a case in which a genomic event leads to a change in immune activity, which then in turn perturbs signalling activity. Model 3 (M3: the independence model) represents a case in which the genomic event influences immune activity and signaling activity independently of each other.

Likelihood functions for each of the three models are derived in section 5.1.3 using standard assumptions of causal inference [200]. To limit the model search space and reduce the number of triplets to test, a multi-step causal model inference framework (CMIF) approach

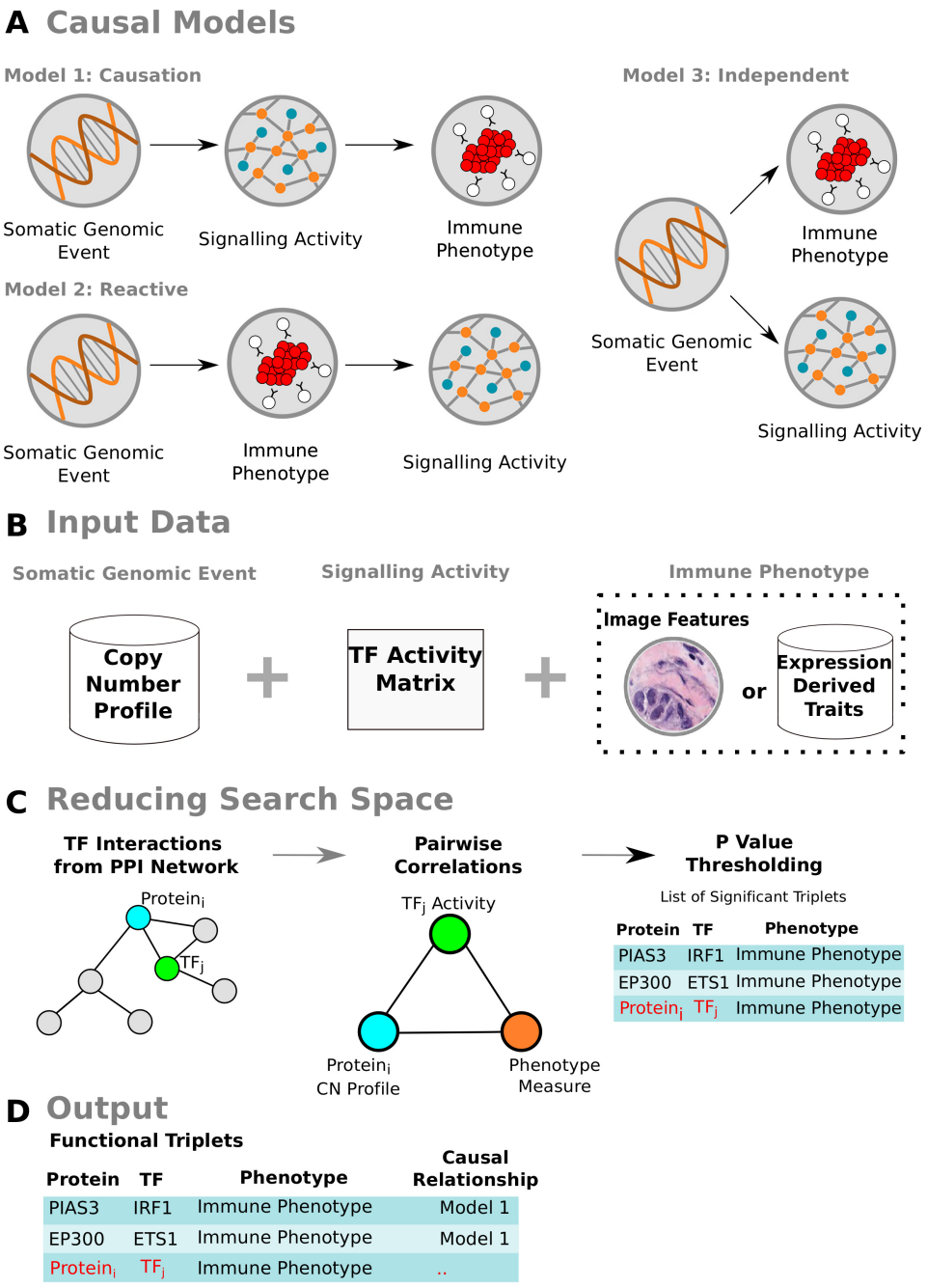


Fig. 5.1 **Description of CMIF A** Directed Acyclic Graphs (DAGs) representing each respective hypothesis evaluated during the analysis. **B** The inputs for CMIF are a matrix of TF activities per sample as measured by VIPER, continuous intensity profiles for the copy number calls and phenotype data that can be either image features or features derived from gene expression data. **C** Using a uniform network prior over interactions between TFs and other proteins, a skeletal association graph is computed between the TF activity, the CN profile of an interacting protein and the immune phenotype. For significant triplets, pre-defined likelihood functions for each model in **A** are maximised over their parameters using maximum likelihood estimation, with the model most likely to be supported by the data determined by the model with the smallest Akaike Information Criterion (AIC). **D** The output of CMIF is a table of functional triplets with their corresponding model classification. This figure has been reproduced from my preprint Chlon *et al.* [2]

was developed as illustrated in Fig. 5.1.

In a first step, CMIF selects genes with an experimentally verified protein-protein interaction (PPI) with any TF of interest (Fig. 5**C**). There are $19,702 \times 788 = 15,525,176$ pairwise associations between copy number profiles and TF activities, and filtering them according to the PPI network from the StringDB database [253] results in just 2,333 potential models. This filtering substantially reduces the search space and enriches for biologically relevant drivers in groups of correlated genes that are jointly amplified or deleted. In a second step, undirected skeletal association graphs are constructed for CNA events underlying both the TF activity and immune phenotype by computing pairwise correlation coefficients between the variables. Benjamini-Hochberg multiple hypothesis testing correction [208] is applied to each p-value, and only complete skeletal graphs are passed to the final step (Fig. 5.1**C**). Finally, the likelihood function of each model is maximised over the parameter space, and the model with the smallest Akaike Information Criterion (AIC) is chosen. After filtering and model selection, CMIF provides as output the regulatory structure between the CNA event, TF activity and lymphocyte phenotype. (Fig. 5.1**D**).

5.4.2 Evaluating CMIF with CS/TCS immune metrics

In the first analysis, the TCS and CS metrics derived from gene expression data were used as phenotypes of immune activity. Applying CMIF to the TCS/CS metrics, TF activities and copy number profiles produced 475 complete skeletal graphs characterised by 111 unique TFs. These models were anchored by CNA events at loci corresponding to 176 unique genes. Each complete skeletal graph was admitted to the next stage of analysis, where the likelihood models M1, M2 and M3 were fit to the data using maximum likelihood estimation.

The CMIF revealed that 344 triplets (72%) were best represented by the causal model (M1) whereas the reactive model (M2) best explained the regulatory structure in the remaining 131 (28%) cases. Strikingly, no M3 models were supported by the data, highlighting the efficiency of the PPI prior (Fig. 5.1**C**) in filtering independent regulatory structures.

Validation in independent cohort The 475 proposed models were validated independently in the TCGA cohort. Of the candidate triplets, 194 (54.6%) M1 and 24 (18.3%) M2 models were reproduced in the TCGA cohort using CMIF (Fig. 5.2**A**). The higher validation percentage of M1 models over M2 models in TCGA data suggests that causal drivers of T cell infiltration are more robust and thus more frequently recapitulated in breast cancer populations. The higher prevalence of M1 over M2 models might be explained by cancer cells being immunoedited [60], a process by which somatic mutations break downstream

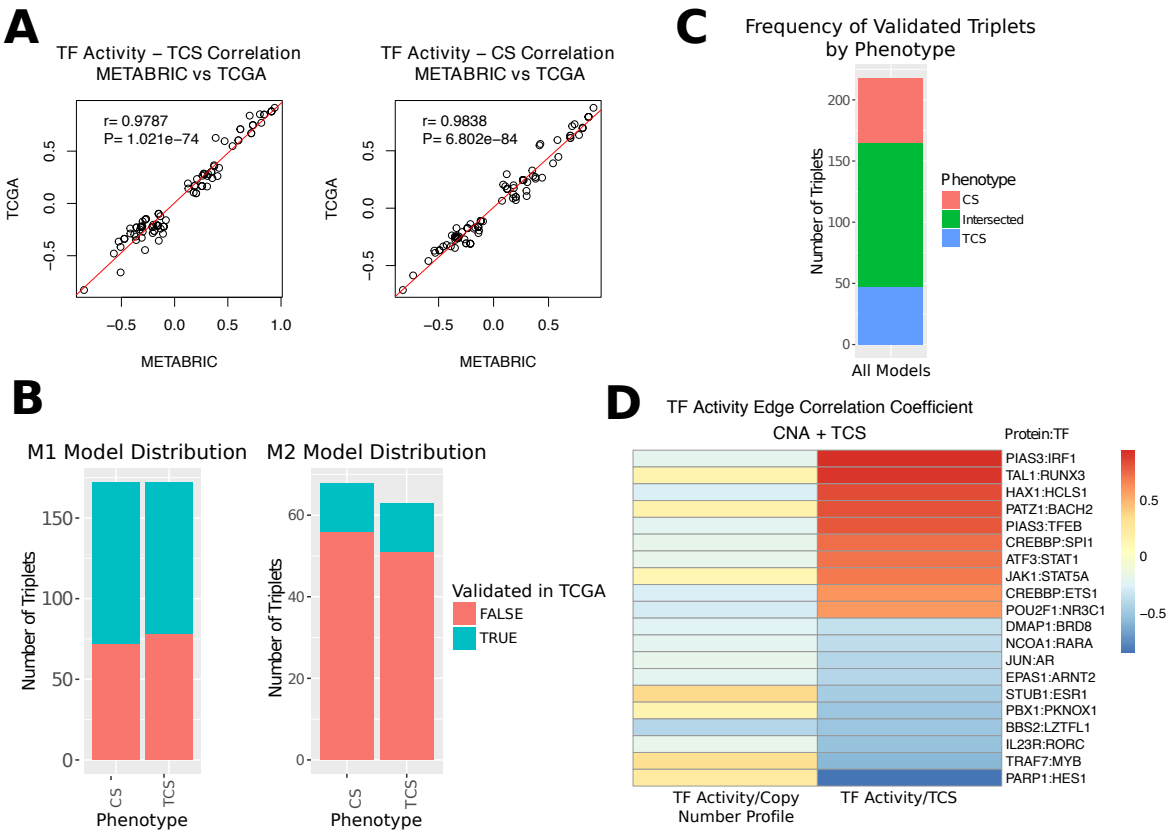


Fig. 5.2 **Model validation in TCGA A.** Scatter plots with fitted regression lines illustrating strong concordance between METABRIC and TCGA when transcription factor activity significantly explains the variance in the TCS and CS immune traits. **B.** The proportion of predicted relationships in METABRIC that validate in TCGA as stratified by model type and lymphocyte trait. **C.** Stacked barchart illustrating the frequency and overlap of models between the different immune traits. **D.** Top and bottom 10 TCGA-validated causal models as ranked by the proportion of TCS variance explained by the transcription factor activity. Y-axis indexing is organised as (Gene at locus of CNA event): (Transcription Factor). Heatmap columns illustrate Pearson's correlation coefficient between the CNA signal and transcription factor activity, and transcription factor activity and TCS measurement (left to right). This figure has been reproduced from my preprint Chlon *et al.* [2]

pathways associated with a normal immune response. Over time, this would enable the tumour to exert more control over the immune system than vice-versa.

The correlation between TF activity and the individual immune traits was well conserved between METABRIC and TCGA (TCS: $\rho = 0.98$, $P < 2.2 \times 10^{-16}$; CS: $\rho = 0.984$, $P < 2.2 \times 10^{-16}$) (Fig. 5.2B). As such, these associations illuminate robust co-dependence relationships between lymphocyte infiltration/activation and TF activity. Of the validated models, 118

were shared between the TCS and CS traits, with 47 unique to the TCS (165 total) and 53 unique to the CS (171 total). (Fig. 5.2**C.**). This high degree of concordance is reassuring considering that lymphocyte recruitment and cytolytic activity are complemetary systems, underpinned by common co-regulators.

Validation by literature Many of the top predictions generated by CMIF are well supported by the literature. When ranked by correlation strength, *IRF*1 activity was highlighted as the most significant causal mediator of the TCS phenotype across both METABRIC and TCGA. CMIF also found that *IRF*1 activity is significantly down-regulated by the amplification of *PIAS*3 (Fig. 5.2**D**). These findings are consistent with reports that *PIAS*3 induces transcriptional repression of *IRF*1 by binding to it as a SUMO-1 ligase [254]. Furthermore, *IRF*1 has been shown to play a crucial role in driving anti-tumour immune response [255] and thus this model's categorisation as causal for TCS is well substantiated by the strong body of literature surrounding the relationship between the variables.

