PLEASE TYPE

THE UNIVERSITY OF NEW SOUTH WALES
Thesis/Dissertation Sheet

Surname or Family name: Pinese

First name: Mark Other name/s:

Abbreviation for degree as given in the University calendar: PhD

10mm

School: St. Vincent's Clinical School Faculty: Medicine

Title: Molecular signatures of survival in pancreatic ductal adenocarcinoma: why do some patients live longer than others, and how can we identify them?

Abstract 350 words maximum: (PLEASE TYPE)

Pancreatic ductal adenocarcinoma (PDAC) is the most deadly common cancer, with fewer than ten percent of patients surviving five years from diagnosis. In contrast to many other cancers, the grim prognosis of PDAC has remained largely the same over the past thirty years, reflecting our relatively poor understanding of the disease. This situation may soon change: recent large-scale efforts have produced detailed molecular and clinical data for almost a thousand cases of PDAC, finally enabling detailed dissection of the cancer's molecular basis. This work used these new data to directly address two questions linked to PDAC's exceptionally poor prognosis: what are the biological processes that drive poor prognosis in PDAC? and can we develop a practical and effective prognostic staging system to better guide PDAC treatment?

Chapter 2 considered the first question, using a gene expression factorisation approach to identify two metagenes that determine survival in PDAC. These metagenes were linked to the biological processes of proliferation, and the EMT, and were also prognostic in a number of other solid tumours, revealing these processes as generally important to cancer patient survival.

Chapter 3 addressed the second question, developing a pilot tool to estimate the prognosis of a PDAC patient, and guide the decision of whether to resect that patient's tumour. The tool is unique in that it can be applied prior to resection, and its validation performance was competitive with gold standard post-resection methods. This performance might be improved through better selection of prognostic biomarkers; this is the subject of Chapter 4, which describes a method for the identification of biomarkers that are well-suited for development into clinical tests.

PDAC has an exceptionally poor prognosis, but better understanding of the disease, and improvements in its management, can yield incrementally better outcomes. The work presented here focused on understanding and predicting prognosis in PDAC. It identified two metagenes that were strongly linked to outcome, indicating valuable therapeutic directions to improve the survival of patients with particularly aggressive tumours; and it provided a proof-of-concept, and a method for development, of a biomarker-based preoperative prognostic tool for the improved management of PDAC.

Declaration relating to disposition of project thesis/dissertation

I hereby grant to the University of New South Wales or its agents the right to archive and to make available my thesis or dissertation in whole or in part in the University libraries in all forms of media, now or here after known, subject to the provisions of the Copyright Act 1968. I retain all property rights, such as patent rights. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

I also authorise University Microfilms to use the 350 word abstract of my thesis in Dissertation Abstracts International (this is applicable to doctoral theses only).

....
Signature Witness Date

The University recognises that there may be exceptional circumstances requiring restrictions on copying or conditions on use. Requests for restriction for a period of up to 2 years must be made in writing. Requests for a longer period of restriction may be considered in exceptional circumstances and require the approval of the Dean of Graduate Research.

FOR OFFICE USE ONLY Date of completion of requirements for Award:

THIS SHEET IS TO BE GLUED TO THE INSIDE FRONT COVER OF THE THESIS

Molecular signatures of survival in pancreatic ductal adenocarcinoma

Why do some patients live longer than others, and how can we identify them?

Mark Pinese

A thesis in fulfilment of the requirements for the degree of
Doctor of Philosophy

St. Vincent's Clinical School
Faculty of Medicine

March 2015

ORIGINALITY STATEMENT

‘I hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by another person, or substantial proportions of material which have been accepted for the award of any other degree or diploma at UNSW or any other educational institution, except where due acknowledgement is made in the thesis. Any contribution made to the research by others, with whom I have worked at UNSW or elsewhere, is explicitly acknowledged in the thesis. I also declare that the intellectual content of this thesis is the product of my own work, except to the extent that assistance from others in the project's design and conception or in style, presentation and linguistic expression is acknowledged.’

Signed …….....

Date …….....

To Emma and Eva –
it's done now!

Acknowledgements

As tempted as I was at times to follow the path of Shivers' outstanding non-acknowledgement [89], in truth this work would simply not be without the myriad contributions of friends, family, and colleagues.

I am indebted to Prof. Andrew Biankin, for teaching me some of the rules of professional science, and for his great accomplishment in piloting the APGI to success; without the data generated by that project, this dissertation would not have been possible. Likewise, Prof. Sean Grimmond and his group at the Institute of Molecular Bioscience were instrumental in generating the lion's share of the expression data upon which I worked.

Survival analysis requires exacting interpretation, coding, and follow-up of clinical records; laborious and difficult work that is too often left unacknowledged. I am very grateful to the entire clinical data team of the APGI/NSWPCN, for thanklessly collating the high-quality data which I required, and for humouring my continual requests for data. In particular, this work would be far weaker without the substantial contributions of Clare Watson, Skye Simpson, Mary-Anne Brancato, and Amber Johns.

Dr. David Chang taught me so much about true translational research; his work and guidance continues to transform the predictor of Chapter 3 from an academic exercise, into something that could truly make a positive difference for patients. His contributions to the greater prognostic predictor project far outweigh mine, and his long efforts on molecular prognostics in pancreas cancer enabled the small work I describe here.

Dr. Mark Cowley, I suspect, was not quite aware of the task ahead when he kindly agreed to supervise me following Prof. Biankin's departure. This dissertation is stronger for his critical input, and his patient help and encouragement at all times was far better than I think I deserved.

Friends and family have given support, both overt and covert, freely through the last four years. In particular, Aaron Statham, Al-

ison Ferguson, and Drs. Hugh French and Brian Gloss, have made contributions to my mental health – usually administered at Darlo Bar – that go beyond the call of duty. I have also been very fortunate to have had the unquestioning support of the Pinese and Scott families.

Scholarships from the Australian government, UNSW Australia, and the Garvan Institute meant that I did not need to cut back on my food habit in order to meet my basic coffee needs.

Too many people have assisted this work, directly or indirectly, for me to list them all here. To all of these, I apologise if I have omitted you – remind me sometime, and I promise you a drink of your choice. Some I suspect have been assisting in secret. To those, you really should have seen this omission coming, but thanks all the same.

It's traditional to save the best to last; in this case it's also amusing. To my truly amazing wife Emma: thank you. Though you didn't write a word, I see this dissertation as much your work as mine. Throughout this PhD, you have shown fortitude and compassion that is inspirational. I probably will never know all the secret machinations that you undertook to make this dissertation happen – at least not until I receive an itemised bill – but I want you to know that they worked, and that this simply wouldn't have come together without you. And to little Eva: you just might read this some day, so you may as well know that you were a very early overachiever – you gave Dad a great big kick up the arse, even when you were less than two feet tall!

Sydney
23 March 2015

Abstract

Pancreatic ductal adenocarcinoma (PDAC) is the most deadly common cancer, with fewer than ten percent of patients surviving five years from diagnosis. In contrast to many other cancers, the grim prognosis of PDAC has remained largely the same over the past thirty years, reflecting our relatively poor understanding of the disease. This situation may soon change: recent large-scale efforts have produced detailed molecular and clinical data for almost a thousand cases of PDAC, finally enabling detailed dissection of the cancer's molecular basis. This work used these new data to directly address two questions linked to PDAC's exceptionally poor prognosis: what are the biological processes that drive poor prognosis in PDAC? and can we develop a practical and effective prognostic staging system to better guide PDAC treatment?

Chapter 2 considered the first question, using a gene expression factorisation approach to identify two metagenes that determine survival in PDAC. These metagenes were linked to the biological processes of proliferation, and the EMT, and were also prognostic in a number of other solid tumours, revealing these processes as generally important to cancer patient survival.

Chapter 3 addressed the second question, developing a pilot tool to estimate the prognosis of a PDAC patient, and guide the decision of whether to resect that patient's tumour. The tool is unique in that it can be applied prior to resection, and its validation performance was competitive with gold standard post-resection methods. This performance might be improved through better selection of prognostic biomarkers; this is the subject of Chapter 4, which describes a method for the identification of biomarkers that are well-suited for development into clinical tests.

PDAC has an exceptionally poor prognosis, but better understanding of the disease, and improvements in its management, can yield incrementally better outcomes. The work presented here focused on understanding and predicting prognosis in PDAC. It identified two metagenes that were strongly linked to outcome, indicating valuable therapeutic directions to improve the survival of patients with particularly aggressive tumours; and it provided a proof-of-concept, and a method for development, of a biomarker-based preoperative prognostic tool for the improved management of PDAC.

Contents

List of Figures

List of Tables

List of Algorithms

Software Versions

The following versions of software were used in all work.

MSigDB	4.0
R	3.1.1
ahaz	1.14
doMC	1.3.3
energy	1.6.2
Exact	1.4
flexsurv	0.5
GSVA	1.14.1
illuminaHumanv4.db	1.24.0
lumi	2.18.0
lumidat	1.2.3
MASS	7.3-35
maxstat	0.7-22
NMF	0.20.5
nnls	1.4
randomForestSRC	1.5.5
risksetROC	1.0.4
robustbase	0.92-3
survival	2.37-7
shiny	0.10.2.2

Conventions

Unless otherwise specified, the following conventions are used throughout this dissertation.

- Indices in algorithm pseudocode are 1-based.
- Logarithms (log) and exponentiations (exp) are to base e.
- Square brackets around a predicate P denote the Iverson bracket: $[P] \Leftrightarrow 1$ if P is true, else 0.
- Round brackets around a function-predicate pair $f(i) \mid P(i)$, indicate tuple builder notation:

  $$(f(i) \mid P(i))_{i=a}^{b} \Leftrightarrow (f(a) \text{ if } P(a), f(a+1) \text{ if } P(a+1), \ldots, f(b) \text{ if } P(b))$$

  where an element $f(i)$ is only included in the tuple if $P(i)$ is true.
- The symbol $\oplus$ is reserved for tuple concatenation:

  $$(a_1, a_2, \ldots, a_n) \oplus (b_1, b_2, \ldots, b_m) \equiv (a_1, a_2, \ldots, a_n, b_1, b_2, \ldots, b_m)$$

- x_+ indicates the value of the ramp function at real x, $x_+ \equiv \max(0, x)$.
- $\mathbf{0}^n$ denotes the vector in $\mathbf{R}^n$ with all entries equal to zero.
- $\mathbb{B}$ denotes the Boolean set $\{true, false\}$. For convenience, Boolean values are also sometimes represented using 0 and 1, as $0 \Leftrightarrow false$, $1 \Leftrightarrow true$.

Chapter One

Foreword

Pancreas cancer is a disease of great importance. Despite being relatively uncommon, the exceptional lethality of most pancreas cancers makes them the 5th most frequent cause of cancer death in Australia [2]. The most common and aggressive form of pancreas cancer is pancreatic ductal adenocarcinoma (PDAC), which carries a five year survival rate of approximately 5% [1]. This dire prognosis has remained largely unchanged over the past thirty years [1, 52], highlighting how little success the tools of modern molecular oncology have had in treating this poorly-understood disease. PDAC is a challenging disease to study, and relatively little is known about its molecular basis. This may soon change: current cancer genome projects have now accrued large cohorts of PDAC samples and clinical data, finally enabling the systematic dissection of this serious malignancy. The work described here used these new data to gain a better understanding of why PDAC is so aggressive, and in particular why some patients survive a relatively long time, whereas others succumb quickly to their disease. Knowledge of the biological processes that mirror patient survival can improve patient staging and disease management in the medium term, and highlights target areas for the development of more effective therapies against this difficult disease.

Pancreas cancer is the most aggressive common malignancy, with a worse survival rate than even mesothelioma [1]. The exceptionally poor prognosis of pancreas cancer means that although it is the 12th

most commonly-diagnosed cancer in Australia, it is the 5th most common cause of cancer death; worldwide, almost 360,000 deaths are expected to occur to pancreas cancer in 2015 [30]. The majority of pancreas cancers are PDAC, a malignancy believed to originate in the exocrine compartment of the pancreas [85]. PDAC is strikingly resistant to treatment, and modern molecular oncology has yet to make much of an impact on the disease: despite extensive research, survival with pancreas cancer has not greatly improved in the last thirty years (Figure 1.1).

The current standard of care for PDAC is based on resection of the primary tumour, with accompanying chemotherapy or chemoradiotherapy [27]. Unfortunately, resection is seldom possible, as approximately 80% of patients present either with metastatic disease, or tumours that are too locally involved to safely resect [85]. Even in the best case of a complete resection of a localised tumour, followed by thorough adjuvant chemotherapy, most patients will relapse with distant metastases that are ultimately fatal [10]. From this, two points are evident: even a complete resection rarely removes all disease, and current chemotherapy regimes are incapable of destroying the remaining transformed cells. Little ground can be made on the first point, as the residual disease is likely due to the presence of undetected micro-metastases, which would be impossible to resect. To address the second point, and improve the outcomes of patients with PDAC, more effective chemotherapy regimes need to be developed. This need is especially acute, as surgery to remove local PDAC is a major procedure with significant associated morbidity [49], and currently many patients who undergo resection receive little benefit from it. To date, efforts to find improved chemotherapy for PDAC have been hampered by a poor understanding of the molecular nature of PDAC, and the challenges particular to the study of this disease.

PDAC is challenging to study. It is a relatively rare cancer, in a tissue that is difficult to access, and unethical to biopsy in healthy people. Additionally, very few samples of early PDAC are available, as the disease is usually only diagnosed at an advanced stage, and

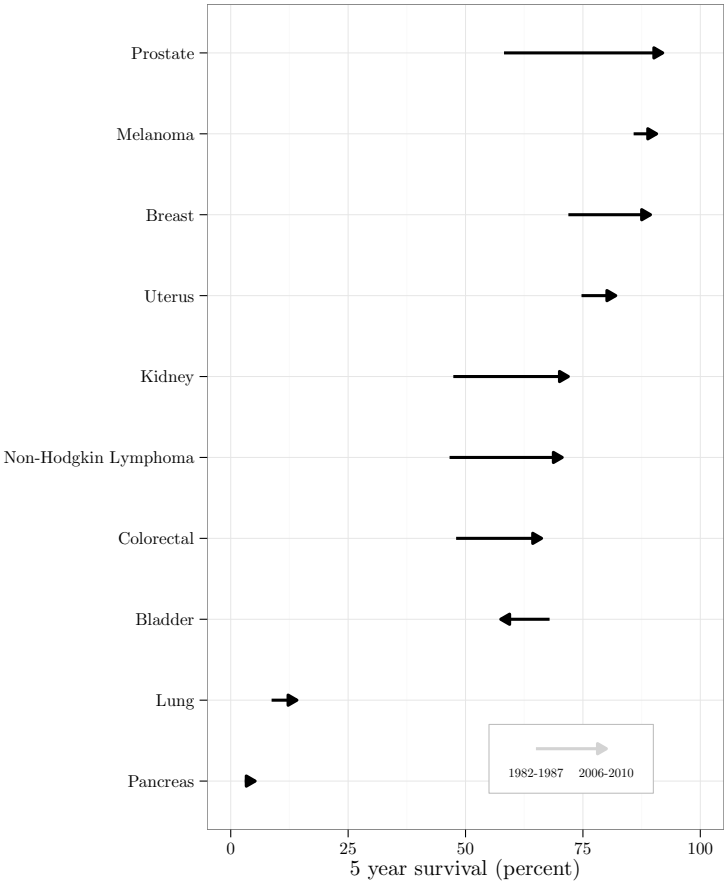


Figure 1.1: Changes in survival rates for common cancers in Australia between 1982-1987, and 2006-2010. The cancers shown are the ten most common by incidence; arrows move from the 1982-1987 survival rates to the 2006-2010 rates. Most cancers had substantial increases in patient survival between the 1980s and the 2000s; pancreas cancer was an exception, with survival rates barely changing over the past 30 years. Source: AIHW cancer survival and prevalence [1].

therefore the early stages of PDAC carcinogenesis have been particularly challenging to map [53]. Molecular investigations are hampered by the presence of large quantities of desmoplastic stroma in bulk tumour samples; this stroma often significantly outnumbers the transformed epithelial cells, and dominates biomarker signals in the tissue [25]. Mouse models of PDAC have been developed that appear to recapitulate aspects of human disease. However, a number of alternative models exist, with differing molecular characteristics [110], and it's unclear which, if any, is the best proxy for the human disease. If molecular subtypes truly exist within PDAC, a number of mouse models may in fact be appropriate simulators of human disease, with each mimicking a different subtype; there is an indication that this may in fact be the case [25].

Despite these challenges to the study of PDAC, painstaking work over the last three decades had developed a mostly complete picture of the pathogenesis and characteristics of the disease. PDAC appeared to be associated with a relatively small defined set of mutations, that arose in a consistent fashion from precursor lesions initiated from an acinar progenitor cell [48, 72, 60]. Clinical prognostic factors of PDAC were well-studied [32, 82], and combined into a prognostic nomogram [18], and a very large number of putative prognostic biomarkers had been reported [43]. However, this great depth of knowledge into the disease did not translate into better outcomes in the clinic, with the most notable development in the treatment of PDAC being the introduction of adjuvant gemcitabine, to modest ultimate effect on survival rates (Figure 1.2).

An enticing hypothesis to explain the poor performance of so many chemotherapeutic strategies is that PDAC is not a homogeneous disease, but rather a complex mix of many subtypes, each with a type-specific drug sensitivity profile. Research through the 2000s largely did not support this notion; although histological and molecular variants of PDAC had been described, the majority of malignancies followed a depressingly consistent clinical path [72]. The challenges inherent in PDAC research in particular, and subtype discov-

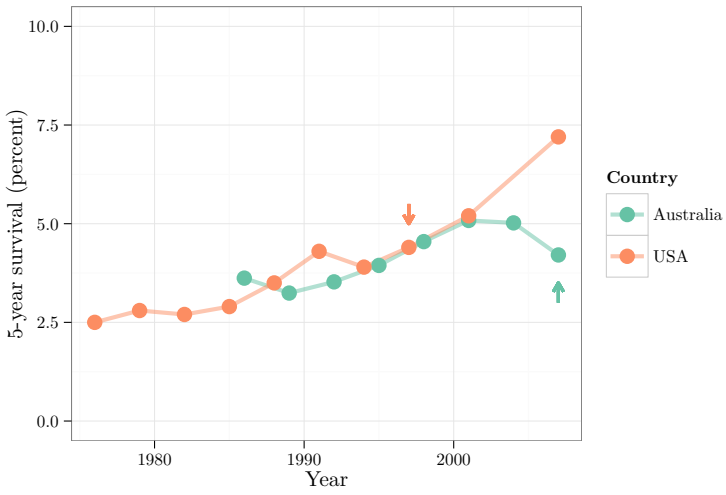


Figure 1.2: Pancreas cancer 5-year survival rates have remained largely stable over time. Arrows indicate the year of regulatory approval for the use of gemcitabine adjuvant chemotherapy in each country. Sources: AIHW cancer survival and prevalence [1], and NCI SEER CSR [52].

ery in general, meant that only in 2011 was a report made describing the systematic discovery of three possible subtypes of PDAC, from a small data set of 63 samples [25]. Tumours from these subtypes were clinically distinct, reinvigorating the search for molecular subtypes of PDAC in larger and better-controlled cohorts.

New large-scale cancer genome efforts promise to finally establish the molecular nature of PDAC in man. Combined, The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC) have accrued approximately one thousand cases of PDAC, with extensive linked molecular measurements and clinical annotation. These projects are still in progress, but early analyses of data to date have indicated that PDAC is a relatively genetically homogeneous disease [14], with the primary genetic variation between tumours being in the structural stability of their DNA [109]. This diversity in genomic stability is an exciting result, which immediately suggests a therapeutic direction [109], but is unlikely to be the last

word in PDAC subtypes.

Subtype information is often challenging to learn from genomic data, as these data can have low sensitivity, and a given change in biological behaviour can be caused by numerous different genetic alterations. Gene expression (GEX), on the other hand, is relatively easy to quantify accurately, and serves as a composite signature of biological state that is particularly well suited to elucidating subtypes and molecular status [80]. Examination of the ICGC genomic alterations [14, 109] suggests that PDAC is either a very homogeneous disease, with only one subtype, or an extraordinarily heterogeneous one, with dozens. However, this conclusion is plausibly due to the focus of those analyses on genetic alterations; to date the GEX data generated by those studies has been neglected. A careful analysis of the GEX data available for PDAC, in the spirit of early very successful work in breast cancer, promises to define disease subtypes of PDAC of biological interest and clinical relevance.

This dissertation considered one aspect of this subtype analysis, using large-cohort GEX data to comprehensively characterise the molecular basis of differential survival in PDAC. Three core questions arose during the work:

- what biology underpins differential survival with PDAC?
- is it practical to predict an individual patient's future survival in a clinical setting? and
- how can we best mine these new large datasets for biomarkers to develop into clinical tools?

These questions, all based on the theme of survival in PDAC, are addressed in the three main chapters of this thesis.

Chapter Two

Signatures of Survival Processes in Pancreas Cancer

Thesis: Specific molecular processes control survival of patients with resectable pancreatic ductal adenocarcinoma, and these processes can be identified using gene expression data.

Summary Very little is known regarding the biological processes that control the survival of patients with PDAC, the most common and aggressive form of pancreas cancer. The range of relative patient survival times that is observed in practice is not well explained by extrinsic factors such as age at diagnosis, and perhaps instead reflects differences in the biological processes operating within each tumour. Recent molecular profiling work [25] has identified possible molecular subtypes within the previously homogeneous group of PDAC, but these subtypes have not achieved the maturity or clinical application of those in breast cancer, and their discovery and validation has been hampered by ad-hoc methodology, and the lack of large, well-curated cohorts of PDAC samples. The recently-compiled Australian Pancreatic Cancer Genome Initiative (APGI) cohort contains the largest group of clinically annotated PDAC samples, with

accompanying GEX and high-quality follow-up data, in the world. It presents a unique opportunity to apply modern techniques for prognostic signature identification to the discovery of biological processes that drive the clinical course of pancreas cancer. These signatures may find application as prognostic tools in their own right, but more importantly can supply much-needed information on the fundamental biology of the one common cancer that has, to date, been almost entirely refractory to all the tools of modern molecular medicine.

2.1 Introduction

Despite extensive research, PDAC remains a poorly-understood disease. Recent genomic profiling has revealed the genetic alterations that accompany the cancer [14], and a huge number of prognostic factors are known [43], but these findings have shed little light on the fundamental disease processes at work in individual tumours. This is a consequence of genetic and biomarker data being poorly-suited for understanding the biological state of a cell: although genetic alterations are central to the etiology of cancer, they give incomplete information on the pathways and systems actually active in a given tumour, and biomarkers supply non-causal readouts of cell state that are difficult to trace back to underlying biological processes.

Sitting between the regulatory function of transcription control, and the effector function of protein expression, GEX data integrate information from all aspects of cell condition, including genetic alterations, signalling pathway activity, and metabolic status. As such, it is unsurprising that GEX data are superior indicators of cell state, better than all other high-throughput measurement methods, such as protein expression or genetic alterations [80]. However, the involvement of GEX with so many biological inputs is also a weakness: typical differential expression studies will identify many hundreds of transcripts that vary between disease states, and the deconvolution of this complex set of hundreds of effects back to a small number of causative molecular processes remains challenging.

Disease GEX profiling studies have largely refrained from attempting to infer the state of a few molecular processes from the many hundreds of differentially-expressed genes identified. A number of factors are likely to have contributed to this reluctance: deconvolution methods require relatively large sets of high-quality measurements [70], early techniques were poorly-suited to the particular requirements of the GEX deconvolution problem, and the signature databases that assist the assignation of a biological annotation to the output from a deconvolution calculation (for example, the MSigDB [95]) are only now reaching maturity, with some areas of biology still under-represented.

A simple synthetic example illustrates the problem and process of GEX deconvolution, and the character of solutions produced by both classical and modern techniques. Consider a group of samples, each of which is in one of three distinct biological states: state A, state B, and an intermediate state. Which state a sample is in affects the expression of two genes, gene 1, and gene 2: state A is associated with higher gene 2 expression than gene 1 expression; state B with higher gene 1 expression than gene 2; and the intermediate state with low expression for both genes (Figure 2.1). From the figure it is apparent that samples lie along two lines in transcription space; these lines I term metagenes.

Accurately knowing the metagenes at work within a biological system considerably simplifies reasoning about transcription within the system. In the example of Figure 2.1, state A is associated with high metagene 1, state B with high metagene 2, and the transition state with low scores of both. Additionally, the loadings of genes on the metagenes themselves (the directions of the metagene arrows) provides information on transcriptional control within the system: metagenes define the axes along which cell state must move, and so provide a simpler and more accurate representation of cell state than the full set of gene expression measurements. Metagenes can also be considered to capture co-expressed modules of genes, with likely biological significance. The advantages of a metagene-centric perspective

to interpreting GEX become increasingly apparent as more genes are considered, and when thousands of genes are measured per sample, deconvolving the highly complex patterns of expression of thousands of genes, to only tens of metagenes, represents a powerful reduction in complexity. However, in practical use deconvolution methods must operate in thousand dimensional spaces, rather than the two dimensions in this example, and the computational and methodological complexities involved, as well as the poor results yielded by traditional approaches, have limited the application of GEX deconvolution.

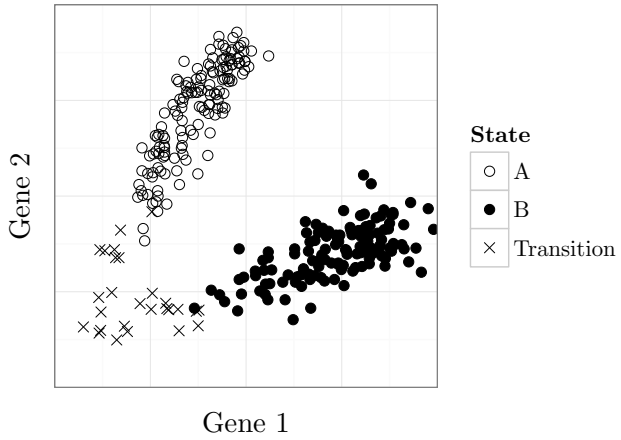
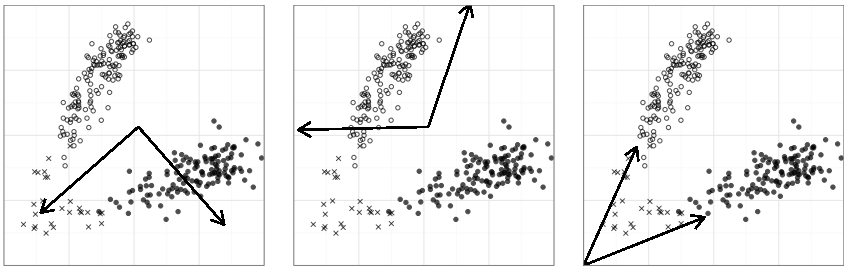


Figure 2.1: The gene deconvolution problem. Shown are the hypothetical expression levels of two genes across three biological states, where each point represents the gene expression of a single sample in one of the three biological states. State A (hollow circles) is characterised by gene 2 > gene 1; state B (solid circles) by gene 1 > gene 2; and the intermediate state (crosses) by low levels of both genes. The challenge of gene deconvolution is to automatically infer, from unlabelled data (ie state is unknown), the dominant lines of gene expression (metagenes) along which most samples lie.

A number of techniques from the field of matrix factorisation have been applied to the GEX deconvolution problem, first principal component analysis (PCA) [5], then independent component analysis (ICA) [66], and more recently various forms of non-negative matrix factorisation (NMF) [19]. A number of reports have highlighted the

unsuitability of PCA for GEX deconvolution, and the relative superiority of ICA [63, 86, 98]; this is primarily due to the PCA requirement that metagenes be orthogonal [65], a situation that is not supported by our knowledge of biology, and results in bizarre artefacts such as PCA metagenes not actually being aligned with the expression pattern of any sample (Figure 2.2(a)). Although the results from ICA are more interpretable than those from PCA, they still do not consider that GEX is a non-negative process: it is impossible to have a concentration of mRNA that is less than zero, and therefore for best interpretability we wish metagenes to have non-negative 'expression' as well. ICA does not produce solutions satisfying this requirement, and more importantly its non-Gaussianity objective is not necessarily optimal for GEX deconvolution (Figure 2.2(b)), reducing its ultimate utility. NMF techniques have the potential to produce excellent GEX decompositions (Figure 2.2(c)), but are relatively new methods that have very high computational requirements, and often require careful tuning, making their effective application challenging.

In addition to the general technical challenges of GEX deconvolution, issues particular to pancreas cancer significantly complicate attempts to identify molecular processes at work within the tumours. Pancreas cancer is challenging to sample, and mRNA in the tissue degrades rapidly once extracted, complicating sample collection. Additionally, a feature of PDAC is the presence of a dense desmoplastic stromal reaction throughout the tumour, that is formed by genetically normal patient stroma cells [71]. The fraction of tumour cells that are actually cancerous varies by more than 10-fold between tumours [14], meaning that without careful correction, gene expression profiles are dominated by stromal cell fraction signals, and not true differential expression within a cell type. Microdissection has been used to separate cancer cells from surrounding stroma in order to simplify analysis [25], but current thought in the field is that the stroma in PDAC is an essential and enabling, if not in itself neoplastic, component of the tumour [71], and that the examination of cancer cell expression in isolation ignores the likely important interplay between



(a) PCA (b) ICA (c) NMF

Figure 2.2: NMF produces a more accurate GEX decomposition than either PCA or ICA. Metagenes found by each method are shown as arrows. PCA (panel a) produces metagenes that don't match the expression pattern seen in any sample; these metagenes do not have a ready biological interpretation. ICA (panel b) accurately identifies one metagene, but the inappropriateness of the non-Gaussianity criterion for these data leads to an incorrect estimate of the other; although this solution is better than that of PCA, not all metagenes align well with biology. NMF (panel c) provides the best deconvolution; the metagenes identified closely match the expression patterns observed, and reflect the true structure of co-expression within the samples.

the two major synergistic components of a tumour: transformed epithelial cells, and genetically normal stroma.

Due to these challenges to GEX deconvolution of PDAC, to date only one study (by Collisson *et al*, published in 2011) has reported a breakdown of PDAC GEX into a small number of biological modules [25]. This study examined microdissected cancer cells only, and found that the transformed epithelial cells of PDAC could be placed into three major categories, based on their patterns of gene expression. Tumours from these three categories followed distinct clinical courses, and cell lines exhibited category-specific sensitivity to therapeutic drugs. As the first report to identify potential clinically relevant molecular subtypes within PDAC, the Collisson study was a significant advance in the understanding of the molecular processes at play within what was previously considered a homogeneous disease. However, it also possesses shortcomings that limit its clinical

utility.

Two main issues complicate the interpretation of the Collisson classes: microdissected cancer cells were used, and therefore stromal effects would be severely attenuated; and the deconvolution technique employed was tuned to achieve sample clustering, rather than GEX deconvolution. Consequently, although the Collisson classes could be a fundamental advance in the understanding of PDAC, they necessarily do not consider the full context of the disease, and potentially have artifically identified subgroups when in reality a smooth continuum of disease types may exist. Additionally, although the Collisson tumour subgroups were observed to follow different clinical courses, they were not explicitly generated to stratify patients by outcome, and so may not have captured the full biology underlying differential survival in PDAC.

A substantial gap remains in our molecular understanding of PDAC: little is known about the core molecular processes at work within both the cancer and stroma of different tumours, and almost nothing on those processes that control patient survival following diagnosis. Such a gap in knowledge is not merely of academic interest: a better understanding of the processes affecting patient survival can lead directly to improved methods for staging, may stratify patients for customised therapies, and even suggest targets for therapeutics capable of transforming a poor-prognosis cancer into a good-prognosis one. The primary obstacle for the identification of these survival-associated processes in PDAC is one of data: a large, high-quality dataset of GEX measurements and associated well-curated clinicopathological variables (CPVs) is needed. The APGI cohort addresses this data problem for the identification of fundamental survival processes in PDAC. As the largest cohort of PDAC samples ($n = 110$ for a homogeneous, well-annotated PDAC subset), with accompanying GEX and curated CPVs, in the world, it can provide the data quality and cohort size required by modern GEX deconvolution techniques.

In this chapter I describe the application of NMF for the GEX deconvolution of genes associated with outcome. The metagenes thus

identified represent orthogonal coordinately-expressed sets of genes which I then map to biological annotations, identifying the fundamental processes that may be involved in controlling the clinical course of a patient's pancreas cancer. The results of this work are directly applicable as signatures of survival time following diagnosis of PDAC, identify discrete biological processes that appear to determine outcome with pancreas cancer, and highlight fertile future avenues for research into this poorly-understood disease.

2.2 Results

Survival-associated metagenes were identified by selecting the set of genes which had GEX associated with outcome in the APGI cohort, and then performing NMF factorisation to deconvolve the full matrix of gene expression signals into a small set of metagenes. Metagenes were found to fall into patterns defining two axes of outcome-associated cell state. These prognostic axes were then tested for association with clinical course and other CPVs, as well as known general prognostic signatures, and their prognostic ability was validated in a range of cancers by testing in separate cohorts. The two prognostic axes were then correlated with biological process signatures to associate axis scores with the activity of biological processes.

Cohort characteristics and subsetting

228 unique patients from the APGI cohort had both GEX and follow-up data; for the discovery of metagenes specifically associated with PDAC survival these were subset to patients with histologically confirmed PDAC, who did not receive treatment prior to resection, did not suffer perioperative mortality, and were treated within Australia. This subsetting produced a homogeneous 110-patient APGI discovery cohort, which was used for all metagene discovery work.

General characteristics of both the full APGI cohort, and the 110-patient PDAC APGI discovery cohort, are summarised in Table 2.1.

Table 2.1: Characteristics of the full APGI patient cohort, and the homogeneous PDAC-only subset used for signature discovery. Ordinal variables are shown as median, with quartiles in parentheses. Categorical variables for which percentages do not add up to 100% indicate the presence of minor unlisted categories. Abbreviations: AAC - ampullary adenocarcinoma; IPMN - intraductal papillary mucinous neoplasm; PNET - pancreatic neuroendocrine tumour; PR - Puerto Rico

Characteristic		Full APGI	Discovery
Number of patients		228	110
Gender	Male	54.8%	54.6%
Ethnicity	Caucasian	92.3%	95.4%
	Asian	6.4%	4.6%
	African	0.9%	0%
Treatment country	Australia	86.0%	100%
	USA / PR	12.7%	0%
Age at diagnosis	(years)	68 (60 - 75)	67 (61 - 73)
Procedure	Whipple	63.2%	71.8%
Excision margin status	R0	76.8%	62.7%
	R1	20.6%	22.7%
	R2	2.6%	14.6%
Histological type	PDAC	61.8%	100%
	AAC	11.0%	0%
	IPMN	5.7%	0%
	PNET	5.7%	0%
Histological grade	1	12.0%	7.3%
	2	55.8%	64.6%
	3	30.1%	27.3%
	4	2.1%	0.8%
Location	Head	64.0%	84.6%
	Ampulla	11.4%	0%
	Tail	11.0%	8.2%
	Body	5.7%	6.4%
Size of longest axis	(mm)	33.0 (24.5 - 45.0)	35.0 (28.0 - 45.0)
Invasion	Perineural	70.3%	88.1%
	Vascular	62.4%	67.9%
Node involvement		69.3%	77.1%
Disease-specific death		52.6%	63.6%
Length of follow-up	(days)	614 (366 - 888)	632 (402 - 912)

Two axes predict survival with resectable pancreatic cancer in multiple cancers

Probe selection In order to focus the GEX deconvolution method on finding outcome-associated metagenes, it was necessary to filter the full set of gene expression data to only contain those genes that were likely to be associated with patient survival.

Unsupervised filtering to remove lowly-expressed, invariant, and redundant probes yielded APGI cohort gene expression measurements for 13,000 genes, of which 361 were identified to be associated with time from diagnosis to disease-specific death (DSD) by sure independence screening (SIS)-feature aberration at survival times (FAST), using a complementary pair subset selection (CPSS) wrapper to reduce false positive rate. The FAST statistic was chosen for its speed and ability to identify quite general relationships between a continuous variable and outcome [34], while avoiding the well-known loss of statistical power that comes from discretising continuous expression values [84].

50 variable selection runs on permuted data gave a median number of selected genes of 87.5, resulting in an estimated false-discovery rate (FDR) for the selection procedure of approximately 25%. Variable selection was tuned to achieve this relatively high FDR (see methods for details on selection thresholds), in an attempt to ensure that even genes for which expression was only weakly prognostic, were included.

Prognostic genes factorised into six metagenes NMF was used to reduce the complex expression patterns of 361 survival-associated genes into a small number of metagenes. NMF aims to approximate a non-negative gene $\times$ sample GEX matrix A by a product of low-rank non-negative matrices W and H, $A \approx WH$. The gene $\times$ metagene matrix W, termed the basis matrix, stores the contribution of each gene's expression to each metagene, whereas the metagene $\times$ sample matrix H, termed the coefficient matrix, contains the 'expression' of each metagene in each sample. The NMF procedure is highly sensi-

tive to the choice of the rank of W and H (the number of metagenes) – an incorrect rank will lead to metagenes inappropriately being either combined, or split.

The expression of the 361 survival-associated genes across the 110 patients of the APGI PDAC cohort was decomposed into metagenes by the sparse non-negative matrix factorisation, long variant (SNMF/L) NMF algorithm. The number of metagenes (factorisation rank) was automatically estimated to be 6, being the lowest rank for which the improvement in estimation error achieved by adding the next rank, was less than that observed for permuted data (Figure 2.3).

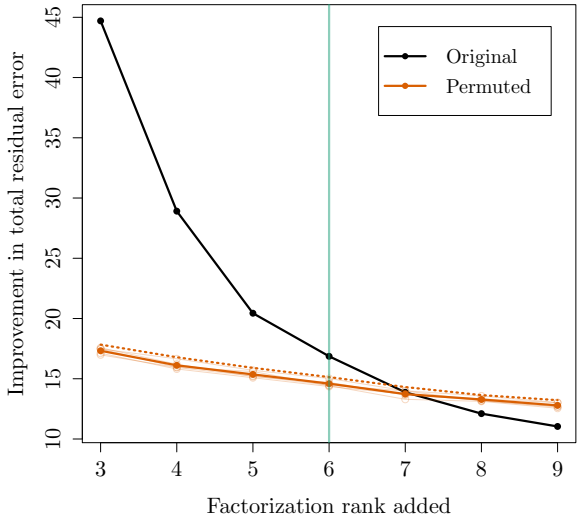


Figure 2.3: Automatic selection of factorisation rank. SNMF/L was performed for varying ranks on either unpermuted data (black line) or data permuted within samples (orange lines), and the improvement in total residual approximation error $\|A - WH\|_F$ calculated. The highest added rank for which the error improvement on unpermuted data exceeded that of permuted data plus two standard deviations (threshold shown by dotted orange line) was the final selected rank (indicated by a green line).

500 random restarts of rank 6 SNMF/L were then performed on the survival-associated gene matrix to yield the final factorisation.

The resultant clustering consensus matrix was stable (Figure 2.4), and the basis matrix W was reasonably sparse (Figure 2.5). Sparsity of the basis matrix is a desirable condition for this analysis, as it indicates that metagenes are largely distinct transcriptional modules, with little overlap in terms of shared transcripts with high loadings; SNMF/L was selected against alternative NMF algorithms as its design favours solutions with sparse W. A table of values of the basis matrix W is available as app:sigs-w-matrix on page 123.

