List of Corrections

Fatal: plus patch vi
Fatal: For the disc: Although the correlation between computed tomography (CT) and endoscopic ultrasound (EUS) estimates of tumour size, and actual size upon resection, is respectable [2], full clinical validation of this prognostic's use in a pre-operative setting will require 7
Fatal: TODO. Waiting on DC's data. 7
Fatal: Something for the discussion here – maybe that's why stroma is a slippery prognostic 37
Fatal: give instantiation values for the algo somewhere 57
Fatal: Add the derivation in somewhere – perhaps an appendix. It's a pain in the arse so probs want to avoid the main text. 62
Fatal: Add specific value of mindepth used 63

Mah Dissertat'n

Mark Pinese

January 20, 2015 Build 0.0.400

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 …….....

Acknowledgements

Abstract

Da abstract.

Contents

Contents i

List of Figures iii

List of Tables iv

1 Introduction 1

2 A Preoperative Molecular Prognostic for Pancreas Cancer 2
- 2.1 Introduction . . . 3
- 2.2 Results . . . 6
- 2.3 Discussion . . . 16
- 2.4 Methods . . . 16

3 Signatures of Survival Processes in Pancreas Cancer 17
- 3.1 Introduction . . . 18
- 3.2 Results . . . 23
- 3.3 Discussion . . . 38
- 3.4 Methods . . . 43

4 Comparative genomics 52
- 4.1 Methods . . . 53

5 Conclusion 67

Appendices 69

A Basis matrix W for the six survival-associated metagenes 69

B MSigDB signatures correlated with axis A1 80

C MSigDB signatures correlated with axis A2 **82**

D Approximate calculation of PARSE scores **83**

Glossary **96**

References **100**

List of Figures

2.1 Prognostic predictor functional forms 9
2.2 Baseline hazard forms differ between patient sexes 10
2.3 Model survival predictions agree with stratified KM estimates . . . 12
2.4 Brier score paths for candidate models on holdout data 14
2.5 Bootstrapped differences in IBS between candidate models and KM0 15

3.1 Illustration of the gene deconvolution problem 20
3.2 Comparison of GEX deconvolution techniques on synthetic data . 21
3.3 Automatic selection of NMF factorization rank 29
3.4 Consensus matrix for the final rank-6 clustering 30
3.5 Basis matrix W of the final SNMF/L factorization 31
3.6 Fit trajectory of the least absolute shrinkage and selection operator (LASSO) predicting DSS from metagene coefficients 32
3.7 Prognostic metagenes form two axes of cell state 33
3.8 Prognostic axes are uncorrelated 33
3.9 Survival subgroups defined by PARSE score axes in different tumours 35
3.10 A1 signal is closely associated with meta-PCNA score 38
3.11 A2 signal is closely associated with meta-EMT score 40

4.1 Locality-sensitive FDR estimation of LOH calls using a Markov chain Beta-Uniform mixture model. 65
4.2 Locality-sensitive FDR estimation of CNV calls using a Markov chain double-Beta-Uniform mixture model. 66

D.1 Performance of the PARSE score approximation 84

List of Tables

2.1 Characteristics of patient cohorts 8

3.1 Characteristics of the Australian Pancreatic Cancer Genome Initiative (APGI) patient cohorts 28
3.2 PARSE score is prognostic in a range of TCGA cancers 34
3.3 Association P-values between metagenes and CPVs 39
3.4 CPVs tested for association with prognostic axis signals. 50
3.5 Subset of MSigDB signatures tested for association with axis activities 51

B.1 MSigDB signatures correlated with axis A1 81

C.1 MSigDB signatures correlated with axis A2 82

List of Algorithms

1 Determine if a locus is heterozygous 58
2 Calculate CNV loss P-values 64

Software versions

Unless otherwise specified, the following versions of software were used in all work.

bamtools	2.2.2
bedtools	2.18.2
cd-hit	4.6.1 **MP Fatal: plus patch**
FastQC	0.10.1
GATK	3.1-1
julia	0.3.2
MSigDB	4.0
muTect	1.1.6-4-g69b7a37
ncbi-blast	2.2.29
picard-tools	1.109
PROVEAN	1.1.5
Python	2.7.8 / 3.4.1
R	3.1.1
ahaz	1.14
depmixS4	1.3-2
doParallelMC	1.0.8
Exact	1.4
GSVA	1.14.1
illuminaHumanv4.db	1.24.0
lumi	2.18.0
lumidat	1.2.3
nleqslv	2.5
NMF	0.20.5
nnls	1.4
org.Hs.eg.db	3.0.0

randomForest	4.6-10
Rsolnp	1.14
survival	2.37-7
samtools	1.0
SHRiMP	2.2.3
strelka	1.0.14
tabix	1.0
vcftools	0.1.10
VEP	76

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 denote the Iverson bracket: $[X] := 1$ if X is true, else 0.
- x_+ indicates the value of the ramp function at real x, $x_+ := \max(0, x)$.

Chapter 1

Introduction

Chapter 2

A Preoperative Molecular Prognostic for Pancreas Cancer

Thesis: A preoperative prognostic tool for pancreas cancer can be developed to discriminate good between 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 cruicially 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 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 pre-operatively assessable, and can assist in making personalised treatment decisions.

2.1 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 [15]. However, pancreas cancer resection is a major procedure, with the potential for serious complications, morbidity, and reduced quality of life following recovery [26]. 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 [15]. 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 [4]. 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 selection 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.

The decision of whether or not to resect will always involve close consultation between the patient and doctors, comparing the costs and benefits of surgery as appropriate to each patient's case. The downsides of resection are well-understood, but the benefit to be gained is challenging to quantify and communicate, being highly dependent on the many particulars of an individual patient's disease. A simple approach to represent benefit from pancreas

cancer resection involves survival curves, which plot the probability that a patient will be alive, at a range of times following resection (TODO example fig). A survival curve distills the information from an arbitrary number of prognostic factors into a single simple figure, to provide an intuitive overview of a patient's expected disease course. If accurate, such curves can provide a concrete measure of benefit from resection, and thus provide invaluable input into the treatment decision process. Survival curves can be involved to calculate, which has likely limited their historical use. However, the modern prevalence of computers removes this barrier, and even a modest device such as a smart phone can easily run prognostic tools capable of generating accurate personalised patient survival curves.