In another example, *RUNX*3, a well known tumour suppressor gene [256], was identified as the second strongest causal modulator of the TCS. This is consistent with reports that *RUNX*3 activity mediates lymphocyte chemotaxis through the TGF-*B* pathway [257]. A positive association found between *TAL*1 and *RUNX*3 has been also been confirmed in studies demonstrating that the *RUNX* genes are direct targets of *TAL*1 [258]. Additionally, CMIF identified *ETS*1 as a causal mediator for the TCS, which is unsurprising given that its activity has been shown to regulate the transcription of chemokines and cytokines directly involved in lymphocyte migration [259].

Interestingly, CMIF discovered mechanisms leveraging TFs associated with the downregulation of T cell proliferation. One notable example is *RORC*, whose activity is known to suppress the expression of *IL*2, a known T cell proliferation cytokine [260, 261]. Furthermore, *HES*1 activity was ranked as the most significant downregulator of the TCS. Extensive literature exists detailing the role of *HES*1 in T cell development and proliferation [262], but its influence over the immune component of TMEs in solid lesions is not fully understood. This observation warrants more extensive investigation in the context of lab-based experiments.

Validation with image-derived features Another way of validating the robustness of a proposed model is testing how well it predicts lymphocyte infiltration in an orthogonal dataset. To this end, this study incorporated paired tumour whole tissue section slides stained with

Haematoxylin and Eosin (H&E) from the METABRIC study [76], enabling an orthogonal estimation of lymphocytic infiltration independent of the gene expression based estimates used in the first analysis.

The list of 165 filtered TCS M1 models was tested for reproducibility in an image dataset consisting of 534 samples. This dataset had previously been processed by an automated pathology pipeline [67], producing sample-wise measurements of absolute lymphocyte count and the lymphocyte density. Further normalisation techniques were applied in this study to generate traits from these features as described in section 5.3. Next, image features, TF activities and copy number profiles were integrated into the CMIF approach to produce a image-specific list of models. The overlap between the image-based causal models and those from the transcriptomic phenotype set revealed that 18 (10.9%) of the initial predictors of the TCS were also predictive of the image features. Interestingly, the majority (15/18) of validated image models belonged to the lymphocyte density trait.

Surprisingly, the extent to which TF activity correlated with both the TCS and the image lymphocyte density was more poorly conserved ($\rho = 0.45$, $P < 2.2 \times 10^{-16}$) than that between the TCS between METABRIC and TCGA. This is not particularly surprising given that transcriptomic T cell features are not perfect proxies for lymphocyte features extracted from images. While the TCS makes measurements about T cells exclusively, H&E image resolution is is not sufficiently descriptive to differentiate T cells from other cells with a similar morphology. Furthermore, there are confounding systematic errors that may have arisen during the segmentation and classification process used to generate the image features that render the correlation with the transcriptomic phenotypes weaker than expected. For consistency, the overlapping model list was filtered for models where the association sign was conserved between TF activity correlations with the TCS/image phenotypes.

The image-validated model list was comprised of 12 triplets, revealing 11 unique genes exerting influence over lymphocyte infiltration by dysregulating the activity of 8 TFs (Table 1). Notably, 8 out of the top 10 strongest causal models for the TCS phenotype (as ranked by association with TF activity) validated for the image lymphocyte density trait, highlighting excellent conservation of predictive utility across orthogonal data types.

Of the validated models, notable examples include a process by which *PIAS*3 copy number amplification was found to attenuate the TCS/lymphocyte density by downregulating $TFEB$ and $IRF1$ activity, both of which positively associate with the mentioned traits (Table

CNA	TF	TF-TCS Correlation	TF-TCS P value	Model	Type
RBBP5	YEATS4	-0.10	0.00	1	TCS
CREBBP	ETS1	0.62	0.00	1	TCS
EP300	ETS1	0.62	0.00	1	TCS
PIAS3	IRF1	0.94	0.00	1	TCS
PIAS3	TFEB	0.80	0.00	1	TCS
POU2F1	NR3C1	0.60	0.00	1	TCS
PARP1	HES1	-0.85	0.00	1	TCS
NR5A2	NFYA	-0.08	0.00	1	TCS
PATZ1	BACH2	0.81	0.00	1	TCS
TAL1	RUNX3	0.91	0.00	1	TCS
CBFB	RUNX3	0.91	0.00	1	TCS
HAX1	HCLS1	0.84	0.00	1	TCS

Table 5.1 CMIF output of genomic drivers and TF perturbations causal for the TCS trait that also predict lymphocyte density in stained tumour sections.

5.1). *CREBBP* and *EP*300 were found to exert similar causal pressure on the traits through their action on the TF *ETS*1. This coincides with experimental evidence demonstrating that *CREBBP* and *EP*300 form a protein complex *CBP*/*p*300 that is recruited by *ETS*1 to facilitate its TF functionality [263].

*NR*3*C*1 was identified in the top 8 candidate TFs (Table 5.1): this gene encodes for the glucocorticoid receptor, and influences immune activity through inflammation [264]. This TF was ranked only 95th of 510 in the list of image associations and 27th of 510 in the TCS associations list. Its function as a consistent driver of immune infiltration was only elucidated once the causal relationships between genome, signaling and immune phenotypes were modeled together, highlighting the advantage of regulatory networks over standard association approaches.

5.4.3 Causal model case studies and mechanisms

These results provide several specific biological examples of causal models of the interaction between cancer signaling and immune activity.

EP300 and NCOR1 modulate cytolytic activity through ETS1/SPI1/TP53 network perturbation Copy number amplification of *EP*300 and *NCOR*1 were found to dysregulate the cytolytic activity trait in both the METABRIC and TCGA cohorts. Interestingly, the original study by Rooney and colleagues [69] found that single nucleotide variants in these genes

correlated positively with cytolytic activity in cancer types other than breast. The CMIF's ability to elucidate these mechanisms in breast cancer may be due the higher prevalence of CNA mutations over SNPs in the disease [230], thus providing our analysis with more statistical power. Furthermore, the CMIF extends our understanding of the association between the list of mutational drivers and cytolytic activity by suggesting they act by dysregulating the activity of $ETS1$, $SPI1$ and $TP53$.

The discovery of a positive association between $SPI1$ activity and the CS ($\rho = 0.7$) is consistent with studies demonstrating that $SPI1$ transcribes $CCL5$, a key player in cytolytic activity [265]. Similarly, $ETS1$ deletion in mice has been linked to decreased cytolytic activity in NK cells [266], consistent with the model's observed positive correlation ($\rho = 0.58$). The association between TP53 activity and cytolysis is poorly understood, although some findings have found associations between mutant TP53 and downregulated cytolytic activity in ovarian and other cancers [69]. Individually, the direct correlations between cytolytic activity and $EP300$ ($\rho = 0.141$) and $NCOR1$ ($\rho = 0.08$) are weak. However, inferring the regulatory structure with respect to TF activity enabled the CMIF to highlight $EP300$ and $NCOR1$ amplification as key drivers of cytolytic activity in breast cancer.

TF drivers of immune localisation regulate adaptive immune pathways Functional annotation of TF transcriptional targets can elucidate which molecular pathways are over- or underrepresented in the presence of an immune phenotype. To investigate this, the model set was partitioned according to the sign of the association between TF activity and lymphocyte infiltration. Across both partitions, the transcription targets of each TF were aggregated, and GO term enrichment was applied to functionally annotate the gene sets.

For TFs positively associated with lymphocyte recruitment, the terms "T cell activation" and "adaptive immune response" were among the top associated pathways (adjusted $P = 1.7 \times 10^{-33}$ and 2.4×10^{-28}) (Fig. 5.3**B**). Interestingly, "antigen processing and presentation" was also ranked highly on the list (adjusted $P = 1.3 \times 10^{-14}$), reinforcing the importance of comprehensive antigen recognition in facilitating lymphocyte recruitment. TFs negatively associated with lymphocyte recruitment displayed disjoint sets of target genes (Fig. 5.3**A(i)**) whose aggregate functional annotation was predominantly associated with pathways involved in innate immune cell regulation (Fig. 5.3**B**). In contrast, all positively associative TFs in the model space demonstrated overlapping regulons (Fig. 5.3**A(ii)**).

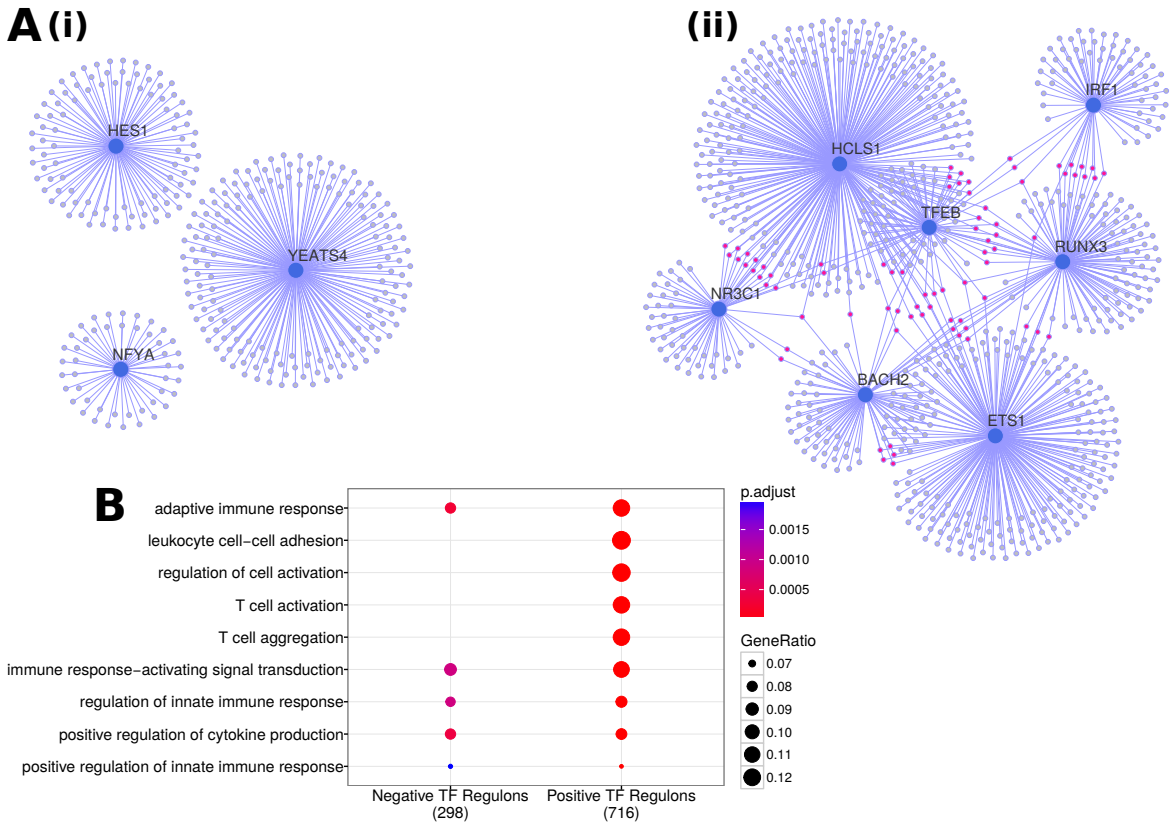


Fig. 5.3 **Analysis of TF targets A** Network visualisation of the inter-regulon overlap (illustrated through purple dots) between causal TFs that **(i)** down-regulate the TCS/lymphocyte density and **(ii)** those that up-regulate it. Evidently, TFs that causally up-regulate the T cell representation have a greater degree of regulon overlap whereas no intersection is observed for TFs that down-regulate the trait. **B** GO term enrichment analysis highlighting the most significantly annotated terms to the gene sets **A(i)** and **A(ii)** respectively. Whereas regulons pertaining to TFs positively associative with lymphocyte infiltration are more enriched for T-cell related pathways, down-regulators the phenotype are more associated with innate immune system pathways. This figure has been reproduced from my preprint Chlon *et al.* [2]

Systems driving lymphocyte recruitment stratify by ER status Differences in magnitude and prognostic relevance of lymphocytic infiltration between ER stratified breast cancer patients have been widely observed [267, 268, 198], but little is known regarding the causal chain of events that gives rise to this discrepancy. It was hypothesised that dysregulated TF activity in ER+ samples could provide a potential mechanism by which ER+ tumours evade immune destruction. To investigate this, the clinical dataset for the METABRIC cohort samples was used to infer whether genomic drivers of lymphocytic recruitment stratify by pathologist-assigned ER status.