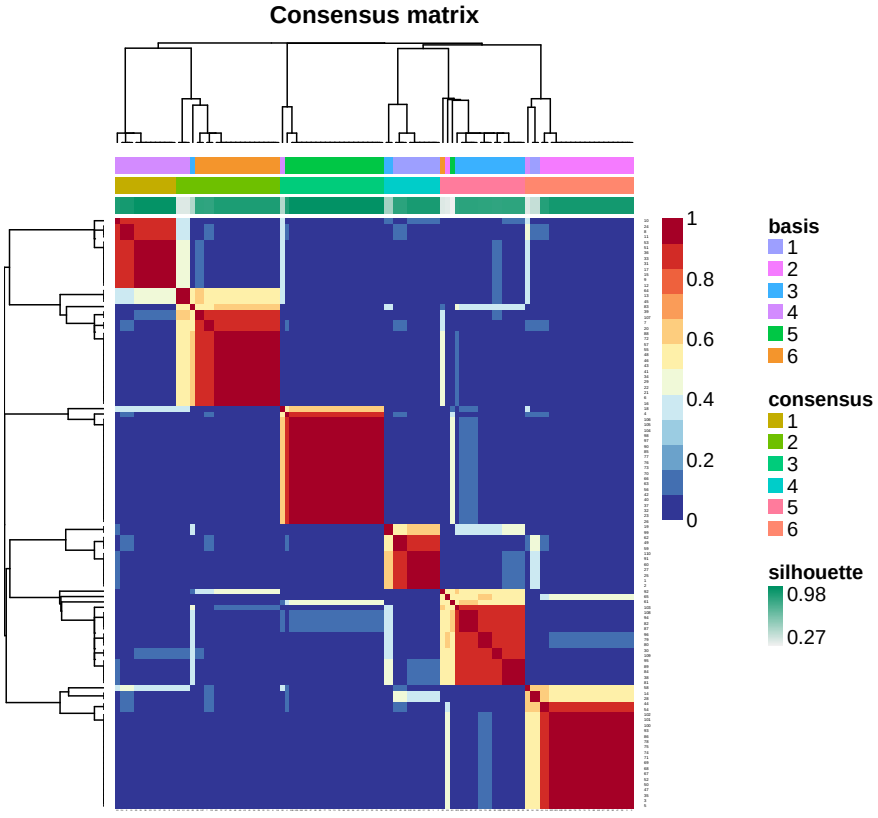


Figure 2.4: Clustering consensus matrix for the final rank-6 clustering. Colours indicate the stability of gene (in rows) and sample (in columns) clusters across random restarts of the factorisation; at rank 6 this factorisation was largely stable, with identical clusters assigned in all 500 random restarts to the majority of genes and samples.

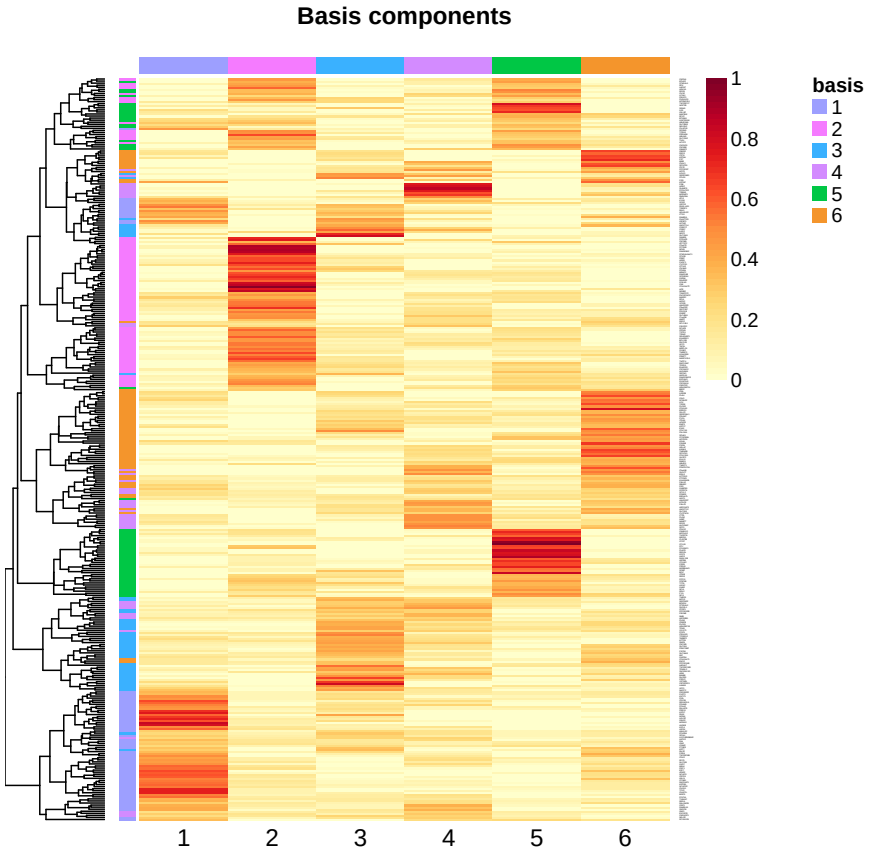


Figure 2.5: Basis matrix W of the final SNMF/L factorisation. Rows represent genes, and columns metagenes, with cell colours proportional to the loading of a given gene on a given metagene. The loadings are sparse within rows, indicating that the metagenes are modular, each affecting the expression of largely distinct sets of target genes. A table of values of this basis matrix is available as app:sigs-w-matrix on page 123.

Three metagenes together formed a prognostic model The transcription patterns of genes associated with survival in the APGI cohort could be decomposed into just six largely distinct metagenes. Due to the presence of false positives in the 361 screened input genes, some of the metagenes will have no strong association with outcome. To identify which of the six metagenes were ultimately predictive of patient survival, I performed least absolute shrinkage and selection operator (LASSO) regression on the 110-patient APGI discovery cohort data, using non-negative least squares (NNLS)-estimated coefficients of each of the six metagenes as marginal predictors of outcome. The LASSO regularisation parameter λ was chosen by 10-fold cross-validation to be the highest value for which the mean test set partial likelihood deviance was within one standard error of the lowest mean value. This resulted in a final model in which three metagenes, MG1, MG2, and MG5, were selected as prognostic (Figure 2.6).

Prognostic metagenes define two axes of cell transcription Further investigation of the three prognostic metagenes revealed that they were associated: APGI patient coefficients for pairs MG1 and MG5, and MG2 and MG6 (the latter not selected by the LASSO), were mutually exclusive (Figure 2.7, Kendall's τ test $P < 1 \times 10^{-6}$ for each pair). This suggested that both metagenes in each pair captured the signal of a single axis of cell behaviour, with one measuring activation of the axis, and the other deactivation. For subsequent work I therefore combined the signals of the metagenes within each axis, to give axis activity summaries: Axis A1 activity = MG1 coefficient − MG5 coefficient; Axis A2 activity = MG6 coefficient−MG2 coefficient. Activation values for axes A1 and A2 were uncorrelated, indicating that these axes were orthogonal processes operating in the APGI cohort tumours (Figure 2.8, Kendall's τ test $P = 0.21$). Metagenes MG3 and MG4 also formed a mutually exclusive pair (not shown), but were not investigated further, as neither was determined to be prognostic by the metagene LASSO.

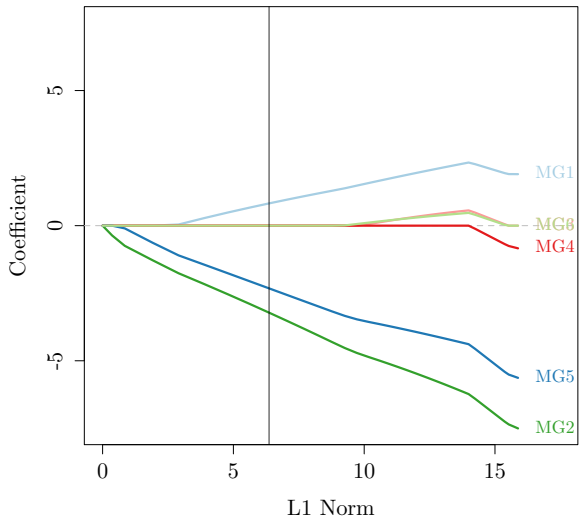


Figure 2.6: Coefficient vs penalty fit trajectories for the LASSO model predicting disease-specific survival (DSS) from metagene expression. Each line represents the model coefficient for a metagene as the model is smoothly varied from a null model (L1 norm = 0), to a full unpenalised Cox fit (L1 norm ≈ 16). The vertical line indicates the optimal value of L1 norm as selected by the 1SE criterion on 10-fold cross-validation; at this point in the trajectory only metagenes MG1, MG2, and MG5 contribute to prognosis estimates.

The PARSE score A repeat of the previous LASSO fit with 10-fold cross-validation (CV), this time using predictors of A1 activity, A2 activity, and the A1:A2 interaction, identified both A1 and A2, but not their interaction, as useful predictors of outcome. Coefficients from the LASSO fit were used to define a new risk score, the prognostic axis risk stratification estimate (PARSE), as PARSE score = $1.354 \times$ A1 activity $+ 1.548 \times$ A2 activity.

Exact calculation of the PARSE score requires the solution of a number of NNLS problems, which presents a potential barrier to use. An approximation to PARSE can be derived by relaxing the non-negative constraint; this approximation requires only a weighted mean of gene expression estimates, and is detailed in app:sigs-parse-

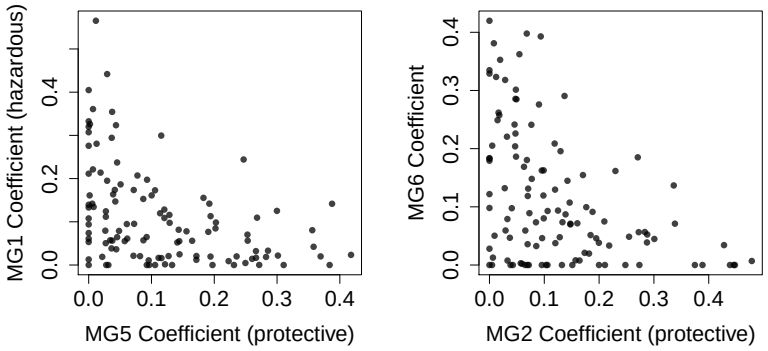


Figure 2.7: Prognostic metagenes form two axes of cell state. Metagene pairs MG1 and MG5, and MG2 and MG6, displayed mutually exclusive coefficient patterns in the APGI cohort, and could be combined to form just two axes of cell state.

approx on page 143.

Validation of the PARSE score External validation confirmed that the PARSE score was prognostic in other cohorts, including in cancers other than PDAC. PARSE score was significantly prognostic in PDAC cohorts GSE28735 [113] (LRT $P = 0.0149$) and TCGA paad (LRT $P = 0.0156$), but not in GSE21501 [94] (LRT $P = 0.115$). When assessed against all TCGA cancers for which at least 50 patients had both an event and complete RNASeq data, the PARSE score was also significantly prognostic for head and neck squamous cell carcinoma, kidney renal clear cell carcinoma, lower grade glioma, and lung adenocarcinoma, at a 5% familywise error rate (FWER) (Table 2.2, column a). This significant result reflected the ability of the PARSE score to stratify patients into risk groups in a range of solid tumours, as illustrated in Figure 2.9.

Meta-PCNA is a 130-gene signature of cell proliferation that has been found to be generally prognostic in a number of cancer cohorts [106]. To exclude the possibility that PARSE score simply recapitulated the known meta-PCNA signature, I examined whether

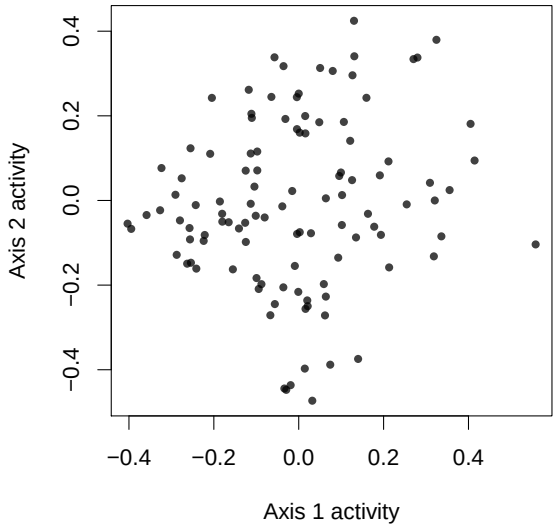
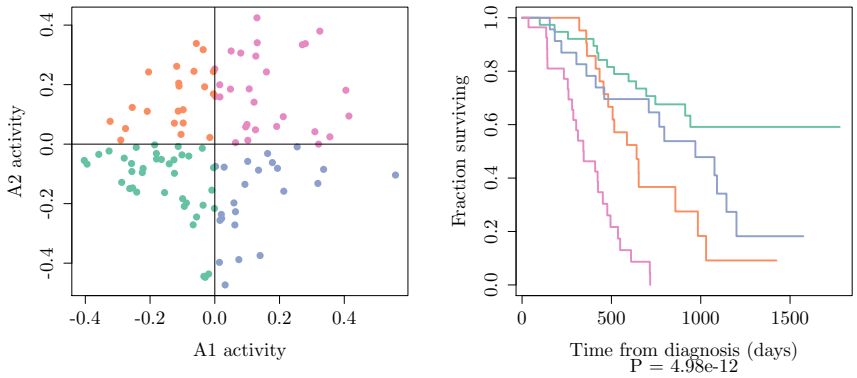


Figure 2.8: Prognostic axis signals are uncorrelated. Activity estimates of axes defined by highly correlated mutually exclusive metagene pairs (Axis A1 = MG1 - MG5, axis A2 = MG6 - MG2) were uncorrelated (Kendall τ test $P = 0.21$), indicating that these axis signals encoded orthogonal outcome-associated processes within tumours.

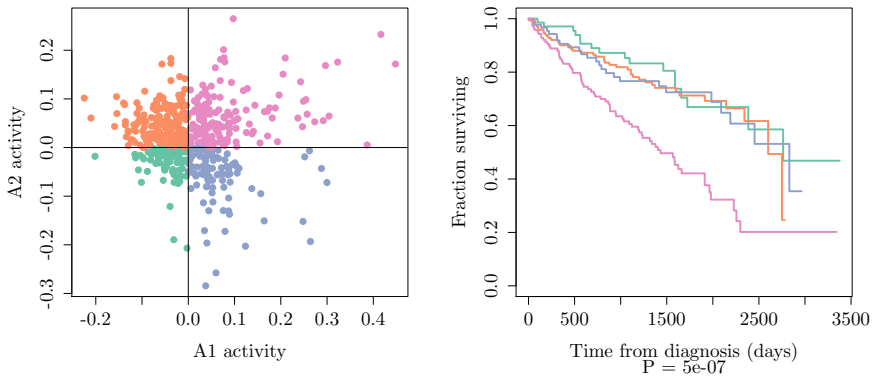
PARSE contributed additional prognostic information to meta-PCNA in the large TCGA cohorts. In TCGA kidney renal clear cell carcinoma, lower grade glioma, and lung adenocarcinoma, there was significant evidence that the PARSE score provided prognostic information beyond that given by meta-PCNA, at a 5% FWER (Table 2.2, column b).

Table 2.2: The PARSE score is prognostic in a range of TCGA cancers. P-values are from likelihood ratio tests either comparing a Cox model with PARSE score as a linear predictor, to a null model (a); or a Cox model with PARSE and meta-PCNA scores as linear predictors, against one with meta-PCNA alone (b). Shaded cells are significant at a 5% FWER following Holm's correction. TCGA study codes: *glm*: glioblastoma multiforme; *hnsc*: head and neck squamous cell carcinoma; *kirc*: clear cell kidney carcinoma; *lgg*: lower grade glioma; *luad*: lung adenocarcinoma; *lusc*: lung squamous cell carcinoma; *ov*: ovarian serous cystadenocarcinoma.

TCGA study	Number of events	Number of patients	Risk score P-value (a)	Improvement P-value (b)
gbm	54	143	0.2287	0.1587
hnsc	124	367	8.08×10^{-3}	0.0108
kirc	153	497	2.03×10^{-12}	2.89×10^{-3}
lgg	53	272	1.49×10^{-5}	7.85×10^{-3}
luad	106	431	8.34×10^{-6}	1.04×10^{-4}
lusc	117	395	0.9624	0.4110
ov	115	251	0.0238	0.0178

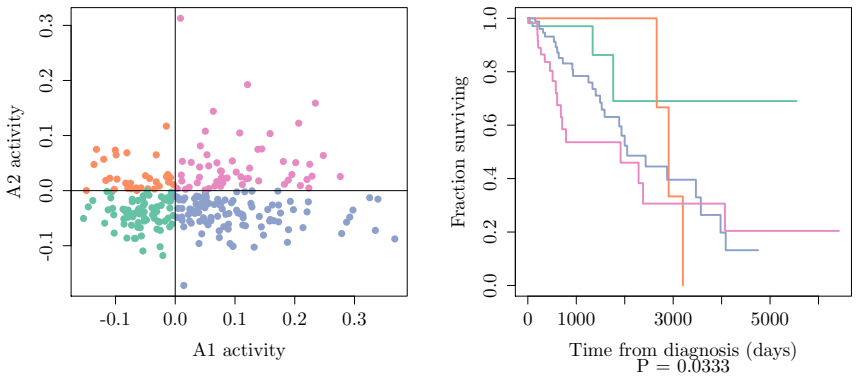


(a) APGI cohort (Resubstituted)

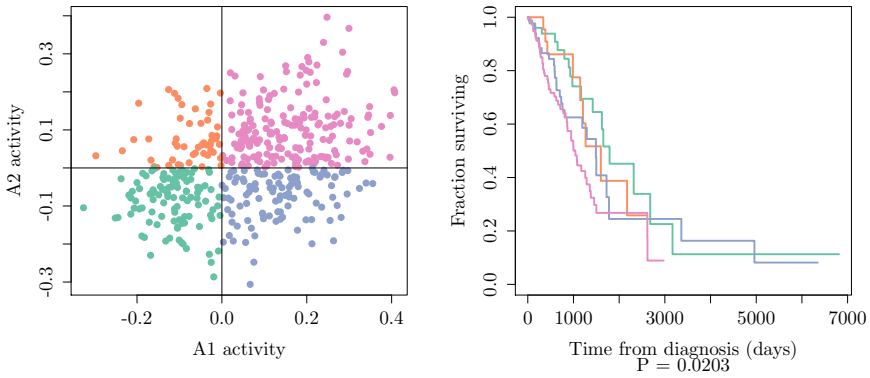


(b) TCGA kirc cohort

Figure 2.9: PARSE score axes define patient subgroups with differing outcome in a range of solid tumours. Activities for axes A1 and A2 of the PARSE score were calculated on the labelled cohorts, and patients split into four subgroups based on the sign of A1 and A2 activities (left panels). The four subgroups thus defined displayed significantly differing clinical courses (right panels). (continued...)



(c) TCGA lgg cohort



(d) TCGA luad cohort

Figure 2.9: (Concluded). PARSE score axes define patient subgroups with differing outcome in a range of solid tumours. Activities for axes A1 and A2 of the PARSE score were calculated on the labelled cohorts, and patients split into four subgroups based on the sign of A1 and A2 activities (left panels). The four subgroups thus defined displayed significantly differing clinical courses (right panels).

PARSE identifies proliferation and EMT as fundamental processes controlling survival in PDAC

To link the two prognostic axes that form the PARSE score with potential underlying biology, axis activities on the APGI discovery cohort were compared to clinical variates, known survival signatures, and scores for signatures from the molecular signatures database (MSigDB) [95].

Axis A1 PARSE axis A1 score (MG1 − MG5) was significant positively correlated with estimates of cancer cell fraction in the tumour as assessed by qPure [91] [1] (Kendall's $\tau = 0.284$, $n = 110$, Table 2.3), although the strength of this association was marginal (linear model $R^2 = 0.155$). No other CPVs were significantly associated with A1 score after correction for multiple testing (Table 2.3).

MSigDB correlations, as well as comparisons to a general proliferative signature, revealed that A1 primarily reflected the proliferative state of cells. A1 signal was very strongly correlated with meta-PCNA [106] score (Kendall's $\tau = 0.663$, $n = 110$, Figure 2.10), a relationship supported by its close association to cell cycle-related MSigDB signatures (app:sigs-msigdb-corrs-axis1 on page 135).

Axis A2 Among the clinical variables tested, PARSE axis A2 (MG6 − MG2) was negatively correlated with qPure tumour cell fraction, and positively associated with higher tumour histological grade (Table 2.3). The negative association between A2 score and tumour cell fraction is the opposite of the positive association seen with A1 score, despite high levels of both A1 and A2 being associated with poor prognosis. This reveals a potential context dependency in the influence of stromal content on survival, where high stromal content of a tumour may indicate either good or poor prognosis, depending

[1] qPure is a tool to determine cancer cell fraction in a mixed tumour DNA sample by quantification of B allele frequency (BAF) separation from single nucleotide polymorphism (SNP) genotyping array data. I contributed to the development of qPure, by proposing and designing the experiments that ultimately led to the creation of the tool, and by designing the final calibration model that links BAF separation to cancer DNA fraction.

on which underlying axis is responsible. Reflecting the poor prognosis associated with high A2, A2 score was also significantly but weakly dependent on grade: on average, A2 signal was 0.1103 higher in grade 3 or 4 tumours over grade 1 or 2, with $R^2 = 0.119$.

A number of MSigDB signatures were associated with A2 signals, among them integrins, extracellular matrix (ECM) processes, and a signature for LEF1-mediated epithelial to mesenchymal transition (EMT) (app:sigs-msigdb-corrs-axis2 on page 141). Prompted by the strong positive correlation between A2 and the LEF1 overexpression signature, I investigated the association between A2 signal and score for a general signature of EMT, meta-EMT [38]. meta-EMT and A2 signals were strongly positively correlated (Kendall's $\tau = 0.568$, $n = 110$, linear model $R^2 = 0.557$, 2.11), even when cancer cell fraction was taken into account (LRT $P = 9.4 \times 10^{-14}$), strongly indicating that A2 signal predominantly encodes EMT activity. A potential link between A2 and inflammation may also be present: A2 signal was strongly positively correlated with the gene set variation analysis (GSVA) score for MSigDB GNF2_PTX3 (Kendall's $\tau = 0.593$, app:sigs-msigdb-corrs-axis2 on page 141), a proxy for expression of the acute phase response protein pentraxin 3.

Link to S100 prognostic biomarkers Chapter 3 uses the tissue levels of proteins S100A2 and S100A4 to predict survival with resected PDAC. S100A2 and S100A4 are both considered to be biomarkers of metastasis [13, 62, 100], and so we expect their expression to be associated with the putative metastasis-related axis A2. Although neither S100A2 nor S100A4 were among the set of 361 survival-associated genes, their mRNA abundance estimates were significantly positively correlated with axis A2 signal, with Kendall's $\tau = 0.30$ ($P = 3.25 \times 10^{-6}$) for S100A2, and $\tau = 0.29$ ($P = 5.67 \times 10^{-6}$) for S100A4.

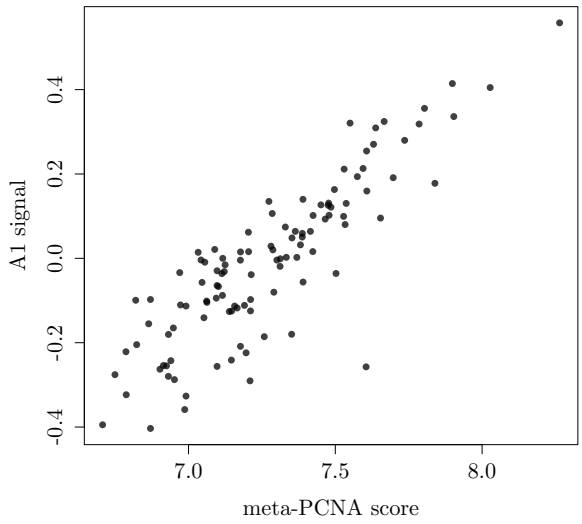


Figure 2.10: Axis A1 signal is closely associated with the meta-PCNA signature. A1 signal and meta-PCNA [106] scores were as evaluated on the APGI training set; Kendall's $\tau = 0.663$, $n = 110$, linear model $R^2 = 0.740$.

2.3 Discussion

At the molecular level, the phenomenon of cancer has long been recognised as a composite of many processes [40], however the relative importance of each process to a particular type of cancer has been largely uncertain. In pancreas cancer, a huge number of individual biomarkers are known [43], and attempts have been made to stratify cancers into empirical molecular subtypes [25], but no studies have yet provided a comprehensive analysis of which basic hallmarks of cancer are actually important in determining patient outcome. This chapter fills that gap in knowledge, by exhaustively identifying proliferation and the EMT as the major molecular processes that control survival of patients with pancreas cancer.

Cancer is fundamentally a proliferative process: it is through inappropriate and continued proliferation, and the consequent destruc-

Table 2.3: Association P-values between metagenes and CPVs. P-values were either from Kendall τ tests, in the case of continuous or large ordinate clinical variates, or from ANOVA, in the case of categorical variates. Only three associations were significant at a 5% FWER level by Holm's correction; these are highlighted. For pathological grade and cancer cell fraction variables, the direction of association is indicated by (+) or (−) annotations.

Variable	Axis 1	Axis 2
Age at diagnosis	0.925	0.666
Ethnicity	0.771	0.113
Gender	0.158	0.010
Histological subtype	0.697	0.157
Invasion		
Perineural	0.095	0.225
Vascular	0.650	0.071
Pack years smoked	0.356	0.275
Pathological grade	2.39×10^{-3} (+)	1.30×10^{-4} (+)
Cancer cell fraction	2.13×10^{-4} (+)	4.11×10^{-4} (−)
Recurrence site		
Bone	0.789	0.413
Brain	0.430	0.062
Liver	0.160	0.105
Lung	0.390	0.713
Lymph nodes	0.933	0.870
Mesentery	0.933	0.121
Omentum	0.139	0.082
Other	0.193	0.161
Pancreatic bed	0.887	0.530
Pancreas remnant	0.534	0.184
Peritoneum	0.916	0.015
Staging: M	0.441	0.425
Staging: N	0.252	0.263
Staging: T	0.264	0.427
Staging: Overall stage	0.061	0.236
Tumour location	0.177	0.139
Tumour longest axis length	0.844	0.171

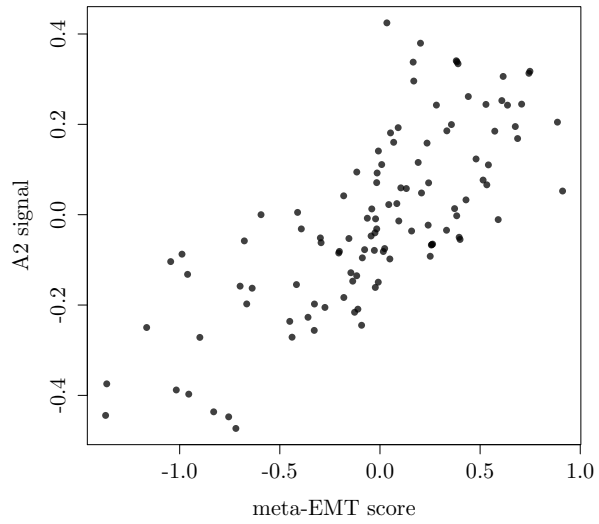


Figure 2.11: Axis A2 signal is closely associated with a signature of the EMT. A2 signal and meta-EMT [38] scores were as evaluated on the APGI training set; Kendall's $\tau = 0.568$, $n = 110$, linear model $R^2 = 0.557$.

tion of normal tissues and disruption of homeostasis, that cancer progressively overwhelms the body. The prognostic axis A1 discovered here was strongly correlated with the meta-PCNA signature of cell proliferation [106] (Figure 2.10), and appears to encode overall proliferative activity in patient tumours. The association between axis A1 activity and outcome was not unique to pancreas cancer: A1 was prognostic in a number of solid tumours, suggesting that proliferative activity is a prognostic marker of wider applicability than originally reported. Interestingly, there is evidence that the effect of proliferative level on outcome can be conditional on other biology: in the TCGA clear cell kidney cohort, high A1 activity was only associated with poorer outcome when axis A2 activity was also high.

Signals of the A2 axis were well-correlated with the meta-EMT signature [38] (Figure 2.11), suggesting that A2 levels reflected the activity of EMT processes within tumours. The EMT is a major

enabling step in metastasis, the process by which most cancers are ultimately lethal [97]. The EMT has a particular importance to pancreas cancer, as it is believed that occult metastases, present at the time of primary tumour resection, are the major cause of recurrence in resected patients (see Chapter 3 for detailed discussion of this point). A2 signal, and by proxy EMT activity, may be acting as a marker of tumour metastatic ability, and indirectly reflect the likelihood that a patient will have metastatic disease at the time of resection. This possibility is supported by the positive correlation between axis A2 signal, and levels of mRNA for S100A2 and S100A4, which are suspected biomarkers of occult metastasis in PDAC. In the presence of such metastases, the effectiveness of primary resection is greatly reduced, and earlier death of the patient is to be expected. The observed worsening prognosis with increasing A2 signal in postoperative patients is consistent with this proposed mechanism, and suggests that A2 loadings could be adapted to identify markers of early metastasis to aid clinical decision making.

Proliferation and the EMT are two of the ten major hallmarks of cancer [40], and so it is unsurprising that they are so closely associated with patient survival. What *is* surprising is that the majority of hallmarks do *not* appear to be strongly associated with outcome in resected PDAC. In particular, the absence of stromal or inflammatory signatures is unexpected given that PDAC cells are almost always surrounded by extensive stroma, which is believed to be a clinically significant component of the disease [69].

Transcription patterns linked to the tumour stroma may form a prognostic module that was missed by this work. Desmoplastic stroma is a pervasive and significant component of PDAC tumours [47], but its relevance to outcome is unclear: high tumour stromal content has been reported to be both harmful [69, 79], and protective [81, 90]; and the association between stroma activity and outcome is similarly ambiguous [11, 90]. This divergence in experimental findings suggests that the effect of the stroma on outcome is modulated by uncontrolled confounding factors. In such a sit-

uation, the approach taken in this work can not reliably identify stroma-associated transcriptional modules, even if these modules are genuinely linked to outcome. Some evidence that this may have occurred is given by the inverse association between tumour stroma content, as measured by $1 - \text{qPurescore}$, and axis A1 and A2 signals (Table 2.3). This work's potential poor sensitivity in the presence of confounding factors is not restricted to the discovery of stromal effects, but is a general consequence of the marginal variable screening approach that was used.

The signature discovery approach undertaken in this work was tuned to detect all major transcriptional modules affecting outcome in pancreas cancer, but may have missed less significant modules that have a minor influence on survival, only affect a relatively small subset of patients, or are masked by interaction effects. The nature of the modules detected by the selection-factorisation approach used here is strongly dependent on the performance of the initial hard-thresholding prognostic gene selection step. This work used a simple marginal screening approach that enjoys performance guarantees for near-orthogonal designs [29], but may be unreliable for the highly correlated measurements seen in gene expression data. In particular, genes with high conditional, but low marginal, associations with survival; or effects on outcome that are weak, or restricted to a small subgroup of patients; may have been missed by the initial screen. Any minor modules encoded by the expression of these missed genes would not have been identified by this work. A related limitation of the work is its use of primarily grade 2 or 3 resected patient samples; it is possible that other biological processes are involved in the survival of patients with higher-grade or unresectable tumours, but the data avilable were insufficient to properly test this hypothesis. Despite these potential issues limiting sensitivity, the simple marginal screening procedure used here nonetheless succeeded in defining the two major signatures that reflect outcome in resected pancreas cancer. Future analyses on larger cohorts may be able to identify additional minor prognostic modules, such as potential stroma-associated mod-

ules, by adjusting for the signals of axes A1 and A2 identified in this work.

The PARSE prognostic score, and its axis A1 and A2 components, were prognostic in a range of validation cohorts, both of PDAC, and in other solid tumours. This latter result was surprising, and suggests the importance of proliferation and the EMT as determinants of differential prognosis in a range of malignancies. The precise nature of association was dependent on cohort: in PDAC and TCGA lung cancer, A1 and A2 signals contributed approximately additively to hazard, whereas in the TCGA kidney and glioma cancer cohorts, evidence of interaction between the axes was observed (Figure 2.9). The positive validation of PARSE in a wide range of solid tumours indicates commonalities in molecular survival mechanisms between disparate cancers, and also suggests a more general application of the signature identification procedure used in this work.

Although the PARSE did not validate in dataset GSE21501, that dataset possesses unusual qualities that call the significance of the negative result into question. The well-established prognostic variables of T stage and node status are only marginally prognostic in GSE21501, and the six gene prognostic signature developed in the publication linked to GSE21501 [94] is not significantly prognostic in either the APGI or TCGA cohorts (data not shown). It is possible that the patients of the GSE21501 cohort are fundamentally different from the majority of PDAC cohorts in some way, and that investigation of the reason for this discrepancy would improve the generalisability of the PARSE score.

The methods used in this chapter are not restricted to the identification of outcome-associated metagenes. By modifying the initial gene selection step, metagenes correlated with any endpoint (for example, disease subtype, or drug response) can be identified, if they are present. Unsupervised metagene identification can also be performed, by performing unsupervised gene selection. By virtue of the SNMF/L decomposition used, the metagenes identified will be sparse, non-negative, component-based representations of the under-

lying transcriptional patterns, greatly facilitating interpretation in the often opaque world of transcriptional signatures. This use of sparse non-negative decompositions of transcription patterns both reflects a physical constraint (mRNA concentrations cannot be negative), and is a tool to break a complex response into discrete, easily-understood parts. This choice of sparse representations is further supported by theoretical indications that transcriptional programs are constrained to be sparse [61], and empirically is justified by Hastie's 'bet on sparsity' principle [44]: we will never be able to model dense systems, so we may as well assume all are sparse, and model them appropriately – the alternative is to simply regard all modelling as futile, and then start searching for a new occupation.

Although transcriptional activation patterns are physically constrained to be positive, and there are good reasons to suppose that they are sparse, there is no requirement for them to be *discrete*. Especially when considering the average transcription level across a heterogeneous tissue, it is not unreasonable to expect the activities of metagenes to lie on a continuum, from no activation to maximal activation. Metagenes that exhibit binary behaviour (that is, the metagene is either fully on or fully off, with no samples lying in between) are also possible, but, in a large population of diverse cells, are likely to be the exception rather than the rule. In this context, the methods developed in this chapter have the advantage of being able to capture both discrete and continuous patterns of metagene activity. This is in contrast to commonly-employed clustering approaches, which force examples into discrete clusters, regardless of whether this treatment is appropriate or not. In analyses of transcriptional patterns that seek to identify disease subtypes, such clustering approaches are very common, yet this work indicates that, at least for PDAC, they are also inappropriate.

The activities of axes A1 and A2 formed a smooth continuum in a number of cohorts, with no indication of clustering into discrete subgroups (Figure 2.9), strongly indicating that, in these cancers, A1 and A2 activity do *not* define clear disease subtypes. This finding was

only possible due to the general nature of the decomposition used, which does not force samples into clusters; the previously-reported Collisson subtypes of PDAC [25] were discovered using a variant of NMF that is tuned to stratify samples into stable subgroups, regardless of whether such a grouping is particularly sensible or not. Such a clustering approach in the Collisson was somewhat justified by its use of microdissected cells, which are more likely to lie in discrete regions of transcriptional space than the bulk tissue used in this work. However, the use of a clustering variant of NMF in the Collisson work presumed the existence, and forced the discovery, of sample clusters, whose existence is not supported by the continuum of A1 and A2 activities seen here. The results of Collisson *et al* and this work are not necessarily incompatible, given the substantial differences between the studies in sample type and endpoint, but in light of the results of this work, a re-examination of the Collisson data using a non-clustering NMF variant would be informative. The issue of artificial clustering in NMF algorithms is a subtle one: for example, had this work not used the SNMF/L decomposition, but instead the closely-related sparse non-negative matrix factorisation, wide variant (SNMF/W), metagene activities, and consequently samples, would have been artifically clustered into a small number of subgroups, and the metagenes themselves would have been far less interpretable.

The work in this chapter ultimately set out to answer a basic biological question – "why do some patients with PDAC live longer than others?" – but its results suggest fruitful areas of research for clinical applications. Most immediately, if a method for the preoperative measurement of tumour A1 and A2 activity could be developed, it would allow more accurate stratification of patients into survival bands, and better disease management overall. Although it is impractical to preoperatively measure levels of all 361 transcripts comprising the PARSE score, in principle the levels of a very small number of genes may accurately approximate the full set, permitting the preoperative estimation of the PARSE. This idea was the one developed in Chapter 3, using the S100A2 and S100A4 proteins as

biomarkers. S100A2 and S100A4 are likely proxies of axis A2 activity: they are significantly correlated with A2 signal, and are thought to act as markers of metastatic ability, which is intimately linked with axis A2 and the EMT. Should a similar marker be identified for axis A1, even more accurate stratification of patients can be expected. Close examination of the A1 and A2 components, or a principled search for high-performance biomarkers following the work of Chapter 4, may even suggest superior biomarkers to S100A2 and S100A4, ultimately producing a preoperative prognostic tool that is more accurate than that developed in Chapter 3.

The idea that the differential survival of patients following PDAC resection reflects differences in the levels of two transcriptional axes suggests a bold approach: can a poor prognosis tumour be transformed into a good prognosis one, by modulation of the prognostic axes? The axes correlated strongly to signatures of proliferation and the EMT, suggesting that interventions to modulate these processes would be the most directly effective methods to improve patient outcome following resection. Of course, this work cannot ultimately establish whether the levels of the A1 and A2 axes, or for that matter proliferation and the EMT, have a causative role in determining patient survival, or are merely markers of more fundamental survival processes. However, given the importance of proliferation and the EMT to cancer biology in general, it seems likely that these processes are the ones truly influencing patient outcome, and suggests that interventions to affect these processes will be a fruitful area of future translational research into PDAC.

In this work, I set out to determine whether specific molecular signatures control the survival of patients with resectable PDAC, and to link these survival signatures to fundamental biological processes. I found that prognostic gene expression signals could be factorised into two orthogonal components, and linked these components to the fundamental cancer processes of proliferation and EMT. These two processes were the dominant determinants of survival in resected PDAC, and a number of other solid tumours. This basic biology re-

sult immediately suggests directions for future translational research, to create more accurate preoperative staging systems, and to develop new therapeutic strategies that directly target the two cancer processes that reflect survival in resected PDAC.

2.4 Attribution

Data for the APGI discovery cohort were generated as part of the APGI project, under the umbrella of the ICGC. The generation of these data was a huge team effort, of which I only played a small part. However, both conception of the project, and all steps subsequent to raw data generation, from low level processing of Illumina data (IDAT) files through to analysis planning, signature development, testing, and interpretation, were performed solely by me.

2.5 Methods

Cohort recruitment and ethics

All samples were prospectively acquired as part of the APGI project, and detailed inclusion criteria and ethics approvals are given in the associated publication [14]. Briefly, samples were of primary, untreated, operable PDAC, collected during resection. For all cases, the diagnosis of PDAC was made by at least two pathologists with expertise in pancreas diseases.

Sample collection, preparation, and gene expression microarrays

Protocols for collection and processing of these samples have been published [14]. In summary, specimens were snap frozen in liquid nitrogen immediately following resection, and RNA extracted using the Qiagen AllPrep DNA/RNA/Protein Mini kit. For each sample, 150 ng of total RNA was amplified using the Life Technologies TotalPrep RNA Amplification Kit, and 750 ng of the resultant amplified

cRNA was hybridised on to Illumina Human HT-12 V4 arrays. Arrays were scanned on an Illumina Bead Array Reader, to yield IDAT scan files. All kit and microarray procedures were performed as per the manufacturer's instructions.

Data preprocessing

Microarray quality control and normalisation IDAT files were read into Bioconductor `lumi` structures using the `lumidat` package. Seven arrays were excluded on the basis of poor signal, due to fewer than 30% of probes on these arrays having detection P-values of less than 0.01. The remaining 234 microarrays represented a range of tumour types, and were normalised as one batch using the `lumi` package. Normalisation proceeded serially as: RMA-like background subtraction (`lumiB` method `"bgAdjust.affy"`), variance stabilising transform (VST) (`lumiT` method `"vst"`), and quantile normalisation (`lumiN` method `"quantile"`).

