Modelling Tumour Heterogeneity with Patient-Specific Networks

John Robert Salamon
BSc. Hons. (Biotechnology)

Thesis submitted to Flinders University
College of Medicine and Public Health
for the degree of Doctor of Philosophy

5 December 2022

Abstract

The molecular heterogeneity of cancers such as colorectal cancer (CRC) hinders the effectiveness of treatments. To overcome this heterogeneity and identify better therapeutic targets, we require patient-specific models which can more accurately predict the characteristics of individual tumours. Despite the availability of large scale patient-specific data from projects such as The Cancer Genome Atlas (TCGA), integrating data of this scale into a biologically meaningful model is a complex task. In this thesis, I investigated multiple patient-specific and network approaches to modelling tumour heterogeneity. I first took the approach of identifying patient-specific differentially expressed genes from TCGA transcriptomics data. From these, I was able to define prognostically relevant patient subgroups. I performed patient-specific pathway enrichment analysis and pathway level patient clustering, identifying novel patient clusters with significant differences in survival. Using the same patient-specific data, I combined transcriptomic and genomic data with protein-protein interaction (PPI) data to create patient-specific network models. I used the epidermal growth factor receptor (EGFR) PPI network, a network critical to the progression of CRC, to demonstrate this approach. I determined that while patient-specific network topology in the EGFR network was not directly linked to patient survival, it did differ significantly between patient subtypes. I developed SIFFIN, a novel tool to simulate the flow of biological information through these patient-specific networks, which predicted substantive alterations between patients. Finally, I explored tumour heterogeneity from a spatial perspective, aiming to develop novel network-based tools to integrate spatially-resolved patient data. I developed InsituNet, a tool for spatial transcriptomics analysis, enabling spatially-resolved analysis of tumour heterogeneity, and further adapted this tool to support spatial metabolomics data.

Acknowledgements

I would like to thank my supervisor, Professor David Lynn, for his guidance, supervision and patience during the entire course of this Ph.D. I would also like to thank Professor Lisa Butler for her feedback and encouragement. I am extremely grateful to all the members of the Lynn lab at SAHMRI both current and past, as well as the wider SAHMRI community, that I have had the great fortune to interact and collaborate with during my candidature. Finally I would like to thank all my family and friends for their love and support, especially Britt, Eli, and dad.

Funding

During the time of writing this thesis I was supported by the Australian Government Research Training Program Scholarship.

Declaration

I certify that this thesis:

1. does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any university; and

2. to the best of my knowledge and belief, does not contain any material previously published or written by another person except where due reference is made in the text.

John Salamon

Contents

1. Literature Review

1.1 Colorectal cancer in summary

1.1.1 A global health burden

As one of the most common forms of cancer globally with 1.8 million new cases and 881,000 mortalities estimated in 2018 (Bray *et al.,* 2018), colorectal cancer (CRC) represents a growing global health burden. CRC is the second most common form of cancer in women and third most common in men, making up 10.2% of all global cancer incidences and 9.2% of mortalities. The incidence and mortality of CRC varies by more than 10-fold globally, with greater than two-thirds of cases occurring in countries with a high or very high human development index (HDI) (Arnold *et al.,* 2017). While the incidence of CRC is relatively stable in most high-HDI countries, its global prevalence continues to accelerate, being strongly linked to rapid societal and economic changes that lead to a more "Westernised" diet and lifestyle (Maule & Merletti, 2012).

1.1.2 Demographic variations

CRC incidence and mortality rates vary across multiple different population groups. For instance, CRC is more prevalent and fatal in men than women in Australia (*Cancer in Australia* 2019), the US and UK (White *et al.,* 2018). Black people in the US have both the highest incidence rates of non-hereditary CRC, and are also more likely to die from CRC than any other group (Siegel *et al.,* 2017). Socio-economic factors which contribute to lower rates of screening are often cited as a driver of such discrepancies (Kwaan & Jones-Webb, 2018), with geographic analysis of CRC hotspots in the US identifying impoverished areas as having increased CRC incidence and mortality, a

pattern which mirrors the increased CRC rates in economically transitioning countries globally (Siegel *et al.,* 2015). In both the US and Australia, the incidence of CRC rises rapidly after the age of 50 (Siegel *et al.,* 2017), but due to the introduction of screening programs and awareness over the last two decades, incidence in over 50s has remained steady. In contrast, incidence rates within those aged <50 has been steadily increasing for reasons that remain unclear (Young *et al.,* 2015; Feletto *et al.,* 2019; Siegel *et al.,* 2019). These data highlight the need for early stage identification of CRC, as diagnosis at later stages results in significantly less effective treatment and poorer patient outcomes (Andrew *et al.,* 2018).

1.1.3 Anatomical distinctions

Cancers of the large intestine include colon cancer (CC) and rectal cancer (RC) which are together referred to as colorectal cancer (CRC). Despite this, proximal (right-sided, consisting of the cecum, ascending colon and transverse colon) colon cancers, distal (left-sided, consisting of the descending and sigmoid colon) colon cancers, and rectal cancers all tend to vary in prognoses (Baran *et al.,* 2018), surgical challenges (RC surgery tends to impose more risk for damage to surrounding tissues (Paschke *et al.,* 2018)), and molecular characteristics (the proximal and distal regions can be accurately classified based on their gene expression profiles (Glebov *et al.,* 2003)) to the point that some advocate for abandoning the term "colorectal cancer" entirely (Paschke *et al.,* 2018). Why these anatomical regions differ so much on a molecular level is not fully understood, but factors such as the difference in bacterial load, gut pH and exposure to nutrients and bile acids (Figure 1.1) have been put forward as hypotheses (Murphy *et al.,* 2011). The proximal and distal colon segments also have distinct embryological origins, with the proximal and distal regions originating from the midgut and hindgut respectively during development, potentially contributing to the molecular differences in proximal and distal CC (Bhatia *et al.,* 2020).

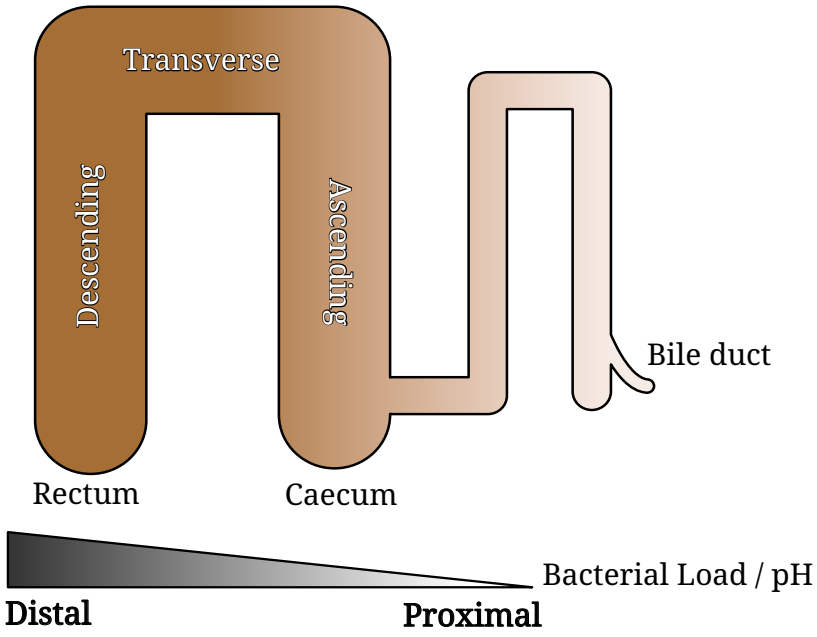


Figure 1.1: *Simplified diagram of the small and large intestine, adapted from Donaldson* et al., *2016. Increased pH and bacterial load towards the distal colon has been highlighted as a potential driver of the molecular differences between proximal and distal cancers.*

Notably, age at diagnosis has a strong effect on whether CRC presents as proximal or distal, with proximal (right sided) CRC more commonly being diagnosed at a later age (Mik *et al.,* 2017). Proximal tumours are more likely to be overlooked during screening, leading to delayed diagnosis and worsened overall prognosis for right-sided tumours (Hansen & Jess, 2012).

The vast majority (around 96%) of CRCs are *adenocarcinomas* (From Greek adēn, meaning 'gland'), developing in the intestinal epithelium from pre-cancerous polyps (adenomas). Besides having different origins, proximal and distal tumours often exhibit different histologies, with proximal tumours more frequently arising from sessile serrated adenomas, a flatter and far more difficult to detect type of adenoma than the more common tubular adenoma (Obuch *et al.,* 2015). Regardless of their origin, if detected, these polyps can be removed during screening via colonoscopy, making screening an extremely powerful preventative measure.

1.2 Molecular development and heterogeneity of CRC

1.2.1 Hypotheses to explain heterogeneity

Cancer cells differ substantially, even within the same tumour, in for example their size, proliferation rates, susceptibility to chemotherapy, and metastatic potential. Two distinct but not necessarily mutually exclusive hypotheses exist which explain this heterogeneity. The clonal evolution hypothesis, first proposed by Nowell (Nowell, 1976), is the result of understanding that cancers arise from successive genetic mutations. Clonal evolution posits that cancer is a product of the Darwinian scenario in which ongoing accumulation of somatic mutations in cells gives rise to subclones with a selective growth advantage within the tumour environment (i.e. better able to compete for resources or evade the immune system). Clonal evolution in tumours has now been well documented (Greaves & Maley, 2012) and the hypothesis further extended. For example, spatial analysis of tumours using whole-exome sequencing (Gerlinger *et al.,* 2012) revealed significant intra-tumour genetic heterogeneity, which led to an extension of clonal evolution in which multiple subclones with similar fitness may stably coexist in a tumour (Snuderl *et al.,* 2011).

Despite its success, the dominance of clonal evolution has been reduced in recent times by the emergence of the cancer stem cell (CSC) model, which proposes that tumour growth is due to a minority of tumourigenic cancer stem cells within a given tumour population, and that the functional heterogeneity between cells is as a result of their differentiation status (J. Peixoto & Lima, 2018). This model implicates stem cells found in intestinal crypts as the origin of CRC (Barker *et al.,* 2009). Originally thought of as a one-way hierarchy in which certain stem cells would act as progenitors to create differentiated cancer cell types, the modern CSC model now incorporates the plasticity of cancer cells by accepting that certain conditions may cause a reversal in differentiated cells causing them to once again become CSCs (Prasetyanti & Medema, 2017). As typical cancer chemotherapy treatments rely on the abnormally high growth rate of most tumour cells, the CSC model provides a compelling explanation for high recurrence rates - such treatments are unable to target the more slowly dividing cancer

stem cells (Khalek *et al.,* 2010). In CRC, certain biomarkers of CSCs have been found to be a positively correlated with patient survival in large cohorts (Badic *et al.,* 2020).

1.2.2 Tumour development in CRC

Molecular development in CRC was described in detail by the adenoma-carcinoma model of tumour progression proposed by Fearon & Vogelstein (Fearon & Vogelstein, 1990). This model describes a multi-step sequence of accumulative genetic aberrations that lead to two main molecular phenotypes of CRC: the chromosomal and microsatellite instability phenotypes. While most certainly an oversimplification of the true molecular nature of CRC, it is nonetheless a useful and clinically relevant model of carcinogenesis. This model falls neatly under Nowell's paradigm of clonal evolution, as it supposes that CRC development is driven mainly by monoclonal expansion within the primary tumour. In their model of CRC, Fearon and Vogelstein predicted that at least 7 distinct mutations were required for complete progression to carcinoma.

Recently, Fearon and Vogelstein's model was challenged by the "Big Bang" model of CRC tumourigenesis (Sottoriva *et al.,* 2015). Based on the observations that clonal selection is infrequent in advanced tumour stages and that spatial constraints in solid tumours limit selective forces, the Big Bang model suggests that tumours primarily grow as a single expansion following initial transformation, and that intra-tumour heterogeneity is primarily generated early during tumour growth. The predictions of this model align more closely to the CSC hypothesis, under which long-lived stem cell lineages are the primary drivers of tumourigenesis. Like the CSC and clonal evolution hypotheses, the Big Bang and Vogelstein models are not necessarily mutually exclusive, and regardless of the exact timeline of development, at least three common mutation pathways have been well described.

1.3 Pathways of CRC development

1.3.1 Common molecular events

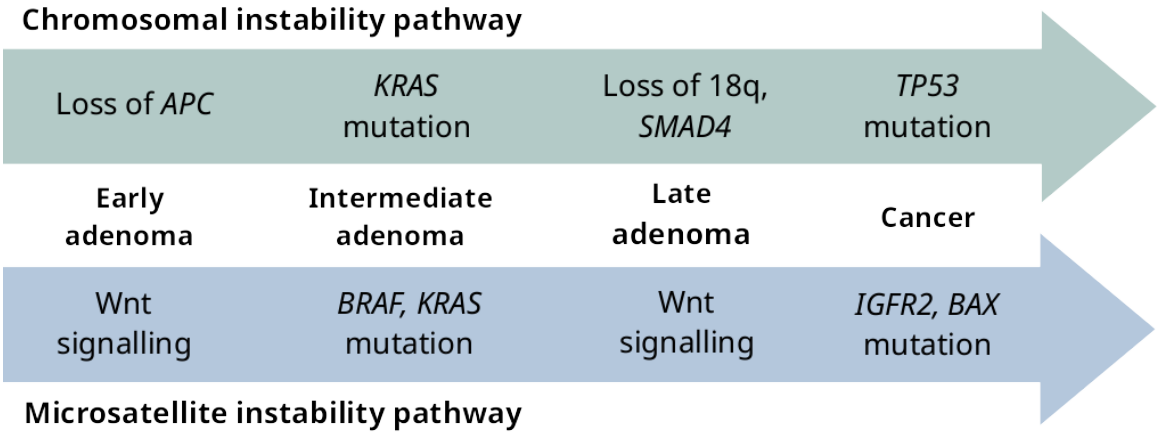


Figure 1.2: *Fearon and Vogelstein's multi-step model of CRC development, from adenoma to carcinoma, following two major pathways of genomic instability, chromosomal instability (CIN) and microsatellite instability (MSI), characterised by a defect in the DNA mismatch repair (MMR) system. Adapted from Walther* et al., *2009.*

In nearly all CRC cases (∼80%), the first event in the development from adenoma to carcinoma is inactivation of the adenomatous polyposis coli (*APC*) tumour suppressor gene on chromosome 5q (Fearnhead *et al.,* 2001). *APC* was first identified and characterised in the context of familial adenomatous polyposis (FAP), a heritable condition that involves germline mutations to *APC* and greatly increased rates of CRC (Groden *et al.,* 1991). APC has long been known as a key regulator of the Wnt signalling cascade through its role in the degradation of β-catenin (Polakis, 1997) and is integral to the regulation of many cellular functions such as cell adhesion and apoptosis. Inactivation of APC will lead to a build-up of β-catenin and constitutive activation of the Wnt pathway, subsequently causing transcriptional dysregulation of various proliferation-associated genes targeted by T-cell factor/lymphoid enhancer factor (TCF/LEF) transcription factors (Morin *et al.,* 1997; Cadigan & Waterman, 2012). APC is however a multi-functional protein and also features microtubule- and

end binding 1 (EB1)- binding domains at its C-terminus (Fearnhead *et al.,* 2001), with EB1 being a highly conserved protein that controls the plus ends of growing microtubules (Tirnauer & Bierer, 2000). APC and EB1 are thought to function together to regulate microtubule function and thus chromosome alignment during mitosis, implicating APC mutations not only in the initiation of tumorigenesis through activation of Wnt signalling, but also in driving chromosomal instability in CRC directly through disruption of the APC-EB1 interaction (Fodde *et al.,* 2001).

1.3.2 Chromosomal instability phenotype

The chromosomal instability (CIN) phenotype is both the most common and best understood molecular pathway of CRC development, sometimes referred to as the aneuploidy pathway (Upper portion of Figure 1.2). Over 70% of CRC tumours exhibit this phenotype to some degree (Pino & Chung, 2010). It is characterised by genetic instability on a chromosomal level, with widespread imbalances in chromosome number and loss of heterozygosity being common. Following the first step in carcinoma development (*APC* inactivation) *KRAS* mutation is typical, and subsequent mutations will accumulate until the inactivation of the *p53* tumour suppressor gene on chromosome 17p. At this point the transformation into carcinoma is complete (Figure 1.2, right). Frequently chromosome 18q will also be lost in CIN tumours, containing the aptly named Deleted in Colorectal Cancer (*DCC*) gene which likely has a role in tumour suppression (Nguyen & Duong, 2018).

1.3.3 Microsatellite instability phenotype

An alternative form of genomic instability found in around 20% of CRC tumours is the microsatellite instability (MSI) phenotype. Characterised by a defect in the DNA mismatch repair system leading to instability in long stretches of DNA microsatellites and global hypermutation, the MSI phenotype is most commonly seen in the proximal colon and often with more difficult to detect forms of polyp, such as the sessile serrated adenoma (which were originally not thought to be a precursor to CRC (Leggett & Whitehall, 2010)). Global hypermutation causes MSI tumours to exhibit a

high number of immunogenic mutations, and so these tumours will typically be found with very high T-cell infiltration due to an increased presence of neoantigens (Baran *et al.,* 2018). This immune infiltration makes MSI tumours particularly susceptible to immunotherapies such as anti PD-1 immune checkpoint inhibitor therapy (Le *et al.,* 2015), and is also associated with a reduced rate of metastasis (Giannakis *et al.,* 2016). MSI tumours are sometimes the result of hereditary defects to mismatch repair genes, i.e. hereditary nonpolyposis colorectal cancer (HNPCC) or Lynch syndrome (Steinke *et al.,* 2013), a condition that can also manifest as various other cancers, but most frequently as CRC.

1.3.4 CpG island methylator phenotype

A newer but increasingly accepted molecular subtype of CRC is the CpG island methylator phenotype (CIMP), described by Toyota *et al.* (Toyota *et al.,* 1999). CpG islands are genomic regions enriched for cytosine-phosphate-guanine (CpG) dinucleotides. Methylation of CpG islands within promoter regions causes transcriptional silencing, an important epigenetic mechanism for preserving genomic stability in normal tissues. Unlike the better established genomic instability phenotypes, CIMP is characterised by epigenetic modifications in the form of aberrant hypermethylation in specific CpG sites, resulting in transcriptional inactivation of tumour suppressors and other tumour-related genes (Mojarad *et al.,* 2013). CIMP is commonly observed within a subset of MSI-high tumours and shares some pathological characteristics with MSI such as proximal location and high frequency of *BRAF* mutations, however studies in both microsatellite stable (MSS) and MSI tumours have confirmed CIMP-high association with these characteristics independently of MSI (Wu & Bekaii-Saab, 2012; Weisenberger *et al.,* 2018).

1.4 Therapeutic approaches

1.4.1 EGFR targeted therapies

For late stage and metastatic CRC, KRAS-mutation status is an important biomarker for targeted treatment. Mutations in KRAS are found in around 38% of colorectal cancer cases (Oliveira *et al.,* 2004), and are predictive of failure of epidermal growth factor receptor (EGFR) targeted therapies. KRAS is a GTPase that operates in a switch-like manner for signal transduction early in the EGFR network, activating downstream effectors and ultimately regulating transcription through downstream transcription factors. Despite considerable research, KRAS has been described as essentially undruggable (Papke & Der, 2017). In CRC there is strong selection for KRAS mutation as a way of essentially bypassing EGFR-dependent signalling entirely. Despite mutant KRAS being an indicator for likely failure of anti-EGFR treatment, even patients with wild-type KRAS are still unresponsive in the majority of cases when undergoing anti-EGFR treatment (Amado *et al.,* 2008). This inefficacy might reasonably be attributed to metastatic subpopulations developing an independent mutant KRAS genotype, however a comprehensive study of KRAS mutations in primary and metastatic tumours in colorectal cancer patients found that KRAS mutation status was discordant between the two in only 2% of cases (Knijn *et al.,* 2011). This finding is supportive of the notion that KRAS mutation is an early driver of CRC progression – as well as highlighting the fact that it is difficult to point to a single mutation as being the sole cause of a targeted treatment failing.

In metastatic CRC some of the most frequently prescribed treatments are EGFR inhibitors, either tyrosine kinase inhibitors (e.g. erlotinib, gefitinib) – which bind to the tyrosine kinase (TK) domain on the receptor, blocking EGFR activity, or monoclonal antibodies (e.g. cetuximab, ecitumumab), which bind to the extracellular domain and prevent EGF or other ligands from binding (Yarden & Pines, 2012). EGFR signalling plays a critical role in tumourigenesis, mediating many processes including transcription, cell cycle (especially through production of cyclin D due to the transcription factor Myc) and proliferation through the RAS/RAF/MEK/ERK

pathway, and metabolism, growth and apoptosis through PI3K/AKT/mTOR (Wee & Z. Wang, 2017), although it has been suggested that some of the efficacy observed from anti-EGFR treatment may in fact be due to indirect effects which alter the immune microenvironment (Giordano *et al.,* 2019).

1.4.2 Angiogenesis targeted therapies

Angiogenesis (blood vessel growth) is dysregulated in CRC development and progression (as tumours require excessive resources to continue to grow). Loss of angiogenesis equilibrium is one of the hallmarks of cancer (Hanahan & Weinberg, 2000), and is regulated to a large degree by vascular endothelial growth factor (VEGF) signalling. VEGF / VEGF receptor (VEGFR) targeted treatments are therefore effective anti-cancer treatments, especially in metastatic CRC (Battaglin *et al.,* 2018). There are five VEGF family members, with VEGF-A being the most important in tumour angiogenesis. VEGF-A binds primarily to VEGFR-2, one of three VEGFR-family receptor tyrosine kinases. VEGF signalling in normal and cancer cells is upregulated due to hypoxia via the hypoxia inducible transcription factors HIF-1 and HIF-2. The first approved treatment targeting this pathway was bevacizumab in 2004, a monoclonal antibody which neutralises the VEGF-A ligand. Other agents which have since been approved include regorafenib, aflibercept, and ramucirumab, all of which function similarly to anti-EGFR treatments, i.e. inhibition strategies which target either the TK or extracellular domains of the involved receptor.

1.4.3 Early and late stage therapies

By far the most important molecular factor for prognosis in early-stage CRC is MSI status (Punt *et al.,* 2017), which is associated with greatly improved outcomes and low risk of recurrence following surgical resectioning. MSI status is used to determine whether resectioning should be accompanied by adjuvant chemotherapy (as surgical interventions for stage II MSI CRC have excellent prognosis without chemotherapy) (Kawakami *et al.,* 2015), and also to predict tumour vulnerability to anti-PD1 immunotherapy (Punt *et al.,* 2017).

The standard treatment for late stage CRC is chemotherapy. Fluorouracil (5-FU) is commonly used in combination with folinic acid, yielding a median survival of approximately 6 months, with response rates to the therapy of around 20%. Use of these traditional therapies in parallel with sequential administration of other chemotherapeutic drugs such as oxaliplatin, leucovorin, and irinotecan has led to improved median survival of around 20 months (Cremolini *et al.,* 2015). Use of these chemotherapies in conjunction with EGFR and VEGF inhibitors improve these figures yet further to around 30 months (Sánchez-Gundín *et al.,* 2018). Despite targeted treatment advances, metastatic CRC is still a devastating disease, with a 5-year survival rate of around 14% (Street, 2020).

1.5 Clinical and molecular subtyping

1.5.1 Tumour, Node, Metastasis

The CIN, MSI, and CIMP forms of genomic and epigenomic instability may be present in varying degrees in a particular tumour, and are sometimes used to assist in prognosis. MSI status is now frequently used for patient stratification due to the improved immunotherapy response for such tumours (de Vries *et al.,* 2016) and is also used to determine whether adjuvant chemotherapy is required after surgical resectioning. In practice however, testing for MSI or CIMP status is used in conjunction with the tumour-node-metastasis (TNM) staging system, probably the most routine clinically used prognostic method for classifying patients. TNM covers three primary factors:

- T: Tumour, relating to the size and extent of the primary tumour. In CRC, this primarily refers to how deeply the tumour has grown into the lining of the bowel.
- N: Node, relating to the extent of spread to regional lymph nodes.
- M: Metastasis, simply describing whether there are distant metastases.

These three factors can be used extremely reliably to stratify patients by their predicted survival, and to some extent by which treatments should be attempted. A

simplified prognostic stage is commonly derived from TNM ranging from stages I-IV, (technically stage 0 for carcinoma *in situ*). The presence of any metastasis will automatically place a patient in stage IV, while the middle stages effectively describe increasing primary tumour growth and spread to lymph nodes. The intent of the prognostic stage is to have fairly homogeneous survival probabilities for patients within each stage, with each increasing stage representing a worsening of prognosis. The TNM classification manuals from the American Joint Committee on Cancer (Edge *et al.,* 2010) and Union for International Cancer Control (Sobin *et al.,* 2011) are frequently updated.

For some cancers such as breast cancer, the most recent 8th edition of TNM (effective as of 2017) includes molecular biomarkers including *HER2* as part of the Oncotype DX multigene panel (J. Koh & M. J. Kim, 2019). Although assays that allow better identification of stage II-III CRC patients with a high post-surgery recurrence probability do exist in the form of the Coloprint 18-gene assay (I. B. Tan & P. Tan, 2011) and the Oncotype DX 12-gene assay for colon cancer (You *et al.,* 2015), neither of these have yet been integrated into TNM.

One proposed molecular addition to TNM for CRC is immunoscore (Ros-Martínez *et al.,* 2020), a measure of T-cell infiltration that would better capture the influence of the tumour microenvironment. Immunoscore increases with the density of lymphocyte populations, and has been found to correlate very well with improved prognosis (Pagès *et al.,* 2018). Currently however, molecular considerations in CRC TNM are quite limited. Fortunately, in recent years many large-scale projects have been able to collect vast quantities of personalised molecular tumour data, leading to significant advancements being made in molecular-based CRC subtyping.

1.5.2 The Cancer Genome Atlas

Efforts to better stratify and classify patients have been greatly aided by large-scale projects such as The Cancer Genome Atlas (TCGA), a project of the US National Cancer Institute and National Human Genome Research Institute. TCGA is a systematic analysis of multiple types of human tumours that provides DNA, RNA, protein

and epigenetic aberration data on an individual patient level (Weinstein *et al.,* 2013). The publicly available data collected from this project have proved invaluable for research into the molecular differences between different cancer types, as well as between-patient differences, especially in CRC (The Cancer Genome Atlas Network, 2012). TCGA studies have made publicly available over 2 petabytes of data across 33 different cancer types, including from 633 colorectal adenocarcinoma patients. These data include in-depth information on individual patients, such as clinical diagnoses, outcomes, sequencing data, biopsy information, and more. Most cases also include results from RNA-sequencing experiments on primary tumour tissue, and some also include matched normal tissue samples.

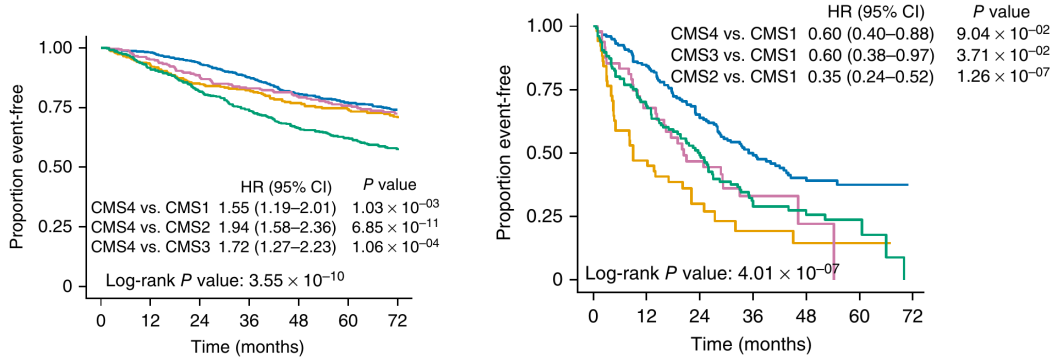
Further extensions and bioinformatic analyses of the TCGA datasets have provided a wealth of supplementary information that further expands the possibilities for using TCGA data, from proteomic studies from the Clinical Proteomic Tumor Analysis Consortium (CPTAC) (B. Zhang *et al.,* 2014) to comprehensive analysis of alternative splicing within TCGA cohorts (Kahles *et al.,* 2018). Recently, TCGA and the International Cancer Genome Consortium (ICGC) have formed the Pan-Cancer Analysis of Whole Genomes (PCAWG) Consortium, which has released an integrative analysis of 2,658 cancer genomes with matched normal samples across 38 tumour types (PCAWG Consortium, 2020), providing detailed insights into the molecular differences between cancers.

1.5.3 Consensus molecular subtypes

The scale of transcriptomics data made available by projects such as TCGA enabled the Colorectal Cancer Subtyping Consortium (CRCSC) to create a unified classification system for CRC (Guinney *et al.,* 2015). This system is based on the transcriptomic analysis of more than 4000 individual tumour samples. Six independently developed algorithms for classifying patients based on transcriptomic data, totalling 27 separate groupings, were combined using a network similarity approach to create a new classification system that defined four significant subtypes of CRC, the Consensus Molecular Subtypes (CMS):

- CMS1: MSI-Immune (Hypermutated, exhibiting high-MSI, high-CIMP, and immune infiltration)
- CMS2: Canonical (exhibiting high-CIN with Wnt and Myc activation)
- CMS3: Metabolic (low-CIN, low-CIMP, KRAS-mutated, epithelial signature)
- CMS4: Mesenchymal (high-CIN, enrichment of epithelial–mesenchymal transition (EMT), stromal infiltration, TGFβ activation)

These transcriptomic subtypes build on the existing understanding of genomic and epigenomic subtypes, and have important implications for patient stratification. They also represent one of the first real efforts to perform molecular stratification of patients with non-MSI CRC tumours. CMS1 tumours have a distinctive pattern of hypermutation and hypermethylation, (being MSI-high and CIMP-high), as well as over-representation of BRAFV600E mutations. CMS1 is the most immunologically active, and tends to exhibit extensive immune cell infiltration, including cytotoxic T lymphocytes, CD3+ T helper cells, and NK cells (Dienstmann *et al.,* 2017). Patients with CMS1 tumours typically have good general prognosis with low probability of relapse, however if relapse occurs prognosis is poor (Figure 1.3b). In comparison, CMS2, 3 and 4 can be well described via Vogelstein's CIN phenotype. CMS3 and CMS4 effectively branch off from CMS2, with CMS3 gaining KRAS mutations, consistently lower copy number alterations, and metabolic reprogramming, versus CMS4's upregulation of TGFβ and EMT. CMS2 samples have upregulation of Wnt and Myc targets, with higher expression of EGFR, ERBB2 and IGF2 (Guinney *et al.,* 2015). CMS4 tumours have a particularly strong angiogenic influence on the tumour micorenvironment, with signalling activation derived from stromal cell infiltration from adjacent cancer-associated fibroblasts (Dunne *et al.,* 2016). In terms of patient outcomes, CMS4 patients have the poorest overall prognosis (Figure 1.3a). Notably, CMS4 tumours have an upregulation of genes involved in the epithelial–mesenchymal transition (EMT) (Guinney *et al.,* 2015). Epithelial cells which undergo the EMT gain resistance to apoptosis and enhanced migratory capacity, greatly increasing the likelihood of metastatic spread (Kalluri & Weinberg, 2009).



(a) *Aggregated CMS survival curves* **(b)** *Post-relapse CMS survival curves*

Figure 1.3: *Kaplan-Meier survival analysis of patients in the cohort used by Guinney* et al. *to define CMS groups (Guinney* et al., *2015). Reproduced with permission from Springer Nature. CMS1 (yellow), CMS2 (blue), CMS3 (pink) and CMS4 (green).* **a)** *Aggregated cohort (n = 2,129), displaying significant differences only for CMS4.* **b)** *Post-relapse survival (n = 405), displaying multiple significant survival differences.*

Currently, the CMS classification system has not been fully translated into clinical use. While CMS-based patient stratification represents the best molecular classification in CRC to date, it still limited in than only four subtypes are discernible - still very far off the promise of personalised or precision medicine. Indeed, ∼13% of samples used to create the CMS system were unable to be classified by it. While efforts to improve the CMS classifications are ongoing (Menter *et al.*, 2019), the extreme molecular heterogeneity of CRC calls for alternative strategies which will enable more patient-specific and clinically relevant classifications.

1.6 Consolidation and integration of multi-omics data

1.6.1 Data dimensionality reduction

Large scale projects such as TCGA have generated an unprecedented depth and breadth of biological data, covering thousands of individual patients across multiple omics sources. To analyse data of such scale, univariate methods such as ANOVA, linear models or t-tests that are the statistical mainstay of conventional biology fall

short, being unable to consider the multi-level relationships that exist between these diverse forms of biological data. Some methods, such as genome-wide association studies (GWAS), have been successfully used to identify thousands of genetic loci associated with diseases. However, the associations these studies provide have proven highly difficult to translate into functionally relevant biology (Tam *et al.,* 2019).

Due to the increased number of data modalities available, a more complete characterisation of the molecular status of individual tumours can be obtained. For example while transcriptomic data provides a snapshot of gene expression, the regulation of this expression is dependent epigenetic changes such as DNA methylation, which was also profiled for TCGA cohorts. It is unsurprising then that multi-omics integration of TCGA datasets has increasingly been found to outperform methods which rely on a single omics type for patient stratification (Boehm *et al.,* 2022). This has been demonstrated for example by application of methods like the Cox proportional hazards (CPH) model in glioblastoma (Network, 2015). More data is not necessarily always better however, with adding additional data to a CPH model sometimes actually reducing stratification performance (Z. Huang *et al.,* 2019). For this reason, research into methods that properly utilise and integrate these data are called for. Given the typical complexity of these omics data, many approaches focus on simplifying or reducing complexity and noise.

Dimensionality reduction is a common multivariate mathematical approach to reducing complexity in high-dimensional omics datasets. Dimensionality reduction methods can take large, complex data (such as the expression of tens of thousands of genes) and reduce them down to two or three dimensions that capture the majority of variance present. These methods are widely used for visualisation and analysis, and include Principal Components Analysis (PCA), multidimensional scaling (MDS), t-SNE (Van Der Maaten & Hinton, 2008), and UMAP (McInnes *et al.,* 2018). Typically dimensionality reduction methods are best applied to a single experimental dataset, and if not, statistical methods to overcome technical differences between experiments must be laboriously applied (i.e. batch effect correction) to integrate multiple independent datasets. Specialised software exists for this express purpose, for example ComBat (Johnson *et al.,* 2007). Another important consideration of such analyses is

how measurements of the same samples collected from different platforms should be combined.

One tool addressing both inter-sample and inter-omic factors is mixOmics (Rohart *et al.,* 2017), a framework for integration of multi-omics that enables statistical combination of heterogeneous omics datasets via data dimensionality reduction methods, mainly for exploration and visualisation. MixOmics applies multiple techniques in order to classify samples, including unsupervised methods such as PCA, as well as supervised learning methods such as partial least squares regression discriminant analysis (PLS-DA).

Dimensionality reduction tools are powerful, but may be insufficient for classification due to their assumption of feature independence. For example, gene expression may be quantified on a gene-by-gene basis using RNA-seq, but in reality individual genes very rarely, if ever, operate in an isolated way from all other genes. This means that true biological signal may be lost among the noise of individual genes when dimensionality reduction approaches are used as a basis for classification. This insight, that genes are typically part of a larger process or pathway, has led to the introduction of pathway-based methods for managing cellular complexity.

1.6.2 Classification and feature identification with PLS-DA

Performing classification of unknown samples or identifying key features that differentiate classes (e.g., genes in genomics data) is a common task in the analysis of omics data. Partial least squares-discriminant analysis (PLS-DA) is a method often recommended for such analyses due to its ability to perform well with high-dimensionality data. In terms of its function, PLS-DA is sometimes described as a supervised version of principal component analysis (PCA) (Ruiz-Perez & Narasimhan, 2017). PLS-DA has been included in software packages for omics data analysis such as mixOmics (Rohart *et al.,* 2017), and is frequently employed in the analysis of genomics, metabolomics, and other data types. Despite widespread use, PLS-DA has received some criticism due to the ease with its models can be overfitted, especially when sample sizes are small, or poor cross-validation methodologies are employed (Rodríguez-Pérez *et al.,* 2018). In a

recent systematic analysis of cross-validation methods, Rodríguez-Pérez *et al.* found that the leave one out (LOO) method for cross-validation which is frequently used produces the worst overfitting, with bootstrap procedures being the most accurate. Bootstrap methods are the most computationally expensive, with LOO being the least, as essentially the worst-case scenario K-Fold validation, in which the training data is divided into k equally sized partitions. For example, mixOmics recommends 5-10 fold. However, even when rigourous cross-validation is employed, if PLS-DA models are trained using small sample sizes caution must be made when making broader predictions beyond the study sample.

1.6.3 Machine learning approaches to data integration

Increasingly, multi-omics integration of TCGA data by more advanced machine learning approaches has been demonstrated (Boehm *et al.,* 2022). Two main architectural approaches for integrating multi-modal data in machine learning are early and late fusion. Cancer Integration via Multikernel LeaRning (CIMLR) (Ramazzotti *et al.,* 2018) is a recent example of an early fusion architecture which combines different data modalities before beginning any model training. A key advantage of early fusion is the ability to learn similarities across different data types simultaneously. CIMLR has been demonstrated to be effective for tumour subtyping across multiple cancer types using TCGA datasets. Late fusion methods in comparison perform individual modelling of different data types, then aggregate these in a later step. Late fusion in comparison is more common for machine learning approaches making use of extremely heterogeneous data types. For example, Shao *et al.* (Shao *et al.,* 2020) used convolutional neural networks to extract features from histology images, and later combined this with transcriptomic data in a late fusion approach for early-stage survival outcome prediction. Although machine learning methods for data integration show promise, one of the biggest challenges facing this research is data availability. For such approaches to be effective, collection and careful curation of multiple data types across multiple patients is required at scale.

1.7 Pathway based analysis

1.7.1 The benefit of a pathway approach

From cell surface receptor molecules to downstream transcription factors within the nucleus, genes and proteins contribute to phenotypic effects through various interlinked molecular signalling cascades, or pathways. For example, the canonical Wnt pathway is involved in a diverse range of biological functions including cell homeostasis and repair, and is nearly always hyperactivated during CRC development (Schatoff *et al.,* 2017). One way of determining whether genes, mutations and proteins may be acting in concert to produce a particular phenotype or disease is to test whether they are members of the same molecular pathways. By "zooming out" to the level of pathway dysregulation rather than focusing too closely on individual genes or proteins, it is possible to find commonalities in otherwise impossibly heterogeneous data. A great variety of methods have been proposed for conducting pathway analysis, however one particularly useful organisation of these methods originated with Khatri *et al.* (Khatri *et al.,* 2012), who grouped pathway approaches into three main classes: over-representation analysis (ORA), functional class sorting (FCS) and pathway topology (PT).

1.7.2 Pathway databases

A prerequisite for any pathway-based analysis is prior knowledge of pathway structure and organisation. As such, significant effort has gone into sorting molecules into separate signalling pathways. For example, at the time of writing the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa & Goto, 2000) contains 537 manually curated pathway maps representing specific biological systems and functions. Other examples of pathway databases include WikiPathways (Pico *et al.,* 2008), a community-maintained pathway resource; Reactome (Jassal *et al.,* 2020), an open source database containing 2,423 human signalling pathways; and Gene Ontology (Ashburner *et al.,* 2000), an initiative to develop a controlled vocabulary for genes and gene products.

Unlike the former databases, Gene Ontology is not strictly a pathway database, rather it consists of three separate annotation hierarchies in which related terms are linked in a directed graph. These hierarchies can be flattened to become equivalent to other pathway annotations. To run pathway enrichment methods, a single database is typically chosen. However, these databases differ in the number of pathways they contain and the size of these pathways (in terms of the level of detail and what types of interactions are incorporated), meaning choice of database can strongly influence the results. To address this concern, several integrative meta-databases have been created, for example Pathway Commons (Rodchenkov *et al.,* 2020) and MSigDB (Liberzon *et al.,* 2015), which consolidate multiple pathway databases in a single resource. Other work has gone further and attempted to integrate pathway databases such that analogous pathways across different databases are combined to further reduce biases when conducting pathway enrichment (as such methods assume that pathways are independent entities) (Mubeen *et al.,* 2019). Despite such efforts, biologically meaningful integration of data across multiple databases remains a significant challenge for the effective use of pathway databases, a problem which can only really be addressed by collaborative annotation efforts between different projects.

1.7.3 Over-representation analysis

To utilise pathway knowledge for analysis and interpretation of omics data, the usual approach is to perform statistical enrichment analysis, i.e. pathway enrichment analysis. Pathways in this context are sets of genes that describe particular biological processes or functions, and typically incorporate some additional information about pathway structure or interactions between items. While more advanced methods incorporate this additional information, at their simplest pathways may be reduced to simple sets of genes to which enrichment analysis methods may be applied. Perhaps the simplest enrichment approach (in terms of ease of implementation) is over-representation analysis (ORA). Given a set of pathway genes and a set of genes that are over-represented in a sample, an ORA will examine the gene overlap between these sets, and using a statistical test such as the hypergeometric test, assign a statistical significance to the overlap. When ORA is used to detect enrichment across many different pathway gene

sets, statistical adjustment for multiple testing such as Bonferroni correction should usually be applied to the results.

ORA has some notable limitations, such as assuming that pathways are all independent, only making use of a subset of genes determined to be differentially expressed, discarding any genes below an arbitrary pre-defined threshold (losing any information on the relative strength of signals), and being sensitive to gene background. The null hypothesis of an ORA may be described as "competitive" (i.e. genes within the pathway do not appear more often than genes outside of the pathway) due to the fact it depends on not only genes within the query set, but also on the genes outside of the set (the background) (Goeman & Bühlmann, 2007). It is a common mistake to provide an inappropriate background set of genes, and often tools will not even state what background is being used, leading to potentially very misleading results (Timmons *et al.,* 2015). Despite its limitations, ORA is still widespread due to its simplicity of implementation and use. Many web-based pathway tools provide ORA functions, for example Enrichr (Kuleshov *et al.,* 2016) or g:Profiler (Reimand *et al.,* 2007).

1.7.4 Functional class scoring

To address some of the issues of ORA, functional class scoring (FCS) pathway methods were developed. FCS posits that coordinated changes in large sets of related genes may have significant effects. FCS methods still suffer from the fact that all gene sets are analysed independently, however unlike ORA, FCS does not require setting an arbitrary threshold to determine differential gene expression, but rather assigns an enrichment score for each pathway based on every gene in the sample. Multiple methods using the core FCS concept have been proposed including PLAGE (Pathway Level Analysis of Gene Expression) (Tomfohr *et al.,* 2005) and GSEA (Gene set enrichment analysis) (A. Subramanian *et al.,* 2005), probably the most widespread FCS approach. GSEA creates a list of every gene ranked by expression, then determines whether a set of pathway genes is distributed at the top or the bottom of the ranked list (leading to a high enrichment score), or randomly throughout (low score). A

significance level is then assigned using a permutation procedure, and adjusted for multiple testing.

1.7.5 Patient-specific FCS

GSEA is in effect a supervised approach - one must provide case/control labels (e.g. tumour/normal). In contrast to the supervised binary case/control GSEA, patient-specific enrichment FCS methods must generate sample-wise enrichment scores. Unsupervised FCS methods in this patient-specific category include gene set variation analysis (GSVA) (Hänzelmann *et al.,* 2013), single sample GSEA as described by Barbie *et al.,* 2009, PLAGE (Tomfohr *et al.,* 2005), and the Z-score method described by E. Lee *et al.,* 2008. More recent FCS methods include Singscore (Foroutan *et al.,* 2018), a computationally simple method with easily interpretable percentile rank scores; LEGO (Dong *et al.,* 2016), a method which also incorporates network information in the form of weights in its scoring, and EGSEA (Alhamdoosh *et al.,* 2017), an ensemble method which combines 12 other FCS algorithms. These patient-specific methods are less popular as they do not produce a result that can immediately be used, their outputs of patient-specific pathway enrichment scores are typically used as the input for further analysis, e.g. for fitting to survival models.

GSVA in particular provides single-sample scores for each query gene set using a random-walk approach which is weighted based on the rank of each gene belonging to the gene set in an individual sample. The result is a single score which gives a measure of how a particular gene set or pathway is behaving within a sample, relative to other samples in the cohort. Conventional statistical methods can then be applied to determine the statistical significance of any change in these scores, as GSVA does not attempt any statistical inference itself. The unsupervised approach used by GSVA is advantageous in that no labels are required (e.g. tumour versus normal), however it also has downsides in terms of requiring at least 10 samples to function correctly. GSVA has two options for calculating scores, either the maximum deviation from zero, or the difference between positive and negative scores for each pathway (such that higher scores would only occur as the result of coordinated changes in expression

within a particular pathway). Which approach is preferred depends on whether the pathway database used has divided its gene sets into "up" and "down", as is the case for MSigDB (Liberzon *et al.,* 2015).

1.7.6 Pathway topology

Yet more sophisticated pathway approaches make use of the network topology of pathways, known as pathway topology (PT) methods. Pathway topology methods more fully utilise pathway information than simple enrichment analyses. ORA and FCS consider only whether sets of genes are involved in pathways, not where these genes exist in the pathway and which other genes they interact with. For example, signalling pathway impact analysis (SPIA) (Tarca *et al.,* 2009) which operates on gene expression data, takes into account the hierarchical position of pathway genes when calculating pathway scores, such that genes that occur "upstream" in a pathway and may cause a larger impact to the pathway are prioritised. Another example is network-based gene set analysis (NetGSA) (Shojaie & Michailidis, 2010; Ma *et al.,* 2016) which computes enrichment scores using a latent variable model of the underlying pathway network. NetGSA may also be used for complex experimental designs beyond the binary tumour vs. normal conditions which are assumed by most methods, such as multiple time point experiments. A systematic comparison of multiple pathway topology based methods (Ma *et al.,* 2019) found that while large pathways tend to perform equally well across methods (as is the case for genomic data), methods that take into consideration topology including NetGSA exhibit superior performance for smaller pathways, as is often the case for metabolomics data. A similar comparison of pathway topology methods by Ihnatova *et al.* (Ihnatova *et al.,* 2018) found that the influence of pathway size typically outweighed that of pathway topology, owing to the higher number of significant DE genes identified in larger pathways. Notably, they found that multivariable methods which work with complete lists of all tested genes, such as Clipper (Martini *et al.,* 2013) have increased sensitivity, identifying more significantly impacted pathways when there are only subtle changes in differentially expressed genes between conditions. The authors suggest different methods which may be appropriate under different conditions, with the strongest recommendation

being for multivariable approaches being used for small sample size, subtle expression change datasets, aand univariable methods for large scale datasets with large changes in gene expression.

1.7.7 Patient-specific PT

In most of the PT methods mentioned so far, the inputs are from a single omics type (generally gene expression), and the outputs are pathway enrichment scores between some set of binary classes (e.g. tumour versus normal samples). For understanding which pathways are enriched in an individual patient sample, more advanced patient-specific pathway approaches must be used. Using an N-of-one for pathway activity inference presents a technical challenge that has been addressed by fewer tools. One method that is capable of integrating multiple levels of omics data into a sample-wise enrichment score is PARADIGM (Vaske *et al.,* 2010), a tool that further extends pathway topology to multiple omics levels with a factor graph approach. In PARADIGM, graph variables represent the state of cellular entities (genes, proteins or complexes) with respect to a control or normal level. PARADIGM outputs an integrated pathway activity score, which is a patient-specific estimation of pathway activity (Figure 1.4).

Figure removed due to copyright restrictions.

Figure 1.4: *An overview of the PARADIGM method (Vaske* et al., *2010). Structural pathway information is combined with genomic data to provide patient-specific measures of pathway activity.*

Analysis using PARADIGM can suffer when the pathway of interest is not well understood or incomplete, however. A different method that is more robust to incomplete data is Pathifier (Drier *et al.,* 2013), which takes the approach of creating a "principal curve" that captures the variation of data within each pathway, then measuring each sample's distance from this curve. Analysing expression data from Sheffer *et al.* (Sheffer *et al.,* 2009), Pathifier was used to identify two pathways, CXCR3-mediated signalling and oxidative phosphorylation, which were significantly associated with survival in CRC. A more recent example of patient-specific pathway based approaches is Pathway RespOnsive GENes (PROGENy) (Schubert *et al.,* 2018), a method that incorporates publicly available drug perturbation experiments to better map the effect of post-translational modifications and accurately infer pathway activity from gene expression data under specific conditions.

Increasingly, the trend in large-scale data analysis in the multi-omics space is towards an integrated, systems approach. Multi-omics exploration and visualisation is now possible using online data portals such as cBioPortal (Cerami *et al.,* 2012), UCSC Xena (Goldman *et al.,* 2020), the Genomic Data Commons (Grossman *et al.,* 2016), and others (I. Subramanian *et al.,* 2020). Such portals consolidate and present multi-omics data in a user friendly way. Some of these platforms also enable patient-specific pathway analysis, for example UCSC Xena provides an interface for PARADIGM.

Pathway-based methods have proved extremely effective in managing the complexity of multi-omics data. They allow quick summation of complex data (e.g. RNA-sequencing) in terms of pre-existing knowledge, and may even be applied in a patient-specific manner. Pathway approaches however are limited in terms of their pre-existing biases, making pathway-based methods less suited for novel computational predictions. However, their reliance on existing knowledge is also what makes pathway approaches so intuitive, as results may immediately be linked into the context of existing literature. In many of the more advanced pathway approaches such as PARADIGM, these pathway analyses begin to focus more upon network-based analysis. Networks represent an extremely powerful tool for analysis and visualisation of multi-omics datasets, and have the potential to overcome many of the limitations of pathway approaches.

1.8 Network and systems biology

Biological systems are organised in a hierarchical fashion, with each level resulting in ever increasing complexity of the whole organism. Individual molecules combine to form genes, proteins and metabolites, which interact in diverse ways to form subcellular structures, processes and pathways (Ladiges *et al.,* 2010). These pathways are regulated by complex webs of interacting molecules to perform functions like cell growth, replication, and apoptosis. Descriptions of these multi-level relationships have been used to create hierarchically linked gene sets, for example gene ontology (GO) (Ashburner *et al.,* 2000). Groups of individual cells acting synergistically make up the tissues comprising the macro-scale organs of the body, which in total consists of trillions of individual cells (Bianconi *et al.,* 2013). The study of molecular biology is therefore fundamentally an attempt to decode these systems of enormous complexity into meaningful information. The dominant approach to this problem has historically been reductionism. The reductionist approach to understand these complex systems is to reduce them into the smallest still-functional component such as a single isolated protein or gene. While reductionism has been extremely effective at describing these individual elements, it is difficult to apply such methodology to diseases that emerge as a result of the complexity of entire systems, rather than from a fault in an individual component (Regenmortel, 2004). The obvious example of such a disease is cancer, the hallmarks of which stem from the countless different ways in which the cell's own molecular machinery can be dysregulated (Hanahan & Weinberg, 2000).

Owing to the sheer quantity of biological data now available due to technological advances such as high throughput sequencing, it is now possible to take a systems approach to studying biology. Under a systems approach, rather than studying isolated components, the structure of entire biological systems is considered. Molecular components are conceptualised in terms of their relationships to other components, i.e. with which other entities do they interact, and how do networks of interacting components together orchestrate cellular functions? When taking a systems approach, some of the fine detail of reductionist biology is lost, as a trade-off is always made between scale and detail (Bornholdt, 2005). The level of this simplification depends

upon the research at hand, e.g. to understand the dynamics of a particular cellular process a model on the scale of atoms may be required, but to understand the function of a complex disease may require an organism scale model. Due to the extreme variability of scale and data in systems biology, network models are a particularly attractive abstraction, as they can model the wiring of interactions between molecules, proteins, cells, or even phenotypes at any scale (Koyutürk *et al.,* 2012).

Networks can be defined from known molecular interactions, such as protein-protein interactions or gene co-expression, such that network topology reflects biology. Defining networks in this manner causes the structure to become meaningful, as vertices which are connected are also likely functionally linked, a phenomenon which has been observed in multiple types of molecular interaction (Barabási *et al.,* 2011). These networks provide a framework for investigating biological function, either through simply organising and annotating groups of interacting components, or through more advanced graph theory applications (Koyutürk *et al.,* 2012). The full network of interacting molecules within the cell is often referred to as the *interactome*.

1.8.1 Network and graph properties

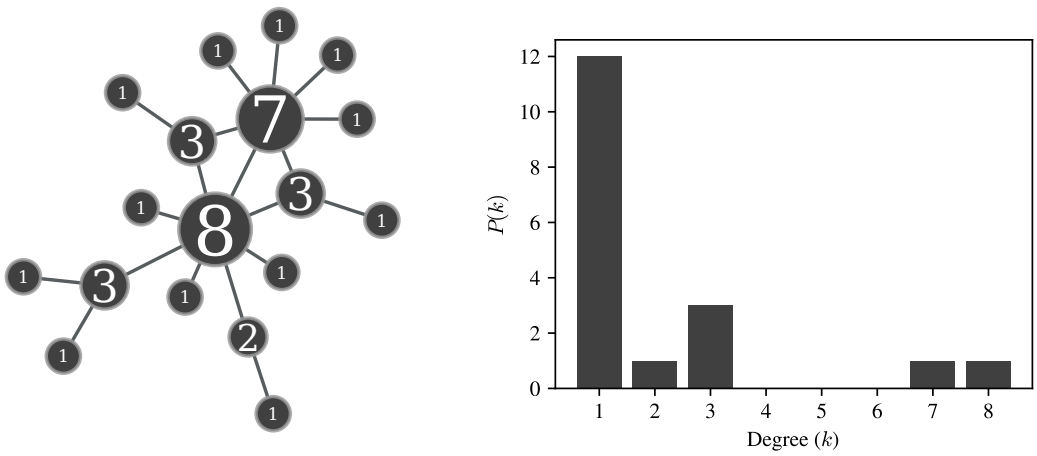
Networks are versatile tools for analysing complex systems. Network theory is effectively an applied subset of mathematical graph theory. A graph consists of a set of vertices V (otherwise known as nodes), which are connected by a set of edges E (otherwise known as links). In a network, vertices typically represent a discrete object or concept, and edges signify a relationship between vertices. Networks are described as either directed or undirected, depending on whether the relationship is symmetric or not. In the case of PPI networks, vertices are proteins, and edges represent a binary PPI. Although biological signalling often occurs with directionality, most PPI networks remain undirected as the high throughput experimental techniques used to determine them cannot detect this.

The defining characteristics of networks across diverse biological, sociological and technological fields have found to be more similar than not, allowing knowledge from better understood systems to be re-applied in the lesser known (Strogatz, 2001).

Despite the diversity and complexity of networks found in nature and society, a key insight into the power of network theory is that the topological characteristics of all networks may be described by relatively simple mathematical principles.

Degree and degree distribution

Complex networks were for many years modelled using Erdos and Rényi's random model (Erdős & Rényi, 1960), assuming that vertices are connected at random, such that the number of connections per vertex follows a Poisson distribution. This statistic (connections per vertex, or number of neighbours), is the *degree* of a vertex. It was found however that many real-world networks, such as the world-wide web or PPI networks, do not follow a Poisson degree distribution. Rather, the degree of these networks tends to approximate a power law (Barabási & Oltvai, 2004). That is, the probability of any particular vertex having a degree k is given by $P(k) \sim k^{-\gamma}$. In biological networks (including PPI networks), the value of γ tends to be between 2 and 3 (Barabási & Albert, 1999). These networks were named by Barabási as scale-free, referring to the ability of power laws to retain their structure at any scale.



(a) *Degree labelled network* **(b)** *Degree distribution*

Figure 1.5: *The degree of a vertex, k, measures how many direct neighbours it has.* **a)** *An example network with scale-free topology (small number of high-degree hub vertices) with vertex degree labelled* **b)***: The degree distribution ($P(k)$) across all vertices of the network in* **a**.

An important property of scale-free networks is the existence of a small number of extremely high degree vertices called hubs. These hubs interact with many other vertices, but the vast majority of their interacting partners will have a much lower degree. Hubs are a defining topological characteristic of scale-free networks, and provide a rationale for why such networks have a robustness to random failure, but a weakness to targeted attacks (Albert *et al.,* 2000). In PPI networks for example, deletion of a hub gene is far more likely to be lethal than a protein with low degree, a phenomenon termed the centrality-lethality rule (He & J. Zhang, 2006; Mw & Ad, 2005).

Modularity

The topology of PPI networks is useful for investigating the functional structure of cells, as well as predicting the function of proteins based upon their location within the network (Barabási & Oltvai, 2004). Proteins which are directly interacting with a protein involved in a specific process have a high probability of also being involved in the same process (Hartwell *et al.,* 1999), and thus will tend to form "modules" of highly interlinked proteins which together share functionality. Identification of functional modules in PPI networks is a difficult (NP-hard in terms of computational complexity) problem for which many network clustering tools have been developed (Dittrich *et al.,* 2008; Reyna *et al.,* 2018; H. W. L. Koh *et al.,* 2019).

Path length, clustering coefficient, and the small world phenomenon

Path length counts how many edges are traversed in a path from vertex A to vertex B. As there are usually multiple possible paths, identifying the shortest path(s) is a common exercise which will determine the topological distance between vertices. The average of all shortest paths between all pairs of vertices in a network is the mean path length, a statistic which describes the overall difficulty of navigating a network. A related statistic is the clustering coefficient, which is a measure of how interlinked a given vertex is. The clustering coefficient C of vertex v is given by $C = 2n/k(k-1)$, where n is the number of edges which connect the k neighbours of v. The probability

distribution of any vertex having a particular clustering coefficient is then given by $C(k)$, where k is the degree. Along with degree distribution ($P(k)$), this property is independent of the network's size and may therefore be used to classify networks of any scale (Barabási & Oltvai, 2004). Complex networks usually have short mean path lengths and high average clustering coefficients, referred to as the small world phenomenon (Watts & Strogatz, 1998). In the context of PPI networks, the small world phenomenon means that most proteins are only a few interactions away from each other, and so network perturbations can quite easily have far-reaching effects.

Bottlenecks

When considering all of the shortest paths between pairs of vertices in a network, the links between network modules have the greatest number of these paths going through them. The statistic of number of shortest paths is known as betweenness centrality, and vertices with high betweenness centrality which link functional modules are often referred to as bottlenecks. These bottleneck vertices have the capability to control signal flow between modules, and very frequently represent essential proteins (H. Yu *et al.,* 2007).

1.8.2 Visualisation capabilities of networks

Layout algorithms

Visualisations of networks can be as varied as the data they represent, but perhaps the most important property to consider in network visualisation is the layout, i.e. the exact placement each vertex and edge. Although layout has no effect on the actual topological structure of a network for analysis, choice of layout can greatly assist (or hinder) interpretation. Manual construction of network layouts is only practical for the smallest of networks, and so there exist a great number of layout algorithms which automatically decide on vertex placement. For example, some classes of network are hierarchical in nature, and thus can be visually represented in a hierarchical fashion. However, the force-directed approach is possibly the most popular among network

layout algorithms due to their simplicity of implementation, rapid execution time and typically impressive and intuitive results (Fruchterman & Reingold, 1991). Force-directed algorithms essentially simulate networks as physical systems in which edges are springs exerting a force upon the network (either pulling it apart or together). Such algorithms typically begin by randomising the position of all vertices, then simulating the physical system over discrete time steps. Due to the initial randomisation, these algorithms are non-deterministic.

Vertex and edge attributes

A wide variety of additional information can be integrated into a network in the form of node or edge attributes such as shape, size, or colour. For example, one might encode the expression level of a gene as a node attribute or the experimental confidence associated with an interaction as an edge attribute. Even more complex types of visualisations such as heatmaps or pie charts may be contained within the network representation (Shannon *et al.,* 2003).

Network visualisation and analysis tools

Software libraries such as iGraph (Csardi & Nepusz, n.d.), NetworkX (Hagberg *et al.,* 2008), and graph-tool (T. P. Peixoto, 2017) allow users with programming experience to construct network visualisations and perform network analyses. Various free software platforms specifically tailored to biological network visualisation and analysis also exist, including NAViGaTOR (Brown *et al.,* 2009), VANTED (Rohn *et al.,* 2012), Cytoscape (Shannon *et al.,* 2003), and many others (Gehlenborg *et al.,* 2010). Cytoscape in particular has been particularly highly cited due to its extensive plugin system which allows integrative use of a diverse range of tools (Figure 1.6). Cytoscape was created by the Institute of Systems Biology in Seattle, and is now maintained by an international group of open source developers. It is implemented as a Java OSGi bundle, which makes it cross platform and highly modular. For end users, this means access to many Cytoscape "apps" which can be used together in any combination. For developers, Cytoscape is a ready to use and open framework upon

which network biology tools can be constructed with relative ease, with distribution to users all integrated within the platform.

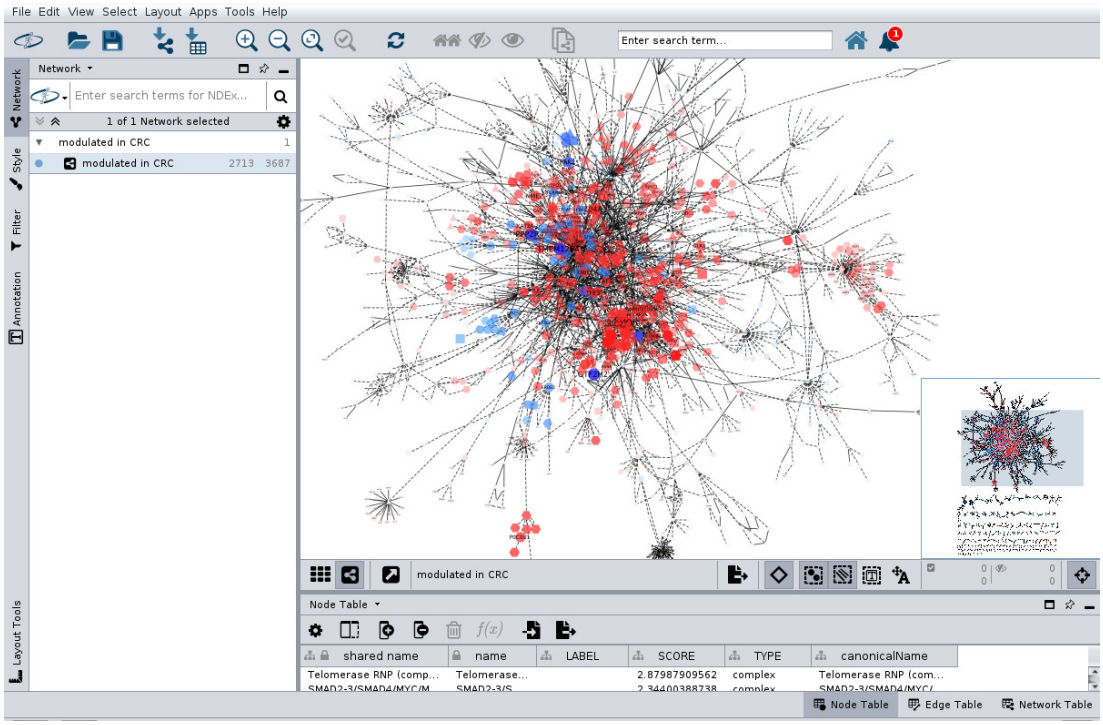


Figure 1.6: *Visualisation of genes significantly modulated in CRC using Cytoscape 3.8 (Shannon* et al., *2003), demonstrating force-directed layout of a large network, as well as customised vertex and edge attributes.*

1.8.3 Protein-protein interaction networks

Networks may be used to model the relationships between any molecular entities, but among the most common of these are protein-protein interaction (PPI) networks. A PPI is generally understood to be direct physical binding of proteins occurring within a particular cell (Rivas & Fontanillo, 2010). Proteins do not usually function in isolation, rather they form intricate and dynamic connections, acting as integrated molecular machines to perform biological functions. Mapping out the complete network of proteins that interact within living cells is therefore a fundamental task required to facilitate systems biology (Rivas & Fontanillo, 2010).

The human PPI network is extremely large. Current estimates put the number of human protein-coding genes at around 20,000 (Salzberg, 2018). These genes do not

necessarily match one-to-one with particular proteins, as the total number of distinct proteins may be further expanded and complicated by post-translational modifications (PTMs) such as phosphorylation, and yet further by alternatively spliced isoforms. Yang *et al.* were able to confirm that alternatively spliced proteins often have entirely distinct interaction profiles (Yang *et al.,* 2016), meaning that a major consequence of alternative gene splicing is widespread expansion of PPI networks. Additionally, certain proteins will only be expressed in particular subcellular contexts, at specific developmental stages, or within certain cell types. The result of all these different variables means there are a combinatorial number of possible states that protein networks may assume, which alter dynamically between different conditions.

1.8.4 Collection of PPI data

High throughput screening of protein-protein interactions is achieved via methods including yeast two-hybrid (Y2H), and affinity purification coupled with a protein identification method like mass spectrometry (together known as AP-MS). The yeast two-hybrid assay is now a 3 decade old technology (Fields & Song, 1989), yet is still an extremely popular way to detect binary protein-protein interactions. It has some major limitations however, such as spurious interactions causing high false positive rates, and requiring a yeast host, meaning that PPIs of other species are not always detectable due to missing PTMs (Ruffner *et al.,* 2007; Snider *et al.,* 2015). The requirement of yeast host cells may be sidestepped however by using more complex mammalian 2-hybrid approaches which better mimic actual in vivo interactions (Y. Luo *et al.,* 1997). AP-MS methods in comparison are based on biochemical purification of bait proteins from cell lysates, followed by mass spectrometry identification of bound preys (Gingras *et al.,* 2007). This enables identification of PPIs within physiological conditions near to those of the original cells, although the lysing process does create an unnatural environment in which unintended disruptions may occur (Snider *et al.,* 2015). Such unwanted interactions may be controlled for experimentally however using empty vector controls, and also computationally with databases of known contaminants and other statistical methods. In contrast to Y2H, AP-MS does not necessarily provide evidence that interaction is direct, as it also captures co-complex associations (I. W.

Taylor & Wrana, 2012).

Various large-scale projects have collected PPI data for public use, for example Rolland *et al.,* who identified an extremely large collection of binary PPIs via Y2H screening (approximately 14,000), more than doubling the previously available number, Huttlin *et al.,* who used AP-MS to create the BioPlex network, which contains over 56,000 protein interactions (Huttlin *et al.,* 2017; Huttlin *et al.,* 2015), and Kennedy *et al.,* who mapped >6000 PPIs specifically within the EGFR network using AP-MS (Kennedy *et al.,* 2020).

1.8.5 Interaction dataset curation

The existence of so many independent PPI datasets has led to multiple curation efforts aiming to make these data publicly available, for example IntAct (Hermjakob *et al.,* 2004) and BioGRID (C. Stark *et al.,* 2006), but vast numbers of additional resources with incompatible interfaces exist (Bader *et al.,* 2006). In order to prevent duplication of curation efforts and provide a common interface of non-redundant PPIs, the International Molecular Exchange (IMEx) consortium was formed (Orchard *et al.,* 2012). The IMEx consortium ensures consistency of data across all member databases by using a controlled interaction vocabulary developed by the Human Proteome Organization Proteomics Standards Initiative (HUPO-PSI) (Sivade (Dumousseau) *et al.,* 2018). This ensures interactions are available in a standardised format, and may be easily filtered using criteria like experimental strategy, host organism, or interaction type. IMEx (and its constituent databases) has made publicly available extremely detailed topological maps of the interactome as uncovered by countless individual experiments. The availability of such data allows complex disease phenotypes to be studied in terms of the dysregulation of PPI networks (Hastings *et al.,* 2020).

1.8.6 Dynamic re-wiring of PPI networks

PPI networks are not static and unchanging, rather they may be driven by multiple intrinsic and extrinsic factors to alter, with particular subnetworks only materialising

in certain tissue and cell-type specific contexts (Yeger-Lotem & Sharan, 2015). Recent studies have suggested that in fact, widespread re-wiring of network topology can occur in cancer in response to seemingly minor genomic alterations. Protein-protein interactions may be disrupted by various different types of mutation, causing widespread interactome rewiring – in cancer, the result of rewiring is typically a bypassing of normal regulatory controls and constitutive activation of signalling pathways (Bowler *et al.,* 2015). Many computational methods have been developed for predicting PPI disrupting mutations. One such example is reKINect (Creixell, Schoof, *et al.,* 2015), which was used to identify network-attacking mutations within phosphorylation-based signalling networks by mapping kinase domains and phosphorylation sites. Creixell et al. then also developed a second tool, KINspect, a computational framework that assists in understanding kinase-substrate interaction specificity (Creixell, Palmeri, *et al.,* 2015), which was used to determine if cancer mutations alter kinase specificity and therefore cause further downstream rewiring.

Systematic experimental studies have also been performed in order to characterise these mutations, and whether they impair PPI networks. In CRC cell lines, the Kennedy et al. demonstrated that the dosage of mutant KRAS has widespread network topology consequences on a physical level (Kennedy *et al.,* 2020). This implies that perhaps a similar effect may also exist between individuals. Sahni et al. (Sahni *et al.,* 2015) profiled several thousand missense mutations found across various Mendelian disorders, using multiple interaction assays to test whether the mutations led to perturbation of protein-protein interactions. The work by Sahni et al., along with other studies, is included in the EBI curation of PPI disrupting mutations, which has curated more than 28,000 instances of experimentally validated mutations (del-Toro *et al.,* 2019). Given the evidence for widespread network rewiring in cancers such as CRC, it is not unreasonable to think network rewiring would also occur and be useful for managing molecular heterogeneity on a patient-specific level.

1.9 Networks in cancer biology

1.9.1 Integration of multi-omics data

Interpretation of large scale omics is frequently informed through use of networks, including for identification of drivers and biomarkers, obtaining insights into the mechanisms of tumour biology through simulation of biological signal transduction, classification and subtyping of tumours, and developing novel therapeutic interventions (Ozturk *et al.,* 2018). Networks are used for these purposes due to several key advantages that they provide. One of these advantages is the consistent underlying mathematical structure upon which network and graph theory algorithms may be developed. Another is the fact that networks are intrinsically well suited to a systems-level approach to investigating tumour heterogeneity. However, one of the most significant advantages of a network approach for studying tumour biology is the ability to integrate diverse data sources into a single model.

Network theory is a unifying framework upon which multi-omics data (e.g. transcriptomics, epigenomics, proteomics) may be integrated from large-scale projects such as TCGA and the ICGA. Gene expression, mutation, or methylation data may be compressed and combined into a single node, or separate network layers may be created for each data type and mapped together. However, one weakness of such an approach is that because interaction data is extremely dense, with the interactome containing many millions of interactions, interactome networks will frequently turn into "hairballs" and become very difficult to interpret. It can therefore be useful to determine which parts of the network are most relevant given a particular cellular context. Omics Integrator (Tuncbag *et al.,* 2016) is an example of a tool designed to use multi-omics data to create high confidence subnetworks which better explain the observed data. Omics Integrator addresses many issues which are infrequently considered in multi-omics integration, for example it is able to make use of chromatin accessibility data, such as from DNase-Seq or ChIP-Seq, to identify transcriptional regulators which may be responsible for observed gene expression levels or alterations between conditions. This can be preferable to mapping gene expression to proteins

directly, as the relationship between mRNA expression and protein levels is not linear. Another tool designed for network integration of multi-omics is iOmicsPASS, which uses a Z-score normalisation approach to combine different data types (H. W. L. Koh *et al.,* 2019). Due to the nature of network approaches however, many integrative tools such as these have overlapping functionality, and integration of multi-omics is often merely the first step of a network based analysis. For example, Omics Integrator and iOmicsPASS both not only integrate multi-omics data, but also offer patient stratification and subnetwork discovery capabilities.

1.9.2 Driver and oncogene prediction

One major application of networks in cancer biology is to the identification of driver genes (Ozturk *et al.,* 2018). In terms of specific somatic mutations, individual tumours tend to be extremely heterogeneous. However, the bulk of these mutations (and therefore much of the apparent genomic heterogeneity) is due to passenger mutations (i.e. mutations which do not contribute to tumour development) (Vogelstein *et al.,* 2013). Non-passenger mutations which are causal for driving tumour development are known as driver mutations, and are targets of significant therapeutic interest. Networks provide significant advantages when searching for driver genes as they allow the problem to be approached as a search for highly mutated modules, rather than individual genes. This is due to the "disease module" hypothesis, based on the observation that driver mutations tend to cluster in the same network neighbourhood (Barabási *et al.,* 2011).

Simplistic methods for determining disease-related vertices may look at direct-interactions, implicating any vertex that is a direct neighbour of a known disease-associated vertex in the same disease (Oti *et al.,* 2006), or may rank potential disease-associated vertices by examining the length of shortest-paths to known disease genes (George *et al.,* 2006). More advanced strategies for detection of disease modules will commonly use a diffusion algorithm to determine a network neighbourhood of interest (Köhler *et al.,* 2008). Köhler *et al.* demonstrated that using a diffusion or random-walk approach was superior to the more simplistic direct interaction or shortest path

methods for detecting potential drivers, using known disease-associated genes from the Online Mendelian Inheritance in Man (OMIM) database (Hamosh *et al.,* 2000). This approach was also used by Network-based integration of multi-omics data (NetICS) (Dimitrakopoulos *et al.,* 2018), a more recent example of using network diffusion to identify drivers and biomarkers. NetICS features integration of genetic and epigenetic alterations to detect mediator genes which are responsible for orchestrating downstream expression changes in an interaction network. NetICS requires a directed interaction network from which sample-specific rankings are determined using a diffusion process, which are later aggregated to gain population-level gene rankings. The method was demonstrated to be effective in identifying infrequently implicated driver genes in 5 different TCGA cancer datasets (Dimitrakopoulos *et al.,* 2018).

Network approaches to driver prediction overcome the limitations of simple mutation frequency tests, which suffer from the "long tail" phenomenon in which most driver mutations are quite rare across a cohort (Kandoth *et al.,* 2013). The ability to detect rare drivers means that network approaches have significant potential for patient-specific driver detection. OncoIMPACT, for example, predicts patient-specific driver genes in multiple cancer types by linking somatic mutation and gene expression within predefined interaction networks (Bertrand *et al.,* 2015). OncoIMPACT aims to explain deregulated genes in tumour samples (detected from differential expression) in terms of their connectivity to mutated genes. Under their model, mutations which are linked to frequently deregulated genes are considered to be indicative of the cancer phenotype, and are thus considered more likely to be potential drivers. A different approach to predicting personalised driver genes is to create individual networks for patients, for example Liu et al. developed a statistical approach for creating sample-specific networks (SSN) (Liu *et al.,* 2016). SSNs were constructed by comparing pairwise correlations of each pair of molecules in each individual sample against a group of control samples, using a perturbation approach to gain statistical significance. In effect, each SSN is a sample-specific gene co-expression network. Liu *et al.* found that functional driver genes could be reliably predicted from hubs in these SSNs, using multiple different TCGA cancer datasets.

1.9.3 Identification of disease modules

Identifying significantly altered subnetworks

Due to the disease module hypothesis, driver and oncogene prediction in networks is effectively a sub-problem of the more general task of defining disease-relevant subnetworks. Identification of subnetworks associated with cancer or other diseases is often accomplished using a diffusion strategy. Using models based on physical heat diffusion, such approaches define a set of "hot" vertices of interest from which diffusion will begin. The "heat" then spreads to surrounding vertices, and can be used to identify new functional modules, implicating nearby vertices in a particular process or disease. One popular algorithm utilising this approach is TieDie (Paull *et al.,* 2013), which uses a tied diffusion process from two separate sets of vertices, and has been applied to sets of transcription factors and patient specific mutations to identify cancer-related networks that can distinguish between major breast cancer subtypes (Paull *et al.,* 2013).

Another widely used diffusion-based algorithm to is Hierarchical HotNet (Reyna *et al.,* 2018), which has been used for instance for identifying protein networks significantly perturbed by post-translational modifications in SARS-CoV-2 infection (Stukalov *et al.,* 2021). Hierarchical HotNet is the most recent and best performing of three similar algorithms designed for subnetwork identification, its predecessors being HotNet (Vandin, Upfal, *et al.,* 2011) and HotNet2 (Leiserson *et al.,* 2015). While all these algorithms are based on heat diffusion from an initial set of vertices of interest, the implementation differs. HotNet uses a continuous diffusion kernel, while HotNet 2 and Hierarchical HotNet utilise a discrete random walk approach. In Hierarchical HotNet, this random walk also incorporates vertex weights and directional edges. After performing this random walk, the distribution of heat scores is used to create a similarity matrix. Hierarchical clustering is performed on this similarity matrix, providing clusters at multiple scales. Hierarchical HotNet then determines whether the vertex weights are higher than would be expected by chance in particular clusters, i.e., it identifies significantly mutated/perturbed subnetworks.

Recently, the PCAWG consortium applied multiple network methods (including Hierarchical HotNet) to perform an integrative analysis of drivers across 27 tumour types from the ICGC/TCGA (Reyna *et al.,* 2020) with a focus on identifying differential functional effects between coding and non-coding mutations. The consortium found that certain pathways and processes are targeted more than others by specific classes of mutation, for example RNA splicing pathways are predominantly altered far more by non-coding mutations than others.

The over-representation of hubs

A common issue with network analyses, including during disease module identification, is the over-representation of hubs. Hubs may link to almost all other nodes in a network, and may therefore appear more important to a particular function than they actually are, biasing results. There are many different approaches to solving the hub weighting issue, some tools (e.g. OncoIMPACT (Bertrand *et al.,* 2015)) may disregard hubs entirely for certain calculations, while others will use algorithms which reduce significance in proportion to the node degree (Tuncbag *et al.,* 2016). An alternative approach is to use additional information like differential gene expression to identify hubs which are relevant only in a particular context, as exemplified by the Cytoscape app CHAT (Muetze *et al.,* 2016).

Patient-specific interaction subnetworks

Some approaches focus on leveraging multi-omics data to identify patient-specific subnetworks within larger interaction networks. For example, recent advances have allowed for clustering patient-specific mutations by their 3D position on protein structures to uncover patient-specific subnetworks in glioblastoma (Dincer *et al.,* 2019). Supervised learning approaches have also been utilised for identification of sample-specific subnetworks. For example the iOmicsPASS tool uses nearest shrunken centroid classification to select predictive features from Z-score normalised multi-omics data, resulting in patient-specific network scores which can be used to define predictive subnetworks for patients or conditions (H. W. L. Koh *et al.,* 2019).

1.9.4 Patient stratification

Non-network sample stratification

To devise targeted treatments, an important task is defining clinically relevant tumour subtypes. Such classification is typically accomplished using a sample-wise clustering algorithm. For example, the Consensus Molecular Subtypes (CMS) of CRC were derived from multiple unsupervised clustering algorithms applied to transcriptomics data (Guinney *et al.,* 2015) (mainly hierarchical agglomerative clustering). When considering multi-omics data however, clustering and stratification based on all available data becomes difficult due to collection bias, noise, and sheer scale. One common approach to overcome issues such as bias and noise is via consensus clustering (Monti *et al.,* 2003), in which different perturbations of the data are used to more robustly cluster samples. Unfortunately this technique suffers from being extremely computationally intensive and does not scale well with increasing numbers of samples or omics types. Alternative approaches sometimes utilise supervised clustering, such as iCluster, a machine learning approach using a latent variable model for integrative clustering (Shen *et al.,* 2009). However, such supervised methods both require and are particularly sensitive to feature preselection, potentially biasing analysis.

Network approaches to sample stratification

Networks may be used as a tool for stratification in multiple ways. One network approach which shows significant utility for robust multi-omics based sample clustering is to use patient similarity networks, as exemplified by Similarity Network Fusion (SNF) (B. Wang *et al.,* 2014). SNF clusters samples by creating sample-similarity networks for each omics type, then integrating those networks into a single similarity network via nonlinear combination, aiming to overcome bias, noise, and scale. The resulting fused similarity networks cluster samples over multiple data types, and have been used to detect subtypes in TCGA cancer datasets. Through analysis of survival differences between predicted clusterings using Cox regression, SNF based clusterings consistently provided more significant differences between groups than clustering based on any

omics data individually (B. Wang *et al.,* 2014). This network-based approach has favourable performance characteristics in comparison to tools like iCluster (in which time complexity scales exponentially with input size).