A number of pancreas cancer grading and schemes and prognostic tools have been described, but inconsistent performance, or a reliance on information that can only be known post-operatively, limits their use in pre-operative decisions. The level of serum carbohydrate antigen 19-9 (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 [30, 3, 4, 35]. CA-19-9 levels are easily determined pre-operatively, 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 [3]. 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 [30].

The current standard prognostic tool for pancreas cancer is the Memorial Sloan-Kettering Cancer Center (MSKCC) nomogram [11], which integrates a number of clinico-pathological variables (CPVs) to arrive at point 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 post-operative context, and cannot assist in pre-operative 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 [7]) can only be reliably measured following resec-

tion. Any prognostic tool for pancreas cancer that relies heavily on classical CPVs will very likely share this same reliance on post-operative variables, and so an effective pre-operatively assessable prognostic will need to shirk classical CPVs, and leverage novel pre-operative 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 epithelial to mesenchymal transition (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 [5, 53, 31]. 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 pre-operative setting, is untested.

Recent techological developments have made possible the pre-operative measurement of tissue biomarkers during 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 [40, 45, 48], and in principle EUS-FNA-IHC could form the basis of a routine pre-operative 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 pre-operatively applicable techniques to measure these biomarkers, and multiple large, clinically annotated cohorts of resected pancreas cancer, 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 will be used to develop 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 initial version of PCOP is based on biomarker measurements made on tissue collected during resection, and thus will not be directly applicable pre-operatively. However, pilot study data will be used to demonstrate that levels of S100A2 and S100A4 measured by pre-operative EUS-FNA-IHC correlate well to tissue levels of the biomarkers measured on operative specimens, indicating that a more refined version of PCOP trained on pre-operative data will be equally effective.

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 post-operatively, and so cannot be used to inform the decision of whether or not to resect. Tissue biomarkers of metastatic potential might identify, pre-operatively, 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 are well-suited to . This chapter describes the use of pre-operatively assessable variables, including biomarker measurements, to create PCOP, a tool that produces estimates of prognosis. 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.

2.2 Results

Data from the large, retrospectively-acquired New South Wales Pancreatic Cancer Network (NSWPCN) cohort were used to derive PCOP, a tool to predict the survival of pancreas cancer patients following curative-intent resection. Discrimination and calibration of PCOP were verified on two independent surgical cohorts. Data from an EUS-FNA-IHC pilot study established that pre-operatively assessed tissue biomarker levels reflected measurements from operative biopsies, and therefore that PCOP could be translated to a pre-operative decision setting.

Prognostic variables and biomarkers

As the aim was to develop a prognostic predictor that could be applied pre-operatively, only factors that could be practically measured pre-operatively were considered for inclusion in the PCOP. The traditional CPVs that were judged to be pre-operatively 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. Pre-operative blood levels of the biomarker CA-19-9 were available for a subset of the training cohort, but none of the validation sets; for this reason, and the marker's generally poor performance in isolation [30], CA-19-9 levels were not considered for inclusion in the PCOP.

Pre-operative measurements of tumour size (for example, by CT X-ray or EUS) were not available in the training and validation sets, and were approximated by post-operative measurements during the development and testing of this nomogram. Similarily, biomarker measurements were approximated using IHC staining of tissue collected during resection, as only very limited pre-operative 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.
[1]

Cohorts and Characteristics

General characteristics of the NSWPCN, Glasgow, and Dresden cohorts are summarised in Table 2.1.[2]

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

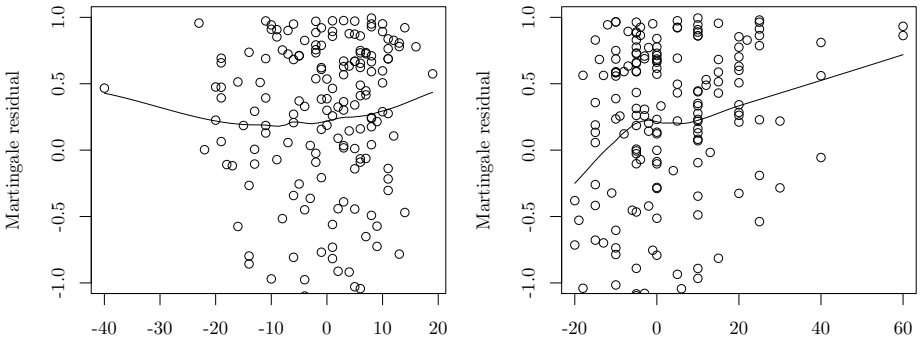
[1] MP Fatal: For the disc: Although the correlation between CT and EUS estimates of tumour size, and actual size upon resection, is respectable [2], full clinical validation of this prognostic's use in a pre-operative setting will require ...

[2] MP Fatal: TODO. Waiting on DC's data.

Table 2.1: Characteristics of the NSWPCN training cohort, and the APGI, Dresden, and Glasgow validation cohorts. Ordinal variables are shown as median, with quartiles in parentheses. Categorical variables for which percentages do not add up to 100% indicate the presence of minor unlisted categories.