Unsupervised probe selection Probes were excluded if they met any of the following criteria: fewer than 10% of samples with expression P-values of less than 0.01, a probe quality (from the `illuminaHumanv4PROBEQUALITY` field in Bioconductor package `illuminaHumanv4.db`) not equal to 'perfect' or 'good', missing gene annotation, or a standard deviation of normalised expression values across all samples of less than 0.03. The choice of this latter threshold is expected to yield approximately a 5% false probe rejection rate, based on an analysis of the variation between technical replicate samples. In cases where multiple post-filter microarray probes mapped to the same gene, only the probe with the highest standard deviation, as evaluated across all samples that passed quality checks, was retained. The effect of these combined filtering steps was to reduce the number of features under consideration from 47,273 probes to 13,000, one per gene.

Sample selection From the full set of 234 tumour samples that passed quality checks, eight were from four samples that had each

been arrayed twice, and two were from patients with multiple conflicting CPV data. The two with conflicting CPV data were excluded from further study, and the eight replicated samples were averaged, after multidimensional scaling (MDS) indicated that each replicate pair had very similar expression.

The 228 APGI patients for which GEX and clinical data were available were subset further to yield a homogeneous PDAC cohort, suitable for the discovery of the survival-associated processes specific to PDAC. 141 of 228 patients had pathologically confirmed PDAC; of these, five were judged to have suffered a perioperative death, and were not considered further. 110 of the 136 remaining patients were treated in hospitals in Australia, 23 in the USA, two in Italy, and one in Puerto Rico. To eliminate the potential for country-specific gene expression patterns to interact with possible differential survival between countries, only the Australian subset of the cohort was retained, resulting in 110 patients in the final APGI discovery cohort.

Summary The above preprocessing steps yielded matched CPV and resected tumour GEX data for 13,000 genes across 110 patients.

Outcome-associated gene selection

Genes that were associated with DSS were identified by SIS-FAST [34], with a CPSS wrapper to reduce the false positive rate [88]. FAST statistics for time from diagnosis to DSD were calculated using R package `ahaz` on standardised log-scale expression values; genes which had an absolute statistic value exceeding 7 were selected by the inner SIS-FAST procedure. The outer CPSS wrapper selected genes which were returned by at least 80% of 100 complementary paired SIS-FAST runs. Gene selection FDR was estimated by permutation: 50 repeats of the full gene selection procedure were performed on data in which patients had been randomly shuffled, and the FDR was estimated as the median number of genes selected in permuted runs, divided by the number of genes selected by the unpermuted procedure.

Rank estimation and metagene factorisation

The gene × patient expression matrix of outcome-associated genes was decomposed into metagenes by the SNMF/L procedure [57], as implemented in R package `NMF`. SNMF/L is a variant of NMF, a class of procedures that decomposes a non-negative matrix A into a product of non-negative matrices W and H, $A \approx WH$. W and H typically have rank much less than A, the effect of NMF then being to effectively reduce a large gene × sample matrix A into smaller matrices, the gene × metagene basis matrix W, and metagene × sample coefficient matrix H. SNMF/L was chosen from the many NMF variants available for its design that favours solutions with sparse W: SNMF/L factorisations tend to associate each gene with a small number of metagenes, a situation that matches our biological expectation that, for most genes, expression of that gene is only associated with a small number of biological processes.

As NMF is a linear factorisation, the VST-transformed expression matrix A was approximately linearised by elementwise exponentiation, $a_{i,j} \leftarrow 2^{a_{i,j}}$. To reduce the influence of large variations in baseline expression on the factorisation, each row (gene) of A was then independently linearly scaled to lie between zero and one, $a_{i,j} \leftarrow (a_{i,j} - \min(a_{i,*})) \div (\max(a_{i,*}) - \min(a_{i,*}))$, where $a_{i,*}$ denotes row i of A.

Factorisation rank was estimated by comparison with a permuted null [31]: for test ranks ranging from 2 to 9, 5 SNMF/L decompositions were performed, each on a version of the transformed expression matrix in which rows (genes) had been independently permuted within each column (sample). Approximation error for each decomposition was calculated as $\|A - WH\|_F$, and the reduction in approximation error with increasing rank was compared between factorisations of the original data, and those of the 5 permuted data matrices. The highest rank for which the improvement in error achieved by adding that rank to the factorisation on the original data, exceeded the improvement seen by adding that rank on the permuted data, taking into account permutation noise, was selected

as the final factorisation rank. Specifically, let the improvement in approximation error that results in choosing a rank i decomposition over a rank $i-1$ decomposition, on the unpermuted data, be $\Delta_i = \|A - W_{i-1}H_{i-1}\|_F - \|A - W_iH_i\|_F$. Equivalently, define Δ_i^{*j} to be the improvement observed when rank i is added to the factorisation of A^{*j}, the j^{th} permutation of the data matrix: $\Delta_i^{*j} = \|A^{*j} - W_{i-1}^{*j}H_{i-1}^{*j}\|_F - \|A^{*j} - W_i^{*j}H_i^{*j}\|_F$. Denote the mean and standard deviation of Δ_i^* across all 5 permutations of the data matrix, for each i, as $\overline{\Delta_i^*}$ and $\text{SD}(\Delta_i^*)$, respectively. Then, the final selected rank k was selected as $k = \max(\{i : \Delta_i > \overline{\Delta_i^*} + 2\text{SD}(\Delta_i^*)\})$.

Following rank estimation, a final factorisation of the data was performed using only the identified rank, and a larger number of random algorithm restarts, as described below. Subsequent work used this final factorisation.

The SNMF/L algorithm requires parameters α and η to control regularisation; for all factorisations $\alpha = 0.01$, and $\eta = \max(A)$.[2] The default convergence criteria of the `NMF` package were used.

SNMF/L may not necessarily find a global optimum factorisation; to address this, multiple random initialisations of matrix W were made from $\text{Uniform}(0, \max(A))$, the SNMF/L procedure was run to convergence, and the result with lowest approximation error was retained. 50 random restarts were used during rank estimation runs, and 500 for the final factorisation; examination of approximation error distributions for these repeated runs indicated that these values were conservative, and factorisations were robust to the choice of random start.

Estimating metagene coefficients on new cohort data

To apply the signatures developed in this work to GEX data other than those from the APGI training set, the following procedure was used. GEX measurements from the new cohort were subset to the 361 outcome-associated genes identified by CPSS-SIS-FAST (these

[2]Note that this parameter α is denoted β in the R `NMF` package; I use the symbol α here for consistency with the original description of SNMF/L [57].

genes are listed in app:sigs-w-matrix on page 123), and transformed to a linear scale if necessary. Linear measurements were then scaled within genes to between zero and one, as was performed for metagene factorisation. Genes for which no expression data were available (the genes being either filtered out in preprocessing or not measured at all) were assigned scaled expression values of zero. These manipulations yielded a gene $\times$ sample matrix A' with rows matching the gene $\times$ metagene basis matrix W from SNMF/L. The metagene $\times$ sample coefficient matrix H' for the new cohort was then estimated by NNLS implemented in R package `nnls`, solving for each column of $a'_{*,i}$ of A' the optimisation problem $h'_{*,i} = \text{argmin}_x \|Wx - a'_{*,i}\|_2$, where $h'_{*,i}$ denotes column i of H'. Values of the W matrix used are available as app:sigs-w-matrix on page 123.

For consistency, the above procedure was used to estimate metagene coefficients H for the discovery APGI cohort, as well as all validation cohorts.

Calculation of the PARSE score on new cohort data

Given metagene coefficients estimated as above, axis activity scores were calculated as Axis A1 activity = MG1 coefficient$-$MG5 coefficient; Axis A2 activity = MG6 coefficient$-$MG2 coefficient. PARSE scores were then made by combining axis activity estimates, as PARSE score = $1.354 \times$ A1 activity $+ 1.548 \times$ A2 activity.

Although not used in this work, a simplified procedure for the approximate calculation of PARSE scores was also developed (app:sigs-parse-approx on page 143). By only considering genes with an absolute app:sigs-parse-approx coefficent exceeding a given threshold, app:sigs-parse-approx can also be used to develop PARSE score approximations that rely on fewer than the full 361 outcome-associated genes. This avenue was not explored in this chapter, which was more concerned with the identification of outcome-associated biological processes, than the production of a minimal-cardinality risk classifier. To address this latter problem, Chapter 4 describes a more direct route for the development of highly robust and very low car-

dinality prognostic classifiers.

External validation of outcome-associated metagenes

Gene expression data for accessions GSE21501 and GSE28735 were downloaded as processed series matrix data from the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO). Survival times, censoring indicators, clinical covariates (for GSE21501), and probe expression estimates were extracted from the series matrix files. Probes were annotated with gene symbols using the associated GPL annotation files, and probes with no gene annotation were discarded. If multiple probes mapped to the same gene symbol, only the probe with the highest standard deviation across all samples in a data set was retained. Finally, only probes with a standard deviation within the top 20^{th} percentile within a data set were kept for metagene scoring.

Gene expression and outcome data for all TCGA cancers were downloaded from the public TCGA open-access repository at `https://tcga-data.nci.nih.gov/tcgafiles/ftp_auth/distro_ftpusers/anonymous/tumor/`, on 18 November 2014. RNASeq Version 2 Level 3 expression estimates (on an approximately linear scale) from Illumina HiSeq machines only were used, without further processing. Expression estimates were scaled within genes to between 0 and 1 separately within each TCGA cancer type. For reasons of statistical power, only TCGA cancers for which at least 50 patients had both complete RNASeq expression data, and an event, were considered in validation. Cohort paad was included despite it not meeting this criterion, to allow validation against another PDAC cohort.

For each validation data set, metagene coefficients, axis activities, and PARSE scores, were calculated as described above. Prognostic performance of the PARSE score was tested within each validation data set using likelihood ratio tests comparing a Cox model using PARSE score as the sole linear covariate, with an intercept-only Cox model.

GSVA scoring

The expression of gene sets from the MSigDB [95] were estimated on the APGI cohort using a modification of the GSVA method [41]. GSVA with default settings was used to estimate expression scores for all MSigDB gene sets in the full 13,000 × 228 VST-scaled APGI GEX data matrix. MSigDB contains both undirected gene sets such as metabolic pathways, in which members of the set are not expected a-priori to move in concert, and directional signatures, with paired `*_UP` and `*_DN` components that would be expected to change in co-ordinated and opposite patterns. Conventional analyses based on MSigDB ignore this distinction, but for this work I combined paired directional signatures to yield an overall signed estimate of signature activity. For undirected signatures, GSVA activity estimates were simply calculated using parameter `abs.ranking=TRUE`. In the case of paired signatures, GSVA scores were estimated separately for the `*_UP` and `*_DN` sets using parameter `abs.ranking=FALSE`, and the signed combined activity `*_SIGNED` was calculated as the `*_DN` score subtracted from the `*_UP` score. This procedure resulted in summarised activity estimates for 8,138 gene sets, many of which were highly correlated.

Gene sets with highly correlated activity scores were collapsed into compound summary sets as follows. Pairwise Pearson correlation distances between all scores were calculated as $d_{i,j} = \frac{1}{2}(1-\text{cor}(s_i, s_j))$, and were used to cluster gene sets using R `hclust` and complete linkage. R `cutree` identified clusters of highly similar gene sets, using a distance threshold of 0.02; gene set activities within each cluster were merged by taking median values across all samples, to form a new merged gene set activity estimate. Following merging, 7,633 single and compound gene set activity estimates remained across 228 samples.

meta-PCNA and meta-ECM score calculation

Scores for the meta-PCNA signature were calculated from GEX data as described in the meta-PCNA publication [106]. To estimate meta-

ECM scores, log-scale GEX data were median centered, and then median values across samples were calculated for all genes in two ECM signature lists [38, Table S3], to yield EMT-overexpressed, and EMT-underexpressed, gene list median expression estimates per sample. The meta-ECM score was then calculated as the EMT-overexpressed median value, less the EMT-underexpressed median value.

Prognostic axis functional characterisation

Clinical variate comparisons Prognostic axis activities calculated on the APGI data were tested for association with a restricted set of the available APGI CPVs, as outlined in Table 2.4. Numeric variables were tested for association with each axis by Kendall's τ test; factor and Boolean variables using ANOVA with the CPV as the explanatory variable. 50 tests in total were performed (25 variables, 2 axes), and P-values were corrected together using the Holm-Bonferroni procedure [50]. Corrected P-values of less than 0.05 were considered significant.

MSigDB signature score comparisons Kendall correlation coefficients were calculated between axis activity estimates and GSVA scores for MSigDB gene sets, on the APGI expression dataset. A subset of the full MSigDB was used, as outlined in Table 2.5. Absolute correlations of greater than 0.5 were deemed substantive and reported for further characterisation.

Table 2.4: CPVs tested for association with prognostic axis signals.

Clinical variate	Type
Age at diagnosis	Ordinal
Ethnicity	Factor
Gender	Boolean
Histological subtype	Factor
Invasion:	
Perineural	Boolean
Vascular	Boolean
Pack years smoked	Ordinal
Pathological grade	Boolean
Recurrence found in:	
Bone	Boolean
Brain	Boolean
Liver	Boolean
Lung	Boolean
Lymph nodes	Boolean
Mesentery	Boolean
Omentum	Boolean
Other	Boolean
Pancreas remnant	Boolean
Pancreatic bed	Boolean
Peritoneum	Boolean
Staging: M	Boolean
Staging: N	Boolean
Staging: T	Factor
Staging: Overall stage	Factor
Tumour location	Boolean
Tumour longest axis length	Ordinal

Table 2.5: The subset of MSigDB signatures tested for association with axis activities. Within each MSigDB class, only those matching the indicated inclusion pattern were tested. * represents a wildcard; ∅ matches nothing.

MSigDB class	Signature name inclusion pattern
c1	∅
c2	KEGG_*, PID_*, REACTOME_*
c3	*
c4	GNF2_*, MORF_*
c5	*
c6	*
c7	*

Chapter Three

A Preoperative Molecular Prognostic for Pancreas Cancer

Thesis: A preoperative prognostic tool for pancreas cancer can be developed to discriminate between good and poor prognosis patients more reliably than current methods.

Summary For those patients fortunate enough to be diagnosed with a resectable tumour, surgical removal of the primary cancer is the best first-line therapy for pancreas cancer. However, the significant morbidity associated with pancreas cancer resection makes it crucially important to only operate on the patients who will derive a net benefit from the procedure. Identifying just those patients who will respond to resection remains a serious challenge in pancreas cancer treatment: current criteria to select patients for resection perform poorly, and consequently many patients undergo a complex procedure, with serious effects on future quality of life, for little benefit. Tumour biomarkers have the potential to dramatically refine current morphology-based staging criteria by supplying a direct readout of tumour biology, and recent technological developments have enabled the preoperative measurement of tissue biomarkers in pancreas

cancer. The ability to measure pancreas cancer tissue biomarker levels preoperatively, combined with the enhanced information on disease state available from tissue biomarkers, finally enables the development of preoperative staging systems that accurately identify pancreas cancer patients for resection. This chapter details the development and validation of the Pancreas Cancer Outcome Predictor (PCOP), a two-biomarker prognostic tool for resectable pancreas cancer, that is in principle preoperatively assessable, and can assist in making personalised treatment decisions.

3.1 Introduction

For patients with a resectable tumour and no known metastases, surgical removal of the primary tumour is the current recommended first-line therapy for pancreas cancer, and the only intervention offering the realistic possibility of a cure [27]. However, pancreas cancer resection is a major procedure, with the potential for serious complications, morbidity, and reduced quality of life following recovery [49]. Due to the substantial negative effects of surgery, the decision of whether or not to perform curative-intent resection should balance the risks of surgery against its expected benefits, for each individual case.

Unfortunately, current practice guidelines recommend that curative-intent surgery be offered to all metastasis-free patients with a resectable tumour, with no consideration of personal benefit [27]. This blanket approach to selecting patients for curative resection has proven to be highly inadequate. Even following pathologically complete tumour removal and adjuvant chemotherapy, more than 70% of current pancreas ductal carcinoma patients will relapse with, and ultimately succumb to, distant metastases [10]. These occult metastases must have been present prior to removal of the primary tumour, yet were undetectable during initial investigations, and their presence means that any curative-intent resection was futile. As a result, the majority of 'curative' resections that are undertaken based on current selec-

tion criteria are performed on patients with occult metastases, have no hope of actually effecting a cure, and would not have been undertaken at all if the presence of metastatic disease had been known prior to surgery. Better methods to select patients for resection are urgently needed.

A number of pancreas cancer grading schemes and prognostic tools have been described, but inconsistent performance, or a reliance on information that can only be known postoperatively, limits their use in preoperative decisions. Two such schemes are based on levels of the biomarker carbohydrate antigen 19-9 (CA-19-9), and the Memorial Sloan-Kettering Cancer Center (MSKCC) prognostic nomogram [18].

The level of serum CA-19-9 is a well-characterised biomarker of pancreas cancer, with high levels correlating with increased tumour burden, lower probability of resectability, increased post-resection recurrence, and worse prognosis [59, 9, 10, 68]. CA-19-9 levels are easily determined preoperatively, but the use of this marker is complicated by a lack of consensus on threshold concentrations, the elevation of CA-19-9 levels by a number of conditions other than pancreas cancer, and the complete absence of this marker in approximately 10% of the general population [9]. Additionally, although CA-19-9 levels are statistically associated with post-resection recurrence by distant metastasis, a very low positive predictive value (PPV) renders the biomarker unhelpful when deciding whether or not to resect [59].

The current standard prognostic tool for pancreas cancer is the MSKCC nomogram [18], which integrates a number of CPVs to arrive at estimates of survival post-resection. Unfortunately, its clinical utility is small: as it relies on information that is only available following resection, the MSKCC nomogram is only useful in a postoperative context, and cannot assist in preoperative decisions to resect. This severely limiting reliance on postoperative variables is made necessary by the fact that all strong classical prognostic factors in pancreas cancer (such as lymph node infiltration, resection margin status, or histological grade [15]) can only be reliably measured fol-

lowing resection. Any prognostic tool for pancreas cancer that relies heavily on classical CPVs will very likely share this same reliance on postoperative variables, and so an effective preoperatively assessable prognostic will need to avoid relying on classical CPVs, and leverage novel preoperative measures of prognosis.

Levels of tissue biomarkers directly reflect cellular state, and thus have the potential to predict cancer behaviour far more reliably than macroscopic CPVs. Given that most pancreas cancer patients who undergo curative resection quickly recur due to occult metastases, biomarkers of metastasis have the potential to identify those patients who are likely to already have occult metastatic disease at the time of surgery, and thus better inform the decision to resect. Two such biomarkers of metastasis are the cancer cell levels of the EMT-related S100A2, and S100A4 proteins, both of which are strongly predictive of outcome following resection, and appear to reflect the presence of a pro-metastatic invasive phenotype in the cancer [13, 100, 62]. Despite this promise, these tissue biomarkers have to date only been assessed in bulk tissue samples collected during surgery, and their utility, or even measurability, in a preoperative setting, is untested.

Recent technological developments have made possible the preoperative measurement of tissue biomarkers during endoscopic ultrasound (EUS), a routine diagnostic modality for pancreas cancer. Immunohistochemical (IHC) staining has been successfully performed on fine needle aspirate (FNA) biopsies of pancreas neoplasms collected during EUS [78, 87, 92], and in principle EUS-FNA-IHC could form the basis of a routine preoperative biomarker measurement methodology in pancreas cancer. This proposed biomarker measurement approach utilises only techniques that are commonly available in pancreas cancer treatment centres, and thus has the potential to be rapidly integrated into current diagnostic workflows, should biomarker measurements prove to be clinically valuable.

The nexus of known biomarkers of metastatic behaviour, new preoperatively applicable techniques to measure these biomarkers, and multiple large, clinically annotated cohorts of resected pancreas can-

cer, presents an opportunity to address the pressing need for better criteria to select patients for pancreas cancer resection. As part of the APGI, as well as other work, the group has collected tissue measurements of S100A2 and S100A4 biomarkers, and detailed patient follow-up, for a large number of cases of pancreas cancer from a range of independent cohorts. These cases were used to develop a pilot version of the the Pancreas Cancer Outcome Predictor (PCOP), a tool to predict outcome following resection, using tissue levels of S100A2 and S100A4 as major prognostic factors. This pilot version of the PCOP is based on biomarker measurements made on tissue collected during resection, and thus is not directly applicable preoperatively. However, it demonstrates the feasibility of a preoperative biomarker-based prognostic tool, and identifies statistical issues involved in the generation of such prognostics, in preparation for a formal prospective preoperative sample collection effort.

The majority of pancreas cancer resection procedures today are performed on patients who should never have been offered surgical resection at all. These patients have undetected metastases at the time of surgery, and will derive little benefit from a major operation, that has serious impacts on quality of life. Current tools for patient staging and estimation of prognosis are either ineffective at identifying patients at risk for occult metastases, or only applicable postoperatively, and so cannot be used to inform the decision of whether or not to resect. Tissue biomarkers of metastatic potential might identify, preoperatively, those patients who have a high likelihood of metastatic disease, greatly assisting disease management decisions. This metastasis prediction can be integrated with other clinical variables to yield personalised estimates of prognosis over time, that can be easily understood by both physicians and laymen. This chapter describes the use of preoperatively assessable variables, including biomarker measurements, to create the first version of the PCOP, a tool that produces estimates of prognosis. The PCOP provides a natural way to show the influence of risk factors on a patient's personalised prognostic path, and thus can assist in making treatment

decisions appropriate for each individual pancreas cancer patient.

3.2 Results

Data from the large, retrospectively-acquired New South Wales Pancreatic Cancer Network (NSWPCN) cohort were used to derive the PCOP, a tool to predict the survival of pancreas cancer patients following curative-intent resection. Discrimination and calibration of the PCOP were tested on three independent surgical cohorts. A simple web interface was constructed to illustrate how a prognostic tool such as the PCOP could be deployed in practice.

Prognostic variables and biomarkers

As the aim was to develop a prognostic predictor that could be applied preoperatively, only factors that could be practically measured prior to resection were considered for inclusion in the PCOP. The traditional CPVs that were judged to be preoperatively assessable were patient sex, patient age at diagnosis, tumour location (dichotomised as head of pancreas vs other location), and size of the tumour's longest pathological axis. In addition to these traditional factors, the dichotomised tissue levels of S100A2 and S100A4 proteins were included as candidate biomarkers in the construction of the PCOP. Preoperative blood levels of the biomarker CA-19-9 were available for a subset of the training cohort, but none of the validation set patients; for this reason, and the marker's generally poor performance in isolation [59], CA-19-9 levels were not considered for inclusion in the PCOP.

Preoperative measurements of tumour size (for example, by computed tomography (CT) X-ray or EUS) were not available in the training and validation sets, and were approximated by postoperative measurements for the development and testing of this nomogram. Similarly, preoperative biomarker measurements were approximated using IHC staining of tissue collected during resection, as only very limited preoperative EUS-FNA samples were available in the cohorts

used. The implications of these approximations for the prognostic tool developed here, as well as for future work, are considered in the discussion.

Cohorts and characteristics

General characteristics of the NSWPCN, Glasgow, APGI, and Dresden cohorts are summarised in Table 3.1. The NSWPCN training cohort contained a small subgroup of patients with abnormally long recorded survival times ($>$ 3,000 days, 7/256 patients), that were strongly suspected to represent data errors, either as a consequence of incorrect coding following loss to follow-up, or misdiagnosis. Given the age of the cohort, it was deemed impractical to revisit the original records to check these patients, and so all patients with recorded survival times exceeding 3,000 days were excluded from the NSWPCN training data. The NSWPCN cohort characteristics in Table 3.1 have been calculated on the 249 patients that passed the 3,000 day data quality cutoff.

The four cohorts had broadly similar marginal survival functions (Figure 3.1), although these were statistically distinct (logrank $P = 5.7 \times 10^{-6}$). There were significant differences between the cohorts in the distribution of prognostic CPVs: large variation was present in the fraction of patients with clear resection margins (range 27%−65%, Fisher exact test $P = 2.2 \times 10^{-15}$), tumours in the head of the pancreas ($81\% - 100\%$, $P = 8.6 \times 10^{-13}$), and lymph node involvement ($66\% - 83\%$, $P = 8.3 \times 10^{-5}$). The variability in margin involvement is plausibly due to differences in the definition of margin clearance over time and between geographical regions [22], but the other discrepancies likely indicate fundamental differences in cohort composition. These covariate differences were not sufficient to explain the observed differences in outcome: after correcting for all available covariates, cohort still had a significant effect on survival (likelihood ratio test $P = 3.8 \times 10^{-8}$), with patients from the NSWPCN training set displaying worse covariate-corrected prognosis than those from other cohorts (hazard ratios for NSWPCN patients over others all

> 1.98).

The differences in prognosis between cohorts may be linked to the greater age of the NSWPCN cohort, the majority of which contains patients diagnosed between 1998 and 2003, over the more modern validation cohorts. Improvements in therapy effectiveness over time, particularly with regards to chemotherapy, may explain the improved overall outcome of the validation patients over the NSWPCN cohort. Unfortunately, as reliable data on chemotherapy were not available in any cohorts, this possibility could not be tested, and could represent a major uncontrolled confounding factor in the data. After controlling for all measured variables, there was no significant difference in baseline survival function between cohorts (Grambsch-Thernau test [36] Holm-corrected $P > 0.23$, 24 tests), indicating that at least the general form of the hazard function was similar across all cohorts. However, despite this similar baseline function, the presence of a strong and significant cohort effect that is independent of all measured variables will limit the maximum possible validation performance of any prognostic predictor on these data.

Biomarker scores were significantly differently distributed between cohorts (S100A2 $15\% - 33\%$, $P = 1.5 \times 10^{-4}$, S100A4 $65\% - 88\%$, $P = 1.3 \times 10^{-4}$). This difference in biomarker scores is likely largely due to cohort-specific technical differences in tissue collection, processing, staining, and scoring, although cohort composition effects may also have contributed.

The large differences between training and validation cohorts provides a strong test of the ability of a prognostic tool to generalise to new cohorts, laboratory processes, and scoring pathologists. Residual unexplained effects of cohort on survival will limit the validation calibration performance attainable on these data, but clinically useful accurate discrimination of good- and poor-prognosis patients may still be achievable.

Table 3.1: Characteristics of the NSWPCN training cohort, and the APGI, Dresden, and Glasgow validation cohorts. Ordinal variables are shown as median, with quartiles in parentheses.

		Training	Validation		
Characteristic		NSWPCN	APGI	Dresden	Glasgow
Number of patients		249	75	150	189
Gender	Male	49.4%	54.7%	54.7%	52.9%
Tumour location	Head	80.7%	85.3%	92.7%	100%
Excision margin status	R0	58.2%	32.0%	65.3%	27.0%
Node involvement		65.8%	78.7%	68.7%	82.5%
S100A2 positive		16.1%	14.7%	25.3%	32.8%
S100A4 positive		75.5%	65.3%	88.0%	70.9%
Disease-specific death event		95.2%	68.0%	74.7%	85.2%
Size of longest axis	(mm)	30 (25 - 40)	35 (28 - 43)	35 (25 - 40)	30 (25 - 40)
Age at diagnosis	(years)	69 (62 - 75)	67 (61 - 74)	68 (59 - 73)	64.0 (57.8 - 69.4)
Length of follow-up	(days)	479 (270 - 851)	655 (362 - 743)	514 (311 - 915	501 (233 - 915)

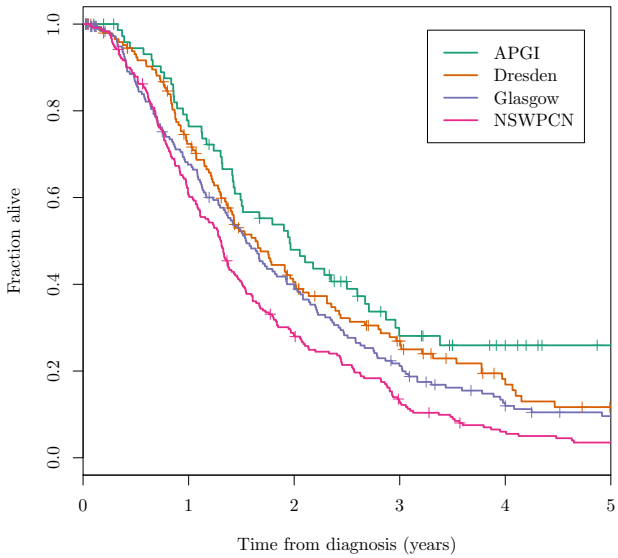


Figure 3.1: Kaplan-Meier marginal survival estimates for the cohorts used in this chapter. Censoring times are indicated by + symbols.

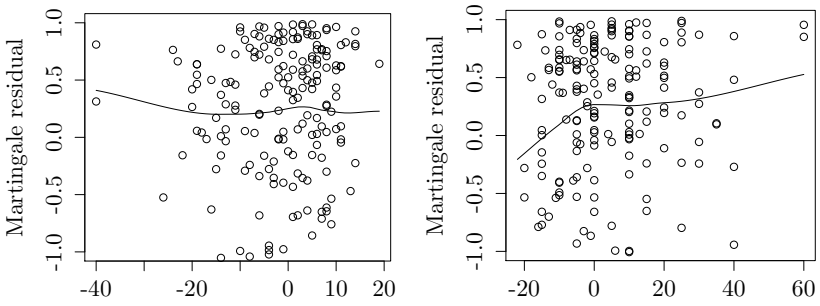
Prognostic model building and selection

Candidate prognostic models were constructed on the NSWPCN training data by iterative model fitting, evaluation, and refinement. To guard against overfitting caused by this iterative process, the NSWPCN cohort was randomly split once, into model building and testing sets. All model fitting and refinement described below was performed on the 200-patient model building set, to yield three final candidate prognostic predictors. The performance of each of these three predictors was then assessed on the 49-patient model test set, and the most parsimonious high-performing model was chosen as the PCOP prognostic predictor, for subsequent external validation.

Cohort shift The NSWPCN training cohort was collected over a long period, with patient diagnosis dates spanning the thirteen years from 1994 to 2006. Over such an extended interval, subtle changes in cohort composition or therapy may cause a shift in cohort characteristics, and reduce the prognostic performance of a model that was built on the historical data, when it is applied to contemporary cases. Cohort shift was investigated by examining the association between date of diagnosis, and all prognostic and outcome variables: in the absence of shift, no variables would be expected to change significantly over time. Date of diagnosis was not significantly associated with any other variable, or outcome (distance correlation [96] and Cox proportional hazard (CPH) regression, 7 tests, lowest $P = 0.35$); there was therefore no indication of cohort shift in the NSWPCN training data.

Model functional form and expanded terms The CPH framework was used to assess functional form for the two continuous covariates: age at diagnosis, and maximum pathological axis size. Local regression (LOESS) smooths of martingale residuals [99] indicated a largely linear relationship for age at diagnosis (Figure 3.2(a)), and a knee-shaped form for size (Figure 3.2(b)), with the knee at approximately 0 in median-centered units. In subsequent fits this potential

nonlinear size effect was modelled by adding a size_+ ramp term. The original set of five linear prognostic terms, plus the additional nonlinear size term, was denoted the expanded term set.



(a) Patient age (centered) (b) Tumour size (centered)

Figure 3.2: NSWPCN prognostic predictor functional forms. Smoothed Cox model martingale residual plots indicate hazard relationships that are approximately linear for centered age (panel a), and piecewise linear for centered tumour size (panel b). For clarity, plots have been restricted to the residual range $[-1, 1]$.

Proportional hazards assumption A Grambsch-Therneau test [36] on the CPH model fit using all expanded terms indicated that patient sex violated the proportional hazards (PHs) assumption ($P = 0.0104$, Figure 3.3) – in other words, the two sexes had significantly different baseline hazard shapes. To account for this effect, all subsequent models were stratified by patient sex, so that the survival of male and female patients was modelled by two different baseline hazard functions. A Grambsch-Thernau test on the stratified model indicated no further significant violations of the PH assumption (global $P = 0.4194$).

Outlier removal Strongly influential or outlying samples from the full marginal Cox fit were removed from the NSWPCN building set. I considered this unusual measure to be necessary given known and unresolvable quality issues in the NSWPCN cohort data. For all subsequent work, patients with full marginal Cox model absolute

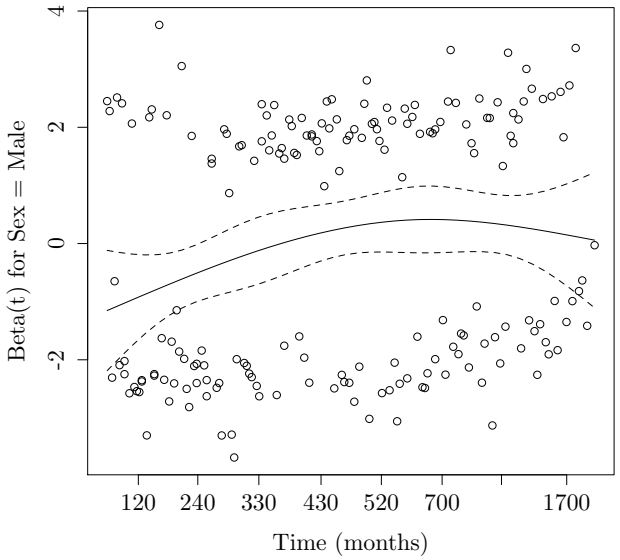


Figure 3.3: NSWPCN baseline hazard differs between patient sexes. A natural spline smooth of scaled Schoenfeld residuals for patient sex has a slope obviously differing from zero, suggesting that the baseline hazard forms differ between the two sexes, and that the combined data violates the PH assumption of Cox regression. Individual residuals are displayed as points, the natural spline smooth (df = 4) as a solid line, and approximate ±1 SE bounds as dashed lines.

deviance residuals exceeding 2.5, or any absolute DFBETAS score exceeding 0.3, were excluded from the original building set. This filter removed seven patients, reducing the size of the model building set to 193 patients.

Variable selection Stepwise variable elimination was used to select an Akaike information criterion (AIC)-optimal model starting from the full marginal CPH model, containing all expanded terms and a sex stratum. The identified optimal CPH model used four variables: tumour location (head vs body), tumour size (linear term only), S100A2 status, and S100A4 status, in addition to the sex stratum. The AIC-selected set of four prognostic terms, and a patient sex stratum, was denoted the reduced term set.

Model CP1 A final prognostic CPH regression model was fit to the outlier-removed NSWPCN model building data using only the reduced term set; this model was termed CP1. CP1 did not violate the PH assumption by the Grambsch-Therneau test (global $P = 0.794$). Predictions from model CP1 were broadly concordant with stratified Kaplan-Meier (KM) estimates across all covariate subgroups, indicating no serious lack of fit of the model (Figure 3.4).

Model GG1 Semiparametric Cox PH models such as CP1 provide a convenient framework for covariate testing and model diagnostics, but their unspecified baseline hazard term significantly complicates their use as prognostic predictors: patients are naturally only scored for relative hazard, and estimates of survival probabilities are unavailable. Although it is possible to approximate the baseline hazard in the Cox model, a more natural alternative is to use fully parametric models, in which the baseline hazard distribution is explicitly specified. The advantages of parametric models in terms of robustness and interpretability are offset by their more stringent assumptions: if the chosen baseline distribution is unsuited to the particular data to be fit, predictions from parametric models can be very poor. Given the potential benefits of parametric models for survival prediction, a parametric alternative to model CP1 was developed, and its fit assessed. This parametric model was termed GG1.

Model GG1, employing a generalised gamma (GG) survival distribution [26], was fit to the outlier-removed NSWPCN model building data by maximum likelihood. Guided by the model functional form and baseline hazard stratification indicated by the Cox model diagnostics, the GG distribution location parameter β was made linearly dependent on all terms in the reduced set, but the shape parameters σ and λ were modelled as dependent on patient sex only. Graphical comparisons between GG1 predictions and KM estimates of survival indicated that GG1 predicted outcome to within error across major patient subgroups (Figure 3.4).

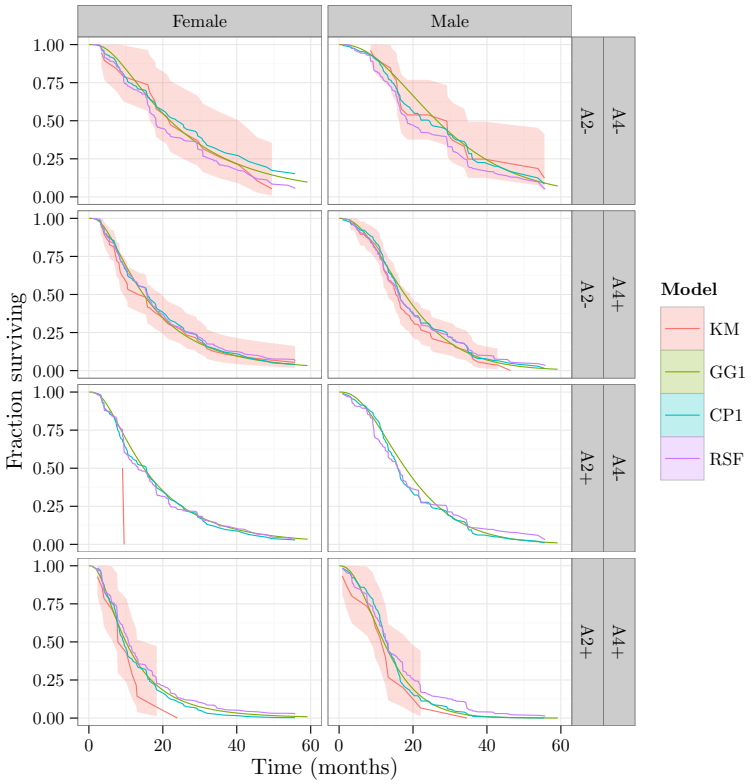


Figure 3.4: Model survival predictions agree with stratified KM estimates. KM estimates of survival probability for each combination of patient sex and biomarker status are shown as solid red lines, with 95% confidence intervals indicated by red ribbons. Estimates of survival probability generated by models CP1 (blue), GG1 (green), and RSF (purple), broadly followed the form of the KM estimator, and lay within its bounds at all times, although all models consistently overestimated survival of the double positive subgroup. Both model fitting and prediction used the NSWPCN model building set, and so these plots illustrate model goodness-of-fit, but cannot indicate possible overfitting. KM traces for the S100A2 positive, S100A4 negative group were omitted, as there were insufficient patients in this group for reliable KM estimates to be available. For all plots, tumour location, and size, were set to cohort median values.

Model RSF Regression models like CP1 and GG1 are familiar and readily interpretable, but are heavily dependent on the analyst identifying appropriate variables and functional forms. Ensemble tree models such as random forests [16] naturally and automatically model nonlinearity and arbitrary level interactions, and are tolerant of large numbers of irrelevant or colinear variables, albeit at the cost of very poor interpretability, and large data and computational requirements. Random forests have been adapted to model censored data [55], and can provide an alternative prognostic predictor that is distinct in behaviour from CP1 and GG1, and may be able to exploit data structure not leveraged by these more classical models.

To investigate whether tree ensemble models could provide improved performance over classical approaches, a random survival forest model, termed RSF, was fit to the outlier-removed NSWPCN model building data. In contrast to CP1 and GG1, which used the reduced set of terms as covariates, RSF was supplied all preoperatively-assessable variables as candidate predictors.

Model selection Predictive performance of the three prognostic models (CP1, GG1, and RSF) was compared on the holdout NSWPCN model test set, to select a single high-performing parsimonious model for external validation. Model discriminatory ability was assessed over time using the area under the curve (AUC) of the incident/dynamic time-dependent receiver operating characteristic (TD-ROC) [45], and overall prognostic accuracy (combining both discrimination and calibration) over time by the Brier score [35]. The integrated Brier score [35] was also used to provide an aggregate measure of overall model accuracy. Performance in the interval from seven to 34 months post-diagnosis was of particular interest, as the majority of patients in the NSWPCN training set died during this period (Figure 3.1).