Other network approaches to subtyping aim to group similar tumour samples based upon molecular interaction network topology, which is possible even if different specific genes are perturbed (Vandin, Clay, *et al.,* 2011). Network Based Stratification (NBS) (Hofree *et al.,* 2013) is an example of how cancer cohorts may be stratified into clinically relevant subtypes in various cancers based on integration of multi-omics with broader interaction networks. NBS overlays patient-specific somatic mutation data on a gene interaction network and applies network propagation (Vanunu *et al.,* 2010) (a method based on random-walks) to smooth patient profiles prior to unsupervised clustering with non-negative matrix factorisation, followed by consensus clustering. NBS was demonstrated to be able to define subtypes which were predictive of survival in ovarian cancer independently of other clinical covariates (such as tumour stage or age) (Hofree *et al.,* 2013).

1.9.5 Simulating signalling networks

Therapeutic target prediction

A key reason for using networks in cancer research is the potential for identifying novel drug targets and combination therapies. Development of detailed models which are able to predict patient-specific response to targeted interventions has been attempted using multiple different mathematical formulations of cellular signalling networks. One such formulation is a logical boolean model, in which each edge can be switched on or off by logical operators (i.e. OR, AND, NOT). This approach has been applied for example by Béal *et al.* with their multi-omics integration tool PROFILE (Béal *et al.,* 2019), or Eduati *et al.* (Eduati *et al.,* 2020), who utilised logical networks in order to identify optimal combination therapies within pancreatic cancer using their tool CellNOptR (Terfve *et al.,* 2012). Eduati *et al.* were able to demonstrate that a small, well described logical apoptosis network could be "trained" using experimentally derived drug perturbation data from biopsies or cell lines, and demonstrated that

these trained networks could be predictive of novel combination therapies within cell lines. Other highly detailed signalling models have been developed based on ordinary differential equations, which model reaction kinetics of each individual interaction (Hastings *et al.,* 2020). Such approaches are difficult to apply on a large scale however, due to the requirement of extensive labour-intensive experimentation to derive reaction kinetics.

Diffusion and random walks

Many network algorithms, especially those for signal transduction simulation and module discovery, make use of diffusion and random walks, two fundamentally similar concepts. In general, diffusion describes the movement of something from an area of high concentration to low concentration (i.e., along a concentration gradient), a concept prevalent across many scientific fields. From a physical perspective, diffusion is the result of random migration of molecules due to thermal energy (i.e, Brownian motion). As particles move, they collide with each other, eventually reaching a state of equilibrium in a closed system. This physical process of diffusion may be described using differential equations (Fick's laws of diffusion (Fick, 1855)), or modelled in a discrete manner using random walks which simulate the seemingly random paths of individual molecules (Berg, 1993). In graphs (and networks), a random walk is a special case of a Markov chain, in which walkers have a certain probability of jumping to the next vertex at every discrete step (Masuda *et al.,* 2017). Variations on random walks form the backbone of many algorithms used in the network analysis space due to their helpfulness in revealing topological properties, such as which vertices are most central given a set of important vertices, or which paths through the network are most relevant to a particular function. Bioinformatic techniques for network analysis including random walks are sometimes termed "network propagation" methods (Cowen *et al.,* 2017), which includes the related concepts of information diffusion, random walks, and electrical resistance approaches.

Biological information flow

Network propagation approaches have been applied to modelling biological signal transduction and information flow within molecular signalling pathways, such as networks of gene regulatory relationships, the physical interactions of PPI networks, or metabolic networks (Y.-A. Kim *et al.,* 2011). Random walk processes are particularly well suited for modelling signalling in these networks, and were employed by Stojmirović & Y.-K. Yu to develop an approach for identifying functional information transduction modules (ITMs) within PPI networks. Their approach initially defined an emitting and absorbing model, in which contextually relevant proteins were selected as either the source or sink (destination) of random walks. An important feature of the ITM approach is the introduction of a damping factor, i.e. a certain probability of a random walk dissipating. This limits how far a walker can proceed through the network, mimicking the natural information loss that occurs in networks. A later update of the concept introduced a channel model to better permit directed information flow, in which both source and sink nodes could be defined (Stojmirović *et al.,* 2012). This makes it particularly attractive for modelling signalling between cell surface receptor molecules and transcription factors in large PPI networks, as was done for example by Kennedy *et al.* to model the dynamic signalling alterations between different CRC cell lines (Kennedy *et al.,* 2020). This kind of modelling of dynamic signalling has not often been applied on a patient-specific basis, however, which is a subject that I aim to explore more thoroughly in this thesis.

1.10 Developments in spatially resolved omics

1.10.1 Spatial transcriptomics

Cancers are not only variable between patients, but also within single tumours (intratumour heterogeneity), with high levels of genomic heterogeneity within a tumour being directly associated with poor patient outcomes (Marusyk *et al.,* 2012). Genotypically unique subpopulations may exist in different geographical regions, and can

evolve over time to have unique properties (Bedard *et al.,* 2013). High-throughput "bulk" RNA sequencing technology (RNA-seq) has been an indispensable tool for examining genomic function for well over a decade, most typically through examining differential gene expression. However, the principle limitation of bulk RNA-seq for tumour biology research is that biopsies must be homogenised prior to sequencing. This homogenisation results in many cells being blended together, discarding spatial information and making information on specific cell types difficult to obtain. Thus, standard RNA-seq experiments are insufficient to decode the complexity of many biological systems, especially in highly structured tissues such as human brain tissue, or extremely heterogeneous tissue such as tumour biopsies (R. Stark *et al.,* 2019). This need is being addressed in two major ways: one is with single-cell transcriptomics, a technology which has surged in popularity in recent years (Suvà & Tirosh, 2019); the other is with *in situ*, spatially-resolved transcriptomics methods which preserve the spatial position of transcripts under investigation (Crosetto *et al.,* 2015; Moor & Itzkovitz, 2017; Burgess, 2019; Strell *et al.,* 2019). Many competing methodologies have been described with varying technical approaches (Lein *et al.,* 2017). The increasing popularity of these spatially-resolved methods is epitomised by the recent awarding by *Nature Methods* of "Method of the Year" to spatial transcriptomics (Marx, 2021), a method originally described by Ståhl *et al.* (Ståhl *et al.,* 2016).

One major class of spatially-resolved transcriptomics technologies is those based on fluorescence in situ hybridisation (FISH). Single molecule FISH (smFISH) may be used to perform microscopy imaging of mRNA using libraries of short, transcript-targeted, fluorophore labelled oligonucleotide probes. This approach has been applied to detect transcripts within individual cells (Itzkovitz & van Oudenaarden, 2011) and mammalian tissues (Lyubimova *et al.,* 2013). However, such applications of smFISH lack effective multiplexing capabilities due to requiring fluorophores which remain uniquely identifiable when used simultaneously. This limitation may be sidestepped by attaching a combination of probes to a single transcript at multiple points, which was demonstrated by Lubeck & Cai to be able to measure mRNA in 32 genes simultaneously with super resolution microscopy (Lubeck & Cai, 2012). Alternatively, an approach which increases experimental complexity but has the potential to vastly

extend the number of combinations is sequential (or temporal) barcoding, in which multiple hybridisation and imaging steps are used to bind different probes along transcript sequences (Lubeck *et al.,* 2014). Still, all spatial transcriptomics methods which rely on optical readout suffer from imaging density issues, as transcript localisation can exceed the physical limits of diffraction. The development of super-resolution microscopy (B. Huang *et al.,* 2009) is one option that may overcome this, however using such technology can be cost prohibitive. An alternative low cost workaround is through physical magnification through expansion microscopy (Wassie *et al.,* 2019), for example exFISH which links molecules in an expanding hydrogel, effectively increasing the potential resolution of smFISH imaging (F. Chen *et al.,* 2016).

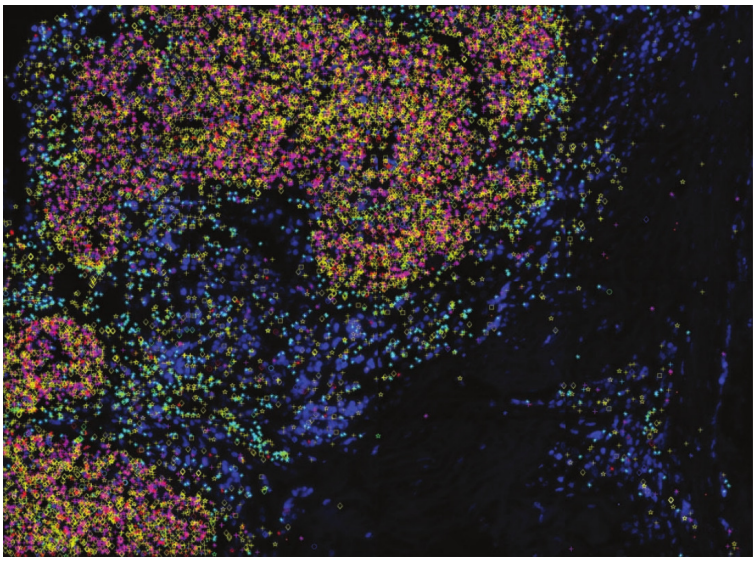


Figure 1.7: *Visualisation of* in situ *sequencing on a HER2-positive breast cancer tissue sample by Ke* et al. *Reproduced with permission from Springer Nature. Each coloured symbol represents a localised mRNA transcript detection, plotted on top of fluorescence microscopy imaging.*

Sequential barcoding has also been utilised by non-smFISH methods which use *in situ* transcriptomic amplification and sequencing. Rather than binding fluorescent probes directly, such methods synthesise cDNA in place, which is then used as a target for amplification and sequencing by ligation. This can then be combined with a sequential fluorescence barcoding approach, as demonstrated by Ke *et al.* who used padlock probes followed by rolling-circle amplification and sequencing by ligation to

generate fluorescently barcoded images of 39 transcripts within breast cancer tissue samples, including 21 transcripts used within the OncoType DX prognostic panel (Ke *et al.,* 2013). This approach is very similar to FISSEQ (J. H. Lee *et al.,* 2014), which maps many more reads, albeit at lower resolution. One drawback of sequential barcoding approaches is the potentially much lower success rate than single-cycled smFISH, as if probes fail to bind at any single step the barcode will be incorrect or invalid. In response to this issue, Chen et al. developed multiplexed error robust FISH (MERFISH) (K. H. Chen *et al.,* 2015) which utilises error-resistant encoding schemes to reduce detection errors. This approach allows error-robust imaging of over 100 distinct RNA species in individual cells. MERFISH is limited by the fact that an increased number of probes must be used for error-robust encoding, so many that only transcripts larger than ∼3kb can be targeted with the method. A more modern sequencing-based method, spatial transcriptomics (Ståhl *et al.,* 2016), uses microarrays with fixed reverse transcription primers to sequence RNA from tissue sections. A recent advance (High definition spatial transcriptomics or HDST) has allowed this method to be extended even further to 2-μm resolution (Vickovic *et al.,* 2019), allowing subcellular investigation of spatial heterogeneity.

Single-cell RNA sequencing (scRNA-seq) technology has also been adapted to obtain spatially resolved data, however this tends to be a lower-resolution approach. For example, a development of the Drop-seq single cell sequencing approach was developed into Slide-seq (Rodriques *et al.,* 2019), in which an array of distinctly barcoded beads is used to resolve spatial features to around 10 μm in preserved tissue sections. A key advantage of Slide-seq is its ease of integration with existing pipelines for single-cell analysis (Navarro *et al.,* 2017). As a still developing area, analysis methods for spatial transcriptomics are relatively limited, and so reuse of existing RNA-seq and scRNA-seq tools is common.

1.10.2 Spatial metabolomics

Spatial metabolomics is a recently emergent field of omics research concerned with detection and analysis of metabolites, drugs, lipids and other small molecules within

the spatial context of cells and tissues (Alexandrov, 2020). Similarly to bulk transcriptomics, bulk metabolomics requires extraction of metabolites from samples in a way that does not preserve tissue or cellular localisation. Mass spectrometry (MS) is the most common and effective approach to bulk metabolomics (Dettmer *et al.,* 2007). In contrast to the extreme diversity of spatial transcriptomics technologies, most spatial metabolomics occurs via some form of mass spectrometry imaging (MSI). MSI allows spatial resolution of metabolites by dividing a sample into a grid of pixels prior to desorption of molecules within the pixel, followed by generation of a mass spectrum for each pixel. While the particular methods of desorption vary, one of the most common methodologies is to use matrix-assisted laser desorption/ionisation (MALDI), in which a laser is used to ionise and ablate molecules at every pixel location (Rohner *et al.,* 2005). This methodology has been applied to gain structural information on tumour biology, including for example mapping lipid composition within the CRC tumour microenvironment (Mirnezami *et al.,* 2014), and tracking localisation of the EGFR-targeted antibody cetuximab in 3D colon cancer cell cultures (Liu *et al.,* 2018).

1.10.3 Analysis of spatially resolved omics

Spatially resolved omics technologies face the same analytical challenges involved with their bulk counterparts, plus an entirely new host of challenges from the new dimensions of data being explored. While they naturally allow for intuitive qualitative visualisations, the statistical and computational methods for analysing spatially resolved omics data are still in their infancy. Basic tools in spatial transcriptomics for converting raw outputs into matrices of gene counts and spot positions exist, like ST Pipeline (Navarro *et al.,* 2017), but tools to analyse information taking into spatial location are sparse. Some of the most promising tools make use of image analysis and tools of geospatial statistics like point pattern analysis, for example STUtility (Bergenstråhle *et al.,* 2020). Highlighting the overlap between the spatial and single-cell fields, STUtility itself is based on the Seurat framework for spatial reconstruction of scRNA-seq data (Satija *et al.,* 2015). A common approach to spatial omics analysis when discrete pixels of data are available is to use unsupervised clustering algorithms like tSNE to classify each spatial point. However, this approach is not applicable if

discrete pixels do not exist, and does not fully utilise the spatial resolution of the data. As a fairly recent and underdeveloped field, there is a demand for new analysis methods. This demand could potentially open up a new space for network analysis, due to the capability of networks to integrate diverse data sources and the need to link spatial data to existing, better established technologies.

Following the literature review performed in this chapter, key algorithmic approaches for analysing tumour heterogeneity, as well as resources for researching colorectal cancer specifically have been identified. In chapter 2, I will investigate the different ways in which patient transcriptomic data can inform patient stratification and be used to predict patient-specific outcomes. In chapter 3, I will integrate my findings from the previous chapters to create patient-specific network models and test methods for using these models to predict patient outcomes, with the goal of creating an approach that integrates multiple sources of data. Finally in chapter 4, I will leverage network analysis to investigate spatial tumour heterogeneity, a field that is becoming more accessible due to emerging spatially-resolved technologies. A summary of these objectives is presented in Figure 1.8.

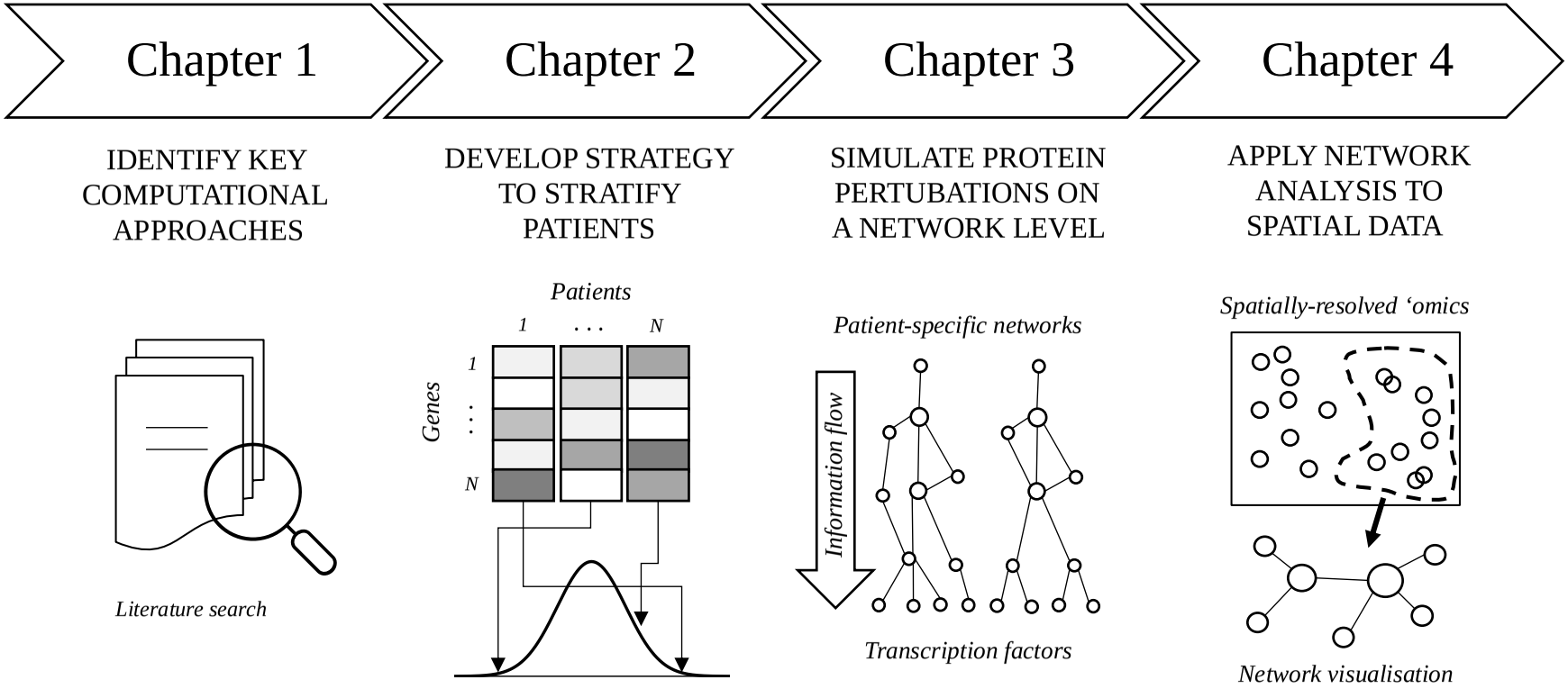


Figure 1.8: *Diagram of overall thesis objectives and research strategies.*

2. Patient-specific gene and pathway analysis in colorectal cancer

2.1 Introduction

For decades our understanding of tumour heterogeneity, and thus stratification of patients, has been guided primarily by pathology and histology. This is reflected in standards such as the tumour node metastasis (TNM) staging system (Edge *et al.,* 2010), in which molecular-based classifications are comparatively rare. Molecular classification systems have been viable in research since the development of microarray technology (Perou *et al.,* 2000), however it is only in recent years that a significant reduction in the cost of transcriptome-wide RNA sequencing technology has made clinical application of molecular classification a realistic possibility for patient-specific treatment stratification and prognosis (Van den Berge *et al.,* 2019).

Projects such as The Cancer Genome Atlas (TCGA) which has published data on thousands of patient samples across 33 tumour types (Hoadley *et al.,* 2018) have made patient transcriptomics data publicly available. These data from TCGA and other projects have been used by the Colorectal Cancer Subtyping Consortium to create a unified classification system for colorectal cancer (CRC) called the Consensus Molecular Subtypes (CMS) (Guinney *et al.,* 2015), based on more than 4000 individual tumour samples. The CMS define four subtypes of CRC, CMS1-4, which represent the best transcriptomics-based classification of CRC to date. CMS1 encompasses the majority of tumours exhibiting microsatellite instability (MSI), and is characterised by expression of genes involved in immune infiltration. CMS2-3 in comparison to CMS1 have been found to exhibit higher chromosomal instability (CIN) as measured by somatic copy-number alterations. CMS2 is the canonical subtype, with enrichment

of *WNT* and *MYC* downstream targets. CMS3 is described as the metabolic subtype, featuring an over-representation of *KRAS* mutations which are potentially the cause of the subtype's characteristic enrichment of metabolism-related signatures. CMS4 is the mesenchymal subtype, which is characterised by the upregulation of genes involved in epithelial-to-mesenchymal transition (EMT), angiogenesis, and stromal infiltration.

Despite the CMS being the best transcriptomic-based CRC subtype classification to date, and the first capable of distinguishing more than the MSI subtype in CRC, much room is left for improvement. The CMS are of limited prognostic value, with the only overall survival difference between subtypes being the notably poorer outcome of the mesenchymal CMS4 subtype (Guinney *et al.,* 2015). In addition, 13% of the samples used to create the CMS were unable to be robustly classified. It has been suggested that to improve the robustness of classifications, transcriptomics data should be combined with other molecular data such as proteomics, metabolomics, genomics and epigenetics, as alterations on all these levels contribute to tumour heterogeneity (Blanco-Calvo *et al.,* 2015). However, other avenues worth exploring may still exist for further increasing the utility of transcriptomics data for patient stratification, as generating such multiomics data for large numbers of samples is unlikely to be feasible clinically.

Transcriptomic studies typically identify differentially expressed genes between normal and tumour tissues. These sets of genes may then be used as a basis for classification. Less commonly examined is the inter-patient heterogeneity of differentially expressed genes between different tumour samples. Comparison to normal samples is generally desirable as this controls for non-disease patient differences and noise. From a systems perspective however, it may not be the case that all molecular differences which influence overall patient survival are specifically differentially expressed in tumour samples. Rather, inter-patient heterogeneity which is observed between different tumour samples may be important to consider. As inter-patient heterogeneity is a powerful driver of differing clinical response and outcomes in cancer (Reuben *et al.,* 2017), being able to identify patient-specific transcriptomic differences within a cohort may be key to improving molecular classifications and understanding the tumour biology of CRC, potentially leading to improved patient-specific outcome predictions

and new therapeutic targets. Furthermore, many publicly available transcriptomics datasets lack matched normal data for the majority of samples (as is the case for TCGA CRC data), meaning that to fully utilise these data it is necessary to identify patient-specific differences between tumour samples.

A systems approach to improving the utility of transcriptomics data could also involve summarising gene activity in terms of sets of functionally related genes, i.e. biological pathways, with tools such as functional enrichment analysis (Jin *et al.,* 2014). While these tools have been a mainstay of bioinformatics for many years, they are less commonly applied on a patient-specific level. When such tools do provide measures of patient-specific pathway activities, they will generally do so with reference to a normal or control level due to the inherent noise between samples, for instance PARADIGM (Vaske *et al.,* 2010). Some patient-specific tools for elucidating pathway enrichment such as Gene Set Variation Analysis (GSVA) (Hänzelmann *et al.,* 2013), single sample GSEA as described by Barbie *et al.,* 2009, or PLAGE (Tomfohr *et al.,* 2005) are capable of producing pathway scores which compare only within tumour samples, but their results are typically used for tasks such as creating survival models, and are not frequently used to assist with patient classification. Given that a pathway-level approach is likely to provide more informative results than expression data in isolation, the combination of focusing on patient-specific heterogeneity and subsequent dimensionality reduction of this information in terms of significantly altered pathways may be an effective way to increase the utility of transcriptomics data and generate clinically relevant classifications in colorectal cancer beyond the CMS.

2.2 Hypothesis and Aims

To elucidate between-patient molecular differences in tumour heterogeneity, novel patient-specific methods which capture the most important differences between individual tumour samples are needed. I hypothesised that individuals within a cohort could be characterised by a small portion of patient-specific differentially expressed genes. I further hypothesised that identifying these patient-specific differentially expressed (PSDE) genes would be useful for stratifying patients into clinically relevant molecular subtypes. To address these hypotheses, I proposed the following aims:

1. Develop a method to identify genes that are differentially expressed at a patient-specific level within a cohort of patients.

2. Test whether PSDE genes are representative of biological characteristics by examining connections to biological pathways.

3. Use PSDEs as a basis for stratification of patients into novel subtypes, and determine whether these subtypes are predictive of patient outcome.

2.3 Methods

2.3.1 Patient-specific differentially expressed genes

I developed a methodology to identify genes specific to each patient tumour sample which are expressed at a significantly different level from other patients within the same cohort, without comparison to normal samples. These patient-specific differentially expressed (PSDE) genes for a given patient therefore reveal the most extreme inter-patient differences in gene expression characterising the individual sample. I accomplished this using the combination of two per-gene thresholds, a fold change and a Z-score threshold. If the expression of a gene was outside of both of these thresholds for a patient, that gene would be defined as a PSDE gene for that patient. The purpose of combining the fold change and Z-score thresholds was to identify genes which increase or decrease in specific samples by an appreciable and biologically significant amount (hence, the fold change), and are also altered in their expression outside of what might be considered to be the normal range of the gene in these samples (hence, the Z-score). This logic aims to prevent genes from being classified as PSDE in too many patients, thus reducing their usefulness. The fold change threshold F for a gene g (F_g), was defined as a 2-fold change from the cohort median expression of that gene. The Z-score threshold Z for a gene g (Z_g) was based on a ± 1.96 standard deviation from the mean of logged counts per million (CPM), corresponding to a two-tailed significance test. A gene g was defined as patient-specific differentially expressed (PSDE) for a particular individual, if both F_g and Z_g held true, i.e.:

$$PSDE_g = F_g \wedge Z_g \tag{2.1}$$

Definition of threshold F_g, two-fold change

For each sample, the fold-change in gene expression (CPM) is given by:

$$FC_g = \frac{x}{\tilde{x}_g} \tag{2.2}$$

Where $\tilde{x}_g$ is the median CPM of gene g. Threshold F_g is then given by:

$$F_g = |\log_2(FC_g)| > 1 \tag{2.3}$$

As on a $\log_2$ scale, a unit increase or decrease corresponds to a two-fold change. The nature of this relative threshold is to scale with the median CPM - that is, the higher the median CPM, the higher the absolute CPM difference must be to satisfy a two-fold change.

Definition of threshold Z_g, Z-score

The gene-wise logCPM Z-score is given by:

$$ZS_g = \frac{x - \mu_g}{\sigma_g} \tag{2.4}$$

Where x is the logged CPM value, μ_g the cohort mean, and σ_g the cohort standard deviation. Threshold Z_g, the Z-score threshold, is then given by:

$$Z_g = |ZS_g| > 1.96 \tag{2.5}$$

An absolute Z-score of >1.96 corresponds to approximately 95% of the area under the normal curve (or $p < 0.05$ on the cumulative normal distribution). It is important to note the different usage of mean for Z-score, and median for fold-change calculation. This decision was made to account for the negative binomial distribution of CPM values, in which the mean tends to be skewed much higher than the median, in comparison to the normal distribution of log-transformed CPM data.

PSDE genes defined in this way could then be divided further into up and down regulated, based on whether their expression was above or below the median expression of that gene in the cohort. The thresholds are demonstrated in Figure 2.1. Expression of gene g in patient A falls below the thresholds, and so the gene is defined as a down-regulated PSDE gene. Expression in patient D falls above the thresholds, and so it is defined as an up-regulated PSDE gene for this patient.

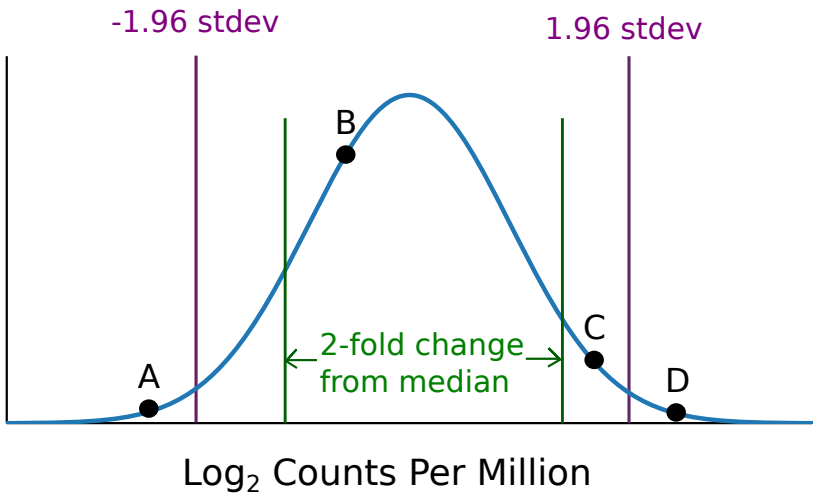


Figure 2.1: *The distribution of logged CPM values for an example gene, g, is shown. The expression of gene g in four patients, A, B, C and D is indicated. Gene g is identified as a down-regulated PSDE gene for patient A, as it is expressed below both the Z-score and fold change thresholds, while it is an up-regulated PSDE gene for patient D. In patients B and C, this gene is not defined as a PSDE gene.*

2.3.2 Patient-specific pathway enrichment

Patient-specific pathway activities were determined by performing pathway enrichment analysis on up-regulated and down-regulated PSDE genes separately. Pathway annotations were sourced from Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa & Goto, 2000), Gene Ontology (GO) (Ashburner *et al.*, 2000), MSigDB (Liberzon *et al.*, 2015), WikiPathways (Slenter *et al.*, 2018), and Reactome (Fabregat *et al.*, 2018). A Fisher's Exact test was used to determine statistical significance. P values were adjusted for multiple testing using the Benjamini-Hochberg procedure, as implemented in the statsmodels Python package (Seabold & Perktold, 2010).

To facilitate more effective visualisation of pathway enrichment scores, I introduced the P-derived enrichment score (ES), a simple metric representing the overall trend in pathway activity based upon two separate P-values obtained from up-regulated and down-regulated PSDE gene sets:

$$ES = (1 - p_{up}) + (p_{down} - 1) \tag{2.6}$$

The ES is bounded between ±1, where values close to +1 and -1 respectively indicate increased or decreased pathway activity. A useful property of this score is to cancel out pathways which are enriched in both up and down gene sets which prevents pathways from appearing simultaneously up and down-regulated. The resulting single value is similar to the enrichment score of methods such as GSEA (A. Subramanian *et al.,* 2005). The ES method was applied to visualise pathway activities for multiple sample groups using heatmaps or other visualisations.

2.3.3 RNA-seq data acquisition

Colorectal cancer (CRC) patient samples from The Cancer Genome Atlas (TCGA) were used as a test case to develop the PSDE method due to the open availability of a large patient cohort (n=633) and because certain molecular phenotypes in CRC such as microsatellite instability (MSI) are well characterised and would be useful to validate findings. TCGA's CRC cohort is comprised of two projects: TCGA-COAD (Colon Adenocarcinoma, 461 unique individuals); and TCGA-READ (Rectum Adenocarcinoma, 172 unique individuals). RNA sequencing read counts as generated by the TCGA using HTSeq2 (Anders *et al.,* 2015) were downloaded from the NCI Genomic Data Commons (GDC) using the GDC data transfer tool[1]. A file manifest for use with the transfer tool was created by browsing the GDC web portal[2] and selecting all gene expression files available from patients in the TCGA-COAD and TCGA-READ cohorts. Corresponding metadata for HTSeq2 counts was downloaded from both the harmonized GDC portal and the legacy archive[3] and merged with a custom script, as not all legacy information was present in the harmonized metadata (of specific note, information on the sequencing platform was absent). All counts were compiled into a single table of samples (identified by TCGA barcode[4]) and genes (identified by versioned Ensembl ID) as seen in Table 2.1. Ensembl ID version numbers were stripped in order to facilitate translation into other identifier types.

[1] https://gdc.cancer.gov/access-data/gdc-data-transfer-tool

[2] https://portal.gdc.cancer.gov

[3] https://portal.gdc.cancer.gov/legacy-archive

[4] https://docs.gdc.cancer.gov/Encyclopedia/pages/TCGA_Barcode/

Table 2.1: *Sample of the TCGA CRC HTSeq table.*

	TCGA-A6-6654-01A-21R-1839-07	TCGA-AA-3972-01A-01R-1022-07
ENSG00000000003.13	2795	2943
ENSG00000000005.5	1	9
ENSG00000000419.11	2901	2020
ENSG00000000457.12	731	332

2.3.4 TCGA metadata analysis

To better understand the composition of the TCGA CRC cohort in preparation for further analysis, associations between different metadata features such as tumour grade and survival were explored through multiple measures of correlation. As both categorical and continuous features were present in the metadata, different association measures were employed when comparing different feature types. For example, tumour stage, a categorical feature, has four main categories (i, ii, iii, iv), while days to death, a continuous feature, could be any positive number. To compare tumour stage and days to death, Pearson correlation could potentially be employed if every category was first expanded into a dummy boolean variable (0 or 1). However, when many features are being compared, the size of the results quickly becomes difficult to interpret. To circumvent this, I applied measures of association that can compare across different feature types, so that a single association matrix comparing all features could be created. In the case that both features were continuous, Pearson correlation was used. In the case of both being categorical, Cramer's V (Cramér, 1999) was employed. To examine the association between a categorical and continuous feature, Pearson's correlation ratio was used. This analysis made use of the dython library for python, which I modified to handle missing values on a per-test basis[5].

[5]The library was previously only able to respond by filling in missing values with a placeholder, or dropping the entire sample or feature, which would produce difficult to interpret results for the TCGA metadata, for which many features and samples have incomplete data.

2.3.5 Preprocessing and normalisation of TCGA RNAseq read counts

Raw counts were prepared for inter-sample analysis by converting HTSeq counts into counts per million (CPM) using EdgeR's cpm() function. CPM is a simple expression unit for normalised counts that may be used to correct for inter-sample sequencing coverage. For each sample, the CPM is given by the following equation, where r_g denotes the number of reads mapped to a gene g, and R is the total reads in the sample:

$$CPM_g = \frac{r_g}{R} \cdot 10^6 \tag{2.7}$$

Genes were filtered to exclude lowly and non-expressed genes. This filtering was achieved by excluding genes that fell below a threshold of 3 CPM in at least 100 samples, as exemplified with the following R code:

```
keep <- rowSums(cpm > 3) >= 100
```

Further count normalisation to compensate for transcriptome composition bias was performed with edgeR's calcNormFactors() function, using the trimmed mean of M values with singleton pairing (TMMwsp) method, an extension of TMM which can better handle a large proportion of zeroes (Robinson & Oshlack, 2010).

2.3.6 Obtaining consensus purity estimates

The variation in tumour purity between samples, i.e. the percentage of a sample actually containing tumour cells, is known to influence gene expression differences between samples (Haider *et al.,* 2020), and so needs to be adjusted for prior to analysis. It is not clear whether tumour purity is actually an intrinsic, biological property of the tumour that should be examined closely, or simply an extrinsic, technical artefact that should be treated as a batch effect. Aran *et al.* make arguments for both the extrinsic and intrinsic case, citing the fact that tumour samples from the same patient have similar purities as evidence for the latter, however by showing tumour purity is

not well correlated with factors such as survival they suggest that it may be more of the former. Of interest, the CMS4 subtype was found to be significantly less pure than the other subtypes in CRC. The reality may be that tumour purity is influenced by both intrinsic and extrinsic features, and certainly more study into this is warranted.

I obtained estimated tumour purity for the TCGA CRC samples from Supplementary Data 2 of Aran *et al.* 2015, in the form of consensus purity estimates (CPE). The CPE as described by Aran *et al.* is a normalised consensus of four different methods for purity estimation, which includes somatic copy-number data, (ABSOLUTE), gene expression profiles of immune and stromal genes (ESTIMATE), methylation of immune-specific CpG sites (LUMP) and image analysis of haematoxylin and eosin stained slides (IHC).

2.3.7 Traditional differential gene expression analysis

Identification of differentially expressed (DE) genes was conducted using edgeR for tumour vs. normal and male vs. female comparisons, both to validate the sample metadata, and to compare to the PSDE results. Testing of DE genes between sexes used only tumour samples, so that the genes identified were only those specifically DE in tumours. These tests were performed using a simple design matrix as follows:

```
design <- model.matrix(~Platform+Purity+Group)
y <- estimateDisp(y, design, robust=TRUE)
fit <- glmQLFit(y, design)
qlf <- glmQLFTest(fit, coef=3)
topTags(qlf)
```

Where Platform is sequencing platform (e.g. Illumina Hiseq or Illumina GA), Purity is the continuous consensus purity estimate, and Group includes either sex or sample type levels. This design models sequencing platform and purity as a batch effect, so that contrasts between the groups of interest can be made with these effects removed. EdgeR's glmQLFit() was used to fit a negative binomial generalised linear model for each gene, after which glmQLFTest() was used to perform a quasi-likelihood (QL) F-test (Lund *et al.,* 2012) to determine significantly differentially expressed genes between

the selected conditions. The QL F-test is suggested for RNAseq datasets due to its consideration of uncertainty of gene dispersion estimates (Robinson *et al.,* 2010).

2.3.8 Batch effect correction and transformation prior to PSDE gene analysis

Preliminary examination of the data revealed that a significant batch effect was present due to the two sequencing platforms used by TCGA (Illumina Hiseq and Genome Analyser). This was corrected for using removeBatchEffect() function in limma (Ritchie *et al.,* 2015) which uses a linear model for adjustments. This approach was also used to adjust for differences in tumour purity. Accounting for tumour purity in differential expression analysis is complicated as not all genes will be DE between the tumour and normal states, so scaling all genes by a constant factor may be inappropriate. TCGAbiolinks for example includes a function TCGAtumor_purity(), which simply filters samples based on whether they pass an arbitrary purity threshold (Mounir *et al.,* 2019), sidestepping the issue entirely. For the purposes of modelling patient-specific gene expression, I decided that tumour purity should be adjusted for as a technical artefact rather than considered as a biological phenomenon. Therefore, purity was added as a continuous numeric batch effect which was compensated for in the matrix provided to limma. The removeBatchEffect() function from limma was used rather than ComBat (from the R *sva* package) (Johnson *et al.,* 2007) due to ComBat's inability to correct for more than one effect at once.

Adjusted CPM values were $\log_2$ transformed so that the distribution of counts for each gene were approximately normal (in comparison to distributions of non-logged counts, which are better approximated by a negative binomial distribution (see Figure 2.2). PSDE genes were then identified as outlined in section 2.3.1.

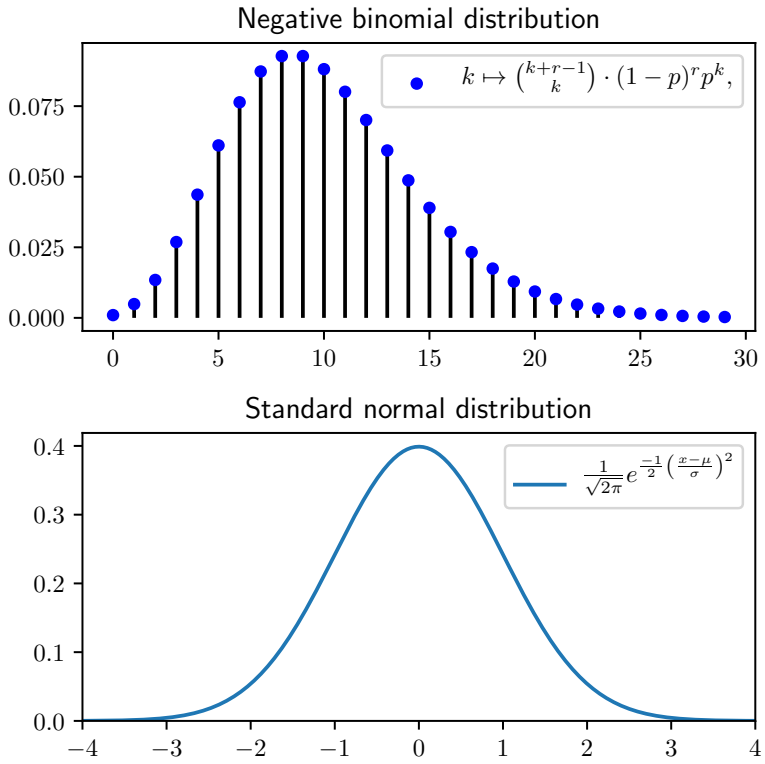


Figure 2.2: *The negative binomial distribution (top) is often used to model CPM. When counts are log-transformed, they more closely approximate and may be modelled by a normal (Gaussian) distribution (bottom). The negative binomial distribution is sampled here at* $r = 10$ *and* $p = 0.5$.

2.3.9 Data exploration and cleaning

Dimensionality reduction methods including Principal Components Analysis (PCA) (from Scikit-learn (Pedregosa *et al.*, 2011)) and Uniform Manifold Approximation and Projection (UMAP) (McInnes *et al.*, 2018) were used to explore the data, identify outliers, and to verify the effectiveness of batch correction. Sample read mapping quality was assessed by summing reads identified as "no feature", "multimapped", or "ambiguous" by HTSeq2 (Figure 2.3). Based on this, low quality samples were defined as those with a high proportion (>50%) of reads assigned to the aforementioned mappings, and were excluded from future analysis.

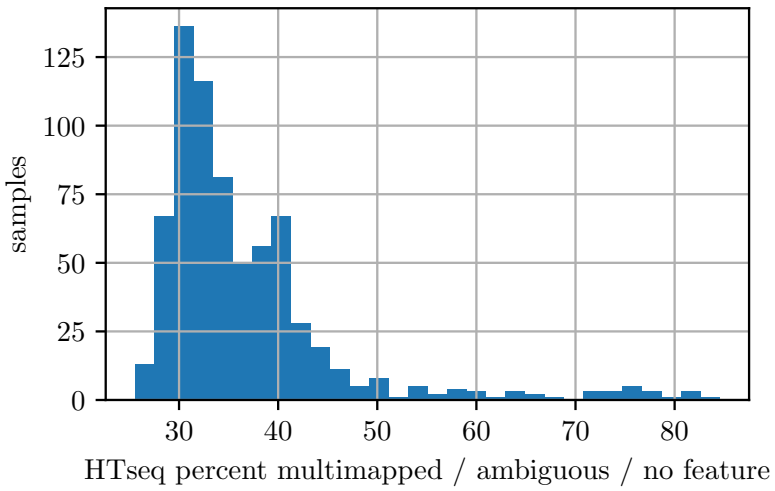


Figure 2.3: *Percentage of reads not mapped to unique gene features across all samples (identified as "no feature", "multimapped", or "ambiguous"). Samples where >50% of the reads were either multimapped, ambiguous, or aligned to no feature were considered low-quality, and discarded.*

2.3.10 Hierarchical clustering of samples based on PSDE genes and pathways

Hierarchical linkages were constructed via agglomerative clustering using the Ward variance minimization method (Ward, 1963) as implemented in SciPy. Hierarchical clustering algorithms may be agglomerative or divisive - in agglomerative clustering, every sample is initially its own cluster, and cluster pairs are iteratively merged to construct the hierarchy. That is, clustering begins by considering local structure. Divisive clustering is simply the inverse, in which the entire dataset starts as one cluster and is continuously divided, i.e., clustering begins by considering global structure.

2.3.11 Consensus clustering

Consensus clustering is a resampling-based methodology to assess the robustness of results obtained from a clustering algorithm. In short, it involves repeated sub-sampling and re-clustering of data to create a consensus matrix representing how frequently each pair of samples is clustered together across multiple analyses. The

consensus matrix M may then be transformed into a distance matrix by taking $1 - M$, which can be used to calculate a final consensus hierarchical clustering.

Consensus clustering was implemented as described by Monti *et al.* (Monti *et al.,* 2003). My multithreaded implementation is written in Python with numpy arrays, and resamples based on a given percentage of samples. Unless specifically mentioned, all consensus clustering was run 1000 times, and was resampled to 75% of the data for each run.

2.3.12 UMAP preprocessing for clustering

UMAP is a non-linear (in comparison to linear methods such as PCA) dimensionality reduction tool which has become extremely popular for clustering of transcriptomics data, especially single-cell (Becht *et al.,* 2019). UMAP reveals global structure without destroying local structure in large datasets, and has been shown to sometimes drastically improve the accuracy of clustering (Allaoui *et al.,* 2020). For the purpose of preprocessing data for consensus clustering, UMAP was run independently for each consensus fold with a different random seed. This controlled for the variability it introduced as a stochastic method which can be sensitive to initial conditions. The official Python implementation of UMAP was used for visualisation purposes (McInnes *et al.,* 2018), however for consensus clustering an alternative R/C++ implementation uwot[6] was used as I found it had better cross-platform reproducibility. For clustering purposes, n_neighbors was set to 30, and n_components reduced to 10, as recommended in the UMAP documentation[7].

2.3.13 Determination of optimal cluster number

The silhouette coefficient (Rousseeuw, 1987) was used for validating cluster composition and to assist with determining an optimal number of clusters during PSDE gene informed clustering. The silhouette_samples() and silhouette_score() functions from

[6] https://github.com/jlmelville/uwot

[7] https://umap-learn.readthedocs.io/en/latest/clustering.html

Scikit-learn were used to determine the silhouette coefficient from consensus clusters at each number of clusters k.

Consensus cluster scores were also used to assist in determining an optimal k. The cumulative area under the curve was used as described by (Monti *et al.,* 2003) to assess the largest improvement in consensus score, and per-item consensus scores were also used to construct "consensus plots" similar to silhouette plots.

2.3.14 Clustering visualisation

Clustered heatmaps with row and sample dendrograms were produced by using seaborn's clustermap() function. Heatmaps visualised $\log_2$ CPM Z-score normalised data on a per-gene basis. Consensus molecular subtype (CMS) labels for the TCGA CRC cohort as defined by Guinney *et al.* were obtained from Synapse[8] and highlighted next to the heatmap for comparison to PSDE gene or pathway defined molecular subtypes. To visualise how different clusterings compared, diagrams in the "alluvial" style were generated using a Python script[9].

2.3.15 Mutation analysis

Mutation data were downloaded as MAF (Mutation Annotation Format) files from the GDC. The GDC offers MAF files generated via multiple different variant caller algorithms, including MuSE (Fan *et al.,* 2016), MuTect, VarScan and SomaticSniper. MAFs as produced by the MuTect algorithm were chosen due to widespread use and recommendation of this algorithm in the literature (Xu, 2018). MAF files for both colorectal cancer TCGA projects (COAD and READ, for colon or rectal adenocarcinoma) were downloaded and concatenated. Mutation enrichment analysis was conducted for clusters using a Fisher's exact test, based on the principles of pathway over-representation analysis, to determine which specific mutations were significantly enriched in clusters of patients compared to the wider cohort.

[8] https://www.synapse.org/#!Synapse:syn4978511

[9] https://github.com/vinsburg/alluvial_diagram

2.3.16 Survival analysis

Kaplan-Meier analysis was employed to assess differences in patient survival. Cox regression analysis was used to assess whether certain metadata variables influenced survival. Both of these types of survival analysis were conducted via the Lifelines package (Davidson-Pilon *et al.,* 2019) for Python (version 0.25.4). Pairwise and multivariate logrank tests were employed to determine statistical significance. In some cases, the assumption of proportional hazards was violated, i.e., survival curves crossed over each other, and so survival was also compared using 5 year restricted mean survival time (RMST) (Royston & Parmar, 2013), and tests of survival differences at the 5 year time point assessed using a Chi-squared test, first applying log(-log) transformation to recover additional power (Klein *et al.,* 2007).

2.3.17 Unsupervised network partitioning of IMEx data

The graph-tool Python library (version 2.37) (T. P. Peixoto, 2017) was used to load the IMEx PPI network as a graph data structure following filtering for human interactions, which enabled simpler modification and filtering. All interactions in IMEx are assigned an MIScore which represents the confidence in the experimental evidence backing each interaction, normalised between 1 and 0 (Kerrien *et al.,* 2012). Edges with an MIscore <0.6 (0.6 being the minimum level regarded as high confidence interactions by IntAct (Villaveces *et al.,* 2015)) were removed so that only high-quality interactions were retained. To detect large-scale topological features, the nested stochastic block algorithm (T. P. Peixoto, 2014) was used to partition the network into unsupervised partitions based on topology. Partitions were analysed via gene set enrichment analysis using pathway databases including Gene Ontology (Ashburner *et al.,* 2000) and KEGG (Kanehisa & Goto, 2000).

2.3.18 Localisation analysis of PSDE genes in the human high-confidence interactome

PSDE genes as determined for each patient in The Cancer Genome Atlas's (TCGA's) CRC cohort were localised to the network partitions as previously detected by graph-tool and visualised. Ratios of up-regulated and down-regulated genes were compared to determine whether the distribution of PSDE genes was consistent across regions detected by unsupervised partitioning. Significance testing of ratios was performed using `prop.test` from R 3.6.3, adjusted for multiple testing using the Benjamini-Hochberg procedure implemented in the statsmodels Python package (Seabold & Perktold, 2010).

2.3.19 Identification of network modules using PSDE genes

Hierarchical HotNet (Reyna *et al.,* 2018) was used to identify topological modules using PSDE genes as prior information. The high-confidence IMEx human interactome was used as the base network from which modules were identified. This was performed for up-regulated and down-regulated genes separately, using the frequency of PSDE gene occurrence as prior information scores for HotNet.

To investigate PSDE genes in the context of patient-specific network topology, subnetwork modules of interest were also examined for individual patients. Hierarchical HotNet was provided with a list of PSDE genes for each patient to obtain statistically significant PPI subnetworks. Alternative methods for module identification were also explored, including providing prior information scores based on the frequency of PSDEs in different subtypes (e.g. from Consensus Molecular Subtypes (Guinney *et al.,* 2015) or PSDE-informed clusters (see Chapter 2)), and providing these results as inputs to the Hierarchical HotNet algorithm.

The resulting network modules were analysed using gene set enrichment analysis using pathway databases including Gene Ontology (Ashburner *et al.,* 2000) and KEGG (Kanehisa & Goto, 2000). Modules were visualised using graph-tool and matplotlib (version 3.1.3) (Hunter, 2007).

2.3.20 PLS-DA model construction and validation

PLS-DA models were constructed using the Python-based Scikit-learn package (version 0.24.1) (Pedregosa *et al.,* 2011). A subset of gene expression data (as informed by PSDE analysis) was used to train models on a binary classification task (four different models for patient survival status at 1, 3, 5 and 10 year time points). Survival data were obtained from TCGA metadata. Cross-validation of the model was done using the K-Fold method with 10 partitions (i.e., 10-fold cross validation). Data were split into 10 partitions at random, and the training repeated 10 times, with each repeat leaving out a different partition to be used as a validation. The performance of the model at classifying the validation partitions was evaluated using area under the receiver operating characteristic (AUROC).

2.3.21 Development environment

The majority of my exploratory code was written using Jupyter notebooks, an evolution of the IPython interactive computing environment (Perez & Granger, 2007). The Jupyter notebook format means that both code and results (e.g. graphs) are preserved in a single document. Code which was reused in multiple notebooks was refactored out into a separate library, biomodule, which contains implementations of consensus clustering, over-representation analysis, id conversion tools, and more. The tools that have been used for data preparation and analysis are mostly from the Python (Python Software Foundation, 2020) data science stack, such as pandas, numpy (Harris *et al.,* 2020), matplotlib (Hunter, 2007), and scipy (Virtanen *et al.,* 2020). Where R (R Development Core Team, 2013) packages were used, they were called into Python using rpy2, so that a consistent environment was maintained even when R-exclusive software such as edgeR (Robinson *et al.,* 2010) was required. Many of the statistical algorithms used (e.g. agglomerative hierarchical clustering, dimensionality reduction algorithms, statistical measures, etc.) were those implemented in scipy. Versions of key Python and R packages used may be found in tables 2.2 and 2.3, while the runtime environment is summarized below:

- CPython version 3.7.5 (2019-11-26)
- R version 3.6.3 (2020-02-29)
- Platform: x86_64-pc-linux-gnu (64-bit)
- Running under: Arch Linux

Table 2.2: *Key Python modules and versions.*

	Version
IPython	7.9.0
numpy	1.17.4
pandas	0.25.3
scipy	1.3.3
matplotlib	3.1.3
lifelines	0.22.8
sklearn	0.21.3
rpy2	3.2.4

Table 2.3: *Key R packages and versions.*

	Version
sva	3.34.0
BiocParallel	1.20.1
genefilter	1.68.0
mgcv	1.8-31
nlme	3.1-144
edgeR	3.28.0
limma	3.42.0

2.3.22 Availability

Full access to code, results, and documentation, including Jupyter notebooks and standalone scripts, is available on the Lynn lab Bitbucket repository, located at

https://bitbucket.org/lynnlab.

2.4 Results

2.4.1 Development of a method to identify patient-specific differentially expressed genes among CRC patients

To investigate the potential of transcriptomic inter-tumour differences to inform patient stratification, I developed a methodology to identify a set of genes for each patient tumour which I described as patient-specific differentially expressed (PSDE) genes. This methodology involved examining the gene expression distribution for each gene across the entire patient cohort, and applying thresholds to these distributions to identify genes that were unusually more highly or lowly expressed compared to other tumour samples. I tested this method using colorectal cancer (CRC) patient RNA-seq samples from The Cancer Genome Atlas (TCGA) (n=550 post quality control). An advantage of using this cohort was that the Consensus Molecular Subtypes (CMS), the current state of the art in transcriptomics-based classification in CRC (Guinney *et al.,* 2015), provided a pre-existing molecular stratification of these patients which was useful for comparison to the results of the PSDE gene method.

Preprocessing of CRC RNA-seq data removes noise and batch effects

Transcriptomics data for TCGA's CRC cohort (combined TCGA-COAD and TCGA-READ, n=633) was downloaded from the NCI Genomic Data Commons (GDC) and preprocessed using a pipeline based on edgeR. The PSDE gene method by its nature is sensitive to technical artefacts and normalisation. To minimise the impact of batch effects and technical noise on the method, stringent quality control of genes and samples was performed, resulting in 13,558 genes across 550 unique patient tumour samples being retained. Visualisation of the total CPM distribution before and after filtering confirmed that this quality control was effective at removing lowly expressed genes which would otherwise bias the PSDE results (Figure 2.4).

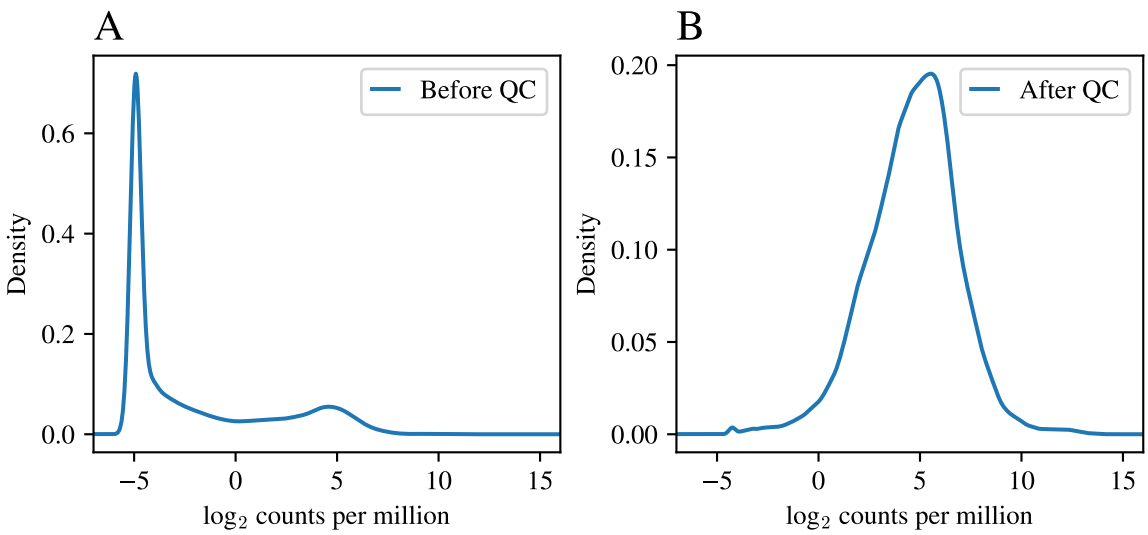


Figure 2.4: *Log_2 CPM distribution of all genes across all samples before (A) and after (B) quality control. Genes which were expressed at <3 CPM in at least 100 samples were excluded.*

As PSDE analysis aimed to detect subtle inter-patient heterogeneity, it was extremely sensitive to noise and batch effects. TCGA sample aliquots are assigned a 22-digit barcode (e.g., *TCGA-A6-6654-01A-21R-1839-07*), which describes tissue source site, participant, sample type, vial number, portion, and plate number. Batch effects within TCGA RNA sequencing samples have frequently been reported in the literature (Lauss *et al.*, 2013), which often include the features described in the barcode. Other factors such as date of sample collection and sequencing platform can also contribute to batch effects. I was able to show, concordant with the findings of Guinney *et al.* during creation of the CMS, that the effect of sequencing platform used (Illumina Hiseq or Genome Analyser) is extremely strong within the TCGA CRC cohort, making up the majority of the observable variation. Using a 2D UMAP embedding to visualise all samples (Figure 2.5), data from the two sequencing platforms were clustered nearly entirely separately.

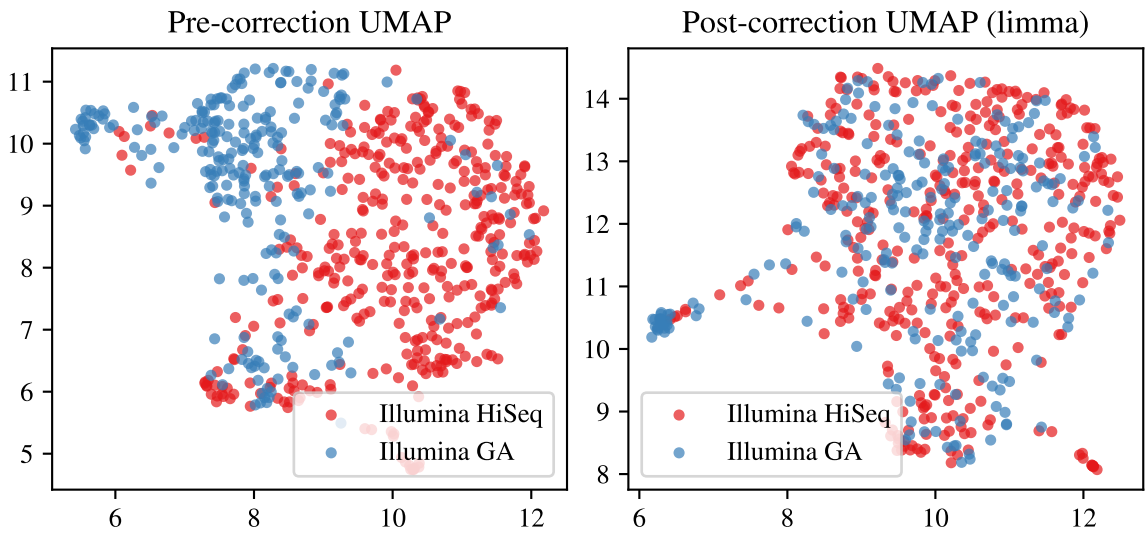


Figure 2.5: *2D UMAP embedding of log_2 CPM transcriptomic data for TCGA CRC samples before (left) and after batch effect correction with limma (right). Sequencing platform for each sample is highlighted in red for Illumina HiSeq and blue for Illumina Genome Analyser.*

I corrected the platform effect using the removeBatchEffect() function from limma (Ritchie *et al.*, 2015) which uses a linear model to perform adjustments. Adjusting for batch effects such as this was essential as they would otherwise strongly influence inter-sample differences. I investigated other variables which are frequently cited as contributing to batch effects within TCGA datasets, such as processing centre, portion, and plate number, however none clearly contributed to technical variation as strongly as the platform effect. I found that the platform effect was somewhat confounded with "portion", a variable indicating how samples were divided by TCGA prior to analysis, which was likely a spurious correlation deriving mainly from the sequencing platform effect.

Cluster analysis identifies outlying groups with high variance

Following batch correction, visualisation via UMAP was performed to ensure that the platform effect was adjusted for appropriately (Figure 2.5). However, this also revealed that there were still two outlier groups in the tumour samples, which were variable enough to cluster separately from the bulk of other tumour and normal samples.

Using UMAP followed by k-means clustering, these outlier groups were selected and labelled as seen in Figure 2.6. PCA revealed that these two groups alone contributed to the majority of variance in the dataset (Figure 2.6, right). For PSDE analysis it was especially important to clean the input data thoroughly by either identifying the cause of these outliers and performing batch correction or removing them from the dataset. There is also precedent for excluding these samples; during creation of the CMS, Guinney *et al.* discarded a similar set of outliers using a PCA-based test (the boundary of this exclusion is shown on the right in Figure 2.6). Interestingly, the random-forest based CMS classifier was still able to classify most of those samples, despite them being excluded from the training data.

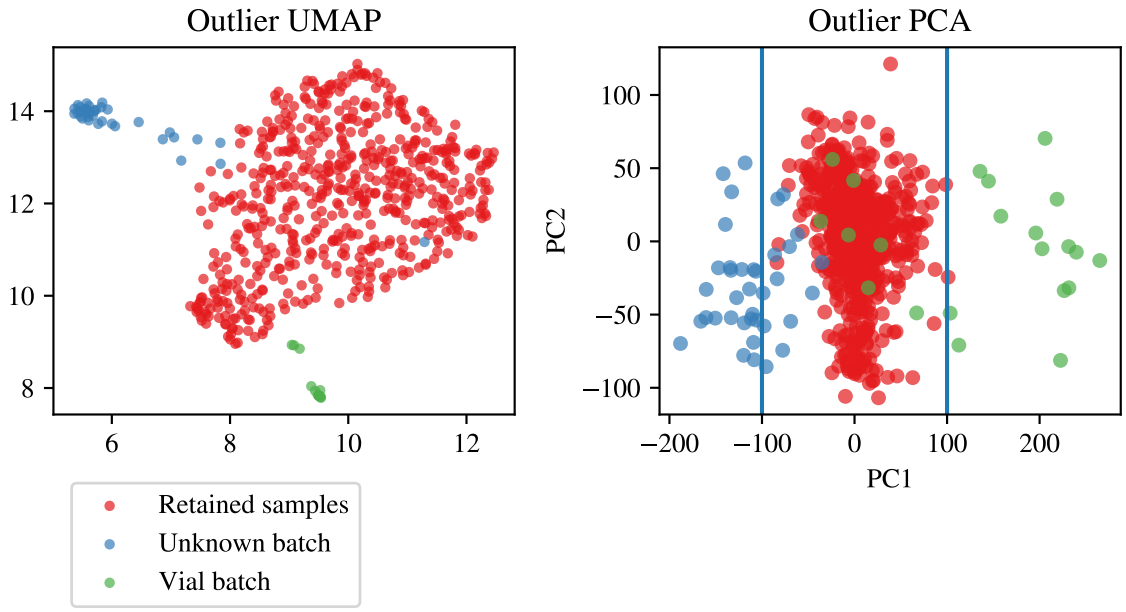


Figure 2.6: *Following gene filtering and batch effect correction, two outlier clusters of patient samples were detected within the TCGA CRC data. These clusters were selected and labelled using k-means clustering on UMAP embeddings (left). PCA revealed that the single largest component of variance in the dataset was due to these outliers (right). Vertical lines indicate where outliers were excluded by Guinney* et al. *for the creation of the Consensus Molecular Subtypes.*

It was apparent that the smaller of the two outlier groups (n=13) had an unusually high percentage of sequencing reads that were labelled as multimapped, ambiguous, or unmappable by HTSeq2 (Appendix Figure 6.4). Furthermore, these samples were

almost entirely of identified as "vial A" samples (as described by the TCGA barcode). These samples were excluded from further analysis. The remaining larger outlier cluster (n=38) was more difficult to interpret, having no apparent explanation based on the metadata. While the vast majority of tumour samples were broadly similar, including the known subtypes such as metastatic and MSI subtypes, these outliers were as distinct from other tumour samples as normal samples. To investigate further, I looked for samples in the outlier cluster which were sequenced multiple times. One sample in the cluster was sequenced in triplicate (*TCGA-A6-2684*), and while one run was found within the outlier cluster, the other samples clustered within the main group. Although it is possible this was due to mislabelling, this difference within the same sample strongly suggests the cluster was result of an unknown technical batch effect. Due to these observations this cluster was excluded.

Differential gene expression analysis validates patient metadata

Using differential gene expression (DGE) analysis to compare normal and tumour samples, I verified that the data labels were correct by identifying genes known to be overexpressed in colorectal cancer. DGE analysis revealed many genes that were significantly up-regulated in tumour samples compared to control tissue (Figure 2.7). Among the most significantly DE genes were *CDH3*, which has been identified as a potential biomarker of CRC progression (Kumara *et al.,* 2017), and *ETV4*, a transcription factor known to promote CRC proliferation when over-expressed (Fonseca *et al.,* 2021).

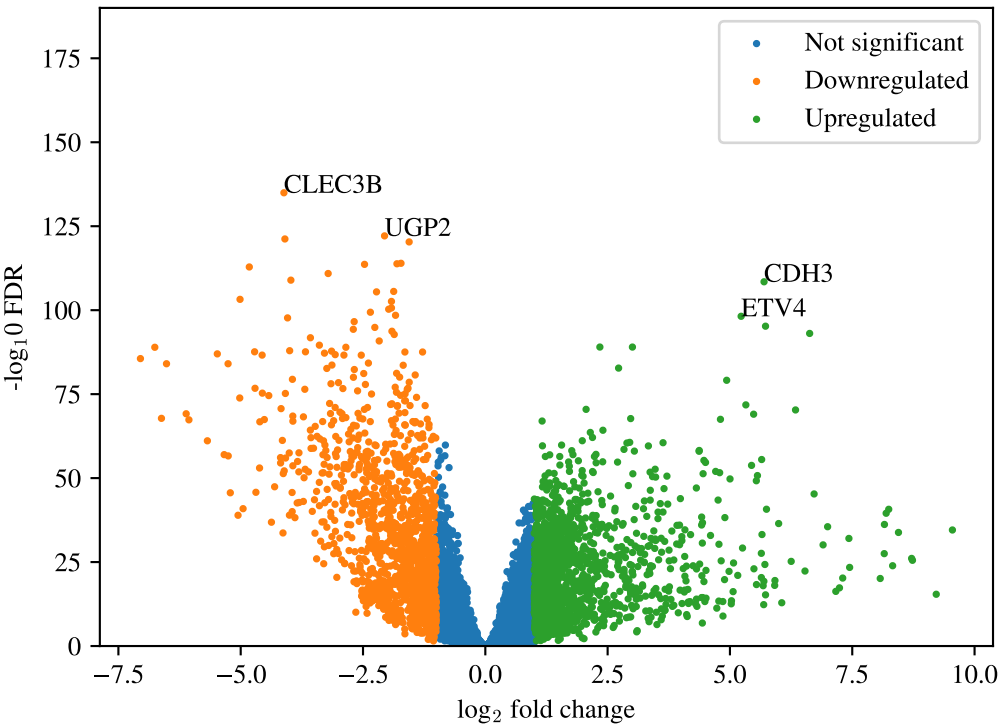


Figure 2.7: *Volcano plot highlighting differentially expressed genes between tumour and normal samples in the TCGA CRC cohort. Some of the most significantly DE genes are annotated.*

DGE analysis was also performed to compare male and female samples. 17 genes in total were identified as being sex-specific, including the X inactive-specific transcript *XIST*. *XIST* is usually an excellent sex-specific marker, and so was examined to verify the sex labelling in the sample metadata. I found *XIST* had near zero expression in males compared to females, at least in normal tissue (Figure 2.8, left), which appeared to validate the metadata. Tumour tissues however displayed much more variance in *XIST* expression (Figure 2.8, right), as did other sex-linked genes such as *ZFY*. As these sex-specific genes could confound later patient-specific analyses they were excluded, ensuring no sex-specific effects contributed to clustering and PSDE gene identification. While they were excluded for this analysis, these sex-implicated genes may in fact contribute to inter-patient tumour heterogeneity. Indeed, the Y-linked genes may be more important to cancer development than previously believed

(Kido & Lau, 2015). Previous research (Weakley *et al.*, 2011) has indicated that gene amplification of *XIST* occurs in microsatellite-unstable CRC tissues, however only 1 of the 21 male TCGA tumour samples with *XIST* expressed at >1 CPM were defined as the microsatellite instability subtype.

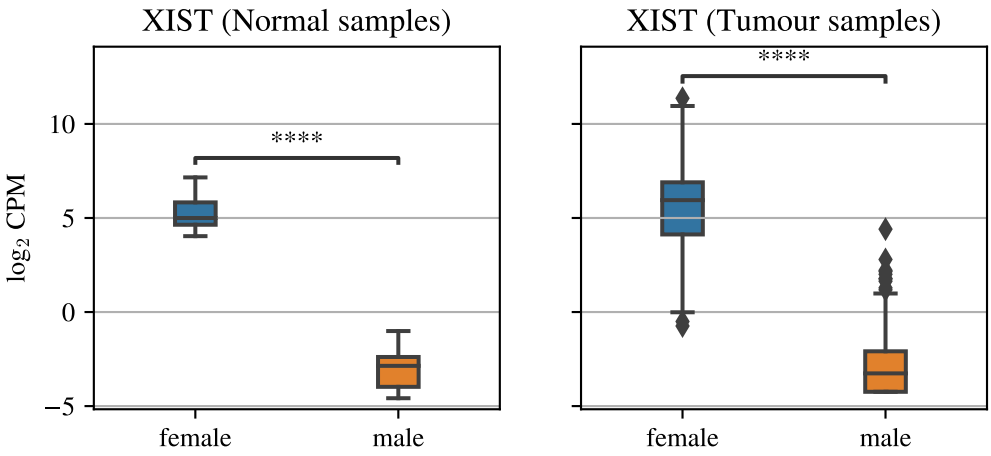


Figure 2.8: *Boxplots of XIST (log_2 CPM) expression in normal and tumour CRC tumour tissues, separated by patient sex. Normal samples had very little variance and were significantly different between sexes ($p = 1.2 \times 10^{-9}$, Mann-Whitney U-test). While the variance in tumour samples was much higher, males and females still exhibited significantly different expression ($p = 2.7 \times 10^{-105}$).*

Other than TNM stage, patient metadata is not predictive of outcome

TCGA metadata features corresponding to the transcriptomic samples were examined in detail to identify any relationships not already controlled by batch effect correction that could influence later interpretation of the data, including information on patient race, sex, survival time from diagnosis, treatment, and also details of sample preparation. Correlations between 23 different metadata features were combined into a single association matrix (Appendix Figure 6.1). This analysis revealed a significant difference in days to sample collection between some metadata categories, including race ($p = 1 \times 10^{-11}$). However, this did not translate into a significant difference in overall survival as measured by a log-rank test (p=0.99), and could potentially be attributable to experimental delays in collecting samples depending on patient condition and demographic.

A correlation (Cramer's V = 0.32) between survival and TNM stage was also identified, as would be expected. This correlation was weak, however, the stages were separated into sub-stages (i.e. iii, iiia, iiib etc.) which may have influenced the strength of the correlation. After compressing stage information into only four main stages, survival at each stage was analysed with a Kaplan-Meier plot (Appendix Figure 6.2) and found to be significantly different between all stages except between stage iv and samples with unreported stage. Later stages indicate increased tumour size, with stage iv meaning distinct metastases are present. As expected, the survival probability decreased for patients with later TNM stages.

2.4.2 Identification of patient-specific differentially expressed (PSDE) genes

I hypothesised that once the batch effects and technical artefacts had been adjusted for, the transcriptomic heterogeneity of each patient tumour in a given cohort could be characterised by a small number of patient-specific differentially expressed (PSDE) genes. PSDE genes were defined based on the intersection of two thresholds, one based on fold change in gene expression from the median, and one based on Z-score (see methods section for formal definition). PSDE genes defined in this way could then be divided further into two categories, up and down regulated, based on whether a PSDE gene's expression was over or under-expressed relative to the median cohort level. I found that the combination of the two thresholds was effective in preventing too many genes per patient being identified as PSDE. For example, I found that some genes had extremely high variance, spanning a wide range of CPM values, such as the *FGR* gene (Figure 2.9). For this gene, fold changes >2 were common, however the Z-score based threshold prevented it from being classified as a PSDE gene in too many samples. This strategy was also effective when CPMs were quite low, where relatively small absolute changes in gene expression could represent large fold changes.

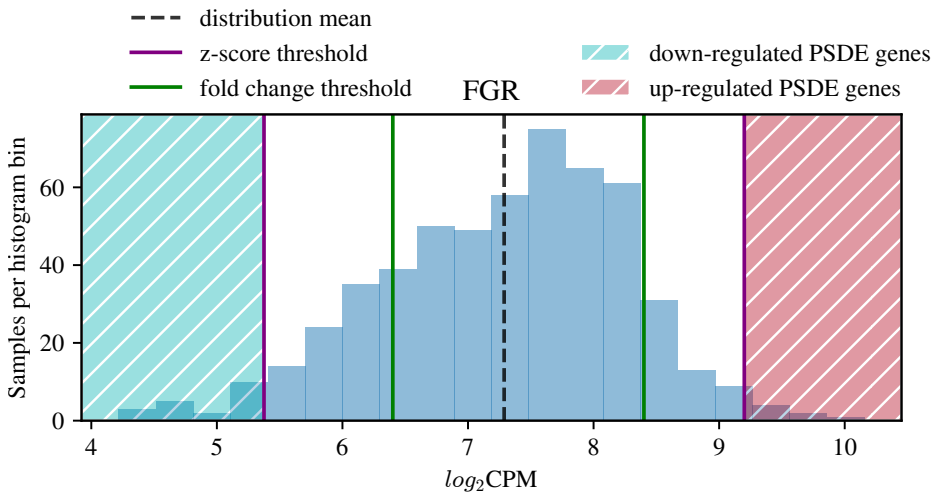


Figure 2.9: *Distribution of FGR gene expression across 550 CRC tumour samples, displaying the thresholds used to define PSDE genes. FGR gene expression is shown as an example of the fold change threshold being less restrictive than the Z-score threshold.*

Conversely, a gene distribution may have very low variance, leading to the z-score thresholds representing relatively small gene expression fold changes. Such was the case for *HNRNPK* (Figure 2.10). In this case, the fold change threshold prevented these Z-score outliers from being classified as PSDE genes.

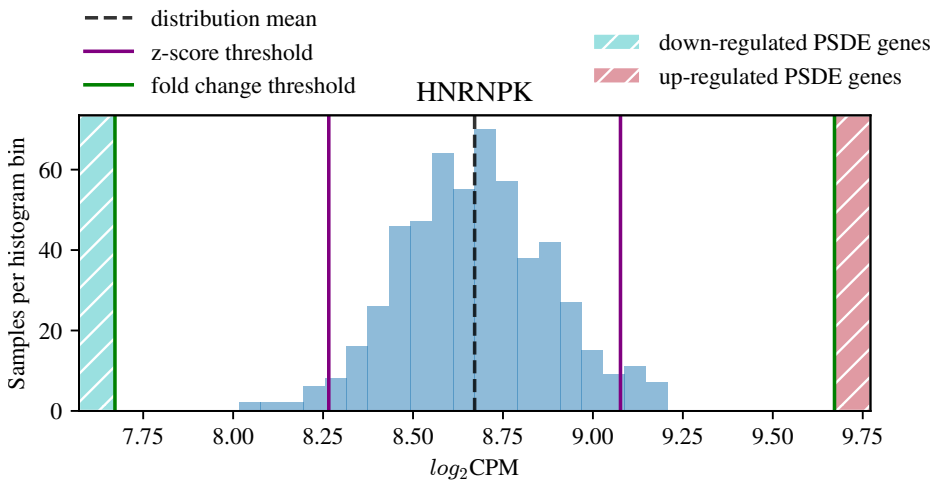


Figure 2.10: *Distribution of HNRNPK gene expression across 550 CRC tumour samples, displaying the thresholds used to define PSDE genes. HNRNPK gene expression is shown as an example of the Z-score threshold being less restrictive than the fold change threshold.*

More commonly however, the two PSDE thresholds were positioned in approx-

imately the same position in the distribution, as was the case for *KRAS* (Figure 2.11).

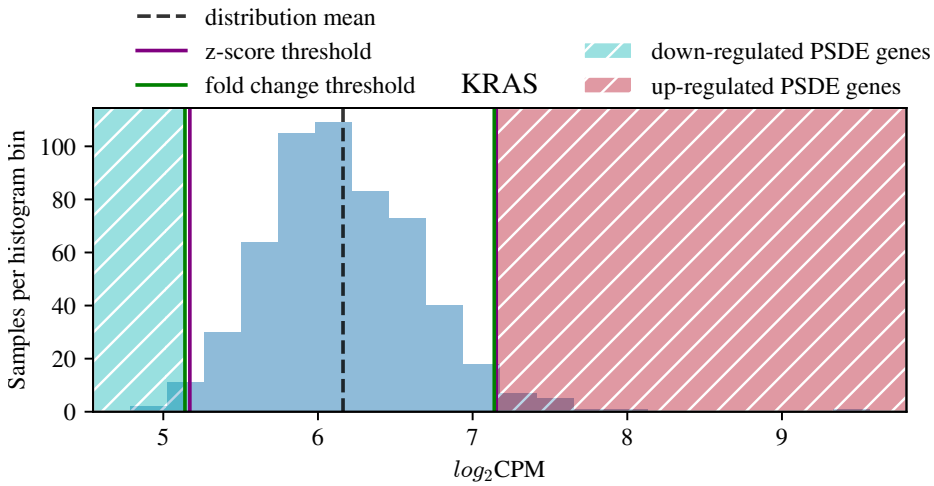


Figure 2.11: *Distribution of KRAS gene expression across 550 CRC tumour samples, displaying the thresholds used to define PSDE genes. KRAS gene expression is shown to demonstrate the thresholds being positioned in approximately the same location in the distribution.*

2.4.3 PSDE genes occur frequently within the CRC cohort

PSDE genes were identified for each of the 550 samples retained from the TCGA CRC cohort after quality control. A median of 403 genes were defined as PSDE per patient, with 176 up-regulated PSDEs and 227 down-regulated PSDEs (by median). The frequency of PSDE genes in tumour samples was compared to that of normal samples (Figure 2.12). As expected, normal samples had far fewer PSDE genes identified than tumour samples. It was also apparent that the number of up and down regulated PSDE genes were approximately the same. Strangely, an unusually high number of PSDEs were defined for some samples. For example, the maximum number of PSDE genes was 3902 for a single patient (*TCGA-CM-5341*). Interestingly, this effect was also seen within the normal samples, however with entirely different samples. 4016 PSDEs were defined for patient *TCGA-AZ-6605* in the normal samples, whereas in the corresponding tumour samples only 40 PSDEs were found. As no specific metadata or biological basis for these extreme outliers could be identified, they were excluded

as outliers from further analyses.

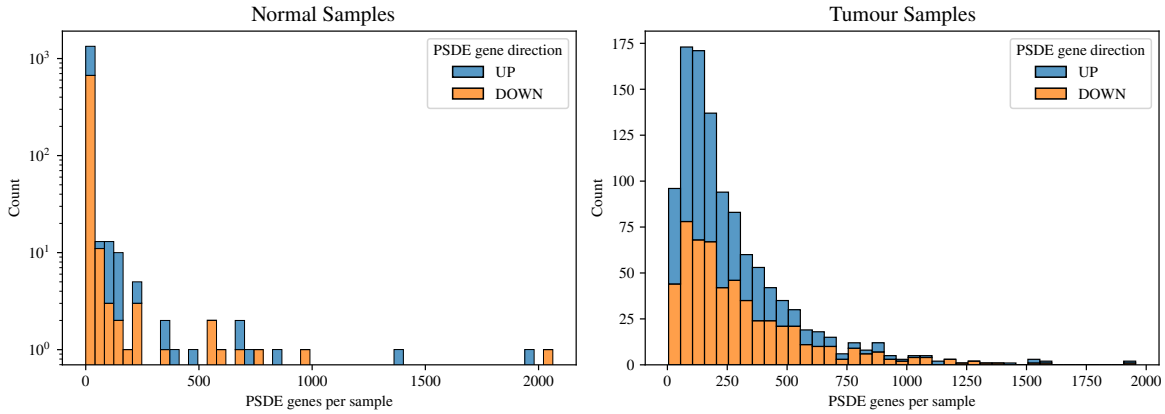


Figure 2.12: *Histogram of genes assigned as PSDE per sample, split into up and down regulated gene sets. Left: PSDE genes identified in normal samples, on a log scale due to a high proportion of zeros (n=51). Right: PSDE genes identified in tumour samples.*

I examined the overall expression of the most frequently identified up and down PSDE genes (95^{th} percentile) to see whether genes with lower or higher overall expression tended to be identified as up or down regulated PSDEs more frequently (Figure 2.13). There were no overlaps between the genes in these two sets, although for each of the frequent PSDE genes, on average 2.7 genes were identified as PSDE in the opposite direction. I found that down-regulated PSDE genes did in fact tend to have lower overall expression than average, however the distribution of up-regulated PSDE genes was not different to the overall distribution. Given the reduced expression level of down-regulated PSDE genes on average, I hypothesised that analyses based on them could be more susceptible to random noise, and therefore focused more heavily on up-regulated PSDE results in subsequent analyses.