Characteristic		NSWPCN	APGI	Dresden	Glasgow
Number of patients		xx	xx	xx	xx
Gender	Male	xx%	xx%	xx%	xx%
Age at diagnosis	(years)	xx (xx - xx)	xx (xx - xx)	xx (xx - xx)	xx (xx - xx)
Tumour location	Head	xx%	xx%	xx%	xx%
Size of longest axis	(mm)	xx (xx - xx)	xx (xx - xx)	xx (xx - xx)	xx (xx - xx)
S100A2 positive		xx (xx - xx)	xx (xx - xx)	xx (xx - xx)	xx (xx - xx)
S100A4 positive		xx (xx - xx)	xx (xx - xx)	xx (xx - xx)	xx (xx - xx)
Excision margin status	R0	xx%	xx%	xx%	xx%
	R1	xx%	xx%	xx%	xx%
	R2	xx%	xx%	xx%	xx%
Node involvement		xx%	xx%	xx%	xx%
Disease-specific death		xx%	xx%	xx%	xx%
Length of follow-up	(days)	xx (xx - xx)	xx (xx - xx)	xx (xx - xx)	xx (xx - xx)

split once into model building and testing sets. All model fitting and refinement described below was performed on the model building set, to yield three final candidate prognostic predictors. The performance of each of these three predictors was then assessed on the model test set, and the most parsimonious high-performing model was chosen as the final prognostic predictor, for subsequent external validation.



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

Figure 2.1: NSWPCN prognostic predictor functional forms. Smoothed Cox model martingale residual plots indicate hazard relationships that are approximately quadratic for centered age (panel a), and piecewise linear for centered tumour size (panel b). For clarity, plots have been restricted to the residual range $[-1, 1]$.

Model functional form and expanded terms The Cox proportional hazard (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 [52] indicated a possible weak U-shaped relationship for age at diagnosis (Figure 2.1(a)), and a knee-shaped form for size (Figure 2.1(b)), with the knee at approximately 0 in median-centered units. In subsequent fits these forms were modelled by adding a quadratic term for centered age at diagnosis, and a size_+ ramp term for centered axis size. The original set of five linear prognostic terms, plus the two additional nonlinear terms, was denoted the expanded term set.

Proportional hazards assumption A Grambsch-Therneau test [19] on the CPH model fit using all expanded terms indicated that patient sex violated the proportional hazards (PHs) assumption (P = TODO, Figure 2.2) – 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 repeated Grambsch-Thernau test on the stratified model indicated no further significant violations of the PH assumption (global P = TODO).

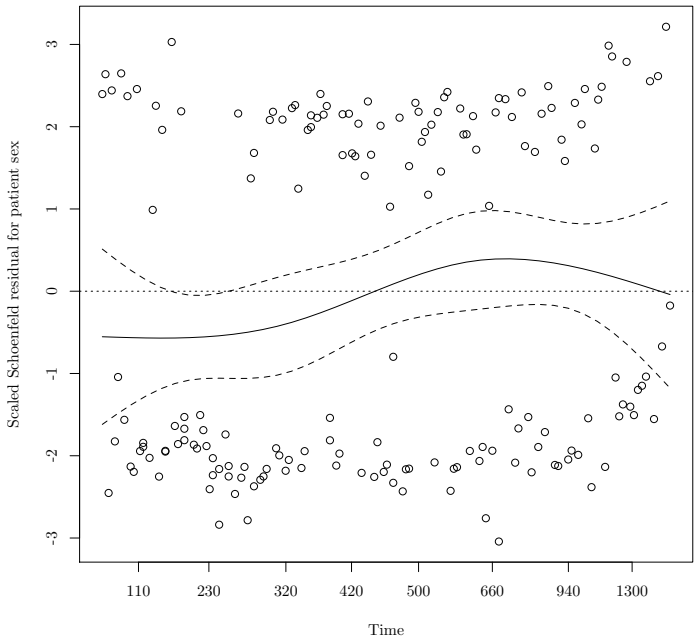


Figure 2.2: NSWPCN baseline hazard differs between patient sexes. A natural spline smooth of scaled Schoenfeld residuals for patient sex has a slope obviously differing from zero, suggesting that the baseline hazard forms differ between the two sexes, and that the combined data violates the PH assumption of Cox regression. Individual residuals are displayed as points, the natural spline smooth ($\mathrm{df} = 4$) as a solid line, and approximate ± 1 SE bounds as dashed lines.

Variable selection Genetic selection was used to identify a model with optimal Bayesian information criterion (BIC) from the set of all CPH models that use any combination of the expanded terms. Models with interactions between terms of up to degree two were considered, and a marginal term constraint was enforced, to ensure that interaction effects were only present in the model specification if the associated main effects were also. Stratification of baseline hazard by patient sex was used in all models. The identified optimal CPH model used three variables: tumour size (linear term only), S100A2 status, and S100A4 status, in addition to the sex stratum. This model was also identified by stepwise backward selection for optimal BIC, starting from the marginal CPH fit using all expanded terms. The final BIC-selected set of prognostic terms (tumour size linear term, S100A2 binary status, S100A4 binary status, and a patient sex stratum) was denoted the reduced term set.

Model CP1 A final prognostic CPH regression model was fit to the 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 = TODO). Model residuals were inspected to identify possible outlier patients, and assess overall fit stability. Deviance residuals indicated no egregious outlier patients, and DFBETAS indicated TODO influential patients with $|\text{DFBETAS}_{i,j}| > \frac{2}{\sqrt{n}}$, where $n = TODO$. TODO: Stuff on these outliers. 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 2.3).