All models had statistically indistinguishable discriminatory power over the $7 - 34$ month period, as assessed by pointwise 95% BCa confidence intervals [28] of the TD-ROC AUC (Figure 3.5). There

Table 3.2: Competing models do not have significantly different integrated Brier score (IBS) performance. The IBS is a combined measure of model predictive ability over a follow-up time interval, which captures both discrimination and calibration; lower numbers are better. Differences in the $7-34$ month IBS of competing models were calculated for each of 1,000 bootstrap samples of the NSWPCN hold-out test set, and 95% BCa confidence intervals [28] calculated. All candidate prognostic models had significantly better IBS than the marginal KM0 model, but there was no significant difference between candidate models. The $7-34$ month region in which most patients die is indicated by vertical black lines.

	Bootstrap	
Comparison	Mean	95% CI
KM0 − GG1	21.1	$[2.5, 39.8]$
KM0 − CPH	20.2	$[4.5, 38.9]$
KM0 − RSF	14.5	$[5.7, 24.6]$
RSF − GG1	6.6	$[-5.6, 17.7]$
RSF − CPH	5.7	$[-2.9, 15.9]$
CPH − GG1	0.9	$[-4.1, 4.3]$

was also no significant difference between candidate models in Brier score, integrated over $7-34$ months, although all models performed significantly better than a marginal Kaplan-Meier prognostic, KM0 (Table 3.2). Despite these non-significant differences, models GG1 and CP1 had consistently superior Brier score to RSF over the period of interest (Figure 3.6). As there was no significant difference in performance between the prognostic models, the simplest model, GG1, was selected to form the PCOP.

Final PCOP fit A final fit of GG1 to the full NSWPCN training data (both model building and validation patients) was made, and is summarised in Table 3.3. This fit defined the PCOP, which predicts

post-resection outcome using a generalised gamma model [26], as

$$\begin{aligned} T \sim GG(\beta = {} & 6.7446 + 0.3732[\text{Sex} = \text{Male}] - 0.2150[\text{Location} = \text{Body}] \\ & - 0.0887\ \text{Size} - 0.3729[\text{S100A2} = \text{Positive}] \\ & - 0.3843[\text{S100A4} = \text{Positive}], \\ \sigma = {} & 0.7503 - 0.2452[\text{Sex} = \text{Male}], \\ \lambda = {} & 0.0288 - 0.7630[\text{Sex} = \text{Male}]) \end{aligned}$$

where T is an individual's failure time, GG is the generalised gamma distribution, Size is in centimetres, and [ ] is the Iverson bracket.

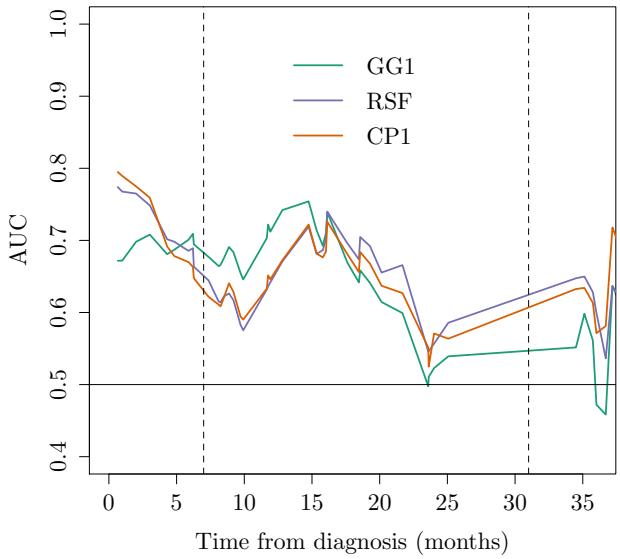


Figure 3.5: Incident / dynamic TD-ROC AUC paths for candidate models on the holdout NSWPCN model test set. Slight differences in performance were evident, with model GG1 providing superior discrimination up to approximately 15 months post-diagnosis, but models RSF and CP1 performing better from approximately 20 months post-diagnosis. These differences were not significant, as assessed by pointwise 95% bootstrap confidence intervals (confidence bands not shown). The $7 - 34$ month interval in which most patients died is indicated by dashed vertical lines.

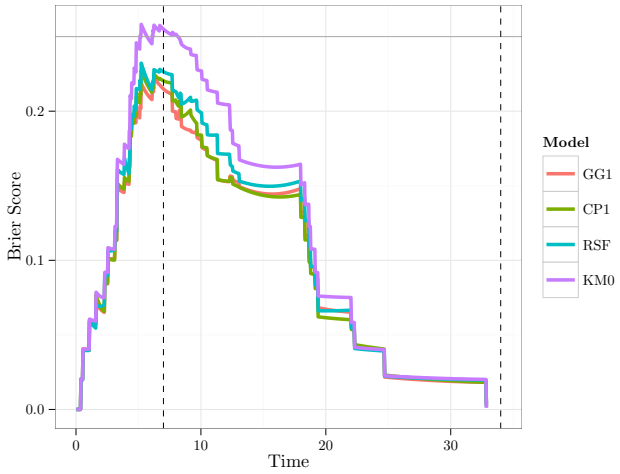


Figure 3.6: Brier score paths for candidate models on the hold-out NSWPCN model test set. All models outperformed the no-information KM0 trace from approximately four months to 21 months post-diagnosis, and no strong differences were apparent between candidate models. The $7 - 34$ month interval in which the majority of patients died is indicated by dashed lines, and the theoretical worst-case Brier score by a horizontal line at 0.25.

External validation

Discrimination, calibration, and overall fit of the PCOP was tested on three independent validation cohorts, following published guidelines [83].

Overall assessment of PCOP fit

Distribution of the PCOP PI Approximate prognostic indexs (PIs) for the PCOP showed broadly similar distributions across the training NSWPCN cohort, and the three validation cohorts (Figure 3.7). Vertical lines denote the empirical 20th, 50th, and 80th percentiles, which were also used to define risk groups to visually evaluate PCOP fit.

Visual assessment of PCOP fit Patients within each cohort were divided into broad risk groups based on their PCOP approximate PI,

Table 3.3: Coefficients of a final GG1 fit to the NSWPCN training data, which defines the PCOP. Coefficient estimates are for a generalised gamma survival model [26].

Term		Estimate	SE
β			
(Intercept)		6.7446	0.1489
Sex	= Male	0.3732	0.1508
Tumour location	= Body	−0.2150	0.1223
Size of longest axis	(cm)	−0.0887	0.0302
S100A2	= Positive	−0.3729	0.1235
S100A4	= Positive	−0.3843	0.1045
σ			
(Intercept)		0.7503	0.0493
Sex	= Male	−0.2452	0.1066
λ			
(Intercept)		0.0288	0.2719
Sex	= Male	0.7630	0.3533

and observed and predicted outcomes within each group were visually compared to evaluate model fit. Three risk groups were defined: a high-risk group of patients with PI less than the empirical cohort 20th percentile; a low-risk group with PI greater than the 80th percentile; and an intermediate risk group with all remaining patients. Overall model fit was good in the Glasgow validation cohort (Figure 3.8(b)), but poor in the APGI and Dresden cohorts (Figure 3.8(c), Figure 3.8(d)).

PCOP Brier score The Brier score summarises overall model prediction error over time, without requiring patients to be divided into arbitrary risk groups based on approximate PI. To further investigate the poor fit observed in some cohorts, Brier score paths were calculated for the PCOP in all three validation cohorts. The performance of the MSKCC nomogram, applied to preoperative variables, was also assessed at its three timepoints to provide a comparison between the

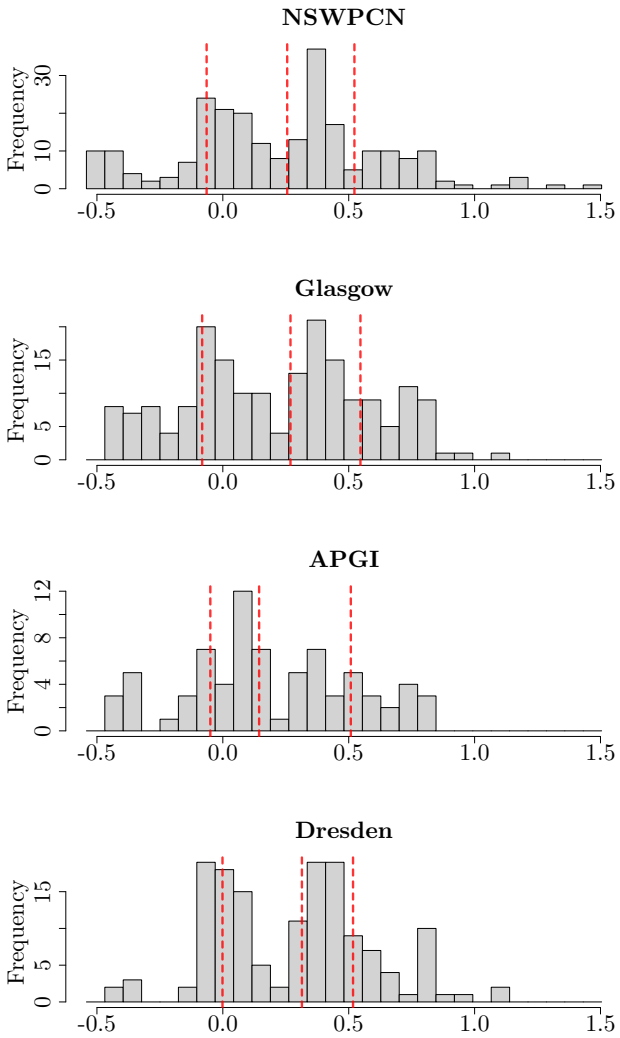
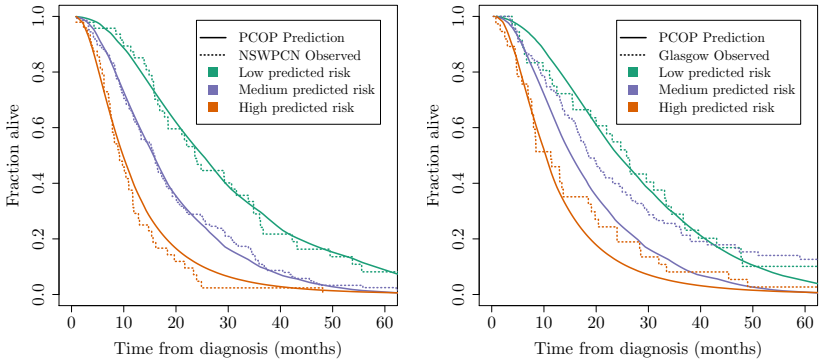
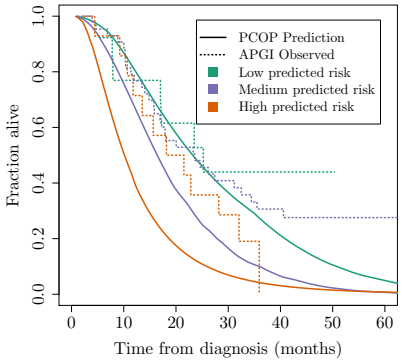


Figure 3.7: Distributions of the PCOP PI in training and validation cohorts. Score distributions were broadly similar in all cohorts, with an approximately bimodal form. Empirical 20th, 50th, and 80th percentiles are indicated by red lines.

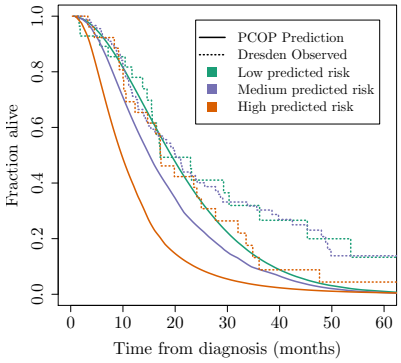


(a) NSWPCN

(b) Glasgow



(c) APGI



(d) Dresden

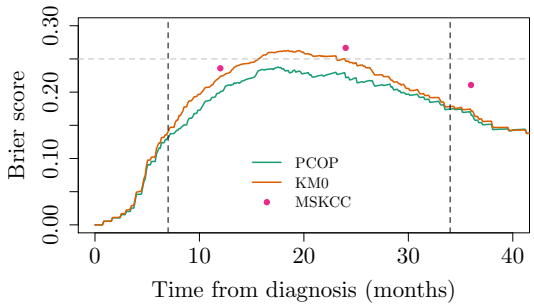
Figure 3.8: Observed and PCOP predicted survival of patient risk groups. Within each cohort, patients were divided into three risk groups: high (red, PCOP PI $<$ 20th percentile), low (green, PI $>$ 80th percentile), and medium (blue, all remaining patients). For each group, a Kaplan-Meier estimate of empirical survival (dotted lines) was compared to the median of PCOP predictions for patients in that group (solid lines). Excellent fit is seen for the NSWPCN training cohort, as expected. Overall fit is poorer for the validation data, with acceptable fit for extreme risk groups in the Glasgow cohort, and generally poor fit in both the APGI and Dresden cohorts.

PCOP and an established prognostic tool.

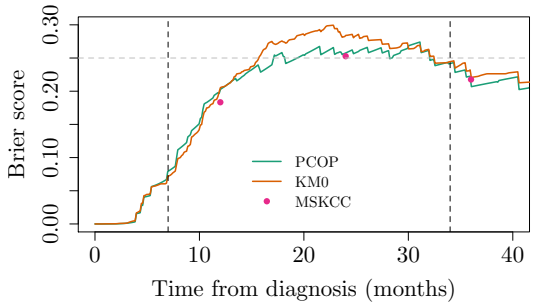
In keeping with the general findings from visual assessment of risk groups, the PCOP was more prognostic than null models in the Glasgow cohort, but not in the APGI or Dresden cohorts, across all times post diagnosis (Figure 3.9). The significance of the difference between PCOP and KM0 Brier score in the Glasgow cohort was assessed by bootstrapping, at 12, 24, and 36 months following diagnosis. The PCOP had a significantly better Brier score than KM0 at 24 months after diagnosis (95% BCa CI $[0.0024, 0.047]$, 500 rounds), but not 12 or 36 months.

The performance of the MSKCC nomogram on preoperative data was poor: in all three cohorts, the performance of MSKCC survival estimate was either not significantly different from, or was significantly worse than, that of the null KM0 model (Figure 3.9). The MSKCC nomogram was designed exclusively for use with postoperative data, and it is reasonable to suppose that this is the reason for its poor performance on preoperative data. However, when all postoperative data were supplied to the MSKCC nomogram, its predictions also were not better than the null KM0 model in both the Glasgow and Dresden cohorts, suggesting either poor general discrimination or calibration of the MSKCC nomogram in these data sets (data not shown).

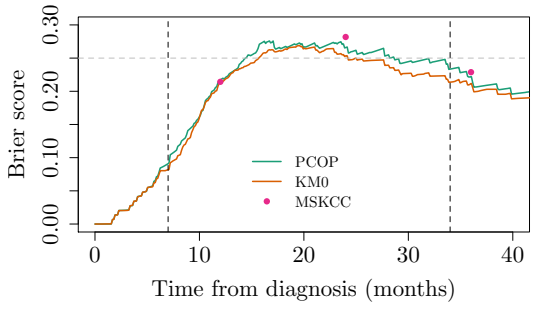
Summary The PCOP demonstrated acceptable overall fit in only one of the three validation cohorts tested. Poor fit can be caused by both model miscalibration, and poor discrimination. Miscalibration of a prognostic model, while not ideal, does not preclude its clinical use [93]: if discrimination remains good, the model can still be used to place patients into risk groups, which inform decision making. On the other hand, if a model shows very poor discrimination, it is generally meaningless. The clinical utility of the PCOP was evaluated by formally testing its discrimination, separately from calibration, in the validation cohorts.



(a) Glasgow



(b) APGI



(c) Dresden

Figure 3.9: Brier score paths for PCOP on validation cohorts. The Brier score measures overall prognostic error, as a combination of calibration and discrimination; lower values are better, and the worst-case theoretical value is 0.25. Brier score paths over time are shown for outcome predictions by PCOP (green line), and the MSKCC nomogram on preoperative data (pink dots). Also shown is a marginal KM prediction of outcome (orange), and the theoretical no-information Brier score limit (horizontal dotted line) – outcome predictors must be substantially better than both of these to be usefully prognostic. The $7 - 34$ month period in which most patients die is delimited by vertical lines. PCOP is substantially better than either the KM or MSKCC predictors in the Glasgow cohort, but all predictors are equally poor in the APGI and Dresden cohorts.

Discrimination

Discrimination of the PCOP on external cohorts was assessed by Harrell's c-index [42], calibration slope tests, and the incident / dynamic TD-ROC [45].

c-index Harrell's c-index is defined as the proportion of patients for which the PI and outcome are concordant: in other words, the empirical probability that, for a randomly chosen pair of patients in a cohort, the one with the higher PI will die sooner. As such, the c-index is an overall measure of the ability of a given PI to discriminate between patients with different outcome. Values of the c-index for the PCOP and preoperative MSKCC PIs were calculated in each validation cohort, and are summarised in table Table 3.4. In the Glasgow cohort, the PCOP was c-index was significantly better than that of a marginal predictor (Penciana test [75] $P = 6.67 \times 10^{-6}$), and was also significantly better than the MSKCC preoperative nomogram c ($P = 0.041$). In the remaining two cohorts, both the PCOP, and the MSKCC preoperative nomogram, showed very poor overall discrimination.

Table 3.4: Harrell's c-indices for PCOP and preoperative MSKCC PIs in validation cohorts. Values of Harrell's c, and associated P-values testing whether c is significantly different from the no-information value of 0.5 ($P_{0.5}$), are given for both the PCOP, and the MSKCC nomogram applied to preoperative data. Both PIs are significantly prognostic in the Glasgow cohort, and only weakly so, if at all, in the APGI and Dresden cohorts.

	PCOP		MSKCC	
Cohort	c	$P_{0.5}$	c	$P_{0.5}$
Glasgow	0.609	6.67×10^{-6}	0.585	6.96×10^{-4}
APGI	0.580	0.045	0.476	0.55
Dresden	0.546	0.124	0.518	0.50

Tests of 'calibration slope' A common approach to test a prognostic model's fit and discrimination on test data is to verify that a PI derived from the model is significantly prognostic when it is used as the sole predictor in a Cox model, and that the PI coefficient is not significantly different from unity. This validation method was applied to both the PCOP and preoperative MSKCC prognostics, on the three validation data sets. The PCOP passed this validation test in both the Glasgow and APGI, but not the Dresden, cohorts, and the preoperative MSKCC did not validate in any cohort (Table 3.5). Despite its validating in two cohorts, the fitted PCOP PI coefficient was consistently less than one, suggesting that overfitting occurred during construction of the PCOP [83].

Table 3.5: Calibration slope tests of the PCOP and MSKCC PI. Coefficients (β) of Cox model fits using either the PCOP or preoperative MSKCC PI as sole predictor are given, along with P-values testing whether the coefficients are significantly different from both zero (P_0), and one (P_1). A PI that passes the test will have results consistent with $\beta = 1$. The PCOP satisfies this requirement for the Glasgow and NSWPCN cohorts, but not the Dresden cohort. The MSKCC nomogram, when applied to preoperative data, fails in all three cohorts. Despite the successful validation of the PCOP in two cohorts, its regression coefficients are consistently smaller than 1, suggesting that the PCOP model was overfit during training.

	PCOP			MSKCC	
Cohort	β	P_0	P_1	β	P_0
Glasgow	0.805	0.001	0.41	0.012	0.284
APGI	0.894	0.036	0.80	0.003	0.634
Dresden	0.527	0.093	0.13	0.003	0.502

Incident / dynamic TD-ROC Both the c-index and calibration slope tests provide single measures of model discrimination, averaged across all observed times. However, a prognostic model's discrimination is not generally constant, but changes over time, and it is possible that a model with poor average performance could have good discrim-

ination at specific follow-up times. To address this, plots of TD-ROC AUCs were used to assess the PCOP's discriminative ability at a range of times after diagnosis.

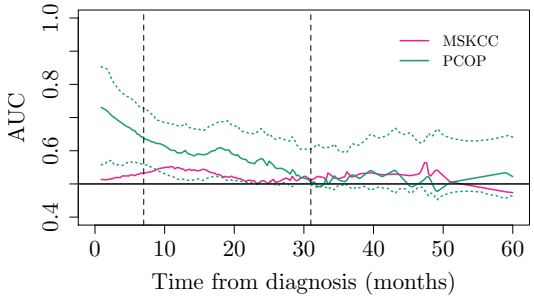
In both the Glasgow and APGI cohorts, the PCOP displayed consistently better discrimination than a null model, and the MSKCC nomogram on preoperative data, for a range of times after diagnosis (Figure 3.10). This result was not repeated in the Dresden cohort, for which both the PCOP and the MSKCC nomogram displayed discrimination that was barely better than baseline.

Validation summary

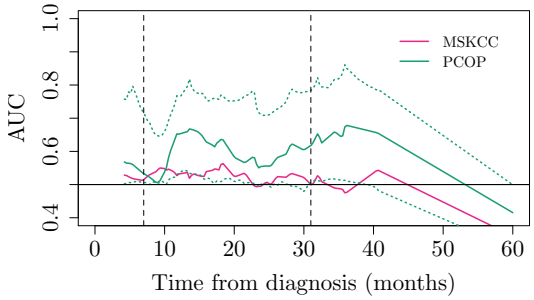
A range of complementary validation approaches were applied to test the performance of the PCOP in external validation cohorts. In terms of overall fit, the PCOP was superior to null models in the Glasgow cohort, but not in the APGI or Dresden cohorts. Specific tests of discrimination were used to check if the poor overall fit was due to poor model discrimination, and indicated that the PCOP discriminated between good- and poor-prognosis patients well in the Glasgow and APGI cohorts, but poorly in the Dresden cohort. The preoperative MSKCC nomogram was a poor prognostic in all data sets.

PCOP web application

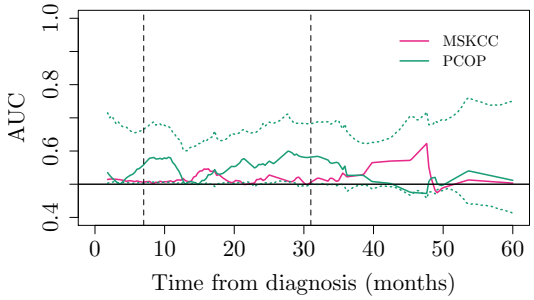
A serious barrier to the use of complex prognostic tools such as the PCOP is the difficulty in their calculation. To address this, a simple web application was created to encapsulate the PCOP, and allow its easy evaluation from any web browser. The PCOP web application interface is illustrated in Figure 3.11. In response to user input of basic clinical parameters, the web interface dynamically recalculates survival estimates, and returns them in graph and table form. Up to two survival curves may be shown simultaneously, to allow the easy comparison of cases. The demonstration PCOP web application resides on the Amazon Web Services (AWS) framework, and so is straightforward to deploy both publicly, and privately, as required.



(a) Glasgow



(b) APGI



(c) Dresden

Figure 3.10: TD-ROC AUC over time in validation cohorts. Bootstrap summaries of the AUC are shown for PIs from the PCOP (mean: green solid lines; BCa 95% confidence intervals: green dotted lines), and the MSKCC nomogram on preoperative data (mean: red lines). The PCOP displays consistently superior discrimination to the no-information level (horizontal line at 0.5) in the Glasgow and APGI cohorts, but performs much more poorly in the Dresden cohort. The MSKCC nomogram PI performs poorly in all cohorts. Vertical dashed lines indicate the $7-34$ month region in which most patients die.

A demonstration instance of the PCOP web application is available at `http://www.pancpredict.net:3277/`.

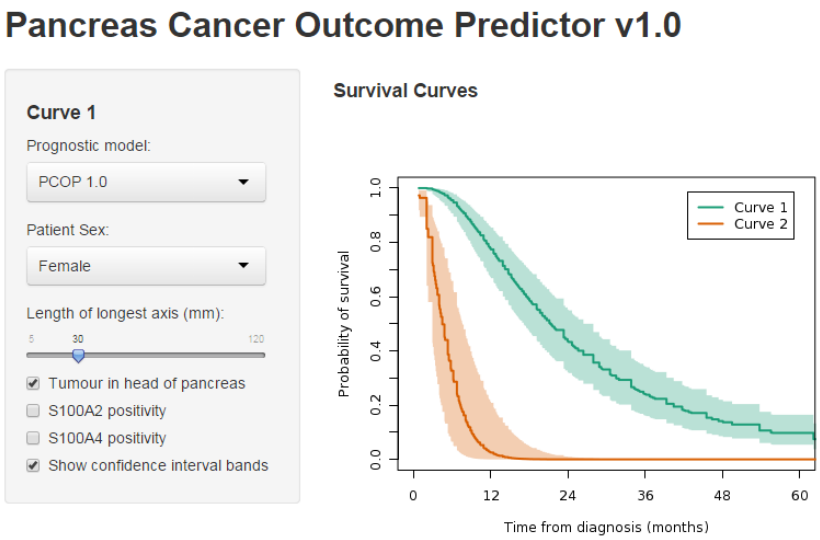


Figure 3.11: Example screenshot of the PCOP web application. The user enters clinical data for up to two cases, and the system dynamically calculates and displays predicted survival curves, optionally with confidence bands.

3.3 Discussion

Is it possible to create a clinically useful preoperative prognostic tool for pancreas cancer? This chapter's work aimed to address part of that question, by developing and validating the PCOP, a prognostic for resected pancreas cancer that uses postoperative measurements as surrogates for preoperative ones. The PCOP validated in one external cohort, but not two others, and the question remains as to whether it, or a true preoperative prognostic based upon it, would ultimately have any clinical utility. Cohort differences and likely confounding were the ultimate cause of the poor validation performance of the PCOP, with the substantial differences between cohorts severely limiting the generalisability of *any* model. Closer examina-

tion of the causes for these cohort differences may identify additional variables that could increase the generality of later iterations of the PCOP. Alternative biomarkers may also yield dividends; the S100A2 and S100A4 markers used here were initially discovered as part of a relatively small candidate screen, and global discovery approaches might be able to identify new biomarkers that substantially improve on them – this direction is explored in Chapter 4. Ultimately, a final version of the PCOP intended for true decision support use will need to be truly preoperatively assessable, and this will require collection of new preoperative-only data, integrating the lessons learned during the development of this pilot version.

The clinical utility of a prognostic tool is a complex function of both the tool's performance, and the clinical context – for example, high-performing prognostics may be clinically useless, and relatively low-performing ones helpful, depending on the particulars of the clinical decision [107]. Certainly, in all cohorts and validation approaches, the PCOP dominated the current standard MSKCC nomogram, when the latter was supplied preoperative data alone. This is unsurprising given that the MSKCC nomogram was never designed to be used preoperatively, but does serve to illustrate that the PCOP is at least as good as the current best available preoperative prognostic option. Whether the PCOP is also ultimately clinically *useful* requires further investigation.

Poor calibration, or even poor discrimination, may not necessarily discount a model from use [93]; to conclusively establish whether a prognostic is helpful in a given clinical situation requires the tools of decision analysis. Unfortunately, a thorough decision analytic treatment of the performance of the PCOP requires accurate, and ideally patient-specific, estimates of the cost and the benefit associated with both the decision to resect and not to resect, or at least an expert consensus on the evidence tipping point for making a decision [108]. These estimates were not available for the particular application of pancreas cancer resection, and so a full decision analytic evaluation of the PCOP could not be performed. There was some indication

that the PCOP would perform favourably in a decision analysis: a Vickers analysis [108] revealed threshold regions for which the use of the PCOP in stratifying treatment did yield a net benefit, in the Glasgow cohort (results not shown). This result was not replicated in the APGI and Dresden validation cohorts, reflecting the general poor performance of the PCOP on these data.

Initial exploratory analysis immediately indicated that there were substantial differences in covariate distribution, and survival, between the cohorts (see, for example, Table 3.1). The cohort covariate differences indicate varying case composition between cohorts, and raise concerns of confounding due to unmeasured variables. This was confirmed by the presence of strong cohort-specific differences in survival that could not be explained using the measured covariates, indicating that important prognostic variables had not been measured. A leading candidate for such an unseen prognostic variable is chemotherapy status, but it is unlikely to be the only such variable. The apparent presence of unmeasured prognostic variables strongly indicates that, in future prognostic development, more covariates should be collected, in the hope of capturing those that are contributing to such strong cohort differences in outcome. The pilot PCOP developed here can be used as an accessory prognostic predictor in such work, reducing the overfitting potential of the expanded model, and improving its validation chances [102]. An additional path to capture more of the cohort-specific outcome differences is by using more, or better, biomarkers of outcome.

Improved biomarkers could be a fruitful avenue to increase PCOP performance across cohorts. The S100A2 and S100A4 markers used in this pilot version of the PCOP were discovered in a small candidate screen [13, 12], and are unlikely to represent the absolute two best protein markers of post-resection outcome in pancreas cancer. A strategy to globally mine the transcriptome for better prognostic biomarkers, ideally-suited for application to clinical cases, is discussed in Chapter 4.

Although this work has only described a pilot version of the PCOP

that relies on postoperative measurements, its results indicate that a preoperative version is likely to be a general, and potentially useful, preoperative prognostic tool for pancreas cancer. Postoperative pathological measurements of tumour size are well-correlated with preoperative imaging estimates [7], and preliminary data on a very small number of patients has indicated that IHC staining of EUS-FNA biopsies for S100A2 and S100A4 is similar to that observed in resected whole tissue (data not presented). Certainly a final fully preoperative version of the PCOP will need to be based on exclusively preoperative measurements, ideally collected prospectively from a number of cohorts. If the cohort confounding issues can be resolved, or given improved biomarkers of outcome, the results of this chapter's work indicate that such a fully preoperative version of the PCOP is very likely to perform well, and certainly better than current best approaches.

3.4 Attribution

The project was initially conceived by myself and Dr David Chang. Dr Chang was responsible for obtaining and curating all raw data for both the NSWPCN and validation cohorts, and performed the biomarker staining and scoring. All subsequent work, from low level quality checks and exploratory analysis, through analysis planning and execution, to interpretation and writing, was done solely by me.

3.5 Methods

Cohort recruitment and ethics

NSWPCN cohort The cohort consisted of consecutive, unselected patients with a diagnosis of PDAC, who underwent curative-intent pancreatic resection at teaching hospitals associated with the Australian Pancreatic Cancer Network. All samples were collected prior to 2003, and are housed at the Garvan Institute of Medical Research, Sydney.

Glasgow cohort PDAC samples were from patients resected at the Glasgow Royal Infirmary from 1997 until 2009. Ethical approval for the use of archival tissue was granted by NHS Research Scotland GGC Biorepository ethical committee Application No. 109 (Southern General Hospital). For the majority of samples, informed consent for storage and research use was obtained before resection; for samples collected prior to the implementation of a prospective consent process, approval for tissue use was granted by the NHS Research Scotland GGC Biorepository.

Dresden cohort Samples were from PDAC resection specimens stored at the Institute of Pathology, University Hospital Dresden, collected between 1993 and 2011. Informed consent for research use was obtained prior to sample collection.

Biomarker staining and scoring

Tissue processing and staining were performed as previously reported [13]. S100A2 was detected using anti-S100A2 mouse monoclonal antibody (clone DAK-S100A2/1, Dako corporation) at 1:50 dilution, and S100A4 using anti-S100A4 rabbit polyclonal antibody (Cat. RB-1804, NeoMarkers USA) at 1:100 dilution. Both primary antibodies were incubated for 60 minutes at room temperature prior to secondary staining, as previously described [13]. Positive S100A2 signal was defined as cytoplasmic staining with intensity 3+ in more than 30% of cancer cells, and positive S100A4 signal was defined as either nuclear or cytoplasmic staining of any intensity in at least 1% of transformed cells.

Model building and selection

All statistical modelling was performed within the `R` environment.

Independence between date of diagnosis and all preoperative variables in the NSWPCN cohort was verified by separate distance covariance [96] permutation tests for each variable, using `R` package `energy` function `dcov.test`, with 499 rounds per test. Date of di-

agnosis was verified to not be prognostic in the NSWPCN samples by a likelihood ratio test on CPH models containing all preoperative prognostic variables and date of diagnosis (as a linear term), or all preoperative prognostic variables alone. Independence and absence of higher order associations was also visually confirmed by examination of smoothed martingale residuals of the CPH model containing all preoperative prognostic variables.

CPH and KM models were fit and analysed using the base package `survival`, and Cox model stepwise variable elimination was performed using the function `stepAIC` from package `MASS`. Generalised gamma survival models were fit using the implementation in package `flexsurv`[1], and package `randomForestSRC` supplied random survival forest functions. The random survival forest model was trained with parameters `splitrule = "logrankscore"`, `nsplit = 2`, and `ntree = 1000`, with all other parameters set to defaults.

Both the incident/dynamic TD-ROC, and the IBS, were used to compare model prognostic performance. TD-ROCs were estimated using R package `risksetROC`, and Brier score paths and IBSs were calculated with custom code, following the Graf derivation [35].

Calculation of a PCOP prognostic index

As the PCOP integrates non-proportional hazards in its survival predictions, it cannot be summarised into a PI as for a proportional hazards Cox model. However, validation methods such as Harrell's c-index, Cox calibration fits, and TD-ROCs, require a patient hazard ranking, as is supplied by the PI. For these methods, each patient's value of the PCOP GG distribution location parameter (β) was used as an approximation to the PI.

[1]Parameter symbols differ between the `flexsurv` package, and this chapter and Cox's formulation of the generalised gamma distribution [26]. In this chapter and Cox's formulation, the generalised gamma location parameter is denoted β, and shape parameters are σ, and λ. In `flexsurv`, these parameters are denoted μ, σ, and Q, respectively.

MSKCC nomogram calculation

The MSKCC prognostic nomogram for resected pancreas cancer [18] was digitised and transformed into `R` code that produced 12-, 24-, and 36-month disease-specific survival estimates given patient CPVs (see app:nomo-mskcc-code on page 153). Predictions for patients with data missing for some nomogram variables were generated by marginalising over the missing predictors, using the reported MSKCC training set variable distributions [18].

PCOP web application

The `R shiny` infrastructure was used to create a simple web application to predict patient outcome using the final PCOP model. This application is available at `http://www.pancpredict.net:3277/`.

Chapter Four

Identifying Optimal Biomarkers for Clinical Tests

Thesis: Large margin binary decision stump classifiers can provide a principled way to translate research biomarker measurement data into simple but high-performance clinical tests.

4.1 Introduction

Research and molecular pathology laboratories take strikingly different approaches to the measurement of biomarkers in patient samples. Research work favours costly manual techniques, which quantify a large number of biomarkers in a relatively small number of samples. Conversely, pathology laboratories make extensive use of highly automated turnkey systems, to robustly measure a relatively small number of biomarkers in a large number of samples. In keeping with this divide, research and pathology laboratories often use very different technologies for the measurement of the same type of biomarker, such as RNA sequencing in research, and quantitative PCR in the clinical realm. This difference in base technology, and the generally increased variability of clinical samples over research ones [46], com-

plicates the translation of discoveries in research into application in the clinic.

Although difficult, this translation of research discoveries into clinical practice is absolutely necessary. Research and pathology techniques are complementary: biomarker discovery requires research techniques capable of interrogating a huge number of potential biomarkers, but the *application* of any discoveries must use pathology techniques that can reliably and economically handle a potentially huge number of patient samples. Finding effective ways to translate research findings into clinical application is critically important. This problem can be approached from two angles: by harmonising measurement technology between research and diagnostic laboratories, or by identifying lead biomarkers in the research laboratory that have a high likelihood of effectively translating to diagnostic measurement methods. The former approach may represent a long-term solution, but does not reflect the current state of technology. To implement the latter approach requires bioinformatic techniques that can search large bodies of research measurement data, to identify a small number of biomarkers that are most likely to make good diagnostic tests.

Effective clinical tests must satisfy a number of requirements, which can be used to guide the translation of a research finding into a clinical tool. Ideally, a clinical diagnostic or prognostic test should be based on the measurement of only a small number of biomarkers [64, 76]. Additionally, it should be highly robust to technical effects, and the inevitable variation in sample quality and handling that comes with clinical specimens [46]. The results of most tests will be interpreted as a simple binary variable[1], and the detection performance of this binary variable should match the particular clinical application [76]. Taking all these requirements into account, a technique to translate discovery biomarker measurements into a clinical test should be able to process measurements of thousands of biomarkers, to identify those single biomarkers that, when their level

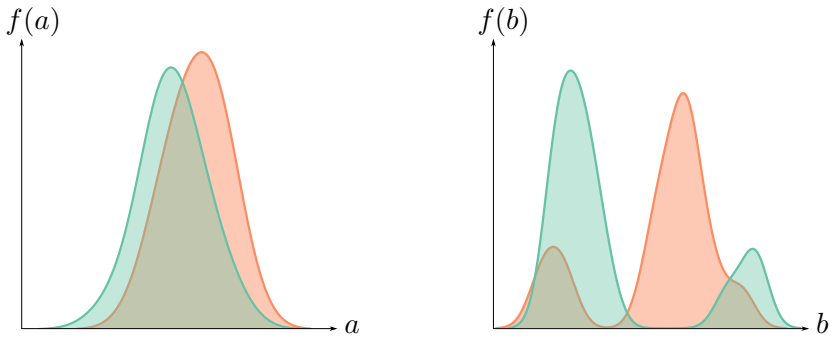
[1]For example, disease versus not diseased, responder versus non-responder, or poor prognosis versus good prognosis.

is thresholded, yield a particular class separation performance, with maximal robustness to technical effects.

Existing methods for the mining of research data for potential biomarkers do not meet these requirements. For both diagnostic and prognostic endpoints, common biomarker prioritisation techniques either do not provide fine control of classification accuracy and robustness, or yield sets of biomarkers that are far too large to practically deploy. Two broad strategies are generally employed to identify biomarkers in large research data sets: univariate statistics, and machine learning. The first strategy encompasses approaches such as per-biomarker regression tests for association between the biomarker and a sample group or outcome, and the second covers the use of algorithms such as support vector machines (SVMs) to construct high-performance predictors of group or prognosis from the measurements of many biomarkers.

Univariate statistic approaches identify single biomarkers of interest, but supply no guarantee as to the suitability of these biomarkers for translation into clinical tests. These statistics are designed to identify differences in average value between two groups, not to construct classifiers. Significance testing and classifier learning are distinct problems, and usually have different solutions: a biomarker that shows a statistically significant difference in signal between two groups may still be a poor classifier; and vice versa (Figure 4.1). Univariate tests also do not take the particular requirements of a clinical problem into account. For example, consider two tests for the same disease – one test is intended to be used as a population screen, and the other to rule out a suspected diagnosis. These two tests would have very different performance requirements, and quite possibly would be best served by two different biomarkers. Univariate test approaches to candidate biomarker selection cannot take this nuance into account, which further limits their usefulness for biomarker selection.

Machine learning algorithms can produce highly accurate classifiers matching given performance requirements [8], but these classi-



(a) Poor classification (b) Good classification

Figure 4.1: Statistical significance does not imply good classification ability. The left and right panels show the distributions of two hypothetical biomarkers, a and b, in two sample classes, shown in orange and green. The intent is to construct a single-feature classifier separating the two sample classes. The levels of biomarker a (left panel) are significantly different between the two classes, but the large overlap in distributions would make it a poor choice for a classifier, with sensitivity and specificity both approximately 0.6. Conversely, biomarker b (right panel) could be used to form an effective threshold-based classifier, with sensitivity and specificity each around 0.8 but it does not show significantly different expression between the groups by the t and Mann-Whitney U tests (although it does by the more general – but seldom-used – Kolmogorov-Smirnov test; data not shown).

fiers generally rely on combining information from a large number of biomarkers. This has not entirely prevented multi-gene tests based on these classifiers from achieving clinical use (an early example of a multi-gene clinical test is MammaPrint™, a simple linear classifier based on measurements of 70 transcripts [103]), but does drastically increase the complexity of translation from the research laboratory to the clinical, and the final cost of the test (for MammaPrint, approximately 4,200 USD per patient). As well as being difficult to translate into a clinical diagnostic, high-performance multi-feature classifiers require care when training [4], their complex workings usually defy

human interpretation [17], and their complexity often simply isn't required for good classification [37]. Many approaches are available to limit the number of biomarkers in classifiers returned by machine learning algorithms [39], but as the number of biomarkers is reduced to one, this generally comes at the cost of performance. For the limiting case of single-biomarker classifiers and prognostics, a machine learning algorithm designed exclusively for the task is needed.