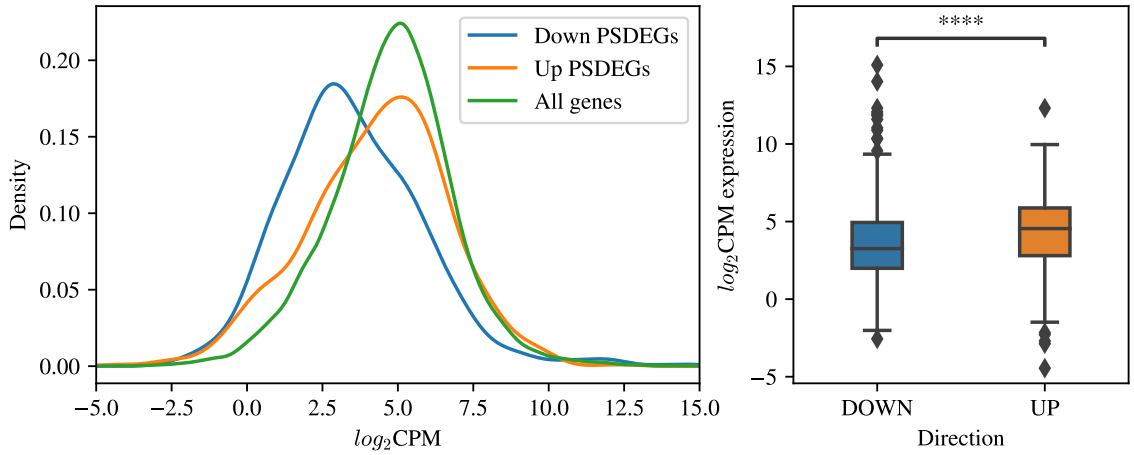


Figure 2.13: *Left: Mean gene expression distributions for the 95th percentile of most frequently up-regulated PSDE genes, 95th percentile of most frequently down-regulated PSDE genes, and all genes. Right: Expression comparison of up and down regulated PSDE genes. A Mann-Whitney U-test was used to assess statistical significance ($p = 4.4^{-14}$)*

2.4.4 PSDE genes are significantly enriched for pathways relevant to CRC

Pathway analysis was used to identify biological pathways and processes that were statistically enriched among PSDE genes. As most genes were identified as up or down regulated PSDE genes at least once, I took the top 95th percentile of PSDE genes by frequency for both up and down regulated PSDE genes to perform pathway over-representation analysis. I examined pathways from multiple databases, including MSigDB Hallmarks (Liberzon *et al.*, 2015), KEGG (Kanehisa & Goto, 2000) and Reactome (Fabregat *et al.*, 2017), plus additional gene sets containing common transcription factor interactors. More enriched pathways were identified for up-regulated PSDE genes (Figure 2.14) than for down-regulated PSDE genes (Figure 2.15). This was likely influenced by the on average higher expression of up-regulated PSDE genes (Figure 2.13) which resulted in a higher signal-to-noise ratio. Some databases were omitted in these figures due to a lack of unique pathway enrichments, i.e., the pathways that were identified were already represented in other databases.

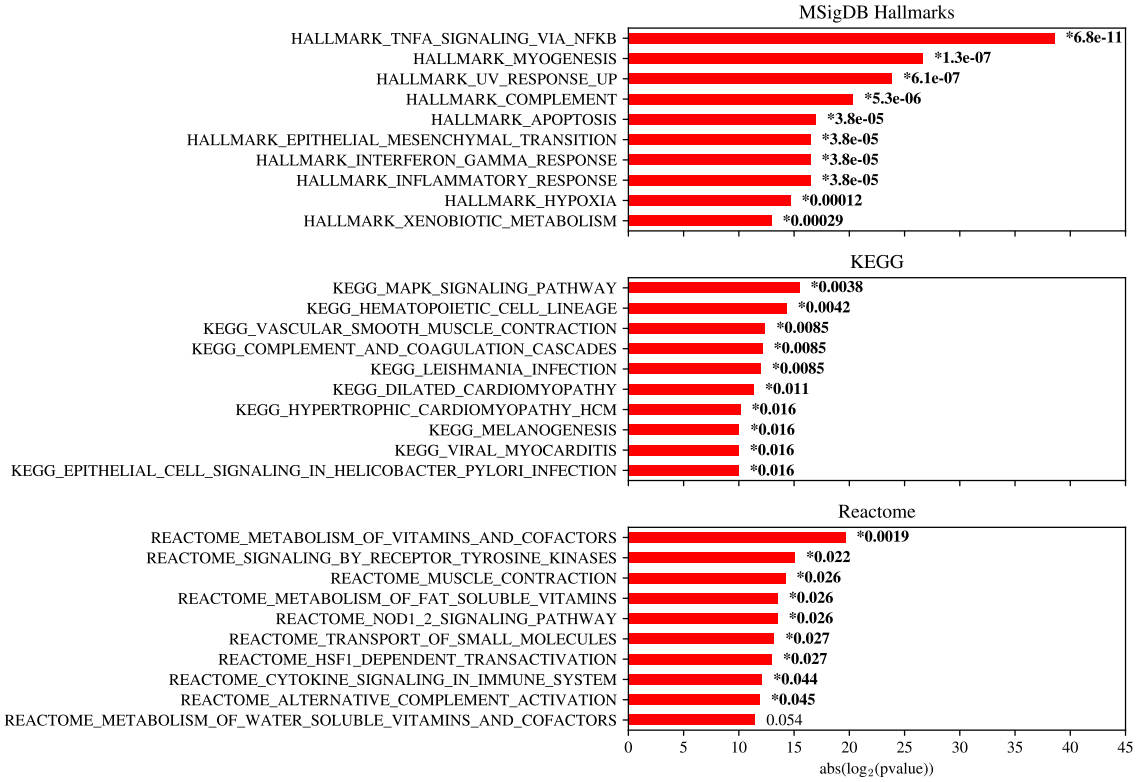


Figure 2.14: *Statistically enriched pathways identified among up-regulated PSDE genes. For each pathway database, the top 10 pathways are shown, ranked by P value. Pathways with FDR < 0.05 are highlighted in bold and with an asterisk.*

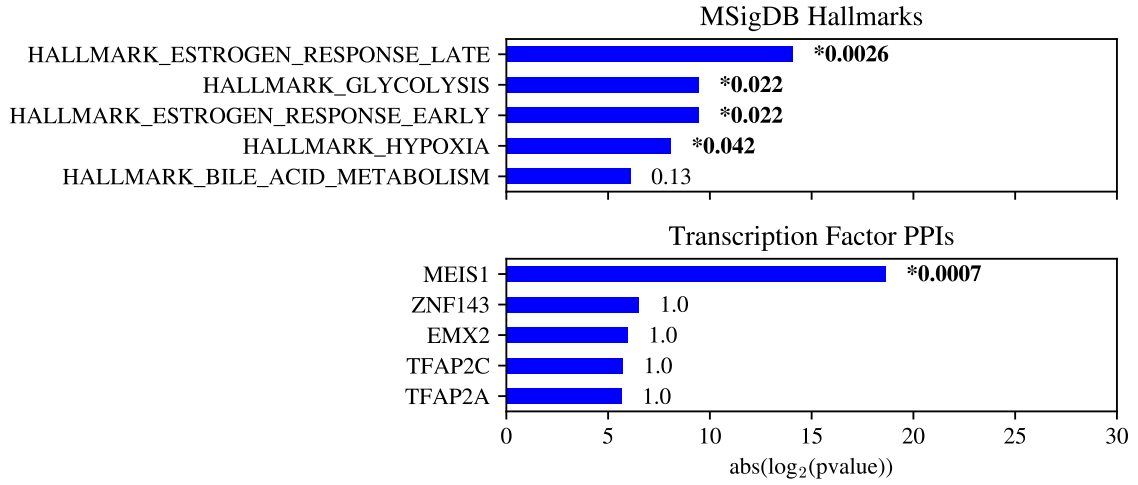


Figure 2.15: *Statistically enriched pathways identified among down-regulated PSDE genes. Only the top 5 pathways are shown due to the small number of significant results. Pathways with FDR < 0.05 are highlighted in bold and with an asterisk.*

Up-regulated PSDE genes were significantly enriched for pathways with roles in colorectal cancer such as MAPK signalling, a critical pathway for tumour development. As has been noted by previous studies using TCGA data (The Cancer Genome Atlas Network, 2012), genetic alterations in the RAS–MAPK pathway are common in CRC. In addition, adhesion and epithelial to mesenchymal transition (EMT) pathways, markers of late-stage metastatic progression, were also strongly enriched, as was the myogenesis pathway which has many genes in common with mesenchymal pathways. Up-regulated PSDE genes were also enriched for inflammation-related pathways, including TNF-α and IFN-γ signalling. Among down-regulated PSDE genes, the enrichment of interactors with the MEIS1 transcription factor was one of the most significant results. MEIS1 is known to be methylated in CRC tumours with the BRAF V600E mutation, which is associated with decreased expression of MEIS1 transcripts (Dihal *et al.,* 2013). Other pathways enriched in down-regulated PSDEs included estrogen signalling. ERβ (the primary estrogen receptor in CRC tissues) expression is typically lost during disease progression in CRC, and has thus been hypothesised to have a protective role in CRC (Caiazza *et al.,* 2015).

A priori, there was no expectation of enrichment for pathways relevant to CRC. The fact that both up and down PSDE genes were enriched for these key pathways suggests that the PSDE approach identifies disease-relevant genes despite the inherent noise of RNA-seq measurements. More importantly these data suggest that those pathways which have known functional roles in CRC are also those with the most inter-tumour heterogeneity. It should be emphasised that the PSDE method does not identify differences in expression found only between tumour and normal tissues, as there is no normal comparison made. For this reason, genes which are universally up or down-regulated across the entire tumour cohort will not be identified as PSDE genes. An example of this is Wnt signalling. This is a key pathway for CRC development and progression, yet it was not identified in the pathway analysis of PSDEs due to the relatively homogeneous activation of this pathway in all tumour samples. To show that this was the case, I again identified PSDE genes, this time using a combined cohort of the normal TCGA CRC samples plus a random sample of 20 tumour samples, with the hypothesis being that from this mixed cohort the expected tumour/normal

pathway differences would be possible to identify in a PSDE analysis. This resulted in an extremely strong enrichment of KRAS signalling, MYC targets, and Wnt / β-catenin signalling within the up-regulated PSDEs of the tumour samples (Figure 2.16), pathways which are not identified in the tumour-only PSDE analysis. This demonstrated that the PSDE method would identify tumour/normal differences in a cohort containing both tumour and normal samples.

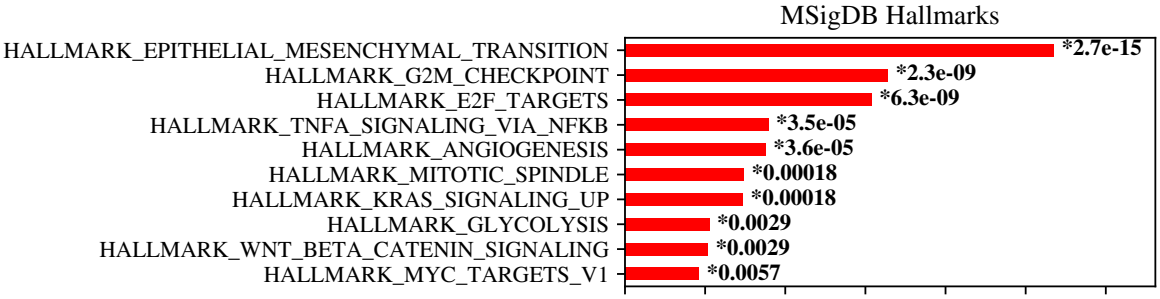


Figure 2.16: *The top 10 most significant pathways (FDR adjusted) found to be enriched among up-regulated PSDE genes defined from a combined analysis of 51 normal and 20 tumour TCGA CRC samples.*

2.4.5 PSDEs reveal novel patient clusters

I next investigated whether PSDE genes could be useful in predicting patient outcomes. Hierarchical clustering was performed on the 75th percentile of the most frequently defined PSDE genes, 2,978 in total, to capture the largest contributors to inter-patient heterogeneity. To ensure that the clustering was robust, 1000-fold consensus clustering was applied to the PSDE-gene informed RNA-seq data, with each iteration sampling 75% of the dataset. For each random sample of the data, UMAP was applied to reduce the dimensionality as a preprocessing step. A consensus matrix was then obtained which was used to construct the final hierarchical clustering.

Consensus and silhouette methods were used to identify the optimal number of clusters, k. I found the simpler "elbow" heuristic too unstable to be useful, with different samplings of the data leading to drastically different results. Consensus (Figure 2.17) and silhouette (Appendix Figure 6.19) methods agreed that $k = 3$ was most optimal for $k <= 4$. However, given that 4 robust divisions already exist

in the CMS groups, this would be insufficient given the existing known variation. In addition, the dramatic increase in maximum consensus score for $k > 4$ (Figure 2.17) indicated the presence of smaller clusters which would be better defined by a higher k. The mean consensus score across all clusters is not necessarily reflective of optimal cluster allocation, i.e., to maximise the mean consensus score, the maximum score for any single cluster is significantly reduced. The increased number of clusters at $k = 5$ led to much higher consensus scores for smaller clusters, which may be observed by the increase in maximum consensus score relative to the mean (Figure 2.17). Given that I would expect the PSDE method to identify rarer subtypes, I chose to prioritise smaller groups and use $k = 5$ for the final clustering, resulting in 5 PSDE-gene informed clusters (PICs) 1-5 (Figure 2.18, A) which included all TCGA CRC samples that passed quality control (n=550). Notably, this included samples which were unable to be clustered by the CMS classifier (Guinney *et al.,* 2015). These samples are referenced in Figure 2.18 A as "NOLBL". CRC samples that Guinney *et al.* did not attempt to classify were also included, and are referenced in Figure 2.18 A as "unknown". Using UMAP to visualise gene expression data (Figure 2.18, C), it was apparent that the variation present in the data was well captured by the PICs.

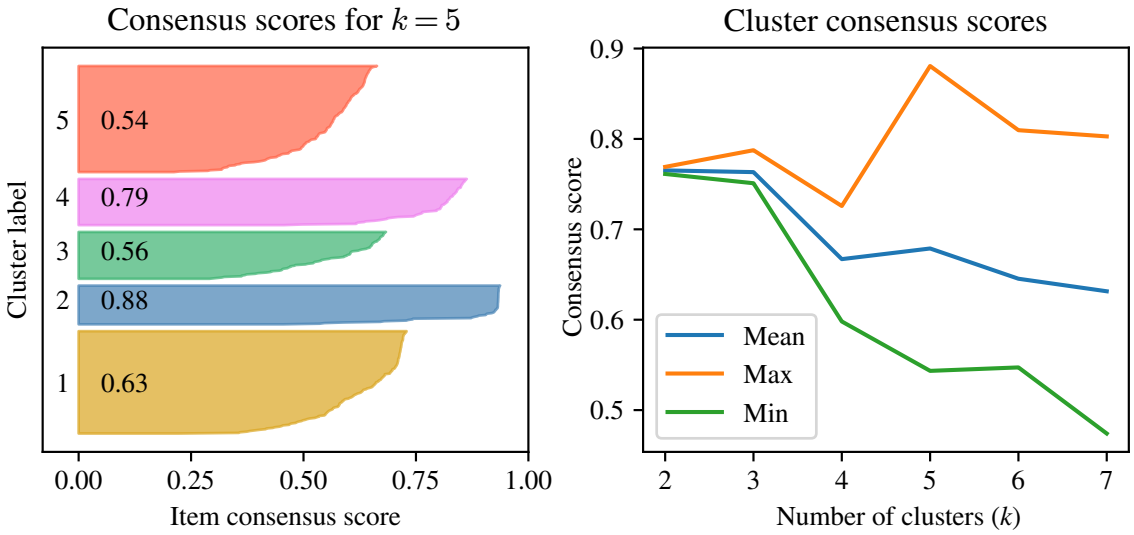


Figure 2.17: Left: *Consensus score visualised as a separated bar plot for $k = 5$ clusters. Each cluster is labelled on the vertical axis and assigned a unique colour. The per-cluster mean consensus score is indicated within each cluster. Each sample within a cluster is sorted in descending order by per-sample consensus score.* **Right**: *Mean, maximum and minimum overall consensus scores for each k evaluated ($k = 2 \ldots 7$). Maximum consensus was reached at $k = 5$, despite mean consensus peaking at $k = 3$.*

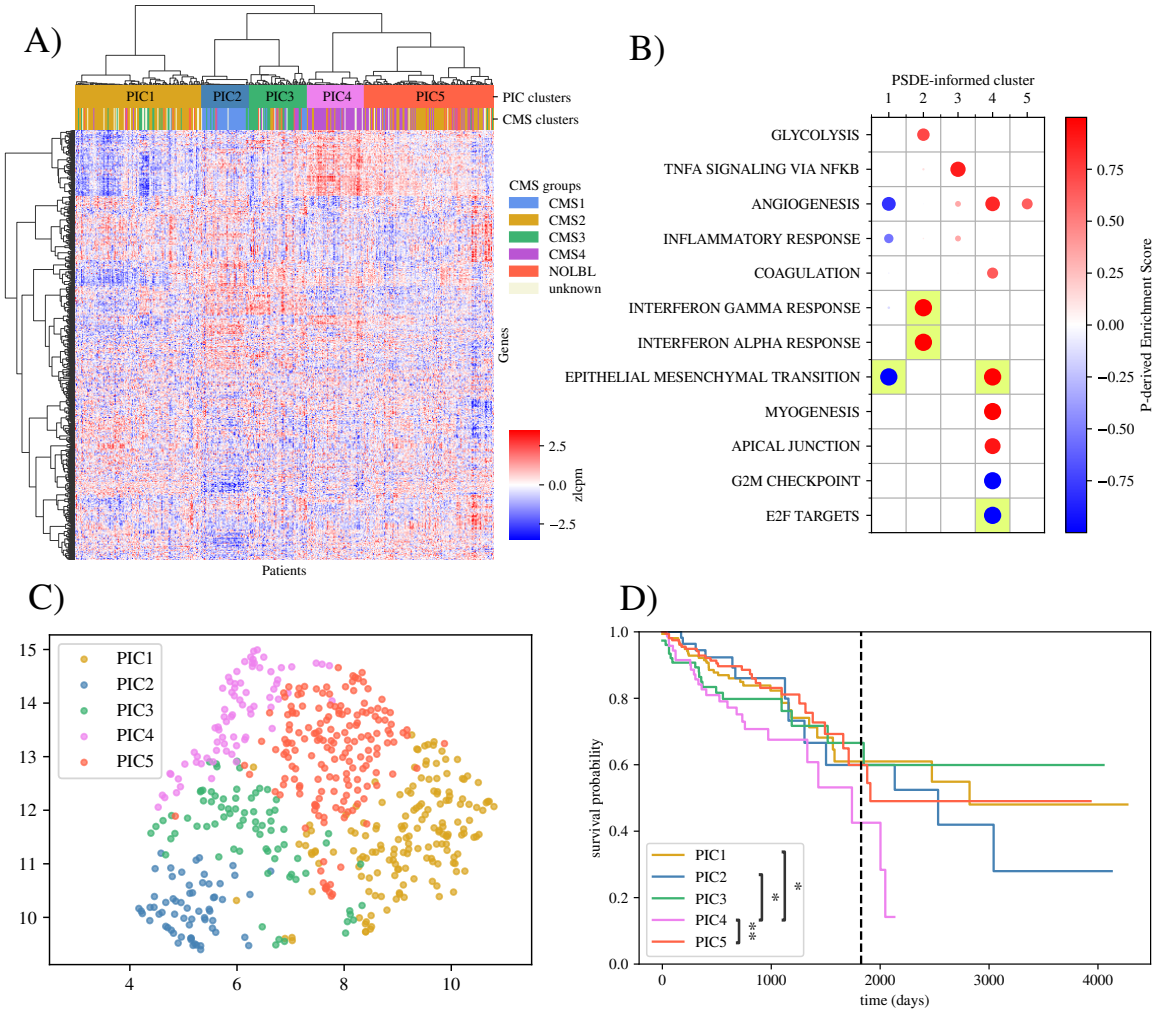


Figure 2.18: ***A)*** *Heatmap of PSDE gene expression. Z-score normalised* log_2 *CPM gene expression. The consensus dendrogram used to determine PSDE gene informed clusters (PICs) is shown on the horizontal axis. For comparison, CMS defined subtypes are annotated on the row below PICs.* ***B)*** *Pathway analysis of genes in each PIC. The MSigDB Hallmarks pathway database was used as the source of pathway annotation. Red dots represent pathways enriched among up-regulated PSDE genes, while blue dots represent enriched pathways among down-regulated PSDE genes. The size of dots is scaled proportionally with statistical significance. Yellow backgrounds indicate FDR* < 0.05*.* ***C)*** *2D UMAP visualisation of all tumour transcriptomic data (post-QC) with PIC clusters highlighted.* ***D)*** *Kaplan-Meier plot for patients as stratified by PICs. Statistical significance was assessed using pairwise logrank tests.*

Pairwise logrank tests found significant differences for PIC4 vs. PIC1 (p=0.004),

PIC4 vs. PIC2 (p=0.017) and PIC4 vs. PIC5 (p=0.034). PIC2 was consistently the most robust cluster in terms of clustering consensus score (Figure 2.17). Samples in PIC2 strongly overlapped with the well-described microsatellite-instability (MSI) subtype of CRC. Consistent with this there was also a strong overlap between PIC2 and the CMS1 subtype. Upon examining specific PSDE genes within this cluster, I found that *MLH1*, an essential component of DNA mismatch repair, was the most commonly identified down-regulated PSDE gene in patients annotated with the PIC2 subtype (occurring in 62% of tumours). PIC4 was a close second for consensus robustness, and aligned closely to the mesenchymal CMS4 subtype. Beyond these two clusters with strong molecular signatures, the remaining PICs had lower consensus scores and were more mixed in terms of their composition of CMS samples.

Patient samples in PICs are significantly enriched for specific pathways

To characterise the different PICs, I identified pathways enriched in each PIC using gene set enrichment analysis and the top 3 MSigDB Hallmarks pathways (by p-value) for each PIC were visualised (Figure 2.18 B). In terms of enriched pathways from the MSigDB Hallmarks database, some of the most significant results were identified within the PIC4 cluster. Many of these pathways relate to metastasis, including significant up-regulation of epithelial to mesenchymal transition (EMT), a process by which cells gain migratory properties required for the metastasis. Another pathway which approached FDR significance in PIC4 was upregulation of the apical junction complex (AJC). Alterations of the AJC may disrupt the intestinal mucosal barrier and are linked to progression and EMT in CRC (Gehren *et al.,* 2015). Angiogeneis and vascular endothelial growth factor (VEGF) signalling was also enriched for PIC4 patients, processes essential for proliferation of metastatic CRC. PIC4 samples also had strong down-regulation of certain cell cycle related pathways. This notably included the G_2-M checkpoint, which prevents activation of mitosis in the presence of DNA damage, as well as downregulation of targets of the E2F transcription factor, which is involved in tumour suppression and in proliferation (Kurayoshi *et al.,* 2018). PIC5 was weakly enriched for similar pathways to PIC4, notably angiogenesis, which were not significant at FDR <0.05.

PIC2 was strongly enriched for immune response related pathways including the interferon alpha and gamma pathways (FDR <0.05). The main pathway enriched for PIC3 samples was tumour necrosis factor alpha (TNF-α) signalling. TNF-α is a cytokine with many roles, including in energy regulation and lipid homeostasis (X. Chen *et al.,* 2009). This signature is reflective of PIC3s overlap with the CMS3 metabolic subtype. PIC1 samples were primarily enriched for down-regulated pathways including EMT. An interesting feature of PIC1 was the downregulation of PD-1 signalling (Appendix Figure 6.7), in comparison with its up-regulation in PIC2. Gene Ontology enrichment analysis (Appendix Figure 6.6) further suggest that PIC1 is a pre-metastatic, immunologically cold group, with the downregulation of cell motility and angiogenesis related processes.

The PIC3 and PIC5 clusters were difficult to characterise using the enrichment score method. For PIC3 this could be due to the relatively small size of the cluster (n=63), however for PIC5 with n=147 patients, it appeared that the cluster was highly heterogeneous on a pathway level. I further examined this heterogeneity by assessing patient-specific (as opposed to PIC-specific) pathway enrichment. This analysis revealed was that each PIC was often heterogeneous on a pathway level. For example enrichment of the E2F targets pathway in PIC4 samples was driven by only a subset of patients (Figure 2.19). These data revealed that even within each PSDE gene defined subtype there is still significant heterogeneity on a gene and pathway level.

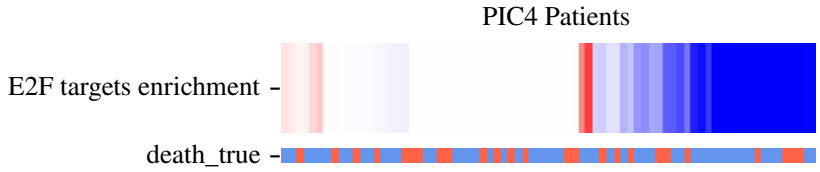


Figure 2.19: *Patient-specific enrichment of the E2F Targets pathway among PIC4 samples. While pathway activity was decreased in many patients (blue), there was still considerable heterogeneity within the PIC4 cluster. Some patients in PIC4 had increased expression of this pathway (red) and for many others, expression was not altered (white).*

Patient samples in PICs are significantly enriched for specific mutations

Next, I investigated whether the PICs could be used to stratify patients prognostically using Kaplan-Meier survival analysis (Figure 2.18, D). This analysis revealed significantly poorer survival of PIC4 patients compared to all other clusters, excluding PIC3. PIC3 is a cluster primarily made up of CMS3 (metabolic subtype) samples, but is also the smallest PIC (n=63), making it more difficult to achieve statistical significance in pairwise comparisons. PIC4 for the most part represents a subset of CMS4 (mesenchymal subtype) samples. The best survival outcomes were found for patients in PIC1 and PIC2, with mean survival times of 1512 and 1519 days, respectively.

I examined the patient metadata of PICs to assess whether any variables may explain the differences observed. For the most part clusters had approximately equal proportion of sex, stage, and other variables. One exception was tissue of origin. PIC2 patients were significantly biased towards tumours originating from the ascending colon and cecum ($p = 2.7 \times 10^{-16}$, chi-squared test). As PIC2 patients correspond mainly to the CMS1, MSI-high subtype (Figure 2.20), these data may be explained by the MSI phenotype being most commonly observed in the proximal colon (Baran *et al.*, 2018). Another notable correlation with PICs was the proportion of stage IV patients present in PIC4, which at roughly 30% was significantly higher (p=0.015) than in other clusters. Given the mesenchymal subtype of PIC4, it is perhaps unsurprising that it would consist of mainly late-stage tumours.

Some of the larger CMS clusters were split into multiple groups by PIC analysis, including CMS2 and CMS4 (Figure 2.20. Interestingly, the CMS4 metastatic subtype was split primarily between PIC4 and PIC5. Despite patients in PIC4 having significantly poorer survival, a difference between PIC4 and PIC5 samples could not be identified when considering only the CMS4 samples.

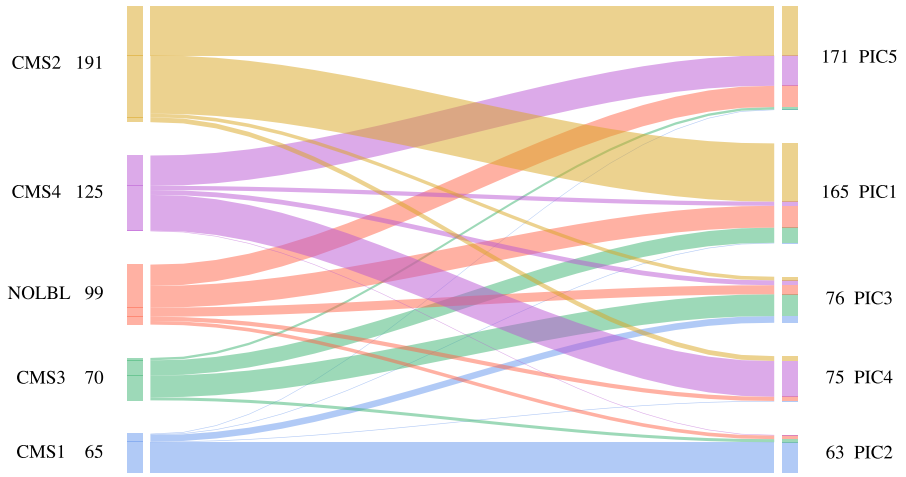


Figure 2.20: *Alluvial plot of CMS cluster compared to PICs, demonstrating patient classification through PSDE-informed clustering produces splits of certain CMS groups.*

In comparison to my PSDE-based approach, only one significant difference in survival was detected among CMS groups (CMS3 vs. CMS4) using pairwise log-rank tests (Figure 2.21). Interestingly, the difference that was detected existed between the metastatic CMS4 subtype and the metabolic CMS3 subtype. PIC analysis of the CMS3 patients grouped the majority of them into PIC3, which was the only cluster not found to have significantly increased survival when compared to PIC4. Applying a log-rank test to CMS3 patients stratified on PIC groups found no difference in survival, potentially indicating that the observed increased survival was merely an effect of the small sample size. It is clear from the increased proportion of stage IV patients in the CMS4 and PIC4 subgroups that the influence of increased progression can be observed on a molecular level, notably through up-regulation of EMT and related processes. However, it appeared that the metastatic samples with poorer survival were less clearly delineated by the CMS. In analysis of CMS4 (Figure 2.21), PIC4 samples were found to have a mean survival time of 1293 days, compared to 1376 for CMS4, indicating poorer survival for PIC4 than the overall CMS4 cohort.

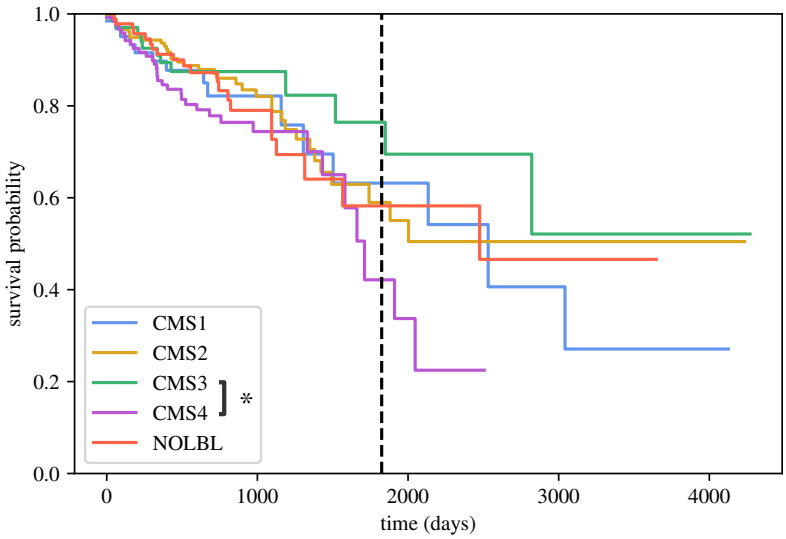


Figure 2.21: *Kaplan-Meier survival analysis of TCGA CRC patients stratified by Consensus Molecular Subtypes.*

Patients in PICs are significantly enriched for different mutations

I next investigated whether samples in different PICs were enriched for specific genomic mutations. Using a Fisher's exact test for over-representation analysis I identified the most enriched deleterious mutations (as determined by SIFT score) within each PIC. I found that *KRAS* mutations were most frequent with PIC1, with the Gly13Asp mutation being the single most strongly enriched ($p = 1.9 \times 10^{-30}$). PIC1 was also strongly enriched for the Glu545Lys mutation in *PIK3CA*, a commonly reported mutation in CRC which has been associated with poorer patient outcome (A.-J. Li *et al.*, 2018). PIC2 patients had a much lower rate of *KRAS* mutation. Rather, the most significantly enriched mutations in PIC2 were Ala668Val in *ITGA1* and Arg124Trp in *GPRC5A*, a gene abundantly expressed in CRC (Zhou & Rigoutsos, 2014). However, PIC2 exhibited a vastly increased mutation level generally, with >10,000 specific mutations being enriched within this cluster. These data are consistent with the global hypermutation associated with the MSI subtype. PIC3 in comparison had the lowest number of enriched mutations, with 3,613 significant mutations, the strongest being the very common mutation Glu545Lys in *PIK3CA*. PIC4 was strongly enriched for *KRAS* mutations ($p = 1.2 \times 10^{-6}$), however the most significant alteration was Arg262Gln

in *TMEM74* ($p = 9.6 \times 10^{-6}$), a gene known to induce autophagy independently of *PI3KC3* (Sun *et al.*, 2017). Autophagy has been demonstrated to be associated with various tumourigenic responses in CRC, including suppressing the immune response and driving a switch to glycolysis (Devenport & Shah, 2019). PIC5 exhibited a similar mutation enrichment pattern to PIC1 with multiple *KRAS* mutations, as well as strong enrichment for Glu545Lys mutations in *PIK3CA* ($p = 4.7 \times 10^{-12}$).

2.4.6 Clustering on a patient-specific pathway level reveals significant survival differences

Pathway and mutation enrichment analysis revealed substantial heterogeneity both between and within PICs. To achieve better stratification of patients based on this pathway heterogeneity, I next performed robust unsupervised clustering of samples based upon pathway enrichment scores, rather than PSDE gene expression. Using enrichment scores obtained for multiple pathway databases (e.g. MSigDB Hallmarks and KEGG) on a patient specific basis, samples were clustered with the same unsupervised consensus hierarchical clustering methodology used for constructing PICs. Using this approach, I found KEGG to be one of the most effective pathway databases for stratifying patients primarily due to having relatively few pathways (increasing clustering efficiency) which still span many biological processes, and used it to define four KEGG-informed patient clusters (KIPCs) (Figure 2.22).

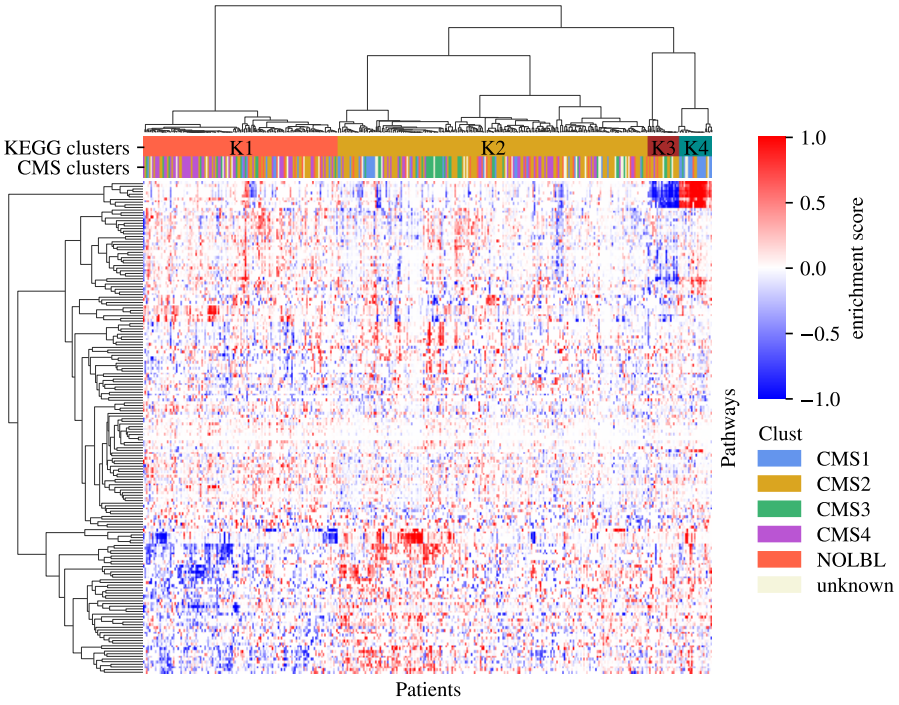


Figure 2.22: *Heatmap of patient-specific enrichment scores. Enrichment analysis was conducted using KEGG as the source database. Four KEGG-informed patient clusters (KIPCs) were identified, K1-4. The consensus dendrogram is shown on the horizontal axis, and KEGG-informed clusters are assigned unique colours. CMS clusters are annotated for comparison below KIPCs.*

KIPCs stratify patients differently to PICs yet still reveal significant survival differences

Kaplan-Meier survival analysis revealed significant differences in survival between at least 3 KIPCs (K1, K2 and K4), with patients in K4 having significantly poorer survival than patients in any other group (Figure 2.23, A). I visualised how these clusters compared to PICs using alluvial plots (Figure 2.23, B). Interestingly, K4 contained a substantial number of samples previously annotated as PIC2. These data demonstrate that patient-specific pathway analysis can identify a subset of patients with significantly poorer survival which are not identified using either a PSDE gene based approach or the CMS subtypes.

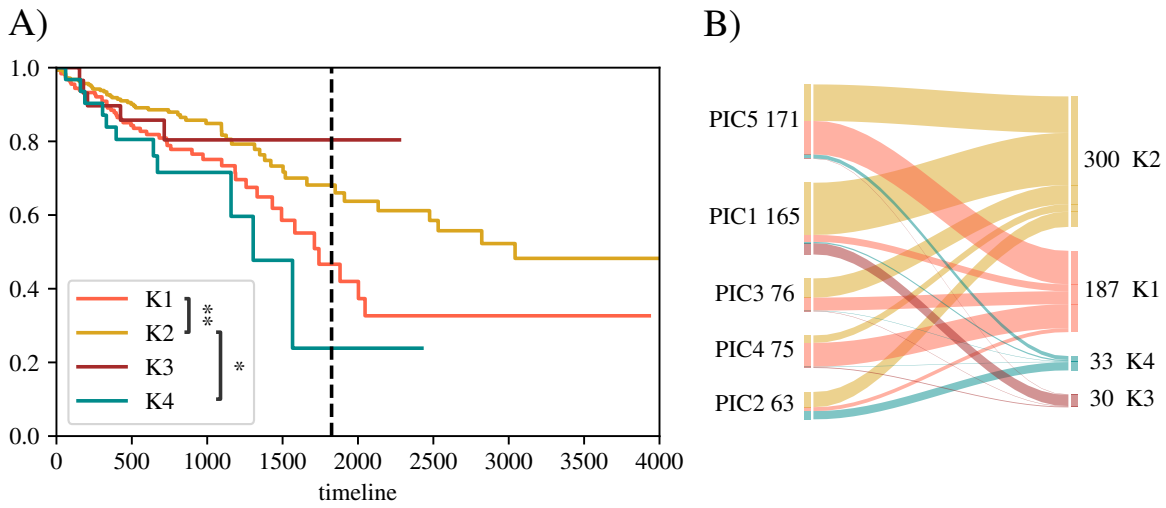


Figure 2.23: A) *Kaplan Meier plot of patient survival in each KEGG-informed patient cluster (KIPC). Statistical significance was assessed using multivariate and pairwise logrank tests.* **B**) *Alluvial-style plot demonstrates the relationship of KIPCs (right) to PICs (left). Clustering based on KEGG pathway enrichment revealed differences which were not seen when clustering based on PSDE gene expression.*

Examining the patient-specific enrichment scores of each KIPC revealed that the two smallest clusters, K3 and K4, had the strongest enrichment scores. K3 was characterised by a down-regulation of immune-related pathways, such as "antigen processing and presentation" and "allograft rejection", while K4 was effectively the inverse, with the same pathways up-regulated. Interestingly, "hematopoetic cell lineage" was down-regulated in K3 patients. I had previously identified the hemostasis Reactome pathway as significantly down in PIC1 (Appendix Figure 6.7). From the alluvial plot (Figure 2.23, B), it can be observed that the K3 cluster is almost entirely a subset of PIC1 patients. Previous investigations have found that hemostatic factor activation is associated with poor prognosis in colon cancer (Ji *et al.*, 2018). These findings corroborate the results found here, in which the downregulation of hemostatic-related functions correlates with increased patient survival.

Patient-specific pathway clustering reveals divisions of CMS1 with significantly different survival probabilites

It was apparent that PSDE gene and patient-specific pathway derived clusters resulted in substantially different stratification of patient survival. While the PICs were broadly similar to the Consensus Molecular Subtypes, the pathway-guided approach identified additional patient subgroups that were dependent on the choice of pathway database. Notably, the patient subgroups informed by patient-specific KEGG (and also MSigDB Hallmarks, see Appendix Figure 6.8) pathway scores resulted in a split of the CMS1 / PIC2 subgroups. The MSI-high CMS1 / PIC2 patient cluster was split into 3 KIPCs with significant differences in survival between CMS1_K3 and CMS1_K4 (p=0.027) (Figure 2.24).

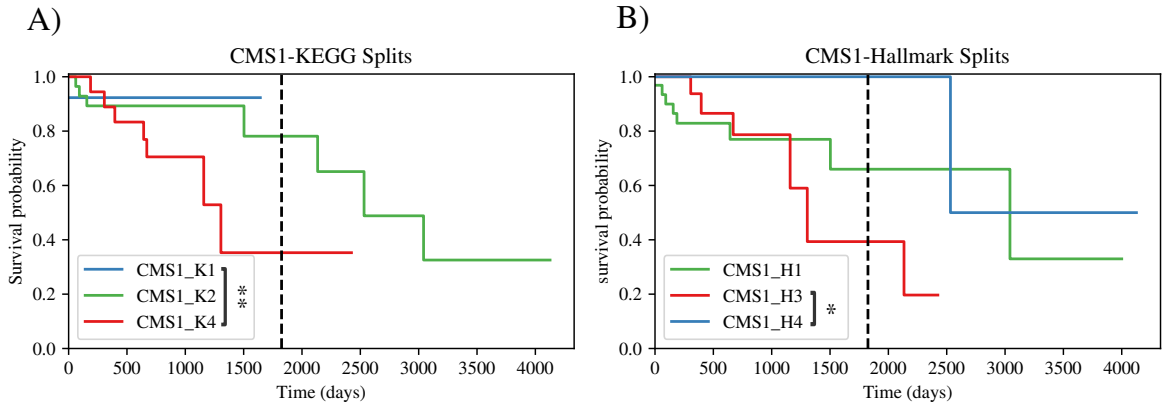


Figure 2.24: *The survival probabilities for CMS1 patients stratified by their patient-specific enrichment scores from KEGG (A) and MSigDB Hallmarks (B) pathway databases.*

In the KIPCs, the majority of CMS1 patients were assigned to K4, which was strongly enriched for immune related pathways, including up-regulation of interferon pathways. Patients in this group had poorer survival relative to other CMS1 patients, which had relatively poor enrichment of immune related pathways. CMS1 samples in K2 were more strongly enriched for metabolic and cell cycle processes, while K1 patients were enriched for pathways including the VEGF, NOD-like, and Toll-like receptor (NLR and TLR) signalling pathways. While bacterial ligands are known to increase the expression of angiogenic factors including VEGF via the TLR pathway

(Bhagwani *et al.,* 2020), it is likely that up-regulation of these pathways occurred due to ongoing immune cell infiltration. This result was interesting as in other analyses MSI-high tumours tend to be relatively homogeneous and thus clustered together, whereas here it is apparent that biologically and perhaps clinically distinct subgroups exist.

2.4.7 A PLS-DA machine learning model can identify features predictive of patient outcome

To further assess the utility of PSDE genes to predict patient outcomes, I developed a Partial Least Squares Discriminant Analysis (PLS-DA) machine learning model. As a quantifiable clinical outcome, the survival time of each patient was used as a dependent variable. I trained the model on the PSDE-informed subset of gene expression data to predict patient survival status at 1, 3, 5 and 10 year timepoints, with accuracy was determined using receiver operating characteristic (ROC) curve analysis. For this binary prediction task the model proved to be most accurate at the 5-year timepoint. Predictions of patient survival at 5 years were consistently more accurate than would be expected by chance, with an AUC of 0.655 (Figure 2.25).

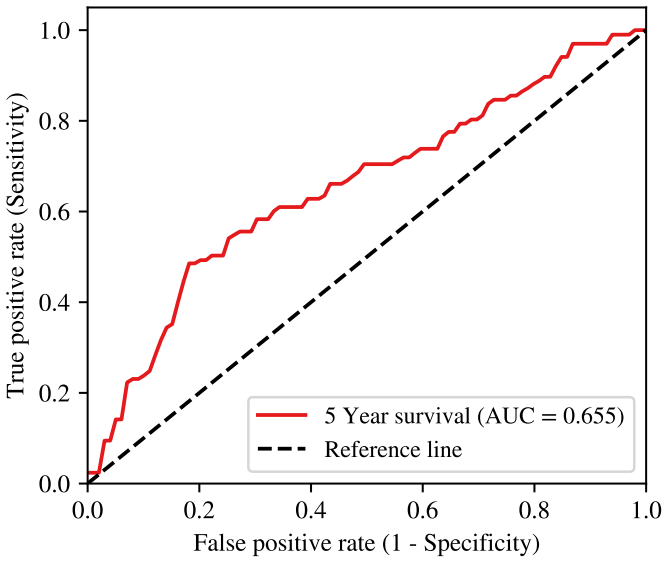


Figure 2.25: *ROC analysis of the accuracy of a PLS-DA model to predict patient survival at 5 years. The model was trained on TCGA CRC transcriptomic data.*

From this model I extracted the 95th percentile of genes most strongly correlated with survival. I ran pathway enrichment analysis on both the negative and positively correlated genes separately, and found that various pathways involved with colorectal cancer progression were positively associated with poorer survival by this model (Figure 2.26), including vascular endothelial growth factor (VEGF) signalling

(WikiPathways: p=0.01, KEGG: p=0.03), key pathways for regulation of angiogenesis, and epithelial mesenchymal transition (EMT) (MSigDB Hallmarks: $p = 1.6 \times 10^{-8}$). This indicated that the model was identifying genes from biological processes which are commonly enriched in metastatic CRC as being predictors of poorer survival.

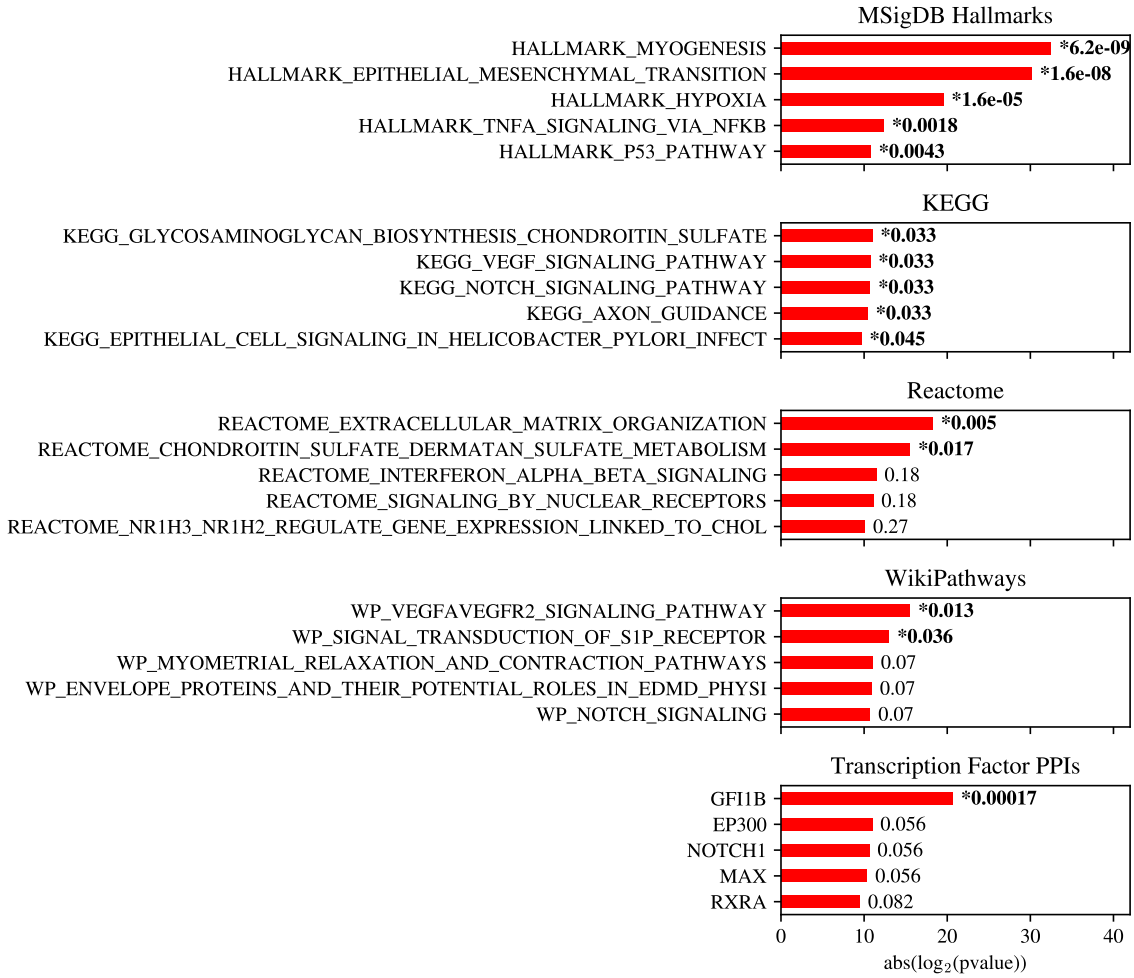


Figure 2.26: *Most significantly enriched pathways for the 95th percentile of genes most strongly correlated with death in a PLS-DA model trained to predict patient outcome. For each pathway database, the top 5 pathways are displayed, ranked by P value. Pathway enrichment at the FDR < 0.05 level is marked with an asterisk.*

The pathways enriched among genes negatively correlated with survival (Figure 2.27) were primarily cell cycle related; including G2M checkpoint, essential for initiation of apoptosis, and E2F transcription factor targets ($p = 1.5 \times 10^{-8}$). DNA damage repair and response pathways were also significantly enriched (Figure 2.18, B). Again,

these data indicated that the model was capable of identifying genes from pathways involved with normal cellular homeostasis and repair as important for patient survival.

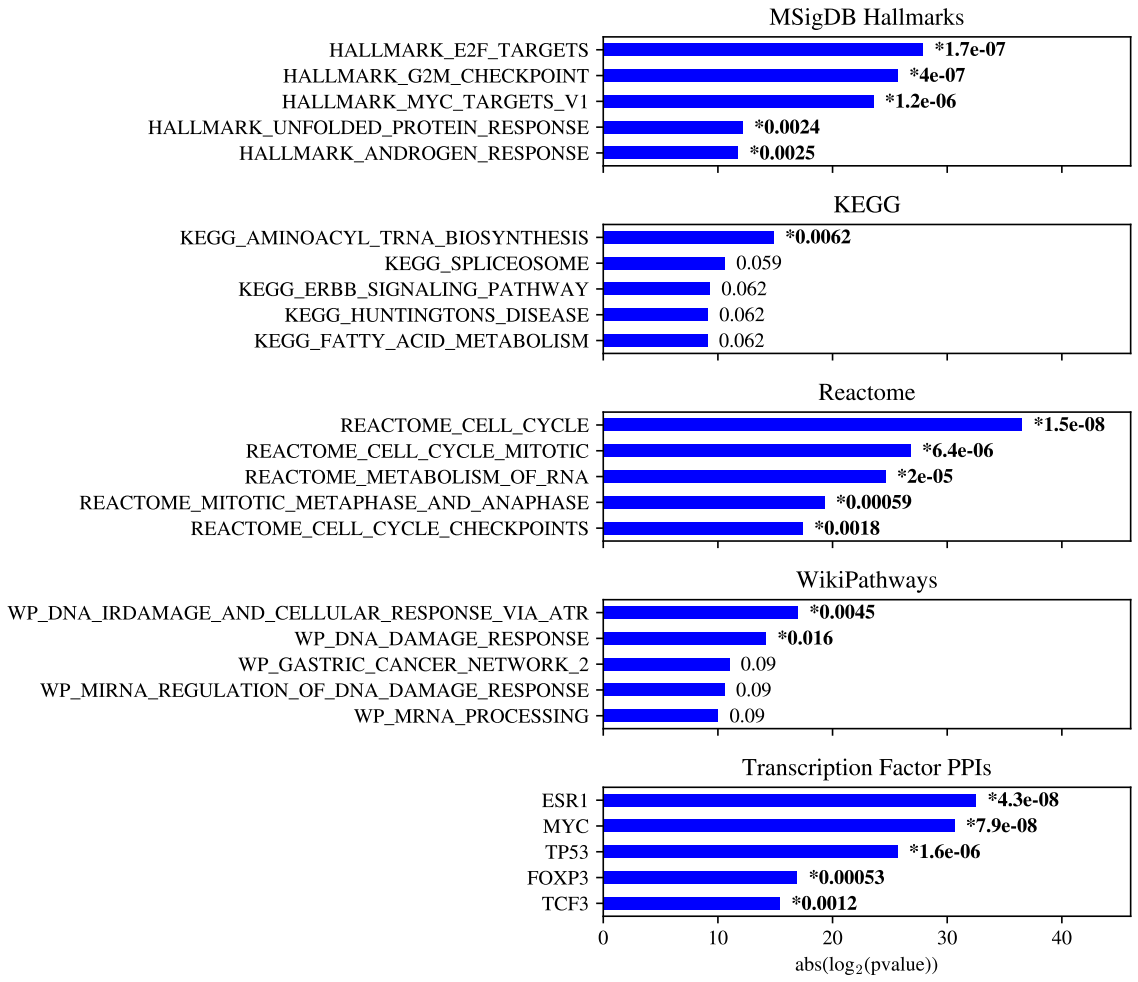


Figure 2.27: *Most significantly enriched pathways for the 95th percentile of genes most strongly correlated with survival in a PLS-DA model trained to predict patient survival. For each pathway database, the top 5 pathways are displayed, ranked by P value. Pathway enrichment at the FDR < 0.05 level is marked with an asterisk.*

A model trained on patient metadata results in increased accuracy

Despite successfully identifying correlations with survival-linked pathways correctly, I hypothesised that the PLS-DA model performance could be further increased by adding more relevant data. I further tested a PLS-DA model trained on various metadata features instead, to see if this would lead to improved accuracy. This metadata model included all available features, including CMS, PIC and KIPC annotations,

excluding metadata directly related to patient survival (i.e., days to death, vital status, etc.). In fact, I found that the metadata based model was capable of much more accurate predictions than the bulk transcriptomics data alone, with the survival predictions now having an AUC of 0.76 as seen in Figure 2.28.

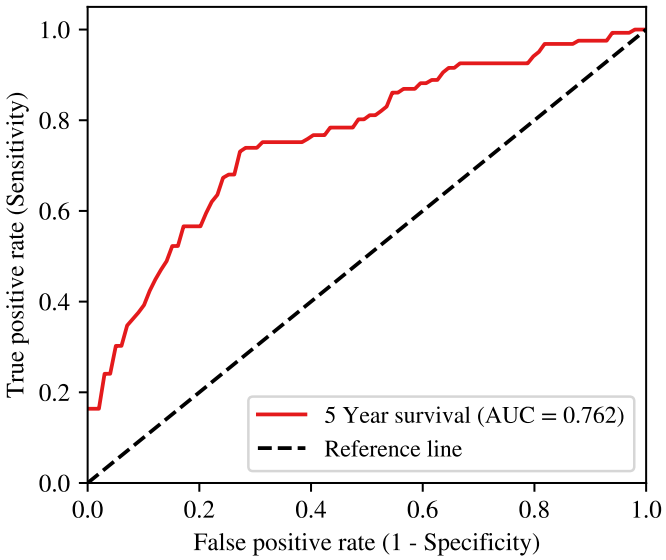


Figure 2.28: *ROC analysis of the accuracy of a PLS-DA model to predict patient survival at 5 years, trained on TCGA metadata and subtype annotations.*

I examined the coefficients of the metadata trained PLS-DA model to determine which features the model most strongly associated with survival (Appendix Figure 6.3). While variables such as weight, height, and age had reasonably strong negative correlations with survival as would be expected, the most powerful predictor in the metadata (once variables which indicated actual survival status were excluded) was TNM stage, with stage IV TNM being strongly correlated with death and stage I strongly with survival, a result that would be expected given the large differences in survival time between these groups (Appendix Figure 6.2). This analysis also revealed that PIC annotation was more strongly associated with survival time than CMS annotation. Drug usage data was also correlated with survival, with use of Avastin or Irinotecan especially (drugs typically employed for metastatic treatment) correlated with poor survival. Of course, these correlations are mostly reflective of clinical stratification and treatment, rather than causative relationships.

2.5 Discussion

Following extensive RNA-seq quality control of the TCGA CRC cohort, including adjusting for tumour purity, sequencing platform batch effects, and removing outliers, I was able to identify patient-specific differentially expressed (PSDE) genes for 550 patients. I found that the up-regulated PSDE genes were enriched for pathways known to be relevant to CRC development and progression, indicating that the PSDE approach was capable of identifying genes relevant to the disease. I then used these PSDE genes as a basis for patient stratification through robust consensus hierarchical clustering, and found prognostically-relevant PSDE-informed cluster (PIC) patient subgroups. These PICs were enriched for specific functional processes, with statistically significant differences in survival in at least one group, PIC4, which was found to have much poorer survival compared to patients in other PICs. Mutation analysis of PIC4 revealed that patients in this cluster frequently had *KRAS* mutations, a mutation predictive of failure of epidermal growth factor receptor (EGFR) targeted therapies found in around 38% of colorectal cancer cases (Oliveira *et al.,* 2004). Pathway analysis revealed that pathways related to metastasis were also strongly up-regulated in these patients, while pathways related to the cell cycle were down-regulated. These data suggest that the PIC4 subtype is a late-stage metastatic CRC subtype. This was corroborated by the significantly higher proportion of stage IV patients within PIC4. Some of the CMS defined subtypes (Guinney *et al.,* 2015) were largely reconstructed by the PIC defined subtypes, especially PIC2 which almost entirely overlapped with CMS1, the microsatellite-instability (MSI), immunologically active subtype. This subtype was the most robust to random resampling during consensus clustering, further increasing confidence in it as a transcriptionally distinct subtype of CRC. One limitation of this approach was the pathway methodology employed, which divided up and downregulated gene sets before determining significantly over-represented pathways. This is a frequently used approach which is often recommended for over-representation analyses (Hong *et al.,* 2014), however it does mean that a particular pathway can be found to be both up- and down-regulated. In my analysis, I chose to ignore such instances, finding them to be relatively rare. In this particular investigation,

I chose to focus specifically on pathways dysregulated in a single direction, however it should be noted that such an analysis could easily lead to a dysregulated pathway being overlooked.

I found that by clustering based upon patient-specific pathway enrichment instead of genes, novel patient clusters with significant differences in survival could be identified. This approach, combining PSDE genes and pathway-level hierarchical clustering, was able to identify several distinct patient groups which were not revealed by gene-level clustering. These clusters were substantially different to CMS groups. This approach applied to KEGG pathways revealed multiple patient groups with significant differences in survival, including the K4 subtype which was enriched for immunological pathways. Patients in K4 had significantly poorer survival than patients in K2. I also identified a larger cluster of patients, K3, with significantly poorer survival than patients in K2. K3 patient samples were significantly enriched for down-regulation of hemostasis, which has previously been identified as associated with poor prognosis in colon cancer (Ji *et al.,* 2018).

I also found that pathway-level clustering could subdivide patients with the MSI subtype. Patients in both the K4 subtype and CMS1 exhibited significantly poorer overall survival than other CMS1 patients. CMS1 samples in K2 were more strongly enriched for metabolic and cell cycle processes, while K1 patients were enriched for pathways including the VEGF, NOD-like, and Toll-like receptor (NLR and TLR) signalling pathways. It is interesting to note that recently the prognostic relevance of MSI status (the defining characteristic of CMS1 patients) in CRC has been debated (B. Wang *et al.,* 2019). My findings suggest that perhaps MSI status is not as biologically homogeneous as is sometimes thought, perhaps supporting the notion that MSI is too broad of a classification to be truly prognostically beneficial. It is important to note that the clusters derived from this pathway-level clustering may over-emphasise certain groups due to the redundancy of pathway members. This is less of an issue for pathways sourced from MSigDB due to the specific focus on non-redundancy in that pathway source, but in KEGG for example, such redundancy is quite possible. For more robust results, a method to reduce redundancy such as SIGORA (Foroushani *et al.,* 2013) (an approach which uses gene pairs, rather than single genes, which tend

to be specific to a single pathway) could be applied. In the absence of this, these results should be interpreted with care.

Finally, I examined how well a machine learning model was able to predict patient survival at 5 years when trained on the expression of PSDE genes, and then for comparison when trained on various metadata factors, including PIC annotations. I found that using the model trained on metadata resulted in more accurate predictions than training on transcriptomics data alone, with PICs being one of the most strongly predictive variables. Of course, this model also incorporated TNM stage, and so is less useful as a prognostic tool and more a way to evaluate how the PIC and pathway-derived clusters compare to CMS and TNM as predictive features.

One major finding of this work was that the heterogeneity in pathway activity in the TCGA CRC patients is not well characterised by the CMS subtypes. While this pathway-based approach to investigating heterogeneity was successful in uncovering previously unknown patient sub-groups, there are many ways in which the method could be improved. The methodology of identifying patient-specific differential expression outlined here was intended to determine the biological properties which best identify an individual patient's cancer given only the results of a cohort of gene expression samples. There were of course many limitations to using these data in this way - each patient represents only a single bulk RNAseq measurement taken at a particular point in time, whereas realistically gene expression can vary quite widely over time (McIntyre *et al.,* 2011). Additionally, the nature of bulk RNAseq is to average the expression of multiple cells, meaning that some of the perhaps clinically important information that defines the heterogeneity of these tumours cannot be captured by this kind of experiment. One of the first pillars of precision medicine is the creation of a taxonomy of human disease - in cancer, this importantly includes creating molecularly-informed disease subtypes (Divaris, 2017). Ideally, this would be accomplished using multiple layers of omics types per patient, rather than the single transcriptomics measurement used here. Some additional information could be collected for these patients - a combination of transcriptomics data with genomic mutation information, protein studies from the CPTAC (B. Zhang *et al.,* 2014), and other types of data may result in more effective clustering when integrated using a

tool such as PARADIGM (Vaske *et al.,* 2010).

My initial attempts at performing patient-specific analysis were hindered due to batch effects and other technical issues with the RNA-seq data. This may be seen as an intrinsic weakness of the method, as results are dependent on the average expression across the cohort of patients, making PSDE analysis highly sensitive to outliers. The way PSDE genes are defined is therefore heavily dependent on both the n patients in the cohort, the composition of the cohort, and the way in which data is preprocessed. This is certainly a limitation, and indeed it meant I had to spend a large amount of time identifying and removing technical effects and outliers before the method would work as intended. Inherently though, this cohort-specific bias is necessary to provide an appropriate population background with which to contrast an individual. In comparison to the differences in survival found within the CMS groups, I found more significant differences between PIC clusters, which may inherently be because the PSDE method is specific to this cohort, and therefore better captures the heterogeneity in these specific individuals.

Regarding the difference between other single-sample transcriptomic analysis tools like GSVA (Hänzelmann *et al.,* 2013) or ssGSEA (Barbie *et al.,* 2009), the PSDE gene approach distinguishes itself in multiple ways. The application of two thresholds is intended to compensate for both absolute and relative variance (with p-value and fold-change thresholds, respectively), and is a relatively unique approach. One limitation of this approach, however is the ultimately arbitrary nature of these thresholds. The thresholds for PSDE gene definition were chosen as to preserve a small but relatively even number of PSDE genes for every patient, however by changing these thresholds, the definition of PSDE genes can be altered. This effect might be circumvented by using a rank-based approach, such as is implemented in tools like GSVA or Singscore (Foroutan *et al.,* 2018). Ranking approaches cannot provide a binary classification as has been shown here, but benefit from retaining potentially relevant genes that may be missed by the set position of a threshold. Although the PSDE approach differs to Singscore and GSVA in this way, it is similar in terms of computational expense, as all of these methods are relatively simple in comparison to more intensive methods like ssGSEA or PARADIGM (Vaske *et al.,* 2010).

Beyond more complex measures of single-sample transcriptomic activity, the PSDE method bears similarities to some pre-existing methods of outlier analysis that can be used for identifying outlier genes and heterogeneity in cancer cohorts, for example COPA (Tomlins *et al.,* 2005). Regarding COPA specifically, the PSDE approach differentiates itself by using a combination of two scaling methods (z-score and fold change), as well as identifying outliers in both directions (the original COPA only identified relative increases in expression). Another unique aspect of the PSDE method is the use of specific threshold to assign significance. However, the stronger rationale for creating the PSDE approach rather than using a pre-existing method was simply that these more comprehensive methods do more than was required for this investigation. The PSDE approach fundamentally differs in terms of its specific goals, which is primarily to create a profile of genes for each sample in a cohort, resulting in a list of specific transcriptomic variations in each individual sample from a cohort that could then be used to compare to individual outcomes. The PSDE method also differs from existing approaches by focusing on tumour-only data. This has the drawback of including genes which may be highly variable in normal samples, which would be undesirable if attempting to identify novel oncogenes or other such tasks. However, it is possible that the activity of such genes might still contribute to the outcomes of an individual, and would therefore be important to include in a more comprehensive model of patient-specific cellular activity, meaning that PSDE genes should be a useful first step for creating such models.

While they may be interesting to use for analysis of individual differences in the CRC cohort, and did indeed assist with finding clusters which were significantly associated with patient survival, I hypothesised that PSDE genes would become more useful if they were linked in a functional context. I tested this here using gene set enrichment analysis within pre-defined biological pathways, and found that the pathway scores could be used to identify in some cases much smaller patient groups which were clinically distinct in terms of their survival. Still, this approach is limited by the available annotations within pathway databases, and indeed I found choice of database strongly influenced the final results. To increase the functional background required and make PSDE genes more informative, a more robust solution that would

overcome this could be to integrate a less biased source of information on functional interactions such as protein-protein interaction data, i.e., a network approach.

3. Patient-specific network analysis reveals a subset of colorectal cancer patients with significantly poorer prognosis

3.1 Background

The reductionist approach to biology supposes that biological systems may be understood by breaking them down into smaller, more comprehensible parts (Regenmortel, 2004). Reductionism implicates a single or very few molecules as being responsible for particular phenotypes, an approach that has proved highly successful in molecular biology research, including for the development of targeted cancer therapies (Boland & Goel, 2005). However, it is now becoming widely understood that diseases such as cancer are to a large degree the result of multiple smaller abnormalities in the complex wiring of the cell, and that to fully understand such systems requires a systems-level approach (Barabási *et al.,* 2011). In contrast to the reductionist viewpoint, a systems approach acknowledges that complex systems have emergent properties which are not evident from their individual components (Gonzalez-Angulo *et al.,* 2010).

Cellular networks are not linear, and seemingly unrelated components may interact with each other under different circumstances, leading to a complex network of interlinked signalling molecules (the interactome) that can alter responses to treatment and disease outcomes in unique ways across patient populations (Y. Li *et al.,* 2018). Genomic information alone has seldom translated into viable new therapeutic strategies, as it does not fully explain the genotype-phenotype relationship which arises from

the complexity of the interactome (Vidal *et al.,* 2011). A priority of modern cancer research has therefore been to map the interactome, identify the mutations that lead to perturbations within it, and to develop models that can predict the link between genotype and phenotype more accurately than "one gene, one function" (Caldera *et al.,* 2017). The difficulty in identifying single targets for pharmaceutical intervention and the fact that many drugs, previously assumed to be targeted treatments, have been found to actually target multiple proteins within particular regions of the interactome (Keiser *et al.,* 2009), as well as the success of combination therapies which target more than a single critical pathway in overcoming drug resistance (Bozic *et al.,* 2013), indicates both the validity and necessity of a systems approach to cancer research (Junttila & de Sauvage, 2013).

Networks as an integrative tool for systems biology

A key tool for systems biology is the network. Networks allow construction of models that incorporate information from diverse sources such as gene expression, mutations, epigenetic markers, protein-protein interactions, and more into a single model (Sonawane *et al.,* 2019). Mathematical network models benefit from the analysis tools of network and graph theory (Pavlopoulos *et al.,* 2011) to assist with generating hypotheses on both the origins of diseases and potential therapeutic interventions in a more integrated fashion (Ozturk *et al.,* 2018). In human cells, phenotypic heterogeneity due to differential gene expression occurs as a result of extrinsic signals communicated by activation or inhibition of receptors at the cell surface. This triggers a network of proteins and metabolites which process signals through a cascade of biochemical reactions occurring between interlinked signalling pathways. The cell responds via activation of downstream DNA-binding transcription factors which regulate the expression of genes (Ma'ayan, 2017). Modelling this complexity is a task that networks are well suited for. Protein-protein interaction (PPI) networks are commonly used to represent the physical association of proteins within a cell (Rivas & Fontanillo, 2010).

Rewiring of the EGFR network due to mutant *KRAS*

PPI networks are dynamic, with some interactions only occurring in specific cellular contexts (Yeger-Lotem & Sharan, 2015). Cancer especially may cause widespread rewiring of the interactome, bypassing normal regulatory pathways (Bowler *et al.*, 2015). This network rewiring as a consequence of mutation can cause a ripple effect, influencing not only a single PPI, but causing further downstream rewiring. This phenomenon was examined by the Protein Interaction Machines in Oncogenic EGF Receptor Signalling (PRIMES) project, which investigated adaptive rewiring of the Epidermal Growth Factor Receptor (EGFR) PPI network due to mutant *KRAS* (Kennedy *et al.*, 2020). EGFR is a member of the ERBB (also known as HER) family of receptor tyrosine kinases, which have been extensively investigated as targets for pharmacological intervention in cancer (Yarden & Pines, 2012). EGFR drives a complex signalling cascade (the EGFR network) that modulates cellular processes critical to cancer progression, such as proliferation and apoptosis (Figure 3.1).

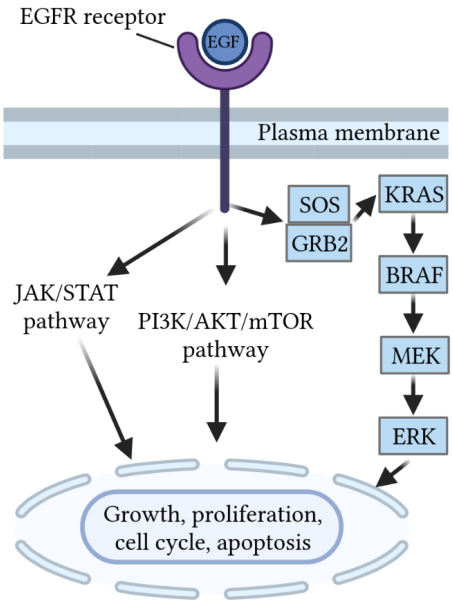


Figure 3.1: *Several canonical signalling pathways downstream of the epidermal growth factor receptor (EGFR) are responsible for mediating essential oncogenic processes. When constitutively activated by mutation, KRAS can bypass targeted inhibition of the EGFR, resulting in tumours which are resistant to anti-EGFR therapy.*

The biological responses to EGFR signalling are ultimately mediated by the activation of downstream transcription factors (Wee & Z. Wang, 2017). EGFR itself is overexpressed due to gene amplification or increased translation in many types of cancer including colorectal cancer (CRC), and is associated with poorer patient outcomes including lower survival rates and higher recurrence (Nicholson *et al.,* 2001). Anti-EGFR treatments (like the monoclonal antibody cetuximab) are commonly used in the treatment of CRC, however the efficacy of anti-EGFR treatments is severely reduced due to mutation of the *KRAS* oncogene, which is found in around 38% of colorectal cancer cases (Oliveira *et al.,* 2004). Mutated KRAS can bypass EGFR inhibition, resulting in secondary resistance (Knickelbein & L. Zhang, 2015). Despite considerable research, KRAS was described as essentially undruggable for many years (Papke & Der, 2017). It is only recently due to the success of allele-specific covalent inhibitors against the $KRAS^{G12C}$ mutant that the designation of "undruggable" has been reconsidered (Moore *et al.,* 2020).

To investigate how *KRAS* mutation rewires the EGFR network, the PRIMES project experimentally mapped the EGFR protein-protein interaction network using affinity purification mass spectrometry (AP-MS) under high and low levels of mutant *KRAS* expression (Kennedy *et al.,* 2020). Systematically mapping >6000 PPIs in the EGFR network, these experiments revealed the structure of the EGFR network in CRC cells and how the network is rewired at a PPI level in cells expressing high levels of mutant *KRAS*.

Prediction of node removals and edgetic effects in networks

Mutations which affect PPIs do not necessarily cause an entire gene to be disrupted, but may instead have a subtle influence on network rewiring across the interactome of cancer cells (Yi *et al.,* 2017). Protein interaction networks may sometimes be altered in a knockout-like fashion (causing all interactions to be lost) but may also cause interaction perturbation (edgetic) effects. Systematic investigations of such mutations (Sahni *et al.,* 2015) have aimed to characterise these mutations, and whether they impair PPI networks. A dataset of experimentally validated PPI disrupting mutations is curated by the EBI (del-Toro *et al.,* 2019), and consists of more than 28,000

specific mutations. Despite the availability of such data, the number of potentially PPI disrupting mutations is vast. Mutations that cause PPI disruption are often localised on the protein interface, a fact that can be exploited to predict potentially disruptive mutations to PPI interfaces computationally (Meyer *et al.,* 2018).

In the context of interactome models, the removal of nodes (i.e. complete loss of gene products) may arise due from nonsense mutations, out-of-frame insertions, deletions that result in a major truncation, gene knockouts, and RNAi-mediated gene expression knockdowns (Charloteaux *et al.,* 2011). In contrast to node removals, edgetic perturbations are those that result in the specificity of interactions between nodes changing. These are likely caused by substitution of a single amino acid in protein-binding sites, or truncations that preserve particular protein domains (Charloteaux *et al.,* 2011). Network models featuring edgetic perturbation have been proposed in order to explain dysfunctions underlying human disease, which has led to the discovery that edgetic perturbations have distinct functional consequences when compared to node removal, as many cases exist in which one gene is linked to multiple disorders (Zhong *et al.,* 2009). Work into making global interactome rewiring analysis in cancer possible has also included annotating cancer alleles by how mutations influence kinase interaction edgetics (Y. Wang *et al.,* 2015). Due to the phenotypic heterogeneity of cancer, how network perturbations and rewiring contribute to oncogenesis is a subject of particular interest.

Alternative splicing is another process that has been shown to lead to edgetic remodelling of PPI networks (Ellis *et al.,* 2012), with splice variants potentially exhibiting altered interaction profiles. Alternatively spliced proteins may be highly functionally divergent, with one study profiling hundreds of alternatively spliced protein isoform pairs finding that the majority of isoforms share less than 50% of their interactions (Yang *et al.,* 2016), meaning that protein interaction capabilities are widely extended by use of alternative splicing.

Patient-specific networks for precision medicine

Due to the availability of large-scale individualised gene expression data, many computational approaches now exist that attempt to capitalise on the potential of network approaches to inform patient-specific outcomes in cancer and other diseases (Hastings *et al.,* 2020). These network-based approaches make use of network and graph theory to predict disease progression, treatment options, and patient outcomes. Patient-specific network models further increase the potential avenues available for re-application and combination of molecular data, and may help to reveal the functional relationships underlying diseases (Barabási *et al.,* 2011).

A common application of the integration of patient-specific data with networks is to improve patient stratification by more accurately modelling inter-patient tumour heterogeneity. For example, network-based stratification (NBS) (Hofree *et al.,* 2013) is a method for the integration of tumour genomes with gene networks, in which patients with mutations in similar network regions are clustered together using an unsupervised method. When tested on ovarian cancer datasets, heterogeneous populations of tumours could be divided into clinically meaningful subtypes determined by molecular profiles. In a related fashion, integrative patient-specific networks are often used to enhance regression and classification tasks. Regularisation of data with respect to network structure (C. Li & H. Li, 2008) has been used to improve the accuracy of disease gene classification tasks. For example, the dgSeq algorithm (P. Luo *et al.,* 2017) combines PPI networks and gene expression data from individuals to train logistic regression models for this purpose. Similar logistic regression approaches which integrate network structure have been proposed for tasks such as biomarker prediction (K. Zhang *et al.,* 2018), or predicting likelihood of metastasis (Chuang *et al.,* 2007).

Patient-specific network approaches also commonly make use of probabilistic models such as bayesian networks and factor graphs (both subclasses of probabilistic graphical models (PGMs)). These structures encode the conditional dependency relationships between random variables (such as the expression of different genes) as a network. These mathematical structures are useful as they facilitate Bayesian inference

of the likelihood of particular outcomes, which is especially useful when the observed data is incomplete. PARADIGM (Vaske *et al.,* 2010), for example, is a popular tool reliant on PGMs to make inferences of patient-specific pathway activities. PARADIGM is able to determine patient-specific gene activity in particular pathways by incorporating curated pathway interactions with gene expression data in a PGM. The unique aspect of PARADIGM is its ability to integrate any number of genomic and functional genomic datasets to infer pathway perturbation in a single patient. A more recent example of the application of PGMs was the model designed by Ha *et al.,* 2018 which performs personalised cancer-specific integrated network estimation (PRECISE) (Ha *et al.,* 2018). Interestingly, PRECISE uses a single topological network structure for different cancer types, using patient-specific information including gene expression to estimate prior probabilities in the model. A major drawback of PGMs is that making inferences using them is computationally expensive, which is only exacerbated when attempting to apply them on a patient-specific basis. Using such models to integrate data on the scale of an entire PPI network would be infeasible.

Logical models are another network formalism which have been successfully adapted to patient-specific data. These networks are generally relatively small but highly detailed mechanistic models of specific biological pathways. They have been adapted to simulate signalling downstream of the EGFR network (Samaga *et al.,* 2009), and have successfully been applied to stratify breast cancer patients (Béal *et al.,* 2019) and to predict possible patient-specific drug targets in brain tumours (Barrette *et al.,* 2018). A weakness of logical models is the small scale upon which they operate. Logical networks tend to be small as a high level of detail is required for each interaction in order to produce a valid mechanistic model.

Topological network properties in cancer

Many measures exist which quantify the topological characteristics of networks, some of which have been directly linked to biological properties. (Breitkreutz *et al.,* 2012) showed that reduced network complexity corresponded to increased survival across different cancer types. Degree distribution entropy (H) may be calculated using the equation $H = -\sum_{k=1}^{N-1} p(k) \log p(k)$, where N is the number of nodes in the given

network, and $p(k)$ is the degree distribution. To calculate degree distribution, with a network of N nodes, n_k of which have degree k, the degree distribution may be obtained by $P(k) = \frac{n_k}{n}$. B. Wang *et al.* used degree distribution entropy as a measure of network heterogeneity, finding that it was an effective measure of resilience to random failures in scale-free networks (B. Wang *et al.,* 2006). Reduced network complexity would mean a more easily disrupted system. In the context of cancer signalling networks, this could mean a network that is more easily targeted by therapeutic interventions.

Node degree, or connectivity, has been used as a measure to demonstrate that proteins mutated in cancer on average tend to have a higher connectivity than other proteins in PPI networks (Jonsson & Bates, 2006). Very highly connected nodes are hubs, and typically biological networks have only a few of these nodes (Barabási, 2016). Related to degree is clustering coefficient, a measure of how likely nodes in the network are to be connected, a metric which is generally very high in biological networks (Barabási & Oltvai, 2004). This may be assessed locally for a given vertex (Watts & Strogatz, 1998) or globally for a network using either the average of local coefficients or by the ratio of closed to open node triplets (groups of three nodes connected by 2 (open triplet) or 3 (closed triplet) edges.

Betweenness centrality is another measure frequently examined in biological networks which describes the number of shortest paths which pass through a given node. It is of interest in the context of cancer networks as potential therapeutic targets and biologically essential nodes often tend to have a high betweenness centrality (Gursoy *et al.,* 2008), identifying them as "bottleneck" nodes. Bottleneck nodes are significantly less well co-expressed with their neighbours than non-bottleneck nodes in PPI networks (H. Yu *et al.,* 2007). There are many measures beyond betweenness which are designed to assess “centrality” in graphs. A benchmarking of different centrality measures by Sharma *et al.* suggested that PageRank may be a more useful metric than betweenness for the task of identifying essential nodes in a network (Sharma *et al.,* 2016). A less common centrality metric is central point dominance, which measures the degree to which a single node can dominate the network (Freeman, 1977).