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 can only be ranked by relative hazard, and absolute estimates of survival probabilities are unavailable. Although it is possible to approximate the baseline hazard in the Cox model, a more robust alternative is to use fully parametric models, in which the baseline hazard distribution is explicity 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 [14], was fit to the 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. The goodness of fit of GG1 was investigated by examination of residuals, and graphical assessment of prediction accuracy. Deviance and DFBETAS residuals indicated no extreme outliers or unduly influential samples, but comparisons between GG1 predictions and KM estimates of survival indicated that the GG1 baseline distribution could not accurately model survival in some

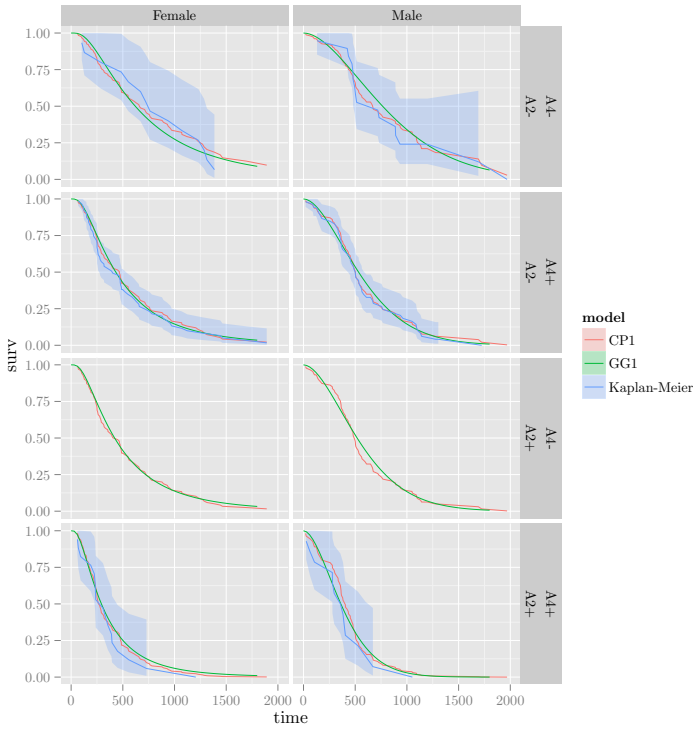


Figure 2.3: Model survival predictions agree with stratified KM estimates. KM estimates of survival probability for each combination of patient sex and biomarker status are shown as solid blue lines, with 95% confidence intervals indicated by blue ribbons. Estimates of survival probability generated by both models CP1 (red) and GG1 (green) broadly followed the form of the KM estimator, and lay within its bounds at all times, although the GG1 fit was relatively poor in some strata. 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.

cohort subsets, particularly the S100A2 negative male group (Figure 2.3)

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 [10] naturally and automatically model nonlinearity and arbitrary level interactions, and are tolerant of large numbers of irrelevant or collinear variables, albeit at the cost of very poor interpretability, and large data and computa-

tional requirements. Random forests have been adapted to model censored data [28], 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 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. For each model, predictor discrimination and overall accuracy over time were assessed by the Brier score for censored data [18], and the area under the curve (AUC) of the cumulative/dynamic time-dependent receiver operating characteristic (TD-ROC) [8]. The Brier score provides an overall measure of predictive performance, but it assesses both model discrimination and calibration, whereas it has been argued that high model discrimination is the far more relevant criterion in most clinical applications (TODO cite). Model discrimination alone can be measured by the TD-ROC-AUC, an extension of the familiar AUC concept to censored data, that is reflective of the use of prognostic models in clinical practice [25].

TODO: I'm approximating GG linpred, so why not RSF as well?

Model discrimination alone over time can be measured by the , which provides an assessment of prognostic model performance that is appropriate to the way these models are applied in clinical practice [9].

In addition, interpretation of the Brier score is complicated by the presence of an unknown model-independent 'data inseparability' term

can be dominated by an unknown model-independent 'data inseparability' term, making it insensitive to predictive differences between models [18]. A relative performance measure that is insensitive to

overall accuracy was measured by the integrated Brier score (IBS). All predictors were also compared against a no-information control predictor, formed as the KM estimate of the marginal survival function in the NSWPCN model building set. This no-information estimate, here denoted KM0, is the opti-

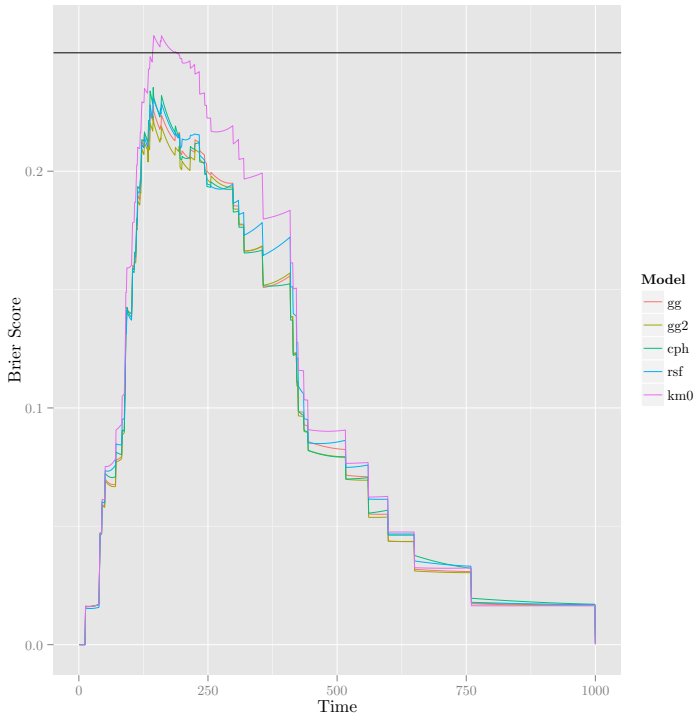


Figure 2.4: Brier score paths for candidate models on the holdout NSWPCN model test set. All models outperformed the no-information KM0 trace from approximately 100 days to 600 days post-diagnosis, and no strong differences were apparent between candidate models.

mal survival predictor in the absence of any prognostic information or patient stratification; the prognostic models developed here must at least outperform KM0 to be of any clinical utility.

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. For each model, prediction accuracy over time was assessed using ,

TODO: Swap out for AUC paths.