To address this need, in 2009 I reported the Messina algorithm, for the efficient cost-sensitive training of single-feature binary decision stump classifiers [77]. The classifiers found by Messina use the abundance of a single biomarker to place samples into two groups: one containing samples with biomarker levels below a threshold, and the other containing samples with biomarker levels at or above the threshold. Given measurements of many biomarkers in a set of samples from two groups, Messina identifies the biomarkers, and their optimal thresholds, that will separate the two groups well. Messina can be used to address two problems: the selection of biomarkers that are well-suited for development into single-measurement diagnostic tests, and the identification of biomarkers with differential signal between two sample groups. The latter functionality was the focus of the original Messina paper, which demonstrated that Messina was a strong alternative to more classical tests of differential expression, that allowed the user to smoothly trade robustness to outliers against sensitivity to subtle changes in expression. Messina was recommended by an independent comparison of outlier differential expression methods [56], but remains a relatively limited tool: it is only available as a standalone program, and its objective function is hard-coded for binary classification, and so Messina can not generate more general classifiers, such as for prognosis.

Tools to identify biomarkers that can yield accurate estimates of prognosis remain relatively crude, compared to those for classification problems. As discussed in Chapter 3, there is a pressing need in PDAC for effective methods to preoperatively stage patients. Biomarkers of prognosis can be measured preoperatively, but the

S100A2 and S100A4 markers tested in Chapter 3 were relatively poor at predicting outcome following resection of PDAC. S100A2 and S100A4 were identified following only a small candidate screen, and it is reasonable to expect that a global search for PDAC biomarkers of prognosis will yield candidates with better performance. Prognostic tests have the same clinical translation requirements as diagnostic tests: they must be based on a small number of biomarker measurements, and by simple thresholding of these measurements, yield a robust prediction of differential outcome. Machine learning algorithms such as random forests and SVMs have been adapted to the prediction of outcome, but suffer the same general problem of machine learning solutions: too many biomarkers must be measured to achieve accurate prediction. Most univariate statistics approaches also suffer the same problem as their classification counterparts, by failing to take the particular class separation needs of the application into account. These problems with conventional approaches have led to the common use of ad-hoc methods when identifying biomarkers of prognosis.

The poor suitability of conventional approaches for identifying biomarkers of prognosis has led to the common use of ad-hoc analyses for this problem, but these usually perform very poorly. Methods in common use are the median cut, in which the cohort is divided into two groups by the median value of each biomarker, and then a log-rank test performed on these two groups; and the so-called 'optimal' cut, in which many cutpoints are tried, and the best log-rank statistic is used as the final test statistic. The median cut approach presupposes that the poor- and good-prognosis subgroups in the data are equal in size, and has poor power if this is not the case, and the 'optimal' cut procedure is decidedly non-optimal, with an unsurprising high false detection rate due to uncorrected multiple testing. Although discussion of the significant deficiencies in these approaches has been in the medical literature for over twenty years [6], they continue to see routine use and further development [20, 21, 24, 112]. A natural approach to correcting the 'optimal' procedure is by in-

corporating multiple testing correction, and software is available to perform this as `R` package `maxstat` [24, 51]. However, in common with more conventional univariate approaches, `maxstat` and similar methods do not allow for control over any desired aspects of the prognostic classifier, such as the hazard ratio between the poor- and good-prognosis groups, or the relative group sizes. As in the classification case, a biomarker selection procedure specifically designed for clinical translation is needed.

In this chapter I present Messina2, a general algorithm for single-feature decision stump classifier learning. Like its predecessor Messina, Messina2 aims to produce maximally robust classifiers that satisfy user-supplied performance requirements. Unlike Messina, Messina2 accepts any performance metric, and so can be used to create threshold classifiers for any task, including diagnosis and prognosis. Messina2's emphasis on maximally robust classifiers means that the biomarkers that it selects are especially well suited for translation to the clinical diagnostic laboratory, helping to address the divide between research discovery and clinical application. To facilitate Messina2's use in bioinformatic workflows, it has been made into the `R` package `messina`, available as part of the Bioconductor project [33][2]. This chapter will describe the general framework of binary decision stump classifiers, detail the Messina2 algorithm and how it differs from Messina, and demonstrate its characteristics and performance relative to other techniques. An application to the problem of identifying better prognostic biomarkers in pancreas cancer will illustrate the utility of Messina2 in real data.

4.2 Messina2

Messina and Messina2 are algorithms that take as input measurements of biomarker abundance in a number of samples, and return as

[2]The version of Messina2 used for this thesis is not yet in the latest Bioconductor release. To access the thesis version, clone the `thesis` branch from the Bitbucket repository `marpin/r-messina`, or from `R`, with `devtools` installed, run the command `devtools::install_bitbucket("marpin/r-messina", ref = "thesis")`.

output decision stump classifiers that can separate the input samples into two groups. For Messina, these groups are based on user-supplied sample labels, such as cancer and non-cancer, and the classifiers are designed to separate these two sample types as robustly as possible, whilst still meeting user-supplied minimum sensitivity and specificity requirements[3]. Messina2 is more general, producing classifiers that satisfy a user-supplied objective function as robustly as possible. One specific form for this objective function yields results very similar to those of Messina, but through use of different objective functions, Messina2 can be used to address effectively any binary separation problem, such as the creation of decision stump classifiers to distinguish between good- and poor-prognosis patients. This chapter will primarily describe Messina2, drawing comparisons to Messina where appropriate.

Messina2 framework

Decision stump classifiers A decision stump classifier is a very simple classifier that is equivalent to a one-level decision tree [54]. In the binary variant employed by Messina2, an unknown sample is assigned to one of two classes, depending on whether that sample's abundance of a certain biomarker is above, or below, a threshold value. More precisely, let the abundance of a specific biomarker in sample i be x_i. Sample i can be a member of one of two classes, here denoted 0 and 1, and the goal is to predict the class of i based on x_i. Then, a decision stump classifier will predict the class of i, w_i, as $w_i = [x_i d \geq td]$, where t is the threshold level at which the predicted

[3]Here sensitivity and specificity are used in their binary classification sense. In a binary classification problem (for example, a diagnostic test to detect the presence of a disease, where the result is either 'diseased,' or 'not diseased,' with no intermediate 'uncertain'), a given classifier tested on a population of samples will have associated rates of true positive, true negative, false positive, and false negative calls. The sensitivity of this classifier on that population is then the true positive call rate divided by the total positive sample rate, and the specificity is the true negative call rate divided by the total negative sample rate. Symbolically, let R_{ab} be the rate at which a sample which is truly of class a is called as class b, $a, b \in \{0, 1\}$, where 1 represents the positive class, and 0 the negative. Then, $\text{Sens} = p_n = R_{11} \div R_{1\bullet}$, and $\text{Spec} = p_c = R_{00} \div R_{0\bullet}$.

class changes, and $d \in -1, 1$ is the classifier direction: if $d = 1$, then values of x at or above t are predicted to be class 1, if $d = -1$, then values of x of at or above t are predicted to be class 0. [] denotes the Iverson bracket (see Conventions on x). This is illustrated for $d = 1$ in the top two lines of Figure 4.2. A decision stump classifier for a given biomarker is completely specified by the values of t and d; finding values for these two variables that in some way meet desired criteria is the goal of classifier training.

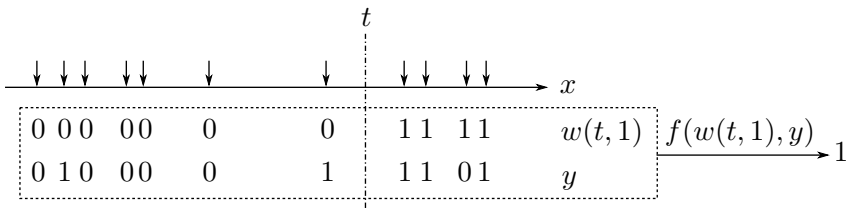


Figure 4.2: Decision stump classifiers and the Messina2 objective. Observed measurements of a hypothetical biomarker in eleven samples are shown as arrows on the x line. Samples are placed into two classes (0 and 1) based on the value of that sample's biomarker relative to a given threshold t; these classes are contained in the tuple $w(t)$. Each sample is also associated with a value y; in this example y is binary (perhaps indicating non-diseased versus diseased samples), but it can be any arbitrary type. Messina2 aims to identify values of t that lead to sample classes $w(t)$ that in some way separate the values of y in a desired fashion. This is formalised in the objective function f, which is supplied $w(t)$ and y, and returns 1 if the sample classes in $w(t)$ lead to a desired separation of y, else 0.

Objective function To identify optimal values of t and d for a given problem, Messina2 employs the concept of an *objective function*. Suppose that we have measurements of a certain biomarker x for each of n samples. For each sample, we also have the value of a dependent variable y; this may be a biological class (eg. cancer or non-cancer), a survival time, or any multidimensional measurement. Our goal is to create a decision stump classifier that operates on x, so that the two groups of samples so created are somehow 'well-separated' in their values for y. The concept of 'well-separated' is

left deliberately vague, because it is under the control of the user, via the objective function. Reasonable examples of good separation would be: for diagnosis, most of the diseased samples are in one group, and most of the non-diseased samples are in the other; and for survival, there is a large difference in prognosis between the two sample groups. We formalise this concept of objective function as follows. Let the objective function $f : (\mathbb{B}^n, \mathbb{Y}^n) \rightarrow \mathbb{B}$ accept an n-tuple of sample groupings (as defined by t and d) w, and an n-tuple of dependent values y, and return a binary value – 1 if the split of y given by w is desirable, otherwise 0. The operation of the objective f for a binary y is illustrated in Figure 4.2. Here w is expressed as a function of t and d, and the particular split given is considered desirable, so that $f(w(t,d),y) = 1$. For a given x, y, and f, it is possible that many values of t and d satisfy the objective function, that is, $f(w(t,d),y) = 1$ does not have a unique solution in (t,d). In this case, a principled way to select a (t,d) that is in some sense optimal, is required.

Margin A concept in machine learning is that of the *margin*. Informally, the margin describes the distance between the decision boundary (here, the threshold t) and the data, and is a measure of the confidence and robustness of a classifier: a classifier with a high relative margin has all observed data points safely 'far away' from the decision boundary, and unlikely to change class following a perturbation. Messina and Messina2 use a slightly different concept of margin, as the smallest distance that the decision boundary must move for the objective function to be violated. Using the heuristic that classifiers with the greatest margin are the most robust, Messina2 chooses the optimal threshold t^* as the midpoint of the largest real interval for which $f(w,y) = 1$, and selects the optimal direction d^* as the one associated with t^*. The intuition behind this selection is that such a choice of threshold and direction maximises the size of the smallest threshold perturbation that is necessary to make the classifier violate

its performance requirements, that is,

$$(t^*, d^*) = \underset{t,d}{\operatorname{argmax}} \min \{|\delta| \ : \ \delta \in \mathbb{R} \wedge f\,(\ (\ [x_i d \geq (t+\delta)\, d]\ )_{i=1}^{n}, y) = 0\}$$

where $f([\ [x_i d \geq (t+\delta) d]\ ]_{i=1}^{n}, y)$ is the value of the objective f when the threshold $t+\delta$ is applied to the data x, with δ the perturbation. This concept is illustrated in Figure 4.3, which shows a wide region of potential thresholds t in which the objective function f is true. The optimal threshold t^* is in the centre of the largest interval for which f is true, thereby maximising the margin Δ.

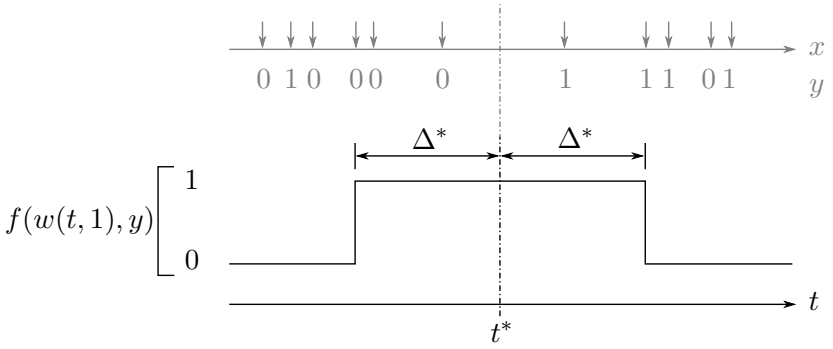


Figure 4.3: Finding the optimal Messina2 threshold. Typically, a range of threshold values t will yield an acceptable y split, as determined by the objective f, and the goal is to select a single optimal threshold t^*. This example illustrates the concept. The value of $f(w(t, 1), y)$ is shown for a range of values of t, with the training data in grey above. A range of t values yields acceptable y separation, as judged by f, and Messina2 selects t^* as the midpoint of the largest such interval. The classifier margin Δ^* is calculated as half the width of this largest interval. This example only considers classifier direction $d = 1$, the case for $d = -1$ is analogous.

Internal validation Messina2 considers all possible thresholds and directions $t \in \mathbb{R}, d \in \{-1, 1\}$ when searching for (t^*, d^*), and as such is at risk of overfitting the data, and reporting spurious classifiers. To overcome this problem, Messina2 by default employs an outer bootstrap loop to verify that a classifier under consideration still satisfies the objective function f, when tested on holdout data. Multiple boot-

strap draws of the training data (x, y) are made, and for each draw a Messina2 classifier is trained on the drawn samples, the trained classifier is used to separate the undrawn (so-called 'out-of-bag') samples into two classes, and the objective f is evaluated on the resulting two classes in the out-of-bag data. Messina2 only returns classifiers for which the objective f is true in at least half of the out-of-bag evaluations. This outer bootstrap layer adds an additional level of stringency to the Messina2 results, but dramatically increases the runtime of the algorithm, and in some applications may not be required: the VapnikChervonenkis dimension of a single-feature threshold classifier is 2, and so Messina2's potential for overfitting is relatively low [105], even before bootstrapping is performed.

Objective functions

Messina2 is designed to accept any objective function f, to allow custom analyses, but is also supplied with two reasonable default objective functions, designed to handle common tasks in classification and survival analysis.

Classification The objective function f_M is designed to mimic the behaviour of the original Messina algorithm for classification and differential expression detection. This function returns true if the data split w separates the classes in y, with sensitivity and specificity at least as good as user-supplied minimum values:

$$f_M(w, y) = [p_n \geq l_n \wedge p_c \geq l_c]$$
$$p_n = \frac{\sum_i [w_i \wedge y_i]}{\sum_i [y_i]}$$
$$p_c = \frac{\sum_i [\neg w_i \wedge \neg y_i]}{\sum_i [\neg y_i]}$$

In the above, p_n and p_c are the observed sensitivity and specificity, respectively, and l_n and l_c are their user-supplied minimum acceptable values. When used with this objective function, Messina2 will return

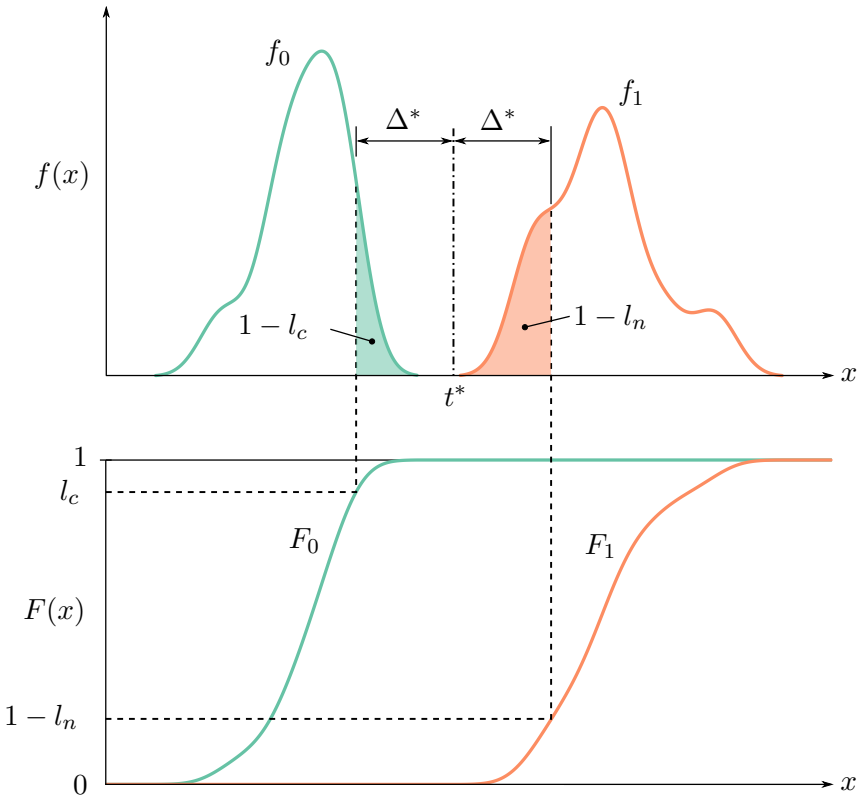
maximum-margin classifiers that separate the two groups defined in y with at least the required minimum sensitivity and specificity. When identifying lead biomarkers for the construction of a clinical test, the analyst can set the sensitivity and specificity requirements to values matching the intended application of the test, and be confident that the biomarkers reported by Messina2 will robustly separate the sample groups with at least this desired level of performance. As sensitivity and specificity are closely related to a variable's cumulative distribution function (CDF), the objective function f_M has close links with the data population CDFs. The optimal threshold t^* is the midpoint between values of t for which the minimum sensitivity and specificity requirements are violated, and if the true CDF were known for positive and negative sample populations, could be directly calculated from the inverse CDFs (Figure 4.4).

Prognosis The objective function f_H is intended for survival analysis, where the goal is to identify biomarkers that can robustly separate a cohort into good- and poor-outcome groups. Messina2 in this mode provides a more principled approach to thresholding biomarker levels for prognosis than median-cut or multiple-cut procedures. $f_H(w, y)$ returns true if the approximate hazard ratio (using survival data in y) between the groups defined by w is at least a user-supplied value. The approximate hazard ratio is calculated from the log-rank test statistic, as:

$$f_H(s, y) = [|\log p_h| \geq \log l_h]$$
$$\log p_h = Z\sqrt{\frac{4}{n_e}}$$

where p_h is the approximate hazard ratio achieved by the data split in w, l_h is the minimum hazard ratio desired, Z is the value of the log-rank test statistic[4], and n_e is the number of events in y.

[4] Z is the signed log-rank test statistic, not to be confused with the χ^2 statistic, which is also encountered. $\chi^2 = Z^2$.



$$t^* = \tfrac{1}{2}\left(F_1^{-1}\left(1 - l_n\right) + F_0^{-1}\left(l_c\right)\right)$$
$$M = \tfrac{1}{2}\left(F_1^{-1}\left(1 - l_n\right) - F_0^{-1}\left(l_c\right)\right)$$

Figure 4.4: Messina2 f_M is closely linked to CDF estimation. Supposing that the joint x, y distribution is exactly known, the optimal threshold and margin t^*, Δ^* of a Messina2 classifier with an f_M objective is a simple function of the target sensitivity and specificity. This is demonstrated above, where it is apparent that the fraction of false positives in class 0 (shaded green) is equal to $1 - F_0(t^* - \Delta^*)$, and the fraction of false negatives in class 1 (shaded red) is $F_1(t^* + \Delta^*)$. These values are equal to $1 -$ specificity and sensitivity, respectively, permitting direct calculation of the optimal classifier if the class CDFs are exactly known. Of course, this is never practically the case, but the link between Messina2's f_M objective, and the joint distribution of x, y, is useful in characterising the performance of this objective.

Messina2 vs Messina

Messina2 using objective function f_M produces results almost identical to those of the original Messina (results not shown). The fundamental point of departure between Messina2 using f_M, and Messina, is in the internal validation step. In Messina2, a classifier is considered to have passed internal validation if more than half of the validation bootstrap iterations had acceptable out-of-bag sensitivity and specificity. Conversely, in Messina, classifiers pass internal validation if the mean out-of-bag sensitivity, and mean out-of-bag specificity, were acceptable (that is, at least as high as the user-supplied limits). If the out-of-bag sensitivity and specificity estimates are well-behaved (in preliminary investigations, this appears to be the case), then these two criteria should produce very similar results.

Algorithm

For brevity, the algorithms for Messina2 are contained in app:messalgos on page 159. Their overall architecture and function is briefly described here.

Messina2 operates on each biomarker in a data set separately, reporting whether that biomarker can be used to build a threshold classifier with the desired performance characteristics, and, if so, the optimal such classifier. The main entry point for Messina2 is Messina2Entry (Algorithm 1), which performs bootstrap validation and final training for a single biomarker. Messina2Entry calls Messina2Core (Algorithm 2), which performs the actual classifier training. Although the preceding text described the search for the optimal t as being over $\mathbb{R}$, this is not required in practice. For given d, $w(t, d)$ only changes at values observed in x, so it is only necessary to check one value of t in each interval spanning consecutive unique values of x. Candidate cutpoints satisfying this requirement are generated by MakeCutpoints (Algorithm 3). The resultant sparse $w(t, d)$ requires some care to recover correct threshold and margin values; the logic for this is in BestInterval (Algorithm 4), the operation of which is illustrated in Figure F.1.

4.3 Results

In the following, the performance of Messina2 is characterised through simulation, and its behaviour compared to alternative techniques. PDAC outcome cohorts are used to demonstrate that prognostic biomarkers identified by Messina2 are more likely to validate in external cohorts than those identified by the `maxstat` method. Finally, Messina2 is applied to PDAC prognostic data from the APGI, in an attempt to identify biomarkers that may form better preoperative measures of prognosis than those used in Chapter 3.

Large-margin classifiers are more robust

We consider a simple technology translation scenario in which a diagnostic classifier developed on one platform is moved to another. The classifier is based on the abundance of biomarker x, which is used to predict disease state $y \in \{0, 1\}$, by comparing x against a threshold t, $\hat{y_i} = [x_i \geq t]$. Suppose that t was originally derived from very many (so that the effect of sampling can be ignored) measurements of x made on one measurement platform, as $x_i = Dy_i + \epsilon_i$, with $\epsilon_i \sim \mathcal{N}(0, 1)$ is the measurement error of the platform. No generality is lost by forcing the error to be standard normal, as this situation can be achieved simply by scaling D. We use the link between margin and the distribution of x, y for objective f_M (Figure 4.4) to derive an expression for the margin of this original classifier. Given a minimum desired sensitivity and specificity of 0.9 each, the Messina2 classifier trained on sufficient samples of x, y will have $t = \frac{1}{2}D$ and $\Delta = \frac{1}{2}(\Phi^{-1}(1 - 0.9) + D - \Phi^{-1}(0.9)) \approx \frac{1}{2}D - 1.28$. Messina2 will not return a classifier with a non-positive margin, so $\Delta > 0 \Rightarrow D > 2.56$ for this example.

The classifier described above is now to be applied on a new measurement platform, and we are interested in determining how its performance will be affected by this change. It is reasonable to expect that the new platform will be calibrated to approximately match the original, so that no significant translation or scaling of the values of

x will occur. However, the post-calibration error of the new platform may be very different from that of the old, and this disparity can't be easily corrected. Let the measurements of x on the new platform be $x'_i = Dy_i + \epsilon'_i$, with $\epsilon'_i \sim \mathcal{N}(0, \sigma'^2)$, and the predictions on the new platform be $\hat{y'_i} = [x'_i \geq t]$. Our intent is to characterise how the error rate of $\hat{y'}$ changes as a function of Δ and σ'. False positives will occur when $\epsilon'_i \geq t$, and false negatives when $D + \epsilon'_i < t$; these two error types together will occur at a rate of

$$\begin{aligned} R_E &= f_0 \left(1 - \Phi\left(\frac{t}{\sigma'}\right)\right) + f_1 \left(\Phi\left(\frac{t-D}{\sigma'}\right)\right) \\ &= f_0 \left(1 - \Phi\left(\frac{D}{2\sigma'}\right)\right) + f_1 \left(\Phi\left(-\frac{D}{2\sigma'}\right)\right) \\ &= \Phi\left(-\frac{D}{2\sigma'}\right) \\ &= \Phi\left(-\frac{\Delta}{\sigma'} - \frac{1.28}{\sigma'}\right) \end{aligned}$$

where f_0 and f_1 are the relative proportions of class 0 and 1 in the test population; these disappear from the expression following simplification.

In this simulation, the total error rate therefore increases with increasing measurement error σ', and decreases with increasing margin Δ: as intuitively expected, classifiers with higher margin are more robust to measurement error. This is shown in Figure 4.5, in which increasing margin can be seen to offset the increase in error rate caused by a high σ'.

Messina2 provides accurate control over classifier margin

High-margin classifiers are desirable due to their robustness, but is Messina2 actually effective in building classifiers with high margin, and how does this compare to other techniques? To address this, Messina2 and representative other techniques were applied to simulated diagnostic and prognostic data. Methods were compared for sensitivity – the rate at which they detected simulated biomarkers

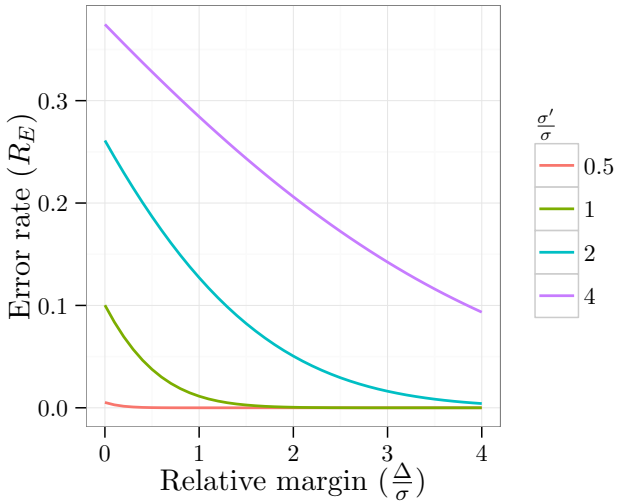


Figure 4.5: High margin increases a classifier's robustness to technology translation. The class prediction error rate of a Messina2-type classifier is shown as a function of the classifier margin Δ, and the noise level of a new measurement technology, σ'. To facilitate dimensionless comparison, Δ and σ' are shown scaled relative to the noise level σ of the technology used to generate the Messina2 training data.

with desired characteristics – and specificity – the rate at which they rejected simulated biomarkers lacking desired characteristics.

Classification (f_M) Messina2 with an f_M objective was compared to the t-test. Simulated x_i were drawn from a normal distribution, as $x_i \sim \mathcal{N}(Dy_i, 1)$, with $D \in [0, 5]$ and $y \in \{0, 1\}$. Classes were balanced, so that near-equal numbers of $y = 0$ and $y = 1$ samples were present in all runs. The total number of samples in a run, n, was either 25, 50, or 100. For each combination of D and n, 1,000 random cohorts were generated and supplied to Messina2 and the t-test, and detection rates recorded. Messina2's f_M parameters were $l_n \geq 0.8$, $l_c \geq 0.8$, and $\Delta \geq 1$, and the t-test size (α) was 0.05.

Across all sample sizes, Messina2 consistently only selected a biomarker when its true margin was near the target value of 1 (Figure 4.6). Conversely, the t-test was far less selective, identifying even biomarkers with a zero margin as being differentially-expressed. This

sensitivity to small differences, particularly in large samples, is a concrete example of the effect illustrated in Figure 4.1: a statistically significant result does not necessarily indicate that a good classifier can be made. The result of this simulation suggests that if the t-test were used to identify single-gene diagnostic biomarkers, a very high false positive rate could be expected, and many of the reported biomarkers would lead to small-margin classifiers that would be unlikely to be robust. This effect only becomes more pronounced as the sample size increases.

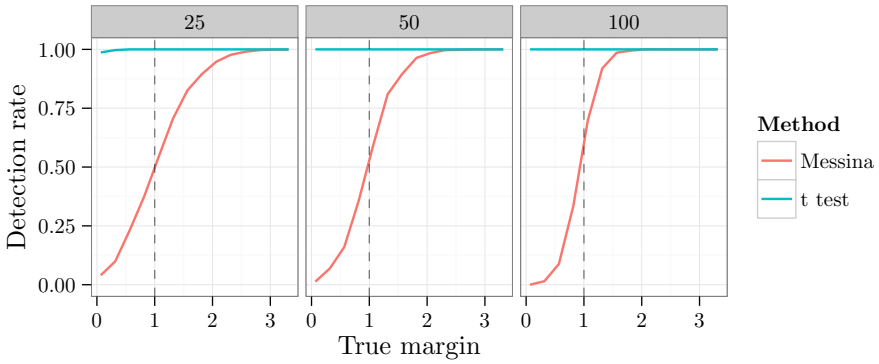


Figure 4.6: Messina2 with f_M is less sensitive, but more specific, than the t-test. Across three sample sizes (top bars), simulated biomarker measurements for two groups were drawn from normal distributions with a range of class separations, and the rates at which Messina2 and the t-test detected a simulated biomarker were compared. The t-test detected almost all simulated biomarkers under all conditions, even those for which the margin was zero, and robustness would likely be poor. In contrast, Messina2 reliably only detected biomarkers with margins at least as high as its target of 1, shown as dashed vertical lines. If the goal is to identify biomarkers that can be made into high-margin diagnostic classifiers, these results indicate that Messina2 is a far more discriminating selection method than the t-test.

Prognosis (f_H) Messina2 using the f_H objective function was compared to the `maxstat` corrected optimal cut procedure, and the simple median-cut log-rank test. Simulated data were generated in a similar fashion to the f_M test above, as follows. Simulated x_i were drawn as

$x_i \sim \mathcal{N}(Ds_i, 1)$, with $D \in [0, 5]$, and $c_i \in \{0, 1\}$ the class membership indicator. The fraction of class 1 in the sample was either 0.2 or 0.5, with number of samples $n \in \{25, 50, 100\}$. Event times were sampled from an exponential distribution with rate $4c_i$, and censoring times from an exponential distribution, with rate calculated to give an expected censoring rate of 0.2 [5]. 500 simulations were performed for each unique combination of parameters. Messina2's f_H arguments were $l_h \geq 2$ and $\Delta \geq 1$, and `maxstat` was used with test `"LogRank"` and correction method `"HL"`. For `maxstat` and the median-cut, a biomarker was considered detected if its corrected P-value was less than 0.05.

The detection results on these simulated data are shown in Figure 4.7. Messina2 was not as reliable at controlling classifier margin in these prognostic data as it was in the classification case (Figure 4.6); this discrepancy may be due to the greater uncertainty inherent in censored data. Despite this, Messina2 remained competitive with the other cut optimisation procedure `maxstat`, particularly in small samples. Messina2 was also substantially faster than `maxstat`, on the test system taking 0.41 seconds per biomarker per core for $n = 100$, versus 4.80 seconds for `maxstat`[6]. As expected, the median cut procedure was very sensitive when the survival subgroups were balanced, but performed poorly otherwise.

Messina2 is more effective than maxstat at finding validating prognostic biomarkers

The preceding simulation studies have indicated that high-margin threshold classifiers are robust to measurement noise, and that Messina2 is efficient at learning such classifiers, but the advantage of Messina2 over other methods on real data remains to be shown. The GEX measurements in PDAC cohorts used in Chapter 2 provide a platform for this comparison, by examining the rate at which a prognos-

[5] A number of values of the censoring fraction were tested, which for reasonable levels was observed to have only minor effects on results (data not shown).

[6] Messina2Core is much faster again; approximately 50 times the speed of Messina2, at default settings.

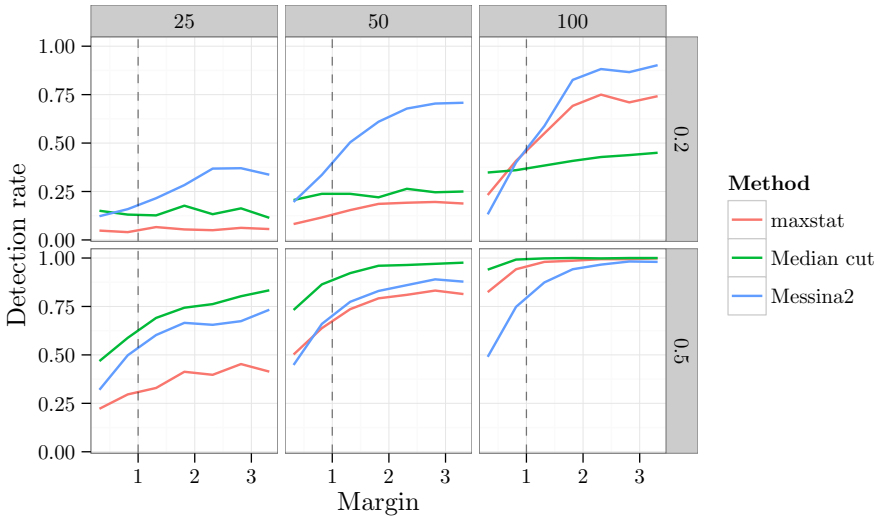


Figure 4.7: Messina2 f_H prognostic performance, compared to `maxstat` and median cut. Detection rate for each method is shown as a function of margin, for sample sizes of 25, 50, and 100 samples (top bars), and class 1 proportions of 0.2 and 0.5 (side bars). The target classification margin of 1 is shown as dashed vertical lines. Although Messina2 (blue) did not control margin as effectively as in the f_M classification case, it performed competitively against `maxstat` (red), particularly in the case of small sample sizes. The median cut procedure (green) was highly sensitive when classes were balanced (class 1 proportion of 0.5), but performed poorly otherwise.

tic biomarker found by a given technique in one cohort, will validate as prognostic in a separate cohort. The PDAC cohorts contain patients from different countries, with differing clinical characteristics, and with biomarkers measured by different technologies, and so provide a strong test of the generalisation ability of classifiers built by Messina2 and `maxstat`.

Pseudo-receiver operating characteristic (ROC) curves were generated to compare the validation performance of default Messina2, Messina2 without bootstrap (Messina2Core), and `maxstat`. Classifiers of each type were trained on data from the APGI PDAC cohort, and validated on data from cohorts GSE28735, and TCGA paad. To generate a pseudo-ROC curve for each method, the trained classifiers

were first ranked by a quality metric (increasing margin for Messina2, decreasing P-value for `maxstat`). Then, for each rank of the metric r, two values were calculated: the proportion of classifiers with rank at least as good as r that did validate (the validation rate, similar to the true positive rate of a conventional ROC), and the proportion of classifiers with rank worse than r that did *not* validate (the non-validation rate, similar to the false positive rate of ROC). The pseudo-ROC is then the curve formed by plotting the validation rate against the non-validation rate, for all ranks of the metric. A detailed description of the validation procedure and pseudo-ROC calculation can be found in the Methods section.

Both Messina2 and Messina2Core produced classifiers that were far more likely to validate than those returned by `maxstat` (Figure 4.8). Messina2Core in particular demonstrated excellent validation performance, suggesting that – at least for the objective f_H – the increased stringency introduced by Messina2's bootstrap validation procedure came at the expense of substantial sensitivity. For the very highest-margin classifiers, Messina2's performance was close to that of Messina2Core, indicating that the sensitivity loss is primarily restricted to lower-margin biomarkers. `maxstat` gave very disappointing performance in general, with validation performance only marginally better than that of random selection, emphasising the robustness of classifiers generated from a machine learning perspective, rather than a statistical one.

Application of Messina2 to identify biomarkers for PDAC prognosis

The very promising results of Messina2 in pseudo-ROC analysis prompted its use in the APGI PDAC cohort, to identify new prognostic biomarkers for use in a preoperative prognostic tool like that developed in Chapter 3. Messina2 was used to identify prognostic biomarkers in the APGI GEX and follow-up data described in Chapter 2, using objective function f_H with $l_h = 2$, and $\Delta \geq 1$. Of 13,000 mRNAs tested, three passed these requirements (Table 4.1, Figure 4.9). The

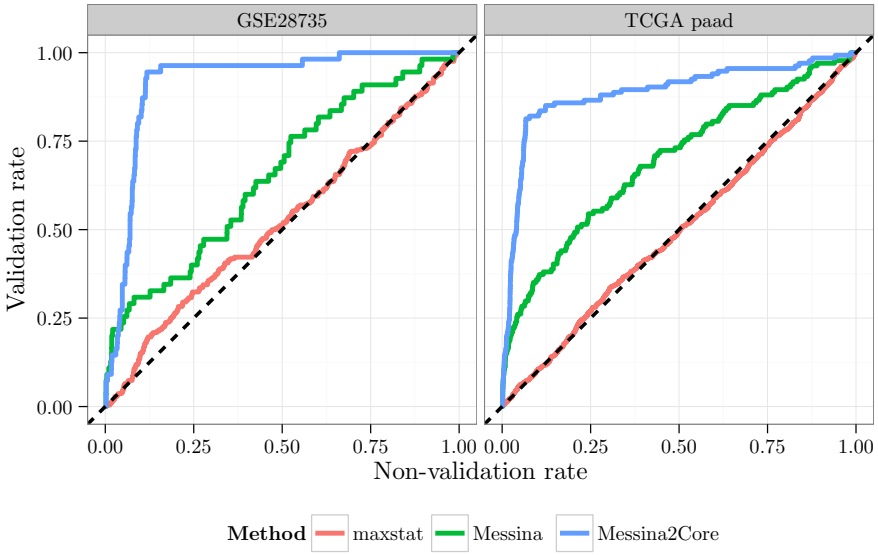


Figure 4.8: Messina2 identifies classifiers much more likely to validate than those found by `maxstat`. Pseudo-ROC curves (see accompanying discussion and methods for a more complete description of these curves) are shown comparing the validation performance of classifiers found by applying Messina2, Messina2Core, and `maxstat`. Classifiers were trained on the APGI PDAC cohort, and then tested in validation cohorts GSE28735, and TCGA paad. Messina2Core demonstrates excellent performance, being able to effectively separate validating classifiers from non-validating classifiers in both cohorts. Messina2 was as sensitive as Messina2Core for high-margin classifers (bottom left of the graphs), but showed greatly diminished sensitivity for lower margin classifiers. `maxstat` was barely better than random, and overall performed very poorly for the generation of classifiers that would be likely to validate in external cohorts measured by different technologies. The dotted line with slope 0.5 indicates the performance expected from random selection of biomarkers.

genes associated with these mRNAs have been reported to be linked to cancer: KRT6A and KRT6C are associated with poor prognosis basal cancers [23, 67], KRT6A in particular has been linked to poor prognosis in PDAC [101], and ANGPTL4 is involved in metastasis in a range of cancers [3, 58, 74, 111].

Table 4.1: High-margin outcome-associated biomarkers found by Messina2 in the APGI PDAC cohort.

Gene	Threshold (t^*)	Direction (d^*)	Margin (Δ^*)
KRT6A	9.50	1	2.00
ANGPTL4	8.90	1	1.36
KRT6C	7.46	1	1.17

The Messina2 classifiers of Table 4.1, trained on the APGI cohort, were tested for prognostic differentiation in the GSE28735 and TCGA paad validation cohorts, but none produced sample groupings with significantly different prognosis (Four log-rank tests, as not all genes of Table 4.1 were present in the validation cohorts, all $\chi^2 < 1.64$). This lack of validation, while disappointing, is quite plausibly due to small cohort size (TCGA), microarray probe dropout (GSE28735), or simply the very small number of genes tested, and more involved validation (by immunohistochemistry, for example) may prove fruitful.