Modelling alterations to network topology in large networks

There are various ways in which information about network topology may be extracted and potentially applied on a patient-specific level to investigate specific hypotheses. Kennedy *et al.* for example used information flow analysis to model alterations in signal flow from EGFR to downstream transcription factors in response to high and low levels of mutant *KRAS* expression (Kennedy *et al.,* 2020). Information flow analysis is of particular interest as it allows assessment of which subnetworks and downstream outputs are preferentially visited, given a particular network topology. This is achieved by the simulation of information via random walks from a source node to downstream sink nodes (Stojmirović & Y.-K. Yu, 2007). Random walkers are bootstrapped multiple times, so that eventually nodes with much higher or lower number of walkers passing through them can be identified. Information flow analysis is a specific example of a concept in network biology sometimes called "network propagation", in which biological signal propagates from prior information nodes to implicate nearby nodes. This approach relies on the key assumption that biological molecules which are related in function tend to interact with one another (Cowen *et al.,* 2017). PageRank is one of the more famous examples of a network propagation algorithm (Brin & Page, 1998). Originally designed for ranking web pages for Google, it has also been applied to similar tasks in biology such as determining the importance of a particular genes within biological pathway (Ozturk *et al.,* 2018). An advantage of network propagation algorithms is that running them even on relatively large networks (of the order of thousands of nodes and edges) is not computationally expensive.

Modelling heat diffusion is another network propagation approach, typically used to identify subnetworks of interest to a given condition in PPI networks. An example of this concept is TieDIE (Paull *et al.,* 2013) which uses a heat diffusion process to infer the gene network structure in individuals, revealing subtype-specific networks. The method is given one interaction network, a set of upstream nodes such as mutation data and also a set of downstream nodes, for example transcription factors, then concurrent diffusion processes are run from each of these sets of nodes – the core set of nodes which are covered by both of these diffusion processes is inferred to be a gene network relevant for the phenotype being studied. TieDIE and related approaches such

as signalling pathway impact analysis (SPIA) (Tarca *et al.,* 2009), HotNet (Reyna *et al.,* 2018) and ITM Probe (Stojmirović & Y. K. Yu, 2009) at their core all make use of some variation of network propagation.

Previous studies have identified topological statistics such as measures of network complexity to be significantly correlated with survival across different cancer types (Breitkreutz *et al.,* 2012), however a similar effect has not been demonstrated using individual patient networks. Many other topological properties which are known to be biologically relevant (for example node connectivity, which is increased in proteins mutated in cancer when compared to other proteins in PPI networks (Jonsson & Bates, 2006)) may also be relevant on a patient-specific basis, however existing studies generally assume the same network topology across different patients. It is apparent that there is a significant gap in the literature with regard to patient-specific network approaches to disease which focus on individual differences in network topology on a large scale, such as in experimentally derived PPI networks. Given the utility of network propagation algorithms for extracting biologically relevant topological information in large networks, applying them on a patient-specific basis may reveal new insights into network level inter-patient heterogeneity.

3.2 Hypothesis and Aims

Previously, I had hypothesised that stratifying patients based on patient-specific differentially expressed (PSDE) genes would be likely to reveal clinically relevant molecular subtypes of colorectal cancer (CRC). Increasing evidence suggests that protein-protein interaction (PPI) networks are altered in diseases including cancer, and that such changes contribute to pathogenesis and patient outcomes. Network and pathway analysis provide frameworks to model the complexity of this cellular dysregulation. To date however, most approaches are applied in a way that assumes the same network topology in different patients.

I hypothesised that creating patient-specific networks of key signalling pathways in CRC would allow network propagation tools such as information flow analysis (Stojmirović *et al.,* 2012) to stratify patients into different subgroups with altered survival outcomes. To address this hypothesis, I proposed the following aims:

1. Combine patient-specific gene expression data from The Cancer Genome Atlas (TCGA) with PPI data to create personalised network models of the EGFR network in each CRC patient.

2. Identify topological properties of these patient-specific networks which are predictive of patient survival and outcomes.

3. Simulate dynamic signal flow within these patient-specific networks to identify differential signalling to downstream transcription factors.

3.3 Methods

3.3.1 Acquisition of protein-protein interactions

To create a network model which could later be personalised for individual patients, a source of known protein-protein interactions (PPIs) was required. The International Molecular Exchange (IMEx) consortium curates a non-redundant set of PPIs which can be accessed from centralised point (Orchard *et al.,* 2012). The entire IMEx database was downloaded (last accessed October 2020) to obtain all publicly available binary PPI interactions[1] in Proteomics Standards Initiative - Molecular Interaction (PSI-MI) TAB format (MITAB). The PSI-MI parser[2] developed by IntAct (Orchard *et al.,* 2014) was used in a custom Java class to extract the human subset of IMEx PPI interactions, defined by interactions in which each protein and the host organism were annotated with the NCBI taxonomy ID 9606 (human). IMEx also includes some non-protein interactors, however these were excluded as the vast majority of curated interactions are between proteins.

3.3.2 Constructing the EGFR PPI network

The Epidermal Growth Factor Receptor (EGFR) and dysregulation of its downstream signalling pathway is critical to the progression of colorectal cancer (CRC). This network was the focus of the PRIMES project, which recently mapped the EGFR pathway in the HCT116 and HKE3 CRC cell lines using affinity purification mass spectrometry (AP-MS) (Kennedy *et al.,* 2020). Due to the availability of these data and the relevance of EGFR in CRC, the PRIMES HCT116 network was used as a base network model to develop a proof-of-concept patient-specific network approach. The PRIMES HCT116 network was used rather than the HKE3 network as HCT116 cells exhibit a transformed phenotype and express higher levels of mutant KRAS than HKE3 cells (Kennedy *et al.,* 2020). All proteins used as baits and identified as preys

[1] ftp://ftp.ebi.ac.uk/pub/databases/intact/current/psimitab/intact.zip

[2] available from https://github.com/MICommunity/psimi

using AP-MS in the PRIMES HCT116 EGFR network formed the initial nodes of a comprehensive EGFR PPI network.

Addition of prey-prey interactions

As AP-MS was used to map bait-prey interactions by the PRIMES project, prey-prey interactions were unable to be identified. To fill in these missing interactions, the network was supplemented with interactions known to occur between the prey proteins, based on other experimental data obtained from IMEx. Quality filtering of IMEx PPIs was performed by retaining interactions with an IntAct MI score (an interaction confidence score normalised between 0 and 1 (Kerrien *et al.,* 2012)) >0.3.

Addition of canonical EGFR proteins and transcription factors

Although the PRIMES network identified preys for almost 100 bait proteins, some known members of the canonical EGFR pathway were not used as baits (due to cost and logistical reasons). A curated list of missing canonical proteins and their interactors in IMEx were added to supplement the PRIMES network (Appendix Table 6.2). Furthermore, 24 transcription factors known to be downstream of the EGFR signalling pathway (Table 3.1) and their interactors were also added. The canonical proteins and transcription factors were integrated into the comprehensive EGFR network by identifying interactions between the additional proteins and PRIMES prey proteins with an MI score >0.3 in the IMEx database. This extended list of interactions was then appended to the existing network, resulting in a high quality EGFR network model (EGFR-HQ) which included all proteins used as baits in the PRIMES HCT116 EGFR network, plus a set of canonical EGFR network proteins and related transcription factors (Figure 3.2).

Table 3.1: *List of the 24 transcription factor proteins that were added to the PRIMES HCT116 EGFR network, many of which were previously identified as relevant in EGFR by the PRIMES project (Kennedy* et al., *2020).*

Symbol	Name	ENSEMBL	UniProtKB
CREB1	cAMP responsive element binding protein 1	ENSG00000118260	P16220
ELK1	ELK1, ETS transcription factor	ENSG00000126767	P19419
FOS	Fos proto-oncogene, AP-1 transcription factor subunit	ENSG00000170345	P01100
FOXO1	Forkhead box protein O1	ENSG00000150907	Q12778
HSF1	Heat shock transcription factor 1	ENSG00000185122	Q00613
JUN	Jun proto-oncogene, AP-1 transcription factor subunit	ENSG00000177606	P05412
MYC	MYC proto-oncogene, bHLH transcription factor	ENSG00000136997	P01106
SMAD2	SMAD family member 2	ENSG00000175387	Q15796
SMAD3	SMAD family member 3	ENSG00000166949	P84022
SMAD4	SMAD family member 4	ENSG00000141646	Q13485
SP1	Sp1 transcription factor	ENSG00000185591	P08047
SRF	Serum response factor	ENSG00000112658	P11831
STAT5A	Signal transducer and activator of transcription 5A	ENSG00000126561	P42229
STAT5B	Signal transducer and activator of transcription 5B	ENSG00000173757	P51692
STAT1	Signal transducer and activator of transcription 1	ENSG00000115415	P42224
STAT3	Signal transducer and activator of transcription 3	ENSG00000168610	P40763
TCF4	Transcription factor 4	ENSG00000196628	P15884
TCF7	Transcription factor 7	ENSG00000081059	P36402
TP53	Tumor protein p53	ENSG00000141510	P04637
FOSB	FosB proto-oncogene	ENSG00000125740	P53539
FOSL1	Fos-like antigen 1	ENSG00000175592	P15407
JUND	JunD proto-oncogene	ENSG00000130522	P17535
EGR1	Early growth response 1	ENSG00000120738	P18146
EGR2	Early growth response 2	ENSG00000122877	P11161

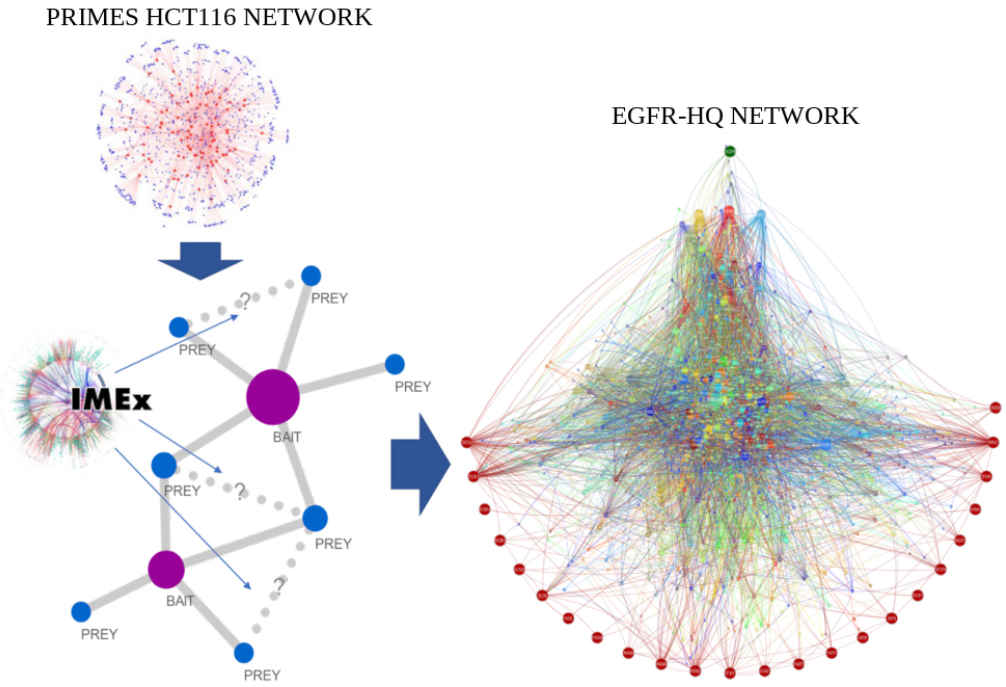


Figure 3.2: *Overview of the process used to construct the high quality EGFR PPI network (EGFR-HQ). Right: Visualisation of the EGFR-HQ network using the graph-tool Python library. To achieve the layout pictured, transcription factors and the EGFR node were fixed in position on the circumference of a circle, then graph-tool's* sfdp_layout *function was used to layout all other nodes. Transcription factors are highlighted in red.*

3.3.3 Performing graph operations on SIF files with Sifter

Networks were saved as simple interaction format (SIF) files due to the simplicity of the format. Editing SIF files as text is simple, however performing graph operations on these files usually requires the use of a heavier graph-oriented library such as NetworkX or graph-tool (T. P. Peixoto, 2017). To enable rapid graph operations on SIF files with minimal effort, I developed a command line tool, Sifter. Sifter is able to perform operations such as taking the union and intersection of multiple graphs, counting nodes and edges, removing duplicate edges and self loops, and other simple tasks. Sifter is available from the Lynn lab bitbucket repository[3].

[3] https://bitbucket.org/lynnlab/sifter

3.3.4 Creation of patient-specific EGFR networks

Node removal approach

To create patient-specific networks, I developed a method which personalised the EGFR-HQ network using individual CRC patient data from The Cancer Genome Atlas (Weinstein *et al.,* 2013). This method used gene expression data from each patient tumour sample to remove specific nodes which were not expressed or under-expressed, thus personalising the topology of the network for each individual. Using the patient-specific gene expression data from 550 TCGA CRC patient tumours as previously described in Chapter 2, genes which were significantly under-expressed in specific patients were identified. For each of these under-expressed genes, the corresponding nodes in the EGFR-HQ network and their interactions were removed, creating 550 different patient-specific EGFR network models.

The approach to detect under-expressed genes to remove from the network built upon the method developed in Chapter 2 (originally for identification of patient-specific differentially expressed (PSDE) genes). PSDE gene identification as previously described used the intersection of two thresholds, fold change (F_g) and Z-score (Z_g), for each gene g in a cohort:

$$PSDE_g = F_g \wedge Z_g \tag{3.1}$$

To identify under-expressed genes for node removal, the fold change threshold for a gene (F_g) was again defined as a 2-fold change from the cohort median expression of that gene, except that only decreases in fold change were considered as only down-regulated genes were of interest. The original Z-score threshold Z_g was a two-tailed test, based on a ± 1.96 standard deviation from the mean of logged counts per million (CPM). For identifying node removals, only the lower tail of the distribution was of interest. To retain a $p < 0.05$ test statistic the Z-score threshold was therefore set to retain genes with $Z_g < -1.65$.

This strategy aimed to identify genes which were expressed in specific patients at a much lower level than the cohort average. To further include all genes which were

essentially not expressed at all, even if they were not included in the modified PSDE approach, an additional criterion was added. If the mean cohort expression for a gene was at least 10 CPM, the gene would automatically be identified as a node removal for any individual in which it was expressed at <3 CPM (the threshold previously defined as "not expressed" during processing of gene expression data in Chapter 2). The rationale for this additional criterion was to remove any genes from individual networks in which the gene fell below the level of detectable expression. In addition, any genes that were completely filtered out during the initial RNASeq quality control steps due to their expression levels being too low were removed from all patient-specific networks. Applying this node removal strategy, 550 patient-specific network models were created by integrating TCGA CRC RNAseq data with the EGFR-HQ network (Figure 3.3).

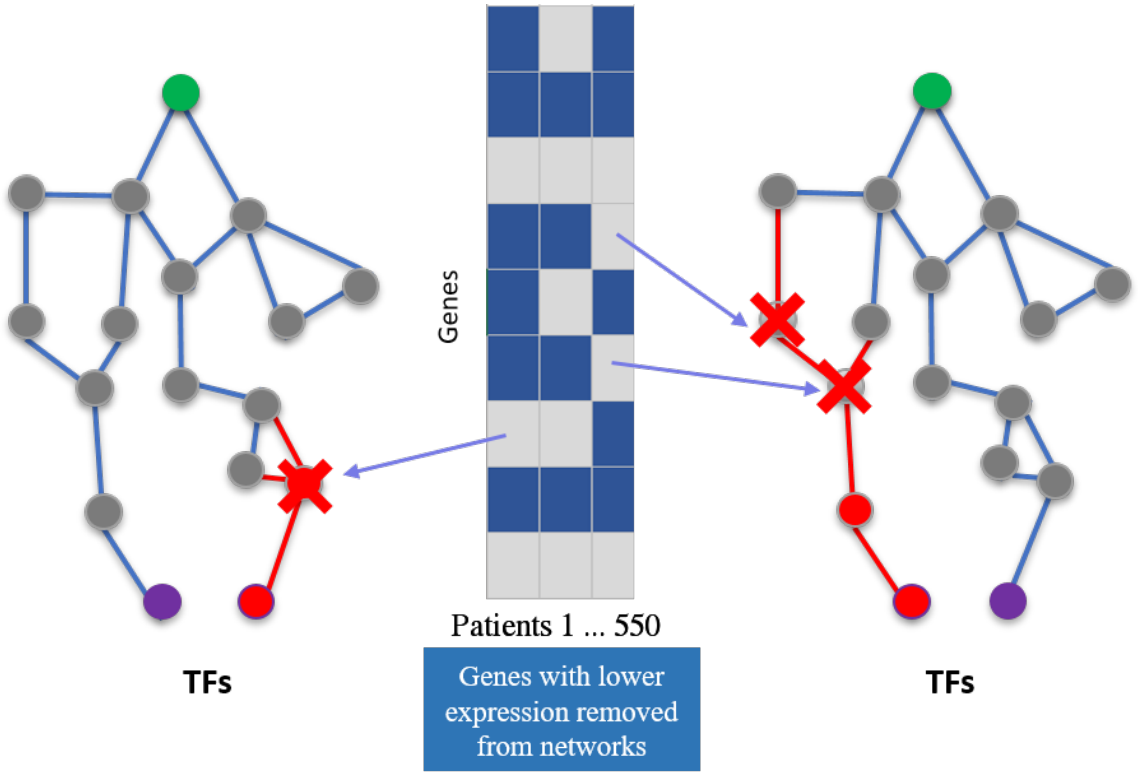


Figure 3.3: *Overview of the node removal strategy used to create patient-specific EGFR network models. CRC tumour gene expression data from TCGA was used to remove nodes from the EGFR PPI network to create networks for each of the 550 patients. These networks were stored as Simple Interaction Format (SIF) files (a plain text tab-delimited list of edges).*

Mapping of gene and protein identifiers

To match genes from TCGA gene expression data to proteins in the EGFR-HQ network, mapping between different gene/protein identifiers was required. Genes in TCGA gene expression data are identified using an Ensembl gene ID (Hubbard *et al.,* 2002). These gene IDs are also appended with a version number, which is incremented when the identifier is revised in some way. This version number was removed in order to map them to other identifiers. Ensembl BioMart[4] was used to identify mappings between Ensembl gene IDs and Uniprot protein identifiers (which are used to identify proteins in IMEx PPI data). To prevent issues such as multiple genes matching to one protein, node removals were only performed when an Ensembl gene ID could be uniquely matched to a single UniProt protein ID. This also applied in the inverse, removals were only performed if a single UniProt ID matched the Ensembl gene ID.

Identification of PPI disrupting mutations

Mutation data were downloaded as MAF (Mutation Annotation Format) files from the GDC. The GDC offers MAF files generated via multiple different variant caller algorithms, including MuSE (Fan *et al.,* 2016), MuTect, VarScan and SomaticSniper. MAFs as produced by the MuTect algorithm were chosen due to widespread use and recommendation of this algorithm in the literature (Xu, 2018), as well as due to its use in other tools such as CBioPortal (Cerami *et al.,* 2012). MAF files for both colorectal cancer TCGA projects (COAD and READ, for colon or rectal adenocarcinoma) were downloaded and concatenated. Within these data, multiple different assessments of mutation impact are provided – including IMPACT from the Ensembl Variant Effect Predictor, SIFT, and PolyPhen. Using the HGVSp field from the GDC MAF files it was possible to link to specific somatic mutations in the (del-Toro *et al.,* 2019) dataset of PPI disrupting mutations. The network edges corresponding to these mutations were removed from the corresponding patient-specific network models.

[4] https://www.ensembl.org/biomart/martview

Identification of domains likely to mediate PPIs

Protein domain annotations were obtained in XML format from the Pfam database (Mistry *et al.,* 2021). The frequency of shared protein domains occurring between binary pairs of interacting human proteins from IMEx was assessed. An over-representation analysis was conducted using a Fisher's exact test to identify domains which were significantly enriched in binary interacting pairs of proteins. The Benjamini-Hochberg procedure was used to adjust for multiple testing. Ranges within proteins containing these domains were considered likely to mediate PPIs and were searched for mutations with a possibly damaging SIFT score in TCGA CRC patients. Where these mutations were identified, corresponding edges were removed from patient-specific network models.

Update of common functions in the biomodule library

Various common functions required for manipulating networks and processing gene expression data were added to my biomodule library[5] to make these tasks more efficient. This functionality included performing the ID conversions between gene and protein identifiers. The list of unique mappings between Ensembl gene IDs in the TCGA data and Uniprot protein IDs in the EGFR-HQ network was cached for rapid conversion. Internally the the PyBiomart library (version 0.2.0) was used to interface with BioMart to obtain mappings.

3.3.5 Topological network analysis and visualisation

To determine whether network topology was directly related to patient outcomes, network properties including degree distribution, clustering coefficient and betweenness centrality were assessed using the graph-tool library (version 2.37) (T. P. Peixoto, 2017). Node-specific network properties were also compared to TCGA RNA expression using volcano plot visualisations. Cox regression analysis was used to assess whether network properties were associated with survival. This analysis was conducted via

[5] https://bitbucket.org/lynnlab/biomodule

the Lifelines package (Davidson-Pilon *et al.,* 2019) for Python (version 0.25.4). Visualisations were also produced using graph-tool, including programmatically laying out and rendering networks, in conjunction with matplotlib (version 3.1.3) (Hunter, 2007). Global network properties for each patient-specific network were compared using ANOVA followed by post-hoc T-tests to identify which groups differed, and visualised using box plots with matplotlib. Statistical analysis was conducted using functions from SciPy (version 1.3.3) (Virtanen *et al.,* 2020).

3.3.6 Simulating biological information flow

To model information flow through the 550 patient-specific network models, the command-line interface of ITM probe[6] was used. ITM probe was previously used by Kennedy *et al.* (Kennedy *et al.,* 2020) to simulate information flow in the EGFR network, and has the advantage of outputting information flow scores for each node in the network. This tool required the network to be in a specific JSON-encoded Compressed Sparse Row (CSR) matrix format. I wrote a Python script to perform this conversion and to run the tool manually. To facilitate easier use of the ITM Probe tool, I also wrote a wrapper library which allows the tool to be invoked directly from Python. This wrapper converts a Simple Interaction Format (SIF) file (or any table with interactions in rows, i.e. a graph in adjacency list format) into the required JSON format, and calls the ITM Probe tool with customisable parameters. The wrapper library can output both text-based and graphical results in the form of a network visualisation (Figure 3.4). This library is available from the Lynn lab Bitbucket repository[7].

[6] ftp://ftp.ncbi.nih.gov/pub/qmbpmn/qmbpmn-tools/src/qmbpmn-tools-1.5.4.tar.gz
[7] https://bitbucket.org/lynnlab/itm_wrapper

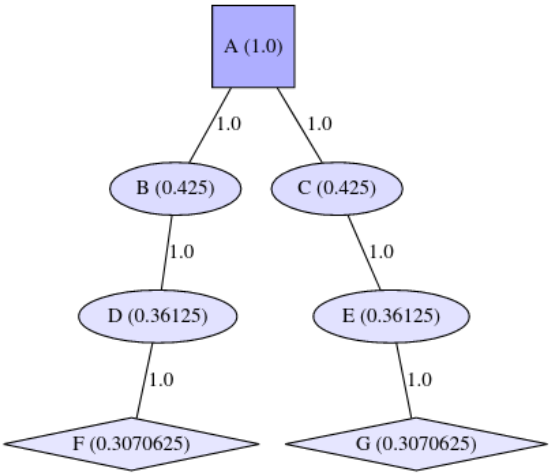


Figure 3.4: *A simple test of the ITM probe wrapper on a small network. Source nodes are represented as squares (A) and sink nodes as diamonds (F and G). Information flow scores are shown on each node, and edge weights on each edge. This test used the emitting model with default dissipation probability (0.15).*

The ITM probe software includes three different models (emitting, absorbing, and channel) that may be used in different contexts. In each model, at least one of either a set of sinks or sources is required. The emitting model simulates information being emitted from a certain set of nodes, and identifies the most frequently visited nodes. The absorbing model, given a set of nodes that will act as absorbing sinks of information, identifies the nodes most likely to send information to them. The channel model identifies the most likely paths between two sets of nodes, the information sources and sinks. In the context of simulating information flow through the EGFR network, EGFR was designated as the source node, while downstream transcription factors were designated as sinks. Therefore, both the channel and emitting modes were possible appropriate models, and so simulations using both modes were assessed. Using the ITM wrapper tool, information flow analysis was run for each of the 550 patient-specific networks previously created using both emitting and channel modes. The default dissipation factor (0.15) was used. When the channel mode was utilised, the transcription factors were selected as sinks. After running information flow analysis, the information flow score (IFS), representing the probability of visiting each node, was extracted.

Determining information flow impact

To assess the change in IFS between individual networks, an impact score was calculated based on the IFS of each node in each patient-specific network. This impact score was defined as the $\log_2$ of the ratio of the information flow score S_n for the patient-specific network n to a baseline score S_b, where the baseline was determined by running ITMProbe on a network with no modifications:

$$Impact = \log_2(S_n/S_b) \tag{3.2}$$

The change in IFS between networks was then used to predict whether patient-specific rewiring altered flow to downstream transcription factors. Cox regression analysis was used to assess whether impact scores for transcription factors were associated with survival, using the Lifelines package for Python. Further investigation of transcription factor activity was performed by using HOMER (Heinz *et al.,* 2010) to detect the enrichment of motifs among up and down-regulated PSDE genes and performing over-representation analysis of motifs in patients with significantly impacted transcription factors.

Determining significant impact scores

To identify nodes which had statistically significant changes in information flow impact score compared to the baseline network, impact scores for each node were compared to an empirical distribution of all impact scores for all nodes across the 550 networks. Values outside of the 95% confidence interval were identified as statistically significant.

Calibrating node removal sensitivity using random removals

A random node removal procedure was implemented to assess on average how many nodes would need to be removed from a network before flow to each transcription factor was significantly altered. An increasing number of nodes were randomly chosen and removed from the EGFR-HQ network, and the process repeated 1000 times. This

simulation resulted in a distribution of expected impact scores for each node in the network, given a certain number of random removals. I determined on average how many random node removals were required before information flow was significantly altered. This information was used to calibrate the actual node removal thresholds.

3.3.7 Development of a novel algorithm for simulating information flow

It became apparent during my use of ITM Probe and investigation of similar tools which use a network propagation approach, that multiple difficulties existed when attempting to apply them on a patient-specific basis. In the case of ITM Probe, this included outputting scores which were not directly comparable when running the algorithm across different networks (Stojmirović *et al.,* 2012), and in other cases not outputting scores at all (Reyna *et al.,* 2018). There are other limitations of existing tools including not being able to set node-specific dissipation probabilities, having results which are difficult to interpret, and also simply being difficult to use, which all result in issues when attempting to apply these approaches on a patient-specific basis.

In the cases where patient-specific analysis is considered, it is generally combined with static network topology, similarly to personalised PageRank or PARADIGM (Vaske *et al.,* 2010). Such tools are not well suited for the comparative analysis of patient-specific networks where actual network topology differs between individuals. Furthermore, most existing implementations of information flow analysis assume an undirected network. However, many protein-protein interactions, such as those involving protein kinases, do have specific directionality. In addition, when the directionality of interactions is not known *a priori*, it may also be inferred from context using network propagation approaches with high reliability (Silverbush & Sharan, 2019).

Simulated Information Flow For Individualised Networks (SIFFIN)

As no currently available implementations of information flow analysis were able to meet all the criteria desirable for patient-specific information flow analysis, I designed

my own network propagation algorithm which would specifically address the patient-specific use case. I called this algorithm Simulated Information Flow For Individualised Networks (SIFFIN)[8]. Similarly to other algorithms such as ITM Probe, SIFFIN uses a network propagation approach to predict information flow.

Information flow, diffusion, Markov chains, random walks, are all distinctly related processes which may be represented by network propagation. The probability / flow at vertex v at time step k is $p_k(v)$, according to:

$$p_k(v) = \sum_{u \in N(v)} p_{k-1}(u) w(u, v) \tag{3.3}$$

Where the normalised weight of vertex u to v is represented by $w(u, v)$. This may be more compactly represented in matrix notation as the following:

$$p_k = W p_{k-1} \tag{3.4}$$

Where W is a normalised adjacency matrix of the network of interest (a transition matrix). From this it follows that the state at a particular time point may be observed by taking $p_k = W^k p_0$, and that a simple algorithm to obtain the summed probability at each node may be obtained by repeated iteration of p_k.

It is also possible to define a random walk with dissipation by adding a smoothing factor α:

$$p_k = (1 - \alpha) W p_{k-1} \tag{3.5}$$

This is the approach used by SIFFIN to simulate random walks with dissipation, as the summed probabilities are very easy to interpret. In this context α is essentially an exact amount of information score lost at each step. With an extra term, this becomes equivalent to "personalised PageRank" (a.k.a., a random walk with restart (RWR)). The role of α is now to determine how much the prior information (source nodes) influence the final result.

[8] https://bitbucket.org/lynnlab/psnr/src/master/scripts/siffin/

$$p_k = \alpha p_0 + (1 - \alpha) W p_{k-1} \tag{3.6}$$

Typical usage of personalised PageRank / RWR is to identify which nodes are most "relevant" to a given query. As such, the probability p_t at the point of convergence is used as a final score, as this combines both the network topology and prior information information into a single score. In comparison, to model how likely it is for information to pass through a given node, the sum of all probabilities may be used.

Inferring edge directionality in SIFFIN

It is important to note that while SIFFIN is intended to run on a directed network, the actual input to SIFFIN can be a partially-directed PPI network. If directed interactions are known *a priori* they may be encoded as directed edges. Otherwise, edges are assumed to be undirected and are encoded as bi-directional directed edges. The directionality of edges is inferred using the approach described by Silverbush & Sharan (Silverbush & Sharan, 2019). First, a RWR is run from both the set of source nodes and set of sink nodes. Using the differential scores for each walk, the probable direction of each edge is inferred (for any edges for which this is not already known), as per the Diffuse2Direct algorithm (Silverbush & Sharan, 2019). An additional property of this algorithm is that it prevents cycles existing in the resulting network, meaning that the algorithm will converge more rapidly than in the undirected case.

Development of an improved information flow impact metric

Impact scores as previously described in section 3.3.6 identify large relative changes in information flow from a baseline level. As this is a relative measure, they are unable to capture large absolute changes in information flow if the baseline level is too high. To compensate for this, I developed a new metric of information flow score changes, scaled percentage change (SPC).

$$SPC = \frac{x - b}{\sqrt[n]{b}} \tag{3.7}$$

where b is the baseline information flow score, x is the altered information flow score, and n is a scaling factor which determines how far the output is scaled between relative and absolute changes (i.e., $n = 1.0$ is simply percentage change, and higher values bring the score asymptotically closer to absolute change). The scaled percentage change was centred on zero by taking the log2 of 1 + the absolute value of SPC.

3.3.8 Clustering and visualisation of network propagation results

Clustering networks using information flow

Hierarchical agglomerative clustering of information flow scores was performed using the SciPy package (version 1.3.3), with the Ward variance minimisation method (Ward, 1963). Hierarchies were constructed separately for both log2SPC for all nodes across all 550 networks, as well as for transcription factor nodes only. Consensus clustering as described by Monti *et al.* was used to obtain robust clusters (Monti *et al.,* 2003). Hierarchical clustering was performed 1000 times for consensus clustering, and was re-sampled to 75% of the data for each run.

3.3.9 Extracting clusters based on hierarchical clustering

The silhouette coefficient (Rousseeuw, 1987) and consensus cluster scores were used to determine an optimal number of clusters (k). The silhouette coefficient (Rousseeuw, 1987) as implemented by silhouette_samples() and silhouette_score() functions from Scikit-learn and consensus scores as described by (Monti *et al.,* 2003) were used to determine an optimal number of clusters (k).

Comparison of information flow clusters to CMS groups

Clusters were compared to consensus molecular subtypes (CMS) (Guinney *et al.,* 2015) using alluvial plots, and relative survival probabilities assessed using Kaplan-Meier plots and pairwise log-rank tests via the Lifelines package (Davidson-Pilon *et al.,* 2019)

for Python (version 0.25.4). Topological properties for information flow clusters were also assessed as per section 3.3.5.

Visualisation and analysis of information flow informed clusters

To visualise information flow scores on the networks, a network layout procedure was developed to prioritise the visualisation of the source and sink nodes. First, all transcription factors and the EGFR node were fixed in position on the circumference of a circle. Next, while these nodes were held in place, graph-tool's *sfdp_layout* was used to layout all other nodes with a force-directed layout. This resulted in a tree-like visualisation with EGFR near the top, and transcription factors arranged in a semicircle at the bottom (Figure 3.5). Network properties such as node colour and size were used to visualise different patient subgroups.

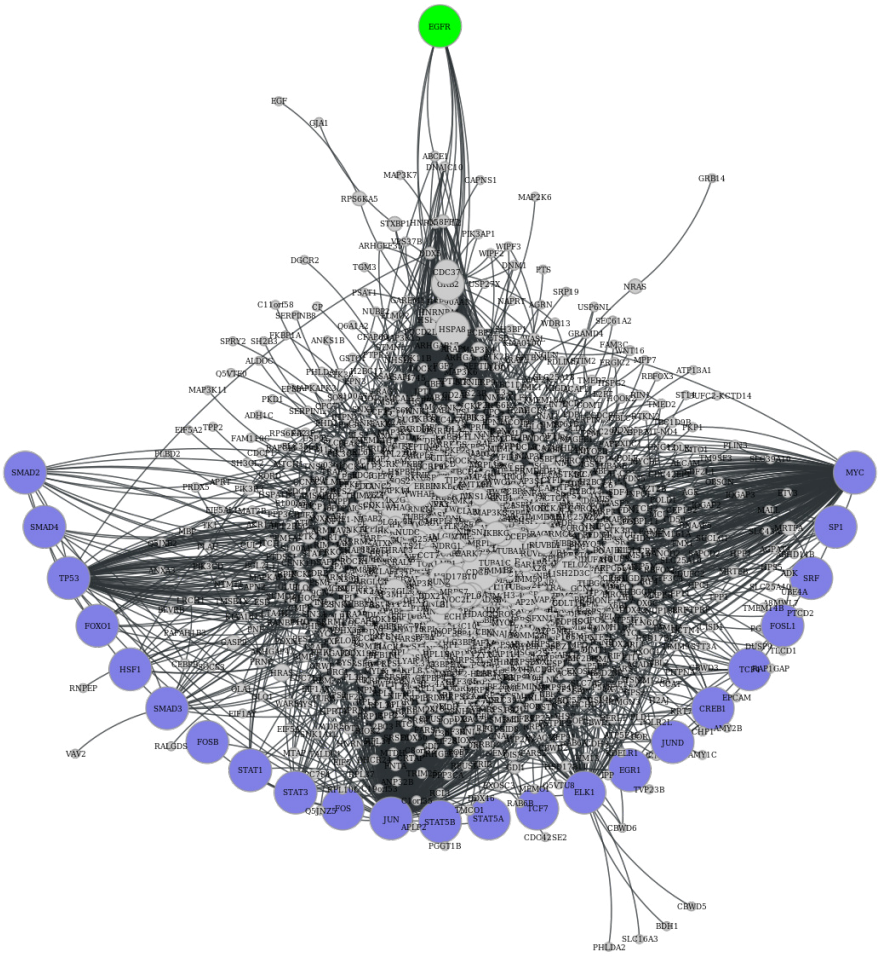


Figure 3.5: *Visualisation of the high quality EGFR network (EGFR-HQ) using a "Christmas tree" style layout. EGFR (green) is at the top of the tree, while the transcription factors (blue) are arranged in a semicircle at the bottom. Node size is scaled in proportion with betweenness centrality.*

3.3.10 Cross-validation of results

Expression of transcription factor target genes

Pypath/Omnipath (Türei *et al.*, 2016) was used as a source of TF to target gene relationships (i.e. TF regulons). DoRothEA (the database from which Omnipath sources its information from) Garcia-Alonso *et al.*, 2019 integrates manually curated interactions, high-throughput TF-DNA measurements (e.g. ChIP), *in silico* predictions, and

predicted interactions from large-scale gene expression profiles. Identifying the genes regulated by each transcription factor downstream of the EGFR network enabled validation of information flow analysis results. Expression of the specific genes that were regulated by each transcription factor was compared to predicted information flow to the given transcription factor. Principal components analysis (PCA) was used to summarise target gene sets, which was compared to TF $\log_2$SPC scores using Pearson correlation. ANOVA and pairwise T-tests were used to compare the expression of TF-coding genes for different patient clusters defined from $\log_2$SPC scores, using functions from SciPy (version 1.3.3) (Virtanen *et al.*, 2020).

3.4 Results

3.4.1 Patient-specific EGFR networks

Constructing a high-quality model of the EGFR PPI network

To investigate my hypothesis that topological heterogeneity in network structure between individual patients is predictive of CRC patient outcomes including survival, I combined patient-specific gene expression data from The Cancer Genome Atlas (TCGA) with human protein-protein interaction (PPI) data from the International Molecular Exchange Consortium (IMEx) to create patient-specific network models. I used the Epidermal Growth Factor Receptor (EGFR) network as a base for these models due to the relevance of the EGFR in CRC and the availability of a high-quality experimentally derived mapping of the network, the PRIMES HCT116 dataset (Kennedy *et al.,* 2020).

To create a high-quality EGFR network (EGFR-HQ) which could then be personalised for each patient, I used the PRIMES HCT116 EGFR network as a framework. Additional nodes were added by selecting known members of the canonical EGFR pathway (Appendix Table 6.2) and 24 transcription factors downstream of EGFR (Table 3.1). Edges were added using experimentally validated PPIs from IMEx (Orchard *et al.,* 2012) between prey proteins, added transcription factors, and canonical nodes in the PRIMES HCT116 EGFR network. The PPI interactions from IMEx are manually curated, and each interaction is assigned an MI score representing confidence in the experimental evidence supporting the interaction, normalised between 1 and 0 (Kerrien *et al.,* 2012). I used a threshold of MI score >0.3 to retain medium confidence interactions (Villaveces *et al.,* 2015) to create the EGFR-HQ network.

The EGFR-HQ network consisted of 1390 vertices and 11,157 edges. In total, 8,134 additional edges from IMEx were added between PRIMES prey proteins, added canonical nodes, and added transcription factors. To ensure that the EGFR-HQ network was a single connected component, any nodes not connected to the largest component were discarded. Following this, only 23 transcription factors remained,

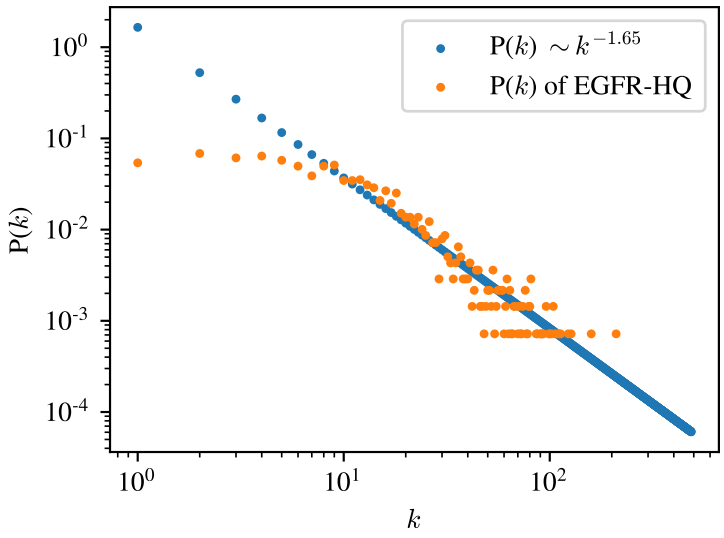


Figure 3.6: *Log-log plot of the degree distribution ($P(k)$) of the EGFR-HQ network, compared to a power law distribution fit of the data (obtained using the powerlaw library for Python (version 1.5) (Alstott* et al., *2014)).*

as EGR2 had no interactions linking it to the rest of the network with an MI score >0.3. The EGFR-HQ network exhibited a scale-free topology typical of PPI networks, with a negative assortativity coefficient ($r=-4\times10^{-3}$) (meaning nodes with similar degrees do not tend to be connected to each other) and a degree distribution which approximated a power law (Figure 3.6). Of interest, the gamma parameter obtained from a power law fit of the degree distribution was 1.65, lower than the usual range of 2-3 for scale-free networks (Barabási & Albert, 1999). This low gamma is likely related to the way the network was constructed. Because edges were only added between prey proteins from the PRIMES HCT116 network, most of which were of low degree, the resulting network inevitably contained fewer nodes of low degree than would be expected, and more nodes of slightly higher degree. Many of the nodes with the highest centralities were PRIMES bait proteins (Table 3.2), an expected result due to high centrality being one of the PRIMES bait protein selection criteria. The most central hubs however were the transcription factors JUN and MYC, as well as CDC5L, up-regulation of which has been shown to contribute to CRC progression via regulation of human telomerase reverse transcriptase (J. Li *et al.*, 2017).

Table 3.2: *The 10 nodes with highest degree (largest hubs) and the 10 nodes with highest betweenness centralities (bottlenecks) in the EGFR-HQ network. PRIMES bait proteins are shown in bold.*

	Degree			Betweenness Centrality	
1	JUN	487	1	JUN	0.2446
2	MYC	210	2	MYC	0.0570
3	CDC5L	159	3	CDC5L	0.0338
4	TUBA1A	127	4	**SH2D3C**	0.0319
5	HNRNPU	123	5	HSPA8	0.0283
6	NPM1	113	6	**GRB2**	0.0255
7	H3C10	111	7	HSP90AB1	0.0246
8	HSPA8	109	8	TUBA1A	0.0221
9	RPL10	106	9	KSR1	0.0206
10	HSP90AB1	104	10	**RAB5A**	0.0187

Creation of 550 patient-specific EGFR network models using a node removal strategy

To personalise the EGFR-HQ network for each individual in the TCGA CRC cohort, I developed a node removal method which identified genes which were significantly under-expressed in individual tumours. The corresponding nodes (and all interactions) were then removed from the network for that patient. This was based on the assumption that the nodes corresponding to genes which were not expressed (or were significantly under-expressed) in patients would not be functional (or have significantly reduced functionality) in terms of capacity to transmit signals in those individuals. Nodes were removed from the EGFR-HQ network for each of the 550 patients. The number of removals varied across networks, with on average 34 nodes (around 2.5% of total) and their interactions being removed from the network per patient (Figure 3.7, A). Which nodes were removed also varied, with each node being removed from different patient-specific networks 15 times on average (Figure 3.7, B). The majority of genes were only removed in a small number of networks. Interestingly, the single

most frequently removed node (removed in 204 networks) was the heat shock protein HSPA6. Previous studies have reported associations between the expression of heat shock protein genes and prognosis in CRC (Chatterjee & Burns, 2017). I did not find an association between *HSPA6* expression and prognosis in this cohort, however I did find a significant association (Cox regression HR=0.72, (95% CI, 0.56 - 0.92), p=0.01) between decreased expression of the related protein HSPA8, the 8th highest degree protein in the EGFR-HQ network (Table 3.2), and poorer patient prognosis.

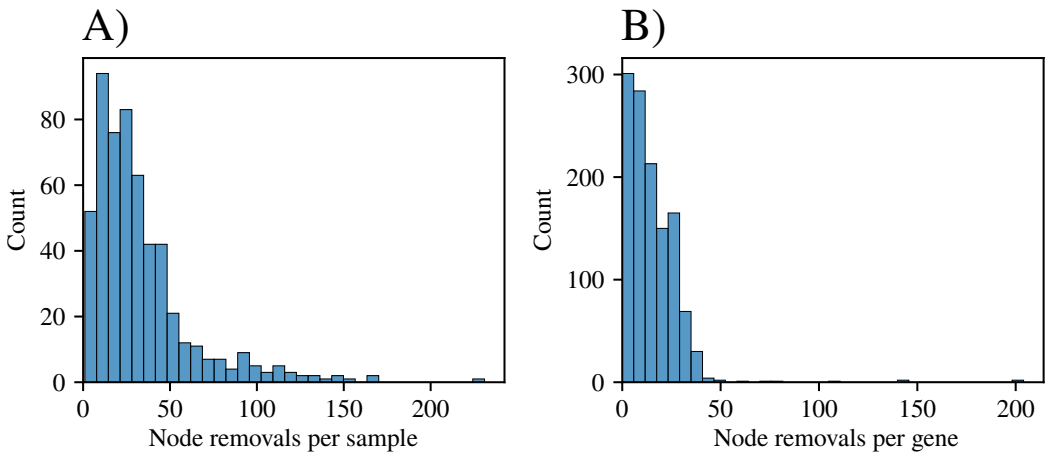


Figure 3.7: *Node removal frequency distributions for A) number of nodes removed per patient-specific network and B) number of times each node was removed across the 550 patient-specific networks.*

Mutations known to disrupt PPIs are present in TCGA CRC patients

Mutations are known to sometimes disrupt PPIs, and systematic studies (Sahni *et al.,* 2015) have been performed to characterise these. One public dataset of experimentally validated PPI disrupting mutations is curated by the EBI (del-Toro *et al.,* 2019) In addition to using transcriptomic data to inform node removal, I made use of this dataset combined with patient-specific genomic mutation data to remove specific edges from the patient-specific networks. Mutations in individual patients (as called by the Mutect variant caller) were matched to the EBI dataset of PPI disrupting mutations (del-Toro *et al.,* 2019). In total, 344 network edges in the EGFR-HQ network were also edges in the del-Toro *et al.* PPI disrupting mutations dataset. Of the mutations present in the dataset, 9 were matched to specific mutations found in the TCGA

CRC data (Figure 3.8). In the TCGA CRC cohort, these mutations occurred most frequently in the *BRAF* and *PIK3CA* genes, mutations which have previously been noted to increase in prevalence in metastatic CRC (Christensen *et al.,* 2018). These matched mutations corresponded to 6 different interactions which could be removed from patient-specific networks. These specific edges were MAP2K2-BRAF, PIK3R1-PIK3CA, GRB14-NRAS, ACTB-ACTG1, CBL-SPRY2, and ERBB3-ERBB2 (Figure 3.8). Examining each group of patients harbouring these mutations did not reveal significant differences in patient survival between them. Interestingly, 20 genes were found to be significantly differentially expressed. Pathway enrichment analysis (using GO Biological Process) of these genes revealed a significant enrichment for mitotic cell cycle phase transition (FDR = 0.04), potentially indicating changes to the cell cycle in patients with these mutations.

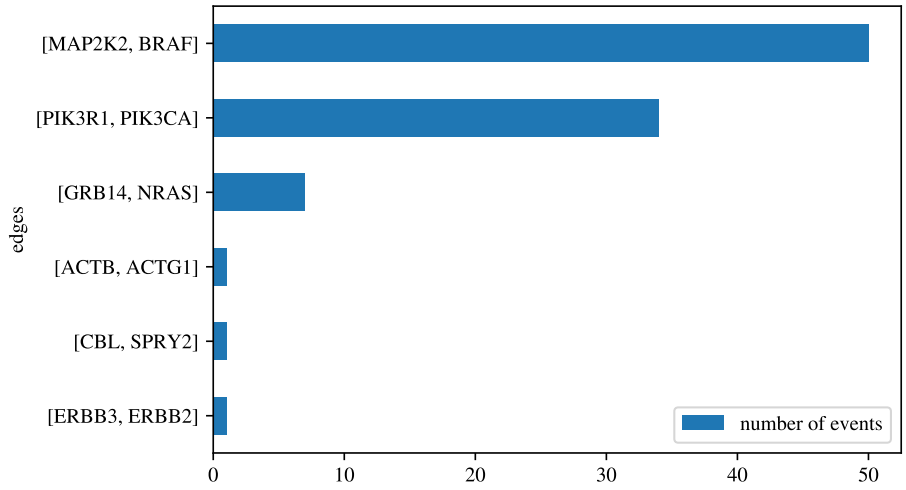


Figure 3.8: *Frequency of EGFR-HQ network edge removal events due to PPI disrupting mutations occurring across the entire cohort of TCGA CRC patients (n=550).*

Specific protein domains are significantly enriched among binary interactions

As there were relatively few specific cases of edges in patient-specific EGFR networks that were annotated in the del-Toro *et al.* dataset to have PPI-disrupting mutations, I implemented another approach to predict potential PPI disrupting mutations. Using over-representation analysis, I first identified specific protein domains which were

over-represented in both protein partners of binary PPIs, hypothesising that these domains likely mediated PPIS. I identified 260 specific domains which, given their prevalence interacting PPIs, were more likely to mediate PPIs. These domains had a relatively high overlap with domains predicted from structural data, with 191 of them being identified in 3DID, a database of domain-domain interactions obtained using high-resolution three dimensional protein structural data (Mosca *et al.,* 2014). The most common of these domains was PF00069, a protein kinase domain containing the catalytic function of protein kinases. The full list of significant domains is available from the Lynnlab bitbucket[9]. While these protein kinase domains likely do not directly mediate interactions, protein kinases are important regulatory components of signal transduction pathways, with phosphorylation-induced conformational changes frequently regulating the activity of PPIs (Wee & Z. Wang, 2017).

I found that 582 nodes in the EGFR-HQ network contained at least one of these PPI-mediating domains, with 502 interacting pairs (435 unique nodes) sharing at least one domain. To link this information on protein domains to patient-specific data, I examined non-synonymous mutations from the TCGA, finding 264 of the nodes containing probable PPI mediating domains had at least one instance of possibly damaging (as determined by SIFT score) mutations within these domains.

[9] https://bitbucket.org/lynnlab/mutation/src/master/significant_domains.csv

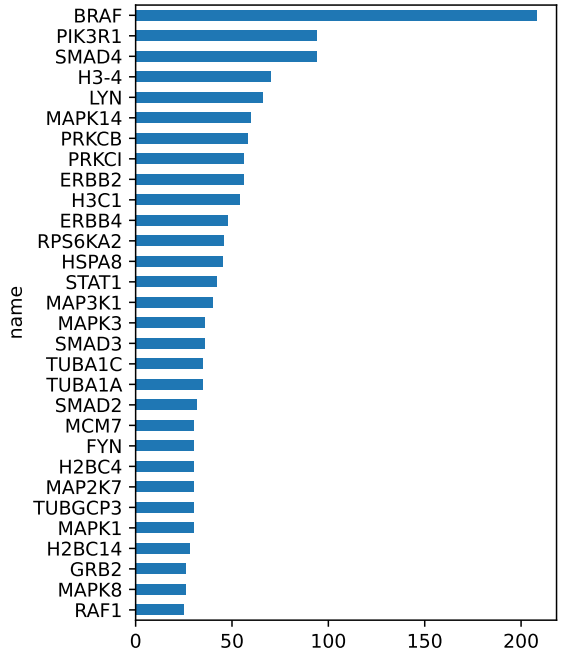


Figure 3.9: *Nodes in the EGFR-HQ network most frequently found to have non-synonymous mutations within domains that potentially mediate PPIs .*

The node most commonly mutated in this manner was *BRAF*, which interestingly was also the most common PPI disrupting mutation found using the del-Toro *et al.* dataset. The full list of these proteins is also available from the Lynnlab bitbucket[10]. Performing pathway enrichment analysis on these nodes revealed a significant enrichment for protein phosphorylation.

I found that the degree of nodes containing domains enriched in binary PPIs was significantly higher than in other nodes ($p=4.4 \times 10^{-15}$). This result was expected, as nodes with more edges had more opportunity for the enrichment analysis to identify them. I further investigated whether the degree of nodes that additionally had TCGA CRC patient mutations within these domains differed to the degree of nodes with domains but no mutations. I found that nodes with mutations in domains that potentially mediate PPIs were also of significantly higher degree than other nodes with these domains without mutations in them ($p=2.8 \times 10^{-6}$) (Figure 3.10). A possible explanation of this result is that it is due to survivorship bias. Higher degree nodes have many interacting partners, and so despite these mutations disrupting certain

[10] https://bitbucket.org/lynnlab/mutation/src/master/significant_proteins.csv

interactions, signalling may continue to propagate on alternative routes. In lower degree, less redundant nodes, mutations in these regions may cause disruptions that are more lethal to the tumour.

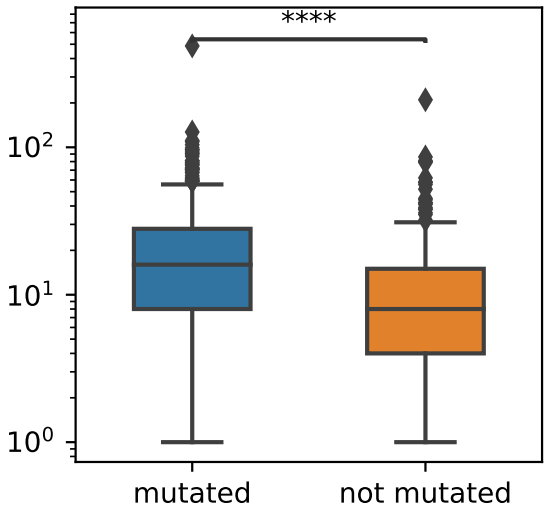


Figure 3.10: *Nodes in the EGFR-HQ network containing domains that likely mediate PPIs were identified. The degree of these nodes in which TCGA CRC patient mutations were identified within these domains, was compared to the degree nodes in which these domains had no mutations.*

More edges are impacted by PPI-disrupting mutations in MSI subtype tumours

Matching both the PPI disrupting mutations and potentially damaging mutations within domains suspected to be involved with PPIs, I found a total of 151 patients with at least one network edge that was disrupted. In 83 of these cases, more than one edge was disrupted. Among these patients, I found that there was a significant enrichment of the hypermutated MSI subtype ($p=1.6 \times 10^{-4}$, Fisher's exact test). In total, 151 patients had edges identified as likely disrupted, combining both the del-Toro *et al.* dataset and predicted disruptions based on shared protein domains. The highest number of disruptions in a single patient was 23, (occurring in a tumour sample from the CMS1, MSI-high subtype). The single most common edge disruption was the *BRAF* - *MAP2K2* interaction, occurring 50 times across the cohort due to the well known BRAFV600E mutation. Patients harbouring this specific mutation are

sometimes considered a unique subgroup of CRC due to its prevalence (Molina-Cerrillo *et al.,* 2020). BRAFV600E is also enriched in CMS1 subtype tumours (Dienstmann *et al.,* 2017). Considering the total size of these networks, the total number of edges removed by this approach was small and did little to alter overall network topology. However, the distinct topological properties of nodes affected by these mutations and the localisation of these mutations within specific patient subtypes suggest that these mutations may be important for understanding the biology of these tumours.

Global network properties vary significantly between CRC patient subtypes

To determine whether the topology of these patient-specific networks was linked to biological variation between patients, I assessed whether various measures of network topology were significantly different between known patient subtypes, both the Consensus Molecular Subtypes (Guinney *et al.,* 2015), and the PSDE-informed clusters (PICs) as defined in Chapter 2. I selected key parameters of network topology based on previous studies which have indicated that topological network properties may be predictive of patient survival. For example, Breitkreutz *et al.* made use of network complexity measures to show that the 5-year patient survival probability with different cancer types was correlated with network complexity (Breitkreutz *et al.,* 2012). Network complexity was assessed using degree distribution entropy, which was found by Breitkreutz *et al.* to be negatively correlated with survival. Given this effect was observed across different cancer types, I hypothesised that a similar effect would likely be observed between individuals with the same cancer. These measures were not significantly associated with patient survival, however I did find that they varied significantly between patient subtypes.

The available literature suggested that measures of network centrality, complexity, connectivity, and size may be linked to patient outcomes. To investigate whether I could reproduce these findings, I calculated mean betweenness, mean PageRank, central point dominance (measures of centrality), degree distribution entropy, assortativity (measures of complexity), global and local clustering coefficient, mean degree (measures of connectivity), and finally mean path per patient-specific network. The number of node removals per network was also assessed as an independent property. I

examined the distributions of these measures, and found that for the most part they were normally distributed and were somewhat variable between individuals, although the amount of variation was sometimes small, especially for measures like betweenness and PageRank (Figure 3.11). As large-scale rewiring of a network would lead to alterations in clustering coefficient, I hypothesised that some of the more highly mutated tumour subgroups such as the microsatellite instability (MSI) subtype might exhibit greater changes in local and global clustering coefficients. I also hypothesised that the removal of bottleneck nodes, which often represent potential drug targets, thus lowering measures of centrality, might be seen in patients with more drug-resistant tumours. Finally, as path lengths and the shortest paths between nodes in networks are properties which capture the way in which information spreads in a given network (Barabási & Albert, 1999), I hypothesised that these measures may be useful to compare to the results of information flow analysis, in which networks which have longer paths between nodes should dissipate information more quickly. The rapid dissipation that occurs in information flow analysis is due to the way dissipation is calculated, as a reduction in total signal that occurs on each step, meaning that if the minimum path length between a source and sink node is large, signal has a chance of dissipating entirely before ever reaching the sink.

Examining the distribution of the chosen network properties across all 550 patient-specific CRC networks (Figure 3.11) revealed consistent inter-patient variability, however most properties were very closely distributed around the median across all networks. Comparison of network properties across CMS groups (Figure 3.12) using ANOVA revealed that two of these measures, central point dominance (CPD, a measure of centrality) and mean local clustering coefficient (MCC, a measure of connectivity) differed significantly between groups. Pairwise T-tests revealed that all of the differences involved CMS4, the mesenchymal subtype which also has the poorest outcomes in terms of survival. In comparison, the analysis of PICs (Figure 3.13) also identified differences in CPD and MCC, with additional differences identified in both assortativity and global clustering coefficient (GCC). Pairwise T-tests of these properties revealed that PIC1 was the group that was most frequently significantly different to other groups, an interesting result given that PIC1 and CMS4 have almost

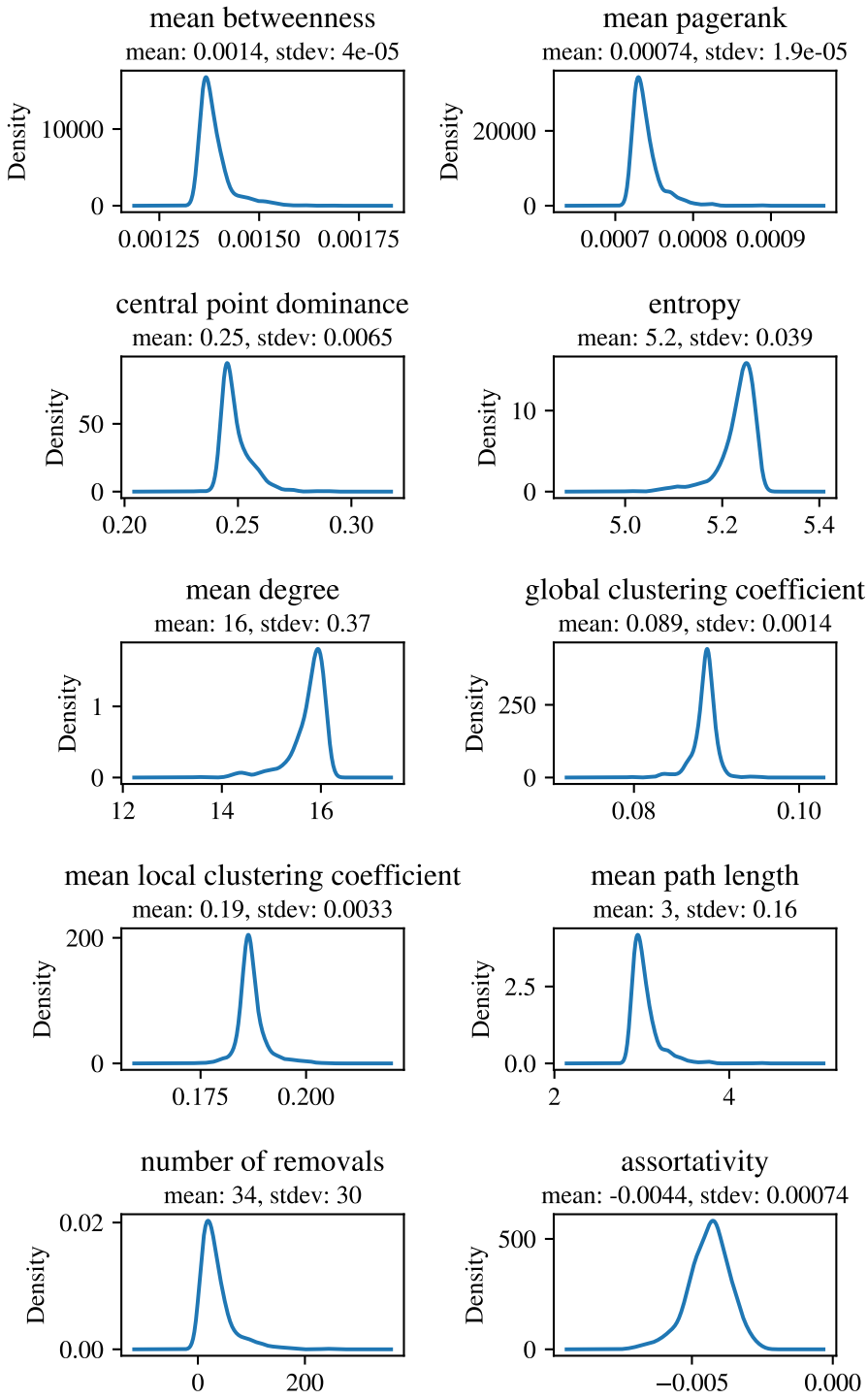


Figure 3.11: *Density distribution plots of network topology measures assessed across 550 patient-specific EGFR networks. Mean and standard deviation (stdev) is reported for each measure.*

no overlap (PIC1 is mostly comprised of CMS2 (canonical subtype) samples).

CMS4 was consistently significantly lower than other groups across the measures of CPD, MCC and GGC, while PIC1 was consistently higher. While the effect size was relatively small (as might be expected as only small alterations were made to a relatively large network) the statistical significance of the differences between patient clusters was high. CPD is a measure which describes the degree to which a single node may dominate communication in a network, in contrast to other centrality measures such as betweenness which measure the centrality of all nodes. For a network in which all nodes have equal centrality, this CPD will be 0, while for a star-shaped network in which all edges connect to a single node, CPD is 1 (Freeman, 1977). A significant reduction of CPD and LCC in the CMS4 subtype indicates the removal of higher-degree nodes in this cluster, a factor which may contribute to the poorer prognosis of this subtype.

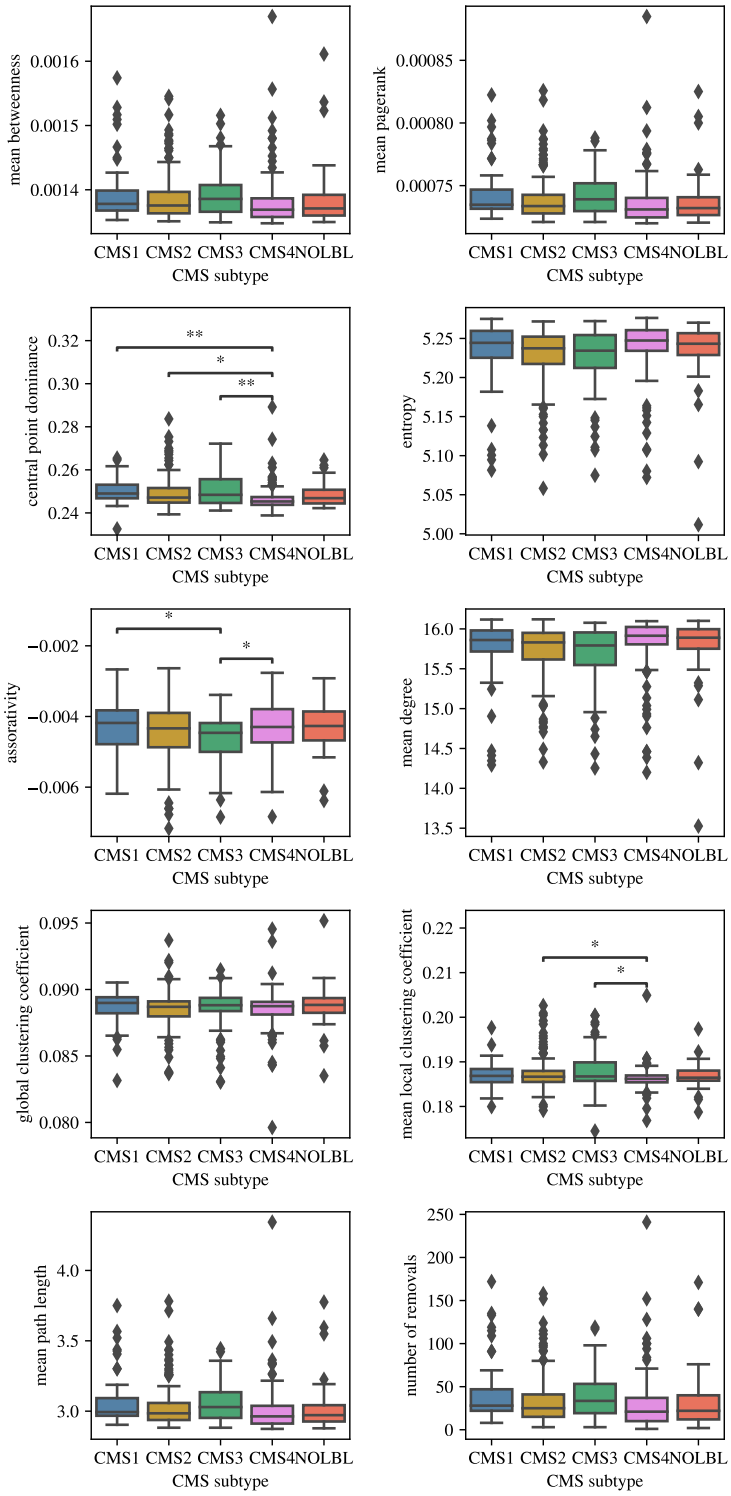


Figure 3.12: *Network properties compared across different Consensus Molecular Subtypes (CMS). One-way ANOVA was used to determine if any properties differed significantly. If statistical significance was reached at* $p < 0.05$ *this was followed by post-hoc pairwise T-tests to identify the different groups* ($* = p < 0.05$, $** = p < 10^{-2}$) *(Bonferroni correction for multiple testing applied).*

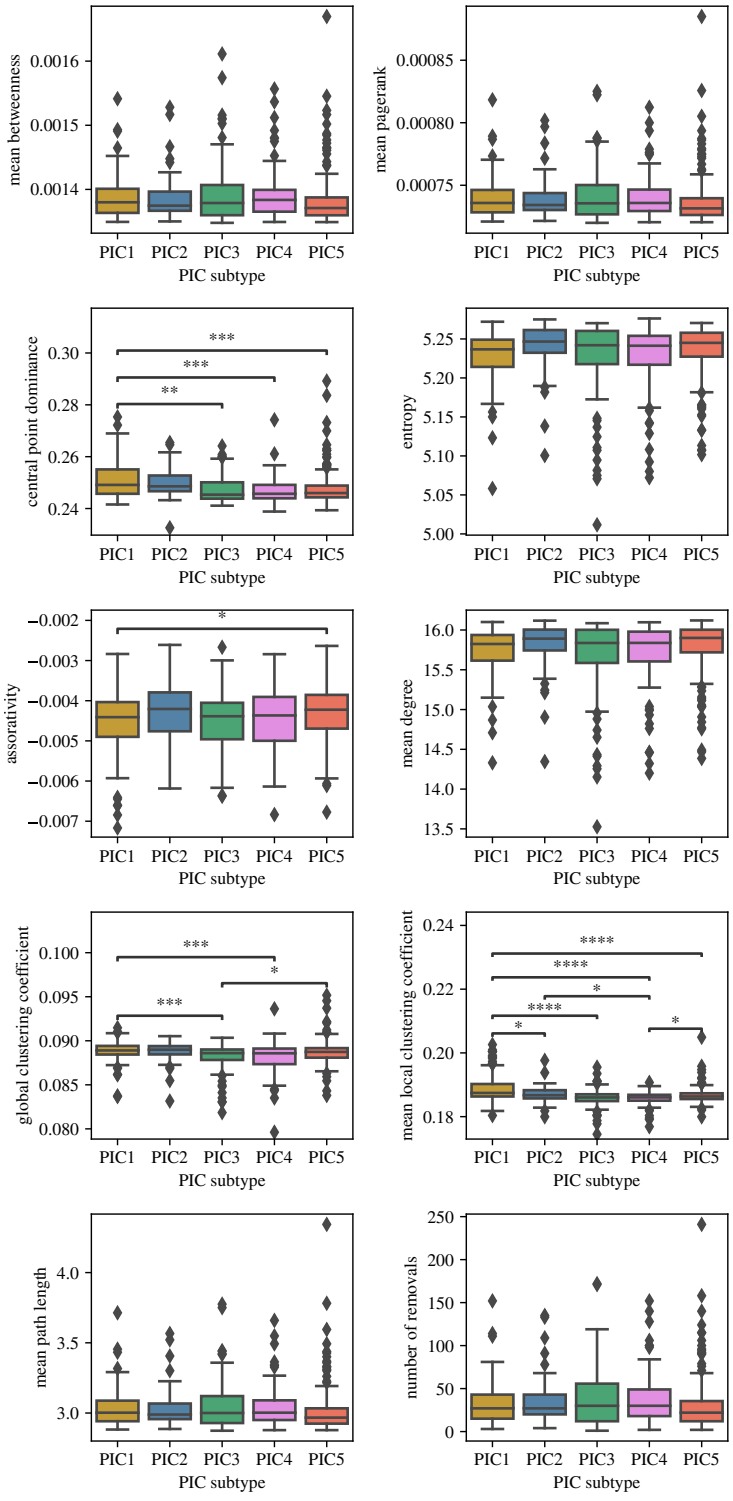


Figure 3.13: *Network properties compared across different PSDE-informed clusters (PICs). One-way ANOVA was used to determine if any properties differed significantly. If statistical significance was reached at* $p < 0.05$ *this was followed by post-hoc pairwise T-tests to identify the different groups* ($* = p < 0.05$, $** = p < 10^{-2}$, $*** = p < 10^{-3}$, $**** = p < 10^{-4}$) *(Bonferroni correction for multiple testing applied).*

Local network properties vary significantly between patient subtypes

While global measures of network topology were variable across patient subtypes, I also examined local node-specific network properties, with the assumption that topological differences may be localised to certain regions of the network. Most of these properties were stable across patients, however certain nodes showed significant variation in topological properties such as degree. Notably the transcription factors MYC and JUN were the most variable in terms of degree. Complete data on node-specific properties for all 550 patient-specific networks may be downloaded from the Lynn lab Bitbucket repository[11]. One notable node variable in degree was GRB2. GRB2 is a key downstream hub protein known to coordinate multiple aspects of EGFR signalling (Bisson *et al.*, 2011). This node was previously found to receive a high level of information flow score by Kennedy *et al.* Comparing the degree of GRB2 in different CMS groups revealed that GRB2 degree was significantly higher in CMS4 patients than most other CMS groups (Figure 3.14).

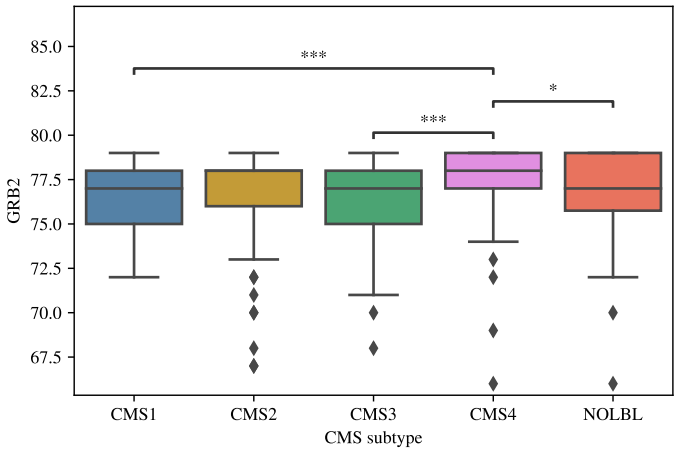


Figure 3.14: *Degree of GRB2 compared across CMS clusters. Pairwise T-tests used to assess statistical significance* ($* = p < 0.05$, $** = p < 10^{-2}$, $*** = p < 10^{-3}$, $**** = p < 10^{-4}$) *(Bonferroni correction for multiple testing applied).*

In comparison to the global trend of decreased centralisation in CMS4 patient networks, as seen in the global CPD and LCC measures, GRB2 degree was instead significantly increased in CMS4 patient networks. However, as CMS4 is the mesenchymal

[11] https://bitbucket.org/lynnlab/psnr/output/nodespec.csv

subtype, typically associated with late-stage metastatic CRC, the increased number of interactions GRB2 has in CMS4 patient networks may be as a result of the critical role GRB2 is known to play in tumour metastasis (Giubellino *et al.*, 2008).

Measures of network topology are not directly associated with patient prognosis

To determine whether patient prognosis could be predicted using these network properties I assessed the association between measures of network topology and patient survival using Cox regression analysis (Figure 3.15). This revealed that there were no significant associations between any of the topological network properties assessed and patient survival time.

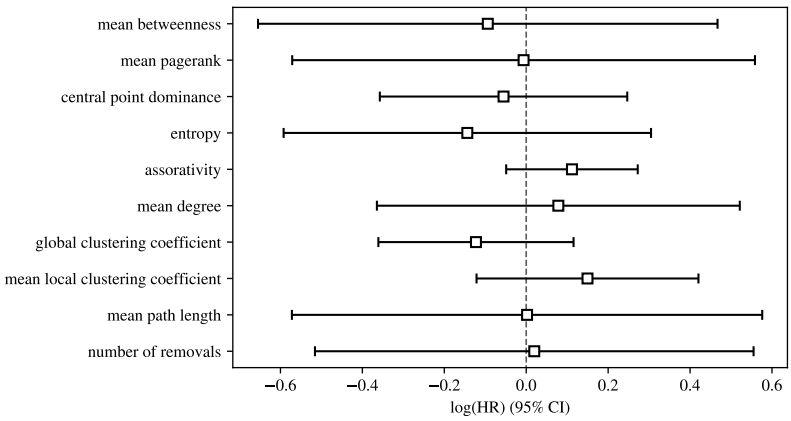


Figure 3.15: *Cox regression analysis of network properties from patient-specific EGFR networks derived from 550 TCGA CRC patients. Log hazard ratios (HR) are shown, with increased HR indicating higher risk.*

While the subtypes themselves were associated with survival differences as described in Chapter 2, Cox regression analysis found that these properties did not independently predict patient outcomes. Still, EGFR network topology in different subtypes was altered enough to change global and local network parameters. This may suggest that substantially differential signalling through these networks should be expected, as these network parameters are related to how signalling changes through different networks.

3.4.2 Modelling information flow through patient-specific networks

Development and validation of a novel information flow tool to investigate between-patient differences

To assess whether patient-specific topological rewiring of the EGFR-HQ network was likely to cause changes in signalling between networks and subsequently lead to differences in downstream transcriptional activation, I designed a network propagation approach to specifically address the patient-specific use case, as described in the Methods section. I called this algorithm Simulated Information Flow For Individualised Networks (SIFFIN). SIFFIN is able to simulate information flow between topologically distinct networks and generate scores that are directly comparable. In addition, it is capable of automatically inferring directionality of edges using an approach similar to Silverbush & Sharan (Silverbush & Sharan, 2019), as well as apply node-specific dissipation probabilities, both things that existing network propagation tools such as ITM Probe (Stojmirović *et al.,* 2012) do not incorporate.

To facilitate between-network comparisons, I also developed a metric to compare information flow score differences called scaled percentage change (SPC), which represented a percentage change which was scaled between the relative and absolute differences in information flow score between individuals using a scaling factor n. Purely relative measures of differences like fold change were not suitable when comparing across all nodes as most baseline information flow scores were very low, such that even a modest absolute change in information flow would lead to an extremely large relative difference. As SPC does not rely purely on relative changes, it was robust for comparing scores across entire networks. To identify nodes which had statistically significant changes in $\log_2$SPC scores compared to the baseline network, the distribution of all $\log_2$SPC scores for all nodes across all networks was determined. Values outside of the 95% confidence interval were considered statistically significant.

In order to test SIFFIN's effectiveness, I directly compared it to ITM Probe (in emitting mode) in a simulated node removal. For this comparison, I made use of a

small, directed model of the EGFR network constructed by Samaga *et al.* (Samaga *et al.,* 2009) originally intended for use as a boolean logic network. In comparison to the EGFR-HQ network, this network was fully directed (Figure 3.16). It also included each of the four ERBB proteins as nodes, which I used as sources. One of the central pathways of this model was the PI3K/AKT/mTOR signalling cascade, which I chose to target via simulated node removal. PI3K/AKT/mTOR is often hyper-activated in CRC, with the expression of PI3KCA, the catalytic subunit of PI3K, commonly being mutated along with KRAS in CRC development (Cathomas, 2014). The PI3K/AKT/mTOR pathway regulates processes such as metabolism, proliferation, and survival (Wee & Z. Wang, 2017), meaning removing PI3K from the network should result in significant downstream changes to signalling. With the removal of PIK3, ITM Probe in emitting mode and SIFFIN were used to simulate information flow (Figure 3.16). Log_2 scaled percentage change (SPC) scores were calculated for each node as described in the Methods section (with a scaling factor of $n = 1.5$). Using ITM Probe, two nodes with significantly different $\log_2$ SPC scores (excluding PI3K itself) were identified, PI3KR and PI34P2. These messenger molecules (directly up and down-stream of PI3K respectively) were predicted to receive significantly reduced flow by this method (Figure 3.16, A). In comparison, SIFFIN predicted significantly reduced flow only to nodes downstream of PI3K, including SOS1/EPS8/E3B1, RAC/CDC42, VAV2, AKT, and AKTD, as well as PI34P2 (Figure 3.16, B.) The most obvious difference between the two methods is that in the SIFFIN approach, as all edges are considered to be directed, nodes directly downstream of removed node are more strongly influenced than in the ITM Probe approach, in which information may flow in both directions. The fact that significant reduction of signal to AKT (as would be expected from removal of PI3K, a major activator of AKT (Wee & Z. Wang, 2017)) did occur in SIFFIN, and not in ITM Probe highlights a distinct advantage of the SIFFIN approach. SIFFIN also predicted more substantial flow increases as a result of the node removal. In SIFFIN, the removal was compensated for by a predicted increase in the signal flow via the RAS/RAF/MEK/ERK pathway to downstream transcription factors, resulting in increases in predicted flow to certain transcription factors including FOS and MYC, while in ITM Probe's results, the net downstream effect was only to reduce total flow. In summary, this simulation indicated that SIFFIN

had clear advantages for predicting downstream signal alterations, largely deriving from the fact that it always modelled information flow as a uni-directional process.