Traces of the Brier score over time indicated that all models produced survival estimates superior to KM0 from approximately 100 days to 600 days post-diagnosis (Figure 2.4). The performance of the models was similar, with RSF displaying slightly higher error rates at longer follow-up times, and models CP1 and GG1 demonstrating comparable performance. Bootstrapped differences in IBS between KM0 and each prognostic model further highlighted

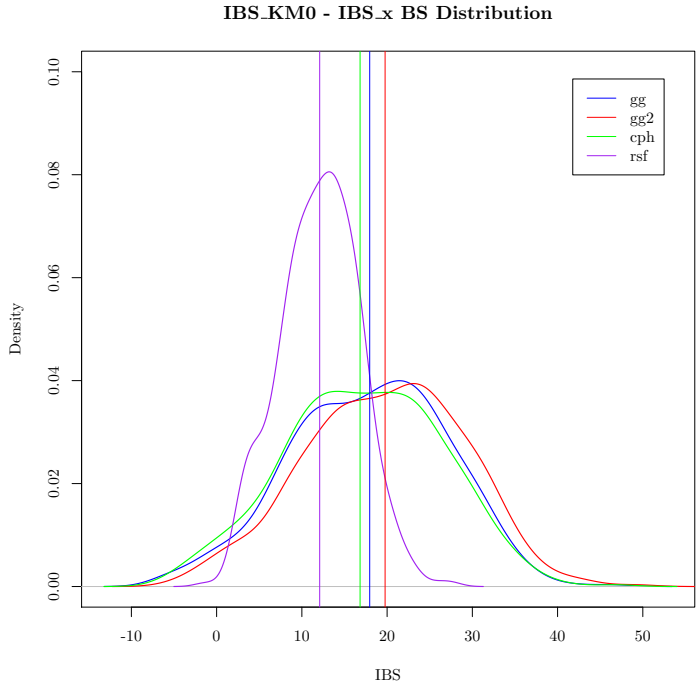


Figure 2.5: Distribution of bootstrapped differences in IBS between candidate models and null-model KM0. For each of 500 bootstrap draws from the NSWPCN holdout model test data, the difference between the IBS of a candidate model, and the IBS of KM0, was calculated as $\Delta\text{IBS}_{\text{Model}} = \text{IBS}_{\text{KM0}} - \text{IBS}_{\text{Model}}$. The smoothed distributions of $\Delta\text{IBS}_{\text{Model}}$ are shown here for each candidate model. All models predicted outcome significantly better than KM0, but beyond that had statistically similar performance.

that although all competing models were superior to KM0, the models displayed statistically similar predictive performance on the test set (Figure 2.5). As there was no substantial difference in performance between the prognostic models, the simplest model GG1 was selected for external validation.

External validation

Web tool

http://54.66.150.159:3277/

TODO: In app, show +/- margin curves, to guide surgeons as to benefit from aggressive surgery.

Cohort characteristics

Development of a preoperative prognostic model

Validation of the prognostic model

Intro here on disc & calib. Also describe cohorts briefly.

Discrimination

Calibration

Web tool TODO – better name

2.3 Discussion

2.4 Methods

Cohort recruitment and ethics

Chapter 3

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 pancreatic ductal adenocarcinoma (PDAC), the most common and aggressive form of pancreas cancer. As discussed in Chapter 2, 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 [13] has identified possible molecular subtypes within the previously homogenous 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 APGI cohort contains the largest group of clinically annotated PDAC samples, with accompanying gene expression (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 fundamen-

tal biology of the one common cancer that has, to date, been almost entirely refractory to all the tools of modern molecular medicine.

3.1 Introduction

Despite extensive research, PDAC remains a poorly-understood disease. Recent genomic profiling has revealed the genetic alterations that accompany the cancer [6], and a huge number of prognostic factors are known [24] (refer to Chapter 1 for further discussion on both points), 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 [42]. 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.

Historically, 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; notable early exceptions are for example [1, 32]. A number of factors are likely to have contributed to this reluctance: deconvolution methods require relatively large sets of high-quality measurements [37], 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 [50]) are only now reaching

maturity, with some areas of biology still underrepresented.

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 3.1). 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 3.1, 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.

A number of techniques from the field of matrix factorization have been applied to the GEX deconvolution problem, first principal component analysis (PCA) [1], then independent component analysis (ICA) [33], and more recently various forms of non-negative matrix factorization (NMF) (first used for GEX in [12]). A number of reports have highlighted the unsuitability of PCA for GEX deconvolution, and the relative superiority of ICA [32, 44, 51];

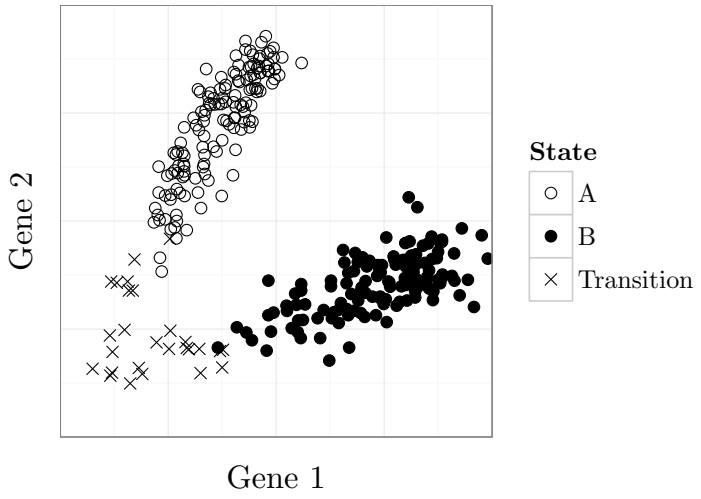
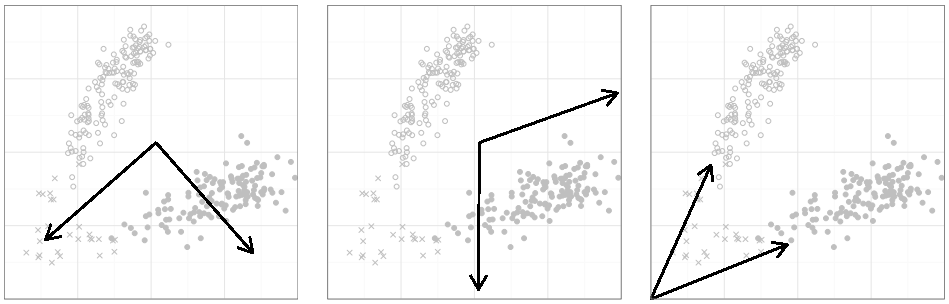


Figure 3.1: The gene deconvolution problem. Shown are the hypothetical expression levels of two genes across three biological states, where each point represents the gene expression of a single sample in one of the three biological states. State A (hollow circles) is characterised by gene 2 > gene 1; 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.

this is primarily due to the PCA requirement that metagenes be orthogonal [34], 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 3.2(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 3.2(b)), reducing its ultimate utility. NMF techniques have the potential to produce excellent GEX decompositions (Figure 3.2(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



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

Figure 3.2: 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.