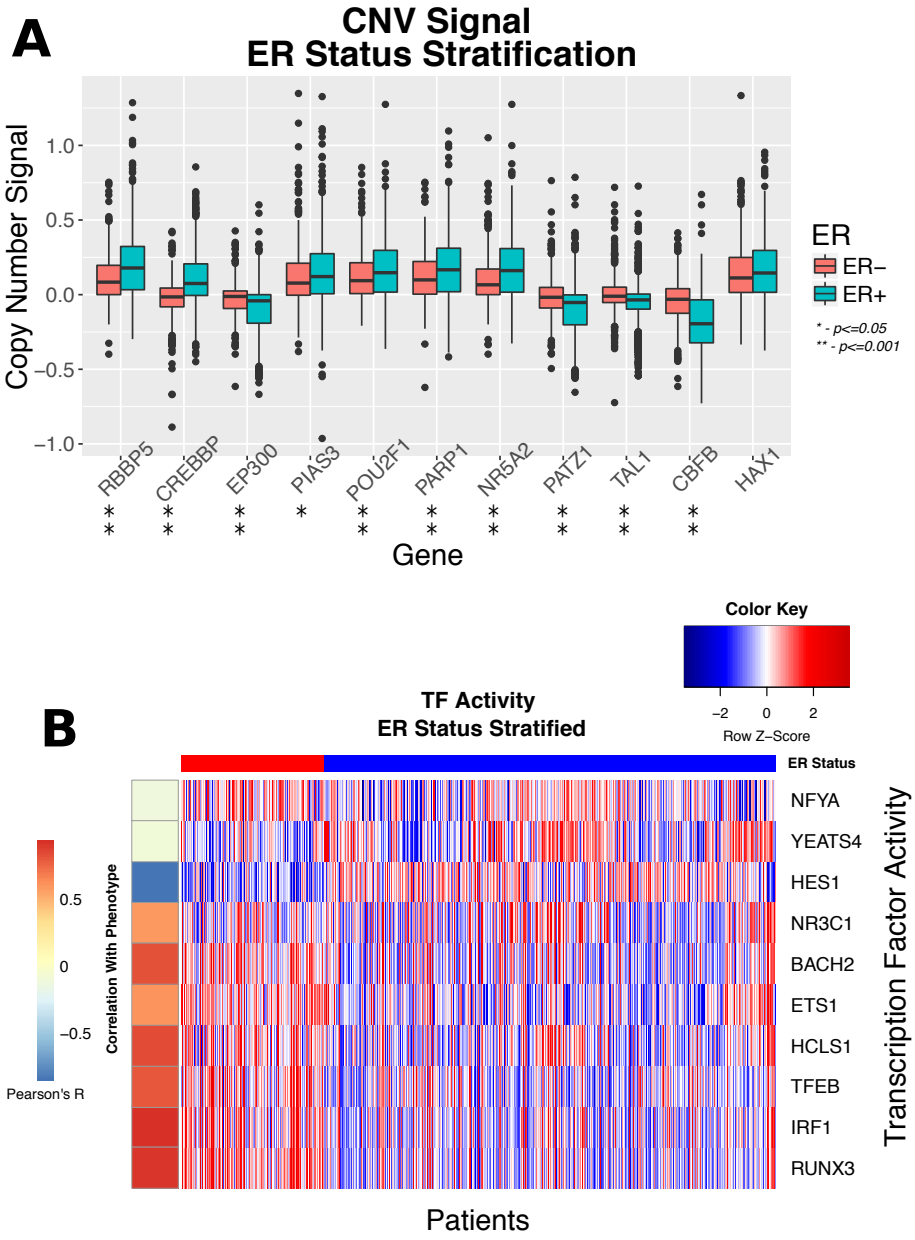


Fig. 5.4 **METABRIC ER Stratification of Causal Models a** Boxplots highlighting the difference in the normalised DNA copy number signal between ER+/ER- cases in the validated triplet list. Population mean rank difference is computed using the Wilcoxon signed-rank test. The plot shows that 10/11 genes are differentially amplified/deleted between ER+/ER- at $P \leqslant 0.05$. **b** Heatmap highlighting the difference in causal transcription factor activity as stratified by ER status. It can be seen that a large proportion TFs positively associated with lymphocyte infiltration have upregulated activity in the majority of ER- samples. Concurrently, TFs inversely or weakly correlated with the phenotype demonstrate stronger representation in ER+ samples over ER-. This figure has been reproduced from my preprint Chlon *et al.* [2]

The copy number profiles of all genomic drivers in the list of models (Table 5.1) significantly stratified by ER status with the exception of HAX1 (Fig. 5.4**A**). For example, genes such as *PIAS*3, *POU*2*F*1 and *CREBBP* were significantly more amplified in ER+ over ER- samples. The amplified states significantly downregulate TFs positively associated with lymphocytic activity such as $TFEB$, *IRF*1, *NR*3*C*1 and ETS1. This leads to significantly downregulated lymphocytic infiltration relative to ER- samples (Fig. 5.4**B**). In contrast, *TAL*1 and *CBFB* were significantly more amplified in ER- over ER+ samples, and significantly higher activity was observed for TFs positively associated with lymphocyte recruitment. Lower cytolytic activity was also observed in ER+ samples relative to ER-, which is unsurprising given that transcriptional targets of TFs positively associated with lymphocytic recruitment were also shown to modulate T cell activation in our GO term analysis (Fig. 5.3**B**).

These observations provide compelling evidence for a genomic basis for the ER stratification of lymphocyte infiltration and activity. These results are difficult to infer from association studies alone, highlighting a chief advantage of a deriving causal frameworks from large datasets.

5.5 Discussion

The aim of this study was to dissect interactions between cancer cells and their microenvironment. To achieve this aim, a multistep methodology was developed for inferring directed relationships between signaling activity and immune infiltration in the tumour microenvironment. This approach overcomes the limitations of conventional association studies by anchoring the analysis on somatic genomic events.

This work uses established methods to estimate TF activity and lymphocyte infiltration (CS) from gene expression data, and proposes a novel score for lymphocytic activity (TCS). Since genes tend to be amplified/deleted together, a PPI network is used as a biological network prior to isolate drivers from a list of genes correlated with an immune trait. Causal inference is achieved using a likelihood test, which returns the most likely relationship between the genotype, TF activity and phenotype given the data. Association based methods are widely used in cancer research and the CMIF methodology is a step towards a causal and mechanistic understanding of these relationships.

This analysis consisted of three steps: identifying models for the TCS/CS traits in a discovery cohort (METABRIC [76]), validating them in a large independent cohort (TCGA [201]) and further evaluating their predictive utility using orthogonal measures of lymphocyte infiltration from H&E images. The final model list revealed 11 driver genes regulating lymphocyte recruitment into the microenvironment by dysregulating the activity of 8 TFs. Whilst the majority TFs in these models have been experimentally linked to lymphocyte infiltration, many of driver genes identified in this study are novel, highlighting a principal advantage of causal driver discovery over standard association studies. This was further realized with the discovery of *EP*300/*NCOR*1 copy number alterations as drivers of cytolytic activity, whereas SNV mutations in these genes were previously found not to correlate with the trait in breast cancer. Drivers of lymphocytic infiltration were found to stratify by ER status, leading to significant stratification of activity profiles of TFs found to be causal for lymphocyte infiltration. This observation provides evidence supporting a genomic basis for the observed stratification of lymphocyte infiltration and prognostic utility by ER status.

This approach has several limitations, many of which are technical and relate to the assumptions to permit the statistical modelling of CMIF's framework. One such limitation involves measurement errors within the individual data inputs to the integrative analysis pipeline. If the margin of error for one variable is wider than that of another, it could potentially lead to the misclassification of the causal-reactive relationship between the two variables. Another limitation arises from the simplicity of the DAGs designed to model the interaction between variables. TFs causal for a trait will regulate genes that are interacting within the context of a much larger network and with feedback controls that need to be accounted for. Although these proposed models successfully predict lymphocyte infiltration in image cohorts, a stronger validation would involve knockdown experiments in mice in order to directly observe changes in TF profile and a trait of interest. Finally, although our analysis proceeded through a likelihood method-based approach, sampling from the precision matrix could potentially serve as a less computationally expensive alternative. For example, Friedman *et al.* propose *Glasso*, a method for learning sparse networks from a precision matrix using a lasso-based approach; their method demonstrates remarkable computational efficiency with respect to large datasets [269]. CMIF's methodology uses well-established regularisation methods to prevent overfitting and optimise model selection, a step not explicitly dealt with by *Glasso*.

A conceptual limitation is that the whole study is based on one major assumption: genomic events in the cell-autonomous compartment drive the development of the cancer and

can thus be used as anchors for causal analysis. This assumption is shared by almost all cancer genomics studies, in particular those that aim to identify genomic drivers of the breast cancer [76, 230]. However, it is acknowledged that in rare occasions, local tissue disruptions such as prolonged period of inflammation could be causal for genomic alterations instead of being caused by them.

Despite these limitations, the CMIF method has demonstrated the ability to recapitulate known mechanisms and has proposed robust models across a variety of independent datasets and orthogonal data types. Given this method's high validation rate between two large and independent datasets and its capacity to predict results supported by the literature, it stands a robust predictor of cancer-immune communication mechanisms from multiomics data.

This analysis was focused on breast cancer, but large efforts like The Cancer Genome Atlas (TCGA) or the International Cancer Genome Consortium (ICGC) provide the same types of data for many other kinds of cancer and thus CMIF's methodological framework can be easily be applied to many other forms of cancer. Additionally, the framework is not confined to the CS/TCS metrics as measures of immune activity and can be applied to any other available feature of the microenvironment.

In summary, my thesis has presented an integrative analysis of genomic events, signaling activity and immune markers, which is flexible and can form the foundation for a more mechanistic understanding of tumour-microenvironment interactions across cancer types.

Chapter 6

Summary and Outlook

Understanding the interaction between cancer and the immune system is pivotal to the development of novel therapeutics. With the advent of high-throughput platforms for collecting biological data, computational immunology can be used to build novel representations of cancer immunity at resolutions ranging from the molecular to sample level. By combining statistical approaches with multiomics data, my thesis tackles two pressing issues in cancer immunology:

1. The association between immune traits and cancer signalling is not fully understood. Dysregulation in the cancer epithelium leads to molecular signalling aberrations associated with local immunosuppression in the TME. In chapter 3, I demonstrate how the differential activity of Hedgehog and Notch signalling in PDAC promote contrasting immune landscapes. Although such associations are well established, my approach links cancer immunity back to the DNA level, by placing both signalling dysregulation and immune features in the context of somatic point mutations at the *KRAS* gene locus.

2. Reconstructing mechanisms underlying immune traits cannot be done using associations alone. Systems biology focuses on reconstructing causal mechanisms underlying a trait by learning dependency structures between molecular components. Recent methods suggest that structures inspired by hierarchical biological processes give rise to more accurate representations of a system. Chapter 5 builds on this observation, where we propose a framework for learning the dependency structure between copy number profiles, transcriptional dysregulation and immune features using Bayesian network approaches. This framework was used to elucidate *de novo* mutations regulating lymphocyte recruitment and cytolytic activity.

Problem 1. requires careful measurements of cancer epithelium signalling, immune features and factors contributed by other TME constituents such as stromal cells. In our

approach, we assume that the aggregate expression of a TF regulon acts as a multiplexed reporter for the activity of the TF protein. Under this assumption, we are able to fit the transcriptional signature of oncogenic *KRAS* to three biological processes characterised by master regulator TFs and their regulons. This approach is particularly powerful since it implicitly describes a system rooted at oncogenic *KRAS*, propagating through layers of transcriptional reprogramming and manifesting in immune response disruption. However, the lack of a formal dependency structure is a major caveat deterring us from confidently labelling our system a "mechanism" of cancer-immune landscaping. Examples from well-established findings may corroborate our discovery of a link between upregulated embryonic development pathways and dysregulated leukocyte behaviour in PDAC, but we cannot make definitive causal claims using correlations alone: *correlation does not imply causation.*

This brings us to problem 2: reconstructing immune regulatory mechanisms. Cancer cells signal to the immune system through complex transcription, translation and transduction regulatory hierarchies. Our assertion that genomic alterations are progenitor events for the causal procession of cancer development enables us to use them as anchors for subsequent events. Linking these processes together into a mechanism is achievable if we account for the regulatory structure within and between data modalities. We facilitate this by correcting gene expression associations for co-regulation and gene-level copy number calls for co-amplification/deletion. The usefulness of my approach was demonstrated by identifying *de novo* regulatory hierarchies for immune features rooted at novel mutations. My method generates consistent models across independent cohorts that are well substantiated by similar *in vivo* and *in vitro* findings from other studies.

Murine Validation Study Our approach in chapter 5 proposed 12 novel mechanisms for CD8+ T cell regulation in BRCA from existing datasets. A gold standard approach for functionally validating these kinds of models involves the *in vivo* administration of carcinoma cell lines genetically modified with respect to the expression of the driver gene, and assessing variations in CD8+ T cell phenotypes and TF activity. At the time of writing, an experimental protocol is underway to validate our predicted mechanisms as predictors for ground truth lymphocyte recruitment. In our *in vivo* validation protocol, we aim to address the following hypotheses:

1. For a mutation g, Is there a quantifiable difference in the number of infiltrating CD8+ cytotoxic T cells between mutated cell lines and controls? Is this difference statistically significant?