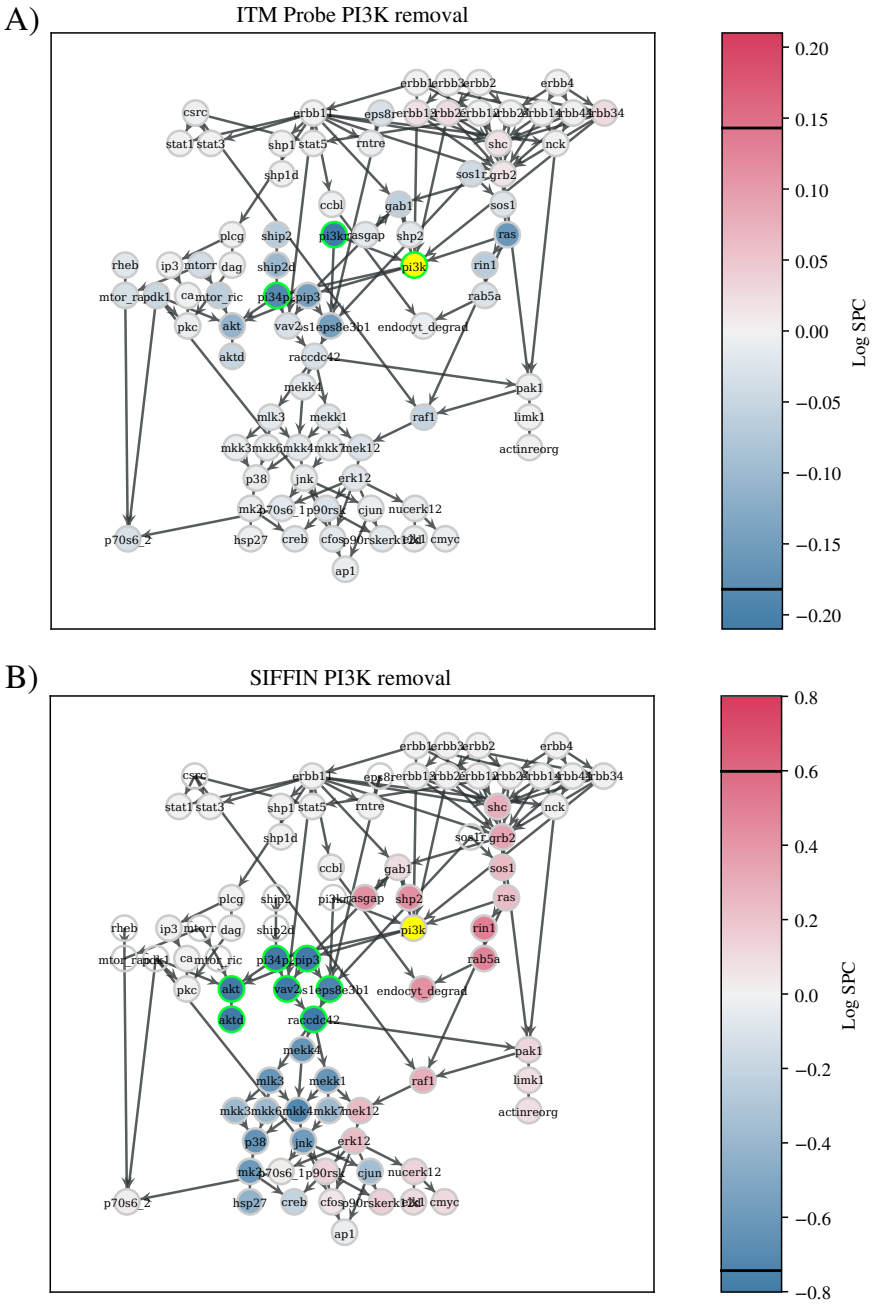


Figure 3.16: *Comparison of the log_2 SPC scores resulting from the removal of the PIK3 node in the model of the EGFR network described by Samaga* et al. *when using A) ITM Probe in emitting mode, and B) SIFFIN to simulate information flow from the four ERBB proteins. The removed node, PI3K, is highlighted in yellow. Nodes with significant log_2 SPC scores are highlighted with a green border.*

Significant SIFFIN scores are correlated with network topology statistics

Using SIFFIN to simulate signal flow from the EGFR node to 23 downstream transcription factors (Table 3.1) in each of the 550 patient-specific EGFR-HQ networks, $\log_2$SPC scores were calculated for each node in each of the 550 networks. For comparison, SIFFIN was also run on the EGFR-HQ network without any node removals to determine baseline score for each node. In addition to node removals, edges were weighted according to the expression of corresponding genes to provide a more accurate model of signal flow through the network. Examining the distribution of all scores, it was apparent that they were approximately normally distributed (Figure 3.17). On average, 64 nodes per network had significantly decreased $\log_2$SPC scores, while on average 30 nodes had significantly increased scores. The number of times a node's score was significantly increased was found to be positively correlated with node PageRank ($p=4.5 \times 10^{-35}$) and degree ($p=5.5 \times 10^{-37}$). These correlations were not identified for significantly decreased nodes, however the number of significantly decreased nodes in a network was significantly negatively correlated with clustering coefficient ($p=3.6 \times 10^{-6}$). These results likely reflect the fact that nodes with a lower clustering coefficient in general would have fewer alternative paths to receive information flow in response to upstream removals.

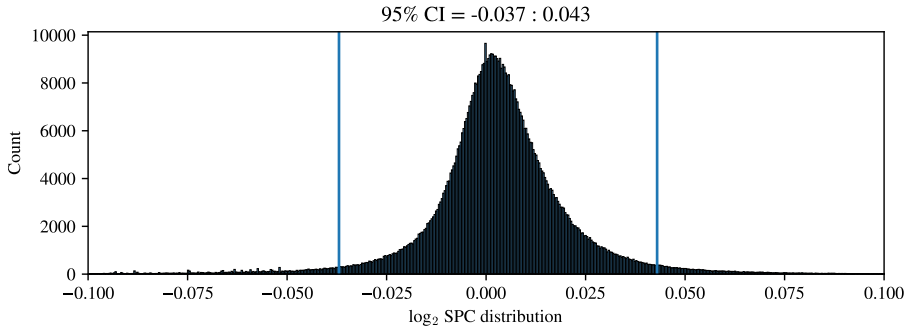


Figure 3.17: *Distribution of log_2SPC scores (with a scaling factor of $n = 1.5$) obtained using SIFFIN for all nodes in 550 CRC patient-specific EGFR networks. The 95% confidence interval is annotated with vertical lines.*

Patient clusters derived from information flow analysis have significant differences in prognosis

To assess whether patient-specific information flow analysis using SIFFIN could identify different groups of patients with altered outcomes, I performed unsupervised consensus hierarchical clustering of all $\log_2$ SPC scores (Figure 3.18, A). Following analysis using consensus scores and silhouette plots to guide optimal cluster number selection (Appendix Figure 6.19), this clustering identified five groups of patients that were labelled flow clusters (FC1-5). These clusters overlapped with the Consensus Molecular Subtypes (CMS) to an extent (Figure 3.19), notably with the largest information flow cluster, FC1, being significantly enriched for the CMS2 canonical subtype ($p=2.1 \times 10^{-7}$). Patients in FC4 were significantly enriched for the CMS1 microsatellite instability (MSI) subtype ($p=3.0 \times 10^{-5}$), whereas patients in FC2 were enriched for the CMS4 mesenchymal subtype ($p=4.4 \times 10^{-5}$). Despite these overlaps, most flow clusters were composed of patients with heterogeneous CMS classification.

Kaplan-Meier analysis revealed that patients in FC2 had significantly poorer survival than patients in FC1 or FC5. Statistically significant differences in survival were identified between FC2 and FC1 ($p=0.0068$) as well as FC2 and FC5 ($p=0.013$). These data are consistent with poorer survival of patients classified as CMS4, the mesenchymal CMS subtype (Figure 3.18, B), but also reveal additional survival differences which were not apparent from the CMS alone, in which only patients in CMS4 had significantly poorer survival. From these data, we can infer that information flow is often as variable within a given subtype as it is across different subtypes. Consistent with analysis of global topological properties, there were significant differences in predicted information flow scores in the patient-specific networks from the perspective of CMS subtypes. Given that these global difference exist, there is strong evidence to suggest that there could be differences in how signals flow through these patient-specific networks to downstream nodes, including transcription factors.

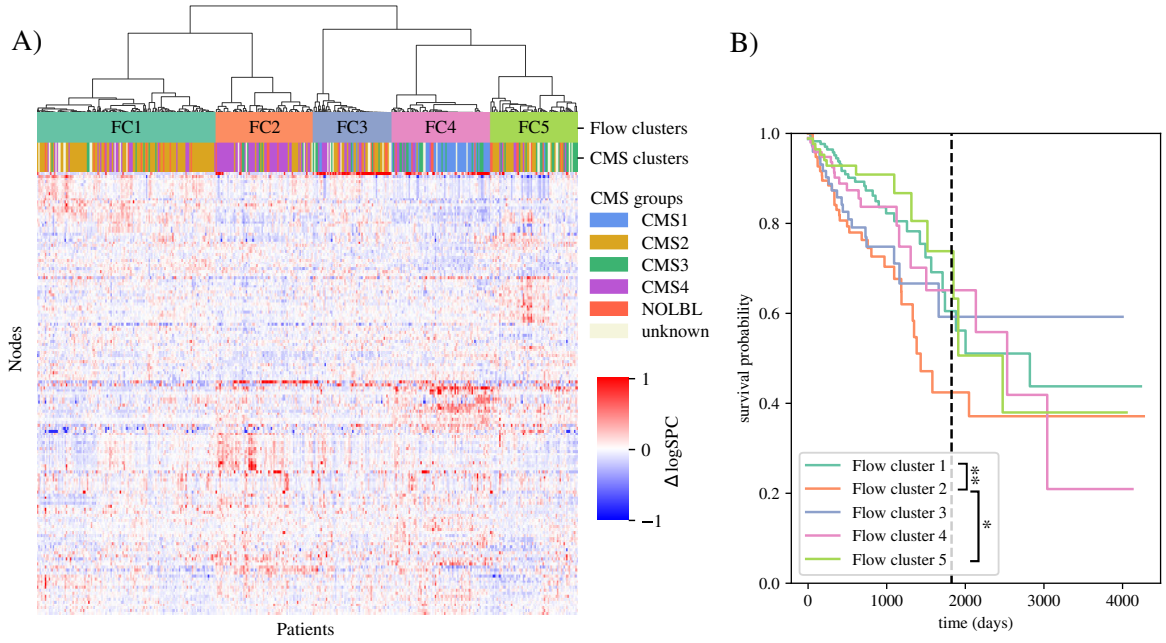


Figure 3.18: *A) Hierarchical consensus clustering of patients using log_2SPC scores for all nodes across 550 patient-specific EGFR networks. Five clusters (FC1-5) were defined. B) Kaplan-Meier plot of patient survival stratified by information flow derived clusters. Statistical significance was assessed using pairwise logrank tests.*

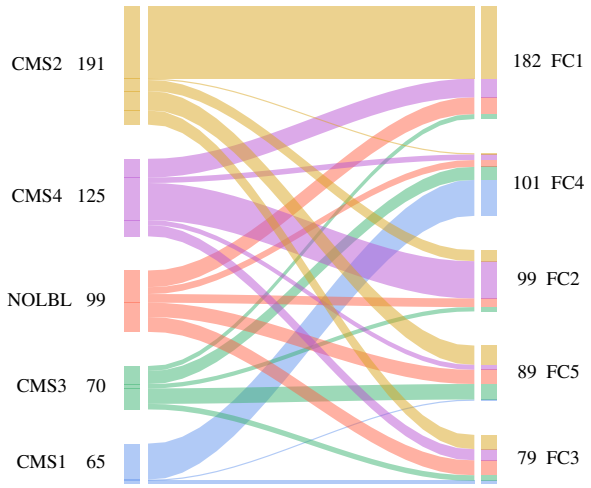


Figure 3.19: *Alluvial plot comparing the stratification of patients by Consensus Molecular Subtypes (left) to flow clusters defined from clustering of log_2SPC scores from SIFFIN (right).*

To highlight the differences between the information flow derived clusters FC1-5, I visualised the EGFR-HQ network using a "Christmas tree" style network layout,

and coloured nodes by the mean $\log_2$ SPC scores for each flow cluster (Figure 3.20). From this visualisation it was apparent that the predicted information flow to transcription factors (which appear at the bottom of the visualisations) was quite variable between different information flow derived clusters, a factor that may be important for explaining the variability in patient survival probabilities.

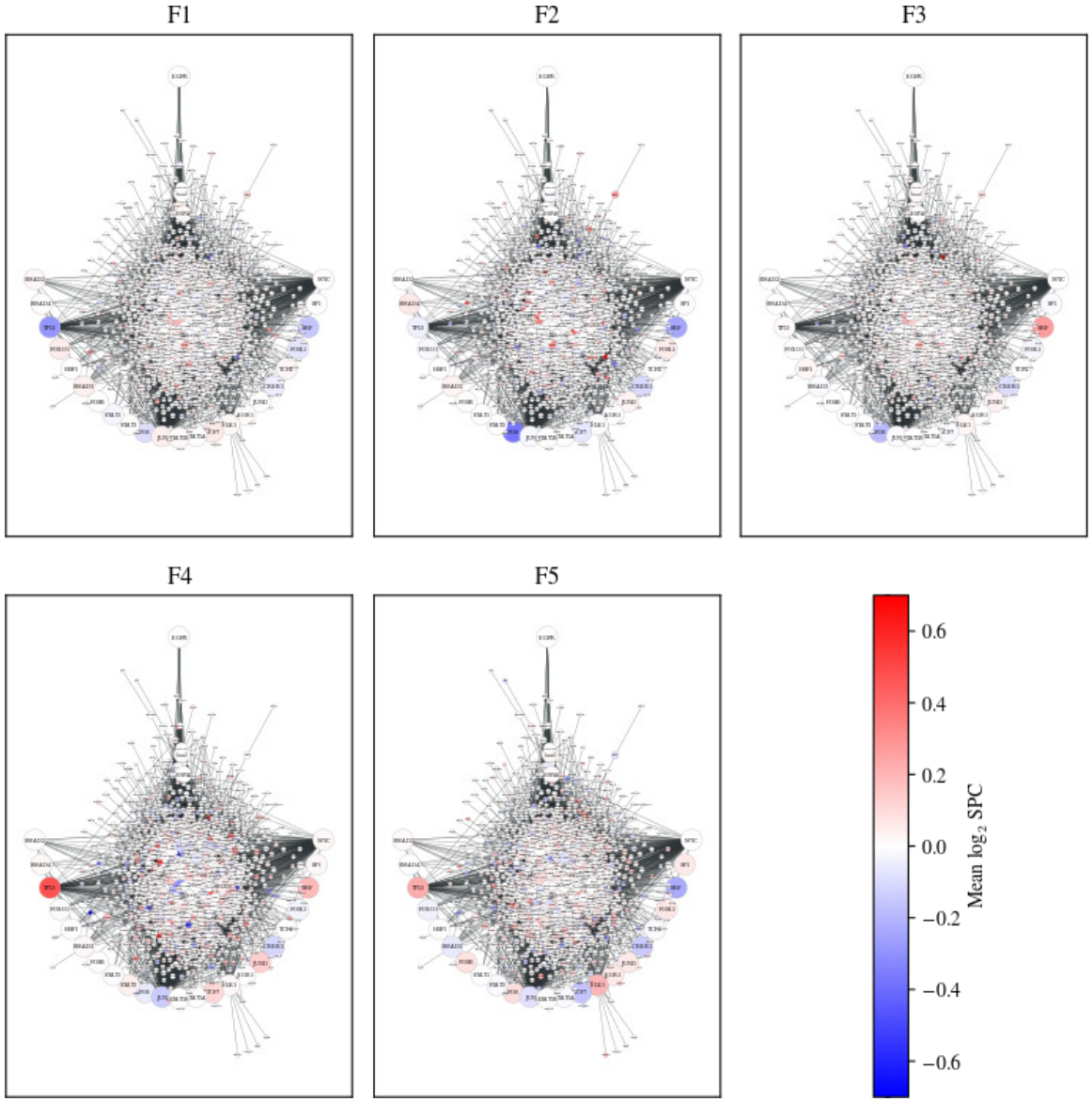


Figure 3.20: *Visualisation of mean log_2SPC scores per information flow derived cluster, using the "Christmas tree" layout style for all nodes. Transcription factors are arranged in a semicircle at the bottom of each network, while EGFR is at the top.*

3.4.3 Modelling the impact of patient-specific network rewiring on information flow to downstream transcription factors

As SIFFIN demonstrated that global changes to information flow in CRC patient-specific networks were prognostically relevant, I next assessed what changes were predicted to downstream transcription factors (TFs), as TFs are the mediators of the cellular responses that likely underlie the heterogeneity of outcomes between patients. A 95% confidence interval was constructed over the distribution of the 23 TF $\log_2$ SPC scores across all 550 networks, and values outside of this interval were considered as statistically significant. The resulting significant scores were then clustered using hierarchical clustering and compared to both Flow Clusters and CMS subtypes (Figure 3.21, A).

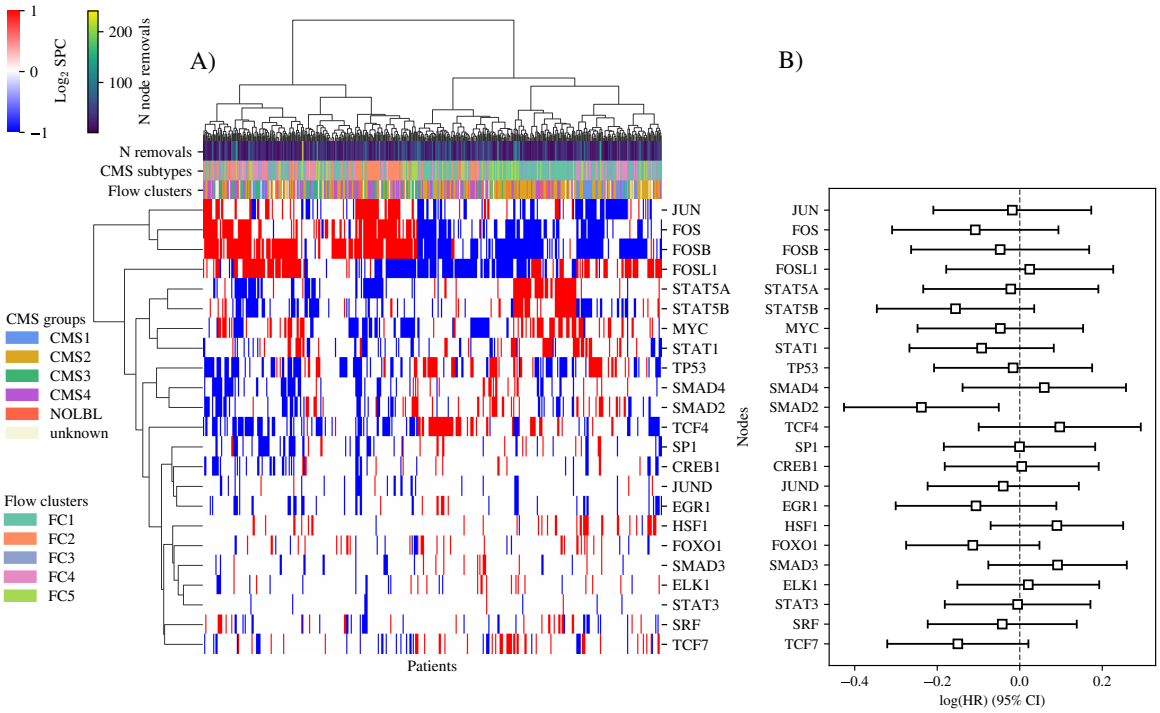


Figure 3.21: *A) Clustering of patients based on $\log_2$ SPC scores ($\log_2$ of individual SPC score to baseline SPC ratio) for each of the connected transcription factors in patient-specific EGFR networks. B) Cox regression analysis of TF $\log_2$ SPC scores obtained from running information flow analysis for 550 patient-specific EGFR networks.*

Significant increases and decreases of information flow to transcription factors were predicted by SIFFIN, suggesting that patient-specific network rewiring alters transcriptional programs downstream of the EGFR network. Interestingly, it appeared that certain transcription factors were more susceptible to changes, such as JUN and FOS which were frequently both significantly increased and decreased relative to baseline information flow. The amount of significant changes also depended on patient groups, increased flow to JUN for example most frequently occurred in the FC2 cluster. Other transcription factors appeared to be more robust to these alterations, such as STAT3, which was only significantly altered in 8 networks.

To investigate whether the differences in flow were predictive of patient outcomes, I next performed Cox regression analysis with patient-specific $\log_2$SPC scores for transcription factors (Figure 3.21, B). This analysis found that increased flow to SMAD2 was positively associated with survival. As a mediator of TGFβ signalling which regulates among other things apoptosis and growth suppression (Wee & Z. Wang, 2017), the association of increased information flow to SMAD2 with better prognosis appears to make biological sense. Following this, I performed Kaplan-Meier survival analysis to compare patients with increased versus decreased predicted signal flow for each transcription factor. This analysis found that survival outcomes were significantly poorer for patients with predicted increased flow to ELK1 (p=0.04) and SMAD3 (p=0.03), while outcomes were significantly better for patients with predicted increased flow to TCF7 (p=0.03) (logrank test). (Figure 3.22). These results were interesting, as they demonstrate that while the specific amount of information flow to TFs is not always directly prognostic, patients may still be stratified in a survival relevant manner based on whether the predicted information flow to a TF increases or decreases relative to the baseline.

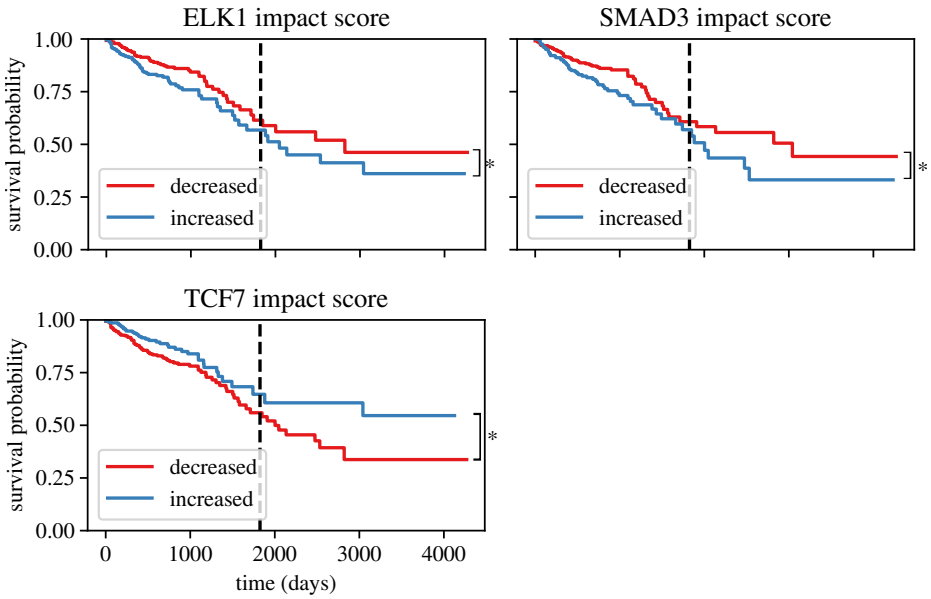


Figure 3.22: *The TCGA CRC cohort was divided into two groups by whether they were found to have increased or decreased transcription factor impact score.*

Random removal analysis demonstrates the significance of patient-specific removals

To assess whether the signal flow to transcription factors downstream of the EGFR network was modified by patient-specific rewiring was not simply due to random noise, I next assessed what level of random rewiring was required to significantly impact information flow. That is, how many nodes on average must be removed before a significant effect is seen on the downstream transcription factors? Literature suggests we should expect that scale-free networks such as PPIs to display a high level of robustness towards random removals, but weakness to targeted attacks (Albert *et al.*, 2000). By randomly removing an increasing number of nodes then modelling the downstream effect on information flow via SIFFIN, I determined the number of node removals required to significantly impact each transcription factor (Figure 3.23). I found that while differences could be observed from even a single removal, statistically significant $\log_2$SPC scores for transcription factors were unlikely to occur by chance alone until around 200 removals, well above the average number of node removals which was used for the 550 patient-specific networks. This indicated that any significant results found were unlikely to be due to random chance. This also demonstrated

that not all transcription factors have the same tolerance against random removal. Interestingly, while each transcription factor had a tendency to either be increased or decreased in information flow compared to baseline, the $\log_2$SPC increased or decreased linearly with an increasing number of random node removals.

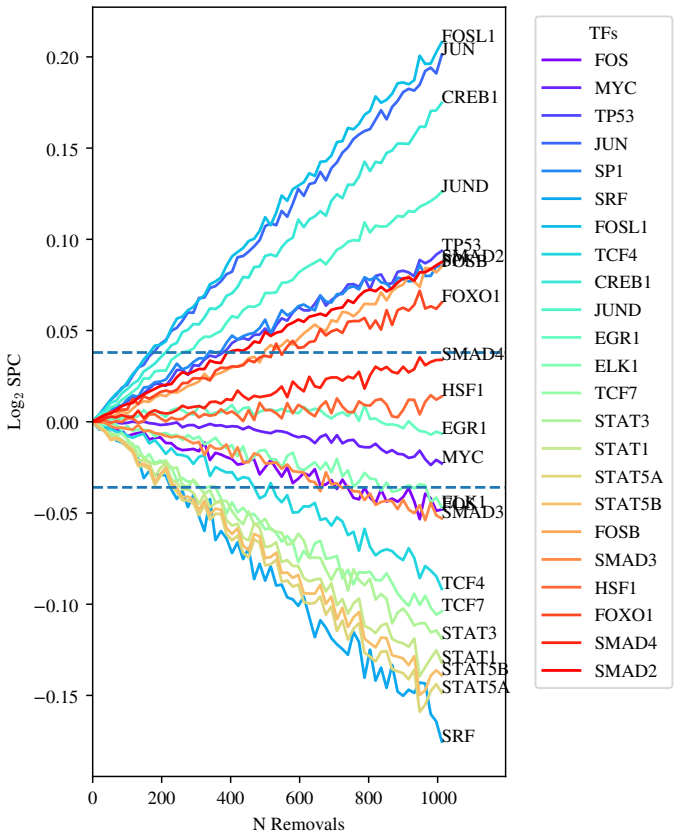


Figure 3.23: *Effect of removing an increasing number of nodes at random on log_2SPC score of selected transcription factors downstream of EGFR. Nodes were randomly removed 1000 times for each number of removals N, and the mean score plotted for each transcription factor. The 95% confidence interval obtained from log_2SPC scores of all nodes obtained across 550 patient-specific CRC networks (Figure 3.17) is annotated with horizontal lines.*

From the information on random node removals, it appeared that the most likely increases in information flow would be to JUN, FOSL1 and CREB1, while the most likely negative changes would be to SRF. Comparing this finding to the information flow results in the 550 patient-specific CRC networks compared to random removals (Figure 3.21, A), while some of the transcription factors with most significant results

were those predicted by the random node removal procedure (e.g., JUN, FOSB), some of the transcription factors predicted to reach significance more easily did not do so very often (e.g., CREB1, SRF). In addition, the changes were not omnidirectional as was predicted by the random node removal. In summary, the transcription factor $\log_2$SPC scores from the patient-specific CRC networks were not the same as those found using the random node removal approach, which suggests that they are a result of actual biological signal rather than simply random chance.

Information flow to transcription factors is correlated with target gene expression

To verify whether the changes predicted by information flow were reflected in the gene expression data, I identified genes known to be regulated by the 23 transcription factors used in my analyses. I sourced these TF-target interactions from the DoRothEA database (Garcia-Alonso *et al.,* 2019). Using principal components analysis (PCA) of the log CPM values of target genes for each transcription factor, I obtained a reduced representation of the activity of those genes. I then correlated these components with $\log_2$ SPC scores (Figure 3.24), finding that many of the target gene sets were significantly correlated when their variance was considered in this way, including SMAD2, flow to which was found to be significantly associated with survival. This is consistent with the expectation that patients with altered information flow to a transcription factor should also have altered expression of genes regulated by that transcription factor.

To further investigate the relationship between information flow and gene expression, I next examined the correlation of predicted information flow ($\log_2$SPC score) to TFs with expression of those same TFs (Figure 3.25). This analysis also identified multiple significant correlations. The correlation coefficient for these correlations tended to be relatively small, indicating that gene expression and information flow score are not directly interchangeable. Further analysis using motif enrichment found few signficantly enriched motifs among patients with significantly impacted transcription factors, with the exception of SP1 and SMAD4 motifs being enriched for patients with significant $\log_2$SPC reductions (Appendix Figure 6.21).

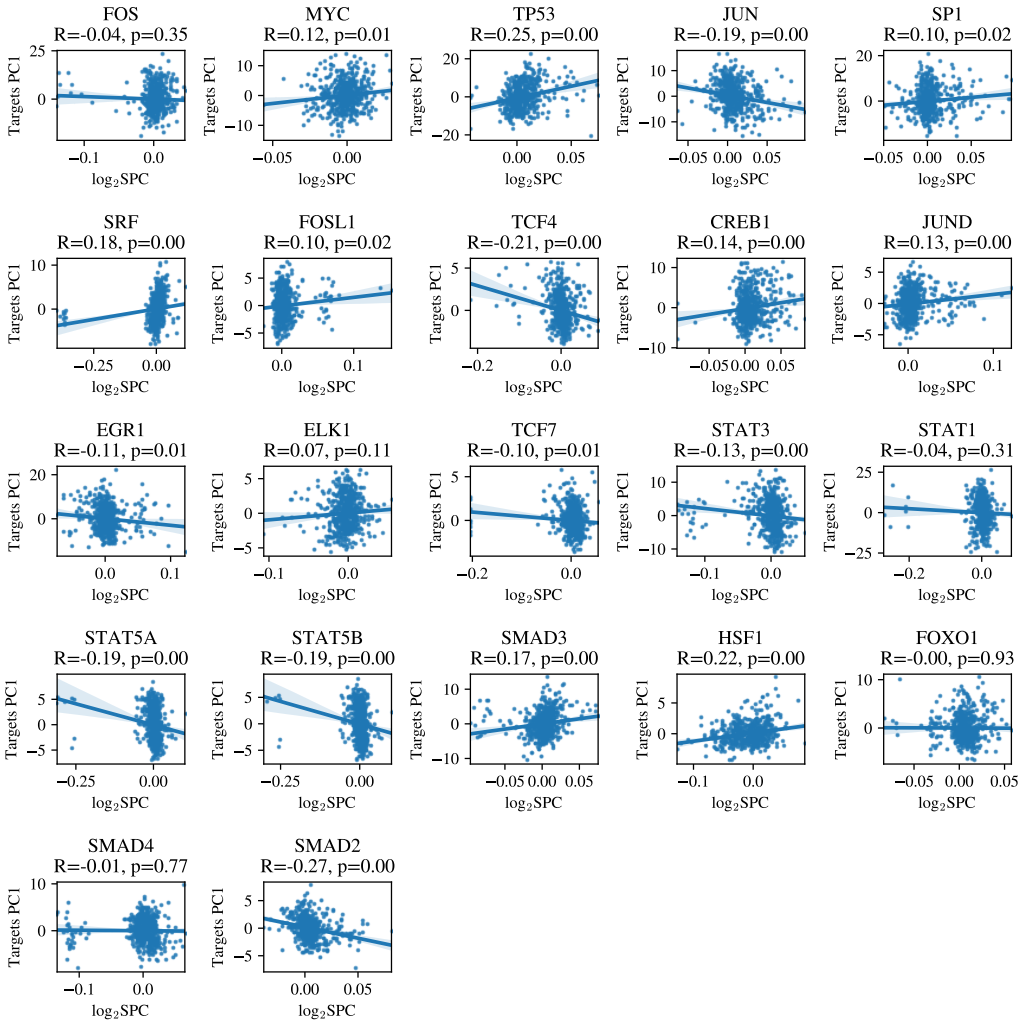


Figure 3.24: *Correlation of information flow score (log_2 SPC) for transcription factors with the first principal component of genes (log_2 CPM) targeted by the corresponding transcription factor.*

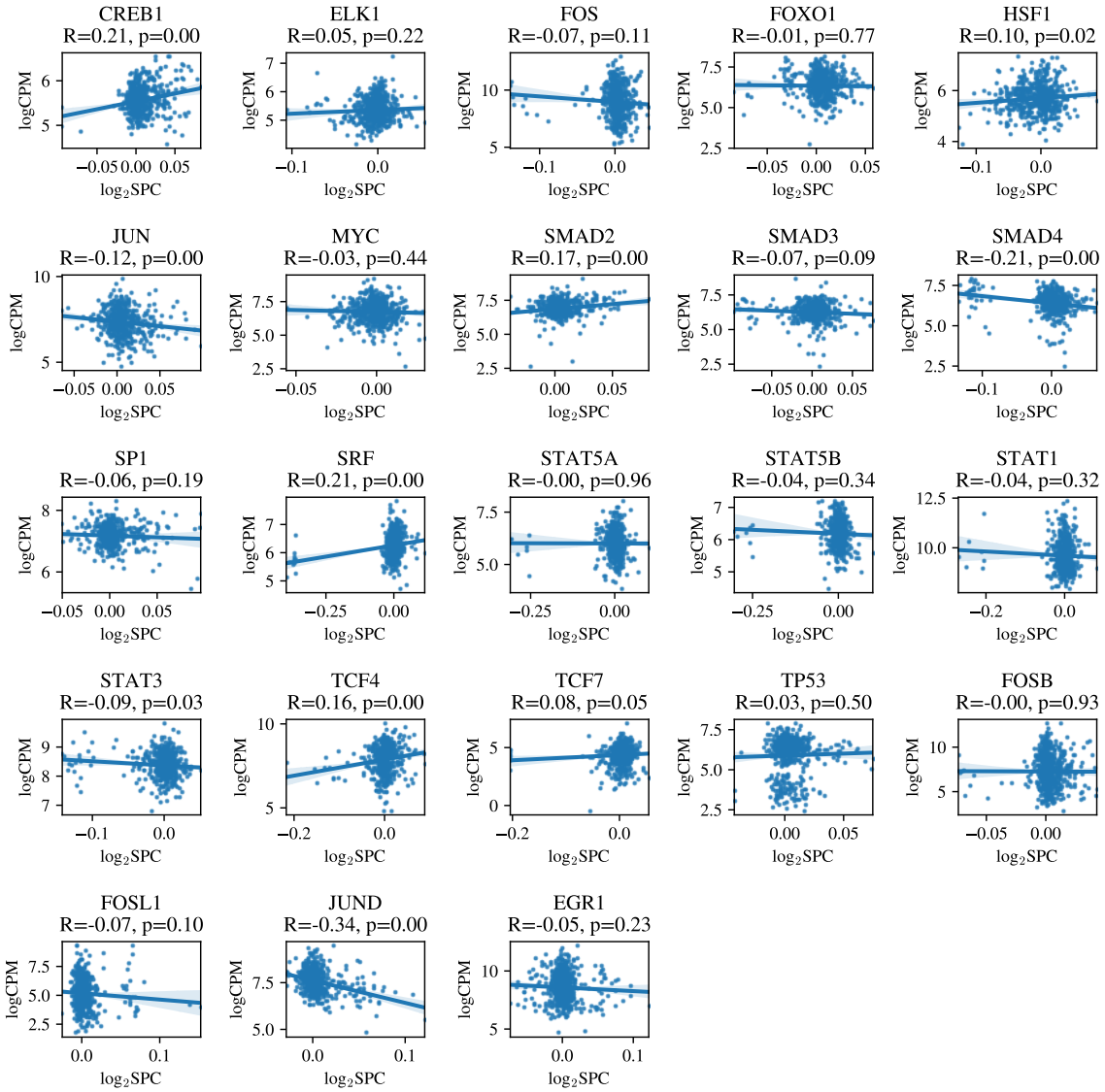


Figure 3.25: *Correlation of predicted information flow (log_2SPC score) for transcription factors in 550 patient networks with corresponding transcription factor gene expression (log_2 CPM).*

3.5 Discussion

Here, the wealth of individualised CRC data produced by The Cancer Genome Atlas (TCGA), was combined with large scale interactome data to create 550 models of cellular signalling in the epidermal growth factor receptor (EGFR) network. I combined patient-specific transcriptomic and genomic data with protein-protein interactions (PPIs) to develop personalised network models for each tumour in each patient. I found that transcriptomic data was relatively easy to adapt to this purpose, both for weighting edges and removing nodes for genes with significantly reduced expression. I also used genomic data to incorporate PPI disrupting mutations, however ultimately found that relatively few of the mutations found in TCGA CRC patients matched the curated PPI disrupting mutation dataset (del-Toro *et al.,* 2019) or mutations that I was able to identify as potentially disruptive by searching for domains enriched in pairs of interacting proteins. Here I presented a node-removal strategy for network creation. However, improvements to the patient-specific tailoring of networks are certainly possible and may increase the biological relevance of results. This may also include incorporating information on alternatively spliced protein isoforms, which are known to occur frequently in TCGA tumour samples (Kahles *et al.,* 2018). This methodology was tested in CRC as a proof of concept. However, in theory, it should be possible to re-apply these methods to other types of cancer and potentially other diseases entirely.

I aimed to predict how these networks were rewired in different patients, and investigated the topology of these 550 personalised models for their prognostic relevance. There were significant differences in these properties between existing CRC subtypes, both in the Consensus Molecular Subtypes (CMS) (Guinney *et al.,* 2015) and PSDE-informed subtypes (PICs). No associations between survival and either global (network level) or local (node level) topological properties were found however, demonstrating that while topological properties were variable between patients they were not a reliable predictor of patient survival. This is in contrast to previous studies in which these properties were significantly associated with particular properties, namely network complexity (Breitkreutz *et al.,* 2012). This inability to reproduce

associations with patient survival previously reported in the literature could be due to differences in network construction, for example the networks of Breitkreutz *et al.* represented multiple different cancer types, and it is entirely possible that these findings simply did not translate well to networks representing individuals with the same cancer. Another possible explanation for these findings is that the updates PPI databases have received over the years since the original findings were made have significantly changed the structure of these networks. Until quite recently with the development of IMEx (Orchard *et al.,* 2014), curation of PPI data was mainly performed independently by multiple groups (Orchard *et al.,* 2012). This fact, combined with the increase in available interactions due to the development of high-throughput screening methods such as affinity purification mass spectrometry, mean that the size and topological properties of these networks have been significantly altered over the last decade.

I developed a novel tool, Simulated Information Flow For Individualised Networks (SIFFIN), which was used to simulate information flow through the 550 individualised CRC networks. In comparison to existing tools such as ITM probe (Stojmirović *et al.,* 2012), SIFFIN infers the directionality of edges in undirected networks using a process described by Silverbush & Sharan, 2019 and then simulates directed information flow with a simple network propagation algorithm (Cowen *et al.,* 2017). I also introduced a new metric for inter-network information flow comparison, scaled percentage change (SPC). I found in simulations of a small-scale EGFR network model (Samaga *et al.,* 2009) that SIFFIN compared favourably with ITM Probe for investigating the downstream consequences of node removals.

SIFFIN was applied to the 550 TCGA CRC networks, and predicted substantive rewiring of signal flow between patients. Stratifying patients on these differences, I found some concordance between previously established subtypes such as the CMS and patterns of information flow, but also found multiple "flow clusters" with heterogeneous composition of CMS classifications. Significant differences in survival were also identified for patients in specific flow clusters, beyond what was observed in the CMS. Given that these SIFFIN-derived clusters exhibited significant differences in patient survival, I expected to find that the signal flow to downstream transcription

factors (TFs) would vary substantially. Examining the predicted information flow to TFs, I found that SIFFIN predicted the preferential activation/repression of specific transcriptional programs in different patient groups. Cox regression analysis on TF $\log_2$ SPC scores revealed that increased flow to SMAD2 (a mediator of TGFβ signalling (Wee & Z. Wang, 2017)) was positively associated with survival. Using a random removal approach, I found evidence that these differences were not simply the result of random chance, as random removals resulted in very distinct downstream alterations to transcription factors. I found specifically that the predicted information flow to SMAD2 was significantly associated with better patient outcomes.

I found that as TFs were relatively infrequently classified as PSDE genes, it was difficult to associate PSDE genes with significant $\log_2$ SPC changes. However, by examining the correlation of the predicted information flow scores with gene expression data, I found that the both the target genes and genes coding for the transcription factors themselves were correlated to varying degrees with predicted information flow. Given that so many correlations between the TF expression and $\log_2$ SPC score were identified, it does suggest that this approach is excessively influenced by expression (as ideally upstream network topology, rather than individual node expression, should influence signal flow). A potential resolution to this issue could be to employ a rank-based edge weighting, as is found in tools such as Gene Set Variation Analysis (Hänzelmann *et al.,* 2013). However, the size of these correlations was overall quite small despite their statistical significance, which demonstrates that the variance in information flow is not entirely due to gene expression alone. This indicates that the information flow to transcription factors was not redundant, but related to gene expression. As both information flow scores and transcription factor activities are directly related to gene expression, it is likely that this correlation is representative of a meaningful biological signal. To provide further support for the hypothesis that information flow is predictive of the activation of transcription factors, one strategy that may be used of is the use of transcription factor binding site analysis to detect the enrichment of motifs among perturbed, which can help identify transcription factors likely responsible. In this instance, motif enrichment did corroborate the information flow results seen from SP1 and SMAD4, but was not statistically significant for other

tested transcription factors.

In conclusion, I found that using patient-specific data to predict differences in network topology can be used to identify novel patient groups, with significant differences in survival between groups defined from $\log_2$ SPC scores identified by SIFFIN not being observed in any previously defined CRC subtypes. This suggests that a network approach to modelling the heterogeneity of cancer may be able to provide unique insights that are overlooked by other approaches that do not consider network topology.

4. Spatially resolved exploration of intra-tumour heterogeneity

Contributions and acknowledgements

Many people contributed to the data presented in this chapter; I would like to acknowledge my principal supervisor Prof. David Lynn for assistance in development of InsituNet and the other methods presented, Dr. Stephen Blake for leading the mouse anti-CD40 experiments, Dr. Xiaoyan Qian, Dr. Thomas Hauling, and Prof. Mats Nilsson for providing *in situ* sequencing data, Mark Van der Hoek for assistance with RNA sequencing, Shadrack Mutuku and Prof. Lisa Butler for providing prostate cancer biopsy data, and Dr. Marten Snel, Dr. Paul Trim and Jacob Truong for mass spectrometry imaging. While work on InsituNet began the year prior to my PhD candidature, development continued during my PhD candidature and the software application was published in 2018 (Salamon *et al.,* 2018). All other work in this chapter was done by me during my PhD candidature.

4.1 Introduction

Tumours are heterogeneous not only between individuals but within a single patient, with different cells of a tumour potentially exhibiting distinct molecular characteristics (Dagogo-Jack & Shaw, 2018) and different tissue regions exhibiting varying levels of immune infiltration (Fridman *et al.,* 2011) and metabolic dysregulation (Vander Heiden & DeBerardinis, 2017). This cellular heterogeneity provides a population from which resistance to treatments may arise, and should therefore be considered in the development of any therapeutic approach. In previous chapters, I have focused on inter-patient heterogeneity primarily due to the limitations of available data, namely bulk RNA-sequencing data which is only capable of producing gene expression averaged across multiple homogenised cells. In contrast, single-cell RNA sequencing (scRNAseq) is capable of profiling the expression of individual cells, overcoming a major limitation of bulk RNA sequencing experiments. Despite this advantage, most applications of scRNAseq still lose the spatial context of the original tissue, making such approaches suboptimal for investigating tumour heterogeneity. In recent years however, multiple spatially-resolved technologies for multi-omics analyses on intact tissue sections have emerged which appear likely to greatly benefit the investigation of tumour heterogeneity.

Spatially resolved transcriptomics (or spatialomics) in particular is a rapidly expanding field, with many different approaches now available to spatially assess gene expression in tissue sections (Liao *et al.,* 2021). Popular approaches to spatially-resolved transcriptomics typically fall broadly under two categories; fluorescence in situ hybridisation (FISH)-like approaches which utilise hybridisation of fluorescent probes followed by optical imaging, and arrayed approaches which typically function more similarly to next-generation sequencing experiments but add spatial information using some form of 2D barcoded array. One hybridisation based method is *in situ* sequencing (ISS) (Ke *et al.,* 2013), an approach capable of detecting individual RNA fragments within sections of preserved tissue at micrometer resolution. *In situ* sequencing may be multiplexed to detect around 40 different types of transcript simultaneously, generating expression and spatial data. This method contrasts to

other approaches such as "spatial transcriptomics" (Ståhl *et al.,* 2016), which takes a whole-transcriptome sequencing approach using a specialised slide containing arrayed oligonucleotides with positional barcodes. This has the advantage of unbiased sampling of the transcriptome, but results in a lower resolution spatial map than approaches like ISS.

Spatially-aware metabolomics is now also possible using matrix-assisted laser desorption ionization (MALDI) mass spectrometry imaging (MALDI-MSI). MALDI-MSI has been applied to various solid tumours (Pirman *et al.,* 2013) enabling the investigation of the metabolomic and lipidomic tumour microenvironment in high spatial detail. MALDI-MSI is capable of imaging across an intact 2D section of tissue, outputting an image in which each pixel contains a record of mass-to-charge (m/z) intensity across the mass spectrum. This technique is capable of identifying various biomolecules, including lipids, and has been proposed as a clinical diagnostic tool for surgical pathology (Basu *et al.,* 2019).

The rise of spatialomics promises to enable studies of tumour heterogeneity and the tumour immune and metabolic microenvironments in far more detail than was previously possible. However, analysis approaches for this kind of spatial data are still in their infancy. Tool such as ST Pipeline (Navarro *et al.,* 2017) enable basic processing of spatial transcriptomics in a similar fashion to scRNAseq, but tools which are built specifically to take into account spatial location are sparse. Some promising approaches utilise image analysis and geospatial statistical methods such as point pattern analysis, for example STUtility (Bergenstråhle *et al.,* 2020). Highlighting the overlap between the spatial and single-cell fields, STUtility itself is based on the Seurat framework for spatial reconstruction of scRNA-seq data (Satija *et al.,* 2015). A common approach to spatial omics analysis when discrete pixels of data are available is to use unsupervised clustering algorithms such as Uniform Manifold Approximation and Projection (UMAP) (McInnes *et al.,* 2018) to classify each spatial point. Supervised approaches can include using partial least square discriminant analysis (PLS-DA), a machine learning approach which performs well with high-dimensional classification tasks, a property which makes it useful for exploratory analysis of metabolomics and mass spectrometry data (Brereton & Lloyd, 2014).

However, this kind of pixel classification approach is not applicable if discrete pixels do not exist, and in addition does not really utilise the spatial aspect of the data. As a fairly recent and underdeveloped field, there is a demand for new analysis methods. This demand could potentially open up a new space for network analysis, due to the capability of networks to integrate diverse data sources and the need to link spatial data to existing, better established technologies.

In this chapter I will discuss some of the software approaches I have developed to address the current gap in spatialomics analysis methods. This includes InsituNet, an application for network visualisation and analysis of spatially resolved transcriptomics data. I will also describe some case studies in which I have applied various methods to spatial data analysis to study topics including tumour heterogeneity and heterogeneity of the tumour immune and metabolic microenvironments on a spatial, intra-tumour level.

4.2 Hypothesis and Aims

I aimed to extend my work on inter-patient tumour heterogeneity to investigate intra-patient heterogeneity, using recent technologies to gain insight into the spatial dimension which is usually lost in conventional omics experiments. By building on simple network principles, I also aimed to build a tool which would be capable of performing integrative network analysis of spatially-resolved data, beginning with in situ sequencing (ISS), but also be capable of analysing other spatially-resolved data types including Spatial Transcriptomics. Furthermore, I aimed to extend this algorithm to encompass spatial lipidomics data collected from matrix-assisted laser desorption ionisation (MALDI) mass spectrometry imaging, in collaboration with colleagues using this approach to assess heterogeneity in prostate cancer.

4.3 Methods

4.3.1 A network-based approach to spatialomics with InsituNet

InsituNet is an application I developed for the investigation of spatially-resolved transcriptomics (Salamon *et al.,* 2018). InsituNet enables interactive exploration and analysis of spatially resolved transcriptomics data and was one of the first tools developed for network-based analysis of this kind of spatially resolved transcriptomics data. InsituNet was originally developed for the analysis of *in situ* sequencing (ISS) data (Ke *et al.,* 2013), a spatially-resolved transcriptomics method which produces extremely dense maps of gene expression, but with relatively few unique transcripts. I had several goals for producing a tool to visualise this type of data:

1. Determine a method of determining which pairs of transcript spatial co-localisations are unexpectedly more or less frequent than expected given the abundance of each transcript.

2. Provide a network-based visualisation of *in situ* sequencing data as a form of data dimensionality reduction, such that the most interesting relationships between transcripts may be identified.

3. Highlight network rewiring in different spatial tissue regions, so that the problem of comparing the transcriptional profile of different 2D areas can be solved.

I developed a simple algorithm to determine spatial co-localisation. Taking as input processed ISS data in which each transcript is represented as a point in 2-dimensional space, two transcripts a and b are defined as spatially co-localised if they are within a given euclidean distance $d = \sqrt{(b_x - a_x)^2 + (b_y - a_y)}$ of each other. This distance may vary based on which features a user is interested in (i.e. intra- or inter-cellular), and so InsituNet provides flexibility to the user to decide what this distance should be. Once spatial co-localisation is defined in this manner, a network is constructed in which each node is a transcript type, and each edge represents spatial co-localisation

between the two transcripts it links. This network can be visualised directly to show which pairs of transcripts are co-localised.

The unfiltered co-localisation networks will generally be too dense to be of practical benefit, given that there are potentially millions of individual co-localisation events and that in such dense data it would be normal to observe a large number of co-localisations simply by chance. To overcome this, the edges are filtered as the network is created in order to only show co-localisations that are statistically significant. Two methods are provided which assess this statistical significance, a label permutation based method and a hypergeometric approach.

Label permutation method

The label permutation method for assessing co-localisation statistical significance is a Monte Carlo method which randomly permutes the transcript labels, aiming to create a co-localisation frequency probability distribution which models how often each possible pair of co-localisations would be expected to occur by chance. The algorithm takes the following steps:

1. Randomly shuffle the labels of the transcripts (i.e., keep transcript locations fixed, only permute transcript names)
2. Find the new number of co-localisations for all transcripts. As node positions are unchanged these are already known, and may be recalculated easily.
3. Repeat steps 1 and 2 until a distribution of co-localisation for each transcript pair is created (e.g., 1000 times).
4. For each each pair of co-localisations which occur with a frequency greater than zero, find the probability of this observation given the generated distribution.

This probability is determined with a 95% confidence interval for each distribution. Co-localisations with frequencies outside of these intervals are considered to be statistically significant.

Hypergeometric method

As an alternative to the label permutation method which is less susceptible to over-representing lowly expressed transcripts, the hypergeometric distribution can also be used to assess the significance of drawing k co-localisations of transcripts a and b, given by:

$$P(X = x) = \frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}}$$

The parameters for this test are N, the total number of co-localisations between all transcripts, k, the total number of co-localisations between two transcripts a and b, n, the number of co-localisations involving any a transcripts, and K, the total number of co-localisations involving any b transcripts. In comparison to the label permutation method, the hypergeometric test fully considers the compound probability of obtaining a certain number fo co-localisations given the abundances of each of the transcripts in a pair.

Correction for multiple testing

Both the permutation and hypergeometric methods for assessing statistical significance require a large number of statistical tests (one for each unique co-localised pair of transcripts). Due to these simultaneous statistical inferences, InsituNet applies the Bonferroni method to adjust P values for multiple testing.

Extension to other data formats

While InsituNet was developed with a focus on analysis of spatially-resolved transcriptomics, specifically ISS, all the algorithm requires as input is a list of 2D points, and therefore it may also be extended to other forms of spatialomics data with relative ease. Many spatialomics data types, for example MALDI-MSI, do not resolve down to individual detections at different locations, but consist of higher dimensional data such as a mass spectrum which localises to a given region. InsituNet was limited in that only one data point can exist at a given 2D location, and so the import of other multi-dimensional data types was enabled by adding an optional input randomisation

procedure which would allow these data to be successfully imported. This allows such data to be analysed as long as the co-localisation distance is set higher than the randomisation distance.

Implementation details and availability

Cytoscape is an open source software platform for network visualisation and analysis with a biological focus (Shannon *et al.,* 2003). InsituNet is implemented as a Java OSGi bundle "app", compatible with Cytoscape version 3.2+ (Java 8+). InsituNet is available as a cross-platform JAR file which may be downloaded from within Cytoscape itself, or from the InsituNet page on the Cytoscape app store[1], from which links to the source code, test datasets, and user documentation can also be found.

Various software implementation details of InsituNet allow it to perform effectively instantaneously for moderately sized datasets (<100,000 points). InsituNet takes a file of comma separated values (csv) as input with three columns for transcript name, x coordinate and y coordinate. This list of points is converted into a kd-tree data structure which allows for much faster spatial operations, including detection of neighbours within a given euclidean distance of a point (time complexity of $O(n)$ in the worst case, compared to $O(n^2)$ for the naive approach of testing the distance between each point).

The tissue viewport in which each transcript is visible was OpenGL-accelerated to enable high performance for interactivity even when thousands of individual transcripts are present. This was implemented using the JOGL library[2].

The Monte Carlo label shuffle eventually converges on a general formula which models the expected co-localisations between a and b as $C(a,b) = \frac{2kn_an_b}{N^2-N}$ if transcript a is not the same as transcript b, or $C(a,b) = \frac{k(n_a^2-n_a)}{N^2-N}$ otherwise, where the total number of co-localisations is given by k, the total number of transcripts is N and n_a and n_b are the number of a and b transcripts respectively (see appendix for details). For efficiency this formula is used by the application rather than actually performing the permutation.

[1] https://apps.cytoscape.org/apps/insitunet

[2] https://jogamp.org

4.3.2 Using spatial transcriptomics to profile immune infiltration into MSS-CRC tumours following immunotherapy

Background

Immune agonist antibodies (IAAs) are a class of cancer immunotherapies which target co-stimulatory receptors on a range of immune cells, inducing immune cell infiltration into the tumour, thus increasing sensitivity to immune checkpoint inhibitors (ICIs) (Vonderheide, 2018; Mayes *et al.,* 2018). ICIs are co-inhibitory targeted immunomodulators which block immune checkpoints such as CTLA-4, PD1 and PDL1 to reverse the immune resistance gained by certain tumours, restoring the anti-tumour function of the immune system. First demonstrated in metastatic melanoma, such ICIs have been demonstrated to be clinically effective in other highly mutated cancers with mismatch DNA repair deficiencies, such as the microsatellite instability (MSI) subtype of colorectal cancer (Robert, 2020). As discussed in previous chapters, the MSI subtype of colorectal cancer (CRC) is more susceptible to ICI immunotherapy due to increased immune cell infiltration into the tumour. However, the majority of CRC tumours are microsatellite stable (MSS), and are therefore usually resistant to such therapies. Therapies which increase immune cell infiltration to sensitise tumours to ICIs could therefore be a viable therapeutic strategy.

Experimental design

To investigate whether treatment with the IAA anti-CD40 increases immune cell infiltration into immunologically cold MSI-CRC tumours, C57BL/6 mice were injected by Dr. Stephen Blake with tumours orthotopically in the colon and treated with anti-CD40 immunotherapy or PBS control once tumours were established. Tumours were a genetically engineered microsatellite stable (MSS) AKP ($Apc^{\Delta/\Delta}$, $Kras^{G12D}$, $Trp53^{\Delta/\Delta}$) organoid model developed by Dr. Susan Woods and Prof. Dan Worthley (SAHMRI). Following treatment tumours were sectioned and prepared for spatial gene expression profiling using the Visium 10X platform, which utilises a spatial transcriptomics method developed by Ståhl *et al.,* which is capable of whole transcriptome sequencing

with 100μm spatial resolution. Spatial transcriptomics was performed by staff at the SA Genomics Centre. I performed the analysis of the spatial transcriptomics data to assess changes in spatial gene expression in MSS-CRC tumours, following anti-CD40 treatment.

Spatial Transcriptomics

The Visium platform requires a spatially-barcoded slide with four main areas to which tissue can be fixed. In this experiment, each of the four areas on the Visium slide contained a separate tumour section from a different mouse, two treated with anti-CD40, plus two untreated controls. Tissues were H&E stained and then imaged. Paired-end sequencing was performed on the Illumina NextSeq 500 platform, generating Illumina base call files (BCLs). Once BCLs were available, I performed initial processing using the Space Ranger tool from 10X Genomics[3] (version 1.2.2), aligning reads to the mm10 (GENCODE vM23/Ensembl 98) mouse reference transcriptome. Using spaceranger mkfastq, the sequencing data were demultiplexed and converted to FASTQ files. With spaceranger count, automatic slide image alignment was performed, and unique molecular identifiers (UMIs) were counted. Finally with spaceranger aggr each individual slide was aggregated into a single normalised dataset. Tissue regions were manually selected with the 10X Loupe Browser, and the Space Ranger pipeline was run with these manual alignments. Read quality metrics including number of genes / counts per spot and mitochondrial content per spot were assessed.

Python scripts were written to automate the visualisation and analysis of the Visium data following processing with Space Ranger. Normalisation of the data to reduce the impact of spot-to-spot technical factors was performed using sctransform (Hafemeister & Satija, 2019). Visualisations of features (e.g. UMI counts per spot) superimposed on the tissue were generated using matplotlib.

Principal components analysis (PCA) from scikit-learn (Pedregosa *et al.,* 2011) (version 0.24.1) and uniform manifold approximation and projection (UMAP) (version 0.5.1) (McInnes *et al.,* 2018) were used for visualisation to identify batch effects

[3] https://support.10xgenomics.com/spatial-gene-expression/software

between the four tissue sections. A batch effect between the four tissues was identified and subsequently adjusted for using the harmonypy version of Harmony (Korsunsky *et al.,* 2019). Harmony requires PCA-transformed counts, so clustering on the batch-adjusted counts was performed by running UMAP on PCA components then clustering the resulting embedding with k-means clustering. Differential gene expression between clusters and treatments was determined using edgeR (version 3.28), using the raw counts prior to processing with Harmony, but with a design matrix to adjust for the tissue batch effect. Pathway enrichment analysis was performed using an over-representation approach as implemented in my biomodule library (available on the Lynn lab Bitbucket repository[4]). The top 400 most significant genes by FDR from differential gene expression were used as a query set for pathway over-representation, using pathway databases including Gene Ontology Biological Process (Ashburner *et al.,* 2000) and KEGG (Kanehisa & Goto, 2000).

4.3.3 Effects of immunotherapies in the liver

Background

IAAs have unfortunately not found widespread clinical translation due to immune-mediated side effects including potentially fatal liver damage and cytokine release syndrome (CRS) (Mayes *et al.,* 2018). Recently, studies have found that the gut microbiota play a very critical role in the efficacy of ICIs (Routy *et al.,* 2018). Given that the liver is a common site of IAA toxicity, it was hypothesised by Blake et al. (Blake *et al.,* 2021) that the gut microbiota may also be critical for mediating the immunotoxicity of IAAs, specifically anti-CD40.

To test whether the gut microbiota mediate the immunotoxicity of IAAs, tumour-bearing mice were untreated or treated with antibiotics (ampicillin and neomycin) during anti-CD40 therapy. To demonstrate that the effects were mediated by the microbiota, responses to anti-CD40 in germ-free mice were also assessed. To investigate the mechanism by which the microbiota altered lipid metabolism in the liver, MALDI-MSI was employed to provide a spatially-resolved profile of the lipidome in antibiotic

[4] https://bitbucket.org/lynnlab

treated, germ-free and control mice. I developed a tool, MSpecView, for identifying specific lipids which were enriched within certain tissues and visualising the resulting spatial MSI datasets.

Experimental details

MC38 tumour cell lines donated by Dr. Susan Woods (derived from C57BL/6 murine colon adenocarcinoma cells) were injected subcutaneously into the flank of mice. Tumour-bearing mice were untreated or were treated with the antibiotics ampicillin (0.5mg/ml) and neomycin (1mg/ml) dissolved in sterile drinking water for the duration of the experiment. Water was replaced three times weekly. I.P. injection of anti-CD4- or PBS control occurred every 4 days, starting at 3 weeks from the initial antibiotics administration. Liver sections were collected and snap frozen at -80°C. Tissues were placed inside a Shandon Cryotome E at -20°C for 30 minutes prior to sectioning. Tissues were mounted with O.C.T and 10 μm thick sections were cut for MALDI MSI. Mass spectrometry imaging was performed by Dr. Paul Trim. Sections were imaged using a timsTOF FleX mass spectrometer operated in negative ion mode. Imaging data was imported into SCiLS Lab and spectra were aligned to local maxima before exporting the data as csv.

Data analysis

I wrote custom Python scripts, available on the Lynn lab Bitbucket repository, to process and clean mass spectra csv files as exported from SCiLS Lab. Peaks with low intensity in all samples (< 100 AU) were excluded from analysis. Principal components analysis (PCA) was used as an exploratory tool to quantify the variance and heterogeneity present across each tissue section. Mean intensity of each section was quantified for each retained m/z peak, and fold change was calculated between each group. Peaks exhibiting a fold change >2 between untreated / antibiotics treated groups or untreated / germ free mice were selected for further analysis. 2D visualization of selected mass peaks in tissue sections was performed using matplotlib 3.4.1.

I also developed an interactive tool I named MSpecView to visualise the three treatment groups which rapidly produced visualisations of specific masses which were up or down-regulated. The code for this tool is available from the Lynn lab Bitbucket. The tool enabled rapid exploration of the mass spectrometry data by first compiling all data into a single HDF5 (hierarchical data format) file which could be explored visually with an interface developed on top of the pyqtgraph[5] library (version 0.11.1). Potential lipid matches to m/z peaks were annotated by searching the Lipid Maps Structure Database (Sud *et al.,* 2007) by mass for negative ions, with +/- 0.5 tolerance.

4.3.4 Lipid composition analysis of prostate tumours

Background

In prostate cancer (PCa), the androgen receptor plays a pivotal role, including driving treatment resistance and dysregulation of lipid metabolism (Zadra *et al.,* 2019). The highly heterogeneous nature of PCa means that homogenising tissues for analysis inevitably mixes multiple cell types in unknown ratios, causing difficulties in analysis and entirely losing information on lipid composition in the tumour microenvironment. Using matrix-assisted laser desorption ionization (MALDI) mass spectrometry imaging (MALDI-MSI), it is possible to retain this spatial information.

Experimental design

MALDI-MSI was applied to assess the lipidome of prostate biopsies collected from 10 patients between the ages of 58 to 70 (Mutuku et al, under review). This study aimed to characterise the lipidome of different tissue types present within these tumours. I constructed a partial least squares discriminant analysis (PLS-DA) machine learning model to integrate lipidomic mass features to test how effectively the detected lipid signatures were for the classification of different tissue types.

[5] https://www.pyqtgraph.org/

PLS-DA model construction and validation

A PLS-DA model was constructed using the Python-based Scikit-learn package (version 0.24.1) (Pedregosa *et al.,* 2011). PCa sections were divided into multiple "spots", each of which consisted of a dataset of 132 m/z intensities. These spots were labelled as tumour, benign and stroma based on pathology annotations and used to train the PLS-DA model. Cross-validation of the model was done using the K-Fold method with 10 partitions (i.e., 10-fold cross validation). Data were split into 10 partitions at random, and the training repeated 10 times, with each repeat leaving out a different partition to be used as a validation. The performance af the model at classifying the validation partitions was evaluated using area under the receiver operating characteristic (AUROC). Due to the small number of independent PCa sections, during each iteration of training at least two of the PCa sections were entirely removed from the training partition to minimise over-fitting.

Biopsy collection and imaging

Human prostate tissue was obtained with written informed consent from participants under the South Australian Prostate Cancer BioResource collection protocol. Tissue sections from 10 patients were collected for haematoxylin and eosin (H&E) staining and MALDI-MSI. MSI was performed by Dr. Paul Trim and Shadrack Mutuku on a Waters SYNAPT HDMS hybrid quadrupole orthogonal acceleration Time-of-Flight Mass Spectrometer (Q-oa-TOF), recorded over a 400-990 m/z range.

Data analysis

Raw data were converted to MSI data files using Waters high definition imaging (HDI) software (version 1.4). Initial processing of these files was performed by Shadrack Mutuku using SCiLS Lab MVS Pro (Bruker Daltoniks GmBH, Germany), in which the identified m/z intervals were aligned to local maxima, then exported as csv files. SCiLS Lab was also used to perform segmentation of the data using k-means clustering. Three distinct m/z segments were identified and annotated as tumour, benign, and

stroma. I constructed a supervised partial least squares discriminant analysis (PLS-DA) model as previously described which included 132 m/z features using MALDI-MSI data exported from SCiLS Lab. The model was trained on regions of tissue annotated as tumour, benign and stroma. Masses determined to contribute to matrix noise were excluded, which were identified as those that were highly over-represented on the outer edges of the tissue sections. The PLS-DA model was constructed using the Python-based Scikit-learn package (version 0.24.1). 10-fold cross-validation was used to assess the model performance using AUROC. During each test/train split, at least two of the PCa sections were entirely different between test and training groups to minimise over-fitting. Masses most strongly associated with a given feature were identified and exported. Classifications of different regions predicted by the PLS-DA model were visualised on images of the original tissue using matplotlib 3.4.1.

4.4 Results

4.4.1 InsituNet

To demonstrate InsituNet, the application was used to analyse a published *in situ* sequencing dataset profiling spatial gene expression in a HER2 transcript–positive human fresh-frozen breast cancer tissue section (Ke *et al.,* 2013). This dataset consisted of 17,722 individual transcript detections for 38 different transcripts, including 21 transcripts used in the OncoType DX breast cancer prognostic expression panel (Sparano & Paik, 2008). InsituNet imports in situ data by parsing an input file of comma delimited values specifying individual transcript names and two-dimensional coordinates. These data are then used to construct a space-partitioning kd-tree to enable rapid range searching. The option of importing corresponding histological images of the tissue section, such as hematoxylin-and-eosin (H&E) stains, is also available to assist in navigating the tissue section (Figure 4.3). After importing data, the user is presented with a visualisation showing the position of each transcript detected in the tissue section (Figure 4.1, A). A unique color/symbol combination is assigned to each uniquely-named transcript. This visualisation is presented within an OpenGL-accelerated window, enabling efficient rendering of potentially millions of transcripts. With a relatively small dataset (130,000 transcript detections, 26 unique types) import and visualisation take around a second on a modern laptop (Intel Core i7 6600U processor, 16GB RAM). InsituNet's efficient performance was confirmed using a larger unpublished in situ sequencing dataset consisting of ∼1.5 million transcript detections, which takes around 17 seconds to import. Transcript spatial expression can be interactively explored using the application to pan and zoom into specific areas of interest. For example, by exploring an example breast cancer tissue section, one can immediately appreciate the heterogeneity in gene expression in different parts of the section, and readily identify distinct regions of the tissue (Figure 4.1, A). However, due to the density of *in situ* sequencing data it is difficult to easily identify which specific transcripts are more expressed in which regions of the tissue section and it is impossible to determine which transcripts are unexpectedly spatially co-expressed.

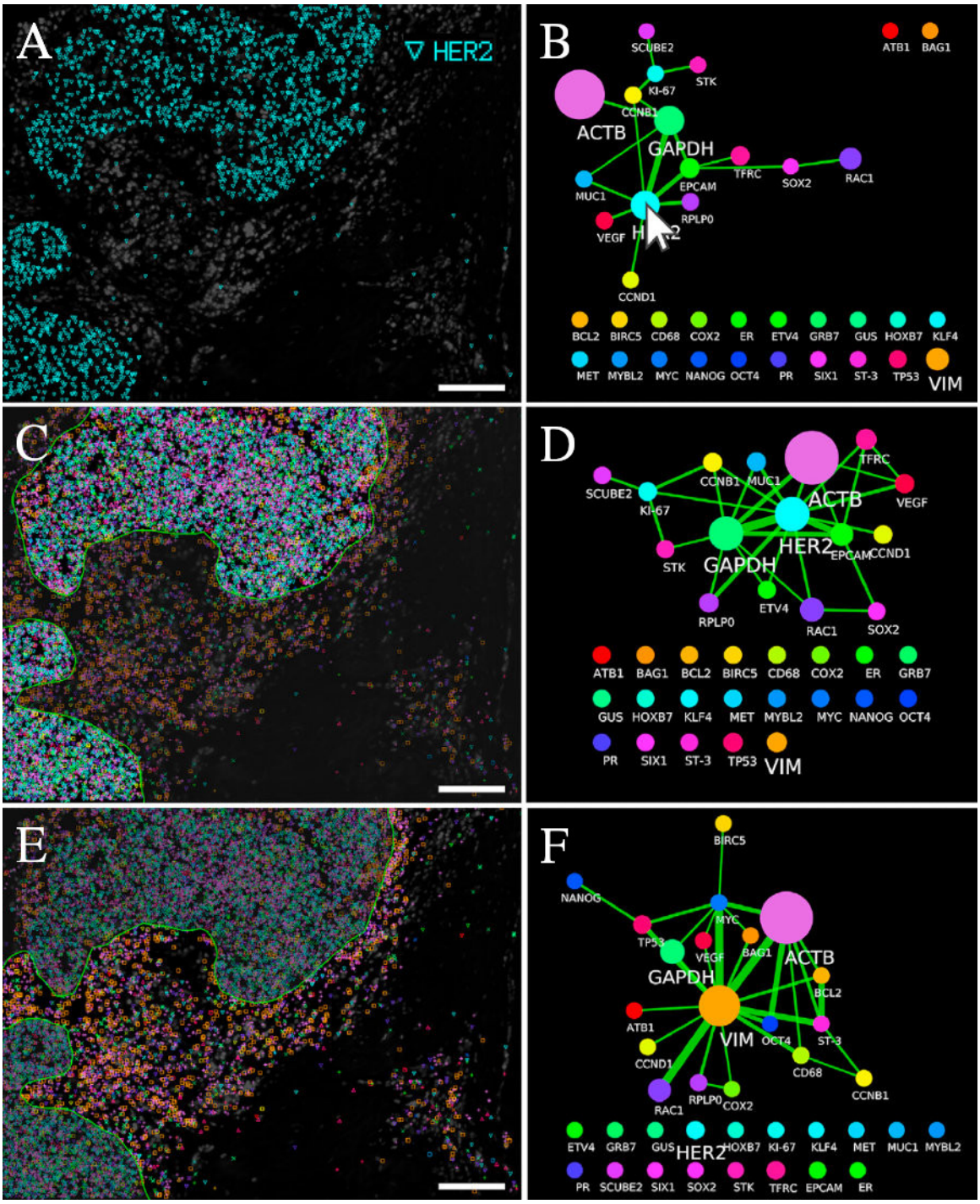


Figure 4.1: *InsituNet analysis of HER2+ breast cancer dataset from Ke* et al., *2013. Networks (B, D and F) were generated from the corresponding tissue regions (A, C and E) (Search distance: 40px/6.6µm). (A and B) HER2 expression in the tissue section was highlighted by selecting the HER2 node in the network. (C and D) Polygonal search region employed to select the HER2+ region. (E and F) The (HER2-) region was selected for comparison. Scale bars: 100µm.*

To address these challenges, InsituNet automatically generates a network-based visualisation of the in situ sequencing data (Figure 4.1, B), which is presented in a new window alongside the initial visualisation of transcript localizations. In the network visualisation, nodes represent each uniquely-named transcript, with the node size proportional to the number of detections of the transcript in the tissue section (or selected region). This approach substantially reduces the dimensionality of the data as potentially millions of transcript detections are represented as a correspondingly much smaller set of nodes in the network, representing each unique transcript. Nodes are coloured concordantly with the colours used to visualise the individual transcript detections in the initial in situ-seq visualisation. Selecting a node(s) will highlight the expression pattern of the corresponding transcript(s) in the in situ sequencing data visualisation panel, allowing users to interactively explore the expression data via the network visualisation. For example, selecting the HER2 node in the example network highlights the HER2 transcript detections in the tissue section (all other transcripts are hidden) (Figure 4.1, A-B).

A key goal of in situ sequencing is to identify transcripts that are preferentially spatially co-expressed in regions of interest in a tissue section and to identify how those co-expression profiles are altered, for example in regions of pathology. Transcripts that are identified as significantly spatially co-expressed in InsituNet are linked by an edge in the network visualisation. The more statistically significant the co-expression is, the greater the weight (thickness) of the edge in the network. To identify spatially co-expressed transcripts, InsituNet analyses the co-occurrence of transcript detections within a user- defined Euclidean distance, for each pairwise combination of transcripts. Transcripts that are co- expressed more than statistically expected are then identified using either a label permutation-based approach or a hypergeometric test. The user can choose which. In both approaches, InsituNet considers the overall abundance of the transcripts (i.e. the number of individual times each transcript is detected) in the selected region and the number of times that the two transcripts spatially co-occur within the defined distance, in the wider context of all co-expressed transcripts. For example, in the example breast cancer tissue section, housekeeping genes, such as beta actin (ACTB) or GAPDH, are expressed throughout the tissue and are represented as

large nodes in the network (since they are highly abundant). However, these nodes still have relatively few edges since they are not surprisingly spatially co- expressed with many other transcripts (given that they are expressed throughout the tissue section) (Figure 4.1, C). InsituNet, also has the option to identify nodes that are co-expressed with other transcripts significantly less than would be expected given their abundance. This identifies transcripts that tend not co-occur together in the same vicinity and these transcripts may represent specific biomarkers of different cell-types or regions of the tissue section.

Network-based visualisations can be generated based on the expression profile of the entire tissue section or on a subsection of the tissue selected by the user, enabling the comparison of expression patterns in different parts of the tissue section (Figure 4.1, E-H). For example, H&E staining of this section (Figure 4.3) shows that the region selected in Figure 4.1, C is the cancerous tissue. This is also evident from the InsituNet analysis as HER2 expression is restricted to these regions (Figure 4.1, D). The ability of the selection tool to select irregularly-shaped regions of the tissue section enables the user to precisely define regions of interest. InsituNet identified a surprisingly strong association ($-\log_{10}(P) > 10$) in this region between the spatial expression of HER2 and GAPDH. This association between GAPDH and HER2 expression is stronger if the HER2+ region is selected in InsituNet (note the increased edge thickness in Figure 4.1, D compared to Figure 4.1, B) and is not evident at all in the HER2- region of the section (Figure 4.1, E-F). Interestingly, GAPDH, although widely used as a reference gene, has been shown to be correlated with ER expression and associated with breast cancer cell proliferation and with the aggressiveness of tumours (Révillion et al., 2000). In contrast, one can observe that vimentin (VIM) expression, a marker of mesenchymally-derived cells, is higher in the non-tumour region of the tissue (based on the size of the node in Figure 4.1, F compared to Figure 4.1, D). One can also observe associations between VIM expression and other transcripts that are not evident outside of this region of the tissue section.

Where regions of interest are not immediately evident, InsituNet also implements an automated sliding window function (the size of the window can be defined by the user). This function enables one to quickly compare expression profiles across

the tissue section by generating network-based visualisations of transcript expression in each window (Figure 4.4). InsituNet also enables network-based visualisations to be generated for different tissue sections, for example, tissue sections from different patients. Where multiple networks are constructed (either from different tissue sections or from different regions within one section), InsituNet spatially synchronises their layout to facilitate comparison. This synchronisation is achieved in a manner similar to DyNet (Goenawan *et al.*, 2016) and any of the layout algorithms available in Cytoscape can be applied as desired by the user. InsituNet also facilitates the management of multiple networks from a unified interface which tracks all networks made using the application. This interface also allows the user to switch quickly between different networks.

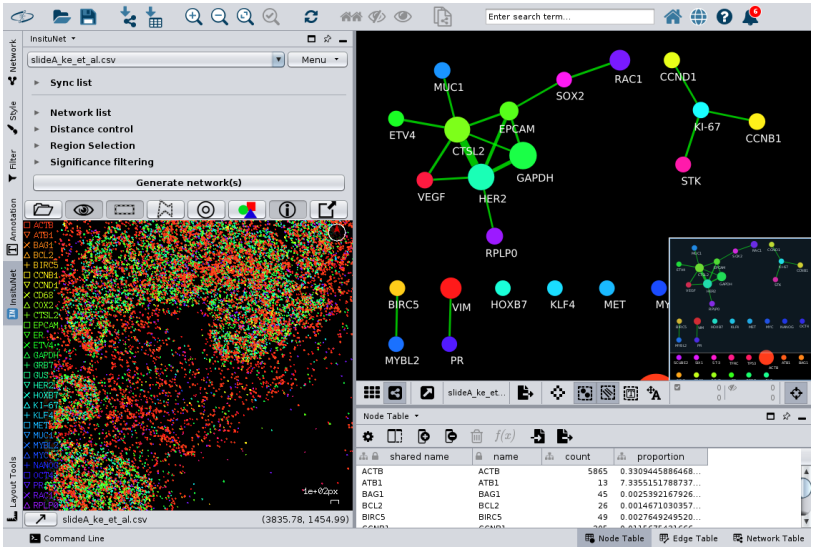


Figure 4.2: *InsituNet with the HER2+ breast cancer dataset from Ke* et al., *2013 imported (30px/5µm search distance, label shuffle significance filtering). The tissue view (left) is an interactive visualisation of the tissue which represents transcripts as coloured symbols. 2D regions can be selected from which co-localisation networks will be constructed. The network view (right) displays a network representation of the significant co-localisations between transcripts. Node size is proportional to abundance within the selected region, and edge width is proportional to significance determined by label permutation or hypergeometric methods.*

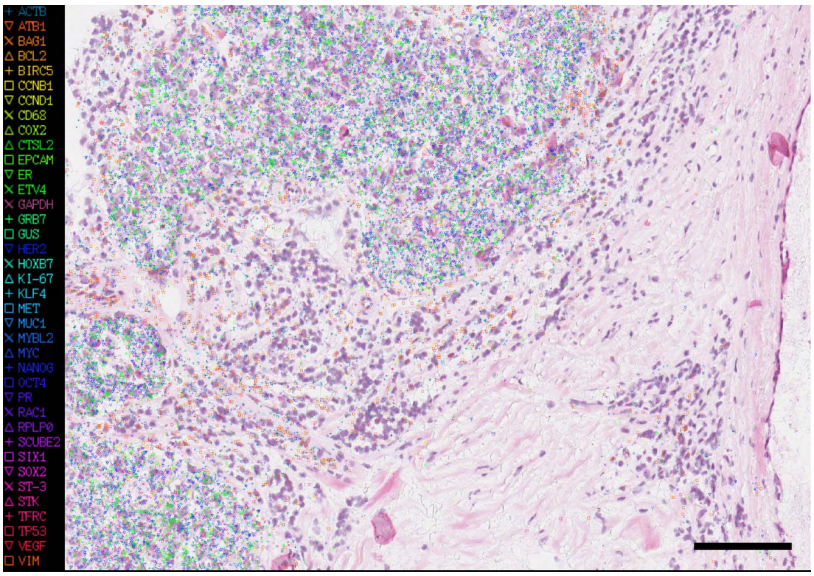


Figure 4.3: *H&E stain of the tissue section shown in Figure 4.2 with in situ sequencing data overlaid using InsituNet.*

InsituNet was developed as an app for the Cytoscape platform, and may be downloaded from the Cytoscape app store[6]. The datasets presented here, including the Ke *et al.* dataset, are available from InsituNet's source code repository[7]. InsituNet's algorithm requires only 2D spatial coordinates. Given this, it is possible to analyse any form of 2D data in this way. Spatially-resolved transcriptomics such as FISSEQ (J. H. Lee *et al.*, 2015) can also be analysed using InsituNet without any modification, as the data already exists as discrete points. Other methods which lack specific transcript coordinates such as Spatial Transcriptomics (Ståhl *et al.*, 2016) instead localise many transcripts within a given array diameter, and so require some reprocessing before they can be analysed. To enable analysis of such data I added functionality to InsituNet which adds a small amount of random noise to the positions of transcripts within each array location, enabling import and analysis. In this manner, InsituNet is capable of performing spatial network analysis on any spatially resolved transcriptomics, and to my knowledge is one of the first applications developed for this purpose.