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 [38]. The fraction of tumour cells that are actually cancerous varies by more than 10-fold between tumours [6], 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 [13], 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 [38], and that the examination of cancer cell expression in isolation ignores the likely important interplay between 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 [13]. This study exam-

ined 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 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.

3.2 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 factorization 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 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 3.1.

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 [17], while avoiding the well-known loss of statistical power that comes from discretising continuous expression values [43].

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%. This relatively high FDR was a consequence of the lenient selection parameters used, in an attempt to ensure that even genes for which expression was only weakly prognostic, were included.

Prognostic genes factorized 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 sensitive 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 factorization, long variant (SNMF/L) NMF algorithm.

The number of metagenes (factorization 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 3.3).

500 random restarts of rank 6 SNMF/L were then performed on the survival-associated gene matrix to yield the final factorization. The resultant clustering consensus matrix was stable (Figure 3.4), and the basis matrix W was reasonably sparse (Figure 3.5). 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 69.

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 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 regularization 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 3.6).

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 3.7, 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 3.8, 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.

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-approx on page 83.

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 [55] (LRT $P = 0.0149$) and The Cancer Genome Atlas (TCGA) paad (LRT $P = 0.0156$), but not in GSE21501 [49] (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 3.2, column a). This significant result reflected the ability of PARSE score to stratify patients into risk groups in a range of solid tumours, as illustrated in Figure 3.9.

Meta-PCNA is a 130-gene signature of cell proliferation that has been found to be generally prognostic in a number of cancer cohorts [54]. To exclude the possibility that PARSE score simply recapitulated the known meta-PCNA signature, I examined whether PARSE contributed additional prognostic in-

formation 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 3.2, column b).

Table 3.1: Characteristics of the full APGI patient cohort, and the homogenous 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)

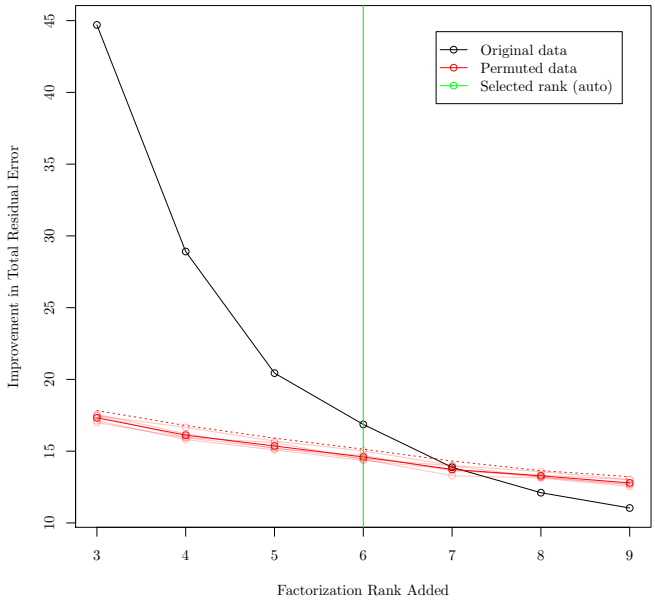


Figure 3.3: Automatic selection of factorization rank. SNMF/L was performed for varying ranks on either unpermuted data (black line) or data permuted within samples (red lines), and the improvement in total residual approximation error $\|A - WH\|_F$ calculated. The highest added rank for which the error improvement on unpermuted data exceeded that of permuted data plus two standard deviations (threshold shown by dotted red line) was the final selected rank (green line).

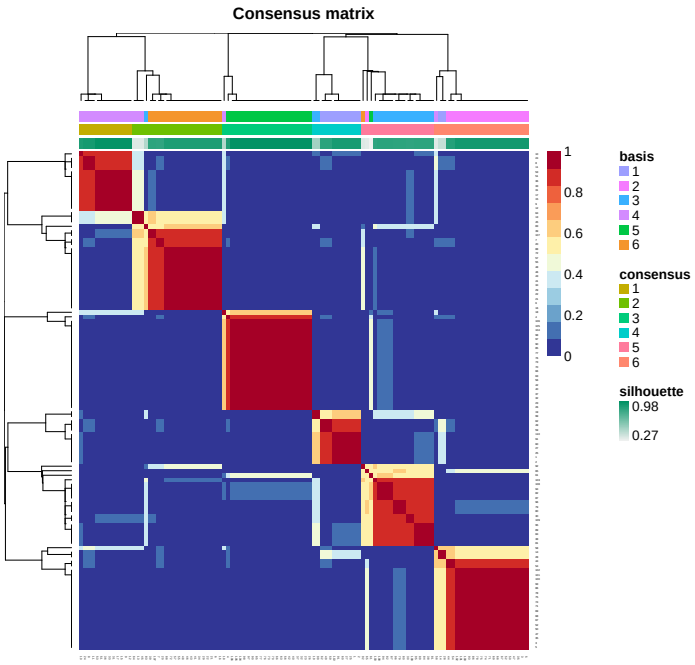


Figure 3.4: Clustering consensus matrix for the final rank-6 clustering. Colours indicate the stability of gene (in rows) and sample (in columns) clusters across random restarts of the factorization; at rank 6 this factorization was largely stable, with identical clusters assigned in all 500 random restarts to the majority of genes and samples.