4.4 Discussion

The thesis driving this work was that "large margin binary decision stump classifiers can provide a principled way to translate research biomarker measurement data into simple but high-performance clinical tests." Confirming this hypothesis, the preceding sections have established the robustness of Messina2's classifiers for both diagnosis and prognosis, showed that Messina2 provides control over classifier margin, and demonstrated on real data that Messina2 represents a principled way to select classifiers with a high likelihood of validating in external cohorts. These capabilities of Messina2 are –

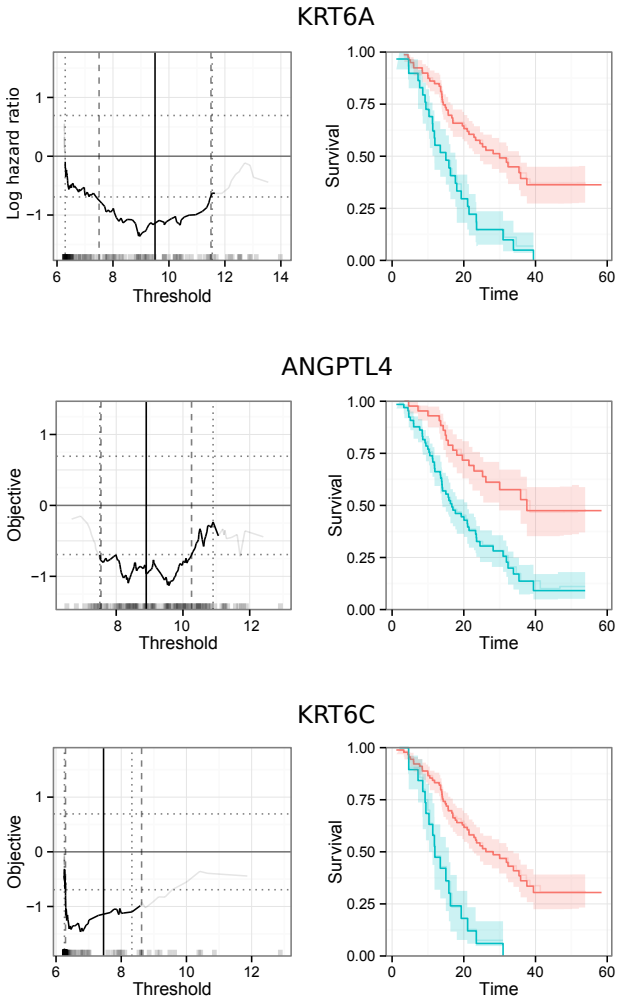


Figure 4.9: High-margin biomarkers of outcome found by Messina2 in the APGI PDAC cohort. Left panels show the log hazard ratio achieved as a function of the threshold; the minimum acceptable hazard ratio of $l_h = 2$ is indicated by horizontal dotted lines, the optimal threshold Δ^* is indicated by a vertical black line, and the distribution of x measurements are represented as a rug on the bottom axis. Right panels display the Kaplan-Meier estimates of the observed survival split at the optimal threshold; red lines are for samples with expression below the threshold, and blue for above. Time is in months after diagnosis, and shaded regions are 90% bootstrap confidence bands from 500 draws.

as far as I am aware – unique, and Messina2 shows promise for the much-needed identification of more effective prognostic biomarkers in PDAC. One of the particular strengths of Messina2 is its great generality, and through user-supplied objective functions Messina can be used to learn effectively any type of binary separation. Messina2's most salient limitation at this time is its restriction to univariate binary classifiers; relaxing these restrictions may improve performance and detection rate, although computational issues may make this infeasible.

Direct simulation and the PDAC validation demonstrated that high-margin Messina2 classifiers were more robust to the inflation in measurement noise that may come with technology translation. In additional work not shown here, classifiers with large margins were also found to be more resistant to the differences in measurement slope and intercept that accompany poor calibration. The low error rate and high robustness of large-margin classifiers is intuitively reasonable, and has strong theoretical support [104], making methods capable of controlling classifier margin important tools in designing robust biomarker-based clinical tests.

In simulation experiments, Messina2 f_M provided highly accurate control of the classifier margin, allowing the analyst to confidently train large-margin diagnostic classifiers with a high likelihood of validating in external data. This capability is unique to Messina2: competing univariate test-based methods of biomarker selection do not control for margin size, and in a classification context are likely to have an unacceptably high non-validation rate (Figure 4.6). In contrast to the f_M case, Messina2 f_H provided poor control over classifier margin, although it was generally competitive with alternate approaches (Figure 4.7). The relatively poor selectivity of Messina2 f_H may simply be a consequence of the greater uncertainty encountered in outcome data, but warrants further investigation. In both the diagnosis and prognosis contexts, it can be argued that univariate statistical tests can still be used to achieve a form of margin control, through increasing the stringency of P-value selection. Although this

will indirectly lead to margin improvements, a remaining issues with univariate test selection methods is that they can not integrate asymmetric costs, or more complex selection criteria, while still optimising margin. Asymmetric cost is very commonly encountered in medical decision making [76], and the utility of univariate test methods in clinical test development is severely limited by their inability to perform cost-sensitive selection.

Messina2's use of general objective functions gives it immense flexibility. Two example objectives have been presented, f_M for cost-sensitive classification, and f_H for prognosis, but any objective that can be adapted into the binary separation framework can be optimised by Messina2. This generality represents a distinct advantage of Messina2 over machine-learning techniques, which usually[7] require bespoke adaptation to a new predicted variable domain.

When applied to real data reflecting a translation scenario, Messina2 demonstrated excellent performance in identifying biomarkers that would validate in external cohorts (Figure 4.8). The sensitivity of both Messina2, and Messina2Core, was far superior to that of the competing biomarker selection technique `maxstat`, with Messina2Core performing particularly well. Messina2Core lacks the bootstrap validation step used by Messina2, and the large difference in sensitivity between the two techniques indicates that Messina2's higher stringency comes at the expense of great sensitivity. However, it is notable that for the most specific bottom-left region of the pseudo-ROC curves of Figure 4.8, Messina2 and Messina2Core show similar performance. This region is the one that is of most interest during classifier development, and so the lower sensitivity of Messina2 away from this region is of little practical consequence. The very poor performance of `maxstat` was surprising, with the technique constructing classifiers that validated little better than chance, and was particularly puzzling given the competitive performance of `maxstat` in simulation studies (Figure 4.7). The huge divide in validation performance between the

[7]The only exception of which the author is aware is nearest neighbour techniques, which can be completely agnostic of the domain $\mathbb{Y}$.

Messina2 classifiers, and the `maxstat` method, may well reflect again the divide between statistical significance and classifier performance (Figure 4.1): although `maxstat` is an effective statistical test for differential survival, it is a poor method for the construction of actual prognostic classifiers.

The work of Chapter 3 investigated the use prognostic biomarkers to predict survival with resected PDAC, with underwhelming results. It could be that PDAC is a disease with a fundamentally unpredictable course – although the results of Chapter 2 suggest that this is not the case – or the biomarkers used in Chapter 3 may simply be poor markers of outcome. To identify potential better biomarkers of PDAC prognosis following resection, Messina2 was applied to data from the APGI PDAC cohort. Three high-margin genes were identified (Table 4.1, Figure 4.9) – KRT6A, ANGPTL4, and KRT6C – all with plausible literature connections to outcome with PDAC [3, 23, 58, 67, 74, 101, 111]. Unfortunately, none of the three biomarkers validated on either of the GSE28735, or the TCGA paad cohorts. This disappointing result may be due to technical effects, and a more careful validation of these three markers is certainly warranted, given their strong literature ties to outcome and metastasis. Although, given more complete data, better single-gene biomarkers of prognosis will likely be found in PDAC, the orthogonal metagene result of Chapter 2 suggests that optimal prognosis will require a variant of Messina that can work with joint biomarker distributions.

Messina2's greatest limitation is its restriction to examining biomarkers singly. This was a practical choice made for a number of reasons: it vastly restricts the classifier search space, reduces the quantity of data required for stable fitting, and matches well the clinical desire for classifiers that use as few biomarkers as possible. However, these benefits come at the cost of prediction performance: with a Vapnik-Chervonenkis dimension [104] of 2, Messina2 classifiers are capable of learning only extremely simple data patterns. There is evidence that excellent classification performance can be achieved from measurement of only a few biomarkers [37]; this is still a small enough

number to be attractive as a clinical diagnostic, and could potentially provide large gains in performance over Messina2. However, the issues of computational tractability, and stability in reasonable-sized data sets, remain. Very little work has been done in this direction, and it remains an area for future development.

4.5 Attribution

The conception and planning of this project, and all work, was performed solely by me.

4.6 Methods

Messina2 version

The version of Messina2 used in this chapter is implemented in the `R` environment, and is available as the `thesis` branch in the Bitbucket repository `marpin/r-messina`. This can be directly installed from an instance of `R` that has the `devtools` package installed, by running the command `devtools::install_bitbucket("marpin/r-messina", ref = "thesis")`.

Simulation studies

Simulation experiments were as described in the text. All calculations were performed in the `R` environment.

PDAC data

PDAC GEX and outcome data were from the APGI, TCGA paad, and GSE28735 cohorts, following unsupervised selection, processed as described in Chapter 2.

Pseudo-ROC generation

A pseudo-ROC curve for a given training technique and validation cohort was calculated as follows.

The training technique was applied to the APGI PDAC GEX-outcome data, to yield a list of threshold prognostic classifiers. This list was then sorted in order of decreasing classifier figure of merit. For Messina2 and MessinaSurv, this figure of merit was the margin, Δ^*; for `maxstat`, this was the negative of the classifier P-value. Classifiers based on transcripts which were not present in the validation data were discarded from the list. Each classifier in the list was then tested for prognostic significance in the validation set. This was achieved by first affinely scaling the optimal classifier threshold t^*, which based on the distribution of biomarker measurements in the APGI training data, to approximately match the distribution of measurements in the validation data, as follows:

$$t'^* = (t^* - \mu(x)) \frac{\sigma(x')}{\sigma(x)} + \mu(x')$$

where t'^* is the threshold to be used on validation data; x is the APGI training data, and x' the validation data, for the biomarker under consideration; and μ and σ are the τ-estimators of location and scale [73], as implemented in the `R` package `robustbase`. The scaled t'^* was then used to partition the validation data into two subgroups, defined by $w'_i = [x'_i \geq t'^*]$, and a log-rank test for differences in outcome between the groups $w = 0$ and $w = 1$ was performed. The classifier was considered to have validated if the uncorrected test P-value was less than 0.05.

To plot the pseudo-ROC curve, for each classifier in the sorted list, two quantities were calculated. The first was the true validation rate, which was the fraction of classifiers with figure of merit at least as good as the classifier under consideration, that did validate; the second was the false validation rate, being the fraction of classifiers with figure of merit at least as good as the classifier under consideration, that did not validate. A plot of the true validation rate versus the false validation rate, for all figures of merit, yielded the pseudo-ROC curve.

Pseudo-ROC curves were calculated for three training techniques: Messina2 (f_H with $l_h \geq 2$ and $\Delta^* \geq 1$), Messina2Core (parameters

as for Messina2), and `maxstat` (test `"LogRank"`, correction method `"HL"`); and two validation cohorts: TCGA paad, and GSE28735.

Chapter Five

Conclusion

Pancreatic ductal adenocarcinoma is a devastating cancer that is remarkably refractory to all current therapies, and there is a real need for better understanding and management of this disease. Recent genomic efforts have begun to exhaustively characterise PDAC at the molecular level, but the analysis of the data generated by these projects has only just begun. This dissertation reported on one aspect of this analysis, by exploring the link between gene expression and PDAC outcome, in ways that were previously not possible. It yielded a better understanding of the basic biological processes that reflect, and possibly control, survival in PDAC, and developed tools to assist the management of the disease, with the goal of ultimately improving patient quality of life. Beyond the application to PDAC presented here, the analysis frameworks used in this dissertation might find use in exploring the links between gene expression and prognosis in other diseases.

Chapter 2: Signatures of survival

Chapter 2 addressed a simple question: what fundamental biological processes are linked to differences in the survival of patients with resected PDAC? By identifying genes with prognostic expression, and

then factorising these genes into coordinately-expressed metagenes, I identified two main orthogonal axes of transcription in PDAC samples, that independently were associated with patient survival time. These two axes were closely correlated with previously-reported signatures of cell proliferation and the EMT, strongly implicating these two processes in the differential survival of patients with PDAC. This result makes sense in light of the known biology of resected PDAC, where it is believed that the majority of resected patients have disseminated undetected metastases at the time of resection, which lead to early relapse and death [10]. In this light, the importance of EMT (causing metastasis), and proliferation (general malignancy), appear reasonable, but the absence of other aspects of PDAC biology from the axes was surprising.

The many aspects of cancer have been conceptually grouped into ten so-called hallmarks [40], with many contributing to a given malignancy; in this light, it is surprising that only two hallmarks were associated with outcome in PDAC. In particular, the absence of any strong stromal signatures, or any significant association between tumour stromal content and expression, was surprising given the importance of the stroma to PDAC pathology [69]. It may be that the stroma is in fact involved in PDAC prognosis, but not in *differential* prognosis; the work of Chapter 2 was powered only to detect the latter.

Both transcriptional axes identified in Chapter 2 were prognostic in a number of other solid tumours, often with distinct patterns. For example, both axes were approximately additive on hazard in PDAC, but interacted in kidney renal clear cell carcinoma, in which high expression on *both* the proliferation and EMT axes was required for poor outcome. This hints at both commonality and differences in the importance of these axes to outcome in cancer in general, and provides further evidence that these axes are truly associated with outcome. It also suggests that the methods developed in Chapter 2 could have broad application.

Chapter 2 developed a template for the principled analysis of

transcription patterns linked to differential outcome, in any sample type. Many of the methods used there are new (for example, CPSS), or difficult to apply correctly (NMF), but the success of the chapter's work suggests that derivative analyses following the template developed in Chapter 2 in other diseases will be equally fruitful. Of particular interest would be the application of the methods of Chapter 2 to the pan-cancer datasets that are now reaching maturity; this endeavour could provide a unique window on those metagenes that are survival-associated across many cancers, and perhaps lead to better understanding of the connections between different malignancies.

Better understanding of the processes that drive survival in cancer could point the way to more effective therapy. The field of PDAC in particular is in desperate need of more effective ways to treat the disease, with all the tools of molecular medicine so far having made only a modest influence on overall survival with the cancer. If the survival-associated axes in PDAC are in fact causative of outcome, then methods to modulate these axes may permit the transformation of a poor-prognosis disease into a good-prognosis one. For a cancer as aggressive as PDAC this manipulation is unlikely to affect a cure, but even small gains are positive news for this dire malignancy.

Chapter 3: Predicting outcome

Chapter 3 addressed a pressing problem in the management of PDAC: resection of the primary tumour is a major operation that carries risk and morbidity, and it is offered to all patients with a resectable tumour, yet relatively few patients derive much benefit from the procedure. Most of the patients who undergo pancreas resection for PDAC will have undetected metastatic disease, and after taking into account the decreased quality of life following their operation, may actually have been better served by undergoing no resection at all. The problem is fundamentally one of staging: current staging methods for PDAC cannot reliably identify pre-resection patients with micro metastases from those without, and so all eligible patients are offered the resection procedure.

As a first step in resolving this serious problem with PDAC management, Chapter 3 described the construction of the PCOP, a tool to *preoperatively* predict the expected survival of a PDAC patient if they were to undergo resection. The PCOP used preoperative measurements of two biomarkers of metastasis, to be collected during EUS-FNA performed prior to resection. Although the prognostic accuracy of the PCOP was modest, it was competitive with a gold-standard PDAC prognostic nomogram that requires *postoperative* data [18], and so may still have clinical utility. In particular, the PCOP may be useful as a clinical 'tie-breaker,' assisting the patient and surgeon to decide on a course to take with a resection that is expected to be particularly challenging. In the case of a poor prognosis from PCOP, the decision may be made to not resect, as little benefit is expected to be gained. On the other hand, should the PCOP predict good survival following resection, the surgeon and patient may consider a more risky and challenging resection to be worthwhile. The PCOP developed in Chapter 3 is a first version, and is expected to improve over time; in particular, the identification of better biomarkers of prognosis may drastically improve the PCOP's accuracy.

Chapter 4: Better biomarkers

Many diagnostic and prognostic classifiers don't successfully validate. This is a serious concern when translating a discovery made in one context, such as a research laboratory, to another, such as a molecular pathology test, with much time and resources wasted on a failed translation project. To assist with this problem, Chapter 4 developed and characterised Messina2, a technique for identifying optimally robust classifiers, that can be used for both diagnostic and prognostic separation of patients. Messina2 significantly outperformed competing methods in both diagnosis and prognosis, and was applied to a resected PDAC cohort to identify three mRNAs that strongly separated good- and poor- prognosis patients. Testing of the expression of protein linked to these mRNAs may reveal improved biomarkers of PDAC outcome to those used in Chapter 3.

Final words

The prognosis of patients with PDAC has remained consistently dire for the last 30 years. This situation may soon change: new genomics efforts are collecting unprecedented numbers of PDAC patients, and characterising their disease in great detail, both at the molecular and clinical level. The challenge will soon be to analyse the huge amount of data from these projects, and the many like them for other diseases, to gain insights on disease that translate to actual improved patient outcomes. This dissertation describes some early efforts in this direction, by analysing the link between gene expression and outcome, at a number of levels. Some results are presented that may add usefully to the knowledge of PDAC, but I believe that the greater value of this work is the methods that are developed within, that I hope will find some use in furthering future knowledge of disease and health.

Appendices

Appendix A

Basis matrix W for the six survival-associated metagenes

	MG1	MG2	MG3	MG4	MG5	MG6
A4GALT	0.0295	0.0000	1.2977	0.0788	0.3625	0.5232
A4GNT	0.0000	0.7419	0.0483	0.0539	0.3720	0.0666
ABHD16A	0.6623	0.7249	0.0000	0.0000	0.5217	0.2210
ABHD5	0.1481	0.7473	0.0000	0.7478	0.3988	1.1727
ABLIM1	0.0145	0.9135	0.3159	0.0000	0.6066	0.3419
ACE	0.0333	0.8332	0.0536	0.0000	0.0000	0.1814
ACKR3	0.0029	0.0000	0.3821	0.3591	0.2080	0.5772
ACYP2	0.2481	0.8949	0.0000	0.2334	0.8454	0.4110
ADH1A	0.0730	0.4440	0.0052	0.1009	0.6614	0.0000
ADM	0.0000	0.0000	0.5168	0.5137	0.0000	0.3570
AGRP	0.0000	0.0000	0.0000	0.6786	0.0000	0.1744
AKIP1	0.6365	0.2394	0.6036	0.7118	0.7849	0.7168
AKR1A1	0.2470	1.0849	0.2633	0.2921	0.6588	0.4524
ALDH5A1	0.0988	0.9930	0.5463	0.0566	0.8968	0.2222
ALOX5AP	0.0525	0.0084	0.0147	1.2654	0.3441	0.7138
AMOT	0.0653	0.8246	0.1374	0.5176	0.4311	0.5705
ANGPTL2	0.0000	0.0000	0.3694	0.8726	0.1807	0.9222

ANGPTL4	0.1789	0.0000	0.4156	0.0461	0.0260	0.3906
ANKLE2	0.7503	0.1422	0.6238	0.5082	0.1879	0.3839
ANKRD22	0.4067	1.3536	0.1731	0.2672	0.0381	0.2229
ANKRD37	0.0562	0.1817	0.2150	0.7249	0.0129	0.5715
ANLN	1.1696	0.2368	0.0796	0.0772	0.0000	0.7203
APCDD1	0.0000	0.1375	0.1494	0.1308	0.5957	0.8366
APCS	0.0000	0.0306	0.1569	0.1001	0.1638	0.3521
ARFGAP3	0.0252	0.2988	0.5370	0.8377	0.4872	0.5353
ARHGAP24	0.0628	1.0614	0.0157	0.7487	1.1007	0.6209
ARHGEF19	0.0837	0.0833	1.2033	0.5242	0.4520	0.5071
ARL4C	0.0000	0.0171	0.3025	0.4910	0.2953	1.2264
ARSD	0.1550	1.2389	0.1919	0.0000	0.2154	0.1439
ASPM	1.1736	0.3897	0.2026	0.1743	0.0380	0.0396
ATAD2	0.9358	0.0696	0.1136	0.0265	0.1092	0.3070
ATF7IP2	0.0000	0.2019	0.1165	0.0000	0.0319	0.0000
ATL3	0.6429	0.0252	0.1566	0.4867	0.2467	0.2863
AURKB	1.0027	0.1107	0.1351	0.0000	0.0096	0.0000
AXIN2	0.0000	0.5221	0.4413	0.1313	0.8077	0.2911
B3GALTL	0.3601	0.3276	0.5636	0.3806	0.4898	0.7750
BAMBI	0.1091	0.0034	0.8430	0.3931	0.2428	0.1686
BBS2	0.2474	1.1417	0.0000	0.2202	1.0006	1.1598
BCKDK	0.2186	0.2923	0.8654	1.0655	0.4050	0.1090
BCL11B	0.1982	0.9231	0.2260	0.2401	0.4151	0.0000
BIRC5	1.3802	0.1694	0.3679	0.5452	0.0000	0.2427
BOC	0.0000	0.0000	0.3211	0.0000	1.6086	0.0000
BTN3A1	0.6641	0.7077	0.0729	0.2544	0.9928	0.2964
C1orf56	0.0000	0.8742	0.0000	0.3677	0.1145	0.3590
C1QTNF6	0.0000	0.0000	0.5885	0.6205	0.2234	0.9726
C2orf70	0.1081	1.0889	0.0206	0.0000	0.0000	0.0000
C5orf46	0.0000	0.0000	0.0000	1.0562	0.1278	1.0438
C9orf152	0.2087	1.3686	0.0000	0.3548	0.0206	0.0000
CA8	0.0000	0.6859	0.0502	0.0094	0.0536	0.0000
CACHD1	0.0000	0.6891	0.0153	0.0000	1.0768	0.4880
CADPS2	0.2591	1.2923	0.0000	0.5506	1.0209	0.5729

CAMK1G	0.0940	0.2377	0.0000	0.0316	0.8847	0.0000
CAPN6	0.0000	0.7541	0.0000	0.2282	0.6418	0.0000
CARHSP1	0.7535	0.5316	0.8652	0.8993	0.2633	0.0000
CATSPER1	0.1179	0.0000	0.9199	0.0000	0.0000	0.1046
CAV1	0.4195	0.0000	0.1925	0.0801	0.2714	0.8420
CCDC88A	0.0000	0.1729	0.4668	0.0109	0.8006	1.0201
CCL19	0.0000	0.0000	0.0000	0.0000	0.9529	0.0000
CCNB1	1.4334	0.4638	0.1274	0.2506	0.0155	0.3645
CCR7	0.0569	0.0000	0.0000	0.0000	1.0524	0.0000
CD70	0.0870	0.0000	0.2096	0.3612	0.0000	0.4343
CDA	0.2927	0.0000	0.3408	0.0000	0.0000	0.6991
CDC45	0.9608	0.0779	0.1086	0.3364	0.0336	0.0000
CDK12	0.1906	0.2755	0.0000	0.0788	0.8330	0.0000
CDK2	1.0635	0.2517	0.0111	0.5230	0.3310	0.3338
CEBPB	0.0729	0.0654	1.2909	0.5287	0.5065	0.8131
CEP55	1.4198	0.3340	0.0000	0.1690	0.0000	0.4555
CFDP1	0.3512	0.5466	0.7440	0.6706	0.0000	0.2594
CHAF1B	0.9890	0.2957	0.1997	0.0187	0.5165	0.0960
CHEK1	1.5161	0.1621	0.0000	0.0034	0.1080	0.2731
CHN2	0.0000	0.4963	0.0000	0.3389	0.4366	0.0000
CIDEC	0.0279	0.0000	0.4258	0.2777	0.0038	0.0000
CIDECP	0.1140	0.0232	0.5161	0.2795	0.1093	0.0000
CKAP2L	1.7829	0.2230	0.2724	0.0319	0.0000	0.0884
CLEC3B	0.0589	0.0691	0.1151	0.0110	0.8063	0.0000
CNIH3	0.0000	0.0591	0.0000	0.3178	0.0000	0.6014
CNNM1	0.0000	0.8666	0.4109	0.0000	0.0897	0.0000
COL12A1	0.0000	0.1328	0.0340	0.5329	0.1874	1.6461
COL5A3	0.0000	0.0000	0.1816	0.0351	0.0660	1.0286
COL7A1	0.0000	0.0000	0.5858	0.0000	0.0000	0.5878
COLGALT1	0.3987	0.1554	0.6227	0.4286	0.1646	0.8792
COLGALT2	0.0000	0.6011	0.0000	0.0199	0.0000	0.0000
COX4I2	0.0000	0.1744	0.0740	0.0000	0.9855	0.3346
CSNK1D	0.2122	0.3756	1.5627	0.4799	0.1570	0.2284
CST6	0.0651	0.0000	0.2022	0.0000	0.0690	0.6328

CTSL	0.3897	0.0000	0.1976	1.1757	0.4702	0.2240
CTSV	0.3015	0.0439	0.2623	0.0203	0.0194	0.1819
CYP2S1	0.3223	1.0232	0.1543	0.0000	0.0927	0.0000
DCAF8	0.0000	1.1369	0.4818	0.1094	0.5277	0.1875
DCBLD2	0.4024	0.0000	0.1236	0.0000	0.1426	0.8437
DCUN1D5	1.3599	0.0751	0.0000	0.8575	0.9561	0.7193
DENND1A	0.8191	0.0000	0.2458	0.1898	0.0000	0.1782
DERA	1.1839	0.1952	0.4571	0.6042	0.2890	0.3195
DHRS9	0.0000	0.0000	0.9957	0.3426	0.0000	0.1699
DKK1	0.4779	0.0000	0.2976	0.1847	0.0000	0.0242
DNAJC9	0.7779	0.1108	0.3734	0.1159	0.1329	0.1528
DPY19L1	0.3414	0.3625	0.2993	0.5360	0.0781	0.5087
DSG2	0.4320	0.5696	0.1794	0.5147	0.0387	0.7066
DSG3	0.1766	0.0000	0.2140	0.0000	0.0000	0.5384
DYNC2H1	0.0000	1.6131	0.1497	0.0000	0.7591	0.6693
E2F7	1.0366	0.0000	0.0315	0.0222	0.0000	0.5360
EDIL3	0.0000	0.0000	0.0000	0.8576	0.0121	0.8163
EIF2AK3	0.1806	1.2690	0.0000	0.3842	0.6143	0.3321
ELMOD3	0.0000	1.1608	0.6902	0.3859	0.5348	0.0874
EMP3	0.2499	0.0000	0.4619	0.1582	0.2170	0.5646
ENO2	0.3608	0.3375	0.7898	0.0339	0.0000	0.9442
EPHX2	0.0000	0.5912	0.1080	0.1660	0.6761	0.0000
ERRFI1	0.1599	0.0301	0.5475	0.3478	0.2866	0.7895
EXOSC8	0.9336	0.6010	0.2789	1.0216	0.3682	0.1481
EYA3	0.0000	0.0869	0.5323	0.0000	0.0000	0.9120
FAH	0.6763	0.4158	0.3555	0.2131	0.3240	0.3914
FAM120AOS	0.1803	1.0488	0.0000	0.2845	0.7143	0.5698
FAM134B	0.0000	0.8232	0.0000	0.2342	0.2083	0.0000
FAM189A2	0.0000	1.0020	0.0000	0.0213	0.1143	0.0000
FAM83A	0.2461	0.0000	0.1165	0.0000	0.0000	0.2211
FAM91A1	0.9811	0.1968	0.1603	0.7865	0.0000	0.2703
FBXO22	0.5017	0.3643	0.0000	0.5761	0.0000	0.3137
FBXW8	0.2492	0.2604	0.6553	0.9331	0.1844	0.3307
FEM1B	0.3031	0.3008	0.0000	0.0017	0.0838	1.4170

FER	0.4975	0.1005	0.1802	0.4440	0.1792	0.8664
FGB	0.0000	0.0000	0.0170	0.3212	0.0000	0.0818
FGD6	0.5544	0.0000	0.1308	0.1418	0.0000	0.4991
FGG	0.0548	0.0379	0.0000	0.1372	0.0068	0.2157
FHDC1	0.1771	1.2361	0.2174	0.0189	0.0000	0.0512
FLRT3	0.7913	0.1342	0.5121	0.2846	0.2220	0.3125
FRZB	0.0889	0.2374	0.0000	0.5404	1.4969	0.0017
FSCN1	0.3709	0.0737	1.0622	0.1342	0.1423	0.7358
FST	0.0000	0.0000	0.1578	0.0000	0.0414	0.4947
FYN	0.0127	0.5194	0.1203	0.1287	1.6862	0.8654
GAB2	0.0435	0.7351	0.3850	0.6361	1.3628	0.2664
GABPB1	0.7363	0.1963	0.0000	0.7422	0.2159	0.6724
GAPDH	0.4758	0.3945	0.8305	0.2369	0.0000	0.7231
GATA6	0.0534	0.8827	0.0860	0.1396	0.1932	0.0000
GATC	1.0220	0.1104	0.0000	0.4818	0.0723	0.4716
GIMAP2	0.1486	0.7215	0.0000	0.6567	0.7701	0.0000
GINS2	1.0803	0.1777	0.3933	0.0729	0.0000	0.0000
GNPAT	0.1710	0.9518	0.1369	0.4352	0.1758	0.1925
GOLM1	0.0000	0.7145	0.1203	0.0488	0.0000	0.0000
GPC3	0.0980	0.2322	0.0000	0.0000	1.2713	0.0000
GPR176	0.4324	0.3072	0.0000	0.7415	0.3745	0.5882
HIPK2	0.2587	1.2502	0.0694	0.2371	0.5213	0.0000
HJURP	1.3269	0.2436	0.2326	0.0210	0.0000	0.0000
HRASLS2	0.3273	0.0000	0.3045	0.2167	0.0000	0.0000
HSP90B1	0.5274	0.4642	0.7758	0.8972	0.2977	0.3795
HSPB6	0.0000	0.1493	0.1298	0.0000	1.3081	0.3131
ICAM2	0.5013	0.1959	0.4755	0.3105	0.4043	0.1342
IDH2	0.7131	0.4322	0.3970	0.2145	0.3314	0.2342
IFT140	0.0000	1.0890	0.5193	0.0000	0.2592	0.0662
IGFBP1	0.2708	0.0000	0.2323	0.0327	0.0000	0.0058
IGLL3P	0.1660	0.1496	0.0000	0.0000	0.7633	0.0000
IKBIP	0.2893	0.0000	0.3028	1.1219	0.1455	0.4694
IL1R2	0.0377	0.2543	0.4285	0.2301	0.0000	0.0605
IL20RB	0.2578	0.0000	0.3094	0.0000	0.0000	0.6805

IL33	0.2369	0.0436	0.0000	0.1304	0.6759	0.0000
ITGA5	0.0000	0.0000	0.4758	0.2666	0.1206	0.6815
ITPKB	0.0000	0.8315	0.6059	0.0000	1.1923	0.6724
KANK4	0.0000	0.0000	0.1981	0.4683	0.0000	1.2292
KCNQ3	0.0000	0.1296	0.1721	0.7768	0.0916	0.5160
KCTD10	0.3776	0.1324	0.2867	0.4387	0.5081	0.7943
KCTD5	0.3848	0.5133	1.1253	0.6056	0.0000	0.0000
KIAA0513	0.0828	1.0351	0.1715	0.3220	0.5910	0.0000
KIAA1549L	0.3755	0.0812	0.2646	0.6647	0.1501	0.6423
KIF14	1.1244	0.3648	0.1952	0.4293	0.0000	0.1264
KIF20A	1.3726	0.2864	0.2082	0.2320	0.0000	0.2888
KIF2C	0.7952	0.1329	0.1096	0.0074	0.0000	0.0000
KLHL5	0.4215	0.1645	0.0000	0.3538	0.6955	1.1410
KNTC1	1.0718	0.1383	0.4419	0.0827	0.1499	0.2787
KRT17	0.2860	0.0000	0.3863	0.1586	0.1201	0.5074
KRT6A	0.1386	0.0000	0.1202	0.0000	0.0000	0.4668
KRT6C	0.1187	0.0000	0.0000	0.0000	0.0000	0.1640
KRT7	0.4597	0.0020	0.5620	0.0000	0.1354	0.4370
KYNU	0.6104	0.0894	0.0693	0.5431	0.0000	0.2790
LAMA5	0.3670	0.0772	1.0234	0.0000	0.3418	0.1832
LCNL1	0.1072	0.2829	0.0115	0.2669	0.5289	0.0000
LDHA	0.6526	0.4664	0.0000	0.3186	0.0504	1.1696
LETM2	0.4402	0.0000	0.3924	0.0000	0.0000	0.2831
LGALS9B	0.1106	1.0239	0.0000	0.0000	0.3463	0.4913
LINC01184	0.6331	0.8045	0.0000	0.3418	0.8076	0.0000
LMO3	0.0000	0.1062	0.0000	0.0090	1.1796	0.0136
LMTK2	0.7364	0.3642	0.3100	0.5254	0.0204	0.2425
LOC100506562	0.5772	0.2935	0.6002	0.6045	0.1075	0.1108
LOX	0.2078	0.0000	0.0806	0.3896	0.0866	0.9212
LYNX1	0.0337	0.0000	0.2575	0.1651	0.0000	0.0951
MAP3K8	0.1984	0.0000	0.0681	0.3075	0.5588	0.4348
MARCKSL1	0.1504	1.3374	0.2978	0.0000	0.0000	0.2627
MARS2	0.7481	1.0181	0.0000	0.4007	0.4981	0.0000
MC1R	0.1042	0.1313	1.0794	0.8656	0.4740	0.1335

MCEMP1	0.0000	0.0000	0.0000	0.6056	0.0000	0.2992
MCM10	1.1446	0.1414	0.0000	0.0141	0.0000	0.0808
MCM4	1.2790	0.1411	0.3090	0.0254	0.0103	0.1276
MCOLN2	0.1988	0.2778	0.0000	0.0000	0.9442	0.0000
MELK	1.0177	0.2864	0.0000	0.2322	0.0133	0.2208
MEOX1	0.0000	0.0536	0.1642	0.0438	0.9639	0.0000
MIF	0.4348	0.3316	0.9576	0.4402	0.0008	0.6845
MIR99AHG	0.0371	0.2791	0.3859	0.4466	1.7947	0.2232
MME	0.0009	0.0000	0.0640	0.4532	0.0419	0.5791
MRAP2	0.0430	0.7825	0.0000	0.2177	0.2314	0.0000
MRPL24	0.1643	1.1324	0.2156	0.1207	0.2213	0.1778
MTRNR2L1	0.2795	0.5589	0.4897	0.0719	0.5523	0.0000
NACC2	0.5312	0.0000	0.7176	0.2474	0.0000	0.1055
NAMPT	0.3355	0.0000	0.0493	0.7543	0.3154	0.3500
NCAPD2	1.3843	0.4110	0.1605	0.1233	0.2041	0.3231
NCAPG	1.6056	0.4449	0.0000	0.0000	0.0000	0.5243
NELFE	0.9382	0.2255	0.5894	0.8561	0.3602	0.0798
NEURL2	0.6888	0.1217	0.0000	0.2556	0.7216	0.4336
NFIA	0.1194	0.8389	0.0000	0.3854	1.5045	0.2708
NFIX	0.0000	0.8819	0.1383	0.0000	1.3919	0.7968
NMB	0.2126	0.1909	0.6634	0.7944	0.0000	0.3640
NPM1	0.0000	1.0465	0.0000	0.0029	0.0826	0.0446
NR0B2	0.0000	0.8362	0.0000	0.0000	0.1422	0.0000
NRP2	0.1462	0.0000	0.4996	0.0000	0.0000	0.0534
NUP155	1.1296	0.4140	0.0620	0.3285	0.2288	0.4554
OAZ1	0.8583	0.5931	0.6573	1.1219	0.5151	0.5871
ORC1	0.9777	0.3231	0.1638	0.9547	0.1157	0.0101
P2RY2	0.1789	0.0331	0.7738	0.2163	0.0000	0.5005
P2RY8	0.2334	0.0728	0.0000	0.2788	1.6555	0.0000
P4HA1	0.0430	0.1009	0.4121	0.8384	0.0000	0.5460
P4HA2	0.3225	0.1659	0.1245	0.5449	0.1088	0.7371
PAX8	0.7680	0.0000	0.5631	0.0000	0.0000	0.0000
PAX8-AS1	0.5656	0.0447	0.3435	0.0750	0.0071	0.0000
PBXIP1	0.0000	0.5144	0.4130	0.0000	0.4392	0.1667

PCDH20	0.0000	0.4318	0.0000	0.1465	0.0000	0.0000
PCF11	0.2613	0.9351	0.2527	0.0950	1.1086	0.4077
PCOLCE2	0.0000	0.0076	0.1188	0.5379	0.0000	0.0542
PDLIM7	0.1954	0.0000	0.4086	0.3731	0.1144	0.6779
PEX11B	0.1066	1.3518	0.0000	0.5264	0.2883	0.2455
PFKFB4	0.5485	0.2199	0.6769	0.4272	0.1428	0.2854
PGAM5	0.9213	0.0000	0.3859	0.4866	0.0000	0.0000
PGBD3	0.6174	0.3626	0.4335	0.2008	0.5630	0.7384
PHACTR3	0.1489	0.0000	0.3225	0.1416	0.0026	0.0728
PHLDA1	0.0838	0.1387	0.7170	0.1250	0.6249	1.5017
PHOSPHO2	0.3445	1.0681	0.0000	0.4652	0.4054	0.0514
PIGL	1.0637	0.1481	0.5587	0.3049	0.2423	0.0000
PLAC9	0.0707	0.0000	0.0000	0.1090	1.2901	0.0766
PLAU	0.2139	0.0000	0.2764	0.0000	0.0249	0.8793
PLEKHS1	0.0000	0.6411	0.3407	0.0862	0.2791	0.0176
PLIN2	0.3057	0.0000	0.0818	1.0167	0.4683	0.2095
PLIN3	0.3365	0.2607	0.9673	0.9320	0.1395	0.4103
PLOD1	0.0595	0.0000	1.2074	0.7504	0.3668	0.8026
PLOD2	0.1489	0.0922	0.2366	0.2919	0.1729	0.8899
POC1A	1.3753	0.3309	0.3179	0.4709	0.0000	0.0000
POLA2	0.8413	0.2234	0.3296	0.1331	0.2137	0.0000
POP5	0.5635	0.5070	1.5160	0.2263	0.1092	0.1799
POU2AF1	0.0611	0.4732	0.0000	0.0007	0.9240	0.0000
PP7080	0.1047	0.9680	0.0000	0.0371	0.0000	0.0000
PPAPDC1A	0.0000	0.0000	0.0000	0.7582	0.0000	1.2230
PPM1H	0.0000	0.8512	0.4600	0.2700	0.2363	0.0000
PPP1R12B	0.1652	0.3193	0.7825	0.6308	0.0253	0.4910
PPP1R14B	0.3673	0.2586	0.7846	0.0000	0.3651	0.5928
PPP1R3C	0.0000	0.0160	0.1325	0.3710	0.0256	0.2554
PPY	0.0000	0.4957	0.0000	0.0805	1.0771	0.0000
PRC1	0.9560	0.3521	0.0407	0.0375	0.0000	0.3200
PRDM16	0.0000	1.1224	0.0000	0.0000	0.5289	0.0867
PREP	0.0587	0.9830	0.3047	0.1977	0.0203	0.0000
PRKCDBP	0.2571	0.0000	1.0161	0.5090	0.2613	0.5936