2. Which transcription factors, if any, are up/down-regulated in the tumour microenvironment of cells with mutation vs control?

3. If hypotheses 1. And 2. fail, can this failure be explained in the context of other infiltrating immune cells or variables unaccounted for by the original model?

The outcome of testing hypotheses 1-3 will inform us how well our approach is able to predict phenotype variance at the *in vivo* level. To address these points, we are in the process of engineering spontaneous murine carcinoma cells with somatic mutations sampled from our list of predicted drivers. To do this, we make use of the EMT6 mus musculus mammary carcinoma cell line, which is typically used to grow tumours in the murine mammary fat pad of immunogenic mice without risk of spontaneous rejection. There is evidence to suggest that EMT6 is more immunogenic than 4T1, another commonly used cell line and most importantly, EMT6 tumours demonstrate robust T cell infiltration in immunogenic mice [173]. Validation work is actively being done in collaboration with the biorepository unit core at Cancer Research UK, Cambridge Institute.

Non-immune phenotypes Given their relation to improved prognostic outcomes, lymphocyte features form extremely popular lines of inquiry for cancer immunologists. As such, our proposed method in chapter 5 looked at lymphocyte features downstream of perturbed transcriptional programmes to reconstruct regulatory hierarchies. On the other hand, our method can be readily extended to any other immunological feature or TME phenotype. The integrative multiomics basis of our framework means that phenotype data originating from a variety of experimental platforms can be incorporated. The only requirement is that these phenotypes can be anchored on a progenitor mutation in the cancer epithelium. This paves the way for reconstructing the regulatory hierarchy involved in non-lymphocyte immune agency, stromal infiltration or even events leading to cancer metastasis.

Extending the Hierarchy Our approach in chapter 4 only accounts for cell operations organised on the genomic, transcriptomic and morphological layers. This is a substantial improvement on earlier approaches that model phenotypes as the function of a single layer. However, there are many more layers connecting genomic aberrations in the cancer epithelium to immune phenotypes, such as the proteome and the metabolome which contains molecules essential for routine cell operations such as amino acids and sugars. Comprehensively characterising mechanisms giving rise to immune phenotypes requires addressing all layers of cellular organisation. This poses a considerable challenge given the limited availability of matched data types. International data gathering consortia are a step in the

right direction, with growing cohorts of matched multimodal datasets paving the way for future models integrating genomic, transcriptomic, proteomic, metabolomic and imageomic data. Less fragmented hierarchical regulatory models are likely to be more robust predictors of phenotypic changes.

References

[1] Shivan Sivakumar, Ines de Santiago, Leon Chlon, et al. "Master Regulators of Oncogenic KRAS Response in Pancreatic Cancer: An Integrative Network Biology Analysis". In: *PLOS Medicine* 14.1 (Jan. 2017). Ed. by Marc Ladanyi, e1002223.

[2] Leon Chlon and Florian Markowetz. "Causal Modeling Dissects Tumour–Microenvironment Interactions In Breast Cancer". In: *bioRxiv* (2017).

[3] Fabrice Andre and Lajos Pusztai. "Molecular classification of breast cancer: implications for selection of adjuvant chemotherapy". In: *Nature Clinical Practice Oncology* 3.11 (Nov. 2006), pp. 621–632.

[4] P C Nowell. "The clonal evolution of tumor cell populations." In: *Science (New York, N.Y.)* 194.4260 (Oct. 1976), pp. 23–8.

[5] Lauren M.F. Merlo, John W. Pepper, Brian J. Reid, and Carlo C. Maley. "Cancer as an evolutionary and ecological process". In: *Nature Reviews Cancer* 6.12 (Dec. 2006), pp. 924–935.

[6] Philippe L. Bedard, Aaron R. Hansen, Mark J. Ratain, and Lillian L. Siu. "Tumour heterogeneity in the clinic". In: *Nature* 501.7467 (Sept. 2013), pp. 355–364.

[7] Nicholas R Bertos, Morag Park, T Fehm, et al. "Breast cancer - one term, many entities?" In: *The Journal of clinical investigation* 121.10 (Oct. 2011), pp. 3789–96.

[8] Andriy Marusyk, Vanessa Almendro, and Kornelia Polyak. "Intra-tumour heterogeneity: a looking glass for cancer?" In: *Nature Reviews Cancer* 12.5 (Apr. 2012), pp. 323–334.

[9] Sohrab P. Shah, Ryan D. Morin, Jaswinder Khattra, et al. "Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution". In: *Nature* 461.7265 (Oct. 2009), pp. 809–813.

[10] Charles M. Perou, Therese Sorlie, Michael B. Eisen, et al. "Molecular portraits of human breast tumours". In: *Nature* 406.6797 (Aug. 2000), pp. 747–752.

[11] Richard M. Durbin, David L. Altshuler, Richard M. Durbin, et al. "A map of human genome variation from population-scale sequencing". In: *Nature* 467.7319 (Oct. 2010), pp. 1061–1073.

[12] Michael Lynch. "Rate, molecular spectrum, and consequences of human mutation." In: *Proceedings of the National Academy of Sciences of the United States of America* 107.3 (Jan. 2010), pp. 961–8.

[13] Tony Hunter. "Signaling—2000 and Beyond". In: *Cell* 100.1 (Jan. 2000), pp. 113–127.

[14] Douglas Hanahan, Robert A. Weinberg, K.H. Pan, et al. “Hallmarks of Cancer: The Next Generation”. In: *Cell* 144.5 (Mar. 2011), pp. 646–674.

[15] Carmen Criscitiello, Angela Esposito, and Giuseppe Curigliano. “Tumor–stroma crosstalk”. In: *Current Opinion in Oncology* 26.6 (Nov. 2014), pp. 551–555.

[16] Federico Garrido, Francisco Ruiz-Cabello, Teresa Cabrera, et al. “Implications for immunosurveillance of altered HLA class I phenotypes in human tumours”. In: *Immunology Today* 18.2 (Feb. 1997), pp. 89–95.

[17] Dass S. Vinay, Elizabeth P. Ryan, Graham Pawelec, et al. “Immune evasion in cancer: Mechanistic basis and therapeutic strategies”. In: *Seminars in Cancer Biology* 35 (Dec. 2015), S185–S198.

[18] Koichi S. Kobayashi and Peter J. van den Elsen. “NLRC5: a key regulator of MHC class I-dependent immune responses”. In: *Nature Reviews Immunology* 12.12 (Nov. 2012), pp. 813–820.

[19] Mingyao Li, Isabel X. Wang, Yun Li, et al. “Widespread RNA and DNA Sequence Differences in the Human Transcriptome”. In: *Science* 333.6038 (2011).

[20] Jonathan W. Yewdell, Ulrich Schubert, Luis C. Anton, et al. “Rapid degradation of a large fraction of newly synthesized proteins by proteasomes”. In: *Nature* 404.6779 (Apr. 2000), pp. 770–774.

[21] Eric A. J. Reits, Jan C. Vos, Monique Gromme, and Jacques Neefjes. “The major substrates for TAP in vivo are derived from newlysynthesized proteins”. In: *Nature* 404.6779 (Apr. 2000), pp. 774–778.

[22] Jacques Neefjes, Marlieke L. M. Jongsma, Petra Paul, and Oddmund Bakke. “Towards a systems understanding of MHC class I and MHC class II antigen presentation”. In: *Nature Reviews Immunology* 11.12 (Nov. 2011), p. 823.

[23] E Gilboa. “The makings of a tumor rejection antigen.” In: *Immunity* 11.3 (Sept. 1999), pp. 263–70.

[24] Ton N. Schumacher and Robert D. Schreiber. “Neoantigens in cancer immunotherapy”. In: *Science* 348.6230 (2015).

[25] J. C. Castle, S. Kreiter, J. Diekmann, et al. “Exploiting the Mutanome for Tumor Vaccination”. In: *Cancer Research* 72.5 (Mar. 2012), pp. 1081–1091.

[26] Y.-C. Lu, X. Yao, J. S. Crystal, et al. “Efficient Identification of Mutated Cancer Antigens Recognized by T Cells Associated with Durable Tumor Regressions”. In: *Clinical Cancer Research* 20.13 (July 2014), pp. 3401–3410.

[27] Nienke van Rooij, Marit M. van Buuren, Daisy Philips, et al. “Tumor Exome Analysis Reveals Neoantigen-Specific T-Cell Reactivity in an Ipilimumab-Responsive Melanoma”. In: *Journal of Clinical Oncology* 31.32 (Nov. 2013), e439–e442.

[28] Ludmil B. Alexandrov, Serena Nik-Zainal, David C. Wedge, et al. “Signatures of mutational processes in human cancer”. In: *Nature* 500.7463 (Aug. 2013), pp. 415–421.

[29] Glenn Dranoff. “Cytokines in cancer pathogenesis and cancer therapy”. In: *Nature Reviews Cancer* 4.1 (Jan. 2004), pp. 11–22.

[30] Jerry L. Adams, James Smothers, Roopa Srinivasan, and Axel Hoos. “Big opportunities for small molecules in immuno-oncology”. In: *Nature Reviews Drug Discovery* 14.9 (July 2015), pp. 603–622.

[31] Brian P Dolan, Kenneth D Gibbs, and Suzanne Ostrand-Rosenberg. “Dendritic cells cross-dressed with peptide MHC class I complexes prime CD8+ T cells.” In: *Journal of immunology (Baltimore, Md. : 1950)* 177.9 (Nov. 2006), pp. 6018–24.

[32] Jr Charles A Janeway, Paul Travers, Mark Walport, and Mark J Shlomchik. “The Humoral Immune Response”. In: (2001).

[33] Michael T. Osterholm. “Preparing for the Next Pandemic”. In: *New England Journal of Medicine* 352.18 (May 2005), pp. 1839–1842.

[34] J. Yokokawa, V. Cereda, C. Remondo, et al. “Enhanced Functionality of CD4+CD25highFoxP3+ Regulatory T Cells in the Peripheral Blood of Patients with Prostate Cancer”. In: *Clinical Cancer Research* 14.4 (Feb. 2008), pp. 1032–1040.

[35] Tyler J Curiel, George Coukos, Linhua Zou, et al. “Specific recruitment of regulatory T cells in ovarian carcinoma fosters immune privilege and predicts reduced survival”. In: *Nature Medicine* 10.9 (Sept. 2004), pp. 942–949.

[36] Makoto Miyara and Shimon Sakaguchi. “Natural regulatory T cells: mechanisms of suppression”. In: *Trends in Molecular Medicine* 13.3 (Mar. 2007), pp. 108–116.

[37] Andrea Facciabene, Xiaohui Peng, Ian S. Hagemann, et al. “Tumour hypoxia promotes tolerance and angiogenesis via CCL28 and Treg cells”. In: *Nature* 475.7355 (July 2011), pp. 226–230.

[38] Bin Shang, Yao Liu, Shu-juan Jiang, and Yi Liu. “Prognostic value of tumor-infiltrating FoxP3+ regulatory T cells in cancers: a systematic review and meta-analysis”. In: *Scientific Reports* 5 (Oct. 2015), p. 15179.

[39] Antonio Sica, Paola Larghi, Alessandra Mancino, et al. “Macrophage polarization in tumour progression”. In: *Seminars in Cancer Biology* 18.5 (Oct. 2008), pp. 349–355.

[40] Jon G Quatromoni and Evgeniy Eruslanov. “Tumor-associated macrophages: function, phenotype, and link to prognosis in human lung cancer.” In: *American journal of translational research* 4.4 (2012), pp. 376–89.

[41] L VANKEMPEN. “The tumor microenvironment: a critical determinant of neoplastic evolution”. In: *European Journal of Cell Biology* 82.11 (Nov. 2003), pp. 539–548.

[42] Alberto Mantovani and Antonio Sica. “Macrophages, innate immunity and cancer: balance, tolerance, and diversity”. In: *Current Opinion in Immunology* 22.2 (Apr. 2010), pp. 231–237.

[43] Dmitry I Gabrilovich and Srinivas Nagaraj. “Myeloid-derived suppressor cells as regulators of the immune system.” In: *Nature reviews. Immunology* 9.3 (Mar. 2009), pp. 162–74.

[44] Frances R. Balkwill, Melania Capasso, and Thorsten Hagemann. “The tumor microenvironment at a glance”. In: *Journal of Cell Science* 125.23 (2013).

[45] Ya-Qing Li, Fang-Fang Liu, Xin-Min Zhang, et al. “Tumor Secretion of CCL22 Activates Intratumoral Treg Infiltration and Is Independent Prognostic Predictor of Breast Cancer”. In: *PLoS ONE* 8.10 (Oct. 2013). Ed. by Hiromu Suzuki, e76379.

[46] Jia Sun, Jintang Sun, Bingfeng Song, et al. "Fucoidan inhibits CCL22 production through NF-κB pathway in M2 macrophages: a potential therapeutic strategy for cancer". In: *Scientific Reports* 6.1 (Dec. 2016), p. 35855.