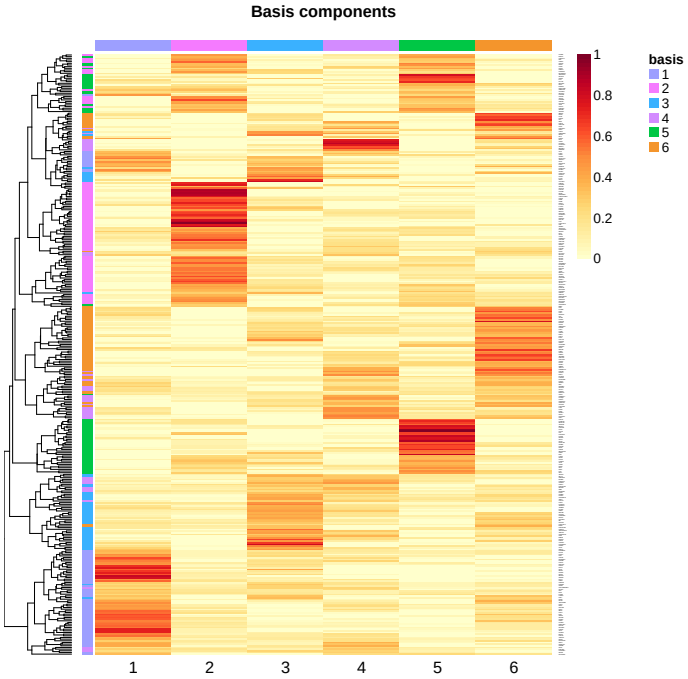


Figure 3.5: Basis matrix W of the final SNMF/L factorization. Rows represent genes, and columns metagenes, with cell colours proportional to the loading of a given gene on a given metagene. The loadings are sparse within rows, indicating that the metagenes are modular, each affecting the expression of largely distinct sets of target genes. A table of values of this basis matrix is available as app:sigs-w-matrix on page 69.

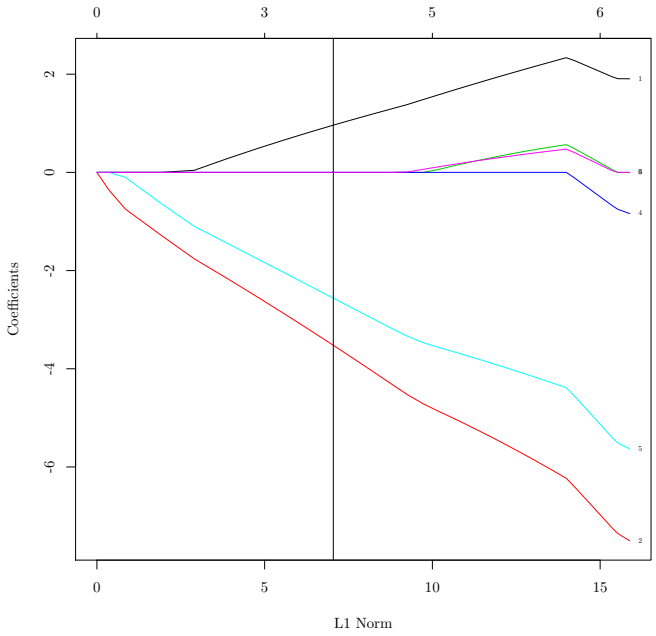


Figure 3.6: Coefficient vs penalty fit trajectories for the LASSO model predicting DSS from metagene expression. Each line represents the model coefficient for a metagene as the model is smoothly varied from a null model (L1 norm = 0), to a full unpenalised Cox fit (L1 norm $\approx$ 16). The vertical 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.

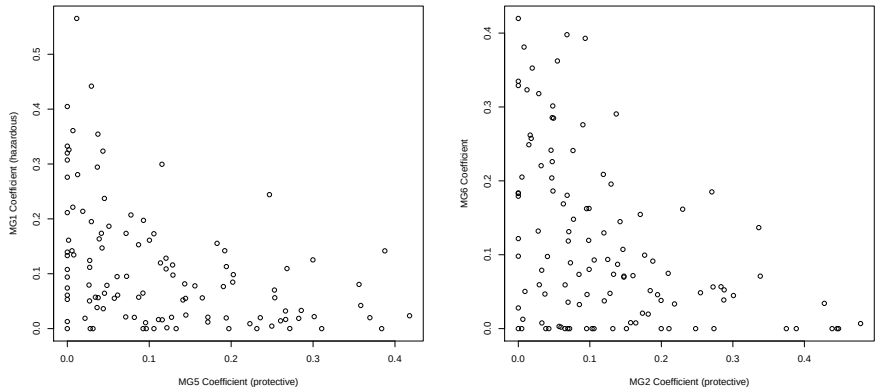


Figure 3.7: Prognostic metagenes form two axes of cell state. Metagene pairs MG1 and MG5, and MG2 and MG6, displayed mutually exclusive coefficient patterns in the APGI cohort, and could be combined to form just two axes of cell state.