[6] https://apps.cytoscape.org/apps/insitunet

[7] https://bitbucket.org/lynnlab/insitunet/src/master/datasets

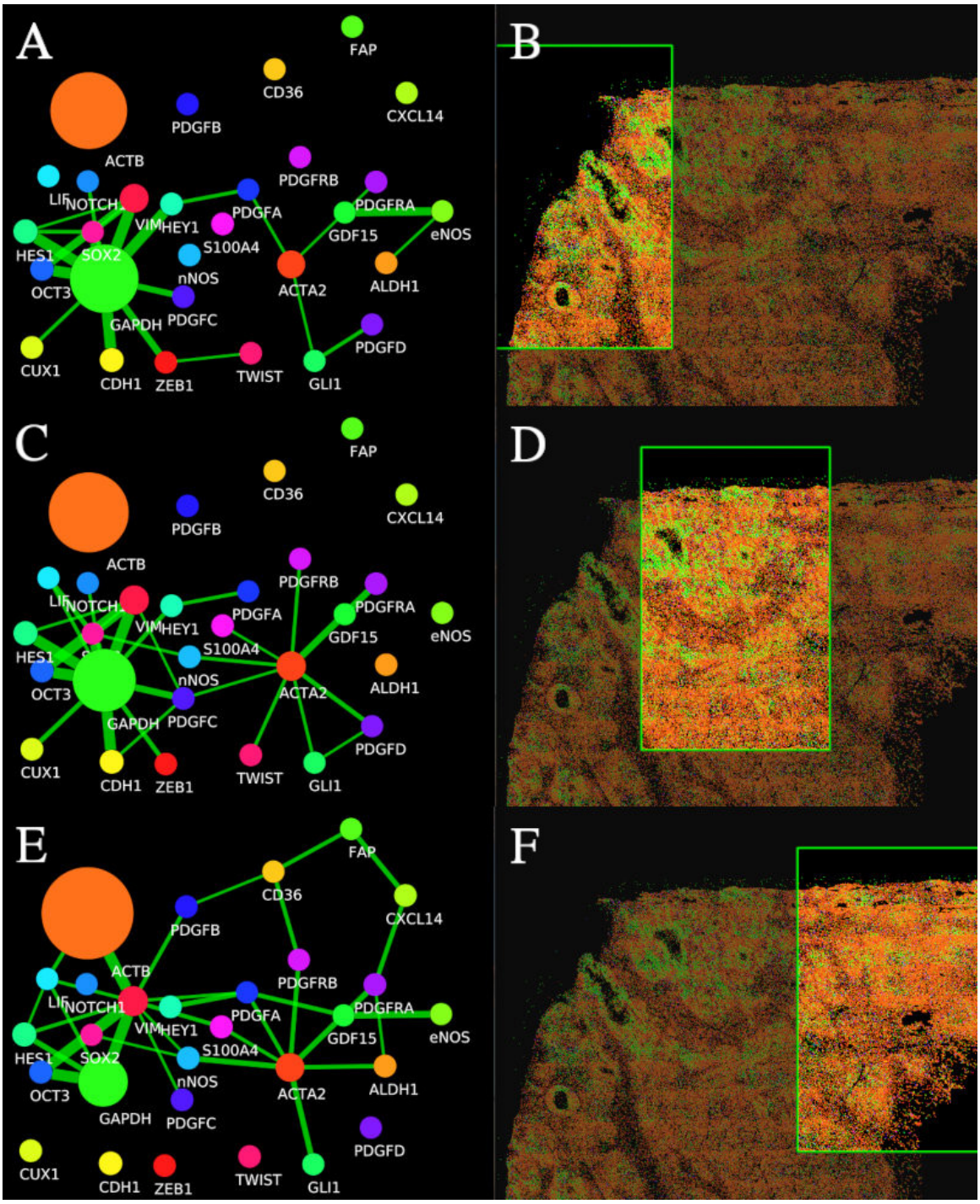


Figure 4.4: *The sliding window function of InsituNet allows users to compare spatial co-expression across the tissue. Three rectangular regions shown represent the sliding window regions of the tissue analysed. Search distance: 30px/5µm). Automatically generated networks are shown on the left (A, C and E) for regions shown on the right (B, D and F). Network node layouts are synchronised, keeping same positions across different networks.*

4.4.2 Profiling immune infiltration of CRC with Spatial Transcriptomics

Immune agonist antibodies (IAAs) target co-stimulatory receptors on a range of immune cells and by inducing immune cell infiltration, increase sensitivity to immune checkpoint inhibitors (ICIs) (Vonderheide, 2018; Mayes *et al.,* 2018). ICIs are clinically effective in the microsatellite instability (MSI) subtype of colorectal cancer (Robert, 2020), however, the majority of CRC tumours are microsatellite stable (MSS). To investigate whether treatment with the IAA anti-CD40 increases immune cell infiltration into immunologically cold MSI-CRC tumours, which would sensitise them to ICIs, a genetically engineered microsatellite stable (MSS) tumour organoid model was injected into mice. Tumour sections from four of these mice were analysed using the spatial transcriptomics method described by Ståhl *et al.* (Ståhl *et al.,* 2016), two of which had been sectioned from mice treated with the immune agonist antibody (IAA) anti-CD40, and two untreated. Approximately 100 million reads were sequenced per tissue section.

Each spot on a Spatial Transcriptomics dataset is an individual scRNAseq dataset, albeit from a single spatial region 55 μm in diameter rather than an actual single cell. Similarly to scRNAseq data however, this generally results in a lower signal-to-noise ratio than bulk RNA-seq, leading to a high abundance of zeroes in the final count tables. Furthermore, the low numbers of transcripts obtained from single cells mean that library preparation requires cDNA amplification, causing amplification bias. This bias is mitigated by use of Unique Molecular Indicators (UMIs), which barcode each transcript prior to amplification. Visualising UMIs superimposed on the tissue images, it was possible to verify that there was not an unusual localisation of UMI counts (Figure 4.5), suggesting that the procedure had been successful. Read quality for spots appeared good, with less than 4% mitochondrial content in all spots.

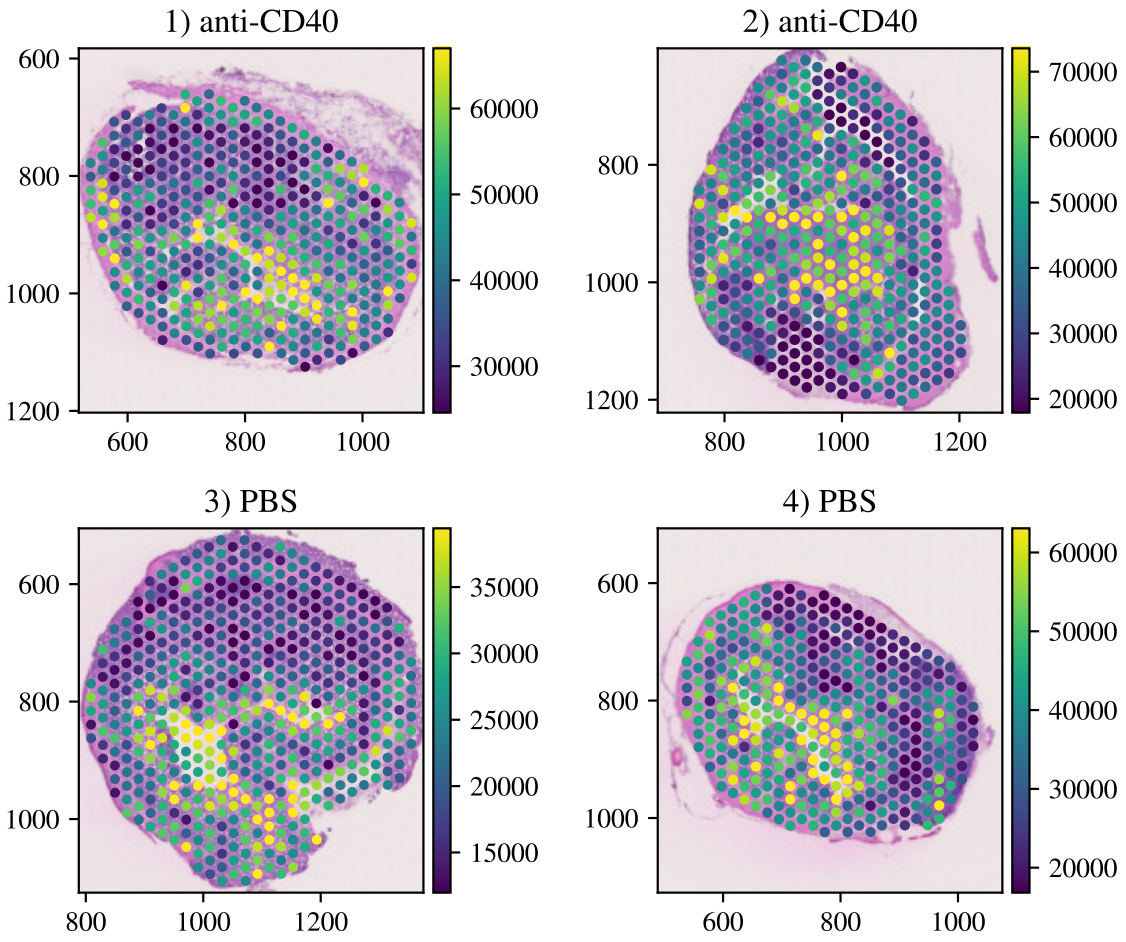


Figure 4.5: *Spatially-localised UMI counts for each tissue section overlaid on H&E stained tissue section. Data were visualised with matplotlib 3.4.1. The intensity colour scale corresponds to absolute UMI counts per 55 µm spatial transcriptomics spot. Each spot represents a spatially localised whole-transcriptome RNA sequencing dataset.*

Following aggregation of the four datasets, data dimensionality reduction for visualisation and clustering was performed using both linear (PCA) and non-linear (UMAP) methods followed by k-means clustering. This analysis revealed that each tissue section clustered almost entirely separately from the others (Figure 4.6, 1). This was undesirable for clustering purposes, as similar areas on each tissue section would not be possible to identify, and was clearly a batch effect. Following adjustment with Harmony, a tool for the integration of single-cell data (Korsunsky *et al.*, 2019), the variation between each tissue section was adjusted for (Figure 4.6, 2 and 3).

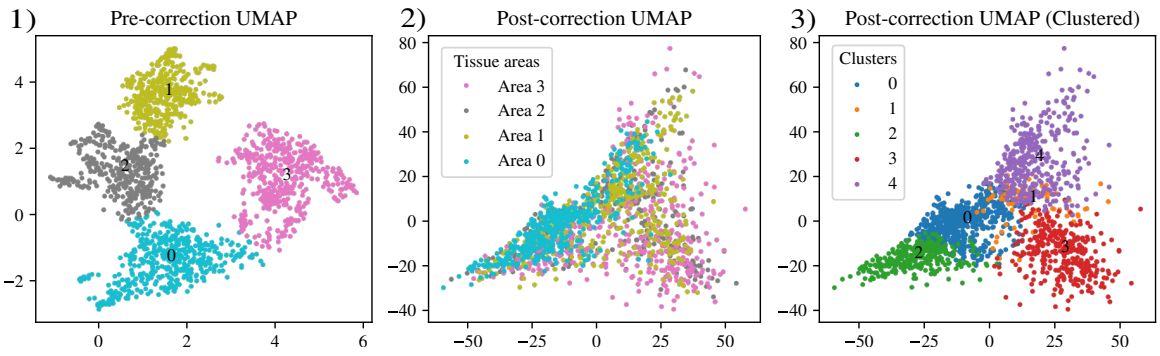


Figure 4.6: *Aggregated data were clustered using k-means clustering. Before adjusting for between-tissue batch effects, each tissue section was clearly identifiable as a separate cluster by this process (1). Following adjustment with Harmony, the previously separated tissue areas were integrated (2), and each cluster detected by k-means clustering was present across multiple tissue sections (3).*

Following batch effect adjustment, clustering of the spatial transcriptomics data was able to identify similar regions of tissue across the different tissue sections (Figure 4.7). Differential gene expression analysis was used to compare the two treatment groups, and also to identify genes specific to the clusters derived from unsupervised k-means clustering. This analysis revealed a large number of immune-related genes which were significantly up-regulated in the two anti-CD40 treated tissue sections (Figure 4.8). Pathway analysis confirmed the enrichment of these genes for roles in the immune system (Figure 4.9).

Five clusters were identified which were present in varying ratios across the four tissue samples which were found to be representative of particular cell and tissue types following pathway enrichment using KEGG and GO Biological Process. Notably, areas of immune infiltration were identified in both cluster 3 (Figure 4.9) (enriched primarily for adaptive immune responses, e.g. T and B cell pathways) and cluster 1 (enriched more for innate responses, e.g. neutrophil infiltration). Cluster 3 for the most part was prevalent within the anti-CD40 treated tissues, and especially within the treated area tissue of area 1 (Figure 4.7, 1). Pathology annotation also indicated that cluster 3 correlated well with areas of apparent immune infiltration. Other clusters were strongly representative of different tissue features; cluster 4 was matched to the muscle tissue, and epithelium for cluster 0. As these were colon sections, the centre

of the tissue was primarily lumen, which was consistently identified as cluster 2. The strong immune signature was present in cluster 3,including T and B cell pathways, was also significantly up-regulated in the anti-CD40 treated tissues, indicating that the IAA treatment had driven increased infiltration of immune cells into the tumour, potentially sensitising it to subsequent treatment with ICIs.

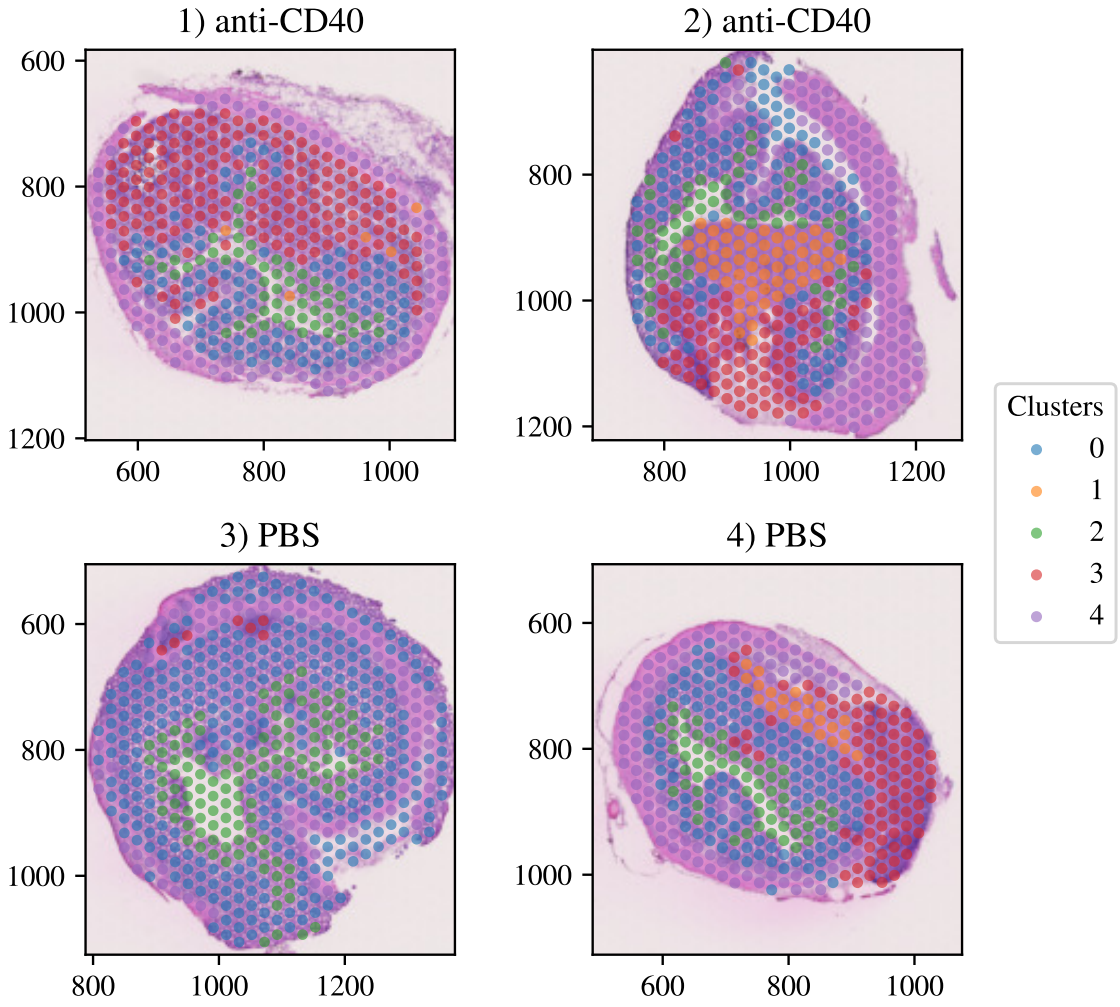


Figure 4.7: *Visualisation of spatial transcriptomics data on four mouse MSS-CRC tumour tissue sections, two anti-CD40 treated (1 and 2), and two untreated (PBS injected, 3 and 4). 5 clusters were identified in varying ratios across each tissue section. Clusters were defined based on gene expression data using k-means clustering. Clusters 3 and 1 were both enriched for roles in the immune system.*

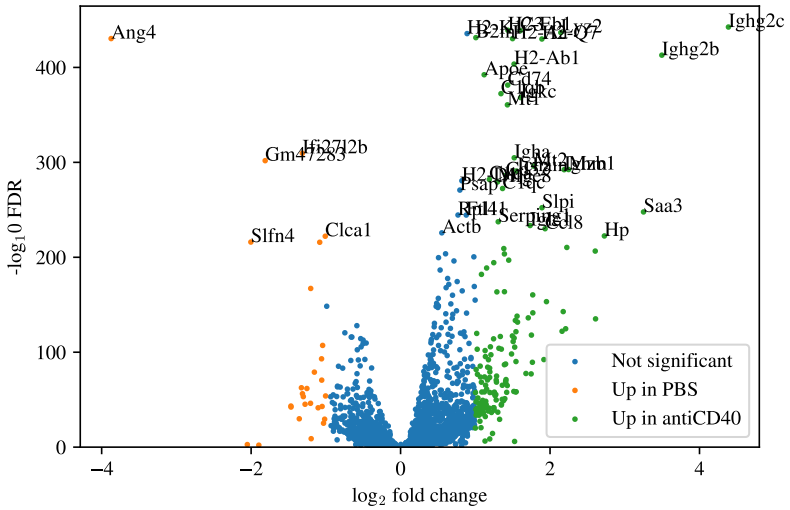


Figure 4.8: *Volcano plot of differentially expressed (DE) genes identified between anti-CD40 treated and control (PBS treated) tumours. The most significant genes by FDR are annotated. DE genes were identified with edgeR (3.28), and visualised with matplotlib (3.4.1).*

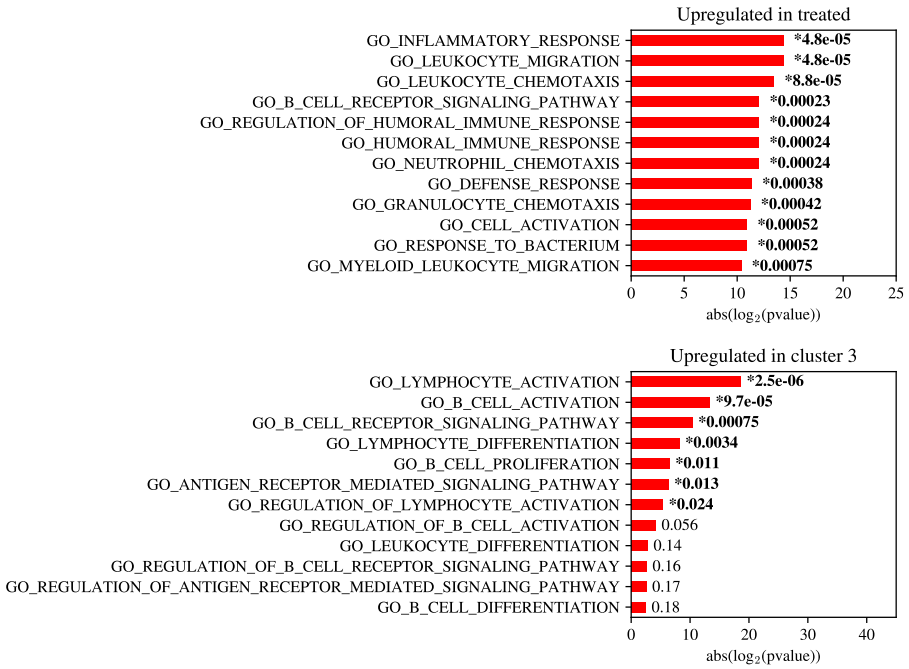


Figure 4.9: *Up-regulated Gene Ontology Biological Process terms in anti-CD40 treated tumour tissues (top) and in cluster 3 (bottom). Pathway enrichment analysis was performed using the top 400 most significant differentially expressed genes for each group.*

To identify spatial modules of gene expression, InsituNet was used to analyse the spatial transcriptomics data from anti-CD40 and control treated tumours. As spatial transcriptomics datasets are extremely dense (whole-transcriptome), a reduction of features was required to facilitate interpretation of the resulting networks. Using PCA, gene expression was reduced into its most variant components and split into up or down-regulated for each component. This resulted in a highly reduced, but difficult to interpret expression map, as each component represented the variation of many genes. To overcome this difficulty while retaining the most spatially variant genes, the top 10 genes comprising each principal component were identified and used directly as input to InsituNet. Each gene was split into lower and higher expression before import based on the distribution of counts for that gene, such that each gene would be represented by two nodes in the network. Networks were created for each tissue section, and their layouts synchronised to facilitate comparison. The networks for anti-CD40 treated tissues were of particular interest, as they formed several treatment-specific modules (Figure 4.10).

Certain network modules were present across treatments, in particular, a module of genes expressed in the smooth muscle (as identified by pathology annotation) (Figure 4.10 A) corresponding strongly with cluster 4 as identified by k-means clustering (Figure 4.7). Interestingly, while this network module was present in both treated and untreated tissues, collagen type 1 α-1 (*COL1A1*) and collagen type 1 α-2 (*COL1A2*) genes were strongly linked with all other members of the module only in the treated tissue. These collagen genes were expressed across all sections, but localised to muscle tissue only in anti-CD40 treated sections. This was potentially the result of anti-CD40 treatment suppressing the expression of these collagen genes. Inhibition of the CD40-CD154 blockade has previously been demonstrated to suppress expression of *COL1A1* in dermal tissues (Kawai *et al.,* 2008). One network module which was only present in treated tissues (Figure 4.10 B) was localised strongly to regions overlapping with cluster 3 (Figure 4.7). This module included high expression of *CCR7*, *IGHM*, (genes commonly expressed by B and T lymphocytes) and *LYZ2* (commonly expressed by macrophages). Perhaps more interesting were the down-regulated genes in this module, which included *ZG16*, *SPINK4*, *MPTX1* and *IGLC1*. Decreased expression of *SPINK4*,

a gastrointestinal peptide, has been shown to be associated with poorer prognosis in CRC patients, as well as being related to higher TNM stage in TCGA patients (X. Wang *et al.*, 2019), with benign tissues exhibiting higher levels of *SPINK4*. *ZG16* has also been shown to be significantly down-regulated in CRC, and is especially correlated with microsatellite instability (Meng *et al.*, 2018), indicating that the anti-CD40 treatment appeared to have initiated molecular changes more consistent with the MSI subtype of CRC.

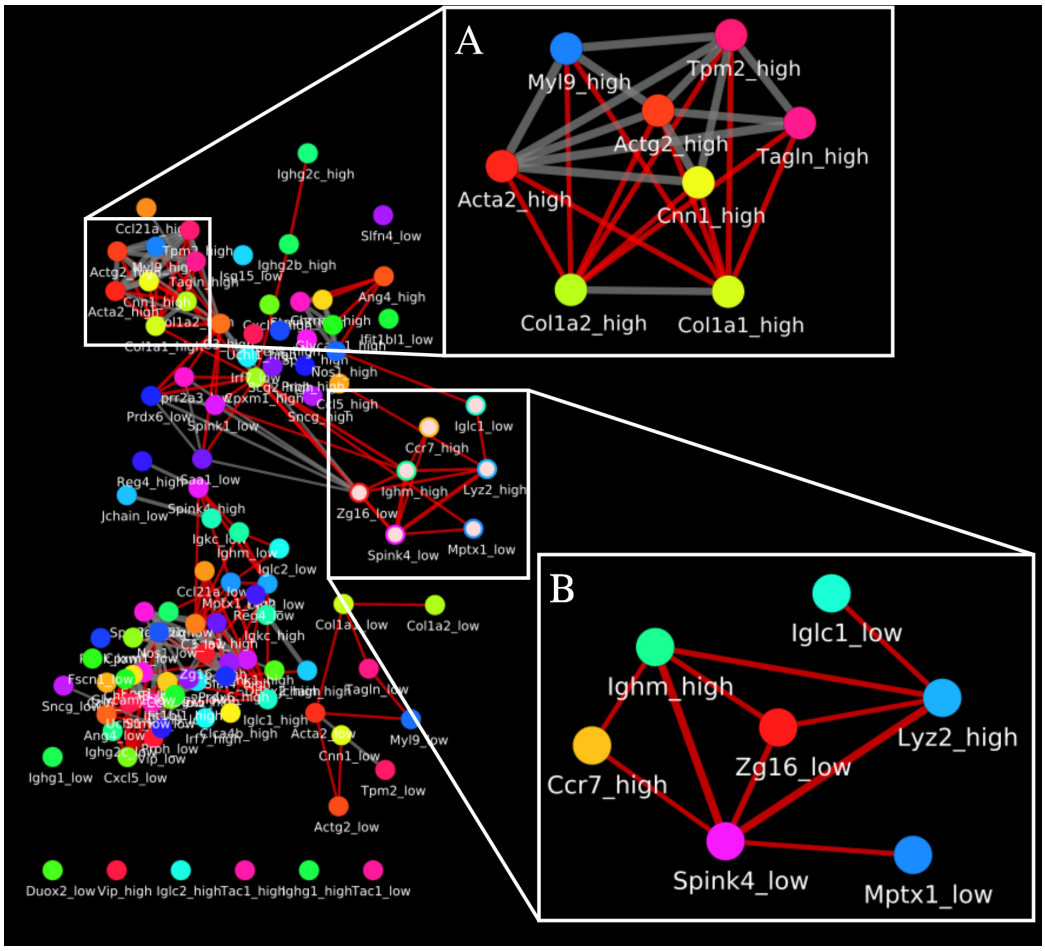


Figure 4.10: *InsituNet network created from anti-CD40 treated tissue sections (distance=100px, label permutation significance). Edges are highlighted in red where relationships only occur in treated tissues. Two network modules of particular interest are annotated; A) a module of genes highly localised to muscle tissue identified during pathology annotation in both treated and untreated sections, and B) a module of genes which only occurs in anti-CD40 treated sections, localised to regions of immune cell infiltration.*

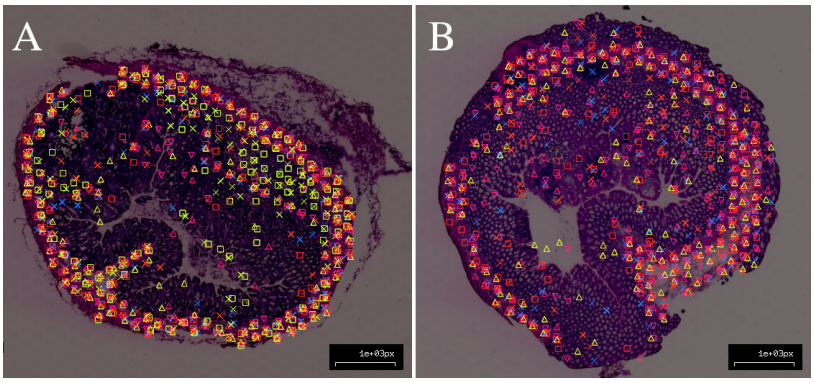


Figure 4.11: *InsituNet tissue view of a muscle tissue related network module (i.e., the result of selecting the nodes in Figure 4.10 A). A) Anti-CD40 treated tumour section B) PBS treated control tumour section. Spatial co-localisation of COL1A1 and COL1A2 was not evident in control sections.*

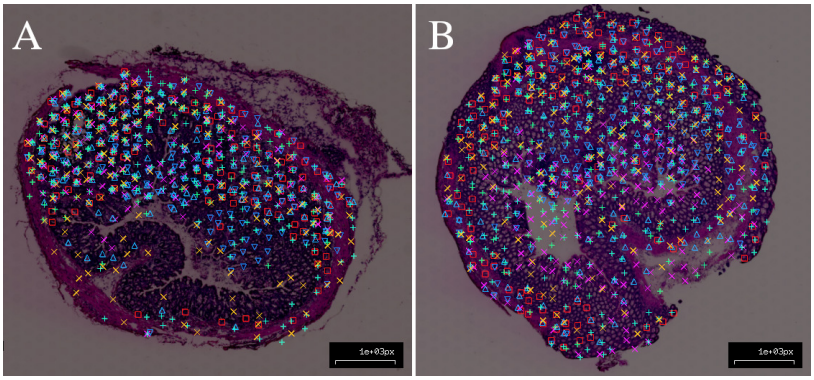


Figure 4.12: *InsituNet tissue view of an immune infiltration related network module (i.e., the result of selecting the nodes in Figure 4.10 B). A) Anti-CD50 treated tumour section. B) PBS treated control tumour section.*

4.4.3 A spatial omics approach to assess anti-CD40 induced immunotoxicity in the liver

As we have shown, immune agonist antibodies (IAAs) such as anti-CD40 are a promising approach for sensitising immunologically cold tumours to ICIs. Their clinical translation, however, has been limited due to dose-limiting toxicity and serious immune-mediated side effects including potentially fatal liver damage and cytokine release syndrome (Mayes *et al.*, 2018). In light of recent studies revealing that the gut micro-

biota are critical for ICI anti-tumour efficacy (Routy *et al.*, 2018), we hypothesised that the gut microbiota may also mediate the immunotoxicity induced by IAAs. As part of this study (Blake *et al.*, 2021), mouse livers were imaged using MALDI mass spectrometry imaging (MSI) to assess microbiota-mediated alterations to lipid metabolism, providing a spatially-resolved map of the liver lipidome. Three anti-CD40 treated mouse liver sections were imaged (one each from the untreated (No ABX), antibiotics treated (ABX), and germ-free (GF) mice) Both negative and positive ion mode were used for imaging, however it was found that most non-noise features were detectable in negative mode, and so only the negative ion results were used. To enable rapid visual exploration of different masses, I developed a visualisation tool, MSpecView, which can scan through each mass spectrum and identify those masses which have a large fold change difference in any of the treatment groups (Figure 4.13).

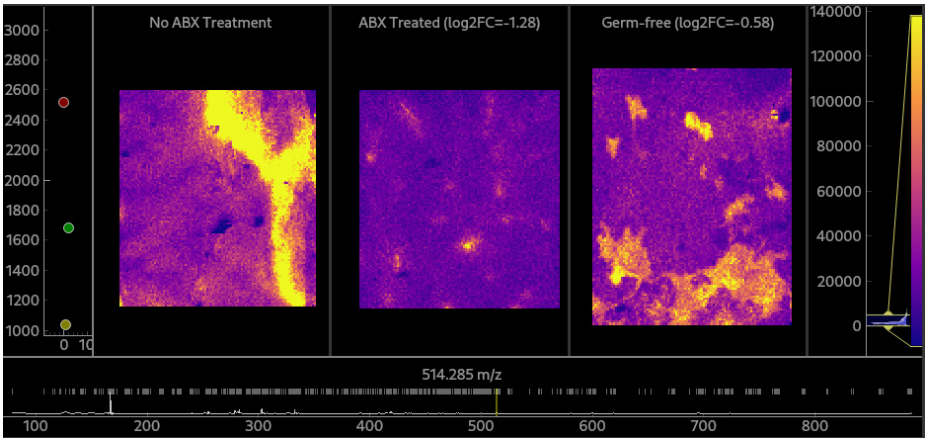


Figure 4.13: *Screenshot of the MSpecView tool, visualising the 514.285 m/z peak (taurallocholic acid). The three panels of the tool show three liver sections from the control, antibiotics treated, and germ-free mice, while an adjustable intensity scale and colourbar can be seen to the right. The mass spectrum may be scanned by holding left and right, or mousing over the spectrum graph seen at the bottom of the screen.*

MSpecView me to scan through the mass spectrum, stopping on masses in which the ABX or GF liver sections displayed a large mean intensity difference compared to control. After identifying masses which exhibited large fold changes in intensity across treatments, specific masses were selected for visualisation (Figure 4.14. Potential lipid identifications were matched to the m/z peaks using the Lipid Maps Structure

Database (Sud *et al.,* 2007).

I found that the livers of GF and ABX treated mice had an accumulation of cholesterol sulfate, which has previously been reported to occur in mice depleted of their gut microbiota (Sayin *et al.,* 2013). These mice also had significantly reduced levels of bioactive lipids which can be produced by gut bacteria and are important for the modulation of inflammation in the liver, including taurallocholic acid, arachnidonic acid, and C17 sphingosine. These data were consistent with transcriptomics analysis of these livers and suggested that both anti-CD40 and antibiotic treatment altered the expression of genes involved in liver bile acid metabolism. Serum levels of the pro-inflammatory cytokines TNFα and IL6 were also significantly lower in ABX mice. To further investigate whether bile acid levels directly influenced anti-CD40 induced liver damage and CRS, a 2% cholestyramine diet was fed to mice to sequester bile acids in the GI tract (Scaldaferri *et al.,* 2013). Interestingly, mice on the cholestyramine diet had significantly reduced levels of IL6 following anti-CD40 treatment, indicating that the mechanism of liver damage may be mediated by TNFα instead, which was not suppressed in the cholestyramine-fed mice. In summary, it was demonstrated that the gut microbiota are critical mediators of the immunotoxicity induced by anti-CD40, and importantly, that targeting the gut microbiota with antibiotics did not disrupt the anti-tumour efficacy of anti-CD40 in combination with anti-PD1. This would suggest that interventions targeting the gut microbiome may be an effective approach to increasing the clinical viability of IAAs.

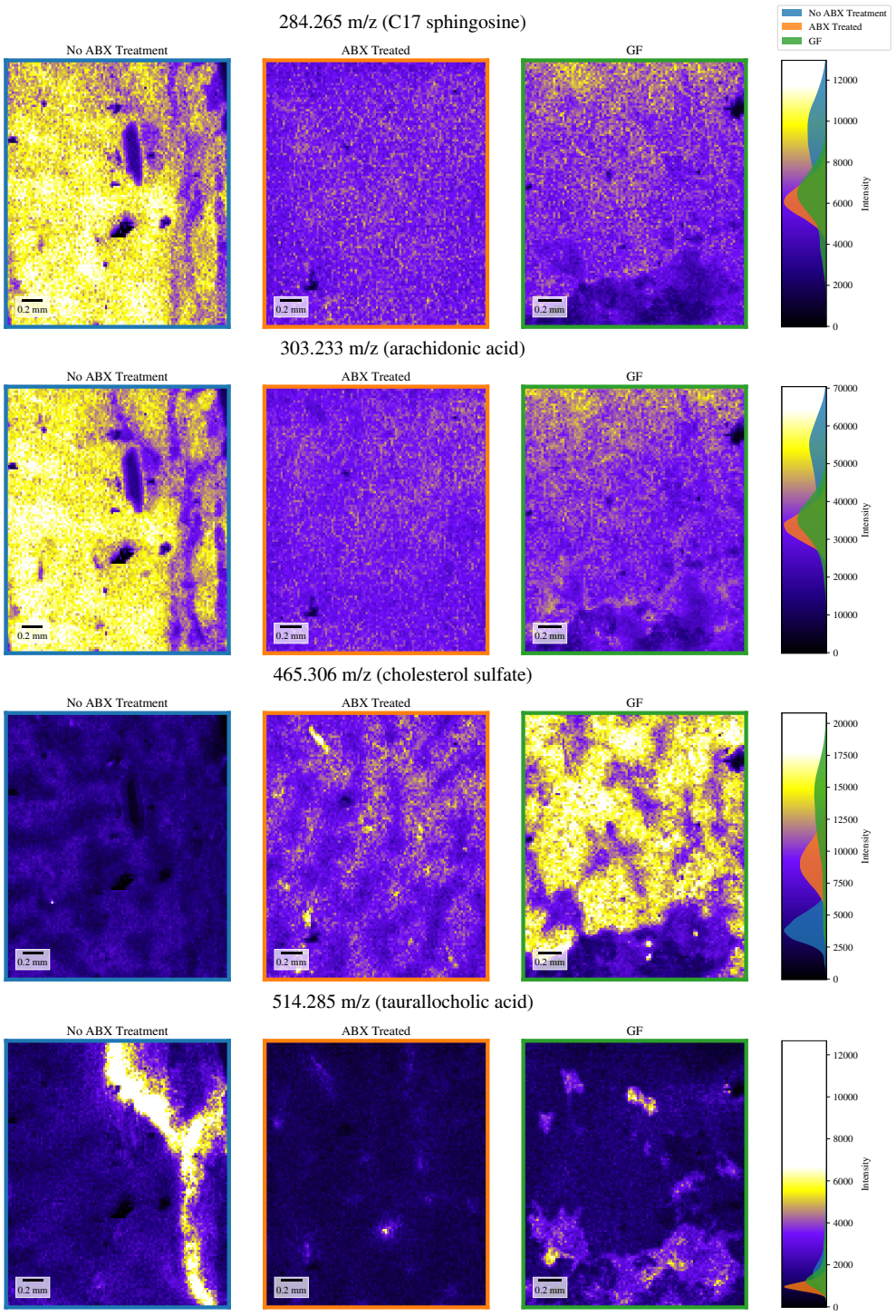


Figure 4.14: *Mass spectrometry imaging of liver sections from control, antibiotic (ABX) treated, and germ-free (GF) mice. The spatial distribution of four different m/z peaks which were identified using the MSpecView tool to be altered between the treatment groups are shown. Intensity histograms are shown to the right, separated into the ABX, no ABX, and GF groups. Lipid identification was performed using the Lipid Maps Structure Database (Sud* et al., *2007).*

4.4.4 Using spatial lipidomics to assess tumour heterogeneity in prostate cancer

As clinical prostate tumours are highly heterogeneous, MALDI-MSI was used to study the spatial chemical profile of 10 prostate cancer (PCa) patient biopsies (Mutuku *et al*, in review). In order to characterise the lipid profiles of different tissue types within the tumours, I constructed a partial least squares discriminant analysis (PLS-DA) classification model from the mass spectra of these biopsies to determine whether the detected lipid signatures would be useful for robust classification of different morphological features. The PLS-DA model was trained on 132 m/z features and contained 3 labels; benign, stroma and tumour, which were labels obtained from pathology annotation. The features were visualised using PCA (Figure 4.15), which resulted in reasonably distinct groupings, but did not separate the 3 groups completely.

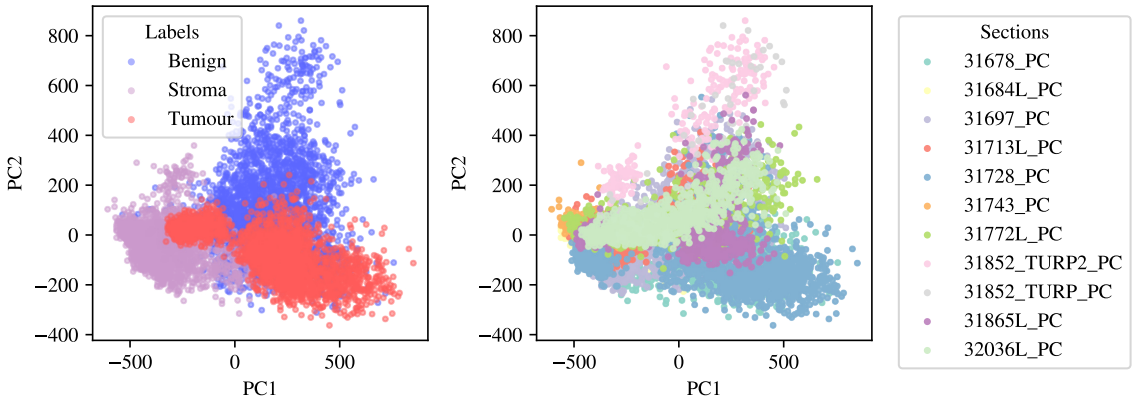


Figure 4.15: *PCA of mass spectrometry imaging features from 10 prostate cancer biopsies. Each point on these plots represents an individual pixel from MSI analysis. Left) Data points are coloured based on their pathology as Benign, Tumour or Stroma. Right) Data points are coloured based on each patient sample.*

It was apparent that an outlier sample was present in the tumour samples (Figure 4.15, right) (32036l_PC). While most of the patient tumour sections were visually distinct in PCA analysis, all of the data from this section were tightly clustered, potentially indicating a batch effect or technical issue with the section. It was therefore excluded from the PLS-DA training data, resulting in a much cleaner separation of

groups (Figure 4.16).

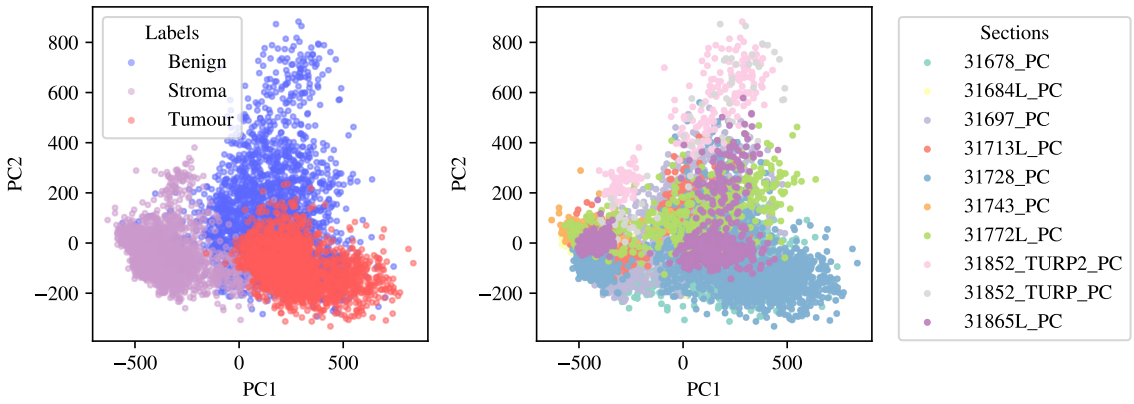


Figure 4.16: *PCA of mass spectrometry imaging features from 9 prostate cancer biopsies, excluding the 32036l_PC section due to batch effects. Left) Data points are coloured based on their pathology as Benign, Stroma or Tumour, as used to train the PLS-DA model. Right) Data points are coloured by tissue section of origin.*

The masses with the highest variance and their specificity for each of the three labels was visualised using the model biplot (Figure 4.17). This revealed that there were specific masses strongly associated with particular labels, for example the 459.246 m/z ion, identified as lysophosphatidic acid (LPA), was most strongly associated with benign regions (regression coefficient = 0.039). 826.563 m/z, identified as phosphatidylcholine (PC) (36:1) was spatially enriched in tumour regions, and strongly correlated with the tumour label in the PLS-DA model (regression coefficient = 0.031). AR dysregulation in PCa has been shown to cause overexpression of elongase enzymes (M. Chen *et al.*, 2018), and PC(36:1) is likely an elongation product of PC(34:1), the most abundant lipid detected, highlighting a possible rationale for its presence in tumour tissues.

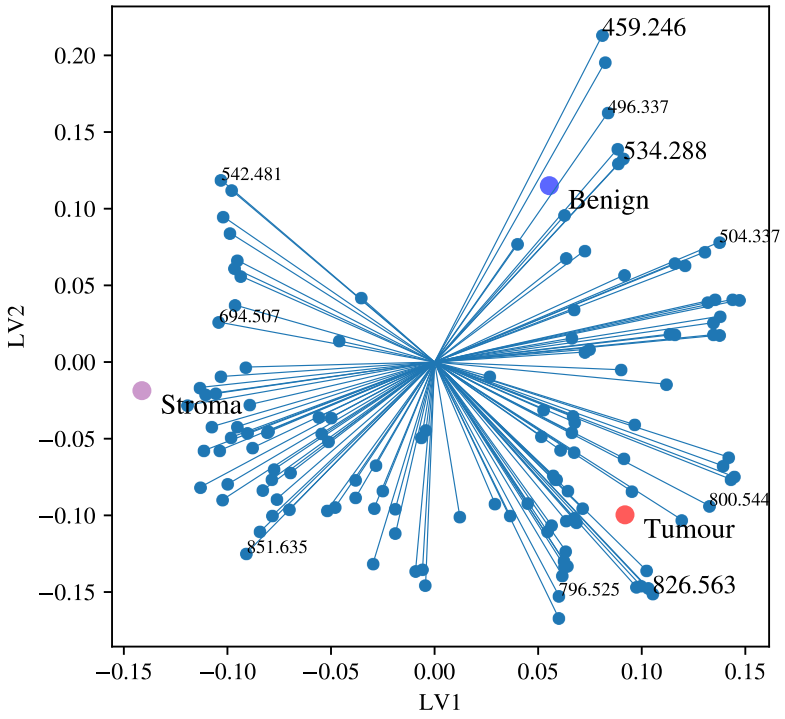


Figure 4.17: *Biplot for the PLS-DA model, including feature weights on latent variables (analogous to components in PCA) 1 and 2 (LV1 and LV2), plus benign, stroma and tumour classifications. The top 10 masses by total contribution to model variance are labelled.*

Using a 10-fold cross-validation procedure, the prediction accuracy of the PLS-DA model was determined. Each randomisation of the dataset ensured that tumour sections were split between test and training sets to ensure that the measured accuracy would be indicative of potential performance on an entirely novel tissue section. The most accurate prediction was of stroma (AUC=0.972), while both tumour (AUC=0.709) and benign (AUC=0.733) prediction performed well but less accurately than stroma (Figure 4.18). Discrimination between benign and tumour regions may have been more difficult due to the limited sample size and the large amount of heterogeneity between individual tumour lipid profiles. Despite reduced accuracy in predicting tumour and benign regions compared to stroma, this model demonstrates that the lipid profile of PCa tumours may be a useful tool to assist in the identification of tumour regions in PCa biopsies, and has identified several lipids that are correlated with tumour regions.

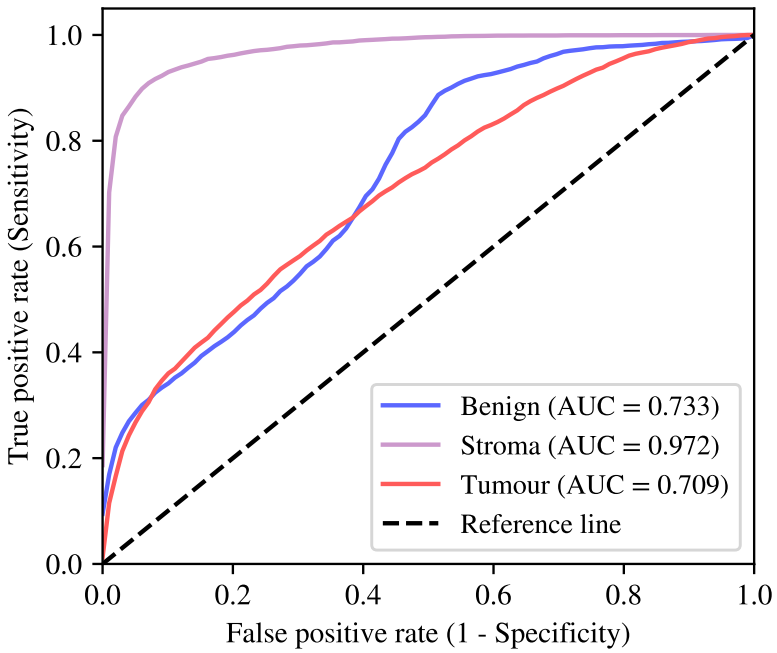


Figure 4.18: *Receiver Operating Characteristic (ROC) curve for cross-validated PLS-DA model predictions of stroma, benign, and tumour regions on 9 prostate tissue sections.*

Finally, the PLS-DA classifier was applied to each 2D tissue section separately to demonstrate the predictive capability of the model (Figure 4.19). Each pixel in these images is coloured by the most likely class as determined by the PLS-DA classifier. Model predictions were compared to histologically annotated regions to confirm that tumour / benign classifications were accurate. The model was most confidently able to distinguish stroma from other regions (AUC = 0.972). Much of this capability was not due to masses uniquely identifiable in the stroma, but rather due to the absence of certain masses. Indeed, the most discriminant mass for stroma was 534.288 m/z, a probable lysophospholipid species which was much more strongly associated with both tumour and benign regions than stroma (regression coefficient = -0.02). Prostate cancer is well known for its multifocality, exhibiting high levels of molecular and spatial heterogeneity (Tolkach & Kristiansen, 2018), meaning identifying lipidomic profiles capable of identifying tumour regions across different patients is an inherently more difficult task than classifying more normal tissue types. Given this, and that the model could perform as accurately at classification of tumour as it did (AUC =

0.709) with tumour and benign areas being quite distinct in their lipidomic profiles is strong evidence that the lipidomic profile of PCa as detected by MALDI-MSI could be a potentially useful clinical tool.

Finally, I investigated whether applying InsituNet to analysing MALDI-MSI data would be possible. I converted the 31678_PC prostate tissue section into a csv file for import into InsituNet, and generated a mass co-localisation network from the section (Figure 4.20). As a proof of concept, this demonstrated that analysis of MALDI-MSI data with InsituNet was possible. I found that this network very robustly separated masses which could be considered noise into their own module (the dense module at the top of Figure 4.20 A and B). These masses are likely an artefact of MALDI-MSI due to only being found around the outer edge of the tissue. These noise artefacts can also be observed in Figure 4.19, in which the outer edge of each section is labelled as stroma with low-confidence by the PLS-DA model. In comparison, the other masses internal to the tissue did in fact localise to pathology annotations, including several masses circled in Figure 4.20 A which surrounded the main tumour area of the tissue, and masses highlighted in Figure 4.20 B which were localised in regions of inflammation. Interestingly the masses identified in Figure 4.20 A near the tumour region were not strongly associated with the tumour label by the PLS-DA model, indicating that this approach may be beneficial for identifying spatial relationships which were not possible to identify using standard classification approaches.

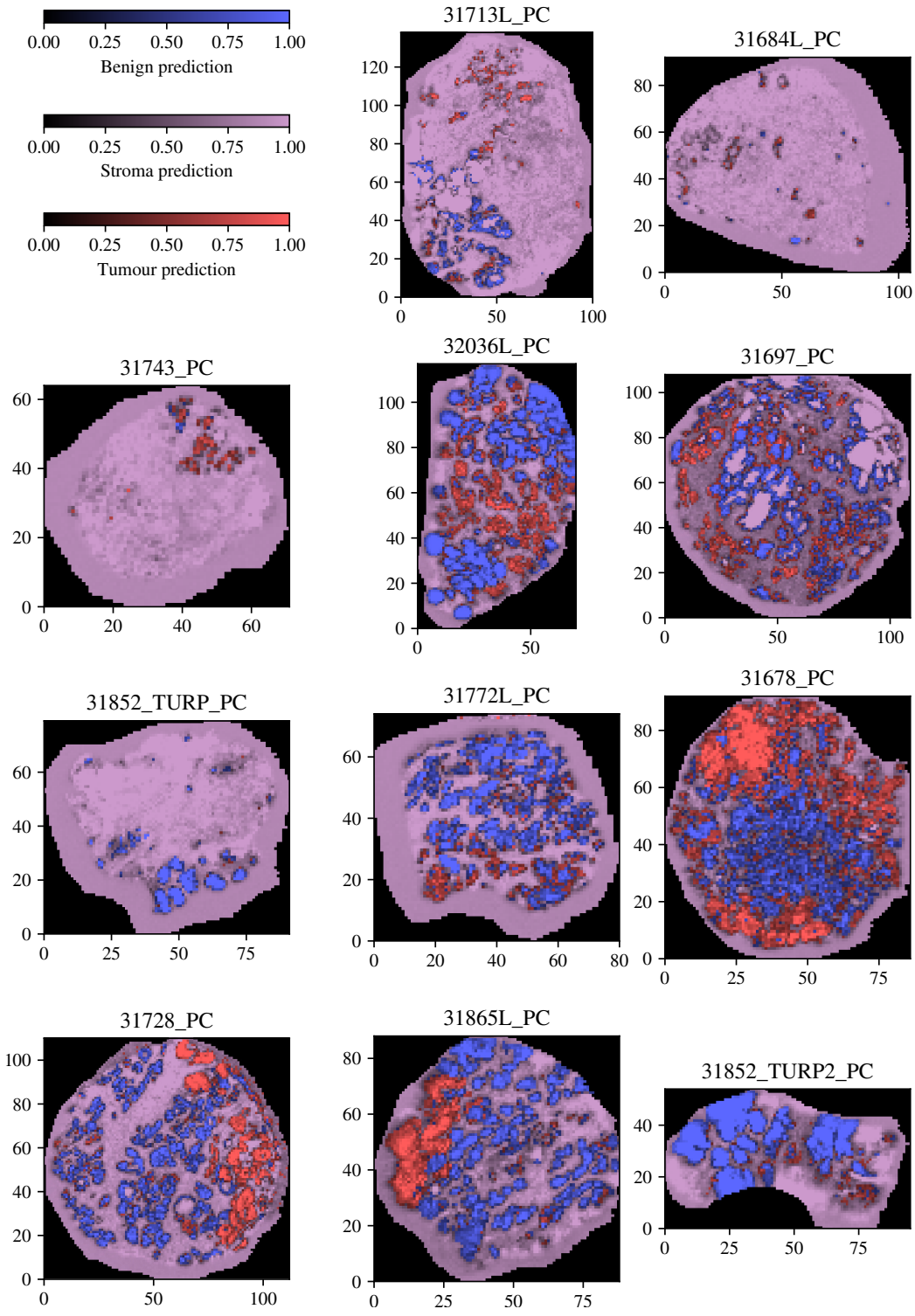


Figure 4.19: *Visualisation of the PLS-DA model predictions for each of the prostate tissue sections from 10 different PCa patients (with two different sections from patient 31852). Each pixel is coloured according to whichever class (benign, stroma or tumour) had the highest probability as determined by the PLS-DA model. Pixel width = 40 μm.*

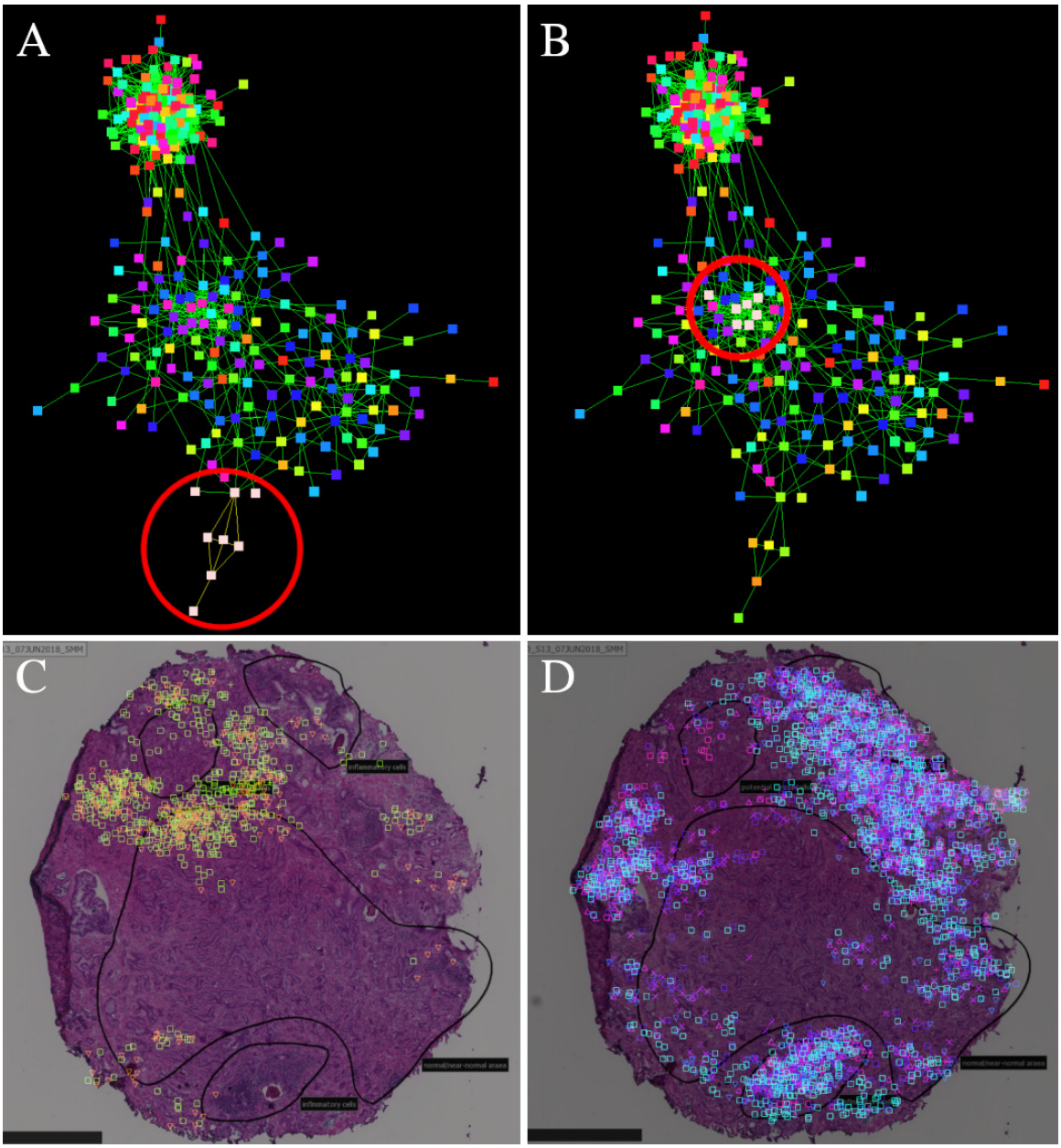


Figure 4.20: *InsituNet applied to the 31678_PC prostate tissue section reveals spatially localised lipidomic profiles. Different regions of the network were highlighted (A and B) to reveal the localisation of masses within different pathology annotations on the tissue (C and D). The masses in A were found to mostly be present in regions annotated as tumour (C), while the masses in B were found to be localised within regions of immune infiltration (D).*

4.5 Discussion

The advent of spatially-resolved 'omics technologies represents a significant opportunity to study biological phenomena such as tumour heterogeneity and tumour immune and metabolic microenvironments. In this series of experiments utilising spatially-resolved 'omics, I aimed to investigate tumour heterogeneity as well as develop tools to better utilise the spatial dimension of these data.

While many existing analysis methods such as those for single cell RNA sequencing (scRNAseq) can be applied to spatially-resolved technologies, they do not truly utilise the spatial dimension of these data. I have developed InsituNet, one of the first network solutions in this space, and demonstrated its application to both fluorescence in situ hybridisation (FISH)-like spatially-resolved transcriptomics such as *in situ* sequencing (Ke *et al.,* 2013), as well as array-based spatially-resolved transcriptomics such as Spatial Transcriptomics (Ståhl *et al.,* 2016). InsituNet is capable of constructing network representations of spatially-resolved transcriptomics data. I found that the resulting networks were effective for exploration and analysis of the otherwise extremely dense and difficult to interpret spatially resolved transcriptomics data, and were capable of highlighting spatial features which would be unclear or missed by conventional methods which are not spatially-aware.

Using spatial transcriptomics, it was possible to see the influence of the immune agonist antibody (IAA) anti-CD40 on microsatellite stable models of colorectal cancer. Spatial transcriptomics analysis revealed that the anti-CD40 treated tumours were strongly enriched for immune signatures, and that these signatures were capable of infiltrating into the tumour, best exemplified by the localisation of cluster 3 infiltration into the treated tissue areas 1 and 2 (Figure 4.7, 1 and 2). Furthermore, with InsituNet a module of spatially co-localised genes within these regions which existed only within treated tissues was identified, which notably included high levels of genes expressed by immune cells including *CCR7*, *IGHM*, and *LYZ2*, but also low levels genes known to be down-regulated in the microsatellite instability subtype of CRC, including *SPINK4* and *ZG16* (Meng *et al.,* 2018; X. Wang *et al.,* 2019), indicating that the treatment was driving the tumours to become more similar to the microsatellite instability subtype.

The localisation of this immune infiltration and the interactions of these genes was only possible to observe using spatially-resolved transcriptomics.

I also demonstrated that the InsituNet approach could be extended to spatially-resolved technologies beyond transcriptomics, with a proof-of-concept of InsituNet analysis of spatially-resolved metabolomic data obtained via matrix-assisted laser desorption ionisation (MALDI) mass spectrometry imaging (MALDI-MSI) (Figure 4.20). In investigating the immune-mediated side effects of IAAs, MALDI-MSI was also applied to generate spatially-resolved metabolic maps of mouse livers. It was found that germ-free mice and mice treated with antibiotics were protected from the liver damage and CRS caused by anti-CD40 treatment, however the dimensionality of mass spectrometry data proved to be difficult to overcome when doing exploratory work. I developed a convenient tool for exploratory purposes named MSpecView for the visualisation and comparison of MALDI-MSI datasets. Using this tool I identified specific masses of interest, especially bile acids, which were subsequently found to directly influence anti-CD40 induced liver damage.

More conventionally, MALDI-MSI data is analysed with tools such as partial least squares discriminant analysis (PLS-DA) for classification purposes. I utilised the PLS-DA approach to develop a tissue / cell type classifier trained on the lipidomic profiles of biopsies from 10 prostate cancer patients, finding that this approach was surprisingly effective (with a cross-validated AUC of 0.709 for classifying tumour tissue, see Figure 4.18), but limited in terms of the spatial conclusions that could be drawn for these highly spatially heterogeneous tumour samples. A limitation of this work is that the small sample size used to train the PLS-DA model. The PLS-DA model is useful as a proof of concept, however given the sample size used here it is likely not useful as a diagnostic tool to identify tumour tissue in other samples, especially given the known susceptibility of PLS-DA to overfitting (Rodríguez-Pérez *et al.,* 2018). In future, a much larger cohort would be desirable to validate the current findings. In comparison to this approach, InsituNet allows immediate feedback as to where genes, metabolites, or other molecules are spatially distributed, and truly derives its information from the spatial dimension of these data, rather than simply assessing each point in isolation, a critical limitation of these other approaches.

The ability of InsituNet to be adapted to multiple forms of spatially-resolved data suggests that re-application to various other spatially-resolved technologies beyond transcriptomics would be possible, and indeed I have demonstrated proof-of-concept of this in lipidomics datasets from MALDI-MSI. The nature of spectral data however means such analyses would greatly benefit from the incorporation of some form of intensity into the algorithm. Currently, InsituNet is only capable of representing these data by thresholding intensities. It is likely that intensity could be incorporated into a more sophisticated algorithm to assess co-localisation significance, which would be an excellent future direction for InsituNet. In future, combining multiple spatially-resolved 'omics with a network framework similar to InsituNet could be a useful systems-level approach to analysing and exploring spatial data from diverse technologies. Potentially as multiple spatialomics technologies become available, many different data types could be integrated onto the same network, with nodes which represent metabolites, proteins, and genes all being presented on different levels of the same co-localisation network, enabling a new level of integrated, spatially-resolved analysis of tumour heterogeneity.

5. Conclusion

The molecular heterogeneity of cancer confounds patient treatment, which is the primary motivator of this thesis. The concepts explored here were tested in the context of colorectal cancer (CRC) as a proof of concept. In recent years, the availability of patient-specific molecular data on a large scale from sources such as The Cancer Genome Atlas (TCGA) (Weinstein *et al.,* 2013) has presented both a challenge and opportunity to construct better predictive models of the biological differences between patient tumours. Current cancer staging systems for CRC such as TNM incorporate relatively little molecular data beyond a few specific biomarkers, which may be severely limiting the scope of interventions which could otherwise be applied.

Efforts to classify tumours based on molecular data in CRC build upon the existing literature on tumourigenesis. At least three common pathways of molecular development have been described in CRC, chromosomal instability (essentially the canonical multi-step process described by Fearon and Vogelstein (Fearon & Vogelstein, 1990)), microsatellite instability (MSI), and CpG island methylation. The state of the art in terms of transcriptomic stratification of patient tumours is the Consensus Molecular Subtypes (CMS) (Guinney *et al.,* 2015), which emphasise the importance of MSI, with the MSI-high CMS1 being one of the most clearly differentiated subtypes, typically presenting with hypermutation and immune cell infiltration. The CMS also highlight the relevance of pathways related to metastasis such as epithelial-mesenchymal transition for defining CMS4, the subtype with poorest patient prognosis. In examining the CMS and their application with within colorectal cancer data from TCGA, it is apparent that differences in patient survival are only significant for patients classified as CMS4, with other CMS subtypes lacking further prognostic resolution. I suggested that further avenues for increasing the utility of transcriptomics data may improve classifications. One of the most important aspects of this was to focus on inter-patient heterogeneity, rather than bulk differences between tumour and normal samples.

I devised a method to identify patient-specific differentially expressed genes (PSDE genes) and applied it to the TCGA CRC cohort. I aimed to use PSDE genes to identify changes between patient tumour samples, and use this information to create a more robust stratification of patients. Bioinformatics tools that focus on patient-specific analysis of transcriptomics and other data such as PARADIGM (Vaske *et al.,* 2010) are not frequently applied for the purposes of patient stratification. Such approaches also generally use normal samples to control for the noise between samples. In my identification of PSDE genes, I opted not to do this, which allowed me to include more samples than some methods (as a minority TCGA CRC tumour samples have paired normal samples), but also made the results more susceptible to noise. Because of these decisions, I spent a great deal of effort attempting to clean technical artifacts and noise from my datasets, which may be a limiting factor of this type of approach. My decision to have a threshold level to assign significance to genes within an individual made analysis simpler, but is perhaps less robust than providing a continuous score for all genes, which is an approach taken by other patient-specific tools such as GSVA (Hänzelmann *et al.,* 2013). I identified PSDE genes in 550 patients from the TCGA CRC cohort. Splitting these into up and down-regulated PSDE genes, I found that the up-regulated PSDE genes tended to be ones essential to CRC development and progression. I used PSDE genes to perform hierarchical clustering of patient samples and found that PSDE-informed clusters were enriched for specific functional processes. Furthermore, I uncovered that classifying patients using this approach identified novel sets of patients with significant differences in survival, who were not identified using the CMS subtyping approach. Specifically, patients in PSDE-informed cluster 4 (PIC4) had significantly poorer survival than patients from multiple other PIC clusters. I next identified pathways that were enriched in individual patients, given their PSDE genes. After performing this analysis, I then performed hierarchical clustering on pathway enrichment scores, finding that this pathway level approach led to novel clusters which were not identified from gene expression alone. This included a division of the otherwise homogeneous MSI-high CMS1 subtype with significantly different survival in patient subgroups, a result which suggested that biologically distinct subgroups of MSI may exist. In summary, I found that a great deal of inter-patient tumour heterogeneity that was predictive of patient survival was still not well captured by

the existing standards of molecular classification.

I followed up the patient-specific pathway analysis approach by investigating whether a network approach to modelling tumour heterogeneity would be a better representation of the varied alterations that occur within individual patients. I had previously observed that taking a pathway-level approach to patient stratification revealed otherwise hidden patient subgroups. These results were interesting, however this pathway-based method of classification was heavily reliant on pathway database annotations. I hypothesised that if I went beyond the available pathway annotations and directly used protein-protein interaction (PPI) networks, I would be able to uncover more otherwise hidden similarities between patients. The application of networks to patient-specific modelling in cancer has often been touted as a way to integrate diverse forms of data, using the network as a structural base (Hastings *et al.,* 2020). Using networks to create probabilistic models which are used to predict outcomes in cancer has also been a major point of interest, but the complexity of such networks makes their application somewhat difficult (Ozturk *et al.,* 2018). More broadly, network topology is suspected of being linked to biological structure. Some authors (Breitkreutz *et al.,* 2012) have attempted to directly link topology to patient outcomes, an area that I investigated in relation to the patient-specific networks I created. The Epidermal Growth Factor Receptor (EGFR) network is critical to tumour progression in CRC, and has previously been shown to be rewired in KRAS mutant CRC (Kennedy *et al.,* 2020). PPI networks have been shown to be dynamically altered in response to mutations in other contexts (Sahni *et al.,* 2015). Mutations may lead to interaction-specific alterations ("edgetic" effects) of variable strength, or disrupt entire gene products, causing complete node removal and disrupting all interactions. Using PPI data obtained from IMEx (primarily from the PRIMES project (Kennedy *et al.,* 2020)), I constructed a high-quality EGFR network model (EGFR-HQ). I then used this model as a base which could be altered using individual patient data. I first used a node-removal approach in which I deleted nodes in individual patient networks corresponding to genes which were expressed significantly lower than the cohort median level. I then used these patient-specific networks as a base for further analysis. I employed information flow analysis, a method of network propagation (Cowen *et*

al., 2017). Using information flow analysis, I aimed to model how signals flow from EGFR through the EGFR network to downstream transcription factors, and whether patient-specific alterations to this network would cause significant modification to this information flow.

I found that the node-removal strategy to remove significantly under-expressed patient-specific genes resulted in substantial modifications to network topology. I was able to identify topological network properties such as clustering coefficient which were significantly different between specific patient subtypes. Patient-specific network properties were not however directly predictive of patient survival, something I had hypothesised might be the case due to previous studies which have identified differing network topology in cancers with different survival rates (Breitkreutz *et al.,* 2012). To better model edgetic effects of mutation, I removed edges in patients in which mutations could be matched from a database of curated PPI disrupting mutations (del-Toro *et al.,* 2019). Due to fairly small numbers of such matches, I aimed to predict which specific protein domains potentially mediated PPIs. I identified 260 such domains, and supposed that edges with mutations within these domains would likely be disrupted. Although the total number of these edge removals was still small, I did find that the degree of nodes containing these mutations within these domains was significantly higher than other nodes. Ultimately, edgetic effects likely contribute to tumour heterogeneity on a network level, however modelling them may be a complex task.

I modelled information flow through these networks from EGFR to downstream transcription factors, initially using an existing tool, ITM Probe (Stojmirović *et al.,* 2012). I found however that this approach was insufficient for the particular use case (modelling many patient-specific networks with slight alterations) and so developed my own tool, Simulated Information Flow For Individualised Networks (SIFFIN). SIFFIN automatically infers directionality of edges and simulates information flow between topologically distinct networks, outputting scores that are directly comparable. Using SIFFIN, I identified significant alterations in information flow between patient networks and stratified patients based on these differences, identifying multiply patient clusters. I found these clusters were distinct from the CMS. I also found

that the alterations to information flow to transcription factors were often correlated with changes in expression of the same transcription factors. However, it was possible that some of this correlation was due to the edge weighting approach used rather than the topology of the entire network. My network-propagation approach is similar to previous models, but has the advantage of both being computationally efficient, as well as enabling comparison between each individual in a cohort. This allows for integration of large-scale data and comparing hundreds of patient-specific datasets. The results of this analysis show that network approaches can uncover a significant amount of otherwise hidden information in large-scale biological data. In conclusion, I found that a network approach to modelling the heterogeneity of cancer may lead to the discovery of biologically relevant insights that are not apparent when using approaches that do not incorporate network topology.

I finally pivoted to another aspect of tumour heterogeneity, spatial heterogeneity. Tumours are heterogeneous not only between individuals, but also within different regions of the tissue. Spatially resolved omics are now becoming commercially available, including spatially-resolved transcriptomics with hybridisation arrays (Ståhl *et al.,* 2016). However, analysis tools and methods in this field are still in their infancy, meaning there is a demand for novel tools which better integrate the spatial dimension. The integrative capability of networks could mean network analysis may answer this demand. I developed a novel network approach to spatially-resolved omic data called InsituNet (Salamon *et al.,* 2018), a tool for analysing spatial data in a network form. I demonstrated that this approach could convert spatial information into a network form that could then be analysed with more conventional network tools. As I case study I was able to use spatial transcriptomics data to analyse the influence of immune agonist antibodies on microsatellite stable preclinical models of CRC (Blake *et al.,* 2021). With InsituNet, it was possible to visualise modules of spatially co-localised genes that were unique to treated tissues. These results suggested that anti-CD40 treatment may be able to drive tumours to become more similar to the MSI subtype, which has implications for potential approaches to otherwise treatment-resistant MSS tumours.

InsituNet represents one of the first user-friendly tools available to map spatially-

resolved omics into a network form, which I have demonstrated is capable of integrating this information into informative models. InsituNet's algorithm is sufficiently generic that as additional spatialomics technologies become available, it should be possible to adapt the program to support these. I demonstrated here that InsituNet could be extended to analyse metabolomics data collected via mass spectrometry imaging. I also investigated the use of other approaches for the analysis of spatial metabolomics data, including machine learning methods. For example, I performed classification of spatial regions within tumour sections using PLS-DA. This form of classification was effective for identifying tumour tissue, but did not provide insights into the spatial co-localisation of masses in the way InsituNet did.

The development of underlying technologies continues to provide yet higher resolutions and quantities of molecular data, further filling in the picture of tumour heterogeneity and allowing for ever more detailed personalised models. Notably, Transcriptomic profiling has advanced significantly in terms of throughput and cost, with single cell and spatial methods both opening up entire new possibilities (Cieślik & Chinnaiyan, 2018). Although the genome-wide transcriptomic data used for constructing patient-specific networks here is not a standard procedure, it now seems possible that routine sequencing and analysis of tumour biopsies could occur in the future. Spatial methods like high definition spatial transcriptomics (Vickovic *et al.,* 2019), single-cell RNAseq and MALDI-MSI add additional dimensions to transcriptomic and metabolomic data that could drastically improve the usability of these omics data, providing crucial spatial context of cells and tissues. Given the significant interest in such methods it seems likely this will extend to other domains in the near future. Of course, all of these diverse and complex data represent a challenge to expand existing predictive models to make best use of the available data.

The ever expanding quantity and quality of data represents a serious challenge for bioinformatics research on multiple levels. In my view, one of the most essential tasks is the collection, management and curation of public data, without which predictive models are limited in how well they can be trained or tested. This is especially important given the current trend towards machine learning approaches that rely on huge sample sizes to be effective. The second crucial step is the development of

effective methods for integrating these diverse data sources. It has been repeatedly demonstrated that predictive models constructed using a combination of multiple complimentary forms of data can outperform approaches that rely on a single omics measurement (Boehm *et al.,* 2022), but this isn't guaranteed. I believe a network-based approach to data integration as demonstrated in this thesis is a viable way of performing this integration and constructing personalised models that are reflective of tumour biology, but the trade-off between incorporating as much data as possible and being precise is a difficult one. One thing I repeatedly noted in this thesis was the disproportionate effect of technical noise, which will continue to be an issue that must be addressed by bioinformatics approaches as the quantity of data increases. The final challenge for integrative models is to provide computational means of stratifying patients, predicting outcomes, and proposing novel therapeutic opportunities. As scale increases, more bioinformatics approaches will be called for which can identify the clinically significant information from these huge datasets. I expect that machine learning methods are likely to continue to increase in prominence due to the sheer scale of data available in the future. However, I think this is still the area most open for innovation and novel bioinformatics approaches that can extract clinically significant information from these huge datasets. Methods like SIFFIN and InsituNet represent my attempts to better utilise the current immense amounts of data in novel ways.

The work presented here investigates less well studied aspects of tumour heterogeneity. The inter-patient variability between molecular data, as modelled by modifications to network structure, resulted in predictions of significant changes downstream of EGFR. I was able to identify novel subtypes which were distinct from existing CRC subtypes, many of which were prognostically relevant. One dimension absent from my investigations in this thesis is time. Spatial and inter-patient molecular heterogeneity is important, but it is crucial to acknowledge that what was investigated here for the most part represented a single snapshot in time. An analysis that incorporated for example the change PSDE genes over time could very well radically alter the results here. From investigating these different aspects of tumour heterogeneity, I found evidence that inter-patient differences may be driving variable responses to cancer therapies, differences which are often not well represented in existing classifications. In

this thesis, I demonstrated that a network approach to integration can be a viable way to construct patient-specific models which can integrate this ever-expanding amount of patient-specific data.

6. Appendix

6.1 TCGA patient metadata analysis

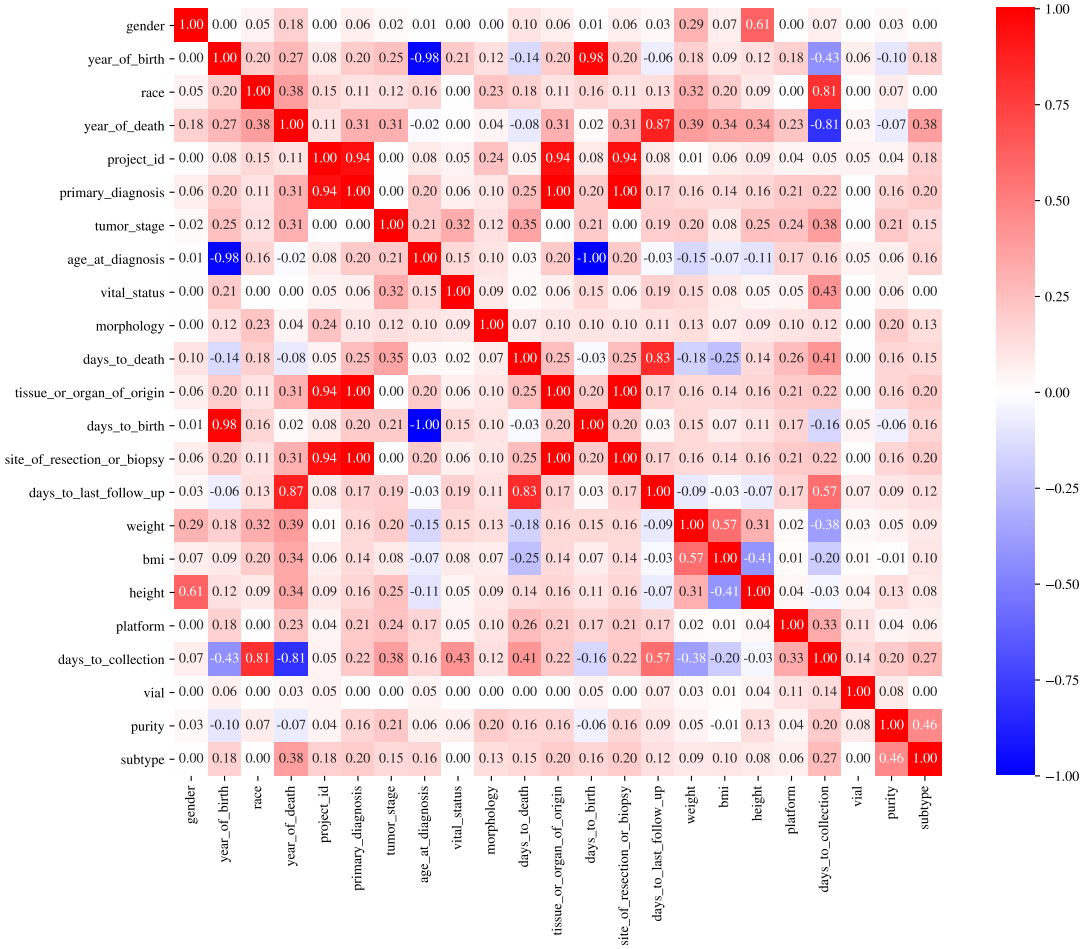


Figure 6.1: *Association matrix for 23 chosen metadata features. Pearson correlation was used to compare continuous data with other continuous data, Cramer's V method was used to compare categorical data with other categorical data, and Pearson's correlation ratio was used for continuous versus categorical data. These measures were normalised on a scale from -1 to 1, where 0 represents no association, and -1 and 1 respectively represent perfect negative or positive correlations.*

To demonstrate that large survival differences exist within this cohort, I stratified patients based on TNM stage (Figure 6.2).