[47] Fernando Oneissi Martinez, Antonio Sica, Alberto Mantovani, and Massimo Locati. "Macrophage activation and polarization." In: *Frontiers in bioscience : a journal and virtual library* 13 (Jan. 2008), pp. 453–61.

[48] Carly Bess Williams, Elizabeth S Yeh, and Adam C Soloff. "Tumor-associated macrophages: unwitting accomplices in breast cancer malignancy." In: *NPJ breast cancer* 2 (2016), p. 15025.

[49] Sandip Pravin Patel and Razelle Kurzrock. "PD-L1 Expression as a Predictive Biomarker in Cancer Immunotherapy". In: *Molecular Cancer Therapeutics* 14.4 (2015).

[50] E John Wherry and Makoto Kurachi. "Molecular and cellular insights into T cell exhaustion." In: *Nature reviews. Immunology* 15.8 (Aug. 2015), pp. 486–99.

[51] Kathleen M. Mahoney, Paul D. Rennert, and Gordon J. Freeman. "Combination cancer immunotherapy and new immunomodulatory targets". In: *Nature Reviews Drug Discovery* 14.8 (July 2015), pp. 561–584.

[52] Naoko Takebe, Lucio Miele, Pamela Jo Harris, et al. "Targeting Notch, Hedgehog, and Wnt pathways in cancer stem cells: clinical update." In: *Nature reviews. Clinical oncology* 12.8 (Aug. 2015), pp. 445–64.

[53] Atsushi Otsuka, Jil Dreier, Phil F Cheng, et al. "Hedgehog Pathway Inhibitors Promote Adaptive Immune Responses in Basal Cell Carcinoma". In: *Clin Cancer Res* 1.9 ().

[54] Derk Amsen, Christina Helbig, and Ronald A. Backer. "Notch in T Cell Differentiation: All Things Considered". In: *Trends in Immunology* 36.12 (Dec. 2015), pp. 802–814.

[55] Debbie Liao, Yunping Luo, Dorothy Markowitz, et al. "Cancer Associated Fibroblasts Promote Tumor Growth and Metastasis by Modulating the Tumor Immune Microenvironment in a 4T1 Murine Breast Cancer Model". In: *PLoS ONE* 4.11 (Nov. 2009). Ed. by Joseph Alan Bauer, e7965.

[56] Wei Tan, Weizhou Zhang, Amy Strasner, et al. "Tumour-infiltrating regulatory T cells stimulate mammary cancer metastasis through RANKL–RANK signalling". In: *Nature* 470.7335 (Feb. 2011), pp. 548–553.

[57] C. Feig, J. O. Jones, M. Kraman, et al. "Targeting CXCL12 from FAP-expressing carcinoma-associated fibroblasts synergizes with anti-PD-L1 immunotherapy in pancreatic cancer". In: *Proceedings of the National Academy of Sciences* 110.50 (Dec. 2013), pp. 20212–20217.

[58] Gavin P. Dunn, Catherine M. Koebel, and Robert D. Schreiber. "Interferons, immunity and cancer immunoediting". In: *Nature Reviews Immunology* 6.11 (Nov. 2006), pp. 836–848.

[59] F M Burnet. "The concept of immunological surveillance." In: *Progress in experimental tumor research* 13 (1970), pp. 1–27.

[60] Gavin P. Dunn, Lloyd J. Old, and Robert D. Schreiber. "The Three Es of Cancer Immunoediting". In: *Annual Review of Immunology* 22.1 (Apr. 2004), pp. 329–360.

[61] M E van den Broek, D Kägi, F Ossendorp, et al. “Decreased tumor surveillance in perforin-deficient mice.” In: *The Journal of experimental medicine* 184.5 (Nov. 1996), pp. 1781–90.

[62] S E Street, E Cretney, and M J Smyth. “Perforin and interferon-gamma activities independently control tumor initiation, growth, and metastasis.” In: *Blood* 97.1 (Jan. 2001), pp. 192–7.

[63] Ryungsa Kim, Manabu Emi, and Kazuaki Tanabe. “Cancer immunoediting from immune surveillance to immune escape.” In: *Immunology* 121.1 (May 2007), pp. 1–14.

[64] L Zitvogel, M Terme, C Borg, and G Trinchieri. “Dendritic cell-NK cell cross-talk: regulation and physiopathology.” In: *Current topics in microbiology and immunology* 298 (2006), pp. 157–74.

[65] Michele W L Teng, Matthew D Vesely, Helene Duret, et al. “Opposing roles for IL-23 and IL-12 in maintaining occult cancer in an equilibrium state.” In: *Cancer research* 72.16 (Aug. 2012), pp. 3987–96.

[66] Deepak Mittal, Matthew M Gubin, Robert D Schreiber, and Mark J Smyth. “New insights into cancer immunoediting and its three component phases–elimination, equilibrium and escape.” In: *Current opinion in immunology* 27 (Apr. 2014), pp. 16–25.

[67] Yinyin Yuan, Henrik Failmezger, Oscar M. Rueda, et al. “Quantitative Image Analysis of Cellular Heterogeneity in Breast Tumors Complements Genomic Profiling”. In: *Science Translational Medicine* 4.157 (2012).

[68] Jérôme Galon, Franck Pagès, Francesco M Marincola, et al. “The immune score as a new possible approach for the classification of cancer.” In: *Journal of translational medicine* 10 (Jan. 2012), p. 1.

[69] Michael S. Rooney, Sachet A. Shukla, Catherine J. Wu, et al. “Molecular and Genetic Properties of Tumors Associated with Local Immune Cytolytic Activity”. In: *Cell* 160.1-2 (Jan. 2015), pp. 48–61.

[70] Shanrong Zhao, Wai-Ping Fung-Leung, Anton Bittner, et al. “Comparison of RNA-Seq and Microarray in Transcriptome Profiling of Activated T Cells”. In: *PLoS ONE* 9.1 (Jan. 2014). Ed. by Shu-Dong Zhang, e78644.

[71] A. H. Beck, A. R. Sangoi, S. Leung, et al. “Systematic Analysis of Breast Cancer Morphology Uncovers Stromal Features Associated with Survival”. In: *Science Translational Medicine* 3.108 (Nov. 2011), pp. 113–108.

[72] Guoan Zhang, Beatrix M Ueberheide, Sofia Waldemarson, et al. “Protein quantitation using mass spectrometry.” In: *Methods in molecular biology (Clifton, N.J.)* 673 (2010), pp. 211–22.

[73] Stefanie Boellner and Karl-Friedrich Becker. “Reverse Phase Protein Arrays-Quantitative Assessment of Multiple Biomarkers in Biopsies for Clinical Use.” In: *Microarrays (Basel, Switzerland)* 4.2 (Mar. 2015), pp. 98–114.

[74] Donna M. Muzny, Matthew N. Bainbridge, Kyle Chang, et al. “Comprehensive molecular characterization of human colon and rectal cancer”. In: *Nature* 487.7407 (July 2012), pp. 330–337.

[75] Roger McLendon, Allan Friedman, Darrell Bigner, et al. “Comprehensive genomic characterization defines human glioblastoma genes and core pathways”. In: *Nature* 455.7216 (Oct. 2008), pp. 1061–1068.

[76] Christina Curtis, Sohrab P. Shah, Suet-Feung Chin, et al. “The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups”. In: *Nature* 486.7403 (Apr. 2012), p. 346.

[77] Alexandra Snyder, Vladimir Makarov, Taha Merghoub, et al. “Genetic Basis for Clinical Response to CTLA-4 Blockade in Melanoma”. In: *New England Journal of Medicine* 371.23 (Dec. 2014), pp. 2189–2199.

[78] Bernhard Mlecnik, Gabriela Bindea, Helen K. Angell, et al. “Integrative Analyses of Colorectal Cancer Show Immunoscore Is a Stronger Predictor of Patient Survival Than Microsatellite Instability”. In: *Immunity* 44.3 (Mar. 2016), pp. 698–711.

[79] Cristina Vilanova and Manuel Porcar. “Are multi-omics enough?” In: *Nature Microbiology* 1.8 (July 2016), p. 16101.

[80] Matteo Bersanelli, Ettore Mosca, Daniel Remondini, et al. “Methods for the integration of multi-omics data: mathematical aspects”. In: *BMC Bioinformatics* 17.S2 (Dec. 2016), S15.

[81] Eric E Schadt, John Lamb, Xia Yang, et al. “An integrative genomics approach to infer causal associations between gene expression and disease”. In: *Nature Genetics* 37.7 (July 2005), pp. 710–717.

[82] Jun Zhu, Matthew C. Wiener, Chunsheng Zhang, et al. “Increasing the Power to Detect Causal Associations by Combining Genotypic and Expression Data in Segregating Populations”. In: *PLoS Computational Biology* 3.4 (2007), e69.

[83] Mandeep Kaur and Luke Esau. “Two-step protocol for preparing adherent cells for high-throughput flow cytometry”. In: *BioTechniques* 59.3 (Sept. 2015), pp. 119–26.

[84] Leandro Luongo de Matos, Damila Cristina Trufelli, Maria Graciela Luongo de Matos, and Maria Aparecida da Silva Pinhal. “Immunohistochemistry as an important tool in biomarkers detection and clinical practice.” In: *Biomarker insights* 5 (Feb. 2010), pp. 9–20.

[85] Aaron M Newman, Chih Long Liu, Michael R Green, et al. “Robust enumeration of cell subsets from tissue expression profiles”. In: *Nature Methods* 12.5 (Mar. 2015), pp. 453–457.

[86] D Venet, F Pecasse, C Maenhaut, and H Bersini. “Separation of samples into their constituents using gene expression data.” In: *Bioinformatics (Oxford, England)* 17 Suppl 1 (2001), pp. 279–87.

[87] Dirk Repsilber, Sabine Kern, Anna Telaar, et al. “Biomarker discovery in heterogeneous tissue samples -taking the in-silico deconfounding approach”. In: *BMC Bioinformatics* 11.1 (Jan. 2010), p. 27.

[88] Timo Erkkilä, Saara Lehmusvaara, Pekka Ruusuvuori, et al. “Probabilistic analysis of gene expression measurements from heterogeneous tissues”. In: *Bioinformatics* 26.20 (Oct. 2010), pp. 2571–2577.

[89] Alexandre Kuhn, Doris Thu, Henry J Waldvogel, et al. “Population-specific expression analysis (PSEA) reveals molecular changes in diseased brain”. In: *Nature Methods* 8.11 (Oct. 2011), pp. 945–947.

[90] Peng Lu, Aleksey Nakorchevskiy, and Edward M Marcotte. “Expression deconvolution: a reinterpretation of DNA microarray data reveals dynamic changes in cell populations.” In: *Proceedings of the National Academy of Sciences of the United States of America* 100.18 (Sept. 2003), pp. 10370–5.

[91] Alexander R. Abbas, Kristen Wolslegel, Dhaya Seshasayee, et al. “Deconvolution of Blood Microarray Data Identifies Cellular Activation Patterns in Systemic Lupus Erythematosus”. In: *PLoS ONE* 4.7 (July 2009). Ed. by Patrick Tan, e6098.

[92] Ting Gong, Nicole Hartmann, Isaac S. Kohane, et al. “Optimal Deconvolution of Transcriptional Profiling Data Using Quadratic Programming with Application to Complex Clinical Blood Samples”. In: *PLoS ONE* 6.11 (Nov. 2011). Ed. by Magnus Rattray, e27156.

[93] R. Gaujoux and C. Seoighe. “CellMix: a comprehensive toolbox for gene expression deconvolution”. In: *Bioinformatics* 29.17 (Sept. 2013), pp. 2211–2212.

[94] Wenlian Qiao, Gerald Quon, Elizabeth Csaszar, et al. “PERT: A Method for Expression Deconvolution of Human Blood Samples from Varied Microenvironmental and Developmental Conditions”. In: *PLoS Computational Biology* 8.12 (Dec. 2012). Ed. by Richard Bonneau, e1002838.

[95] David A. Liebner, Kun Huang, and Jeffrey D. Parvin. “MMAD: microarray microdissection with analysis of differences is a computational tool for deconvoluting cell type-specific contributions from tissue samples”. In: *Bioinformatics* 30.5 (Mar. 2014), pp. 682–689.

[96] D. Ghosh. “Mixture models for assessing differential expression in complex tissues using microarray data”. In: *Bioinformatics* 20.11 (July 2004), pp. 1663–1669.

[97] Etienne Becht, Nicolas A. Giraldo, Laetitia Lacroix, et al. “Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression”. In: *Genome Biology* 17.1 (Dec. 2016), p. 218.

[98] Bo Li, Eric Severson, Jean-Christophe Pignon, et al. “Comprehensive analyses of tumor immunity: implications for cancer immunotherapy”. In: *Genome Biology* 17.1 (Dec. 2016), p. 174.