PRMT7	0.1393	1.5003	0.4373	0.0000	0.1793	0.2230
PROSER2	0.9335	0.1760	0.4026	0.3736	0.2680	0.3965
PRR11	0.8207	0.0503	0.2272	0.0000	0.0000	0.0934
PTGES	0.5703	0.0160	0.5702	0.0681	0.0000	0.5634
PTPN21	0.2722	0.1714	0.3219	0.4864	0.2674	0.8423
PXDN	0.0000	0.0000	0.3795	0.5917	0.3108	1.1884
PYGL	0.0808	0.0000	0.3079	0.3384	0.1413	0.7445
RAB31	0.1110	0.0000	0.2586	0.8745	0.7552	1.1882
RACGAP1	1.3720	0.3729	0.1382	0.1936	0.0734	0.3348
RALGAPB	0.9974	0.5032	0.2879	0.7587	0.2585	0.7977
RAP1GAP	0.0000	1.0067	0.4657	0.2773	0.7542	0.0000
RASL11B	0.0000	0.1852	0.0682	0.2236	1.2121	0.3095
RAVER2	0.1985	0.9070	0.0534	0.0890	0.2667	0.0577
RBMS2	0.6118	0.1541	0.0000	0.4022	0.3184	0.8946
RERE	0.0485	0.7372	0.6212	0.0026	0.9874	0.4207
RERGL	0.2378	0.0000	0.0000	0.1054	1.1842	0.0000
RFC5	1.0809	0.2444	0.0000	0.5248	0.1556	0.3147
RFK	0.0000	0.6594	0.1169	0.0000	0.4342	0.2100
RFX2	0.0000	0.2219	0.2372	0.0000	0.4551	0.2959
RGS3	0.2370	0.1243	0.0000	0.8096	0.2269	0.3212
RGS5	0.0000	0.4317	0.0455	0.0788	0.5794	0.0934
RHOF	0.7466	0.1749	0.4760	0.1428	0.0000	0.5878
RMND5A	0.2696	0.1188	0.2601	0.7065	0.0000	0.0750
RNF103	0.0344	1.2504	0.1672	0.5545	0.2894	0.0635
RPA2	0.4727	0.6964	0.7005	0.4129	1.4239	0.2443
RPIA	0.4609	1.3515	0.2200	0.1918	0.4584	0.0000
SAMD5	0.1340	0.5397	0.0000	0.0000	0.0860	0.0000
SCGB2A1	0.0000	0.8288	0.0000	0.1826	0.1547	0.0000
SCYL2	0.7048	0.3901	0.0000	0.9782	0.4060	0.9614
SDIM1	0.0000	0.0455	0.2422	0.0000	0.5017	0.0000
SEC23IP	0.3380	1.2955	0.0000	0.5310	0.3578	0.4605
SELENBP1	0.0000	1.2032	0.3621	0.2011	0.2603	0.0000
SEPW1	0.0349	0.9518	1.2360	0.0000	0.6293	0.5568
SERPINB3	0.0000	0.0000	0.1755	0.1787	0.0000	0.0506

SERPINH1	0.0000	0.0115	0.3898	0.2169	0.4300	1.0203
SERTAD2	0.2931	0.1441	0.8991	0.9858	0.4859	0.4437
SGSM1	0.0000	0.9290	0.0817	0.0211	0.8410	0.0000
SH3GL1	0.1173	0.1075	1.0090	1.2494	0.2155	0.0000
SLAMF9	0.0435	0.0000	0.0000	0.6663	0.0000	0.0657
SLC12A2	0.0380	0.9089	0.3449	0.0968	0.4855	0.1821
SLC15A1	0.0000	0.0000	0.4779	0.0000	0.0569	0.0565
SLC16A3	0.1282	0.3828	1.1047	0.4222	0.0000	0.9957
SLC2A1	0.1786	0.1209	0.9980	0.4099	0.0000	0.7045
SLC2A3	0.0000	0.0000	0.3369	0.7592	0.3268	0.7204
SLC30A3	0.4502	0.5017	0.0822	0.2136	0.6568	0.0654
SLC40A1	0.0000	0.8927	0.0000	0.5789	0.2440	0.1550
SMOX	0.3692	0.2900	1.4313	0.9987	0.1840	0.0000
SNORA11D	0.0849	0.2729	0.4795	0.4375	0.0039	0.2687
SNRPB	0.9900	0.0786	0.4143	0.9037	0.0238	0.0000
SOBP	0.0000	0.1979	0.8103	0.1044	1.3581	0.0039
SOD2	0.5780	0.1207	0.0000	0.4656	0.4023	0.1652
SPHK1	0.2590	0.0000	0.2748	0.0907	0.6221	1.4095
SPIN4	0.8495	0.3236	0.7960	0.3855	0.2224	0.3985
SPOCD1	0.0000	0.0000	0.1782	0.2094	0.0000	0.7594
SPOCK1	0.1196	0.0000	0.0293	0.5189	0.3390	1.2727
SPP1	0.0294	0.0805	0.0000	1.0413	0.3073	0.7357
ST3GAL2	0.3414	0.0000	0.8015	1.0746	0.4432	0.0000
ST6GAL1	0.1717	0.8423	0.0000	0.2289	0.6651	0.0916
ST6GALNAC1	0.0396	0.9957	0.0803	0.1154	0.0000	0.1050
STAT5B	0.0000	0.9053	0.3202	0.0618	1.3050	0.2213
STK39	0.1526	0.9966	0.2351	0.1373	0.0838	0.1226
SUGCT	0.0000	0.0321	0.0000	0.6297	0.1256	0.9331
SULF2	0.1725	0.1513	0.4552	0.1878	0.3858	0.7665
SYNE2	0.0000	0.8824	0.2432	0.0000	0.2767	0.2763
TAF5L	0.2232	1.0626	0.1753	0.2440	0.2327	0.2249
TARBP2	0.6779	0.3829	1.2178	0.6116	0.1843	0.0000
TCEA3	0.0000	0.8898	0.2645	0.0922	0.6204	0.0000
TCTA	0.0000	0.7508	0.8167	0.0875	0.9836	0.0178

TGFBI	0.1874	0.0000	0.1522	0.1879	0.0548	0.9986
THSD7B	0.0859	0.2031	0.0000	0.2900	0.9574	0.1114
TLE4	0.0509	0.8787	0.0746	0.3315	0.8984	0.4660
TM9SF3	0.0000	1.0785	0.2190	0.0000	0.1641	0.2114
TMED1	0.2561	0.3378	1.1457	0.8311	0.4929	0.2755
TMEM26	0.0407	0.0237	0.1028	0.4886	0.2223	1.4490
TMTC4	0.0000	1.2865	0.3348	0.2090	0.1995	0.2756
TNFRSF10D	0.1474	0.1117	0.6603	0.4579	0.0000	0.1751
TNFRSF17	0.0258	0.0455	0.0000	0.0803	0.5772	0.0000
TNFRSF6B	0.6268	0.0000	0.0684	0.1841	0.0000	0.3940
TOM1	0.0000	0.1032	1.4892	0.8140	0.6813	0.5236
TOM1L2	0.1892	0.0000	0.6276	0.3305	0.0489	0.2346
TOR2A	0.0000	0.9859	0.4755	0.2012	0.5273	0.0000
TPD52L2	0.6311	0.1617	1.3107	0.6501	0.4351	0.2322
TPX2	1.3192	0.1540	0.0351	0.1488	0.0392	0.1087
TRAPPC2	0.5080	1.0792	0.0000	0.4917	0.6155	0.1418
TREM1	0.0472	0.0000	0.0870	0.7055	0.0000	0.3006
TRERF1	0.4920	0.2861	0.3810	0.1345	0.0517	0.1346
TRIM2	0.1310	1.1544	0.3127	0.3092	0.3595	0.0000
TSTD1	0.1685	1.2229	0.4834	0.0685	0.4502	0.0191
TUBA1C	1.3100	0.5454	0.5360	0.5305	0.2711	0.5032
TWIST1	0.0000	0.0000	0.1970	0.9070	0.1202	1.2015
UFC1	0.0000	1.1861	0.2466	0.4651	0.2997	0.0000
UHRF2	0.1520	0.2931	0.3251	0.4968	0.6565	1.1025
UPP1	0.5505	0.0000	0.7864	0.4294	0.1567	0.1100
USP30	0.5449	0.1353	0.3862	0.0000	0.0771	0.0000
VPS35	0.3941	1.3902	0.0000	0.5311	0.0000	0.2457
VSTM2L	0.3176	0.0000	0.9398	0.0000	0.0509	0.0656
WNT2B	0.0885	0.1107	0.0000	0.0139	0.4530	0.0000
XXYLT1	0.2408	0.0000	1.0488	1.0782	0.4595	0.8654
ZBED2	0.1569	0.0000	0.1800	0.0000	0.0000	0.6435
ZFPM1	0.0000	1.2172	0.2917	0.0000	0.4340	0.1504
ZNF185	0.2542	0.1747	1.0210	0.4834	0.0000	0.7221
ZNF565	0.0701	0.2851	0.0717	0.0569	0.2393	0.0768

ZNF658	0.0000	0.8769	0.0000	0.0000	0.9099	0.2753
ZPLD1	0.0000	0.0000	0.1873	0.0325	0.0294	0.1074
ZSCAN16	0.3012	1.4502	0.0000	0.0175	0.5146	0.5090
ZSCAN32	0.3467	1.1558	0.4982	0.3027	0.7286	0.2378

Appendix B

MSigDB signatures correlated with axis A1

Table B.1: MSigDB signatures substantially correlated with activity of the prognostic axis A1.

MSigDB set	A1 correlation
c5.M_PHASE / c5.MITOSIS / c5.M_PHASE_OF_MITOTIC_CELL_CYCLE	0.689
c5.REGULATION_OF_MITOSIS	0.682
c4.GNF2_RFC3 / c4.GNF2_RFC4 / c4.GNF2_SMC2L1 / c4.GNF2_CKS1B / c4.GNF2_CKS2 / c4.GNF2_TTK	0.664
c5.CELL_CYCLE_PROCESS / c5.MITOTIC_CELL_CYCLE / c5.CELL_CYCLE_PHASE	0.653
c5.SPINDLE	0.644
c4.MORF_BUB1B	0.631
c6.CSR_LATE_UP.V1_SIGNED	0.630
c5.SPINDLE_POLE	0.628
c2.PID_PLK1_PATHWAY	0.626
c5.ORGANELLE_PART / c5.INTRACELLULAR_ORGANELLE_PART	0.624
c2.REACTOME_CELL_CYCLE /	

Continued on next page

Table B.1 – continued from previous page

MSigDB set	A1 correlation
c2.REACTOME_CELL_CYCLE_MITOTIC	0.622
c2.REACTOME_CYCLIN_A_B1_ASSOCIATED_ EVENTS_DURING_G2_M_TRANSITION	0.604
c2.REACTOME_MITOTIC_PROMETAPHASE	0.596
c2.KEGG_CELL_CYCLE	0.588
c5.CHROMOSOME_SEGREGATION	0.588
c4.MORF_FEN1	0.586
c2.REACTOME_G1_S_SPECIFIC_TRANSCRIPTION	0.585
c2.REACTOME_ACTIVATION_OF_THE_ PRE_REPLICATIVE_COMPLEX / c2.REACTOME_ACTIVATION_OF_ATR_ IN_RESPONSE_TO_REPLICATION_STRESS / c2.REACTOME_G2_M_CHECKPOINTS	0.583
c2.REACTOME_E2F_ENABLED_INHIBITION_OF_ PRE_REPLICATION_COMPLEX_FORMATION	0.581
c2.REACTOME_E2F_MEDIATED_REGULATION_OF_ DNA_REPLICATION	0.577
c5.CELL_CYCLE_GO_0007049	0.576
c2.REACTOME_KINESINS	0.575
c3.V$ELK1_02	0.574
c5.SPINDLE_MICROTUBULE	0.573
c5.MITOTIC_CELL_CYCLE_CHECKPOINT	0.569
c2.REACTOME_CELL_CYCLE_CHECKPOINTS / c2.REACTOME_G1_S_TRANSITION / c2.REACTOME_SYNTHESIS_OF_DNA / c2.REACTOME_MITOTIC_G1_G1_S_PHASES / c2.REACTOME_MITOTIC_M_M_G1_PHASES / c2.REACTOME_DNA_REPLICATION / c2.REACTOME_S_PHASE	0.566
c4.MORF_ESPL1	0.566
c4.MORF_BUB1	0.565

Continued on next page

Table B.1 – continued from previous page

MSigDB set	A1 correlation
c4.MORF_BUB3/c4.MORF_RAD23A	0.563
c5.CONDENSED_CHROMOSOME	0.562
c4.MORF_RFC4/c4.MORF_RRM1	0.561
c2.BIOCARTA_G2_PATHWAY	0.559
c3.SCGGAAGY_V$ELK1_02	0.558
c2.PID_AURORA_A_PATHWAY	0.556
c5.MITOTIC_SISTER_CHROMATID_SEGREGATION / c5.SISTER_CHROMATID_SEGREGATION	0.555
c4.MORF_UNG	0.554
c2.PID_FOXM1PATHWAY	0.551
c4.MORF_GSPT1	0.550
c2.REACTOME_METABOLISM_OF_NUCLEOTIDES	0.550
c2.PID_ATR_PATHWAY	0.547
c2.BIOCARTA_MCM_PATHWAY	0.546
c4.MORF_CCNF	0.544
c5.CELL_CYCLE_CHECKPOINT_GO_0000075	0.543
c5.MITOTIC_SPINDLE_ORGANIZATION_AND_BIOGENESIS / c5.SPINDLE_ORGANIZATION_AND_BIOGENESIS	0.542
c4.MORF_EI24	0.538
c5.DOUBLE_STRAND_BREAK_REPAIR	0.537
c4.GNF2_PA2G4/c4.GNF2_RAN	0.531
c2.REACTOME_G2_M_DNA_DAMAGE_CHECKPOINT	0.531
c2.KEGG_PYRIMIDINE_METABOLISM	0.531
c4.MORF_GMPS	0.528
c4.MORF_PRKDC	0.528
c2.PID_BARD1PATHWAY	0.528
c4.GNF2_MCM5	0.525
c4.MORF_DNMT1	0.524
c2.REACTOME_POL_SWITCHING	0.523
c4.GNF2_MSH2	0.521

Continued on next page

Table B.1 – continued from previous page

MSigDB set	A1 correlation
c4.MORF_CSNK2B	0.520
c2.PID_AURORA_B_PATHWAY	0.520
c2.REACTOME_DESTABILIZATION_OF_MRNA_BY_KSRP	0.517
c5.DNA_METABOLIC_PROCESS	0.517
c4.MORF_PTPN11	0.516
c5.REGULATION_OF_MITOTIC_CELL_CYCLE	0.516
c5.RESPONSE_TO_ENDOGENOUS_STIMULUS / c5.RESPONSE_TO_DNA_DAMAGE_STIMULUS	0.515
c5.CHROMOSOMEPERICENTRIC_REGION / c5.KINETOCHORE	0.514
c6.MTOR_UP.V1_SIGNED	0.512
c2.REACTOME_APOPTOSIS	0.510
c4.MORF_PPP1CC	0.509
c5.PORE_COMPLEX/c5.NUCLEAR_PORE	0.508
c5.DNA_REPAIR	0.506
c2.REACTOME_CHROMOSOME_MAINTENANCE / c2.REACTOME_TELOMERE_MAINTENANCE	0.506
c5.MACROMOLECULAR_COMPLEX / c5.PROTEIN_COMPLEX	0.506
c4.MORF_XRCC5/c4.MORF_GNB1	0.504
c5.INTERPHASE / c5.INTERPHASE_OF_MITOTIC_CELL_CYCLE	0.503
c5.NON_MEMBRANE_BOUND_ORGANELLE / c5.INTRACELLULAR_NON_MEMBRANE_BOUND_ORGANELLE	0.503
c6.GCNP_SHH_UP_EARLY.V1_SIGNED	0.503
c2.BIOCARTA_RANMS_PATHWAY	0.502
c2.KEGG_DNA_REPLICATION / c2.REACTOME_DNA_STRAND_ELONGATION	0.502
c4.MORF_SOD1	0.502
c5.NUCLEAR_MEMBRANE /	

Continued on next page

Table B.1 – continued from previous page

MSigDB set	A1 correlation
c5.NUCLEAR_MEMBRANE_PART	0.501
c4.MORF_HDAC1	0.501
c2.REACTOME_HIV_LIFE_CYCLE / c2.REACTOME_LATE_PHASE_OF_HIV_LIFE_CYCLE	0.500
c5.CHROMOSOMAL_PART/c5.CHROMOSOME	0.500
c5.PHOSPHORIC_DIESTER_HYDROLASE_ACTIVITY	−0.502
c3.CTGCAGY_UNKNOWN	−0.505
c3.V$OCT1_01	−0.509
c3.V$GATA_Q6	−0.515
c5.CELL_SURFACE_RECEPTOR_LINKED_SIGNAL_TRANSDUCTION_GO_0007166	−0.518
c4.GNF2_MAPT	−0.526
c3.V$OCT1_04	−0.531
c2.REACTOME_G_ALPHA_S_SIGNALLING_EVENTS	−0.539
c3.V$OCT_C	−0.544

Appendix C

MSigDB signatures correlated with axis A2

Table C.1: MSigDB signatures substantially correlated with activity of the prognostic axis A2.

MSigDB set	A2 correlation
c2.PID_INTEGRIN1_PATHWAY	0.654
c2.PID_INTEGRIN3_PATHWAY	0.637
c2.PID_UPA_UPAR_PATHWAY	0.597
c4.GNF2_PTX3	0.593
c2.KEGG_ECM_RECEPTOR_INTERACTION	0.582
c2.PID_INTEGRIN5_PATHWAY	0.577
c4.GNF2_MMP1	0.575
c2.REACTOME_EXTRACELLULAR_MATRIX_ ORGANIZATION / c2.REACTOME_COLLAGEN_FORMATION	0.572
c5.AXON_GUIDANCE	0.571
c2.KEGG_FOCAL_ADHESION	0.567
c2.PID_SYNDECAN_1_PATHWAY	0.552
c2.REACTOME_CELL_EXTRACELLULAR_ MATRIX_INTERACTIONS	0.538
c2.PID_INTEGRIN_CS_PATHWAY	0.536

Continued on next page

Table C.1 – continued from previous page

MSigDB set	A2 correlation
c5.TISSUE_DEVELOPMENT	0.536
c5.COLLAGEN	0.531
c6.CORDENONSI_YAP_CONSERVED_SIGNATURE	0.526
c6.LEF1_UP.V1_SIGNED	0.519
c2.REACTOME_INTEGRIN_CELL_SURFACE_ INTERACTIONS	0.518
c5.AXONOGENESIS / c5.CELLULAR_MORPHOGENESIS_ DURING_DIFFERENTIATION	0.515
c6.STK33_NOMO_SIGNED	0.507
c7.GSE17721_CTRL_VS_CPG_12H_BMDM_SIGNED	−0.508
c7.GSE1460_INTRATHYMIC_T_PROGENITOR_VS_ THYMIC_STROMAL_CELL_SIGNED	−0.508

Appendix D

Approximate calculation of PARSE scores

Exact calculation of PARSE score requires the solution of a number of NNLS problems, which complicates application. The NNLS solutions can be approximated with conventional least squares solutions, ultimately transforming the calculation of an approximate PARSE score into a simple weighted sum of gene expression measurements.

Recall that NMF finds factorisations of the form $A = WH$, with all elements of A, W, and H, being non-negative. In the reverse problem of PARSE calculation, A and $\widehat{W}$ are supplied, and H is to be estimated. I propose an approximation that removes the requirement that H be non-negative, $H \approx \widehat{W}^{+}A$, where $\widehat{W}^{+}$ is the Moore-Penrose pseudoinverse of $\widehat{W}$. By combining this approximation with the linear combination of metagene coefficients that forms the PARSE score, we can approximate PARSE as a simple weighted sum of gene expression measurements:

$$P = LH \tag{D.1}$$
$$\approx L\widehat{W}^{+}A \tag{D.2}$$
$$= kA \tag{D.3}$$

where P is the vector of PARSE score values, L is the metagene loadings for the PARSE score, $L = \begin{pmatrix} 1.354 & -1.548 & 0 & 0 & -1.354 & 1.548 \end{pmatrix}$, and k is a

row vector of gene loadings for calculation of an approximate PARSE score. Approximation of P by kA appears excellent; when tested on APGI gene expression measurements, the approximation closely matched the more laborious exact NNLS solution (Figure D.1).

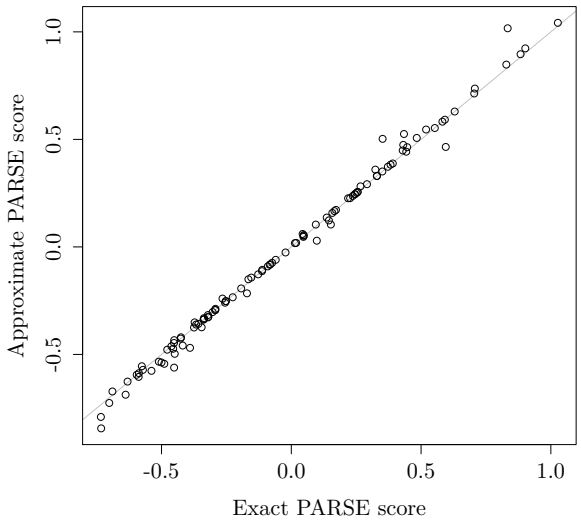


Figure D.1: The linear PARSE score approximation $P \approx kA$ closely matches the exact version calculated using NNLS, when evaluated on APGI GEX data.

To use the approximation in practice, perform the following steps:

1. Prepare a gene $\times$ sample matrix of linear expression estimates A, in which values for each row (gene) have been scaled to encompass the range 0 to 1.

2. Subset A to only the genes present in the k vector (Table D.1), and arrange rows of A so that they exactly match the order of k. If genes present in k are missing from A, insert all-zero rows for these genes into A.

3. Calculate approximate PARSE scores P as $P = kA$. This is equivalent to, for each column (sample) of A, multiplying each

entry of the column of A with the corresponding entry of k, and summing the results.

The loading vector for the calculation of approximate PARSE score, k^T, is given in Table D.1.

Table D.1: Loading vector for the approximate PARSE score. For brevity and to assist interpretation, this has been split by sign into two columns, but in use both columns should be combined to form a single row vector k.

Gene symbol	Loading	Gene symbol	Loading
FEM1B	0.04785	GAB2	−0.03742
NCAPG	0.04487	FRZB	−0.03715
ANLN	0.04364	MIR99AHG	−0.03712
COL12A1	0.04098	RAP1GAP	−0.03483
LDHA	0.04004	NFIA	−0.03387
E2F7	0.03923	TCTA	−0.03326
SPHK1	0.03861	ELMOD3	−0.03300
CEP55	0.03755	SOBP	−0.03269
CHEK1	0.03669	GIMAP2	−0.03176
TMEM26	0.03659	STAT5B	−0.03172
CKAP2L	0.03545	UFC1	−0.03123
DCBLD2	0.03351	BOC	−0.03047
PHLDA1	0.03330	P2RY8	−0.03043
KANK4	0.03261	RNF103	−0.03019
TGFBI	0.03259	KIAA0513	−0.02989
PLAU	0.03213	SGSM1	−0.02933
COL5A3	0.03177	TOR2A	−0.02926
CCNB1	0.03071	PPY	−0.02787
SPOCK1	0.03046	SH3GL1	−0.02784
ENO2	0.02998	RPA2	−0.02756
CAV1	0.02989	SELENBP1	−0.02707
KIF20A	0.02967	TRIM2	−0.02689
RACGAP1	0.02957	TCEA3	−0.02679

Continued on next page

Table D.1 – continued from previous page

Gene symbol	Loading	Gene symbol	Loading
PPAPDC1A	0.02867	HIPK2	−0.02620
RBMS2	0.02834	CAPN6	−0.02615
RHOF	0.02828	ARHGAP24	−0.02524
CDA	0.02792	TSTD1	−0.02503
NCAPD2	0.02756	ALDH5A1	−0.02452
MCM4	0.02708	BCKDK	−0.02452
LOX	0.02695	GPC3	−0.02419
PTGES	0.02681	EPHX2	−0.02392
FER	0.02675	DCAF8	−0.02374
EYA3	0.02671	PPM1H	−0.02311
IL20RB	0.02671	PRDM16	−0.02289
GATC	0.02661	MC1R	−0.02281
KLHL5	0.02641	PEX11B	−0.02280
ARL4C	0.02609	SMOX	−0.02258
ATAD2	0.02602	LMO3	−0.02246
TPX2	0.02590	RPIA	−0.02226
FGD6	0.02545	POU2AF1	−0.02222
PRC1	0.02492	ST3GAL2	−0.02187
MCM10	0.02451	ZSCAN32	−0.02184
BIRC5	0.02419	ZFPM1	−0.02180
ZBED2	0.02396	BCL11B	−0.02161
KNTC1	0.02375	C9orf152	−0.02152
NUP155	0.02330	SLC40A1	−0.02146
TNFRSF6B	0.02308	CADPS2	−0.02136
HJURP	0.02296	PHOSPHO2	−0.02129
PXDN	0.02281	ST6GAL1	−0.02118
COLGALT1	0.02272	PLAC9	−0.02093
PLOD2	0.02261	EIF2AK3	−0.02073
TWIST1	0.02246	IFT140	−0.02068
RALGAPB	0.02214	CHN2	−0.02051
FSCN1	0.02159	ZNF658	−0.01988

Continued on next page

Table D.1 – continued from previous page

Gene symbol	Loading	Gene symbol	Loading
SPOCD1	0.02117	MEOX1	−0.01961
SERPINH1	0.02086	FAM134B	−0.01945
GAPDH	0.02073	THSD7B	−0.01931
DSG3	0.02070	TRAPPC2	−0.01920
MELK	0.02067	ADH1A	−0.01845
DCUN1D5	0.02056	LINC01184	−0.01837
TUBA1C	0.02053	SLC12A2	−0.01821
CST6	0.02032	MRAP2	−0.01810
GABPB1	0.01929	RASL11B	−0.01808
KRT7	0.01916	RERGL	−0.01801
DENND1A	0.01898	PREP	−0.01799
AURKB	0.01869	TMTC4	−0.01797
PRR11	0.01859	TMED1	−0.01796
RFC5	0.01848	TLE4	−0.01794
SLC16A3	0.01842	CAMK1G	−0.01790
SUGCT	0.01833	GATA6	−0.01780
SCYL2	0.01826	CCR7	−0.01775
KRT6A	0.01795	SCGB2A1	−0.01773
P4HA2	0.01770	CCL19	−0.01715
PROSER2	0.01761	PCF11	−0.01710
PTPN21	0.01723	FAM189A2	−0.01692
PYGL	0.01714	MCOLN2	−0.01684
GINS2	0.01713	PLEKHS1	−0.01672
PGBD3	0.01700	PRMT7	−0.01665
COL7A1	0.01688	AXIN2	−0.01658
LETM2	0.01687	TOM1	−0.01640
PDLIM7	0.01678	RERE	−0.01635
KRT17	0.01644	A4GNT	−0.01632
ERRFI1	0.01597	CDK12	−0.01624
ASPM	0.01593	CNNM1	−0.01611
C1QTNF6	0.01572	HSPB6	−0.01586

Continued on next page

Table D.1 – continued from previous page

Gene symbol	Loading	Gene symbol	Loading
DERA	0.01568	LCNL1	−0.01571
MIF	0.01560	MTRNR2L1	−0.01563
C5orf46	0.01559	DYNC2H1	−0.01537
EMP3	0.01550	NPM1	−0.01520
CDK2	0.01546	CARHSP1	−0.01515
POC1A	0.01507	RGS5	−0.01505
FST	0.01504	CLEC3B	−0.01500
KCTD10	0.01501	NR0B2	−0.01468
SULF2	0.01494	ARSD	−0.01466
CCDC88A	0.01480	GNPAT	−0.01458
KIF14	0.01477	MARS2	−0.01442
DSG2	0.01463	KCTD5	−0.01440
UHRF2	0.01445	MRPL24	−0.01395
ZNF185	0.01435	ABLIM1	−0.01392
SLC2A1	0.01424	ITPKB	−0.01390
KIF2C	0.01417	FHDC1	−0.01380
FLRT3	0.01416	C2orf70	−0.01360
CNIH3	0.01413	RAVER2	−0.01352
ITGA5	0.01407	AKR1A1	−0.01321
DNAJC9	0.01385	CACHD1	−0.01313
ANGPTL4	0.01365	ACYP2	−0.01298
KIAA1549L	0.01354	CTSL	−0.01263
PPP1R14B	0.01352	TM9SF3	−0.01255
PAX8	0.01350	PP7080	−0.01242
FAM91A1	0.01341	IGLL3P	−0.01241
EDIL3	0.01326	ST6GALNAC1	−0.01232
RAB31	0.01316	VPS35	−0.01219
P2RY2	0.01288	TAF5L	−0.01213
CDC45	0.01256	STK39	−0.01196
SPIN4	0.01254	NFIX	−0.01186
APCDD1	0.01244	TNFRSF17	−0.01180

Continued on next page

Table D.1 – continued from previous page

Gene symbol	Loading	Gene symbol	Loading
ABHD5	0.01227	PBXIP1	−0.01174
ANKLE2	0.01205	PLIN2	−0.01174
FAM83A	0.01202	GOLM1	−0.01171
KYNU	0.01181	SEPW1	−0.01161
ANGPTL2	0.01178	FYN	−0.01133
B3GALTL	0.01113	CA8	−0.01129
MME	0.01102	CSNK1D	−0.01128
FAH	0.01035	SLC30A3	−0.01126
NEURL2	0.01012	SEC23IP	−0.01125
CTSV	0.00987	RFK	−0.01090
PGAM5	0.00973	SDIM1	−0.01083
ATL3	0.00972	ARFGAP3	−0.01070
CD70	0.00954	CYP2S1	−0.01044
CHAF1B	0.00920	TARBP2	−0.01019
PIGL	0.00833	SERTAD2	−0.00995
PAX8-AS1	0.00830	IL33	−0.00991
LMTK2	0.00804	FAM120AOS	−0.00980
ACKR3	0.00802	SYNE2	−0.00968
KRT6C	0.00798	COX4I2	−0.00943
PRKCDBP	0.00755	ANKRD22	−0.00941
DPY19L1	0.00749	COLGALT2	−0.00903
NACC2	0.00733	FBXW8	−0.00891
POLA2	0.00692	MARCKSL1	−0.00884
DKK1	0.00649	BTN3A1	−0.00868
FBXO22	0.00649	C1orf56	−0.00865
USP30	0.00629	PCDH20	−0.00861
APCS	0.00602	EXOSC8	−0.00850
BBS2	0.00587	AMOT	−0.00825
TRERF1	0.00581	WNT2B	−0.00812
GPR176	0.00563	SLAMF9	−0.00761
FGG	0.00548	PCOLCE2	−0.00752

Continued on next page

Table D.1 – continued from previous page

Gene symbol	Loading	Gene symbol	Loading
AKIP1	0.00545	ZSCAN16	−0.00720
IDH2	0.00528	CIDECP	−0.00684
PFKFB4	0.00525	BAMBI	−0.00680
ANKRD37	0.00474	IL1R2	−0.00660
SLC2A3	0.00438	SAMD5	−0.00655
IGFBP1	0.00427	HSP90B1	−0.00641
A4GALT	0.00418	CFDP1	−0.00617
CEBPB	0.00404	RMND5A	−0.00614
PLOD1	0.00369	CIDEC	−0.00596
VSTM2L	0.00352	TPD52L2	−0.00579
XXYLT1	0.00341	ZNF565	−0.00565
MAP3K8	0.00338	ACE	−0.00556
SNRPB	0.00276	AGRP	−0.00509
TOM1L2	0.00266	PLIN3	−0.00506
NRP2	0.00250	ARHGEF19	−0.00476
P4HA1	0.00225	DHRS9	−0.00454
HRASLS2	0.00196	ATF7IP2	−0.00405
UPP1	0.00182	NELFE	−0.00390
SPP1	0.00175	RGS3	−0.00319
LAMA5	0.00174	TNFRSF10D	−0.00315
PHACTR3	0.00172	LOC100506562	−0.00290
ZPLD1	0.00165	RFX2	−0.00264
CATSPER1	0.00163	SNORA11D	−0.00256
ABHD16A	0.00143	FGB	−0.00252
PPP1R3C	0.00125	ICAM2	−0.00232
ADM	0.00122	LGALS9B	−0.00232
SOD2	0.00120	POP5	−0.00224
PPP1R12B	0.00096	NMB	−0.00205
NAMPT	0.00071	SERPINB3	−0.00201
KCNQ3	0.00040	ORC1	−0.00199
MCEMP1	0.00025	ALOX5AP	−0.00179

Continued on next page

Table D.1 – continued from previous page

Gene symbol	Loading	Gene symbol	Loading
LYNX1	0.00001	SLC15A1	−0.00139
		OAZ1	−0.00134
		TREM1	−0.00073
		IKBIP	−0.00033

Appendix E

R code to calculate MSKCC nomogram survival estimates

```
fit.mskcc = list(
        inputs = list(
        History.Diagnosis.AgeAt = list(
                margins = data.frame(value = 65, fraction
                     = 1),
                scorefunc = function(x) { x = x; -2/15*
                    pmin(pmax(x, 0), 90) + 12 }),
        Patient.Sex = list(
                margins = data.frame(value = c("M", "F"),
                     fraction = c(0.501, 1-0.501)),
                scorefunc = function(x) { 3*I(x == "M")
                    }),
        Portal.Vein = list(
                margins = data.frame(value = c(TRUE,
                    FALSE), fraction = c(0.144, 1-0.144)),
                scorefunc = function(x) { 10*I(x == TRUE)
                     }),
        Splenectomy = list(
                margins = data.frame(value = c(TRUE,
                    FALSE), fraction = c(0.099, 1-0.099)),
                scorefunc = function(x) { 62*I(x == TRUE)
                     }),
```

```
Treat.MarginPositive = list(
        margins = data.frame(value = c(TRUE,
            FALSE), fraction = c(0.207, 1-0.207)),
        scorefunc = function(x) { 4*I(x == TRUE)
            }),
Path.LocationBody = list(
        margins = data.frame(value = c(FALSE,
            TRUE), fraction = c(0.894, 1-0.894)),
        scorefunc = function(x) { 51*I(x == TRUE)
             }),
Path.Differentiation = list(
        margins = data.frame(value = c("1", "2",
            "3", "4"), fraction = c(0.142, 0.564,
            1-0.142-0.564, 0)),
        scorefunc = function(x) { 14*I(x == "2")
            + 35*I(x == "3") + 35*I(x == "4") }),
                    # Undifferentiated (4) not
            covered by the MSKCC nomogram; here
            assign the same score as poorly
            differentiated (3)
Posterior.Margin = list(
        margins = data.frame(value = c(TRUE,
            FALSE), fraction = c(0.86, 1-0.86)),
        scorefunc = function(x) { 22*I(x == TRUE)
             }),
Path.LN.Involved = list(
        margins = data.frame(value = 2.1,
            fraction = 1),
        scorefunc = function(x) {
                x = pmin(40, pmax(x, 0))
                fitfun = splinefun(c(0, 1, 2, 3,
                    4, 10, 15, 20, 25, 30, 35, 40)
                    , c(0, 14.56, 24.64, 30.28,
                    33.00, 39.05, 43.89, 48.83,
                    53.77, 58.61, 63.55, 68.49),
                    method = "natural")
                fitfun(x)
        }),
Path.LN.Negative = list(
        margins = data.frame(value = 16.9,
            fraction = 1),
        scorefunc = function(x) { (pmin(pmax(x,
```

```
            0), 90)-90)*-11/90 }),
Back.pain = list(
        margins = data.frame(value = c(TRUE,
            FALSE), fraction = c(0.137, 1-0.137)),
        scorefunc = function(x) { 15*I(x == TRUE)
             }),
Stage.pT.Simplified = list(
        margins = data.frame(value = c("T1", "T2"
            , "T34"), fraction = c(0.037, 0.119,
            1-0.037-0.119)),
        scorefunc = function(x) { 36*I(x == "T1")
             + 11*I(x == "T34") }),
        # The following matches the original
            Brennan nomogram, but was not used as
            there are too few T4
        # tumours in either the NSWPCN *or* the
            MSKCC cohorts -- how the T4
            coefficient was ever estimated,
        # I'll never know.  The T34 coefficient
            of 11 was arrived at as (0.828*
            10+(1-0.037-0.119-0.828)*63)/
            (1-0.037-0.119),
        # being a frequency-weighted average of
            the T3 and T4 coefficients.
        # margins = data.frame(value = c("T1", "
            T2", "T3", "T4"), fraction = c(0.037,
            0.119, 0.828, 1-0.037-0.119-0.828)),
        # scorefunc = function(x) { 36*I(x == "T1
            ") + 10*I(x == "T3") + 63*I(x == "T4")
             }),
Weight.loss = list(
        margins = data.frame(value = c(TRUE,
            FALSE), fraction = c(0.537, 1-0.537)),
        scorefunc = function(x) { 3*I(x == TRUE)
            }),
Path.Size = list(
        margins = data.frame(),
        scorefunc = function(x) {
                x = pmin(16, pmax(x, 0))
                fitfun = splinefun(c(0, 1, 2, 3,
                    4, 6, 8, 10, 12, 14, 16), c(0,
                     29.74, 59.48, 86.70, 100,
```

```
                        97.29, 90.03, 82.77, 75.51,
                        68.25, 61.10), method = "
                        natural")
                fitfun(x)
        }) ),
outputs = list(
        DSS12mo = function(s) {
                x = pmax(50, pmin(350, s))
                fitfun = splinefun(c(79.0323,
                        115.02, 165.524, 197.278,
                        221.774, 242.339, 261.089,
                        279.839, 299.194, 323.992,
                        337.298), c(0.94, 0.9, 0.8,
                        0.7, 0.6, 0.5, 0.4, 0.3, 0.2,
                        0.1, 0.06))
                y = fitfun(x)
                pmax(0, pmin(1, y))
        },
        DSS24mo = function(s) {
                x = pmax(50, pmin(350, s))
                fitfun = splinefun(c(71.1694,
                        97.7823, 129.536, 153.73,
                        174.294, 193.347, 211.794,
                        231.452, 255.645, 303.125), c
                        (0.86, 0.8, 0.7, 0.6, 0.5,
                        0.4, 0.3, 0.2, 0.1, 0.01))
                y = fitfun(x)
                pmax(0, pmin(1, y))
        },
        DSS36mo = function(s) {
                x = pmax(50, pmin(350, s))
                fitfun = splinefun(c(69.3548,
                        101.109, 125.302, 145.867,
                        164.919, 183.367, 202.722,
                        226.915, 274.093), c(0.8, 0.7,
                         0.6, 0.5, 0.4, 0.3, 0.2, 0.1,
                         0.01))
                y = fitfun(x)
                pmax(0, pmin(1, y))
        })
)
```

```
applyNomogram = function(nomogram, data)
{
        scores = rowSums(sapply(names(nomogram$inputs),
            function(input) {
                if (input %in% colnames(data)) {
                        return(nomogram$inputs[[input]]$
                            scorefunc(data[,input]))
                }
                warning(sprintf("Marginalising␣missing␣
                    variable:␣%s", input))
                margin_score = sum(nomogram$inputs[[input
                    ]]$scorefunc(nomogram$inputs[[input]]$
                    margins$value) * nomogram$inputs[[
                    input]]$margins$fraction)
                return(rep(margin_score, nrow(data)))
        }))

        outputs = sapply(nomogram$outputs, function(f) f(
            scores))
        cbind(Score = scores, outputs)
}
```

Appendix F

Messina2 Algorithms

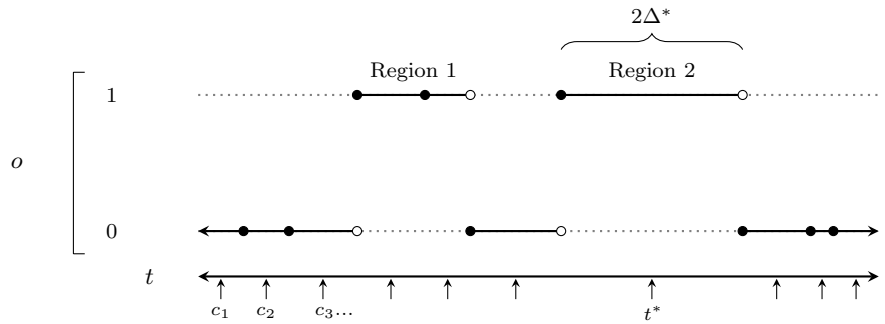


Figure F.1: Operation of the BestInterval algorithm. Example values of a binary objective o are shown for a range of input thresholds t. o is a sparse sample of the objective function f, evaluated at candidate cutpoints c, $o_i = f(w(c_i, d), y)$. At discrete points defined by observed data values (shown as dots), the objective function f can transition, as an observed data point changes its value relative to t, and therefore its assigned class. Two regions in which $o = 1$ are shown. BestInterval locates all such regions, selects the one with largest measure on t (margin), and returns its centre and margin as (t^*, Δ^*). In this example, the centre and margin of region 2 would be returned. To ensure that o is sampled at sufficient density, candidate thresholds $c_1, c_2, \ldots$ are defined between all consecutive values, and beyond the extrema, of x; these are indicated by small arrows.