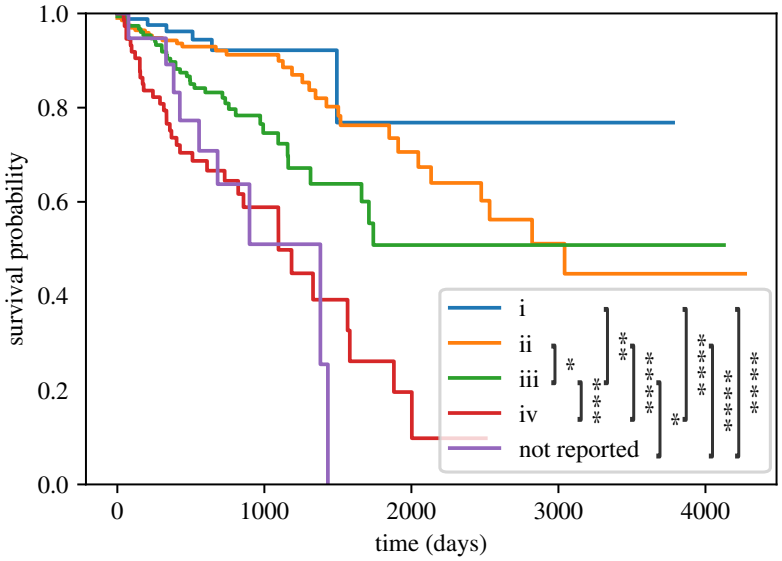


Figure 6.2: *Kaplan-Meier survival plot for patients classified by tumour stage. Using a pairwise logrank test, differences in survival time for all stages were found to be statistically significant ($p < 0.05$), except for between stage iv and samples with unreported stage. RMST at 5 years (shown in legend) was found to decrease with each progressive stage.*

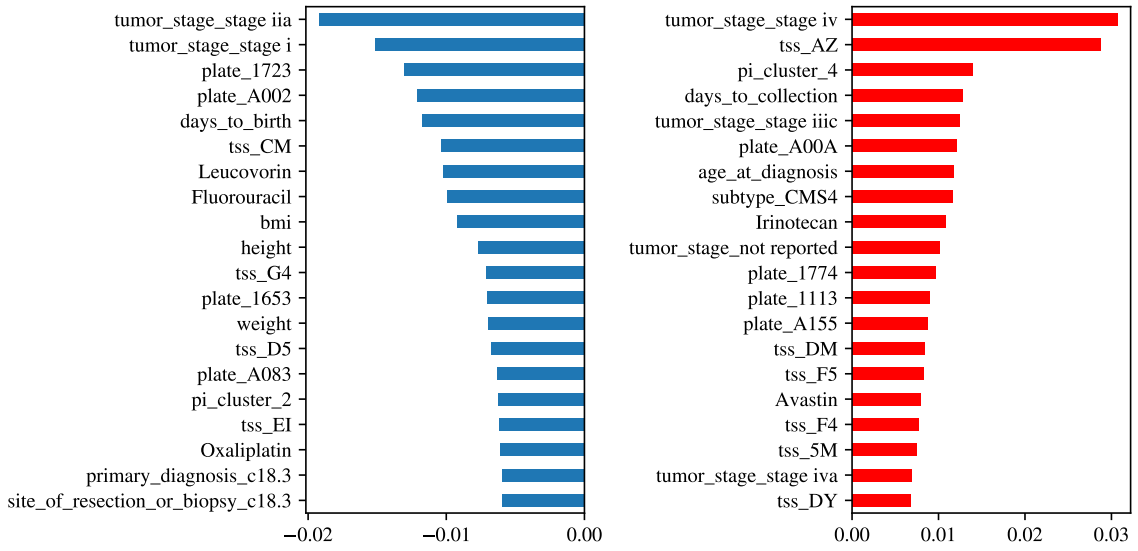


Figure 6.3: *Top 25 features of a PLS-DA model trained on TCGA metadata in terms of negative (left) and positive (right) correlation with patient survival.*

6.2 Read mapping

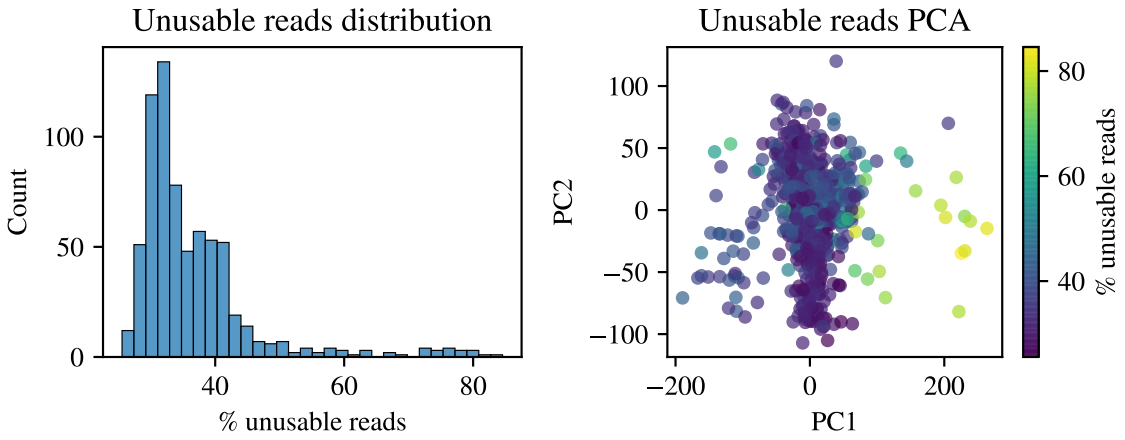


Figure 6.4: *Multimapped, ambiguous, and unmappable sequencing reads contribute to the unusable reads percentage. The total distribution of such reads is visualised as a histogram on the left. When visualising the percentage of these reads on a PCA plot (right), it is apparent that they contribute greatly to the total variance of the dataset.*

6.3 Samples per gene

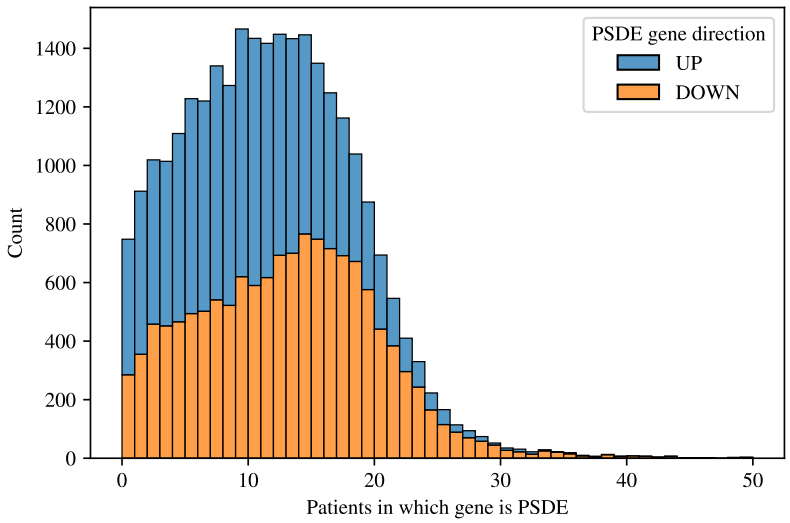


Figure 6.5: *Histogram of how frequently particular genes are assigned as PSDE across the TCGA CRC cohort per gene, split into up and down gene sets.*

6.4 PSDE Cluster Enrichment Analysis

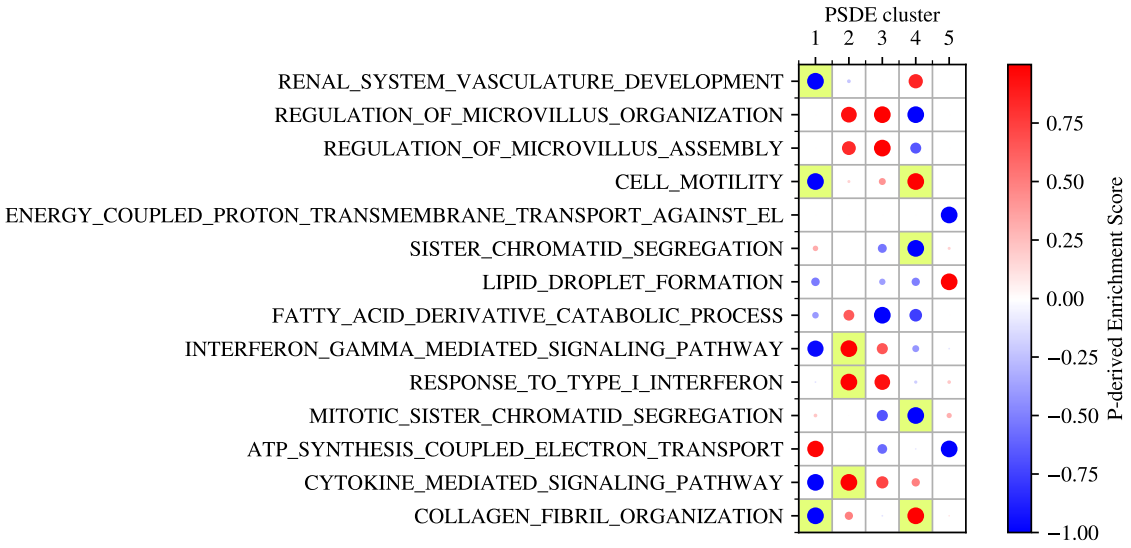


Figure 6.6: *Most significantly enriched pathways per PSDE cluster for GO Biological Process. Dot radius is proportional to FDR significance, while pathway enrichment significance at the FDR < 0.05 level is marked by yellow background.*

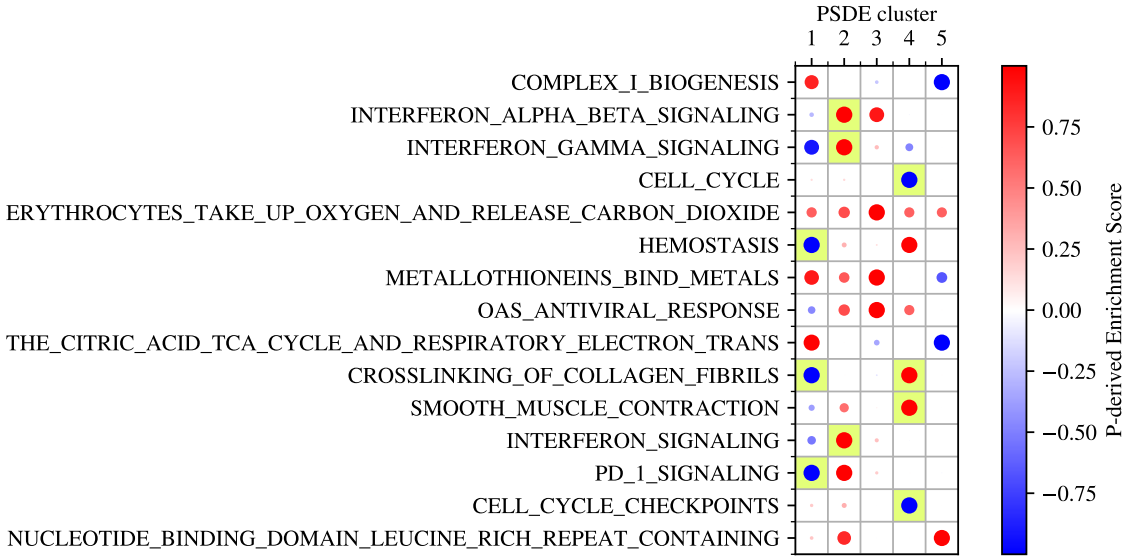
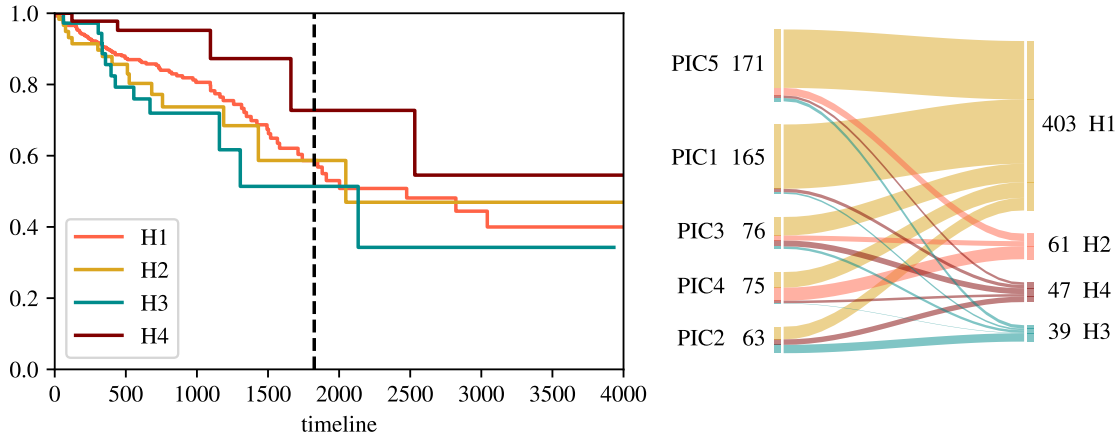


Figure 6.7: *Most significantly enriched pathways per PSDE cluster for Reactome. Pathways are selected by the top 3 most significant pathways for each cluster. Dot radius is proportional to FDR significance, while pathway enrichment significance at the FDR < 0.05 level is marked by yellow background.*

6.5 MSigDB Hallmarks pathway clustering



a) Kaplan-Meier plot of Hallmarks clustering

b) PI-Hallmarks alluvial comparison

Figure 6.8: *Novel clustering of patients based upon patient-specific MSigDB Hallmarks enrichments produced clusters which varied substantially from PI clusters (***b***). Kaplan-Meier analysis in (***a***) determined cluster H4 was significantly different to clusters H2 (p=0.04) and H3 (p=0.012) (pairwise logrank tests).*

A database which produced interesting results for pathway-level clustering was MSigDB Hallmarks (Figure 6.9). As with KEGG, 4 pathway clusters were obtained, H1-4. Cluster H4 patients in particular had significantly better survival probabilities than patients in the other clusters (Figure 6.8). This cluster featured a significant upregulation of hypoxia and glycolysis, pathways, among others (Figure 6.9). As hypoxia is known to initiate glycolytic metabolism in tumours (Kierans & C. T. Taylor, 2021), it is possible that these two pathways are causally linked. It is interesting that this cluster appears to have favourable survival outcomes, as hypoxia generally indicates an immunosuppressive environment and has been linked to poorer patient outcomes in immune-cold tumours (Craig *et al.*, 2020). However, the MSI-high subtype of CRC is known to develop sporadically via promoter hypermethylation of the mismatch repair gene *Mlh1* (Kawakami *et al.*, 2015), and it has been demonstrated *in vitro* that hypoxia can induce epigenetic inactivation of *Mlh1* (Weisenberger *et al.*, 2006), pointing to a possible molecular explanation for this MSI-high cluster.

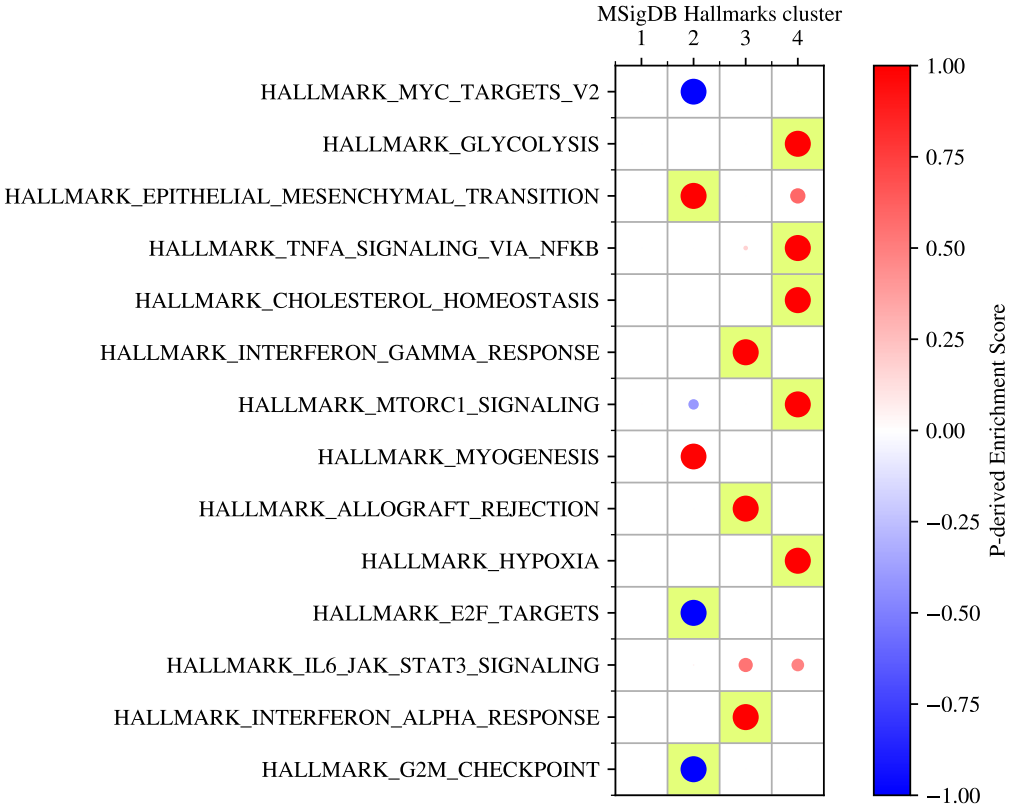


Figure 6.9: *Pathway enrichment analysis of MSigDB Hallmarks clusters. Pathways are selected by the top 3 most significant pathways for each cluster. Dot radius is proportional to FDR significance, while pathway enrichment significance at the FDR < 0.05 level is marked by yellow background.*

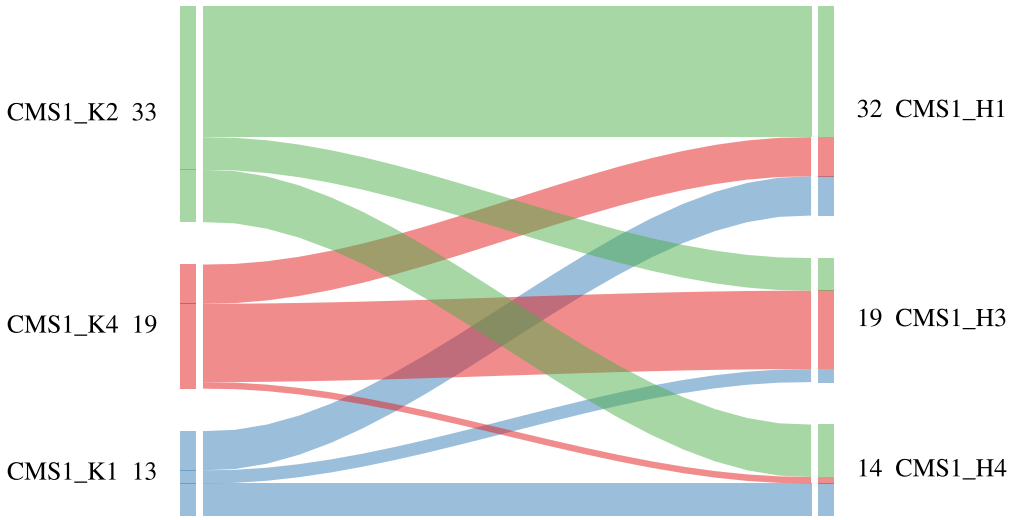


Figure 6.10: *Alluvial plot comparison of CMS1 sub-clusters as determined by patient specific KEGG (left) and MSigDB Hallmarks (right) pathway scores.*

6.6 Hierarchical clustering

Hierarchical agglomerative clustering algorithms follow a bottom-up approach: start assuming all samples are separate clusters, then link similar clusters together until a full hierarchy tree is produced. This type of algorithm can tend to scale poorly with increasing data sizes, and tends to be biased towards the identification of smaller-scale groups than larger groups. Divisive hierarchical clustering methods would have better performance and may prove more effective when trying to identify larger clusters. Alternative clustering methods include non-hierarchical partitioning methods such as k-means. These are generally much faster and simple to use, however they do not provide information on the relationship between all samples in a hierarchy. However, by use of k-means clustering we can provide very quick assessments of cluster tendencies. In the gene clusters defined here, a common issue encountered was small cluster sizes. Potentially by using other clustering methods, larger clusters could be elucidated. However, as can be seen from the various heatmaps of gene expression produced, there really are quite a few small clusters of seemingly related patients. It is possible that some of these are due to technical artefacts that were not fully corrected or erased

by the filtering methods employed here. As an unsupervised approach, hierarchical agglomerative clustering it was chosen as it has a high chance of identifying novel clusters without any assumptions on what different clusters should look like. However, this comes at the cost of sensitivity to such potential technical artefacts.

6.7 Alternative clustering methodologies

Rather than using hierarchical clustering, an alternative method is simple k-means partitioning. This can be done exceedingly efficiently for a large number of genes by firstly running principal components analysis (PCA), then applying k-means clustering to the resulting principal components. Simply applying to two dimensions (Figure 6.11) allows for a simplistic overview of this variation, but when applied to multiple further components (Figure 6.12), k-means yields quite similar results to hierarchical clustering when there large sources of variation among samples.

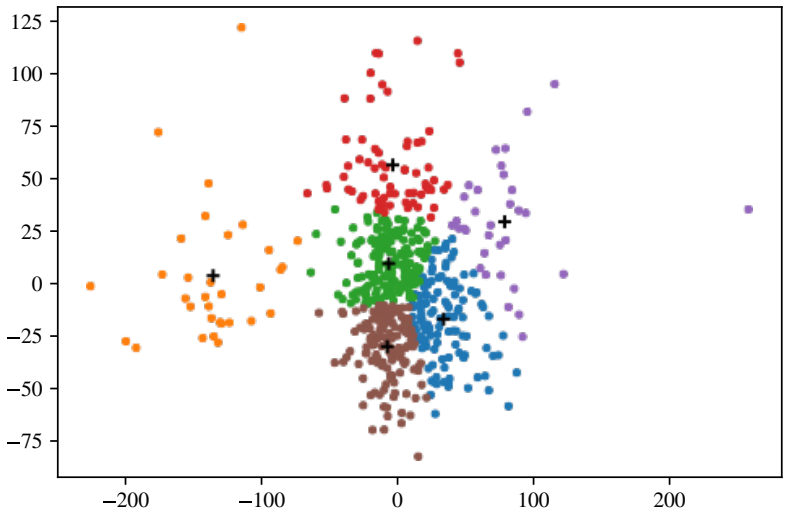


Figure 6.11: *K-means clustering on normalised gene expression data following principal components analysis. Only the first two components of variation were used for clustering. Crosses indicate cluster centers.*

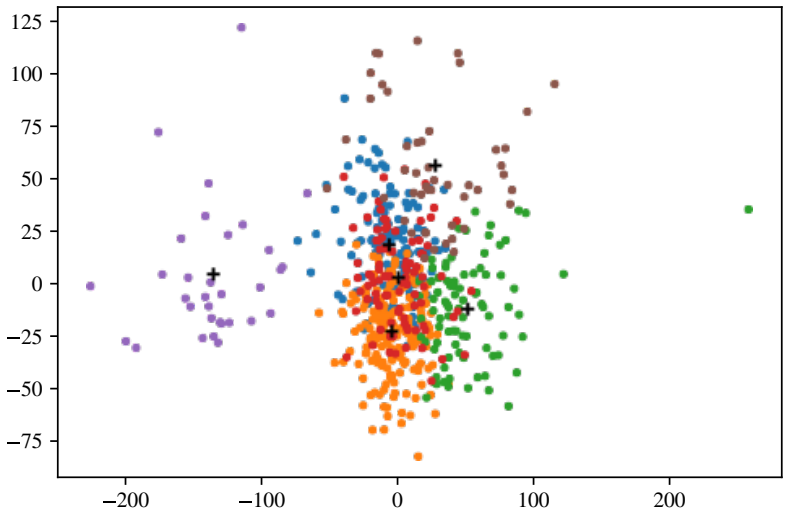


Figure 6.12: *K-means clustering on normalised gene expression data following principal components analysis. Ten principal components of variation were used for clustering. Crosses indicate cluster centres.*

6.7.1 On tumour heterogeneity

The concept of inter-tumour heterogeneity may sometimes refer to the differences between a primary cancer and its related metastases, or to inter-patient heterogeneity between similar tumours in different individuals, whereas intra-tumour heterogeneity refers to the spatial and temporal alterations within a single tumour (La Rosa *et al.,* 2019). Inter-tumour heterogeneity exists between tumours within the same individual (intra-patient), but this type of heterogeneity is less pronounced than the heterogeneity between tumours from different individuals (inter-patient) (Reuben *et al.,* 2017). Finally, as cancer arises and evolves through accumulation of genetic aberrations in individual cells (Hanahan & Weinberg, 2011), heterogeneity also exists within single tumours (intra-tumour heterogeneity). Analysis of the phylogenetic relationships between primary tumours and metastases has revealed that inter-tumour heterogeneity also arises in the same evolutionary manner as intra-tumour heterogeneity, with acquisition of metastatic potential likely being a late-evolving trait (Sanborn *et al.,* 2015). CRC tumours are among some of the most heavily mutated (PCAWG Consortium, 2020), and so management of the extreme heterogeneity this represents is an essential task when considering new strategies for classification.

6.8 GDC mutation calling

All somatic mutations are theoretically detectable in NGS data, given sufficient read depth. However in practice, due to noise in reads, considerable effort goes into confident variant calling. The Genomic Data Commons (GDC)'s pipeline (which is how TCGA data is processed) takes each tumour/normal pair and produces a VCF (Variant Call Format) file, using four different variant callers. Per-patient VCFs are then aggregated into a MAF (Mutation Annotation Format) file for each variant caller used. Publicly available somatic MAFs filter out lower quality calls and potential germline variants (Figure 6.13).

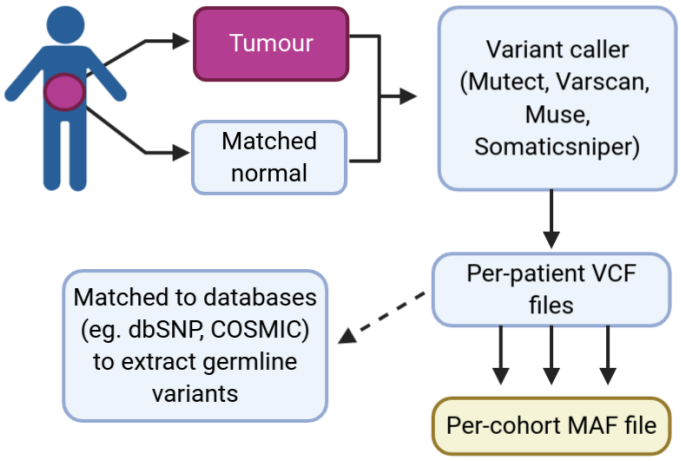


Figure 6.13: *Overview of the methodology used by the GDC for creation of public mutation annotation (MAF) files.*

The reasoning of the GDC behind using multiple variant callers is due to the lack of consensus on what the best strategy for variant calling is in the literature[1]. Because of this, a choice of four different variant callers is provided to users (Table 6.1). A review of these and other algorithms for variant calling was recently conducted by Xu, 2018.

[1] https://gdc.cancer.gov/about-gdc/variant-calling-gdc

Table 6.1: *The four variant callers used by the GDC and the core algorithm implemented by each for determining somatic mutations.*

Name	Core algorithm	Citation	Institute
MuSE	Markov chain model	Fan *et al.*, 2016	MD Anderson Cancer Center
MuTect	Allele frequency analysis	Cibulskis *et al.*, 2013	Broad Institute
VarScan	Heuristic threshold	Koboldt *et al.*, 2012	Washington University St. Louis
SomaticSniper	Joint genotype analysis	Larson *et al.*, 2012	Washington University St. Louis

6.9 Patient-specific network analysis appendices

Table 6.2: *The names, descriptions and UniProt IDs of the 102 additional proteins that were used to supplement the PRIMES HCT116 network, due to their relevance within the canonical EGFR signalling pathway.*

Symbol	Name	UniProtKB
1433B	14-3-3 protein beta/alpha	P31946
AP2A1	AP-2 complex subunit alpha-1	O95782
APLP2	Amyloid-like protein 2	Q06481
ARF4	ADP-ribosylation factor 4	P18085
ARHG7	Rho guanine nucleotide exchange factor 7	Q14155
ASAP1	Arf-GAP with SH3 domain ANK repeat and PH domain-containing protein 1	Q9ULH1
ATF1	Cyclic AMP-dependent TF ATF-1	P18846
BCAR1	Breast cancer anti-estrogen resistance protein 1	P56945
CBLB	E3 ubiquitin-protein ligase CBL-B	Q13191
CBLC	Signal transduction protein CBL-C	Q9ULV8
CBL	E3 ubiquitin-protein ligase CBL	P22681
CEAM1	Carcinoembryonic antigen-related cell adhesion molecule 1	P13688
CEBPB	CCAAT/enhancer-binding protein beta	P17676
CHIO	Beta-chimaerin	P52757
CRKL	Crk-like protein	P46109
CTND1	Catenin delta-1	O60716
CXA1	Gap junction alpha-1 protein	P17302
DP13A	DCC-interacting protein 13-alpha	Q9UKG1
DYN1	Dynamin-1	Q05193

EF1A1	Elongation factor 1-alpha 1	P68104
EGF	Pro-epidermal growth factor	P01133
ELF3	ETS-related TF Elf-3	P78545
EP15R	Epidermal growth factor receptor substrate 15-like 1	Q9UBC2
EPN1	Epsin-1	Q9Y6I3
EPS15	Epidermal growth factor receptor substrate 15	P42566
EPS8	Epidermal growth factor receptor kinase substrate 8	Q12929
ERBB2	Receptor tyrosine-protein kinase erbB-2	P04626
ERBB3	Receptor tyrosine-protein kinase erbB-3	P21860
ERBB4	Receptor tyrosine-protein kinase erbB-4	Q15303
ERRFI	ERBB receptor feedback inhibitor 1	Q9UJM3
FAK2	Protein-tyrosine kinase 2-beta	Q14289
FOXN1	Forkhead box protein N1	O15353
GIT1	ARF GTPase-activating protein GIT1	Q9Y2X7
GNDS	Ral guanine nucleotide dissociation stimulator	Q12967
GRB14	Growth factor receptor-bound protein 14	Q14449
HD	Huntingtin	P42858
HIP1	Huntingtin-interacting protein 1	O00291
ITCH	E3 ubiquitin-protein ligase Itchy homolog	Q96J02
JAK1	Tyrosine-protein kinase JAK1	P23458
JAK2	Tyrosine-protein kinase JAK2	O60674
JUND	TF jun-D	P17535
K1C17	Keratin type I cytoskeletal 17	Q04695
K1C18	Keratin type I cytoskeletal 18	P05783
K2C7	Keratin type II cytoskeletal 7	P08729
K2C8	Keratin type II cytoskeletal 8	P05787
KS6A5	Ribosomal protein S6 kinase alpha-5	O75582
KSR1	Kinase suppressor of Ras 1	Q8IVT5
LIMK1	LIM domain kinase 1	P53667
M3K11	Mitogen-activated protein kinase kinase kinase 11	Q16584
M3K1	Mitogen-activated protein kinase kinase kinase 1	Q13233
M3K2	Mitogen-activated protein kinase kinase kinase 2	Q9Y2U5
M3K3	Mitogen-activated protein kinase kinase kinase 3	Q99759
M3K4	Mitogen-activated protein kinase kinase kinase 4	Q9Y6R4
MP2K4	Dual specificity mitogen-activated protein kinase kinase 4	P45985
MP2K6	Dual specificity mitogen-activated protein kinase kinase 6	P52564
MP2K7	Dual specificity mitogen-activated protein kinase kinase 7	O14733
NCK2	Cytoplasmic protein NCK2	O43639

NDUAD	NADH dehydrogenase [ubiquinone] 1 alpha subcomplex subunit 13	Q9P0J0
P3C2B	Phosphatidylinositol 4-phosphate 3-kinase C2 domain-containing subunit beta	O00750
P55G	Phosphatidylinositol 3-kinase regulatory subunit gamma	Q92569
P85A	Phosphatidylinositol 3-kinase regulatory subunit alpha	P27986
P85B	Phosphatidylinositol 3-kinase regulatory subunit beta	O00459
PIPNA	Phosphatidylinositol transfer protein alpha isoform	Q00169
PK3CA	Phosphatidylinositol 4 5-bisphosphate 3-kinase catalytic subunit alpha isoform	P42336
PK3CB	Phosphatidylinositol 4 5-bisphosphate 3-kinase catalytic subunit beta isoform	P42338
PK3CD	Phosphatidylinositol 4 5-bisphosphate 3-kinase catalytic subunit delta isoform	O00329
PK3CG	Phosphatidylinositol 4 5-bisphosphate 3-kinase catalytic subunit gamma isoform	P48736
PKN2	Serine/threonine-protein kinase N2	Q16513
PLCG1	1-phosphatidylinositol 4 5-bisphosphate phosphodiesterase gamma-1	P19174
PLCG2	1-phosphatidylinositol 4 5-bisphosphate phosphodiesterase gamma-2	P16885
PLD1	Phospholipase D1	Q13393
PLD2	Phospholipase D2	O14939
PLEC	Plectin	Q15149
PLS1	Phospholipid scramblase 1	O15162
PTN11	Tyrosine-protein phosphatase non-receptor type 11	Q06124
RALA	Ras-related protein Ral-A	P11233
RASH	GTPase HRas	P01112
RASK	GTPase KRas	P01116
RASN	GTPase NRas	P01111
RBBP7	Histone-binding protein RBBP7	Q16576
REPS1	RalBP1-associated Eps domain-containing protein 1	Q96D71
REPS2	RalBP1-associated Eps domain-containing protein 2	Q8NFH8
RGS16	Regulator of G-protein signaling 16	O15492
ROCK1	Rho-associated protein kinase 1	Q13464
SH3G2	Endophilin-A1	Q99962
SHIP2	Phosphatidylinositol 3 4 5-trisphosphate 5-phosphatase 2	O15357
SIN3A	Paired amphipathic helix protein Sin3a	Q96ST3
SMD2	Small nuclear ribonucleoprotein Sm D2	P62316
SOCS3	Suppressor of cytokine signaling 3	O14543
SOS1	Son of sevenless homolog 1	Q07889

SOS2	Son of sevenless homolog 2	Q07890
SPY2	Protein sprouty homolog 2	O43597
STAT2	Signal transducer and activator of transcription 2	P52630
STXB1	Syntaxin-binding protein 1	P61764
TGIF1	Homeobox protein TGIF1	Q15583
TNIP1	TNFAIP3-interacting protein 1	Q15025
US6NL	USP6 N-terminal-like protein	Q92738
VAV2	Guanine nucleotide exchange factor VAV2	P52735
VAV3	Guanine nucleotide exchange factor VAV3	Q9UKW4
WASL	Neural Wiskott-Aldrich syndrome protein	O00401
WNK1	Serine/threonine-protein kinase WNK1	Q9H4A3
ZPR1	Zinc finger protein ZPR1	O75312

6.9.1 Excessive removal of ADH1C

When investigating the frequency of node removals, it became apparent a small number of genes were removed an excessive amount, notably ADH1C (alcohol dehydrogenase). This gene displays an extremely wide distribution of counts, and when visualising the thresholds it is apparent that many samples fall below the 3CPM detection limit. A plausible explanation for such a pattern may be copy number variation alterations. Checking the CNVs for ADH1C, I found nothing out of the ordinary, a range of 1-6. The literature on ADH1C suggests this pattern of expression may be to be expected in CRC - in one investigation, later stage carcinomas were found to have significant drops in ADH1C expression compared to normal tissues (Kropotova *et al.,* 2014).

6.9.2 Network and graph file formats

Many different network file formats exist, some more general, some extremely specific to certain domains. In terms of the data structures used, however, there are two main ways in which networks are represented: as an adjacency list, or an adjacency matrix.

Adjacency list

The adjacency list is a single list of binary pairs, in the format A interacts B. This format is perhaps the most common and intuitive representation of a network, and underlies common format such as Simple Interaction Format (SIF), Graphviz Dot format, and even more complex specifications like eXtensible Graph Markup and Modelling Language (XGMML). Such formats which are effectively plain text adjacency lists are generally easy to parse and edit manually, and are quite space efficient. From a computational perspective however the adjacency list format is not particularly efficient, and so many analysis tools will convert networks into a matrix format.

Adjacency matrix

A network with n nodes can be represented as an adjacency matrix M with dimensions $n \times n$, in which $M_{A,B}$ is zero if there is no edge between nodes A and B. In the case that there is an edge, $M_{A,B}$ is a value corresponding to the edge weight. This representation has many advantages.

Sparse matrix

If most nodes in a network are not connected (as is usually the case), most elements within the adjacency matrix will be zero. Storing these sparse matrices is extremely inefficient if there are a large number of nodes but very few edges, and so various compressed formats exist for storage of sparse matrices.

Information flow analysis with ITM probe

Multiple tools using random walks to simulate information transduction on PPI networks exist, such as Hierarchical HotNet, and ITM Probe. ITM probe was deemed the most suitable as its command-line version can easily output flow scores for each node - in comparison to other tools such as HotNet which simply output the subnetworks identified. I used the default dissipation probability (0.15), and the normalised-channel

mode, in which walkers are only counted if they reach sinks before dissipation (effectively normalising flow to sum to 1.0 across the sinks). By comparing information flow scores to the baseline EGFR-HQ network (in which no patient-specific modification were made), impact scores for each transcription factor were calculated. These impact scores were clustered and assessed for their association with patient survival using Cox regression analysis (Appendix Figure 6.14, B).

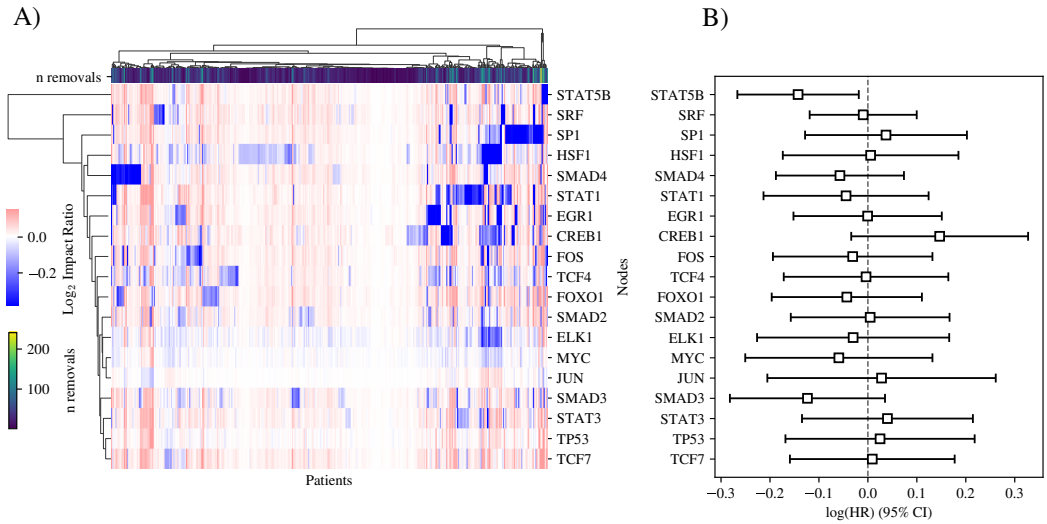


Figure 6.14: *A) Heatmap of the information flow impact scores for transcription factors downstream of EGFR. The normalised-channel model of ITM Probe was used to determine IF scores. B) Log hazard ratios (HR) from Cox regression analysis of transcription factor impact scores. FOSB, FOSL1, JUND, and STAT5A were excluded from this analysis due to low inter-patient variance.*

In terms of patient stratification (Appendix Figure 6.14, A), this was driven to a large degree by decreased impacts, with larger impact scores correlating mostly with higher numbers of removals. Cox regression analysis revealed that the impact score of signal transducer and activator of transcription 5B (STAT5B) was significantly positively associated with survival. In addition, IFA predicted that signalling to STAT5B was much more variable than the closely related STAT5A. This was primarily due to the neighbours of STAT5A also being transcription factors (and thus, random walks ceased before reaching the node). The predicted positive association of STAT5B impact score with patient survival was consistent with the role of the TF as a regulator of apoptosis, the expression of which has previously been identified as both being corre-

lated with TNM stage and to be more strongly associated with regulation of apoptosis than STAT5A in CRC (Du *et al.,* 2012). It has also been shown that STAT5B may be activated by epidermal growth factor in the presence of overexpressed *EGFR*, as is common in CRC (Kloth *et al.,* 2002). I investigated whether the impact score to STAT5B correlated with *EGFR* expression, but found that the correlation was not statistically significant (Pearson's r=0.07, p=0.12). *STAT5B* expression however was significantly correlated with *EGFR* expression (Pearson's r=0.23, p=4.7×10^{-8}).

While the ITM Probe's channel model was able to simulate information flow to sink nodes well, it was difficult to make use of the resulting information flow scores for other nodes in the network. When examining the distribution of all information flow scores to all nodes across all networks, I found that the scores were distributed very heavily near to zero, and that higher scores were extremely rare (Appendix Figure 6.18). While this was not unexpected, what was less desirable was that higher scores tended only to be found at the source and sink nodes (EGFR and the transcription factors) due to the design of the model. Another limitation was that node removals increased the total information flow to the rest of the network, resulting in impact scores for almost all nodes being increased from the baseline (Appendix Figure 6.16). This effect was also apparent when examining the distribution of all impact scores (Appendix Figure 6.15), in which the mean of the distribution skewed noticeably above zero.

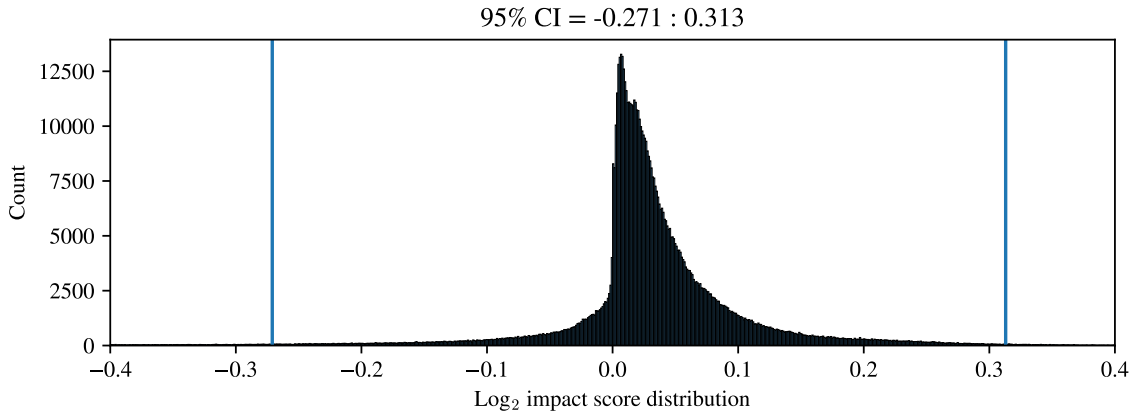


Figure 6.15: *Distribution of information flow scores for all nodes in 550 CRC patient-specific EGFR networks. To facilitate presentation, the x axis does not encompass the entire distribution due to a long tail of negative scores. While nodes close to the source (EGFR) tended to have higher information flow scores, most nodes have a very low score in comparison. The 95% confidence interval is annotated with vertical lines.*

Assessment of significant scores using this impact score distribution was difficult, as while there were many cases of removals resulting in extremely low and significant negative impact scores (for the removed nodes and any immediately downstream), the opposite was not true, as despite most scores being slightly above zero, only a small number of positive impact scores exceeded the confidence interval threshold.

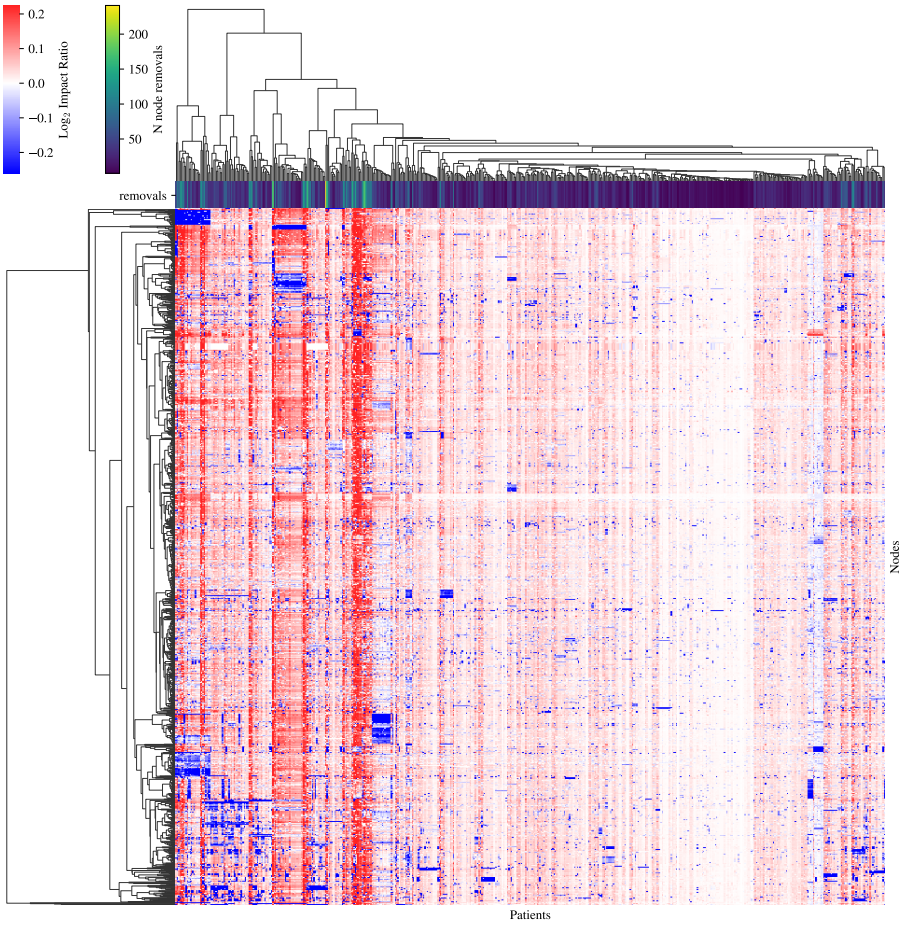


Figure 6.16: *Clustered impact scores of information flow to all network nodes across 550 patient-specific CRC networks as predicted using the normalised-channel model of ITM Probe. Node removal frequency per network is visualised along the top row.*

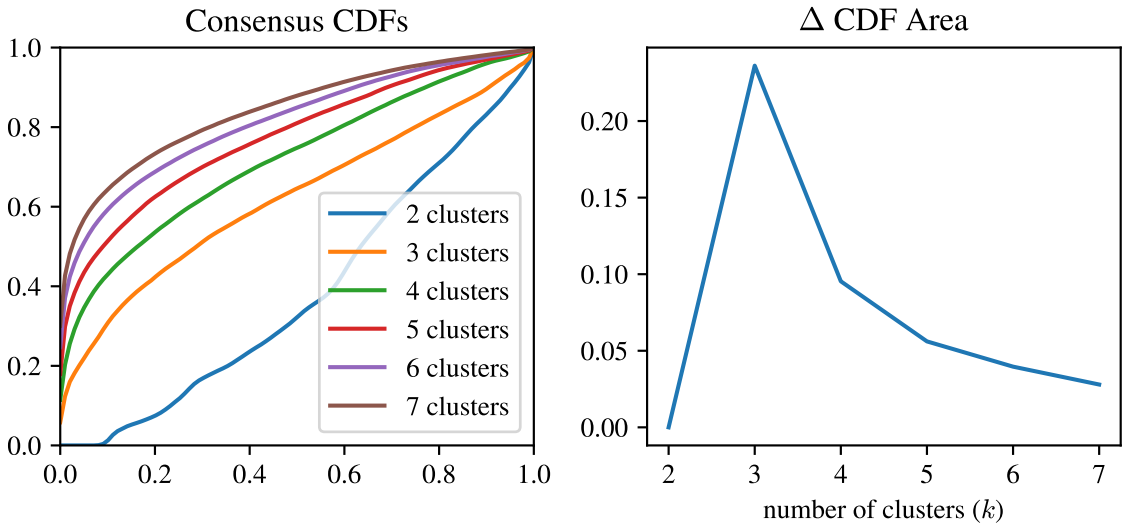


Figure 6.17: *The cumulative distribution of consensus scores (left) was used to calculate the change in area under the curve (right), with the peak change occurring at k=3 clusters.*

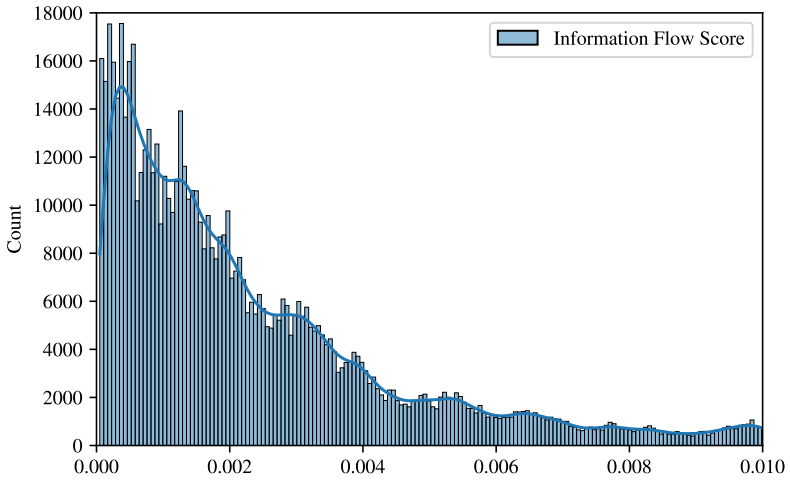


Figure 6.18: *Distribution of information flow scores for all nodes in 550 CRC patient-specific EGFR networks. To facilitate presentation the x axis does not encompass the entire distribution due to a long tail of scores. While nodes close to the source (EGFR) tended to have higher information flow scores, most nodes have a very low score in comparison.*

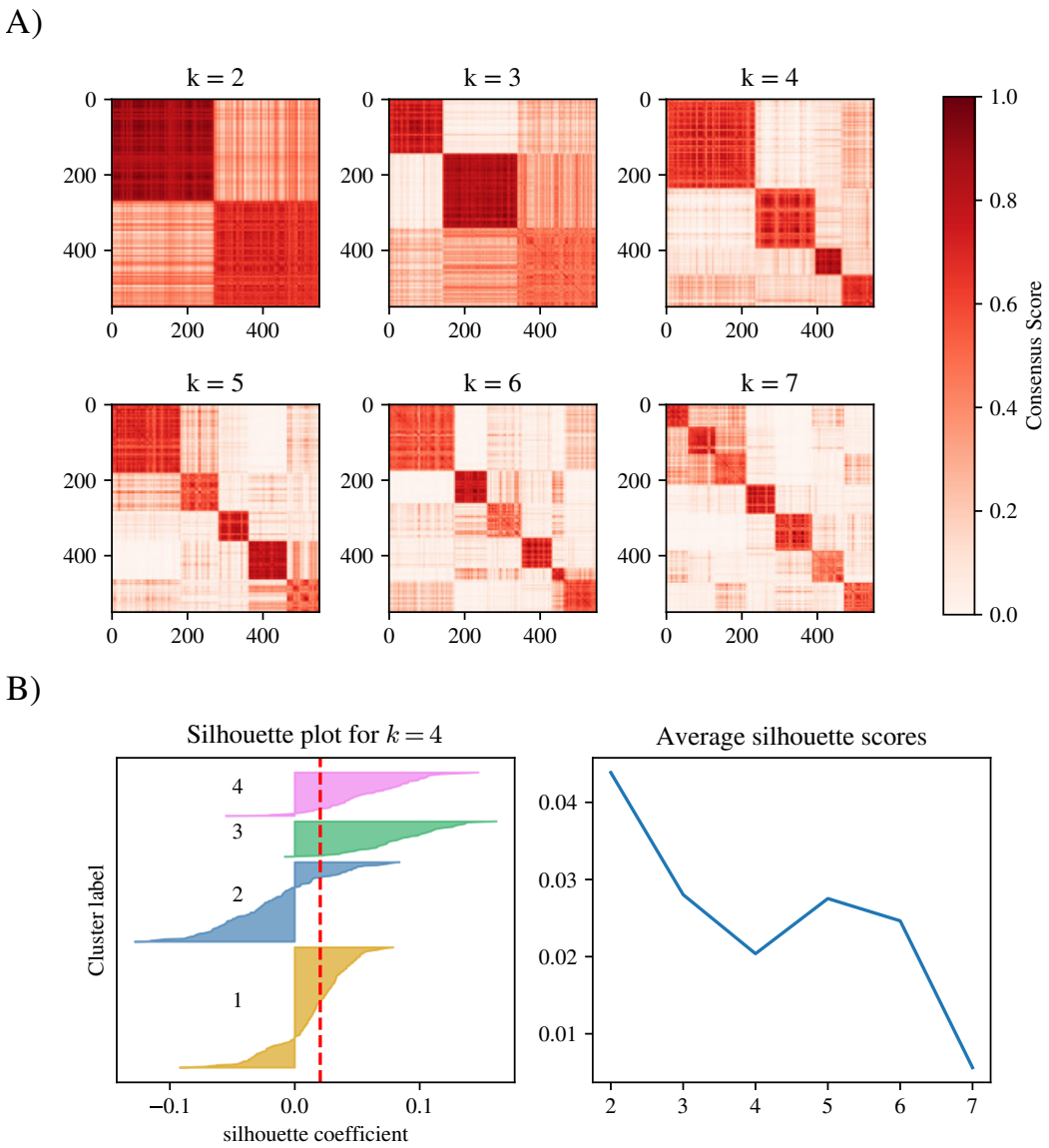


Figure 6.19: *A) Consensus matrices for full-network unsupervised hierarchical clustering of log_2SPC scores obtained with SIFFIN. B) Silhouette score plots used to decide on an optimal cluster number.*

6.10 PRIMES baits

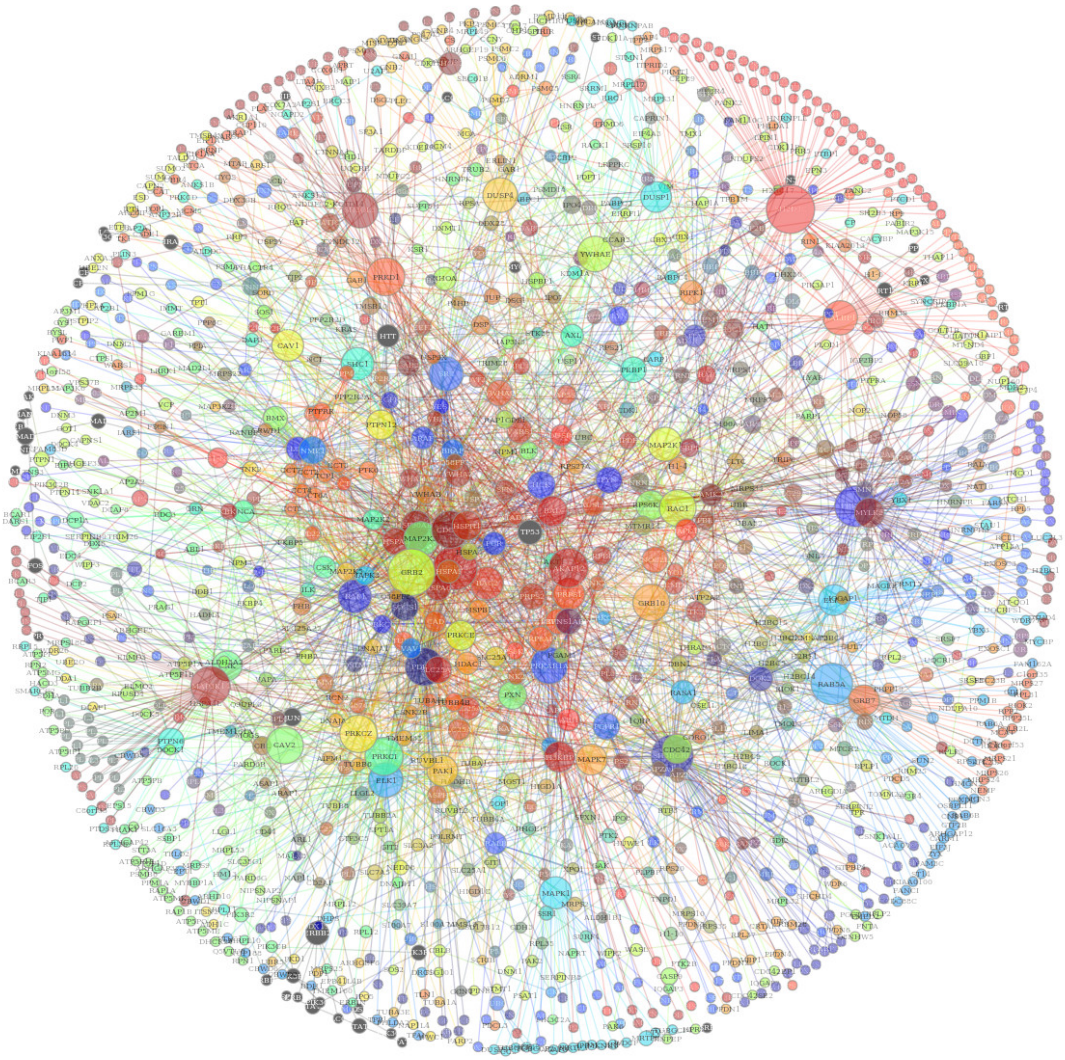


Figure 6.20: *Visualisation of the high quality EGFR network (EGFR-HQ) in which PRIMES baits are highlighted with unique colours. Nodes connected directly to baits are assigned the same colour, or a mix of colours if connected to multiple baits. Edges are coloured using a gradient between the two endpoints. Node size is scaled in proportion to betweenness centrality.*

Validation using transcription factor binding site analysis

Transcription factors bind specific DNA sequence motifs in order to modulate transcription, thus target genes of particular transcription factors are enriched for these motifs. Using HOMER (Heinz *et al.,* 2010) to detect the enrichment of motifs among up and down-regulated PSDE genes for each patient, I determined which transcription factors would most likely be responsible for regulating the up and down-regulated PSDE genes (Figure 6.21).

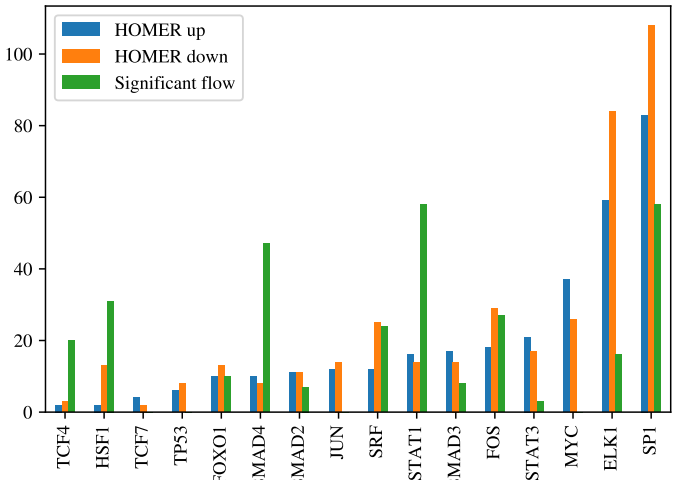


Figure 6.21: *Sums of patients in which each transcription factor is significantly enriched in up-regulated and down-regulated PSDE genes, as well as the number of significant impact scores found for each transcription factor.*

All of the transcription factors retained in the EGFR-HQ network were successfully matched to HOMER motifs. A Fisher's exact test was used to assess whether significant transcription factor flow corresponded to significant downregulation or upregulation of corresponding motifs. This analysis revealed that motifs for SP1 ($p=6.67 \times 10^{-4}$) and SMAD4 ($p=0.02$) were significantly enriched in up-regulated PSDE genes among patients with significant reductions in flow impact more often than would be expected by chance.

6.10.1 Additional network analysis of PSDE genes

While patient-specific differentially expressed (PSDE) genes were a useful basis for stratifying patients in a clinically relevant manner, far more was revealed when they were combined with pathway information to provide biological context. However, this approach is limited by the completeness of available pathway databases such as KEGG. One way to escape this limitation would be to use network analysis, for which I required a source of biological network information, i.e. experimentally determined protein-protein interactions (PPIs). Sourcing PPI data from IMEx, (Orchard *et al.,* 2012), I obtained 232,167 binary human PPIs. Further filtering of these PPIs to remove duplicate edges and retain only high-quality interactions (MI score >0.6) resulted in a network model of the human interactome consisting of 4,747 proteins and 10,845 interactions (Figure 6.22).

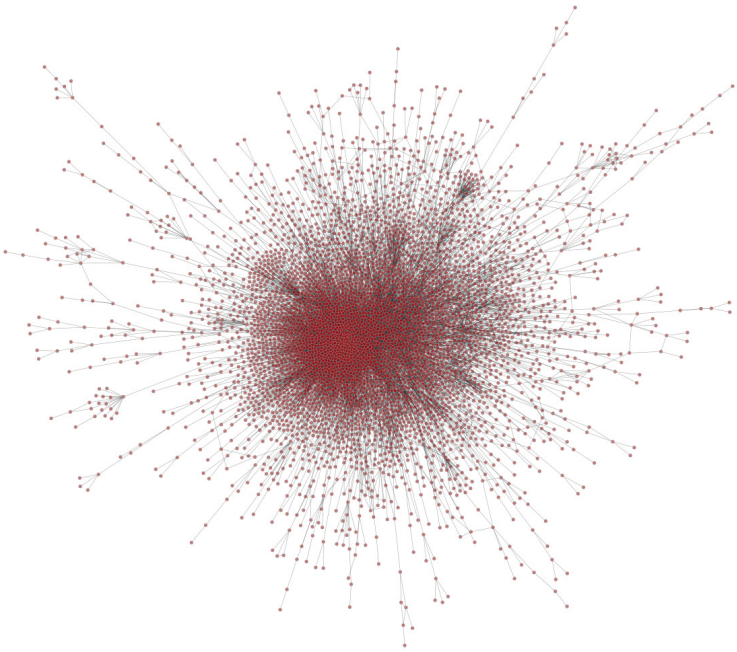


Figure 6.22: *Visualisation all high quality (MI score >0.6) human protein-protein interactions (PPIs) publicly available from the IMEx database. Each node (red) represents a unique protein, and each connected edge represents a PPI. The network layout visualisation was produced using the Scalable Force Directed Placement (sfdp) layout algorithm.*

PSDE genes in the wider interactome

Using a force-directed algorithm to produce a network layout of the interactome network resulted in a "hairball" layout (Figure 6.22). To create an alternative visualisation of the interactome network which would also be useful for analysis purposes, I applied the nested stochastic block algorithm (T. P. Peixoto, 2014) to perform unsupervised hierarchical partitioning, resulting in 16 network partitions.

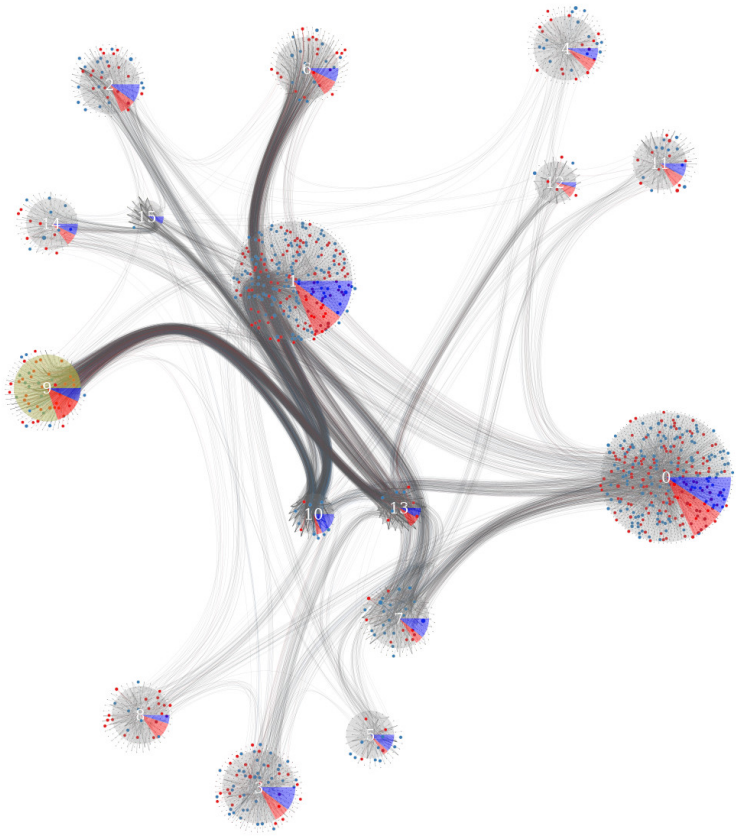


Figure 6.23: *Network of high-quality protein-protein interactions from IMEx, clustered into partitions using hierarchical partitioning. Visualisation and partitioning was performed using graph-tool (T. P. Peixoto, 2017). The proportion of frequently (95th percentile) up-regulated (red) and down-regulated (blue) PSDE genes is annotated as a pie chart for each partition. Clusters statistically enriched in up-regulated or down-regulated PSDE genes (Chi-squared tests, Benjamini-Hochberg FDR adjusted) were highlighted in yellow.*

I found that only one of the 16 network partitions contained a significantly larger number of up-regulated PSDEs than down-regulated PSDEs. The proportion of up-regulated to down-regulated PSDE genes was equal across the remaining 15 partitions. Next, pathway enrichment analysis was used to characterise each of the 16 partitions using KEGG pathways. Partition number 9 (the only partition found to have a significantly higher proportion of up-regulated PSDE genes) was found to be highly enriched for ribosomal proteins, and in fact most of the up-regulated PSDE genes in this network partition encoded ribosomal proteins. Some partitions were significantly enriched for many pathways, including the ERBB signalling pathway in the case of partition 2 (Figure 6.24), while no significant enrichments could be found for others.

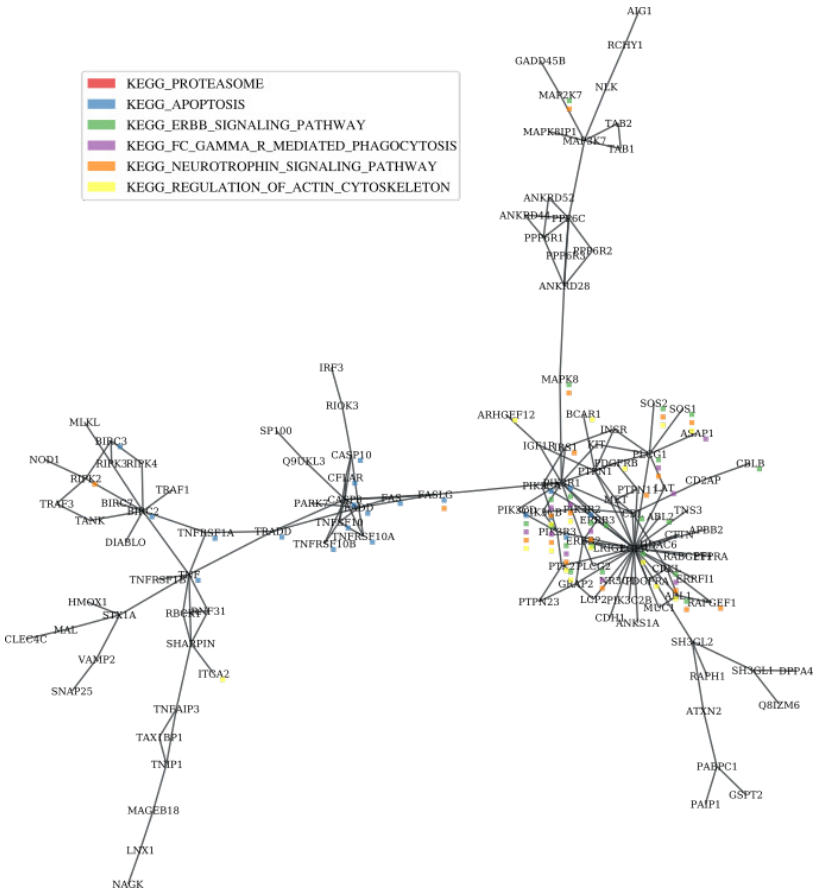


Figure 6.24: *Visualisation of the largest connected component of the partition 2 sub-network. Genes in significantly enriched KEGG pathways are annotated with coloured squares, including the KEGG ERBB signalling pathway (green) and apoptosis pathway (blue). Network layout was produced using the sfdp layout algorithm.*

Network module discovery using PSDE genes

Given that PSDE genes did not localise specifically to any one network partition, I took the approach of searching for network modules using PSDE genes as prior information for input into Hierarchical HotNet (Reyna *et al.*, 2018), a heat-diffusion based algorithm for identifying subnetworks significantly associated with sets of input nodes. Prior information scores were created separately for all up-regulated and down-regulated PSDE genes, using the number of times the gene was identified across the 550 patients in the TCGA CRC cohort. Hierarchical HotNet unexpectedly identified much larger modules when using down-regulated PSDE genes as input, compared to up-regulated PSDE genes. The largest module found for down-regulated PSDE genes was significantly enriched for many pathways, (Figure 6.25), including the KEGG cell cycle and DNA replication pathways.

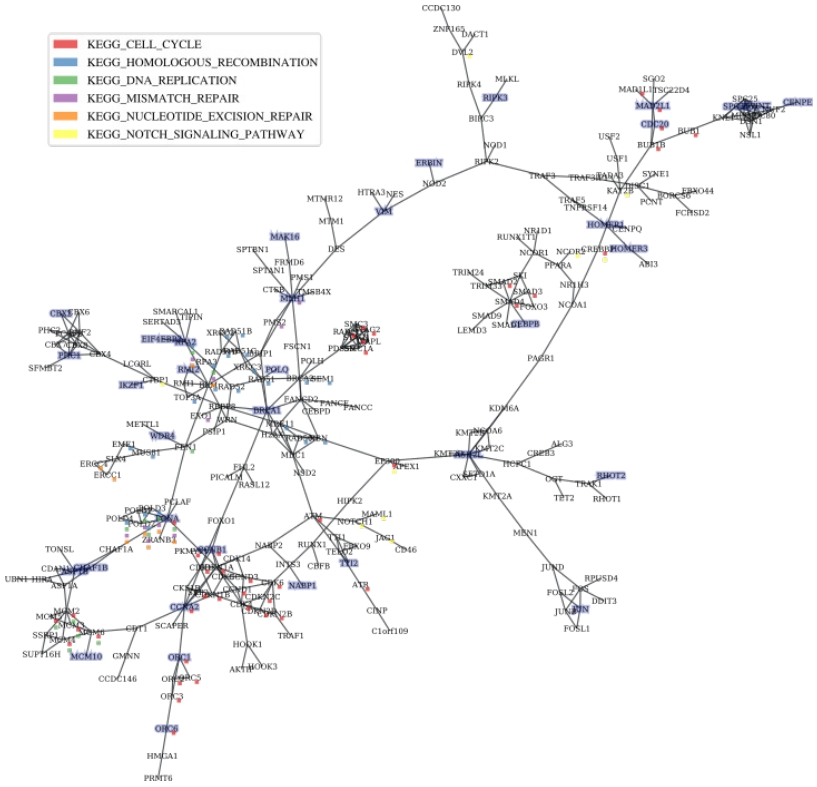


Figure 6.25: *The largest PPI module (253 nodes) discovered with Hierarchical HotNet using down-regulated PSDE genes as prior information. Genes frequently identified as down-regulated PSDE genes in different samples (95th percentile in the cohort of 550 patients) are highlighted in blue. The most significantly enriched KEGG pathways are indicated, with annotations on specific nodes indicating membership of the specified pathway.*

Patient-specific network modules

I next aimed to identify patient-specific network modules, using PSDE genes for each patient as weights for Hierarchical HotNet (if a gene was PSDE in a patient then weight was set to 1.0, otherwise it was 0.0). I found that the modules identified were quite sparse and differed substantially between patients. I retained modules with >10 vertices, and then used hierarchical clustering across all patients to identify clusters of similar modules (Figure 6.26). This revealed that there were relatively few modules with high similarity. From this clustering, I extracted the 3 most distinct clusters of nodes and performed pathway enrichment analysis.

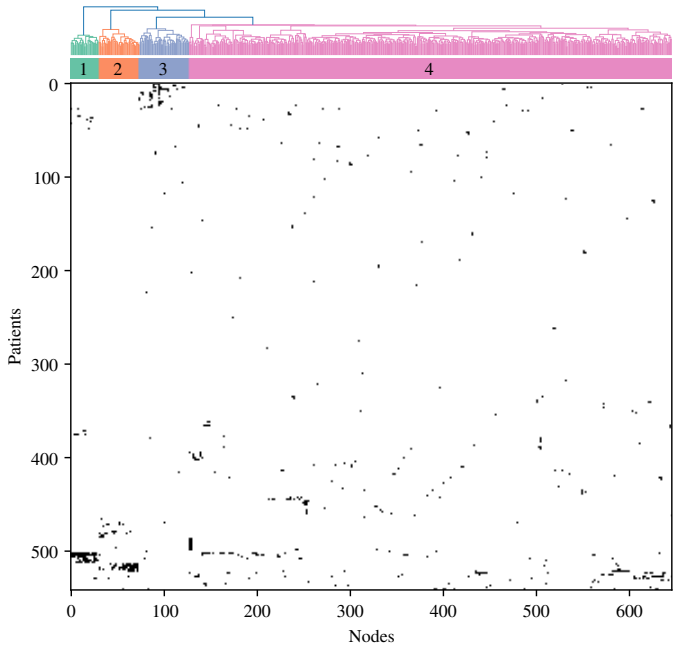


Figure 6.26: *Hierarchical clustering of nodes and patients based on patient-specific network modules. Modules were detected using Hierarchical HotNet with PSDE genes as input. Only modules with >10 nodes were retained. For each patient, all nodes found to be part of PSDE-informed modules are shown in black. Cluster annotation is shown below the dendrogram.*

Cluster 1 proteins were very strongly enriched for the KEGG ribosome pathway (p=0.0). This cluster was largely the same as partition 9 (Figure 6.23), mainly made up of ribosomal subunit proteins. I had previously determined that many up-regulated PSDEs were present in this cluster. Cluster 2 genes in comparison were enriched for roles in the cell cycle ($p=2 \times 10^{-6}$) and the P53 signalling pathway (p=0.017), while cluster 3 proteins were enriched for focal adhesion (p=0.009). I next examined whether the patients comprising these clusters corresponded to PSDE-informed cluster (PIC) subtypes, however there was no apparent overlap of these clusters.

References

1. Albert, R., Jeong, H. & Barabási, A.-L. Error and Attack Tolerance of Complex Networks. *Nature* **406,** 378–382. ISSN: 1476-4687 (July 2000).

2. Alexandrov, T. Spatial Metabolomics and Imaging Mass Spectrometry in the Age of Artificial Intelligence. *Annual Review of Biomedical Data Science,* 29 (2020).

3. Alhamdoosh, M. *et al.* Combining Multiple Tools Outperforms Individual Methods in Gene Set Enrichment Analyses. *Bioinformatics* **33,** 414–424. ISSN: 1367-4803 (Feb. 2017).

4. Allaoui, M., Kherfi, M. L. & Cheriet, A. *Considerably Improving Clustering Algorithms Using UMAP Dimensionality Reduction Technique: A Comparative Study* in *Image and Signal Processing* (eds El Moataz, A., Mammass, D., Mansouri, A. & Nouboud, F.) (Springer International Publishing, Cham, 2020), 317–325. ISBN: 978-3-030-51935-3.

5. Alstott, J., Bullmore, E. & Plenz, D. Powerlaw: A Python Package for Analysis of Heavy-Tailed Distributions. *PLOS ONE* **9,** e85777. ISSN: 1932-6203 (Jan. 2014).

6. Amado, R. G. *et al.* Wild-Type KRAS Is Required for Panitumumab Efficacy in Patients with Metastatic Colorectal Cancer. *JOURNAL OF CLINICAL ONCOLOGY* **26,** 1626–1634. ISSN: 0732-183X (2008).

7. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python Framework to Work with High-Throughput Sequencing Data. *Bioinformatics* **31,** 166–169. ISSN: 1367-4803 (Jan. 2015).

8. Andrew, A. S. *et al.* Risk Factors for Diagnosis of Colorectal Cancer at a Late Stage: A Population-Based Study. *Journal of General Internal Medicine* **33,** 2100–2105. ISSN: 1525-1497 (Dec. 2018).

9. Aran, D., Sirota, M. & Butte, A. J. Systematic Pan-Cancer Analysis of Tumour Purity. *Nature Communications* **6.** ISSN: 2041-1723 (Dec. 2015).

10. Arnold, M. *et al.* Global Patterns and Trends in Colorectal Cancer Incidence and Mortality. *Gut* **66,** 683–691. ISSN: 0017-5749, 1468-3288 (Apr. 2017).

11. Ashburner, M. *et al.* Gene Ontology: Tool for the Unification of Biology. *Nature Genetics* **25,** 25–29. ISSN: 1061-4036 (May 2000).

12. Bader, G. D., Cary, M. P. & Sander, C. Pathguide: A Pathway Resource List. *Nucleic Acids Research* **34,** D504–D506. ISSN: 0305-1048 (Jan. 2006).

13. Badic, B. *et al.* Prognostic Impact of Cancer Stem Cell Markers ABCB1, NEO1 and HIST1H2AE in Colorectal Cancer. *American Journal of Translational Research* **12,** 5797–5807. ISSN: 1943-8141 (Sept. 2020).

14. Barabási, A.-L. *Network Science* (2016).

15. Barabási, A.-L. & Albert, R. Emergence of Scaling in Random Networks. *Science* **286,** 509–512. ISSN: 0036-8075, 1095-9203 (Oct. 1999).

16. Barabási, A.-L., Gulbahce, N. & Loscalzo, J. Network Medicine: A Network-Based Approach to Human Disease. *Nature Reviews Genetics* **12,** 56–68. ISSN: 1471-0064 (Jan. 2011).

17. Barabási, A.-L. & Oltvai, Z. N. Network Biology: Understanding the Cell's Functional Organization. *Nature Reviews Genetics* **5,** 101–113. ISSN: 1471-0064 (Feb. 2004).

18. Baran, B. *et al.* Difference Between Left-Sided and Right-Sided Colorectal Cancer: A Focused Review of Literature. *Gastroenterology Research* **11,** 264–273. ISSN: 1918-2805 (Aug. 2018).

19. Barbie, D. A. *et al.* Systematic RNA Interference Reveals That Oncogenic KRAS -Driven Cancers Require TBK1. *Nature* **462,** 108–112. ISSN: 1476-4687 (Nov. 2009).

20. Barker, N. *et al.* Crypt Stem Cells as the Cells-of-Origin of Intestinal Cancer. *Nature* **457,** 608–611. ISSN: 1476-4687 (Jan. 2009).

21. Barrette, A. M., Bouhaddou, M. & Birtwistle, M. R. Integrating Transcriptomic Data with Mechanistic Systems Pharmacology Models for Virtual Drug Combination Trials. *ACS chemical neuroscience* **9,** 118–129. ISSN: 1948-7193 (Jan. 2018).

22. Basu, S. S. *et al.* Rapid MALDI Mass Spectrometry Imaging for Surgical Pathology. *npj Precision Oncology* **3,** 1–5. ISSN: 2397-768X (July 2019).

23. Battaglin, F. *et al.* The Role of Tumor Angiogenesis as a Therapeutic Target in Colorectal Cancer. *Expert Review of Anticancer Therapy* **18,** 251–266. ISSN: 1744-8328 (Mar. 2018).

24. Béal, J., Montagud, A., Traynard, P., Barillot, E. & Calzone, L. Personalization of Logical Models With Multi-Omics Data Allows Clinical Stratification of Patients. *Frontiers in Physiology* **9.** ISSN: 1664-042X (Jan. 2019).

25. Becht, E. *et al.* Dimensionality Reduction for Visualizing Single-Cell Data Using UMAP. *Nature Biotechnology* **37,** 38–44. ISSN: 1546-1696 (Jan. 2019).

26. Bedard, P. L., Hansen, A. R., Ratain, M. J. & Siu, L. L. Tumour Heterogeneity in the Clinic. *Nature* **501,** 355–364. ISSN: 1476-4687 (Sept. 2013).

27. Berg, H. C. *Random Walks in Biology* ISBN: 978-0-691-00064-0 (Princeton University Press, 1993).

28. Bergenstråhle, J., Larsson, L. & Lundeberg, J. Seamless Integration of Image and Molecular Analysis for Spatial Transcriptomics Workflows. *BMC Genomics* **21,** 482. ISSN: 1471-2164 (July 2020).

29. Bertrand, D. *et al.* Patient-Specific Driver Gene Prediction and Risk Assessment through Integrated Network Analysis of Cancer Omics Profiles. *Nucleic Acids Research* **43,** e44–e44. ISSN: 0305-1048 (Apr. 2015).

30. Bhagwani, A., Thompson, A. A. R. & Farkas, L. When Innate Immunity Meets Angiogenesis—The Role of Toll-Like Receptors in Endothelial Cells and Pulmonary Hypertension. *Frontiers in Medicine* **7.** ISSN: 2296-858X (2020).

31. Bhatia, A., Shatanof, R. A. & Bordoni, B. in *StatPearls* (StatPearls Publishing, Treasure Island (FL), 2020).

32. Bianconi, E. *et al.* An Estimation of the Number of Cells in the Human Body. *Annals of Human Biology* **40,** 463–471. ISSN: 0301-4460 (Nov. 2013).

33. Bisson, N. *et al.* Selected Reaction Monitoring Mass Spectrometry Reveals the Dynamics of Signaling through the GRB2 Adaptor. *Nature Biotechnology* **29,** 653–658. ISSN: 1546-1696 (July 2011).

34. Blake, S. J. *et al.* The Immunotoxicity, but Not Anti-Tumor Efficacy, of Anti-CD40 and Anti-CD137 Immunotherapies Is Dependent on the Gut Microbiota. *Cell Reports Medicine* **2,** 100464. ISSN: 2666-3791 (Dec. 2021).

35. Blanco-Calvo, M., Concha, Á., Figueroa, A., Garrido, F. & Valladares-Ayerbes, M. Colorectal Cancer Classification and Cell Heterogeneity: A Systems Oncology Approach. *International Journal of Molecular Sciences* **16,** 13610–13632 (June 2015).