[99] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, et al. “Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.” In: *Proceedings of the National Academy of Sciences of the United States of America* 102.43 (Oct. 2005), pp. 15545–50.

[100] David A Barbie, Pablo Tamayo, Jesse S Boehm, et al. “Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1.” In: *Nature* 462.7269 (Nov. 2009), pp. 108–12.

[101] Kosuke Yoshihara, Maria Shahmoradgoli, Emmanuel Martínez, et al. “Inferring tumour purity and stromal and immune cell admixture from expression data”. In: *Nature Communications* 4 (Oct. 2013), p. 2612.

[102] Gabriela Bindea, Bernhard Mlecnik, Marie Tosolini, et al. “Spatiotemporal Dynamics of Intratumoral Immune Cells Reveal the Immune Landscape in Human Cancer”. In: *Immunity* 39.4 (Oct. 2013), pp. 782–795.

[103] Yasin Şenbabaoğlu, Ron S Gejman, Andrew G Winer, et al. "Tumor immune microenvironment characterization in clear cell renal cell carcinoma identifies prognostic and immunotherapeutically relevant messenger RNA signatures." In: *Genome biology* 17.1 (Nov. 2016), p. 231.

[104] Maxim V Kuleshov, Matthew R Jones, Andrew D Rouillard, et al. "Enrichr: a comprehensive gene set enrichment analysis web server 2016 update." In: *Nucleic acids research* 44.W1 (July 2016), pp. 90–7.

[105] Vamsi K Mootha, Cecilia M Lindgren, Karl-Fredrik Eriksson, et al. "PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes". In: *Nature Genetics* 34.3 (July 2003), pp. 267–273.

[106] David J Lynn, Geoffrey L Winsor, Calvin Chan, et al. "InnateDB: facilitating systems-level analyses of the mammalian innate immune response." In: *Molecular systems biology* 4 (2008), p. 218.

[107] Duccio Cavalieri, Damariz Rivero, Luca Beltrame, et al. "DC-ATLAS: a systems biology resource to dissect receptor specific signal transduction in dendritic cells". In: *Immunome Research* 6.1 (Nov. 2010), p. 10.

[108] M Kanehisa and S Goto. "KEGG: kyoto encyclopedia of genes and genomes." In: *Nucleic acids research* 28.1 (Jan. 2000), pp. 27–30.

[109] Cedric Simillion, Robin Liechti, Heidi E.L. Lischer, et al. "Avoiding the pitfalls of gene set enrichment analysis with SetRank". In: *BMC Bioinformatics* 18.1 (Dec. 2017), p. 151.

[110] A. Alexa, J. Rahnenfuhrer, and T. Lengauer. "Improved scoring of functional groups from gene expression data by decorrelating GO graph structure". In: *Bioinformatics* 22.13 (July 2006), pp. 1600–1607.

[111] Zhen Jiang and Robert Gentleman. "Extensions to gene set enrichment". In: *Bioinformatics* 23.3 (Feb. 2007), pp. 306–313.

[112] Yong Lu, Roni Rosenfeld, Itamar Simon, et al. "A probabilistic generative model for GO enrichment analysis". In: *Nucleic Acids Research* 36.17 (Oct. 2008), e109–e109.

[113] Jennifer E Smith-Garvin, Gary A Koretzky, and Martha S Jordan. "T cell activation." In: *Annual review of immunology* 27 (2009), pp. 591–619.

[114] James Robinson, Jason A Halliwell, James D Hayhurst, et al. "The IPD and IMGT/HLA database: allele variant databases." In: *Nucleic acids research* 43.Database issue (Jan. 2015), pp. 423–31.

[115] Sung Yoon Choo. "The HLA system: genetics, immunology, clinical testing, and clinical implications." In: *Yonsei medical journal* 48.1 (Feb. 2007), pp. 11–23.

[116] S C L Gough and M J Simmonds. "The HLA Region and Autoimmune Disease: Associations and Mechanisms of Action." In: *Current genomics* 8.7 (Nov. 2007), pp. 453–65.

[117] Xin Li, Nahla Ghandri, Daniela Piancatelli, et al. "Associations between HLA class I alleles and the prevalence of nasopharyngeal carcinoma (NPC) among Tunisians." In: *Journal of translational medicine* 5 (May 2007), p. 22.

[118] M P Lefranc. "IMGT, the international ImMunoGeneTics database." In: *Nucleic acids research* 29.1 (Jan. 2001), pp. 207–9.

[119] Sebastian Boegel, Martin Lower, Michael Schafer, et al. “HLA typing from RNA-Seq sequence reads”. In: *Genome Medicine* 4.12 (2012), p. 102.

[120] András Szolek, Benjamin Schubert, Christopher Mohr, et al. “OptiType: precision HLA typing from next-generation sequencing data”. In: *Bioinformatics* 30.23 (Dec. 2014), pp. 3310–3316.

[121] Scott D Brown, Rene L Warren, Ewan A Gibb, et al. “Neo-antigens predicted by tumor genome meta-analysis correlate with increased patient survival.” In: *Genome research* 24.5 (May 2014), pp. 743–50.

[122] N. McGranahan, A. J. S. Furness, R. Rosenthal, et al. “Clonal neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade”. In: *Science* 351.6280 (Mar. 2016), pp. 1463–1469.

[123] Rong-Fu Wang and Helen Y Wang. “Immune targets and neoantigens for cancer immunotherapy and precision medicine”. In: *Cell Research* 27.1 (Jan. 2017), pp. 11–37.

[124] N. A. Rizvi, M. D. Hellmann, A. Snyder, et al. “Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer”. In: *Science* 348.6230 (Apr. 2015), pp. 124–128.

[125] A Sette, A Vitiello, B Reherman, et al. “The relationship between class I binding affinity and immunogenicity of potential cytotoxic T cell epitopes.” In: *The Journal of Immunology* 153.12 (1994).

[126] Morten Nielsen, Claus Lundegaard, Thomas Blicher, et al. “NetMHCpan, a Method for Quantitative Predictions of Peptide Binding to Any HLA-A and -B Locus Protein of Known Sequence”. In: *PLoS ONE* 2.8 (Aug. 2007). Ed. by Esper Kallas, e796.

[127] Huynh-Hoa Bui, John Sidney, Bjoern Peters, et al. “Automated generation and evaluation of specific MHC binding predictive tools: ARB matrix applications”. In: *Immunogenetics* 57.5 (June 2005), pp. 304–314.

[128] Edita Karosiene, Michael Rasmussen, Thomas Blicher, et al. “NetMHCIIpan-3.0, a common pan-specific MHC class II prediction method including all three human MHC class II isotypes, HLA-DR, HLA-DP and HLA-DQ”. In: *Immunogenetics* 65.10 (Oct. 2013), pp. 711–724.

[129] C. Lundegaard, O. Lund, and M. Nielsen. “Accurate approximation method for prediction of class I MHC affinities for peptides of length 8, 10 and 11 using prediction tools trained on 9mers”. In: *Bioinformatics* 24.11 (June 2008), pp. 1397–1398.

[130] Paul A. Roche and Kazuyuki Furuta. “The ins and outs of MHC class II-mediated antigen processing and presentation”. In: *Nature Reviews Immunology* 15.4 (Feb. 2015), pp. 203–216.

[131] Ole Lund, Edita Karosiene, Claus Lundegaard, et al. “Bioinformatics Identification of Antigenic Peptide: Predicting the Specificity of Major MHC Class I and II Pathway Players”. In: Humana Press, Totowa, NJ, 2013, pp. 247–260.

[132] Morten Nielsen and Ole Lund. “NN-align. An artificial neural network-based alignment algorithm for MHC class II peptide binding prediction”. In: *BMC Bioinformatics* 10.1 (Sept. 2009), p. 296.

[133] Michal Bassani-Sternberg, Sune Pletscher-Frankild, Lars Juhl Jensen, and Matthias Mann. "Mass Spectrometry of Human Leukocyte Antigen Class I Peptidomes Reveals Strong Effects of Protein Abundance and Turnover on Antigen Presentation". In: *Molecular & Cellular Proteomics* 14.3 (Mar. 2015), pp. 658–673.

[134] Michael E. Birnbaum, Juan L. Mendoza, Dhruv K. Sethi, et al. "Deconstructing the Peptide-MHC Specificity of T Cell Recognition". In: *Cell* 157.5 (May 2014), pp. 1073–1087.

[135] Shu-Qi Zhang, Patricia Parker, Ke-Yue Ma, et al. "Direct measurement of T cell receptor affinity and sequence from naïve antiviral T cells". In: *Science Translational Medicine* 8.341 (2016).

[136] Sébastien Apcher, Rodrigo Prado Martins, and Robin Fåhraeus. "The source of MHC class I presented peptides and its implications". In: *Current Opinion in Immunology* 40 (June 2016), pp. 117–122.

[137] J W Yewdell, L C Antón, and J R Bennink. "Defective ribosomal products (DRiPs): a major source of antigenic peptides for MHC class I molecules?" In: *Journal of immunology (Baltimore, Md. : 1950)* 157.5 (Sept. 1996), pp. 1823–6.

[138] JL; Connolly. "Role of the Surgical Pathologist in the Diagnosis and Management of the Cancer Patient". In: *Holland-Frei Cancer Medicine. 6th edition.* Hamilton (ON): BC Decker; 2003.

[139] Marc Macenko, Marc Niethammer, J. S. Marron, et al. "A method for normalizing histology slides for quantitative analysis". In: *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*. IEEE, June 2009, pp. 1107–1110.

[140] Nobuyuki Otsu. "A Threshold Selection Method from Gray-Level Histograms". In: *IEEE Transactions on Systems, Man, and Cybernetics* 9.1 (Jan. 1979), pp. 62–66.

[141] Mitko Veta, Paul J. van Diest, Robert Kornegoor, et al. "Automatic Nuclei Segmentation in H&E Stained Breast Cancer Histopathology Images". In: *PLoS ONE* 8.7 (July 2013). Ed. by Konradin Metze, e70221.

[142] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep learning". In: *Nature* 521.7553 (May 2015), pp. 436–444.

[143] Dan C Cirean, Alessandro Giusti, Luca M Gambardella, and Jürgen Schmidhuber. "Mitosis Detection in Breast Cancer Histology Images with Deep Neural Networks". In: ().

[144] Yann Lecun, Yann Lecun, Léon Bottou, et al. "Gradient-based learning applied to document recognition". In: *PROCEEDINGS OF THE IEEE* 86 (1998), pp. 2278–2324.

[145] Ting Chen and Christophe Chefd'hotel. "Deep Learning Based Automatic Immune Cell Detection for Immunohistochemistry Images". In: Springer, Cham, 2014, pp. 17–24.

[146] Christine A. Iacobuzio-Donahue, Victor E. Velculescu, Christopher L. Wolfgang, and Ralph H. Hruban. "Genetic Basis of Pancreas Cancer Development and Progression: Insights from Whole-Exome and Whole-Genome Sequencing". In: *Clinical Cancer Research* 18.16 (2012).

[147] Eric A Collisson, Anguraj Sadanandam, Peter Olson, et al. “Subtypes of pancreatic ductal adenocarcinoma and their differing responses to therapy”. In: *Nature Medicine* 17.4 (Apr. 2011), pp. 500–503.

[148] Richard A Moffitt, Raoud Marayati, Elizabeth L Flate, et al. “Virtual microdissection identifies distinct tumor- and stroma-specific subtypes of pancreatic ductal adenocarcinoma”. In: *Nature Genetics* 47.10 (Sept. 2015), pp. 1168–1178.

[149] Peter J. Campbell, Shinichi Yachida, Laura J. Mudie, et al. “The patterns and dynamics of genomic instability in metastatic pancreatic cancer”. In: *Nature* 467.7319 (Oct. 2010), pp. 1109–1113.

[150] V T Smit, A J Boot, A M Smits, et al. “KRAS codon 12 mutations occur very frequently in pancreatic adenocarcinomas.” In: *Nucleic acids research* 16.16 (Aug. 1988), pp. 7773–82.

[151] Shinichi Yachida, Catherine M White, Yoshiki Naito, et al. “Clinical significance of the genetic landscape of pancreatic cancer and implications for identification of potential long-term survivors.” In: *Clinical cancer research : an official journal of the American Association for Cancer Research* 18.22 (Nov. 2012), pp. 6339–47.

[152] Marina Pasca di Magliano and Craig D Logsdon. “Roles for KRAS in pancreatic tumor development and progression.” In: *Gastroenterology* 144.6 (June 2013), pp. 1220–9.

[153] Ralph H Hruban, Anirban Maitra, Richard Schulick, et al. “Emerging molecular biology of pancreatic cancer.” In: *Gastrointestinal cancer research : GCR* 2.4 Suppl (July 2008), pp. 10–5.

[154] Robb E. Wilentz, Christine A. Iacobuzio-Donahue, Pedram Argani, et al. “Loss of Expression of Dpc4 in Pancreatic Intraepithelial Neoplasia: Evidence That DPC4 Inactivation Occurs Late in Neoplastic Progression”. In: *Cancer Research* 60.7 (2000).