Data: Covariate measurements $x \in \mathbb{R}^n$, associated dependent values $y \in \mathbb{Y}^n$, objective function $f : (\mathbb{B}^n, \mathbb{Y}^n) \to \mathbb{B}$, minimum subclass fraction $b \in [0, \frac{1}{2})$, and a number of bootstrap rounds r. x is to be sorted in ascending order. $\mathbb{Y}$ varies between modes of Messina2.

Result: A 3-tuple. If the fit failed, $(0, 0, 0)$. Otherwise, optimal classifier parameters (t^*, d^*, Δ^*), with threshold $t^* \in \mathbb{R}$, direction $d^* \in \{-1, 1\}$, and margin $\Delta^* \in \mathbb{R}^+$.

begin
  $n_{\text{pass}} \longleftarrow 0$;
  for $i \leftarrow 1$ **to** r **do**
    `// Generate a bootstrap sample of (x, y)`
    $(x_{\text{in}}, y_{\text{in}}, x_{\text{out}}, y_{\text{out}}) \longleftarrow \text{BootstrapSample}\,(x, y)$;
    `// Train Messina2Core on this bootstrap sample`
    $(t_{\text{in}}, d_{\text{in}}, \Delta_{\text{in}}) \longleftarrow \text{Messina2Core}\,(x_{\text{in}}, y_{\text{in}}, b, f)$;
    `// Assess performance of the Messina2Core classifier on the out-of-bag samples`
    if $d_{\text{in}} \neq 0$ **then**
      $w_{\text{out}} \longleftarrow \left(\left[x_{\text{out}_j} d_{\text{in}} \geq t_{\text{in}} d_{\text{in}} \right] \right)_{j=1}^{|x_{\text{out}}|}$;
      if $f\,(w_{\text{out}}, y_{\text{out}}) = \text{true}$ **then**
        $n_{\text{pass}} \longleftarrow n_{\text{pass}} + 1$;
      end
    end
  end
  `// Did the in-bag trained classifiers satisfy out-of-bag performance requirements in at least half of the bootstrap rounds?`
  if $r = 0 \vee n_{\text{pass}} \geq \frac{1}{2} r$ **then**
    `// Yes; return the fit on the full data`
    return $\text{Messina2Core}\,(x, y, b, f)$;
  else
    `// No; this fit failed.`
    return $(0, 0, 0)$;
  end
end

Algorithm 1: Messina2Entry, a bootstrap validation wrapper around Messina2Core. BootstrapSample is a function that generates a bootstrap sample of its matched tuple arguments x and y, returning both the in-bag and out-of-bag x and y values as tuples. If the number of bootstrap iterations r is zero, no bootstrap validation is performed and this algorithm falls through to Messina2Core.

Data: Covariate measurements $x \in \mathbb{R}^n$, associated dependent values $y \in \mathbb{Y}^n$, minimum subclass fraction $b \in [0, \frac{1}{2})$, and objective function $f : (\mathbb{B}^n, \mathbb{Y}^n) \to \mathbb{B}$. x is to be in ascending order. $\mathbb{Y}$ varies between modes of Messina2.

Result: A 3-tuple. If the fit failed, $(0, 0, 0)$. If the fit succeeded, optimal classifier values (t^*, d^*, Δ^*), with threshold $t^* \in \mathbb{R}$, direction $d^* \in \{-1, 1\}$, and margin $\Delta^* \in \mathbb{R}^+$.

begin

`// Define candidate thresholds` c `as the midpoints between consecutive unique values of` x

$c \longleftarrow \text{MakeCutpoints}\,(x, b)$;

$m \longleftarrow |c|$;

`// Test the objective` f `on each threshold in` c

for $i \leftarrow 1$ **to** m **do**

$w^+ \longleftarrow (\ [x_j \geq c_i]\ )_{j=1}^n$;

$w^- \longleftarrow (\ [x_j < c_i]\ )_{j=1}^n$;

$o_i^+ \longleftarrow f\,(w^+, y)$;

$o_i^- \longleftarrow f\,(w^-, y)$;

end

`// If no threshold passed` f`, return` $\varnothing$

if $o^+ \vee o^-$ *is all* false **then**

return $\varnothing$;

end

`// Search` o^+ `and` o^- `for the widest margin contiguous interval that passes` f

$(t^+, \Delta^+) \longleftarrow \text{BestInterval}(o^+, c)$;

$(t^-, \Delta^-) \longleftarrow \text{BestInterval}(o^-, c)$;

`// Return the best of the` o^+ `and` o^- `results`

if $\Delta^+ \geq \Delta^-$ **then**

return $(t^+, +1, \Delta^+)$;

else

return $(t^-, -1, \Delta^-)$;

end

end

Algorithm 2: Messina2Core. Searches the real line for the largest interval for which the objective $f(w, y)$ is true, and returns the centre and half-width of that interval.

Data: Covariate measurements $x \in \mathbb{R}^n$, and a minimum subclass fraction $b \in [0, \frac{1}{2})$. x is to be sorted in ascending order.
Result: A tuple of candidate cutpoints c, with values sorted in ascending order.

begin
  $x' \longleftarrow \text{unique}(x)$;
  $c \longleftarrow ()$;
  for $i \leftarrow 1$ **to** $|x'| - 1$ **do**
    $p \longleftarrow \frac{1}{2}(x'_i + x'_{i+1})$;
    $s \longleftarrow \frac{1}{n}\sum_{i=1}^{n}[x_i < p]$;
    if $s \geq b \wedge s \leq 1 - b$ **then**
      $c \longleftarrow c \oplus p$;
    end
  end
  if $b = 0$ **then**
    $c \longleftarrow -\infty \oplus c \oplus \infty$;
  end
  return c;
end

Algorithm 3: MakeCutpoints. Returns a sorted list of candidate cutpoints, each dividing the values in x into two groups, with the restriction that all groups must contain at least fraction b of the original values of x. The function unique accepts a tuple argument, and returns a tuple containing the unique values of its argument, in ascending order. The symbol $\oplus$ is used to represent tuple concatenation, and () is the empty tuple.

Data: $o \in \mathbb{B}^m$, $c \in \mathbb{R}^m$, $x \in \mathbb{R}^n$
Result: $(t^* \in \mathbb{R}, \Delta^* \in \mathbb{R}^+)$
begin
  $\Delta^* \longleftarrow 0$;
  $t^* \longleftarrow 0$;
  $i \longleftarrow 1$;
  while $i \leq m$ **do**
    if o_i *is* true **then**
      $r_L \longleftarrow \sup\{x_k \mid x_k \leq c_i \wedge k \in \mathbb{N}^+ \wedge k \leq n\}$;
      for $j \leftarrow i$ **to** m **do**
        if o_j *is* true **then**
          $r_R \longleftarrow \sup\{x_k \mid x_k \leq c_j \wedge k \in \mathbb{N}^+ \wedge k \leq n\}$;
        else
          break;
        end
      end
      $\Delta \longleftarrow \frac{1}{2}(r_R - r_L)$;
      if $\Delta > \Delta^*$ **then**
        $\Delta^* \longleftarrow \Delta$;
        $t^* \longleftarrow r_L + \frac{1}{2}\Delta$;
      end
      $i \longleftarrow j$;
    end
    $i \longleftarrow i + 1$;
  end
  return (t^*, Δ^*);
end

Algorithm 4: BestInterval. Identifies the centre (t^*) and half-width (Δ^*) of the largest interval on t for which $o(t) = \mathit{true}$. o is only sampled at values in c, but changes value only at values in x; the process by which this algorithm translates between these representations is illustrated in Figure F.1.

Glossary

AIC Akaike information criterion. 60

APGI Australian Pancreatic Cancer Genome Initiative. iii, vii, 7, 13–17, 20, 22, 25, 27, 29, 31, 34, 38, 40, 42, 43, 45, 46, 53, 55, 67, 69–75, 78, 98, 103–107, 110–112, 144

AUC area under the curve. v, vi, 63, 65, 74, 75

AWS Amazon Web Services. 74

BAF B allele frequency. 27

CA-19-9 carbohydrate antigen 19-9. 51, 54

CDF cumulative distribution function. vi, 95, 96

CPH Cox proportional hazard. 58–61, 81

CPSS complementary pair subset selection. 16, 40, 42, 117

CPV clinico-pathological variable. vii, 13, 14, 27, 30, 40, 46, 47, 51, 52, 54, 55, 82

CT computed tomography. 54

CV cross-validation. 21

DSD disease-specific death. 16, 40

DSS disease-specific survival. v, 21, 40

ECM extracellular matrix. 28

EMT epithelial to mesenchymal transition. i, 27–29, 31, 32, 34, 37, 52, 116

EUS endoscopic ultrasound. 52, 54, 79, 118

FAST feature aberration at survival times. 16, 40, 42

FDR false-discovery rate. 16, 40

FNA fine needle aspirate. 52, 54, 79, 118

FWER familywise error rate. 22–24

GEO Gene Expression Omnibus. 44

GEX gene expression. v, 6–14, 16, 40, 42, 45, 46, 102, 104, 111, 112, 144

GG generalised gamma. 61, 81

GSVA gene set variation analysis. ii, 28, 45, 46

IBS integrated Brier score. vii, 64, 81

ICA independent component analysis. 10–12

ICGC International Cancer Genome Consortium. 5, 6, 38

IDAT Illumina data. 38, 39

IHC immunohistochemical. 52, 54, 79

IHC immunohistochemistry. 52

KM Kaplan-Meier. v, 61, 62, 71, 81

LASSO least absolute shrinkage and selection operator. v, 20, 21

LOESS local regression. 58

MDS multidimensional scaling. 40

MSigDB molecular signatures database. iii, vii, 9, 27, 28, 45, 46, 48, 135, 136, 138, 141, 142

MSKCC Memorial Sloan-Kettering Cancer Center. ii, iii, 51, 67, 70–75, 77, 82, 153, 154, 156

NCBI National Center for Biotechnology Information. 44

NMF non-negative matrix factorisation. v, 10–14, 16–18, 36, 41, 117, 143

NNLS non-negative least squares. 20, 21, 43, 143, 144

NSWPCN New South Wales Pancreatic Cancer Network. iii, 54–59, 61–67, 69, 73, 79–81

PARSE prognostic axis risk stratification estimate. i–iii, v–vii, 21–27, 34, 36, 43, 44, 143–146, 148, 150

PCA principal component analysis. 10–12

PCOP the Pancreas Cancer Outcome Predictor. ii, v–vii, 50, 53, 54, 58, 64, 66–79, 81, 82, 118

PDAC pancreatic ductal adenocarcinoma. i, iii, vii, 1, 2, 4–8, 11–15, 17, 22, 27, 28, 32, 34–38, 40, 44, 79, 80, 87, 88, 98, 102–108, 110–112, 115–119

PH proportional hazard. 59–61

PI prognostic index. ii, v, 66–69, 72, 73, 75, 81

PPV positive predictive value. 51

ROC receiver operating characteristic. iii, 103, 104, 109, 111, 112

SIS sure independence screening. 16, 40, 42

SNMF/L sparse non-negative matrix factorisation, long variant. v, 17–19, 34, 36, 41–43

SNMF/W sparse non-negative matrix factorisation, wide variant. 36

SNP single nucleotide polymorphism. 27

SVM support vector machine. 85, 88

TCGA The Cancer Genome Atlas. vii, 5, 22–26, 31, 34, 44, 103, 105, 106, 110, 111, 113

TD-ROC time-dependent receiver operating characteristic. vi, 63, 65, 72–75, 81

VST variance stabilising transform. 39, 41, 45

References

[1] Cancer survival and prevalence in Australia: period estimates from 1982 to 2010 - cancer series no. 69, 2012.

[2] Cancer in Australia: An overview 2014 - cancer series no. 90, 2014.

[3] T. Adhikary, D. T. Brandt, K. Kaddatz, J. Stockert, S. Naruhn, W. Meissner, F. Finkernagel, J. Obert, S. Lieber, M. Scharfe, M. Jarek, P. M. Toth, F. Scheer, W. E. Diederich, S. Reinartz, R. Grosse, S. Müller-Brüsselbach, and R. Müller. Inverse PPARβ/δ agonists suppress oncogenic signaling to the ANGPTL4 gene and inhibit cancer cell invasion. *Oncogene*, 32:5241–52, 2013.

[4] C. F. Aliferis, A. Statnikov, I. Tsamardinos, J. S. Schildcrout, B. E. Shepherd, and F. E. Harrell. Factors influencing the statistical power of complex data analysis protocols for molecular signature development from microarray data. *PLoS ONE*, 4(3):e4922, 2009.

[5] O. Alter, P. O. Brown, and D. Botstein. Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Sciences*, 97(18):10101–10106, 2000.

[6] D. G. Altman, B. Lausen, W. Sauerbrei, and M. Schumacher. Danger of using "optimal" cutpoints in the evaluation of prognostic factors. *Journal of the National Cancer Institute*, 86(11):829–835, 1994.

[7] N. D. Arvold, A. Niemierko, H. J. Mamon, C. Fernandez-del Castillo, and T. S. Hong. Pancreatic cancer tumor size on CT scan versus pathologic specimen: implications for radiation treatment planning. *International Journal of Radiation Oncology, Biology, Physics*, 80(5):1383–1390, 2011.

[8] F. R. Bach. Considering cost asymmetry in learning classifiers. *Journal of Machine Learning Research*, 7:1713–1741, 2006.

[9] U. K. Ballehaninna and R. S. Chamberlain. The clinical utility of serum CA 19-9 in the diagnosis, prognosis and management of pancreatic adenocarcinoma: An evidence based appraisal. *Journal of Gastrointestinal Oncology*, 3(2):105–119, 2012.

[10] G. Barugola, M. Falconi, R. Bettini, L. Boninsegna, A. Casarotto, R. Salvia, C. Bassi, and P. Pederzoli. The determinant factors of recurrence following resection for ductal pancreatic cancer. *Journal of the Pancreas (Online)*, 8(1):132–140, 2007.

[11] K. M. Bever, E. A. Sugar, E. Bigelow, R. Sharma, and D. Laheru. The prognostic value of stroma in pancreatic cancer in patients receiving adjuvant therapy. *HPB*, 17(4):292–298, 2014.

[12] A. V. Biankin. Personal communication. Feb 2009.

[13] A. V. Biankin, J. G. Kench, E. K. Colvin, D. Segara, C. J. Scarlett, N. Q. Nguyen, D. K. Chang, A. L. Morey, C.-S. Lee, M. Pinese, and Others. Expression of S100A2 calcium-binding protein predicts response to pancreatectomy for pancreatic cancer. *Gastroenterology*, 137(2):558–568, 2009.

[14] A. V. Biankin, N. Waddell, K. S. Kassahn, M.-C. Gingras, L. B. Muthuswamy, A. L. Johns, D. K. Miller, P. J. Wilson, A.-M. Patch, J. Wu, and Others. Pancreatic cancer genomes reveal aberrations in axon guidance pathway genes. *Nature*, 491(7424):399–405, 2012.

[15] A. Bilici. Prognostic factors related with survival in patients with pancreatic adenocarcinoma. *World Journal of Gastroenterology*, 20(31):10802–10812, 2014.

[16] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[17] L. Breiman. Statistical modeling: The two cultures. *Statistical Science*, 16(3):199–215, 2001.

[18] M. F. Brennan, M. W. Kattan, D. Klimstra, and K. Conlon. Prognostic nomogram for patients undergoing resection for adenocarcinoma of the pancreas. *Annals of Surgery*, 240(2):293–298, 2004.

[19] J.-P. Brunet, P. Tamayo, T. R. Golub, and J. P. Mesirov. Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the National Academy of Sciences*, 101(12):4164–4169, 2004.

[20] J. Budczies, F. Klauschen, B. V. Sinn, B. Gyorffy, W. D. Schmitt, S. Darb-Esfahani, and C. Denkert. Cutoff Finder: A comprehensive and straightforward web application enabling rapid biomarker cutoff optimization. *PLoS ONE*, 7(12):1–7, 2012.

[21] R. L. Camp, M. Dolled-Filhart, and D. L. Rimm. X-tile: A new bio-informatics tool for biomarker assessment and outcome-based cut-point optimization. *Clinical Cancer Research*, 10:7252–7259, 2004.

[22] D. K. Chang. Personal communication. Mar 2015.

[23] W. Choi, S. Porten, S. Kim, D. Willis, E. Plimack, J. Hoffman-Censits, B. Roth, T. Cheng, M. Tran, I. L. Lee, J. Melquist, J. Bondaruk, T. Majewski, S. Zhang, S. Pretzsch, K. Baggerly, A. Siefker-Radtke, B. Czerniak, C. N. Dinney, and D. McConkey. Identification of distinct basal and luminal subtypes

of muscle-invasive bladder cancer with different sensitivities to frontline chemotherapy. *Cancer Cell*, 25(2):152–165, 2014.

[24] S. D. Chorlton, R. M. Hallett, and J. a. Hassell. A program to identify prognostic and predictive gene signatures. *BMC Research Notes*, 7(1):546, 2014.

[25] E. A. Collisson, A. Sadanandam, P. Olson, W. J. Gibb, M. Truitt, S. Gu, J. Cooc, J. Weinkle, G. E. Kim, L. Jakkula, H. S. Feiler, A. H. Ko, A. B. Olshen, K. L. Danenberg, M. A. Tempero, P. T. Spellman, D. Hanahan, and J. W. Gray. Subtypes of pancreatic ductal adenocarcinoma and their differing responses to therapy. *Nature Medicine*, 17(4):500–503, 2011.

[26] C. Cox, H. Chu, M. Schneider, and A. Muñoz. Parametric survival analysis and taxonomy of hazard functions for the generalized gamma distribution. *Statistics in Medicine*, 26:4352–4374, 2007.

[27] Editors. NCCN Guidelines v1.2015: Pancreatic Adenocarcinoma, 2015.

[28] B. Efron. Better bootstrap confidence intervals. *Journal of the American Statistical Association*, 82(397):171–185, 1987.

[29] J. Fan and J. Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B*, 70(5):849–911, 2008.

[30] J. Ferlay, I. Soerjomataram, M. Ervik, R. Dikshit, S. Eser, C. Mathers, M. Rebelo, D. M. Parkin, D. Forman, and F. Bray. Globocan 2012 v1.0, cancer incidence and mortality worldwide: Iarc cancerbase no. 11, 2013. Accessed on 25/3/2015.

[31] A. Frigyesi and M. Höglund. Non-negative matrix factorization for the analysis of complex gene expression data: Identification of clinically relevant tumor subtypes. *Cancer Informatics*, 6:275–292, 2008.

[32] G. Garcea, A. R. Dennison, C. J. Pattenden, C. P. Neal, C. D. Sutton, and D. P. Berry. Survival following curative resection for pancreatic ductal adenocarcinoma: A systematic review of the literature. *Journal of the Pancreas*, 9(2):99–132, 2008.

[33] R. C. Gentleman, V. J. Carey, D. M. Bates, and others. Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology*, 5:R80, 2004.

[34] A. Gorst-Rasmussen and T. Scheike. Independent screening for single-index hazard rate models with ultrahigh dimensional features. *Journal of the Royal Statistical Society: Series B*, 75(2):217–245, 2013.

[35] E. Graf, C. Schmoor, W. Sauerbrei, and M. Schumacher. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine*, 18(17-18):2529–2545, 1999.

[36] P. M. Grambsch and T. M. Therneau. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, 81(3):515–526, 1994.

[37] L. R. Grate. Many accurate small-discriminatory feature subsets exist in microarray transcript data: biomarker discovery. *BMC Bioinformatics*, 6:97, 2005.

[38] C. J. Gröger, M. Grubinger, T. Waldhör, K. Vierlinger, and W. Mikulits. Meta-analysis of gene expression signatures defining the epithelial to mesenchymal transition during cancer progression. *PLoS ONE*, 7(12):e51136, 2012.

[39] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

[40] D. Hanahan and R. a. Weinberg. Hallmarks of cancer: the next generation. *Cell*, 144(5):646–674, 2011.

[41] S. Hänzelmann, R. Castelo, and J. Guinney. GSVA: gene set variation analysis for microarray and RNA-Seq data. *BMC Bioinformatics*, 14(1):e7, 2013.

[42] F. E. Harrell, R. M. Califf, D. B. Pryor, K. L. Lee, and R. a. Rosati. Evaluating the yield of medical tests. *JAMA*, 247(18):2543–2546, 1982.

[43] H. C. Harsha, K. Kandasamy, P. Ranganathan, S. Rani, S. Ramabadran, S. Gollapudi, L. Balakrishnan, S. B. Dwivedi, D. Telikicherla, L. D. N. Selvan, R. Goel, S. Mathivanan, A. Marimuthu, M. Kashyap, R. F. Vizza, R. J. Mayer, J. a. Decaprio, S. Srivastava, S. M. Hanash, R. H. Hruban, and A. Pandey. A compendium of potential biomarkers of pancreatic cancer. *PLoS Medicine*, 6(4):e1000046, 2009.

[44] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.

[45] P. J. Heagerty and Y. Zheng. Survival model predictive accuracy and ROC curves. *Biometrics*, 61(1):92–105, 2005.

[46] S. M. Hewitt, F. a. Lewis, Y. Cao, R. C. Conrad, M. Cronin, K. D. Danenberg, T. J. Goralski, J. P. Langmore, R. G. Raja, P. M. Williams, J. F. Palma, and J. a. Warrington. Tissue handling and specimen preparation in surgical pathology: Issues concerning the recovery of nucleic acids from formalin-fixed, paraffin-embedded tissue. *Archives of Pathology and Laboratory Medicine*, 132:1929–1935, 2008.

[47] M. Hidalgo. Pancreatic cancer. *New England Journal of Medicine*, 362(17):1605–1617, 2010.

[48] W. Hilgers and S. E. Kern. Molecular genetic basis of pancreatic adenocarcinoma. *Genes, Chromosomes and Cancer*, 26(1):1–12, 1999.

[49] C.-K. Ho, J. Kleeff, H. Friess, and M. W. Büchler. Complications of pancreatic surgery. *HPB*, 7(2):99–108, 2005.

[50] S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

[51] T. Hothorn and B. Lausen. On the exact distribution of maximally selected rank statistics. *Computational Statistics & Data Analysis*, 43:121–137, 2003.

[52] N. Howlader, A. M. Noone, M. Krapcho, J. Garshell, D. Miller, S. F. Altekruse, C. L. Kosary, M. Yu, J. Ruhl, Z. Tatalovich, A. Mariotto, D. R. Lewis, H. S. Chen, E. J. Feuer, and K. A. Cronin. Seer cancer statistics review, 1975-2011, nov 2013.

[53] R. H. Hruban, M. Goggins, J. Parsons, and S. E. Kern. Progression model for pancreatic cancer. *Clinical Cancer Research*, 6:2969–2972, 2000.

[54] W. Iba and P. Langley. Induction of one-level decision trees. In M. Kaufmann, editor, *ML92: Proceedings of the Ninth International Conference on Machine Learning, Aberdeen, Scotland, 13 July 1992*, pages 233–240, 1992.

[55] H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer. Random survival forests. *The Annals of Applied Statistics*, 2(3):841–860, 2008.

[56] S. Karrila, J. H. E. Lee, and G. Tucker-Kellogg. A comparison of methods for data-driven cancer outlier discovery, and an application scheme to semisupervised predictive biomarker discovery. *Cancer Informatics*, 10:109–20, 2011.

[57] H. Kim and H. Park. Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. *Bioinformatics*, 23(12):1495–1502, 2007.

[58] S.-H. Kim, Y.-Y. Park, S.-W. Kim, J.-S. Lee, D. Wang, and R. N. DuBois. ANGPTL4 induction by prostaglandin E2 under hypoxic conditions promotes colorectal cancer progression. *Cancer Research*, 71(22):7010–7020, 2011.

[59] T. H. Kim, S.-S. Han, S.-J. Park, W. J. Lee, S. M. Woo, T. Yoo, S. H. Moon, S. H. Kim, E. K. Hong, D. Y. Kim, and J.-W. Park. CA 19-9 level as indicator of early distant metastasis and therapeutic selection in resected pancreatic cancer. *International Journal of Radiation Oncology, Biology, Physics*, 81(5):e743–748, 2011.

[60] J. L. Kopp, G. von Figura, E. Mayes, F. F. Liu, C. L. Dubois, J. P. Morris, F. C. Pan, H. Akiyama, C. V. E. Wright, K. Jensen, M. Hebrok, and M. Sander. Identification of Sox9-dependent acinar-to-ductal reprogramming as the principal mechanism for initiation of pancreatic ductal adenocarcinoma. *Cancer Cell*, 22(6):737–750, 2012.

[61] R. D. Leclerc. Survival of the sparsest: robust gene networks are parsimonious. *Molecular Systems Biology*, 4(213):213, 2008.

[62] S. H. Lee, H. Kim, J.-H. Hwang, E. Shin, H. S. Lee, D. W. Hwang, J. Y. Cho, Y.-S. Yoon, H.-S. Han, and B. H. Cha. CD24 and S100A4 expression in resectable pancreatic cancers with earlier disease recurrence and poor survival. *Pancreas*, 43(3):380–388, 2014.

[63] S.-I. Lee and S. Batzoglou. Application of independent component analysis to microarrays. *Genome Biology*, 4(11):R76, 2003.

[64] L. J. Lesko and A. J. J. Atkinson. Use of biomarkers and surrogate endpoints in drug development and regulatory decision making: Criteria, validation, strategies. *Annual Review of Pharmacology and Toxicology*, 41:347–366, 2001.

[65] M. S. Lewicki and T. J. Sejnowski. Learning overcomplete representations. *Neural Computation*, 12(2):337–365, 2000.

[66] W. Liebermeister. Linear modes of gene expression determined by independent component analysis. *Bioinformatics*, 18(1):51–60, 2002.

[67] C. a. Livasy, G. Karaca, R. Nanda, M. S. Tretiakova, O. I. Olopade, D. T. Moore, and C. M. Perou. Phenotypic evaluation of the basal-like subtype of invasive breast carcinoma. *Modern Pathology*, 19:264–271, 2006.

[68] J. Lundin, P. J. Roberts, P. Kuusela, and C. Haglund. The prognostic value of preoperative serum levels of CA 19-9 and CEA in patients with pancreatic cancer. *British Journal of Cancer*, 69:515–519, 1994.

[69] G. Luo, J. Long, B. Zhang, C. Liu, J. Xu, Q. Ni, and X. Yu. Stroma and pancreatic ductal adenocarcinoma: An interaction loop. *Biochimica et Biophysica Acta - Reviews on Cancer*, 1826(1):170–178, 2012.

[70] R. C. MacCallum, K. F. Widaman, S. Zhang, and S. Hong. Sample size in factor analysis. *Psychological Methods*, 4(1):84–99, 1999.

[71] D. Mahadevan and D. D. Von Hoff. Tumor-stroma interactions in pancreatic ductal adenocarcinoma. *Molecular Cancer Therapeutics*, 6(4):1186–1197, 2007.

[72] A. Maitra and R. H. Hruban. Pancreatic cancer. *Annual Review of Pathology*, 3(2):157–188, 2008.

[73] R. a. Maronna and R. H. Zamar. Robust estimates of location and dispersion for high-dimensional datasets. *Technometrics*, 44:307–317, 2002.

[74] D. Padua, X. H. F. Zhang, Q. Wang, C. Nadal, W. L. Gerald, R. R. Gomis, and J. Massagué. TGFβ primes breast tumors

for lung metastasis seeding through angiopoietin-like 4. *Cell*, 133:66–77, 2008.

[75] M. J. Penciana and R. B. D'Agostino. Overall C as a measure of discrimination in survival analysis: Model specific population value and confidence interval estimation. *Statistics in Medicine*, 23(December 2003):2109–2123, 2004.

[76] M. S. Pepe, R. Etzioni, Z. Feng, J. D. Potter, M. Lou, M. Thornquist, M. Winget, Y. Yasui, and I. Ntroduction. Phases of biomarker development for early detection of cancer. *Cancer*, 93(14):1054–1061, 2001.

[77] M. Pinese, C. J. Scarlett, J. G. Kench, E. K. Colvin, D. Segara, S. M. Henshall, R. L. Sutherland, and A. V. Biankin. Messina: A novel analysis tool to identify biologically relevant molecules in disease. *PLoS ONE*, 4(4):e5337, 2009.

[78] A. Popescu, A.-M. Ciocâlteu, D. I. Gheonea, S. Iordache, C. F. Popescu, A. SÄ$f$ftoiu, and T. Ciurea. Utility of endoscopic ultrasound multimodal examination with fine needle aspiration for the diagnosis of pancreatic insulinoma - A case report. *Current Health Sciences Journal*, 38(1):36–40, 2012.

[79] P. P. Provenzano, C. Cuevas, A. E. Chang, V. K. Goel, D. D. Von Hoff, and S. R. Hingorani. Enzymatic targeting of the stroma ablates physical barriers to treatment of pancreatic ductal adenocarcinoma. *Cancer Cell*, 21(3):418–429, 2012.

[80] B. Ray, M. Henaff, S. Ma, E. Efstathiadis, E. R. Peskin, M. Picone, T. Poli, C. F. Aliferis, and A. Statnikov. Information content and analysis methods for multi-modal high-throughput biomedical data. *Scientific Reports*, 4:e4411, 2014.

[81] A. D. Rhim, P. E. Oberstein, D. H. Thomas, E. T. Mirek, C. F. Palermo, S. a. Sastra, E. N. Dekleva, T. Saunders, C. P. Becerra, I. W. Tattersall, C. B. Westphalen, J. Kitajewski, M. G. Fernandez-Barrena, M. E. Fernandez-Zapico,

C. Iacobuzio-Donahue, K. P. Olive, and B. Z. Stanger. Stromal elements act to restrain, rather than support, pancreatic ductal adenocarcinoma. *Cancer Cell*, 25(6):735–747, 2014.

[82] H. Riediger, T. Keck, U. Wellner, A. zur Hausen, U. Adam, U. T. Hopt, and F. Makowiec. The lymph node ratio is the strongest prognostic factor after resection of pancreatic cancer. *Journal of Gastrointestinal Surgery*, 13(7):1337–44, 2009.

[83] P. Royston and D. G. Altman. External validation of a Cox prognostic model: principles and methods. *BMC Medical Research Methodology*, 13:e33, 2013.

[84] P. Royston, D. G. Altman, and W. Sauerbrei. Dichotomizing continuous predictors in multiple regression: A bad idea. *Statistics in Medicine*, 25(1):127–141, 2006.

[85] D. P. Ryan, T. S. Hong, and N. Bardeesy. Pancreatic adenocarcinoma. *The New England Journal of Medicine*, 371(11):1039–1049, sep 2014.

[86] S. A. Saidi, C. M. Holland, D. P. Kreil, D. J. C. MacKay, D. S. Charnock-Jones, C. G. Print, and S. K. Smith. Independent component analysis of microarray data in the study of endometrial cancer. *Oncogene*, 23(39):6677–6683, 2004.

[87] C. Salla, P. Konstantinou, and P. Chatzipantelis. CK19 and CD10 expression in pancreatic neuroendocrine tumors diagnosed by endoscopic ultrasound-guided fine-needle aspiration cytology. *Cancer*, 117(6):516–521, 2009.

[88] R. D. Shah and R. J. Samworth. Variable selection with error control: another look at stability selection. *Journal of the Royal Statistical Society: Series B*, 75(1):55–80, 2013.

[89] O. Shivers. *Scsh Reference Manual*, 0.6.7 edition, sep 1994.

[90] M. Sinn, C. Denkert, J. K. Striefler, U. Pelzer, J. M. Stieler, M. Bahra, P. Lohneis, and B. Do. α-smooth muscle actin ex-

pression and desmoplastic stromal reaction in pancreatic cancer: Results from the CONKO-001 study. *British Journal of Cancer*, 111:1917–1923, 2014.

[91] S. Song, K. Nones, D. Miller, I. Harliwong, K. S. Kassahn, M. Pinese, M. Pajic, A. J. Gill, A. L. Johns, M. Anderson, and Others. qpure: A tool to estimate tumor cellularity from genome-wide single-nucleotide polymorphism profiles. *PLoS ONE*, 7(9):e45835, 2012.

[92] E. B. Stelow, C. Woon, S. E. Pambuccian, M. Thrall, M. W. Stanley, R. Lai, S. Mallery, and H. E. Gulbahce. Fine-needle aspiration cytology of pancreatic somatostatinoma: the importance of immunohistochemistry for the cytologic diagnosis of pancreatic endocrine neoplasms. *Diagnostic Cytopathology*, 33(2):100–105, 2005.

[93] E. Steyerberg and A. Vickers. Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology*, 21(1):128–138, 2010.

[94] J. K. Stratford, D. J. Bentrem, J. M. Anderson, C. Fan, K. a. Volmar, J. S. Marron, E. D. Routh, L. S. Caskey, J. C. Samuel, C. J. Der, L. B. Thorne, B. F. Calvo, H. J. Kim, M. S. Talamonti, C. a. Iacobuzio-Donahue, M. a. Hollingsworth, C. M. Perou, and J. J. Yeh. A six-gene signature predicts survival of patients with localized pancreatic ductal adenocarcinoma. *PLoS Medicine*, 7(7):e1000307, 2010.

[95] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005.

[96] G. J. Székely and M. L. Rizzo. Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143(8):1249–1272, 2013.

[97] J. E. Talmadge and I. J. Fidler. AACR centennial series: The biology of cancer metastasis: Historical perspective. *Cancer Research*, 70:5649–5669, 2010.

[98] A. E. Teschendorff, M. Journée, P. a. Absil, R. Sepulchre, and C. Caldas. Elucidating the altered transcriptional programs in breast cancer using independent component analysis. *PLoS Computational Biology*, 3(8):e161, 2007.

[99] B. Y. T. M. Therneau, P. M. Grambsch, D. Biostatistics, M. Clinic, and T. R. Fleming. Martingale-based residuals for survival models. *Biometrika*, 77(1):147–160, 1990.

[100] N. Tsukamoto, S. Egawa, M. Akada, K. Abe, Y. Saiki, N. Kaneko, S. Yokoyama, K. Shima, A. Yamamura, F. Motoi, H. Abe, H. Hayashi, K. Ishida, T. Moriya, T. Tabata, E. Kondo, N. Kanai, Z. Gu, M. Sunamura, M. Unno, and A. Horii. The expression of S100A4 in human pancreatic cancer is associated with invasion. *Pancreas*, 42(6):1027–1033, 2013.

[101] A. Van den Broeck, H. Vankelecom, R. Van Eijsden, O. Govaere, and B. Topal. Molecular markers associated with outcome and metastasis in human pancreatic cancer. *Journal of Experimental & Clinical Cancer Research*, 31(1):68, 2012.

[102] H. C. van Houwelingen. Validation, calibration, revision and combination of prognostic survival models. *Statistics in Medicine*, 19:3401–3415, 2000.

[103] L. J. van't Veer, H. Dai, M. J. van de Vijver, Y. D. He, A. A. M. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards, and S. H. Friend. Gene expression

profiling predicts clinical outcome of breast cancer. *Nature*, 415(6871):530–6, 2002.

[104] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag New York, 2nd edition, 2000.

[105] V. N. Vapnik. An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, 10(5):988–999, 1999.

[106] D. Venet, J. E. Dumont, and V. Detours. Most random gene expression signatures are significantly associated with breast cancer outcome. *PLoS Computational Biology*, 7(10):e1002240, 2011.

[107] A. J. Vickers and A. M. Cronin. Everything you always wanted to know about evaluating prediction models (but were too afraid to ask). *Urology*, 76(6):1298–1301, 2010.

[108] A. J. Vickers, A. M. Cronin, E. B. Elkin, and M. Gonen. Extensions to decision curve analysis, a novel method for evaluating diagnostic tests, prediction models and molecular markers. *BMC medical informatics and decision making*, 8:e53, 2008.

[109] N. Waddell, M. Pajic, A.-m. Patch, D. K. Chang, K. S. Kassahn, P. Bailey, A. L. Johns, D. Miller, K. Nones, K. Quek, M. C. J. Quinn, A. J. Robertson, M. Z. H. Fadlullah, T. J. C. Bruxner, A. N. Christ, I. Harliwong, S. Idrisoglu, S. Manning, C. Nourse, E. Nourbakhsh, S. Wani, P. J. Wilson, E. Markham, N. Cloonan, M. J. Anderson, J. L. Fink, O. Holmes, S. H. Kazakoff, C. Leonard, F. Newell, B. Poudel, S. Song, D. Taylor, N. Waddell, S. Wood, Q. Xu, J. Wu, M. Pinese, M. J. Cowley, H. C. Lee, M. D. Jones, A. M. Nagrial, J. Humphris, L. A. Chantrill, V. Chin, A. M. Steinmann, A. Mawson, E. S. Humphrey, E. K. Colvin, A. Chou, C. J. Scarlett, A. V. Pinho, M. Giry-laterriere, I. Rooman, J. S. Samra, J. G. Kench, J. A. Pettitt, N. D. Merrett, C. Toon, K. Epari, N. Q. Nguyen, A. Barbour, N. Zeps, N. B. Jamieson, J. S. Graham, S. P.

Niclou, R. Bjerkvig, D. Aust, R. H. Hruban, A. Maitra, C. A. Iacobuzio-donahue, R. Gru, C. L. Wolfgang, R. A. Morgan, R. T. Lawlor, V. Corbo, C. Bassi, M. Falconi, A. J. Gill, J. R. Eshleman, C. Pilarsky, A. Scarpa, E. A. Musgrove, J. V. Pearson, A. V. Biankin, and S. M. Grimmond. Whole genomes redefine the mutational landscape of pancreatic cancer. *Nature*, 518:495–501, 2015.

[110] C. B. Westphalen and K. P. Olive. Genetically engineered mouse models of pancreatic cancer. *Cancer Journal*, 18(6):502–510, 2012.

[111] Q. Xiao and G. Ge. Lysyl oxidase, extracellular matrix remodeling and cancer metastasis. *Cancer Microenvironment*, 5:261–273, 2012.

[112] C. Yau, L. Esserman, D. H. Moore, F. Waldman, J. Sninsky, and C. C. Benz. A multigene predictor of metastatic outcome in early stage hormone receptor-negative and triple-negative breast cancer. *Breast Cancer Research*, 12(5):R85, 2010.

[113] G. Zhang, P. He, H. Tan, A. Budhu, J. Gaedcke, B. M. Ghadimi, T. Ried, H. G. Yfantis, D. H. Lee, A. Maitra, N. Hanna, H. R. Alexander, and S. P. Hussain. Integration of metabolomics and transcriptomics revealed a fatty acid network exerting growth inhibitory effects in human pancreatic cancer. *Clinical Cancer Research*, 19(18):4983–4993, 2013.