36. Boehm, K. M., Khosravi, P., Vanguri, R., Gao, J. & Shah, S. P. Harnessing Multimodal Data Integration to Advance Precision Oncology. *Nature Reviews Cancer* **22,** 114–126. ISSN: 1474-1768 (Feb. 2022).

37. Boland, C. R. & Goel, A. Somatic Evolution of Cancer Cells. *Seminars in Cancer Biology. Somatic Evolution of Cancer Cells* **15,** 436–450. ISSN: 1044-579X (Dec. 2005).

38. Bornholdt, S. Less Is More in Modeling Large Genetic Networks. *Science* **310,** 449–451. ISSN: 0036-8075, 1095-9203 (Oct. 2005).

39. Bowler, E. H., Wang, Z. & Ewing, R. M. How Do Oncoprotein Mutations Rewire Protein–Protein Interaction Networks? *Expert Review of Proteomics* **12,** 449–455. ISSN: 1478-9450 (Sept. 2015).

40. Bozic, I. *et al.* Evolutionary Dynamics of Cancer in Response to Targeted Combination Therapy. *eLife* **2.** ISSN: 2050-084X (June 2013).

41. Bray, F. *et al.* Global Cancer Statistics 2018: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: A Cancer Journal for Clinicians* **68,** 394–424. ISSN: 1542-4863 (2018).

42. Breitkreutz, D., Hlatky, L., Rietman, E. & Tuszynski, J. A. Molecular Signaling Network Complexity Is Correlated with Cancer Patient Survivability. *Proceedings of the National Academy of Sciences* **109,** 9209–9212. ISSN: 0027-8424, 1091-6490 (June 2012).

43. Brereton, R. G. & Lloyd, G. R. Partial Least Squares Discriminant Analysis: Taking the Magic Away. *Journal of Chemometrics* **28,** 213–225. ISSN: 1099-128X (2014).

44. Brin, S. & Page, L. The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer Networks and ISDN Systems. Proceedings of the Seventh International World Wide Web Conference* **30,** 107–117. ISSN: 0169-7552 (Apr. 1998).

45. Brown, K. R. *et al.* NAViGaTOR: Network Analysis, Visualization and Graphing Toronto. *Bioinformatics* **25,** 3327–3329. ISSN: 1367-4803 (Dec. 2009).

46. Burgess, D. J. Spatial Transcriptomics Coming of Age. *Nature Reviews Genetics* **20,** 317–317. ISSN: 1471-0064 (June 2019).

47. Cadigan, K. M. & Waterman, M. L. TCF/LEFs and Wnt Signaling in the Nucleus. *Cold Spring Harbor Perspectives in Biology* **4.** ISSN: 1943-0264 (Nov. 2012).

48. Caiazza, F., Ryan, E. J., Doherty, G., Winter, D. C. & Sheahan, K. Estrogen Receptors and Their Implications in Colorectal Carcinogenesis. *Frontiers in Oncology* **5.** ISSN: 2234-943X (Feb. 2015).

49. Caldera, M., Buphamalai, P., Müller, F. & Menche, J. Interactome-Based Approaches to Human Disease. *Current Opinion in Systems Biology. • Mathematical Modelling • Mathematical Modelling, Dynamics of Brain Activity at the Systems Level • Clinical and Translational Systems Biology* **3,** 88–94. ISSN: 2452-3100 (June 2017).

50. *Cancer in Australia* tech. rep. 119 (Australian Institute of Health and Welfare, Canberra, 2019).

51. Cathomas, G. PIK3CA in Colorectal Cancer. *Frontiers in Oncology* **4,** 35. ISSN: 2234-943X (2014).

52. Cerami, E. *et al.* The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data. *Cancer Discovery* **2,** 401–404. ISSN: 2159-8274, 2159-8290 (May 2012).

53. Charloteaux, B. *et al.* in *Yeast Systems Biology* 197–213 (Humana Press, 2011). ISBN: 978-1-61779-172-7 978-1-61779-173-4.

54. Chatterjee, S. & Burns, T. F. Targeting Heat Shock Proteins in Cancer: A Promising Therapeutic Approach. *International Journal of Molecular Sciences* **18,** 1978. ISSN: 1422-0067 (Sept. 2017).

55. Chen, F. *et al.* Nanoscale Imaging of RNA with Expansion Microscopy. *Nature Methods* **13,** 679–684. ISSN: 1548-7105 (Aug. 2016).

56. Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S. & Zhuang, X. Spatially Resolved, Highly Multiplexed RNA Profiling in Single Cells. *Science* **348,** aaa6090. ISSN: 0036-8075, 1095-9203 (Apr. 2015).

57. Chen, M. *et al.* An Aberrant SREBP-dependent Lipogenic Program Promotes Metastatic Prostate Cancer. *Nature Genetics* **50,** 206–218. ISSN: 1546-1718 (Feb. 2018).

58. Chen, X., Xun, K., Chen, L. & Wang, Y. TNF-α, a Potent Lipid Metabolism Regulator. *Cell Biochemistry and Function* **27,** 407–416. ISSN: 1099-0844 (2009).

59. Christensen, T. D. *et al.* Associations between Primary Tumor RAS, BRAF and PIK3CA Mutation Status and Metastatic Site in Patients with Chemo-Resistant Metastatic Colorectal Cancer. *Acta Oncologica* **57,** 1057–1062. ISSN: 0284-186X (Aug. 2018).

60. Chuang, H.-Y., Lee, E., Liu, Y.-T., Lee, D. & Ideker, T. Network-based Classification of Breast Cancer Metastasis. *Molecular Systems Biology* **3,** 140. ISSN: 1744-4292, 1744-4292 (Jan. 2007).

61. Cibulskis, K. *et al.* Sensitive Detection of Somatic Point Mutations in Impure and Heterogeneous Cancer Samples. *Nature biotechnology* **31,** 213–219. ISSN: 1087-0156 (Mar. 2013).

62. Cieślik, M. & Chinnaiyan, A. M. Cancer Transcriptome Profiling at the Juncture of Clinical Translation. *Nature Reviews Genetics* **19,** 93–109. ISSN: 1471-0064 (Feb. 2018).

63. Cowen, L., Ideker, T., Raphael, B. J. & Sharan, R. Network Propagation: A Universal Amplifier of Genetic Associations. *Nature Reviews Genetics* **18,** 551–562. ISSN: 1471-0056, 1471-0064 (Sept. 2017).

64. Craig, S. G. *et al.* Immune Status Is Prognostic for Poor Survival in Colorectal Cancer Patients and Is Associated with Tumour Hypoxia. *British Journal of Cancer* **123,** 1280–1288. ISSN: 1532-1827 (Oct. 2020).

65. Cramér, H. *Mathematical Methods of Statistics* ISBN: 978-0-691-00547-8 (Princeton University Press, 1999).

66. Creixell, P., Palmeri, A., *et al.* Unmasking Determinants of Specificity in the Human Kinome. *Cell* **163,** 187–201. ISSN: 0092-8674 (Sept. 2015).

67. Creixell, P., Schoof, E. M., *et al.* Kinome-Wide Decoding of Network-Attacking Mutations Rewiring Cancer Signaling. *Cell* **163,** 202–217. ISSN: 0092-8674 (Sept. 2015).

68. Cremolini, C. *et al.* First-Line Chemotherapy for mCRC—a Review and Evidence-Based Algorithm. *Nature Reviews Clinical Oncology* **12,** 607–619. ISSN: 1759-4782 (Oct. 2015).

69. Crosetto, N., Bienko, M. & van Oudenaarden, A. Spatially Resolved Transcriptomics and Beyond. *Nature Reviews Genetics* **16,** 57–66. ISSN: 1471-0064 (Jan. 2015).

70. Csardi, G. & Nepusz, T. The Igraph Software Package for Complex Network Research, 10.

71. Dagogo-Jack, I. & Shaw, A. T. Tumour Heterogeneity and Resistance to Cancer Therapies. *Nature Reviews. Clinical Oncology* **15,** 81–94. ISSN: 1759-4782 (Feb. 2018).

72. Davidson-Pilon, C. *et al. CamDavidsonPilon/Lifelines: V0.21.0* Zenodo. Apr. 2019.

73. de Vries, N. L., Swets, M., Vahrmeijer, A. L., Hokland, M. & Kuppen, P. J. K. The Immunogenicity of Colorectal Cancer in Relation to Tumor Development and Treatment. *International Journal of Molecular Sciences* **17.** ISSN: 1422-0067 (June 2016).

74. del-Toro, N. *et al.* Capturing Variation Impact on Molecular Interactions in the IMEx Consortium Mutations Data Set. *Nature Communications* **10,** 10. ISSN: 2041-1723 (Jan. 2019).

75. Dettmer, K., Aronov, P. A. & Hammock, B. D. Mass Spectrometry-Based Metabolomics. *Mass Spectrometry Reviews* **26,** 51–78. ISSN: 1098-2787 (2007).

76. Devenport, S. N. & Shah, Y. M. Functions and Implications of Autophagy in Colon Cancer. *Cells* **8.** ISSN: 2073-4409 (Oct. 2019).

77. Dienstmann, R. *et al.* Consensus Molecular Subtypes and the Evolution of Precision Medicine in Colorectal Cancer. *Nature Reviews Cancer* **17,** 79–92. ISSN: 1474-1768 (Feb. 2017).

78. Dihal, A. A. *et al.* The Homeobox Gene MEIS1 Is Methylated in BRAF p.V600E Mutated Colon Tumors. *PLoS ONE* **8.** ISSN: 1932-6203 (Nov. 2013).

79. Dimitrakopoulos, C. *et al.* Network-Based Integration of Multi-Omics Data for Prioritizing Cancer Genes. *Bioinformatics* **34,** 2441–2448. ISSN: 1367-4803 (July 2018).

80. Dincer, C., Kaya, T., Keskin, O., Gursoy, A. & Tuncbag, N. 3D Spatial Organization and Network-Guided Comparison of Mutation Profiles in Glioblastoma Reveals Similarities across Patients. *PLOS Computational Biology* **15,** e1006789. ISSN: 1553-7358 (Sept. 2019).

81. Dittrich, M. T., Klau, G. W., Rosenwald, A., Dandekar, T. & Müller, T. Identifying Functional Modules in Protein–Protein Interaction Networks: An Integrated Exact Approach. *Bioinformatics* **24,** i223–i231. ISSN: 1367-4803 (July 2008).

82. Divaris, K. Fundamentals of Precision Medicine. *Compendium of continuing education in dentistry (Jamesburg, N.J. : 1995)* **38,** 30–32. ISSN: 1548-8578 (Sept. 2017).

83. Donaldson, G. P., Lee, S. M. & Mazmanian, S. K. Gut Biogeography of the Bacterial Microbiota. *Nature Reviews Microbiology* **14,** 20–32. ISSN: 1740-1534 (Jan. 2016).

84. Dong, X., Hao, Y., Wang, X. & Tian, W. LEGO: A Novel Method for Gene Set over-Representation Analysis by Incorporating Network-Based Gene Weights. *Scientific Reports* **6,** 18871. ISSN: 2045-2322 (Jan. 2016).

85. Drier, Y., Sheffer, M. & Domany, E. Pathway-Based Personalized Analysis of Cancer. *Proceedings of the National Academy of Sciences* **110,** 6388–6393. ISSN: 0027-8424, 1091-6490 (Apr. 2013).

86. Du, W. *et al.* STAT5 Isoforms Regulate Colorectal Cancer Cell Apoptosis via Reduction of Mitochondrial Membrane Potential and Generation of Reactive Oxygen Species. *Journal of Cellular Physiology* **227,** 2421–2429. ISSN: 1097-4652 (2012).

87. Dunne, P. D. *et al.* Challenging the Cancer Molecular Stratification Dogma: Intratumoral Heterogeneity Undermines Consensus Molecular Subtypes and Potential Diagnostic Value in Colorectal Cancer. *Clinical Cancer Research* **22,** 4095–4104. ISSN: 1078-0432, 1557-3265 (Aug. 2016).

88. Edge, S. B. *et al. AJCC Cancer Staging Manual* (Springer New York, 2010).

89. Eduati, F. *et al.* Patient-Specific Logic Models of Signaling Pathways from Screenings on Cancer Biopsies to Prioritize Personalized Combination Therapies. *Molecular Systems Biology* **16,** e8664. ISSN: 1744-4292 (Feb. 2020).

90. Ellis, J. D. *et al.* Tissue-Specific Alternative Splicing Remodels Protein-Protein Interaction Networks. *Molecular Cell* **46,** 884–892. ISSN: 1097-2765 (June 2012).

91. Erdős, P. & Rényi, A. On the Evolution of Random Graphs. *Publ. Math. Inst. Hung. Acad. Sci* **5,** 17–60 (1960).

92. Fabregat, A. *et al.* Reactome Pathway Analysis: A High-Performance in-Memory Approach. *BMC Bioinformatics* **18,** 142. ISSN: 1471-2105 (Mar. 2017).

93. Fabregat, A. *et al.* The Reactome Pathway Knowledgebase. *Nucleic Acids Research* **46,** D649–D655. ISSN: 1362-4962 (Jan. 2018).

94. Fan, Y. *et al.* MuSE: Accounting for Tumor Heterogeneity Using a Sample-Specific Error Model Improves Sensitivity and Specificity in Mutation Calling from Sequencing Data. *Genome Biology* **17,** 178. ISSN: 1474-760X (Aug. 2016).

95. Fearnhead, N. S., Britton, M. P. & Bodmer, W. F. The ABC of APC. *Human Molecular Genetics* **10,** 721–733. ISSN: 0964-6906 (Apr. 2001).

96. Fearon, E. R. & Vogelstein, B. A Genetic Model for Colorectal Tumorigenesis. *Cell* **61,** 759–767. ISSN: 0092-8674 (June 1990).

97. Feletto, E. *et al.* Trends in Colon and Rectal Cancer Incidence in Australia from 1982 to 2014: Analysis of Data on Over 375,000 Cases. *Cancer Epidemiology and Prevention Biomarkers* **28,** 83–90. ISSN: 1055-9965, 1538-7755 (Jan. 2019).

98. Fick, A. Ueber Diffusion. *Annalen der Physik* **170,** 59–86. ISSN: 1521-3889 (1855).

99. Fields, S. & Song, O.-k. A Novel Genetic System to Detect Protein-Protein Interactions. *Nature* **340,** 245–246. ISSN: 1476-4687 (July 1989).

100. Fodde, R. *et al.* Mutations in the APC Tumour Suppressor Gene Cause Chromosomal Instability. *Nature Cell Biology* **3,** 433–438. ISSN: 1465-7392 (Apr. 2001).

101. Fonseca, A. S. *et al.* ETV4 Plays a Role on the Primary Events during the Adenoma-Adenocarcinoma Progression in Colorectal Cancer. *BMC Cancer* **21,** 207. ISSN: 1471-2407 (Mar. 2021).

102. Foroushani, A. B. K., Brinkman, F. S. L. & Lynn, D. J. Pathway-GPS and SIGORA: Identifying Relevant Pathways Based on the over-Representation of Their Gene-Pair Signatures. *PeerJ* **1,** e229. ISSN: 2167-8359 (Dec. 2013).

103. Foroutan, M. *et al.* Single Sample Scoring of Molecular Phenotypes. *BMC Bioinformatics* **19,** 404. ISSN: 1471-2105 (Nov. 2018).

104. Freeman, L. C. A Set of Measures of Centrality Based on Betweenness. *Sociometry* **40,** 35–41. ISSN: 0038-0431 (1977).

105. Fridman, W. H. *et al.* Immune Infiltration in Human Cancer: Prognostic Significance and Disease Control. *Current Topics in Microbiology and Immunology* **344,** 1–24. ISSN: 0070-217X (2011).

106. Fruchterman, T. M. J. & Reingold, E. M. Graph Drawing by Force-Directed Placement. *Software: Practice and Experience* **21,** 1129–1164. ISSN: 1097-024X (1991).

107. Garcia-Alonso, L., Holland, C. H., Ibrahim, M. M., Turei, D. & Saez-Rodriguez, J. Benchmark and Integration of Resources for the Estimation of Human Transcription Factor Activities. *Genome Research* **29,** 1363–1375. ISSN: 1088-9051, 1549-5469 (Aug. 2019).

108. Gehlenborg, N. *et al.* Visualization of Omics Data for Systems Biology. *Nature Methods* **7,** S56–S68. ISSN: 1548-7105 (Mar. 2010).

109. Gehren, A. S., Rocha, M. R., de Souza, W. F. & Morgado-Díaz, J. A. Alterations of the Apical Junctional Complex and Actin Cytoskeleton and Their Role in Colorectal Cancer Progression. *Tissue Barriers* **3.** ISSN: 2168-8362 (Feb. 2015).

110. George, R. A. *et al.* Analysis of Protein Sequence and Interaction Data for Candidate Disease Gene Prediction. *Nucleic Acids Research* **34,** e130–e130. ISSN: 0305-1048 (Oct. 2006).

111. Gerlinger, M. *et al.* Intratumor Heterogeneity and Branched Evolution Revealed by Multiregion Sequencing. *New England Journal of Medicine* **366,** 883–892. ISSN: 0028-4793 (Mar. 2012).

112. Giannakis, M. *et al.* Genomic Correlates of Immune-Cell Infiltrates in Colorectal Carcinoma. *Cell Reports* **15,** 857–865. ISSN: 2211-1247 (Apr. 2016).

113. Gingras, A.-C., Gstaiger, M., Raught, B. & Aebersold, R. Analysis of Protein Complexes Using Mass Spectrometry. *Nature Reviews Molecular Cell Biology* **8,** 645–654. ISSN: 1471-0080 (Aug. 2007).

114. Giordano, G., Remo, A., Porras, A. & Pancione, M. Immune Resistance and EGFR Antagonists in Colorectal Cancer. *Cancers* **11.** ISSN: 2072-6694 (July 2019).

115. Giubellino, A., Burke, T. R. & Bottaro, D. P. Grb2 Signaling in Cell Motility and Cancer. *Expert opinion on therapeutic targets* **12,** 1021–1033. ISSN: 1472-8222 (Aug. 2008).

116. Glebov, O. K. *et al.* Distinguishing Right from Left Colon by the Pattern of Gene Expression. *Cancer Epidemiology and Prevention Biomarkers* **12,** 755–762. ISSN: 1055-9965, 1538-7755 (Aug. 2003).

117. Goeman, J. J. & Bühlmann, P. Analyzing Gene Expression Data in Terms of Gene Sets: Methodological Issues. *Bioinformatics* **23,** 980–987. ISSN: 1367-4803 (Apr. 2007).

118. Goenawan, I. H., Bryan, K. & Lynn, D. J. DyNet: Visualization and Analysis of Dynamic Molecular Interaction Networks. *Bioinformatics* **32,** 2713–2715. ISSN: 1367-4803 (Sept. 2016).

119. Goldman, M. J. *et al.* Visualizing and Interpreting Cancer Genomics Data via the Xena Platform. *Nature Biotechnology* **38,** 675–678. ISSN: 1546-1696 (June 2020).

120. Gonzalez-Angulo, A. M., Hennessy, B. T. & Mills, G. B. Future of Personalized Medicine in Oncology: A Systems Biology Approach. *Journal of Clinical Oncology* **28,** 2777–2783. ISSN: 0732-183X (June 2010).

121. Greaves, M. & Maley, C. C. CLONAL EVOLUTION IN CANCER. *Nature* **481,** 306–313. ISSN: 0028-0836 (Jan. 2012).

122. Groden, J. *et al.* Identification and Characterization of the Familial Adenomatous Polyposis Coli Gene. *Cell* **66,** 589–600. ISSN: 00928674 (Aug. 1991).

123. Grossman, R. L. *et al.* Toward a Shared Vision for Cancer Genomic Data. *New England Journal of Medicine* **375,** 1109–1112. ISSN: 0028-4793 (Sept. 2016).

124. Guinney, J. *et al.* The Consensus Molecular Subtypes of Colorectal Cancer. *Nature Medicine* **21,** 1350–1356. ISSN: 1546-170X (Nov. 2015).

125. Gursoy, A., Keskin, O. & Nussinov, R. Topological Properties of Protein Interaction Networks from a Structural Perspective. *Biochemical Society Transactions* **36,** 1398–1403. ISSN: 0300-5127, 1470-8752 (Dec. 2008).

126. Ha, M. J. *et al.* Personalized Integrated Network Modeling of the Cancer Proteome Atlas. *Scientific Reports* **8,** 14924. ISSN: 2045-2322 (Oct. 2018).

127. Hafemeister, C. & Satija, R. Normalization and Variance Stabilization of Single-Cell RNA-seq Data Using Regularized Negative Binomial Regression. *Genome Biology* **20,** 296. ISSN: 1474-760X (Dec. 2019).

128. Hagberg, A., Swart, P. & S Chult, D. *Exploring Network Structure, Dynamics, and Function Using Networkx* tech. rep. LA-UR-08-05495; LA-UR-08-5495 (Los Alamos National Lab. (LANL), Los Alamos, NM (United States), Jan. 2008).

129. Haider, S. *et al.* Systematic Assessment of Tumor Purity and Its Clinical Implications. *JCO Precision Oncology,* 995–1005 (Sept. 2020).

130. Hamosh, A., Scott, A. F., Amberger, J., Valle, D. & McKusick, V. A. Online Mendelian Inheritance In Man (OMIM). *Human Mutation* **15,** 57–61. ISSN: 1098-1004 (2000).

131. Hanahan, D. & Weinberg, R. A. Hallmarks of Cancer: The Next Generation. *Cell* **144,** 646–674. ISSN: 0092-8674, 1097-4172 (Mar. 2011).

132. Hanahan, D. & Weinberg, R. A. The Hallmarks of Cancer. *Cell* **100,** 57–70. ISSN: 0092-8674, 1097-4172 (Jan. 2000).

133. Hansen, I. O. & Jess, P. Possible Better Long-Term Survival in Left versus Right-Sided Colon Cancer – a Systematic Review, 6 (2012).

134. Hänzelmann, S., Castelo, R. & Guinney, J. GSVA: Gene Set Variation Analysis for Microarray and RNA-Seq Data. *BMC Bioinformatics* **14,** 7. ISSN: 1471-2105 (Jan. 2013).

135. Harris, C. R. *et al.* Array Programming with NumPy. *Nature* **585,** 357–362. ISSN: 1476-4687 (Sept. 2020).

136. Hartwell, L. H., Hopfield, J. J., Leibler, S. & Murray, A. W. From Molecular to Modular Cell Biology. *Nature* **402,** C47–C52. ISSN: 1476-4687 (Dec. 1999).

137. Hastings, J. F., O'Donnell, Y. E. I., Fey, D. & Croucher, D. R. Applications of Personalised Signalling Network Models in Precision Oncology. *Pharmacology & Therapeutics* **212,** 107555. ISSN: 0163-7258 (Aug. 2020).

138. He, X. & Zhang, J. Why Do Hubs Tend to Be Essential in Protein Networks? *PLOS Genetics* **2,** e88. ISSN: 1553-7404 (June 2006).

139. Heinz, S. *et al.* Simple Combinations of Lineage-Determining Transcription Factors Prime Cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Molecular Cell* **38,** 576–589. ISSN: 1097-4164 (May 2010).

140. Hermjakob, H. *et al.* IntAct: An Open Source Molecular Interaction Database. *Nucleic Acids Research* **32,** D452–D455. ISSN: 0305-1048 (Jan. 2004).

141. Hoadley, K. A. *et al.* Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. *Cell* **173,** 291–304.e6. ISSN: 0092-8674 (Apr. 2018).

142. Hofree, M., Shen, J. P., Carter, H., Gross, A. & Ideker, T. Network-Based Stratification of Tumor Mutations. *Nature Methods* **10,** 1108–1115. ISSN: 1548-7105 (Nov. 2013).

143. Hong, G., Zhang, W., Li, H., Shen, X. & Guo, Z. Separate Enrichment Analysis of Pathways for Up- and Downregulated Genes. *Journal of the Royal Society Interface* **11.** ISSN: 1742-5689 (Mar. 2014).

144. Huang, B., Bates, M. & Zhuang, X. Super-Resolution Fluorescence Microscopy. *Annual Review of Biochemistry* **78,** 993–1016. ISSN: 0066-4154, 1545-4509 (June 2009).

145. Huang, Z. *et al.* SALMON: Survival Analysis Learning With Multi-Omics Neural Networks on Breast Cancer. *Frontiers in Genetics* **10,** 166. ISSN: 1664-8021 (2019).

146. Hubbard, T. *et al.* The Ensembl Genome Database Project. *Nucleic Acids Research* **30,** 38–41. ISSN: 0305-1048 (Jan. 2002).

147. Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Computing in Science Engineering* **9,** 90–95. ISSN: 1558-366X (May 2007).

148. Huttlin, E. L. *et al.* Architecture of the Human Interactome Defines Protein Communities and Disease Networks. *Nature* **545,** 505–509. ISSN: 1476-4687 (May 2017).

149. Huttlin, E. L. *et al.* The BioPlex Network: A Systematic Exploration of the Human Interactome. *Cell* **162,** 425–440. ISSN: 0092-8674 (July 2015).

150. Ihnatova, I., Popovici, V. & Budinska, E. A Critical Comparison of Topology-Based Pathway Analysis Methods. *PLOS ONE* **13,** e0191154. ISSN: 1932-6203 (Jan. 2018).

151. Itzkovitz, S. & van Oudenaarden, A. Validating Transcripts with Probes and Imaging Technology. *Nature Methods* **8,** S12–S19. ISSN: 1548-7105 (Apr. 2011).

152. Jassal, B. *et al.* The Reactome Pathway Knowledgebase. *Nucleic Acids Research* **48,** D498–D503. ISSN: 1362-4962 (Jan. 2020).

153. Ji, R., Ren, Q., Bai, S., Wang, Y. & Zhou, Y. Prognostic Significance of Pretreatment Plasma Fibrinogen in Patients with Hepatocellular and Pancreatic Carcinomas: A Meta-Analysis. *Medicine (United States)* **97** (2018).

154. Jin, L. *et al.* Pathway-Based Analysis Tools for Complex Diseases: A Review. *Genomics, Proteomics & Bioinformatics. Special Issue: Translational Omics* **12,** 210–220. ISSN: 1672-0229 (Oct. 2014).

155. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting Batch Effects in Microarray Expression Data Using Empirical Bayes Methods. *Biostatistics* **8,** 118–127. ISSN: 1465-4644 (Jan. 2007).

156. Jonsson, P. F. & Bates, P. A. Global Topological Features of Cancer Proteins in the Human Interactome. *Bioinformatics* **22,** 2291–2297. ISSN: 1367-4803 (Sept. 2006).

157. Junttila, M. R. & de Sauvage, F. J. Influence of Tumour Micro-Environment Heterogeneity on Therapeutic Response. *Nature* **501,** 346–354. ISSN: 1476-4687 (Sept. 2013).

158. Kahles, A. *et al.* Comprehensive Analysis of Alternative Splicing Across Tumors from 8,705 Patients. *Cancer Cell* **34,** 211–224.e6. ISSN: 1535-6108, 1878-3686 (Aug. 2018).

159. Kalluri, R. & Weinberg, R. A. The Basics of Epithelial-Mesenchymal Transition. *The Journal of Clinical Investigation* **119,** 1420–1428. ISSN: 0021-9738 (June 2009).

160. Kandoth, C. *et al.* Mutational Landscape and Significance across 12 Major Cancer Types. *Nature* **502,** 333–339. ISSN: 1476-4687 (Oct. 2013).

161. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* **28,** 27–30. ISSN: 0305-1048 (Jan. 2000).

162. Kawai, M., Masuda, A. & Kuwana, M. A CD40–CD154 Interaction in Tissue Fibrosis. *Arthritis & Rheumatism* **58,** 3562–3573. ISSN: 1529-0131 (2008).

163. Kawakami, H., Zaanan, A. & Sinicrope, F. A. MSI Testing and Its Role in the Management of Colorectal Cancer. *Current treatment options in oncology* **16,** 30. ISSN: 1527-2729 (July 2015).

164. Ke, R. *et al.* In Situ Sequencing for RNA Analysis in Preserved Tissue and Cells. *Nature Methods* **10,** 857–860. ISSN: 1548-7091 (July 2013).

165. Keiser, M. J. *et al.* Predicting New Molecular Targets for Known Drugs. *Nature* **462,** 175–181. ISSN: 1476-4687 (Nov. 2009).

166. Kennedy, S. A. *et al.* Extensive Rewiring of the EGFR Network in Colorectal Cancer Cells Expressing Transforming Levels of KRASG13D. *Nature Communications* **11,** 499. ISSN: 2041-1723 (Dec. 2020).

167. Kerrien, S. *et al.* The IntAct Molecular Interaction Database in 2012. *Nucleic Acids Research* **40,** D841–D846. ISSN: 0305-1048 (Jan. 2012).

168. Khalek, F. J. A., Gallicano, G. I. & Mishra, L. Colon Cancer Stem Cells. *Gastrointestinal Cancer Research : GCR,* S16–S23. ISSN: 1934-7820 (2010).

169. Khatri, P., Sirota, M. & Butte, A. J. Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges. *PLOS Computational Biology* **8,** e1002375. ISSN: 1553-7358 (Feb. 2012).

170. Kido, T. & Lau, Y.-F. C. Roles of the Y Chromosome Genes in Human Cancers. *Asian Journal of Andrology* **17,** 373–380. ISSN: 1008-682X (2015).

171. Kierans, S. J. & Taylor, C. T. Regulation of Glycolysis by the Hypoxia-Inducible Factor (HIF): Implications for Cellular Physiology. *The Journal of Physiology* **599,** 23–37. ISSN: 1469-7793 (2021).

172. Kim, Y.-A., Przytycki, J. H., Wuchty, S. & Przytycka, T. M. Modeling Information Flow in Biological Networks. *Physical Biology* **8,** 035012. ISSN: 1478-3975 (2011).

173. Klein, J. P., Logan, B., Harhoff, M. & Andersen, P. K. Analyzing Survival Curves at a Fixed Point in Time. *Statistics in Medicine* **26,** 4505–4519. ISSN: 02776715, 10970258 (Oct. 2007).

174. Kloth, M. T., Catling, A. D. & Silva, C. M. Novel Activation of STAT5b in Response to Epidermal Growth Factor *. *Journal of Biological Chemistry* **277,** 8693–8701. ISSN: 0021-9258, 1083-351X (Mar. 2002).

175. Knickelbein, K. & Zhang, L. Mutant KRAS as a Critical Determinant of the Therapeutic Response of Colorectal Cancer. *Genes & Diseases* **2,** 4–12. ISSN: 2352-3042 (Mar. 2015).

176. Knijn, N. *et al.* KRAS Mutation Analysis: A Comparison between Primary Tumours and Matched Liver Metastases in 305 Colorectal Cancer Patients. *British Journal of Cancer* **104,** 1020–1026. ISSN: 1532-1827 (Mar. 2011).

177. Koboldt, D. C. *et al.* VarScan 2: Somatic Mutation and Copy Number Alteration Discovery in Cancer by Exome Sequencing. *Genome Research* **22,** 568–576. ISSN: 1088-9051, 1549-5469 (Mar. 2012).

178. Koh, H. W. L. *et al.* iOmicsPASS: Network-Based Integration of Multiomics Data for Predictive Subnetwork Discovery. *npj Systems Biology and Applications* **5,** 1–10. ISSN: 2056-7189 (July 2019).

179. Koh, J. & Kim, M. J. Introduction of a New Staging System of Breast Cancer for Radiologists: An Emphasis on the Prognostic Stage. *Korean Journal of Radiology* **20,** 69–82. ISSN: 1229-6929 (Jan. 2019).

180. Köhler, S., Bauer, S., Horn, D. & Robinson, P. N. Walking the Interactome for Prioritization of Candidate Disease Genes. *The American Journal of Human Genetics* **82,** 949–958. ISSN: 0002-9297, 1537-6605 (Apr. 2008).

181. Korsunsky, I. *et al.* Fast, Sensitive and Accurate Integration of Single-Cell Data with Harmony. *Nature Methods* **16,** 1289–1296. ISSN: 1548-7105 (Dec. 2019).

182. Koyutürk, M., Subramaniam, S. & Grama, A. in *Functional Coherence of Molecular Networks in Bioinformatics* (eds Koyutürk, M., Subramaniam, S. & Grama, A.) 1–13 (Springer, New York, NY, 2012). ISBN: 978-1-4614-0320-3.

183. Kropotova, E. S. *et al.* Altered Expression of Multiple Genes Involved in Retinoic Acid Biosynthesis in Human Colorectal Cancer. *Pathology & Oncology Research* **20,** 707–717. ISSN: 1532-2807 (July 2014).

184. Kuleshov, M. V. *et al.* Enrichr: A Comprehensive Gene Set Enrichment Analysis Web Server 2016 Update. *Nucleic Acids Research* **44,** W90–W97. ISSN: 0305-1048 (July 2016).

185. Kumara, H. S. *et al.* P-Cadherin (CDH3) Is Overexpressed in Colorectal Tumors and Has Potential as a Serum Marker for Colorectal Cancer Monitoring. *Oncoscience* **4,** 139–147. ISSN: 2331-4737 (Oct. 2017).

186. Kurayoshi, K. *et al. The Key Role of E2F in Tumor Suppression through Specific Regulation of Tumor Suppressor Genes in Response to Oncogenic Changes* ISBN: 978-953-51-3868-6 (IntechOpen, 2018).

187. Kwaan, M. R. & Jones-Webb, R. Colorectal Cancer Screening in Black Men: Recommendations for Best Practices. *American Journal of Preventive Medicine* **55,** S95–S102. ISSN: 0749-3797, 1873-2607 (Nov. 2018).

188. La Rosa, S., Rubbia-Brandt, L., Scoazec, J.-Y. & Weber, A. Editorial: Tumor Heterogeneity. *Frontiers in Medicine* **6.** ISSN: 2296-858X (2019).

189. Ladiges, P., Evans, B., Knox, B. & Saint, R. *Biology: An Australian Focus* ISBN: 978-0-07-027440-2 (McGraw-Hill, 2010).

190. Larson, D. E. *et al.* SomaticSniper: Identification of Somatic Point Mutations in Whole Genome Sequencing Data. *Bioinformatics* **28,** 311–317. ISSN: 1367-4803 (Feb. 2012).

191. Lauss, M. *et al.* Monitoring of Technical Variation in Quantitative High-Throughput Datasets. *Cancer Informatics* **12,** CIN.S12862. ISSN: 1176-9351 (Jan. 2013).

192. Le, D. T. *et al.* PD-1 Blockade in Tumors with Mismatch-Repair Deficiency. *New England Journal of Medicine* **372,** 2509–2520. ISSN: 0028-4793 (June 2015).

193. Lee, E., Chuang, H.-Y., Kim, J.-W., Ideker, T. & Lee, D. Inferring Pathway Activity toward Precise Disease Classification. *PLOS Computational Biology* **4,** e1000217. ISSN: 1553-7358 (Nov. 2008).

194. Lee, J. H. *et al.* Highly Multiplexed Subcellular RNA Sequencing in Situ. *Science* **343,** 1360–1363. ISSN: 0036-8075 (2014).

195. Lee, J. H. *et al.* Fluorescent in Situ Sequencing (FISSEQ) of RNA for Gene Expression Profiling in Intact Cells and Tissues. *Nature Protocols* **10,** 442–458. ISSN: 1750-2799 (Mar. 2015).

196. Leggett, B. & Whitehall, V. Role of the Serrated Pathway in Colorectal Cancer Pathogenesis. *Gastroenterology* **138,** 2088–2100. ISSN: 0016-5085, 1528-0012 (May 2010).

197. Lein, E., Borm, L. E. & Linnarsson, S. The Promise of Spatial Transcriptomics for Neuroscience in the Era of Molecular Cell Typing. *Science (New York, N.Y.)* **358,** 64–69. ISSN: 1095-9203 (Oct. 2017).

198. Leiserson, M. D. M. *et al.* Pan-Cancer Network Analysis Identifies Combinations of Rare Somatic Mutations across Pathways and Protein Complexes. *Nature Genetics* **47,** 106–114. ISSN: 1061-4036, 1546-1718 (Feb. 2015).

199. Li, A.-J. *et al.* PIK3CA and TP53 Mutations Predict Overall Survival of Stage II/III Colorectal Cancer Patients. *World Journal of Gastroenterology* **24,** 631–640. ISSN: 1007-9327 (Feb. 2018).

200. Li, C. & Li, H. Network-Constrained Regularization and Variable Selection for Analysis of Genomic Data. *Bioinformatics* **24,** 1175–1182. ISSN: 1367-4803 (May 2008).

201. Li, J. *et al.* CDC5L Promotes hTERT Expression and Colorectal Tumor Growth. *Cellular Physiology and Biochemistry* **41,** 2475–2488. ISSN: 1015-8987, 1421-9778 (2017).

202. Li, Y. *et al.* Pathway Perturbations in Signaling Networks: Linking Genotype to Phenotype. *Seminars in Cell & Developmental Biology.* ISSN: 1084-9521 (May 2018).

203. Liao, J., Lu, X., Shao, X., Zhu, L. & Fan, X. Uncovering an Organ's Molecular Architecture at Single-Cell Resolution by Spatially Resolved Transcriptomics. *Trends in Biotechnology* **39,** 43–58. ISSN: 01677799 (Jan. 2021).

204. Liberzon, A. *et al.* The Molecular Signatures Database (MSigDB) Hallmark Gene Set Collection. *Cell systems* **1,** 417–425. ISSN: 2405-4712 (Dec. 2015).

205. Liu, X., Wang, Y., Ji, H., Aihara, K. & Chen, L. Personalized Characterization of Diseases Using Sample-Specific Networks. *Nucleic Acids Research* **44,** e164–e164. ISSN: 0305-1048 (Dec. 2016).

206. Liu, X. *et al.* MALDI-MSI of Immunotherapy: Mapping the EGFR-Targeting Antibody Cetuximab in 3D Colon-Cancer Cell Cultures. *Analytical Chemistry* **90,** 14156–14164. ISSN: 0003-2700 (Dec. 2018).

207. Lubeck, E. & Cai, L. Single-Cell Systems Biology by Super-Resolution Imaging and Combinatorial Labeling. *Nature Methods* **9,** 743–748. ISSN: 1548-7105 (July 2012).

208. Lubeck, E., Coskun, A. F., Zhiyentayev, T., Ahmad, M. & Cai, L. Single-Cell in Situ RNA Profiling by Sequential Hybridization. *Nature Methods* **11,** 360–361. ISSN: 1548-7091, 1548-7105 (Apr. 2014).

209. Lund, S. P., Nettleton, D., McCarthy, D. J. & Smyth, G. K. Detecting Differential Expression in RNA-sequence Data Using Quasi-Likelihood with Shrunken Dispersion Estimates. *Statistical Applications in Genetics and Molecular Biology* **11.** ISSN: 1544-6115 (Oct. 2012).

210. Luo, P., Tian, L. P., Ruan, J. & Wu, F. Disease Gene Prediction by Integrating PPI Networks, Clinical RNA-Seq Data and OMIM Data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics,* 1–1. ISSN: 1545-5963 (2017).

211. Luo, Y., Batalao, A., Zhou, H. & Zhu, L. Mammalian Two-Hybrid System: A Complementary Approach to the Yeast Two-Hybrid System. *BioTechniques* **22,** 350–352. ISSN: 0736-6205 (Feb. 1997).

212. Lyubimova, A. *et al.* Single-Molecule mRNA Detection and Counting in Mammalian Tissue. *Nature Protocols* **8,** 1743–1758. ISSN: 1754-2189, 1750-2799 (Sept. 2013).

213. Ma, J., Shojaie, A. & Michailidis, G. A Comparative Study of Topology-Based Pathway Enrichment Analysis Methods. *BMC Bioinformatics* **20,** 546. ISSN: 1471-2105 (Nov. 2019).

214. Ma, J., Shojaie, A. & Michailidis, G. Network-Based Pathway Enrichment Analysis with Incomplete Network Information. *Bioinformatics* **32,** 3165–3174. ISSN: 1367-4803 (Oct. 2016).

215. Ma'ayan, A. Complex Systems Biology. *Journal of The Royal Society Interface* **14,** 20170391 (Sept. 2017).

216. Martini, P., Sales, G., Massa, M. S., Chiogna, M. & Romualdi, C. Along Signal Paths: An Empirical Gene Set Approach Exploiting Pathway Topology. *Nucleic Acids Research* **41,** e19. ISSN: 0305-1048 (Jan. 2013).

217. Marusyk, A., Almendro, V. & Polyak, K. Intra-Tumour Heterogeneity: A Looking Glass for Cancer? *Nature Reviews Cancer* **12,** 323–334. ISSN: 1474-1768 (May 2012).

218. Marx, V. Method of the Year: Spatially Resolved Transcriptomics. *Nature Methods* **18,** 9–14. ISSN: 1548-7105 (Jan. 2021).

219. Masuda, N., Porter, M. A. & Lambiotte, R. Random Walks and Diffusion on Networks. *Physics Reports. Random Walks and Diffusion on Networks* **716–717,** 1–58. ISSN: 0370-1573 (Nov. 2017).

220. Maule, M. & Merletti, F. Cancer Transition and Priorities for Cancer Control. *The Lancet Oncology* **13,** 745–746. ISSN: 1470-2045 (Aug. 2012).

221. Mayes, P. A., Hance, K. W. & Hoos, A. The Promise and Challenges of Immune Agonist Antibody Development in Cancer. *Nature Reviews Drug Discovery* **17,** 509–527. ISSN: 1474-1784 (July 2018).

222. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv:1802.03426 [cs, stat].* arXiv: 1802.03426 [cs, stat] (2018).

223. McIntyre, L. M. *et al.* RNA-seq: Technical Variability and Sampling. *BMC Genomics* **12,** 293. ISSN: 1471-2164 (June 2011).

224. Meng, H., Li, W., Boardman, L. A. & Wang, L. Loss of ZG16 Is Associated with Molecular and Clinicopathological Phenotypes of Colorectal Cancer. *BMC Cancer* **18,** 433. ISSN: 1471-2407 (Apr. 2018).

225. Menter, D. G. *et al.* Back to the Colorectal Cancer Consensus Molecular Subtype Future. *Current gastroenterology reports* **21,** 5. ISSN: 1522-8037 (Jan. 2019).

226. Meyer, M. J. *et al.* Interactome INSIDER: A Structural Interactome Browser for Genomic Studies. *Nature Methods* **15,** 107–114. ISSN: 1548-7105 (Feb. 2018).

227. Mik, M., Berut, M., Dziki, L., Trzcinski, R. & Dziki, A. Right- and Left-Sided Colon Cancer – Clinical and Pathological Differences of the Disease Entity in One Organ. *Archives of Medical Science : AMS* **13,** 157–162. ISSN: 1734-1922 (Feb. 2017).

228. Mirnezami, R. *et al.* Chemical Mapping of the Colorectal Cancer Microenvironment via MALDI Imaging Mass Spectrometry (MALDI-MSI) Reveals Novel Cancer-Associated Field Effects. *Molecular Oncology* **8,** 39–49. ISSN: 1574-7891 (Feb. 2014).

229. Mistry, J. *et al.* Pfam: The Protein Families Database in 2021. *Nucleic Acids Research* **49,** D412–D419. ISSN: 0305-1048 (Jan. 2021).

230. Mojarad, E. N., Kuppen, P. J., Aghdaei, H. A. & Zali, M. R. The CpG Island Methylator Phenotype (CIMP) in Colorectal Cancer. *Gastroenterology and Hepatology From Bed to Bench* **6,** 120–128. ISSN: 2008-2258 (2013).

231. Molina-Cerrillo, J. *et al.* BRAF Mutated Colorectal Cancer: New Treatment Approaches. *Cancers* **12** (June 2020).

232. Monti, S., Tamayo, P., Mesirov, J. & Golub, T. Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data. *Machine Learning* **52,** 91–118. ISSN: 1573-0565 (July 2003).

233. Moor, A. E. & Itzkovitz, S. Spatial Transcriptomics: Paving the Way for Tissue-Level Systems Biology. *Current Opinion in Biotechnology. Systems Biology • Nanobiotechnology* **46,** 126–133. ISSN: 0958-1669 (Aug. 2017).

234. Moore, A. R., Rosenberg, S. C., McCormick, F. & Malek, S. RAS-targeted Therapies: Is the Undruggable Drugged? *Nature Reviews Drug Discovery* **19,** 533–552. ISSN: 1474-1784 (Aug. 2020).

235. Morin, P. J. *et al.* Activation of β-Catenin-Tcf Signaling in Colon Cancer by Mutations in β-Catenin or APC. *Science* **275,** 1787–1790. ISSN: 0036-8075, 1095-9203 (Mar. 1997).

236. Mosca, R., Céol, A., Stein, A., Olivella, R. & Aloy, P. 3did: A Catalog of Domain-Based Interactions of Known Three-Dimensional Structure. *Nucleic Acids Research* **42,** D374–D379. ISSN: 0305-1048 (Jan. 2014).

237. Mounir, M. *et al.* New Functionalities in the TCGAbiolinks Package for the Study and Integration of Cancer Data from GDC and GTEx. *PLoS Computational Biology* **15.** ISSN: 1553-734X (Mar. 2019).

238. Mubeen, S. *et al.* The Impact of Pathway Database Choice on Statistical Enrichment Analysis and Predictive Modeling. *Frontiers in Genetics* **10.** ISSN: 1664-8021 (2019).

239. Muetze, T. *et al.* Contextual Hub Analysis Tool (CHAT): A Cytoscape App for Identifying Contextually Relevant Hubs in Biological Networks. *F1000Research* **5.** ISSN: 2046-1402 (Aug. 2016).

240. Murphy, G. *et al.* Sex Disparities in Colorectal Cancer Incidence by Anatomic Subsite, Race and Age. *International Journal of Cancer* **128,** 1668–1675. ISSN: 1097-0215 (2011).

241. Mw, H. & Ad, K. *Comparative Genomics of Centrality and Essentiality in Three Eukaryotic Protein-Interaction Networks* 2005.

242. Navarro, J. F., Sjöstrand, J., Salmén, F., Lundeberg, J. & Ståhl, P. L. ST Pipeline: An Automated Pipeline for Spatial Mapping of Unique Transcripts. *Bioinformatics* **33** (ed Berger, B.) 2591–2593. ISSN: 1367-4803, 1460-2059 (Aug. 2017).

243. Network, T. C. G. A. R. Comprehensive, Integrative Genomic Analysis of Diffuse Lower-Grade Gliomas. *The New England journal of medicine* **372,** 2481–2498. ISSN: 0028-4793 (June 2015).

244. Nguyen, H. T. & Duong, H.-Q. The Molecular Characteristics of Colorectal Cancer: Implications for Diagnosis and Therapy (Review). *Oncology Letters* **16,** 9–18. ISSN: 1792-1074 (July 2018).

245. Nicholson, R. I., Gee, J. M. W. & Harper, M. E. EGFR and Cancer Prognosis. *European Journal of Cancer* **37,** 9–15. ISSN: 0959-8049 (Sept. 2001).

246. Nowell, P. C. The Clonal Evolution of Tumor Cell Populations. *Science* **194,** 23–28. ISSN: 0036-8075, 1095-9203 (Oct. 1976).

247. Obuch, J. C., Pigott, C. M. & Ahnen, D. J. Sessile Serrated Polyps: Detection, Eradication, and Prevention of the Evil Twin. *Current treatment options in gastroenterology* **13,** 156–170. ISSN: 1092-8472 (Mar. 2015).

248. Oliveira, C. *et al.* Distinct Patterns of KRAS Mutations in Colorectal Carcinomas According to Germline Mismatch Repair Defects and hMLH1 Methylation Status. *Human Molecular Genetics* **13,** 2303–2311. ISSN: 0964-6906 (Oct. 2004).

249. Orchard, S. *et al.* Protein Interaction Data Curation: The International Molecular Exchange (IMEx) Consortium. *Nature Methods* **9,** 345–350. ISSN: 1548-7105 (Apr. 2012).

250. Orchard, S. *et al.* The MIntAct Project–IntAct as a Common Curation Platform for 11 Molecular Interaction Databases. *Nucleic acids research* **42,** D358–63. ISSN: 1362-4962 (Jan. 2014).

251. Oti, M., Snel, B., Huynen, M. A. & Brunner, H. G. Predicting Disease Genes Using Protein–Protein Interactions. *Journal of Medical Genetics* **43,** 691–698. ISSN: 0022-2593, 1468-6244 (Aug. 2006).

252. Ozturk, K., Dow, M., Carlin, D. E., Bejar, R. & Carter, H. The Emerging Potential for Network Analysis to Inform Precision Cancer Medicine. *Journal of Molecular Biology. Theory and Application of Network Biology Toward Precision Medicine* **430,** 2875–2899. ISSN: 0022-2836 (Sept. 2018).

253. Pagès, F. *et al.* International Validation of the Consensus Immunoscore for the Classification of Colon Cancer: A Prognostic and Accuracy Study. *The Lancet* **391,** 2128–2139. ISSN: 0140-6736 (May 2018).

254. Papke, B. & Der, C. J. Drugging RAS: Know the Enemy. *Science* **355,** 1158–1163. ISSN: 0036-8075, 1095-9203 (Mar. 2017).

255. Paschke, S. *et al.* Are Colon and Rectal Cancer Two Different Tumor Entities? A Proposal to Abandon the Term Colorectal Cancer. *International Journal of Molecular Sciences* **19.** ISSN: 1422-0067 (Aug. 2018).

256. Paull, E. O. *et al.* Discovering Causal Pathways Linking Genomic Events to Transcriptional States Using Tied Diffusion Through Interacting Events (TieDIE). *Bioinformatics* **29,** 2757–2764. ISSN: 13674803 (Nov. 2013).

257. Pavlopoulos, G. A. *et al.* Using Graph Theory to Analyze Biological Networks. *BioData Mining* **4,** 10. ISSN: 1756-0381 (Apr. 2011).

258. PCAWG Consortium. Pan-Cancer Analysis of Whole Genomes. *Nature* **578,** 82–93. ISSN: 1476-4687 (Feb. 2020).

259. Pedregosa, F. *et al.* Scikit-Learn: Machine Learning in Python. *Journal of Machine Learning Research* **12,** 2825–2830. ISSN: 1533-7928 (Oct. 2011).

260. Peixoto, J. & Lima, J. Metabolic Traits of Cancer Stem Cells. *Disease Models & Mechanisms* **11,** dmm033464. ISSN: 1754-8403, 1754-8411 (Aug. 2018).

261. Peixoto, T. P. Hierarchical Block Structures and High-Resolution Model Selection in Large Networks. *Physical Review X* **4,** 011047 (Mar. 2014).

262. Peixoto, T. P. *The Graph-Tool Python Library* 2017.

263. Perez, F. & Granger, B. E. IPython: A System for Interactive Scientific Computing. *Computing in Science Engineering* **9,** 21–29. ISSN: 1558-366X (May 2007).

264. Perou, C. M. *et al.* Molecular Portraits of Human Breast Tumours. *Nature* **406,** 747–752. ISSN: 1476-4687 (Aug. 2000).

265. Pico, A. R. *et al.* WikiPathways: Pathway Editing for the People. *PLOS Biology* **6,** e184. ISSN: 1545-7885 (July 2008).

266. Pino, M. S. & Chung, D. C. The Chromosomal Instability Pathway in Colon Cancer. *Gastroenterology* **138,** 2059–2072. ISSN: 00165085 (May 2010).

267. Pirman, D. A. *et al.* Changes in Cancer Cell Metabolism Revealed by Direct Sample Analysis with MALDI Mass Spectrometry. *PloS One* **8,** e61379. ISSN: 1932-6203 (2013).

268. Polakis, P. The Adenomatous Polyposis Coli (APC) Tumor Suppressor. *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer* **1332,** F127–F147. ISSN: 0304-419X (June 1997).

269. Prasetyanti, P. R. & Medema, J. P. Intra-Tumor Heterogeneity from a Cancer Stem Cell Perspective. *Molecular Cancer* **16,** 41. ISSN: 1476-4598 (Feb. 2017).

270. Punt, C. J. A., Koopman, M. & Vermeulen, L. From Tumour Heterogeneity to Advances in Precision Treatment of Colorectal Cancer. *Nature Reviews Clinical Oncology* **14,** 235–246. ISSN: 1759-4782 (Apr. 2017).

271. Python Software Foundation. *The Python Language Reference* (2020).

272. R Development Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria, 2013).

273. Ramazzotti, D., Lal, A., Wang, B., Batzoglou, S. & Sidow, A. Multi-Omic Tumor Data Reveal Diversity of Molecular Mechanisms That Correlate with Survival. *Nature Communications* **9,** 4453. ISSN: 2041-1723 (Oct. 2018).

274. Regenmortel, M. H. V. Reductionism and Complexity in Molecular Biology. *EMBO Reports* **5,** 1016–1020. ISSN: 1469-221X (Nov. 2004).

275. Reimand, J., Kull, M., Peterson, H., Hansen, J. & Vilo, J. G:Profiler—a Web-Based Toolset for Functional Profiling of Gene Lists from Large-Scale Experiments. *Nucleic Acids Research* **35,** W193–W200. ISSN: 0305-1048 (July 2007).

276. Reuben, A. *et al.* Genomic and Immune Heterogeneity Are Associated with Differential Responses to Therapy in Melanoma. *NPJ genomic medicine* **2.** ISSN: 2056-7944 (2017).

277. Reyna, M. A., Leiserson, M. D. M. & Raphael, B. J. Hierarchical HotNet: Identifying Hierarchies of Altered Subnetworks. *Bioinformatics* **34,** i972–i980. ISSN: 1367-4803 (Sept. 2018).

278. Reyna, M. A. *et al.* Pathway and Network Analysis of More than 2500 Whole Cancer Genomes. *Nature Communications* **11,** 1–17. ISSN: 2041-1723 (Feb. 2020).

279. Ritchie, M. E. *et al.* Limma Powers Differential Expression Analyses for RNA-sequencing and Microarray Studies. *Nucleic Acids Research* **43,** e47–e47. ISSN: 0305-1048 (Apr. 2015).

280. Rivas, J. D. L. & Fontanillo, C. Protein–Protein Interactions Essentials: Key Concepts to Building and Analyzing Interactome Networks. *PLOS Computational Biology* **6,** e1000807. ISSN: 1553-7358 (June 2010).

281. Robert, C. A Decade of Immune-Checkpoint Inhibitors in Cancer Therapy. *Nature Communications* **11,** 3801. ISSN: 2041-1723 (July 2020).

282. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data. *Bioinformatics* **26,** 139–140. ISSN: 1367-4803 (Jan. 2010).

283. Robinson, M. D. & Oshlack, A. A Scaling Normalization Method for Differential Expression Analysis of RNA-seq Data. *Genome Biology* **11,** R25. ISSN: 1474-760X (Mar. 2010).

284. Rodchenkov, I. *et al.* Pathway Commons 2019 Update: Integration, Analysis and Exploration of Pathway Data. *Nucleic Acids Research* **48,** D489–D497. ISSN: 0305-1048 (Jan. 2020).

285. Rodríguez-Pérez, R., Fernández, L. & Marco, S. Overoptimism in Cross-Validation When Using Partial Least Squares-Discriminant Analysis for Omics Data: A Systematic Study. *Analytical and Bioanalytical Chemistry* **410,** 5981–5992. ISSN: 1618-2650 (Sept. 2018).

286. Rodriques, S. G. *et al.* Slide-Seq: A Scalable Technology for Measuring Genome-Wide Expression at High Spatial Resolution. *Science* **363,** 1463–1467. ISSN: 0036-8075, 1095-9203 (Mar. 2019).

287. Rohart, F., Gautier, B., Singh, A. & Cao, K.-A. L. mixOmics: An R Package for 'omics Feature Selection and Multiple Data Integration. *PLOS Computational Biology* **13,** e1005752. ISSN: 1553-7358 (Nov. 2017).

288. Rohn, H. *et al.* VANTED v2: A Framework for Systems Biology Applications. *BMC Systems Biology* **6,** 139. ISSN: 1752-0509 (Nov. 2012).

289. Rohner, T. C., Staab, D. & Stoeckli, M. MALDI Mass Spectrometric Imaging of Biological Tissue Sections. *Mechanisms of Ageing and Development. Functional Genomics of Ageing II* **126,** 177–185. ISSN: 0047-6374 (Jan. 2005).

290. Rolland, T. *et al.* A Proteome-Scale Map of the Human Interactome Network. *Cell* **159,** 1212–1226. ISSN: 0092-8674 (Nov. 2014).

291. Ros-Martínez, S., Navas-Carrillo, D., Alonso-Romero, J. L. & Orenes-Piñero, E. Immunoscore: A Novel Prognostic Tool. Association with Clinical Outcome, Response to Treatment and Survival in Several Malignancies. *Critical Reviews in Clinical Laboratory Sciences* **0,** 1–12. ISSN: 1040-8363 (Mar. 2020).

292. Rousseeuw, P. J. Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Journal of Computational and Applied Mathematics* **20,** 53–65. ISSN: 0377-0427 (Nov. 1987).

293. Routy, B. *et al.* Gut Microbiome Influences Efficacy of PD-1–Based Immunotherapy against Epithelial Tumors. *Science* **359,** 91–97. ISSN: 0036-8075, 1095-9203 (Jan. 2018).

294. Royston, P. & Parmar, M. K. Restricted Mean Survival Time: An Alternative to the Hazard Ratio for the Design and Analysis of Randomized Trials with a Time-to-Event Outcome. *BMC Medical Research Methodology* **13,** 152. ISSN: 1471-2288 (Dec. 2013).

295. Ruffner, H., Bauer, A. & Bouwmeester, T. Human Protein–Protein Interaction Networks and the Value for Drug Discovery. *Drug Discovery Today* **12,** 709–716. ISSN: 1359-6446 (Sept. 2007).

296. Ruiz-Perez, D. & Narasimhan, G. *So You Think You Can PLS-DA?* Preprint (Bioinformatics, Oct. 2017).

297. Sahni, N. *et al.* Widespread Macromolecular Interaction Perturbations in Human Genetic Disorders. *Cell* **161,** 647–660. ISSN: 0092-8674 (Apr. 2015).

298. Salamon, J., Qian, X., Nilsson, M. & Lynn, D. J. Network Visualization and Analysis of Spatially Aware Gene Expression Data with InsituNet. *Cell Systems* **6,** 626–630.e3. ISSN: 2405-4712 (May 2018).

299. Salzberg, S. L. Open Questions: How Many Genes Do We Have? *BMC Biology* **16.** ISSN: 1741-7007 (Aug. 2018).

300. Samaga, R., Saez-Rodriguez, J., Alexopoulos, L. G., Sorger, P. K. & Klamt, S. The Logic of EGFR/ErbB Signaling: Theoretical Properties and Analysis of High-Throughput Data. *PLOS Computational Biology* **5,** e1000438. ISSN: 1553-7358 (Aug. 2009).

301. Sanborn, J. Z. *et al.* Phylogenetic Analyses of Melanoma Reveal Complex Patterns of Metastatic Dissemination. *Proceedings of the National Academy of Sciences of the United States of America* **112,** 10995–11000. ISSN: 1091-6490 (Sept. 2015).

302. Sánchez-Gundín, J., Fernández-Carballido, A. M., Martínez-Valdivieso, L., Barreda-Hernández, D. & Torres-Suárez, A. I. New Trends in the Therapeutic Approach to Metastatic Colorectal Cancer. *International Journal of Medical Sciences* **15,** 659–665. ISSN: 1449-1907 (Apr. 2018).

303. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial Reconstruction of Single-Cell Gene Expression Data. *Nature Biotechnology* **33,** 495–502. ISSN: 1546-1696 (May 2015).

304. Sayin, S. I. *et al.* Gut Microbiota Regulates Bile Acid Metabolism by Reducing the Levels of Tauro-Beta-Muricholic Acid, a Naturally Occurring FXR Antagonist. *Cell Metabolism* **17,** 225–235. ISSN: 1932-7420 (Feb. 2013).

305. Scaldaferri, F. *et al.* Gut Microbial Flora, Prebiotics, and Probiotics in IBD: Their Current Usage and Utility. *BioMed Research International* **2013,** 435268. ISSN: 2314-6141 (2013).

306. Schatoff, E. M., Leach, B. I. & Dow, L. E. Wnt Signaling and Colorectal Cancer. *Current colorectal cancer reports* **13,** 101–110. ISSN: 1556-3790 (Apr. 2017).

307. Schubert, M. *et al.* Perturbation-Response Genes Reveal Signaling Footprints in Cancer Gene Expression. *Nature Communications* **9,** 20. ISSN: 2041-1723 (Jan. 2018).

308. Seabold, S. & Perktold, J. *Statsmodels: Econometric and Statistical Modeling with Python* in *9th Python in Science Conference* (2010).

309. Shannon, P. *et al.* Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research* **13,** 2498 (Nov. 2003).

310. Shao, W. *et al.* Integrative Analysis of Pathological Images and Multi-Dimensional Genomic Data for Early-Stage Cancer Prognosis. *IEEE Transactions on Medical Imaging* **39,** 99–110. ISSN: 1558-254X (Jan. 2020).

311. Sharma, P., Bhattacharyya, D. K. & Kalita, J. K. *Centrality Analysis in PPI Networks* in *2016 International Conference on Accessibility to Digital World (ICADW)* (Dec. 2016), 135–140.

312. Sheffer, M. *et al.* Association of Survival and Disease Progression with Chromosomal Instability: A Genomic Exploration of Colorectal Cancer. *Proceedings of the National Academy of Sciences* **106,** 7131–7136. ISSN: 0027-8424, 1091-6490 (Apr. 2009).

313. Shen, R., Olshen, A. B. & Ladanyi, M. Integrative Clustering of Multiple Genomic Data Types Using a Joint Latent Variable Model with Application to Breast and Lung Cancer Subtype Analysis. *Bioinformatics (Oxford, England)* **25,** 2906–2912. ISSN: 1367-4811 (Nov. 2009).

314. Shojaie, A. & Michailidis, G. Network Enrichment Analysis in Complex Experiments. *Statistical Applications in Genetics and Molecular Biology* **9.** ISSN: 1544-6115 (May 2010).

315. Siegel, R. L., Sahar, L., Robbins, A. & Jemal, A. Where Can Colorectal Cancer Screening Interventions Have the Most Impact? *Cancer Epidemiology and Prevention Biomarkers* **24,** 1151–1156. ISSN: 1055-9965, 1538-7755 (Aug. 2015).

316. Siegel, R. L. *et al.* Colorectal Cancer Statistics, 2017. *CA: A Cancer Journal for Clinicians* **67,** 177–193. ISSN: 1542-4863 (2017).

317. Siegel, R. L. *et al.* Global Patterns and Trends in Colorectal Cancer Incidence in Young Adults. *Gut* **68,** 2179–2185. ISSN: 0017-5749, 1468-3288 (Dec. 2019).

318. Silverbush, D. & Sharan, R. A Systematic Approach to Orient the Human Protein–Protein Interaction Network. *Nature Communications* **10,** 3015. ISSN: 2041-1723 (July 2019).

319. Sivade (Dumousseau), M. *et al.* Encompassing New Use Cases - Level 3.0 of the HUPO-PSI Format for Molecular Interactions. *BMC Bioinformatics* **19,** 134. ISSN: 1471-2105 (Apr. 2018).

320. Slenter, D. N. *et al.* WikiPathways: A Multifaceted Pathway Database Bridging Metabolomics to Other Omics Research. *Nucleic Acids Research* **46,** D661–D667. ISSN: 0305-1048 (Jan. 2018).

321. Snider, J. *et al.* Fundamentals of Protein Interaction Network Mapping. *Molecular Systems Biology* **11,** 848. ISSN: 1744-4292, 1744-4292 (Dec. 2015).

322. Snuderl, M. *et al.* Mosaic Amplification of Multiple Receptor Tyrosine Kinase Genes in Glioblastoma. *Cancer Cell* **20,** 810–817. ISSN: 1878-3686 (Dec. 2011).

323. Sobin, L. H., Gospodarowicz, M. K. & Wittekind, C. *TNM Classification of Malignant Tumours* ISBN: 978-1-4443-5896-4 (John Wiley & Sons, Aug. 2011).

324. Sonawane, A. R., Weiss, S. T., Glass, K. & Sharma, A. Network Medicine in the Age of Biomedical Big Data. *Frontiers in Genetics* **10.** ISSN: 1664-8021 (2019).

325. Sottoriva, A. *et al.* A Big Bang Model of Human Colorectal Tumor Growth. *Nature Genetics* **47,** 209–216. ISSN: 1546-1718 (Mar. 2015).

326. Sparano, J. A. & Paik, S. Development of the 21-Gene Assay and Its Application in Clinical Practice and Clinical Trials. *Journal of Clinical Oncology* **26,** 721–728. ISSN: 0732-183X (Feb. 2008).

327. Ståhl, P. L. *et al.* Visualization and Analysis of Gene Expression in Tissue Sections by Spatial Transcriptomics. *Science (New York, N.Y.)* **353,** 78–82. ISSN: 1095-9203 (July 2016).

328. Stark, C. *et al.* BioGRID: A General Repository for Interaction Datasets. *Nucleic Acids Research* **34,** D535–D539. ISSN: 0305-1048 (Jan. 2006).

329. Stark, R., Grzelak, M. & Hadfield, J. RNA Sequencing: The Teenage Years. *Nature Reviews Genetics* **20,** 631–656. ISSN: 1471-0064 (Nov. 2019).

330. Steinke, V. *et al.* Hereditary Nonpolyposis Colorectal Cancer (HNPCC)/Lynch Syndrome. *Deutsches Ärzteblatt International* **110,** 32–38. ISSN: 1866-0452 (Jan. 2013).

331. Stojmirović, A., Bliskovsky, A. & Yu, Y.-K. CytoITMprobe: A Network Information Flow Plugin for Cytoscape. *BMC Research Notes* **5,** 237. ISSN: 1756-0500 (May 2012).

332. Stojmirović, A. & Yu, Y.-K. Information Flow in Interaction Networks. *Journal of computational biology : a journal of computational molecular cell biology* **14,** 1115–43. ISSN: 1066-5277. arXiv: 1112.3988 (Oct. 2007).

333. Stojmirović, A. & Yu, Y. K. ITM Probe: Analyzing Information Flow in Protein Networks. *Bioinformatics* **25,** 2447–2449. ISSN: 13674803. arXiv: 0904.2770 (Sept. 2009).

334. Street, W. Cancer Facts & Figures 2020, 76 (2020).

335. Strell, C. *et al.* Placing RNA in Context and Space – Methods for Spatially Resolved Transcriptomics. *The FEBS Journal* **286,** 1468–1481. ISSN: 1742-4658 (2019).

336. Strogatz, S. H. Exploring Complex Networks. *Nature* **410,** 268–276. ISSN: 0028-0836. arXiv: cond-mat/0102091 (Mar. 2001).

337. Stukalov, A. *et al.* Multilevel Proteomics Reveals Host Perturbations by SARS-CoV-2 and SARS-CoV. *Nature* **594,** 246–252. ISSN: 1476-4687 (June 2021).

338. Subramanian, A. *et al.* Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles. *Proceedings of the National Academy of Sciences of the United States of America* **102,** 15545–15550. ISSN: 0027-8424 (Oct. 2005).

339. Subramanian, I., Verma, S., Kumar, S., Jere, A. & Anamika, K. Multi-Omics Data Integration, Interpretation, and Its Application. *Bioinformatics and Biology Insights* **14.** ISSN: 1177-9322 (Jan. 2020).

340. Sud, M. *et al.* LMSD: LIPID MAPS Structure Database. *Nucleic Acids Research* **35,** D527–532. ISSN: 1362-4962 (Jan. 2007).

341. Sun, Y. *et al.* TMEM74 Promotes Tumor Cell Survival by Inducing Autophagy via Interactions with ATG16L1 and ATG9A. *Cell Death & Disease* **8,** e3031–e3031. ISSN: 2041-4889 (Aug. 2017).

342. Suvà, M. L. & Tirosh, I. Single-Cell RNA Sequencing in Cancer: Lessons Learned and Emerging Challenges. *Molecular Cell* **75,** 7–12. ISSN: 1097-2765 (July 2019).

343. Tam, V. *et al.* Benefits and Limitations of Genome-Wide Association Studies. *Nature Reviews Genetics* **20,** 467–484. ISSN: 1471-0064 (Aug. 2019).

344. Tan, I. B. & Tan, P. An 18-Gene Signature (ColoPrint ®) for Colon Cancer Prognosis. *Nature Reviews Clinical Oncology* **8,** 131–133. ISSN: 1759-4782 (Mar. 2011).

345. Tarca, A. L. *et al.* A Novel Signaling Pathway Impact Analysis. *Bioinformatics* **25,** 75–82. ISSN: 1367-4803 (Jan. 2009).

346. Taylor, I. W. & Wrana, J. L. Protein Interaction Networks in Medicine and Disease. *PROTEOMICS* **12,** 1706–1716. ISSN: 1615-9861 (2012).

347. Terfve, C. *et al.* CellNOptR: A Flexible Toolkit to Train Protein Signaling Networks to Data Using Multiple Logic Formalisms. *BMC Systems Biology* **6,** 133. ISSN: 1752-0509 (Oct. 2012).

348. The Cancer Genome Atlas Network. Comprehensive Molecular Characterization of Human Colon and Rectal Cancer. *Nature* **487,** 330–337. ISSN: 1476-4687 (July 2012).

349. Timmons, J. A., Szkop, K. J. & Gallagher, I. J. Multiple Sources of Bias Confound Functional Enrichment Analysis of Global -Omics Data. *Genome Biology* **16,** 186. ISSN: 1474-760X (Sept. 2015).

350. Tirnauer, J. S. & Bierer, B. E. Eb1 Proteins Regulate Microtubule Dynamics, Cell Polarity, and Chromosome Stability. *The Journal of Cell Biology* **149,** 761–766. ISSN: 0021-9525 (May 2000).

351. Tolkach, Y. & Kristiansen, G. The Heterogeneity of Prostate Cancer: A Practical Approach. *Pathobiology* **85,** 108–116. ISSN: 1015-2008, 1423-0291 (2018).

352. Tomfohr, J., Lu, J. & Kepler, T. B. Pathway Level Analysis of Gene Expression Using Singular Value Decomposition. *BMC Bioinformatics* **6,** 225. ISSN: 1471-2105 (Sept. 2005).

353. Tomlins, S. A. *et al.* Recurrent Fusion of TMPRSS2 and ETS Transcription Factor Genes in Prostate Cancer. *Science* **310,** 644–648 (Oct. 2005).

354. Toyota, M. *et al.* CpG Island Methylator Phenotype in Colorectal Cancer. *Proceedings of the National Academy of Sciences* **96,** 8681–8686. ISSN: 0027-8424, 1091-6490 (July 1999).

355. Tuncbag, N. *et al.* Network-Based Interpretation of Diverse High-Throughput Datasets through the Omics Integrator Software Package. *PLOS Computational Biology* **12,** e1004879. ISSN: 1553-7358 (Apr. 2016).

356. Türei, D., Korcsmáros, T. & Saez-Rodriguez, J. OmniPath: Guidelines and Gateway for Literature-Curated Signaling Pathway Resources. *Nature Methods* **13,** 966–967. ISSN: 1548-7105 (Dec. 2016).

357. Van den Berge, K. *et al.* RNA Sequencing Data: Hitchhiker's Guide to Expression Analysis, 37 (2019).

358. Van Der Maaten, L. J. P. & Hinton, G. E. Visualizing High-Dimensional Data Using t-Sne. *Journal of Machine Learning Research* **9,** 2579–2605. ISSN: 1532-4435. arXiv: 1307.1662 (2008).

359. Vander Heiden, M. G. & DeBerardinis, R. J. Understanding the Intersections between Metabolism and Cancer Biology. *Cell* **168,** 657–669. ISSN: 1097-4172 (Feb. 2017).

360. Vandin, F., Clay, P., Upfal, E. & Raphael, B. J. in *Biocomputing 2012* 55–66 (WORLD SCIENTIFIC, 2011). ISBN: 978-981-4596-37-4.

361. Vandin, F., Upfal, E. & Raphael, B. J. Algorithms for Detecting Significantly Mutated Pathways in Cancer. *Journal of Computational Biology* **18,** 507–522 (Mar. 2011).

362. Vanunu, O., Magger, O., Ruppin, E., Shlomi, T. & Sharan, R. Associating Genes and Protein Complexes with Disease via Network Propagation. *PLOS Computational Biology* **6,** e1000641. ISSN: 1553-7358 (Jan. 2010).

363. Vaske, C. J. *et al.* Inference of Patient-Specific Pathway Activities from Multi-Dimensional Cancer Genomics Data Using PARADIGM. *Bioinformatics* **26,** i237–i245. ISSN: 1367-4803 (June 2010).

364. Vickovic, S. *et al.* High-Definition Spatial Transcriptomics for in Situ Tissue Profiling. *Nature Methods,* 1–4. ISSN: 1548-7105 (Sept. 2019).

365. Vidal, M., Cusick, M. E. & Barabási, A.-L. Interactome Networks and Human Disease. *Cell* **144,** 986–998. ISSN: 0092-8674 (Mar. 2011).

366. Villaveces, J. M. *et al.* Merging and Scoring Molecular Interactions Utilising Existing Community Standards: Tools, Use-Cases and a Case Study. *Database* **2015** (Jan. 2015).

367. Virtanen, P. *et al.* SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods,* 1–12. ISSN: 1548-7105 (Feb. 2020).

368. Vogelstein, B. *et al.* Cancer Genome Landscapes. *Science* **339,** 1546–1558. ISSN: 0036-8075, 1095-9203 (Mar. 2013).

369. Vonderheide, R. H. The Immune Revolution: A Case for Priming, Not Checkpoint. *Cancer Cell* **33,** 563–569. ISSN: 1878-3686 (Apr. 2018).

370. Walther, A. *et al.* Genetic Prognostic and Predictive Markers in Colorectal Cancer. *Nature Reviews Cancer* **9,** 489–499. ISSN: 1474-1768 (July 2009).

371. Wang, B., Tang, H., Guo, C. & Xiu, Z. Entropy Optimization of Scale-Free Networks' Robustness to Random Failures. *Physica A: Statistical Mechanics and its Applications* **363,** 591–596. ISSN: 0378-4371 (May 2006).

372. Wang, B., Li, F., Zhou, X., Ma, Y. & Fu, W. Is Microsatellite Instability-High Really a Favorable Prognostic Factor for Advanced Colorectal Cancer? A Meta-Analysis. *World Journal of Surgical Oncology* **17,** 169. ISSN: 1477-7819 (Oct. 2019).

373. Wang, B. *et al.* Similarity Network Fusion for Aggregating Data Types on a Genomic Scale. *Nature Methods* **11,** 333–337. ISSN: 1548-7105 (Mar. 2014).

374. Wang, X. *et al.* Downregulated SPINK4 Is Associated with Poor Survival in Colorectal Cancer. *BMC Cancer* **19,** 1258. ISSN: 1471-2407 (Dec. 2019).

375. Wang, Y., Sahni, N. & Vidal, M. Global Edgetic Rewiring in Cancer Networks. *Cell Systems* **1,** 251–253. ISSN: 2405-4712 (Oct. 2015).

376. Ward Jr, J. H. Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association* **58,** 236–244. ISSN: 0162-1459 (Mar. 1963).

377. Wassie, A. T., Zhao, Y. & Boyden, E. S. Expansion Microscopy: Principles and Uses in Biological Research. *Nature Methods* **16,** 33–41. ISSN: 1548-7105 (Jan. 2019).

378. Watts, D. J. & Strogatz, S. H. Collective Dynamics of 'Small-World' Networks. *Nature* **393,** 440–442. ISSN: 1476-4687 (June 1998).

379. Weakley, S. M., Wang, H., Yao, Q. & Chen, C. Expression and Function of a Large Non-coding RNA Gene XIST in Human Cancer. *World Journal of Surgery* **35,** 1751–1756. ISSN: 0364-2313 (Aug. 2011).

380. Wee, P. & Wang, Z. Epidermal Growth Factor Receptor Cell Proliferation Signaling Pathways. *Cancers* **9.** ISSN: 2072-6694 (May 2017).

381. Weinstein, J. N. *et al.* The Cancer Genome Atlas Pan-Cancer Analysis Project. *Nature genetics* **45,** 1113–1120. ISSN: 1061-4036 (Oct. 2013).

382. Weisenberger, D. J., Liang, G. & Lenz, H.-J. DNA Methylation Aberrancies Delineate Clinically Distinct Subsets of Colorectal Cancer and Provide Novel Targets for Epigenetic Therapies. *Oncogene* **37,** 566–577. ISSN: 1476-5594 (Feb. 2018).

383. Weisenberger, D. J. *et al.* CpG Island Methylator Phenotype Underlies Sporadic Microsatellite Instability and Is Tightly Associated with BRAF Mutation in Colorectal Cancer. *Nature Genetics* **38,** 787–793. ISSN: 1061-4036 (July 2006).

384. White, A. *et al.* A Review of Sex-Related Differences in Colorectal Cancer Incidence, Screening Uptake, Routes to Diagnosis, Cancer Stage and Survival in the UK. *BMC Cancer* **18.** ISSN: 1471-2407 (Sept. 2018).

385. Wu, C. & Bekaii-Saab, T. CpG Island Methylation, Microsatellite Instability, and BRAF Mutations and Their Clinical Application in the Treatment of Colon Cancer. *Chemotherapy Research and Practice* **2012.** ISSN: 2090-2107 (2012).

386. Xu, C. A Review of Somatic Single Nucleotide Variant Calling Algorithms for Next-Generation Sequencing Data. *Computational and Structural Biotechnology Journal* **16,** 15–24. ISSN: 2001-0370 (Jan. 2018).

387. Yang, X. *et al.* Widespread Expansion of Protein Interaction Capabilities by Alternative Splicing. *Cell* **164,** 805–817. ISSN: 0092-8674 (Feb. 2016).

388. Yarden, Y. & Pines, G. The ERBB Network: At Last, Cancer Therapy Meets Systems Biology. *Nature Reviews Cancer* **12,** 553–563. ISSN: 1474-1768 (Aug. 2012).

389. Yeger-Lotem, E. & Sharan, R. Human Protein Interaction Networks across Tissues and Diseases. *Frontiers in genetics* **6,** 257. ISSN: 1664-8021 (2015).

390. Yi, S. *et al.* Functional Variomics and Network Perturbation: Connecting Genotype to Phenotype in Cancer. *Nature Reviews Genetics* **18,** 395–410. ISSN: 1471-0056 (Mar. 2017).

391. You, Y. N., Rustin, R. B. & Sullivan, J. D. Oncotype DX® Colon Cancer Assay for Prediction of Recurrence Risk in Patients with Stage II and III Colon Cancer: A Review of the Evidence. *Surgical Oncology* **24,** 61–66. ISSN: 0960-7404 (June 2015).

392. Young, J. P. *et al.* Rising Incidence of Early-Onset Colorectal Cancer in Australia over Two Decades: Report and Review. *Journal of Gastroenterology and Hepatology* **30,** 6–13. ISSN: 1440-1746 (2015).

393. Yu, H., Kim, P. M., Sprecher, E., Trifonov, V. & Gerstein, M. The Importance of Bottlenecks in Protein Networks: Correlation with Gene Essentiality and Expression Dynamics. *PLoS Computational Biology* **3.** ISSN: 1553-734X (Apr. 2007).

394. Zadra, G. *et al.* Inhibition of de Novo Lipogenesis Targets Androgen Receptor Signaling in Castration-Resistant Prostate Cancer. *Proceedings of the National Academy of Sciences of the United States of America* **116,** 631–640. ISSN: 1091-6490 (Jan. 2019).

395. Zhang, B. *et al.* Proteogenomic Characterization of Human Colon and Rectal Cancer. *Nature* **513,** 382–387. ISSN: 1476-4687 (Sept. 2014).

396. Zhang, K., Geng, W. & Zhang, S. Network-Based Logistic Regression Integration Method for Biomarker Identification. *BMC Systems Biology* **12,** 135. ISSN: 1752-0509 (Dec. 2018).

397. Zhong, Q. *et al.* Edgetic Perturbation Models of Human Inherited Disorders. *Molecular Systems Biology* **5,** 321. ISSN: 1744-4292, 1744-4292 (Jan. 2009).

398. Zhou, H. & Rigoutsos, I. The Emerging Roles of GPRC5A in Diseases. *Oncoscience* **1,** 765–776. ISSN: 2331-4737 (Nov. 2014).