[155] K. Kojima, S. M. Vickers, N. V. Adsay, et al. “Inactivation of Smad4 Accelerates KrasG12D-Mediated Pancreatic Neoplasia”. In: *Cancer Research* 67.17 (Sept. 2007), pp. 8121–8130.

[156] M Schutte, R H Hruban, J Geradts, et al. “Abrogation of the Rb/p16 tumor-suppressive pathway in virtually all pancreatic carcinomas.” In: *Cancer research* 57.15 (Aug. 1997), pp. 3126–30.

[157] Angela L McCleary-Wheeler, Robert McWilliams, and Martin E Fernandez-Zapico. “Aberrant signaling pathways in pancreatic cancer: a two compartment view.” In: *Molecular carcinogenesis* 51.1 (Jan. 2012), pp. 25–39.

[158] Christine A. Iacobuzio-Donahue, Anirban Maitra, Mari Olsen, et al. “Exploration of Global Gene Expression Patterns in Pancreatic Adenocarcinoma Using cDNA Microarrays”. In: *The American Journal of Pathology* 162.4 (Apr. 2003), pp. 1151–1162.

[159] Marina Pasca di Magliano and Craig D Logsdon. “Roles for KRAS in pancreatic tumor development and progression.” In: *Gastroenterology* 144.6 (June 2013), pp. 1220–9.

[160] Zhihong Xu, Alain Vonlaufen, Phoebe A. Phillips, et al. “Role of Pancreatic Stellate Cells in Pancreatic Cancer Metastasis”. In: *The American Journal of Pathology* 177.5 (Nov. 2010), pp. 2585–2596.

[161] Michael A Jacobetz, Derek S Chan, Albrecht Neesse, et al. “Hyaluronan impairs vascular function and drug delivery in a mouse model of pancreatic cancer”. In: *Gut* 62.1 (Jan. 2013), pp. 112–120.

[162] K. P. Olive, M. A. Jacobetz, C. J. Davidson, et al. “Inhibition of Hedgehog Signaling Enhances Delivery of Chemotherapy in a Mouse Model of Pancreatic Cancer”. In: *Science* 324.5933 (June 2009), pp. 1457–1461.

[163] Marina Lesina, Magdalena U. Kurkowski, Katharina Ludes, et al. “Stat3/Socs3 Activation by IL-6 Transsignaling Promotes Progression of Pancreatic Intraepithelial Neoplasia and Development of Pancreatic Cancer”. In: *Cancer Cell* 19.4 (Apr. 2011), pp. 456–469.

[164] Chantale Charo, Vijaykumar Holla, Thiruvengadam Arumugam, et al. “Prostaglandin E2 regulates pancreatic stellate cell activity via the EP4 receptor.” In: *Pancreas* 42.3 (Apr. 2013), pp. 467–74.

[165] Hiroshi Kurahara, Hiroyuki Shinchi, Yuko Mataki, et al. “Significance of M2-Polarized Tumor-Associated Macrophage in Pancreatic Cancer”. In: *Journal of Surgical Research* 167.2 (May 2011), e211–e219.

[166] Ash A. Alizadeh, Michael B. Eisen, R. Eric Davis, et al. “Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling”. In: *Nature* 403.6769 (Feb. 2000), pp. 503–511.

[167] Therese Sorlie, Robert Tibshirani, Joel Parker, et al. “Repeated observation of breast tumor subtypes in independent gene expression data sets.” In: *Proceedings of the National Academy of Sciences of the United States of America* 100.14 (July 2003), pp. 8418–23.

[168] Jacques Lapointe, Chunde Li, John P Higgins, et al. “Gene expression profiling identifies clinically relevant subtypes of prostate cancer.” In: *Proceedings of the National Academy of Sciences of the United States of America* 101.3 (Jan. 2004), pp. 811–6.

[169] T. R. Golub, D. K. Slonim, P. Tamayo, et al. “Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring”. In: *Science* 286.5439 (1999).

[170] Celine Lefebvre, Presha Rajbhandari, Mariano J Alvarez, et al. “A human B-cell interactome identifies MYB and FOXM1 as master regulators of proliferation in germinal centers”. In: *Molecular Systems Biology* 6 (June 2010).

[171] Mauro A A Castro, Ines de Santiago, Thomas M Campbell, et al. “Regulators of genetic risk of breast cancer identified by integrative network analysis.” In: *Nature genetics* 48.1 (Jan. 2016), pp. 12–21.

[172] Michael N C Fletcher, Mauro A A Castro, Xin Wang, et al. “Master regulators of FGFR2 signalling and breast cancer risk.” In: *Nature communications* 4 (2013), p. 2464.

[173] L Johnson, K Mercer, D Greenbaum, et al. “Somatic activation of the K-ras oncogene causes early onset lung cancer in mice.” In: *Nature* 410.6832 (Apr. 2001), pp. 1111–6.

[174] E L Jackson, N Willis, K Mercer, et al. “Analysis of lung tumor initiation and progression using conditional expression of oncogenic K-ras.” In: *Genes & development* 15.24 (Dec. 2001), pp. 3243–8.

[175] Matthew E Ritchie, Belinda Phipson, Di Wu, et al. “limma powers differential expression analyses for RNA-sequencing and microarray studies.” In: *Nucleic acids research* 43.7 (Apr. 2015), e47.

[176] Jianjiong Gao, Bülent Arman Aksoy, Ugur Dogrusoz, et al. “Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal.” In: *Science signaling* 6.269 (Apr. 2013), pl1.

[177] Laurent Gautier, Leslie Cope, Benjamin M Bolstad, and Rafael A Irizarry. “affy–analysis of Affymetrix GeneChip data at the probe level.” In: *Bioinformatics (Oxford, England)* 20.3 (Feb. 2004), pp. 307–15.

[178] Peter Bailey, David K. Chang, Katia Nones, et al. “Genomic analyses identify molecular subtypes of pancreatic cancer”. In: *Nature* 531.7592 (Feb. 2016), pp. 47–52.

[179] Robert Gentleman. *Bioinformatics and computational biology solutions using R and Bioconductor*. Springer Science+Business Media, 2005, p. 473.

[180] K. Strimmer. “fdrtool: a versatile R package for estimating local and tail area-based false discovery rates”. In: *Bioinformatics* 24.12 (June 2008), pp. 1461–1462.

[181] S. A.; Stouffer, E. A; Suchman, L. C.; DeVinney, et al. *Studies in Social Psychology in World War II: The American Soldier. Vol. 1, Adjustment During Army Life.* Princeton University Press., 1949.

[182] Mariano J Alvarez, Yao Shen, Federico M Giorgi, et al. “Functional characterization of somatic mutations in cancer using network-based inference of protein activity”. In: *Nature Genetics* 48.8 (June 2016), pp. 838–847.

[183] Richard Bourgon, Robert Gentleman, and Wolfgang Huber. “Independent filtering increases detection power for high-throughput experiments.” In: *Proceedings of the National Academy of Sciences of the United States of America* 107.21 (May 2010), pp. 9546–51.

[184] Darryl Nishimura. “BioCarta”. In: *Biotech Software & Internet Report* 2.3 (June 2001), pp. 117–120.

[185] Sonja Hanzelmann, Robert Castelo, and Justin Guinney. “GSVA: gene set variation analysis for microarray and RNA-Seq data”. In: *BMC Bioinformatics* 14.7 (2013).

[186] Seongho Kim. “ppcor: An R Package for a Fast Calculation to Semi-partial Correlation Coefficients.” In: *Communications for statistical applications and methods* 22.6 (Nov. 2015), pp. 665–674.

[187] Robert Tibshirani, Trevor Hastie, Balasubramanian Narasimhan, and Gilbert Chu. “Class Prediction by Nearest Shrunken Centroids, with Applications to DNA Microarrays”. In: *Statistical Science* 18.1 (2003), pp. 104–117.

[188] Hua Tian, Christopher A Callahan, Kelly J DuPree, et al. “Hedgehog signaling is restricted to the stromal compartment during pancreatic carcinogenesis.” In: *Proceedings of the National Academy of Sciences of the United States of America* 106.11 (Mar. 2009), pp. 4254–9.

[189] J. M. Bailey, B. J. Swanson, T. Hamada, et al. “Sonic Hedgehog Promotes Desmoplasia in Pancreatic Cancer”. In: *Clinical Cancer Research* 14.19 (Oct. 2008), pp. 5995–6004.

[190] S. A. Blaine, K. C. Ray, K. M. Branch, et al. “Epidermal growth factor receptor regulates pancreatic fibrosis”. In: *AJP: Gastrointestinal and Liver Physiology* 297.3 (Sept. 2009), G434–G441.

[191] Tanapat Palaga, Lucio Miele, Todd E Golde, and Barbara A Osborne. “TCR-mediated Notch signaling regulates proliferation and IFN-gamma production in peripheral T cells.” In: *Journal of immunology (Baltimore, Md. : 1950)* 171.6 (Sept. 2003), pp. 3019–24.

[192] Yoichi Maekawa, Yoshiaki Minato, Chieko Ishifune, et al. “Notch2 integrates signaling by the transcription factors RBP-J and CREB1 to promote T cell cytotoxicity”. In: *Nature Immunology* 9.10 (Oct. 2008), pp. 1140–1147.

[193] Rosa A Sierra, Paul Thevenot, Patrick L Raber, et al. “Rescue of notch-1 signaling in antigen-specific CD8+ T cells overcomes tumor-induced T-cell suppression and enhances immunotherapy in cancer.” In: *Cancer immunology research* 2.8 (Aug. 2014), pp. 800–11.

[194] Mélissa Mathieu, Natacha Cotta-Grand, Jean-François Daudelin, et al. “Notch signaling regulates PD-1 expression during CD8+ T-cell activation”. In: *Immunology and Cell Biology* 91.1 (Jan. 2013), pp. 82–88.

[195] Karin E. de Visser, Alexandra Eichten, and Lisa M. Coussens. “Paradoxical roles of the immune system during cancer development”. In: *Nature Reviews Cancer* 6.1 (Jan. 2006), pp. 24–37.

[196] Filipe C Martins, Ines de Santiago, Anne Trinh, et al. “Combined image and genomic analysis of high-grade serous ovarian cancer reveals PTEN loss as a common driver event and prognostic classifier.” In: *Genome biology* 15.12 (Dec. 2014), p. 526.

[197] Valeria Orrù, Maristella Steri, Gabriella Sole, et al. “Genetic Variants Regulating Immune Cell Levels in Health and Disease”. In: *Cell* 155.1 (Sept. 2013), pp. 242–256.

[198] H. Raza Ali, Leon Chlon, Paul D. P. Pharoah, et al. “Patterns of Immune Infiltration in Breast Cancer and Their Clinical Implications: A Gene-Expression-Based Retrospective Study”. In: *PLOS Medicine* 13.12 (Dec. 2016). Ed. by Marc Ladanyi, e1002194.

[199] Eun-Yeong Oh, Stephen M Christensen, Sindhu Ghanta, et al. “Extensive rewiring of epithelial-stromal co-expression networks in breast cancer”. In: *Genome Biology* 16.1 (Dec. 2015), p. 128.

[200] Judea. Pearl and Judea. *Causality : models, reasoning, and inference*. Cambridge University Press, 2000, p. 384.

[201] Daniel C. Koboldt, Robert S. Fulton, Michael D. McLellan, et al. "Comprehensive molecular portraits of human breast tumours". In: *Nature* 490.7418 (Sept. 2012), pp. 61–70.

[202] Serena Nik-Zainal, Helen Davies, Johan Staaf, et al. "Landscape of somatic mutations in 560 breast cancer whole-genome sequences". In: *Nature* 534.7605 (May 2016), pp. 47–54.

[203] Ron Edgar, Michael Domrachev, and Alex E Lash. "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository." In: *Nucleic acids research* 30.1 (Jan. 2002), pp. 207–10.

[204] M B Eisen, P T Spellman, P O Brown, and D Botstein. "Cluster analysis and display of genome-wide expression patterns." In: *Proceedings of the National Academy of Sciences of the United States of America* 95.25 (Dec. 1998), pp. 14863–8.

[205] Harald Steck and Tommi S Jaakkola. "Predictive Discretization during Model Selection". In: ().

[206] A J Butte and I S Kohane. "Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements." In: *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* (2000), pp. 418–29.

[207] Bernhard. Scholkopf and Alexander J. Smola. *Learning with kernels : support vector machines, regularization, optimization, and beyond*. MIT Press, 2002, p. 626.

[208] Yoav Benjamini and Yosef Hochberg. *Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing*. 1995.

[209] Steffen L. Lauritzen. *Graphical Models*. Clarendon Press, Oxford, 1996.

[210] Hiroyuki Toh and Katsuhisa Horimoto. "Inference of a genetic network by a combined approach of cluster analysis and graphical Gaussian modeling". In: *BIOINFORMATICS* 18.2 (2002), pp. 287–297.

[211] J. Schafer and K. Strimmer. "An empirical Bayes approach to inferring large-scale gene association networks". In: *Bioinformatics* 21.6 (Mar. 2005), pp. 754–764.

[212] Finn V. Jensen and Finn V. *Bayesian networks and decision graphs*. Springer, 2001, p. 268.

[213] Thomas Verma and Judea Pearl. *Uncertainty in artificial intelligence 6*. North-Holland, 1991, p. 528.

[214] Andrew Gelman. *Bayesian data analysis*, p. 661.

[215] David Heckerman, Dan Geiger, and David M Chickering. "Learning Bayesian Networks: The Combination of Knowledge and Statistical Data". In: *Machine Learning* 20 (1995), pp. 197–243.

[216] S. Bulashevska and R. Eils. "Inferring genetic regulatory logic from expression data". In: *Bioinformatics* 21.11 (June 2005), pp. 2706–2713.

[217] R. W. Robinson. "Counting unlabeled acyclic digraphs". In: Springer, Berlin, Heidelberg, 1977, pp. 28–43.

[218] David Maxwell Chickering. "Learning Equivalence Classes of Bayesian-Network Structures". In: *Journal of Machine Learning Research* 2 (2002), pp. 445–498.

[219] Robert Castelo and Tomas Kocka. “On Inclusion-Driven Learning of Bayesian Networks”. In: *Journal of Machine Learning Research* 4 (2003), pp. 527–574.

[220] Nir Friedman and Daphne Koller. “Being Bayesian About Network Structure. A Bayesian Approach to Structure Discovery in Bayesian Networks”. In: *Machine Learning* 50 (2003), pp. 95–125.

[221] Gal Elidan, Iftach Nachman INACHMAN, and Nir Friedman NIR. “" Ideal Parent " Structure Learning for Continuous Variable Bayesian Networks”. In: *Journal of Machine Learning Research* 8 (2007), pp. 1799–1833.

[222] W. K. Hastings. “Monte Carlo Sampling Methods Using Markov Chains and Their Applications”. In: *Biometrika* 57.1 (Apr. 1970), p. 97.

[223] Daniel E. Zak, Daniel E. Zak, Francis J. Doyle Iii, et al. “Simulation studies for the identification of genetic networks from cDNA array and regulatory activity data”. In: *PROCEEDINGS OF THE SECOND INTERNATIONAL CONFERENCE ON SYSTEMS BIOLOGY* (2001), pp. 231–238.

[224] D. Husmeier. “Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks”. In: *Bioinformatics* 19.17 (Nov. 2003), pp. 2271–2282.

[225] V Anne Smith, Erich D Jarvis, and Alexander J Hartemink. “Evaluating functional network inference using simulations of complex biological systems.” In: *Bioinformatics (Oxford, England)* 18 Suppl 1 (2002), pp. 216–24.

[226] Daniel E Zak, Gregory E Gonye, James S Schwaber, et al. “Importance of input perturbations and stochastic gene expression in the reverse engineering of genetic regulatory networks: insights from an identifiability analysis of an in silico network.” In: *Genome research* 13.11 (Nov. 2003), pp. 2396–405.

[227] Karen Sachs, Omar Perez, Dana Pe’er, et al. “Causal Protein-Signaling Networks Derived from Multiparameter Single-Cell Data”. In: *Science* 308.5721 (2005).

[228] Laura E. MacConaill and Levi A. Garraway. “Clinical Implications of the Cancer Genome”. In: *Journal of Clinical Oncology* 28.35 (Dec. 2010), pp. 5219–5228.

[229] Levi A. Garraway and Eric S. Lander. “Lessons from the Cancer Genome”. In: *Cell* 153.1 (Mar. 2013), pp. 17–37.

[230] Giovanni Ciriello, Martin L Miller, Bülent Arman Aksoy, et al. “Emerging landscape of oncogenic signatures across human cancers”. In: *Nature Genetics* 45.10 (Sept. 2013), pp. 1127–1133.

[231] Feng Zhang, Wenli Gu, Matthew E Hurles, and James R Lupski. “Copy number variation in human health, disease, and evolution.” In: *Annual review of genomics and human genetics* 10 (2009), pp. 451–81.

[232] James D. Watson. *Recombinant DNA : genes and genomes : a short course*. W.H. Freeman, 2007, p. 474.

[233] Patricia J. Wittkopp and Gizem Kalay. “Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence”. In: *Nature Reviews Genetics* 13.1 (Dec. 2011), p. 59.

[234] Walid D Fakhouri, Fedik Rahimov, Catia Attanasio, et al. "An etiologic regulatory mutation in IRF6 with loss- and gain-of-function effects." In: *Human molecular genetics* 23.10 (May 2014), pp. 2711–20.

[235] D. S. (Douglas Scott) Falconer and Trudy F. C. Mackay. *Introduction to quantitative genetics*. Longman, 1996, p. 464.

[236] Daniel R Rhodes, Shanker Kalyana-Sundaram, Vasudeva Mahavisno, et al. "Mining for regulatory programs in the cancer transcriptome". In: *Nature Genetics* 37.6 (June 2005), pp. 579–583.

[237] Noora Kotaja, Ulla Karvonen, Olli A Jänne, and Jorma J Palvimo. "PIAS proteins modulate transcription factors by functioning as SUMO-1 ligases." In: *Molecular and cellular biology* 22.14 (July 2002), pp. 5222–34.

[238] Logan Everett, Matthew Hansen, and Sridhar Hannenhalli. "Regulating the Regulators: Modulators of Transcription Factor Activity". In: *Methods in molecular biology (Clifton, N.J.)* Vol. 674. 2010, pp. 297–312.

[239] A. Breitkreutz, H. Choi, J. R. Sharom, et al. "A Global Protein Kinase and Phosphatase Interaction Network in Yeast". In: *Science* 328.5981 (May 2010), pp. 1043–1046.

[240] Mathew E. Sowa, Eric J. Bennett, Steven P. Gygi, and J. Wade Harper. "Defining the Human Deubiquitinating Enzyme Interaction Landscape". In: *Cell* 138.2 (July 2009), pp. 389–403.

[241] Arunachalam Vinayagam, Jonathan Zirin, Charles Roesel, et al. "Integrating protein-protein interaction networks with phenotypes reveals signs of interactions". In: *Nature Methods* 11.1 (Nov. 2013), pp. 94–99.

[242] C. Stark, B.-J. Breitkreutz, A. Chatr-aryamontri, et al. "The BioGRID Interaction Database: 2011 update". In: *Nucleic Acids Research* 39.Database (Jan. 2011), pp. D698–D704.

[243] S. Kerrien, B. Aranda, L. Breuza, et al. "The IntAct molecular interaction database in 2012". In: *Nucleic Acids Research* 40.D1 (Jan. 2012), pp. D841–D846.

[244] L. Salwinski, Christopher S Miller, Adam J Smith, et al. "The Database of Interacting Proteins: 2004 update". In: *Nucleic Acids Research* 32.90001 (Jan. 2004), pp. 449D–451.

[245] David Warde-Farley, Sylva L. Donaldson, Ovi Comes, et al. "The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function". In: *Nucleic Acids Research* 38.suppl_2 (July 2010), W214–W220.

[246] Zhenjun Hu, Jui-Hung Hung, Yan Wang, et al. "VisANT 3.5: multi-scale network visualization, analysis and inference based on the gene ontology". In: *Nucleic Acids Research* 37.suppl_2 (July 2009), W115–W121.

[247] Andrea Franceschini, Damian Szklarczyk, Sune Frankild, et al. "STRING v9.1: protein-protein interaction networks, with increased coverage and integration." In: *Nucleic acids research* 41.Database issue (Jan. 2013), pp. 808–15.

[248] Sahar M.A. Mahmoud, Emma Claire Paish, Desmond G. Powe, et al. "Tumor-Infiltrating CD8 ⁺ Lymphocytes Predict Clinical Outcome in Breast Cancer". In: *Journal of Clinical Oncology* 29.15 (May 2011), pp. 1949–1955.

[249] Drew M Pardoll. “The blockade of immune checkpoints in cancer immunotherapy.” In: *Nature reviews. Cancer* 12.4 (Mar. 2012), pp. 252–64.

[250] Craig H Mermel, Steven E Schumacher, Barbara Hill, et al. “GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers”. In: *Genome Biology* 12.4 (2011), R41.

[251] R Core Team. *R language definition*. 2000.

[252] Maria Stella Carro, Wei Keat Lim, Mariano Javier Alvarez, et al. “The transcriptional network for mesenchymal transformation of brain tumours”. In: *Nature* 463.7279 (Jan. 2010), pp. 318–325.

[253] Damian Szklarczyk, Andrea Franceschini, Stefan Wyder, et al. “STRING v10: protein-protein interaction networks, integrated over the tree of life.” In: *Nucleic acids research* 43.Database issue (Jan. 2015), pp. 447–52.

[254] Koji Nakagawa and Hideyoshi Yokosawa. “PIAS3 induces SUMO-1 modification and transcriptional repression of IRF-1.” In: *FEBS letters* 530.1-3 (Oct. 2002), pp. 204–8.

[255] Tomohiko Tamura, Hideyuki Yanai, David Savitsky, and Tadatsugu Taniguchi. “The IRF Family Transcription Factors in Immunity and Oncogenesis”. In: *Annual Review of Immunology* 26.1 (Apr. 2008), pp. 535–584.

[256] Suk-Chul Bae and Joong-Kook Choi. “Tumor suppressor activity of RUNX3”. In: *Oncogene* 23.24 (May 2004), pp. 4336–4340.

[257] Takashi Yano, Kosei Ito, Hiroshi Fukamachi, et al. “The RUNX3 Tumor Suppressor Upregulates Bim in Gastric Epithelial Cells Undergoing Transforming Growth Factor Beta-Induced Apoptosis”. In: *MOLECULAR AND CELLULAR BIOLOGY* 26.12 (2006), pp. 4474–4488.

[258] J.-R. Landry, S. Kinston, K. Knezevic, et al. “Runx genes are direct targets of Scl/Tal1 in the yolk sac and fetal liver”. In: *Blood* 111.6 (Mar. 2008), pp. 3005–3014.

[259] Lisa Russell and Lee Ann Garrett-Sinha. “Transcription factor Ets-1 in cytokine and chemokine gene regulation”. In: *Cytokine* 51.3 (Sept. 2010), pp. 217–226.

[260] Y W He, M L Deftos, E W Ojala, and M J Bevan. “RORgamma t, a novel isoform of an orphan receptor, negatively regulates Fas ligand expression and IL-2 production in T cells.” In: *Immunity* 9.6 (Dec. 1998), pp. 797–806.

[261] Martin F Bachmann and Annette Oxenius. “Interleukin 2: from immunostimulation to immunoregulation and back again.” In: *EMBO reports* 8.12 (Dec. 2007), pp. 1142–8.

[262] Aradhana Rani, Roseanna Greenlaw, Richard A Smith, and Christine Galustian. “HES1 in immunity and cancer”. In: *Cytokine & Growth Factor Reviews* 30 (Aug. 2016), pp. 113–117.

[263] C. E. Foulds, M. L. Nelson, A. G. Blaszczak, and B. J. Graves. “Ras/Mitogen-Activated Protein Kinase Signaling Activates Ets-1 and Ets-2 by CBP/p300 Recruitment”. In: *Molecular and Cellular Biology* 24.24 (Dec. 2004), pp. 10954–10964.

[264] D. Pan, M. Kocherginsky, and S. D. Conzen. “Activation of the Glucocorticoid Receptor Is Associated with Poor Prognosis in Estrogen Receptor-Negative Breast Cancer”. In: *Cancer Research* 71.20 (Oct. 2011), pp. 6360–6370.

[265] Dilip Kumar, Judith Hosse, Christine von Toerne, et al. “JNK MAPK pathway regulates constitutive transcription of CCL5 by human NK cells through SP1.” In: *Journal of immunology (Baltimore, Md. : 1950)* 182.2 (Jan. 2009), pp. 1011–20.

[266] K Barton, N Muthusamy, C Fischer, et al. “The Ets-1 transcription factor is required for the development of natural killer cells in mice.” In: *Immunity* 9.4 (Oct. 1998), pp. 555–63.

[267] Shuzhen Liu, Jonathan Lachapelle, Samuel Leung, et al. “CD8+ lymphocyte infiltration is an independent favorable prognostic indicator in basal-like breast cancer”. In: *Breast Cancer Research* 14.2 (Apr. 2012), R48.

[268] Yan Mao, Qing Qu, Xiaosong Chen, et al. “The Prognostic Value of Tumor-Infiltrating Lymphocytes in Breast Cancer: A Systematic Review and Meta-Analysis”. In: *PLOS ONE* 11.4 (Apr. 2016). Ed. by Elda Tagliabue, e0152500.

[269] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. “Sparse inverse covariance estimation with the graphical lasso”. In: (2007).