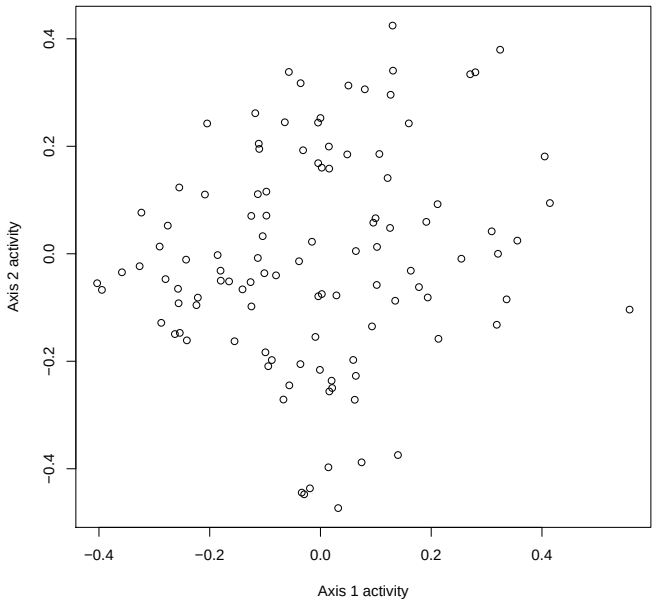
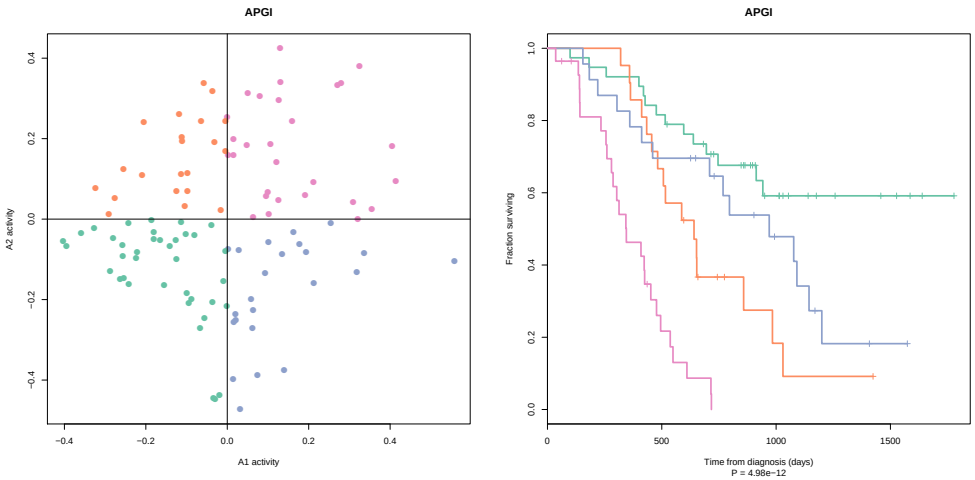


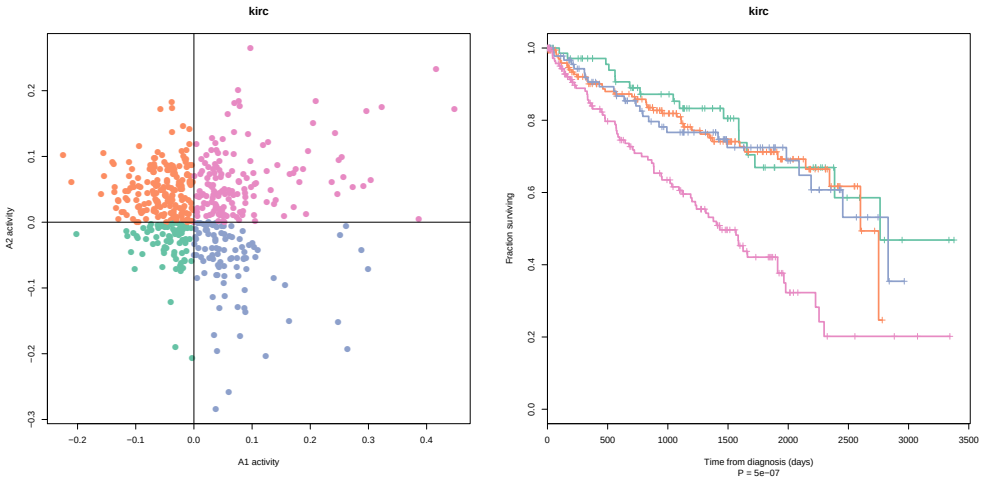
Figure 3.8: Prognostic axis signals are uncorrelated. Activity estimates of axes defined by highly correlated mutually exclusive metagene pairs (Axis A1 = MG1 - MG5, axis A2 = MG6 - MG2) were uncorrelated (Kendall τ test $P = 0.21$), indicating that these axis signals encoded orthogonal outcome-associated processes within tumours.

Table 3.2: 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	8.08E-3	0.0108
kirc	153	497	2.03E-12	2.89E-3
lgg	53	272	1.49E-5	7.85E-3
luad	106	431	8.34E-6	1.04E-4
lusc	117	395	0.9624	0.4110
ov	115	251	0.0238	0.0178

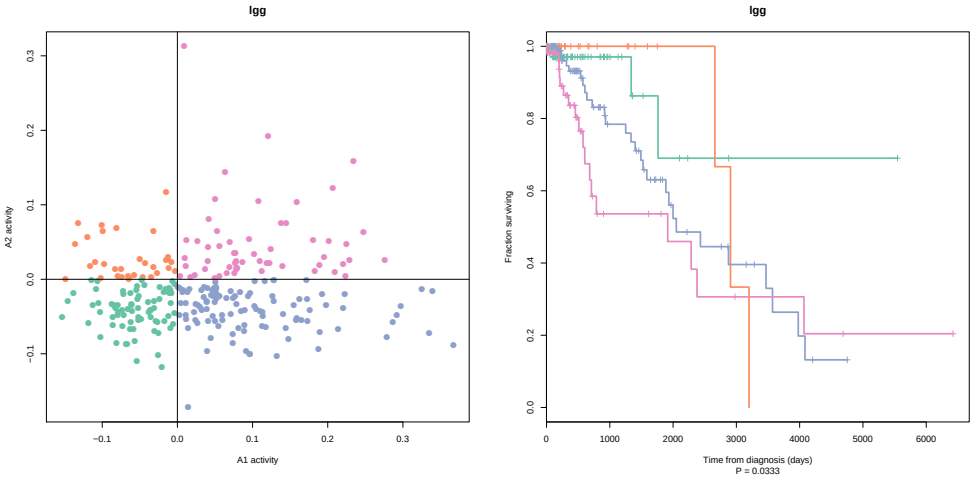


(a) APGI cohort

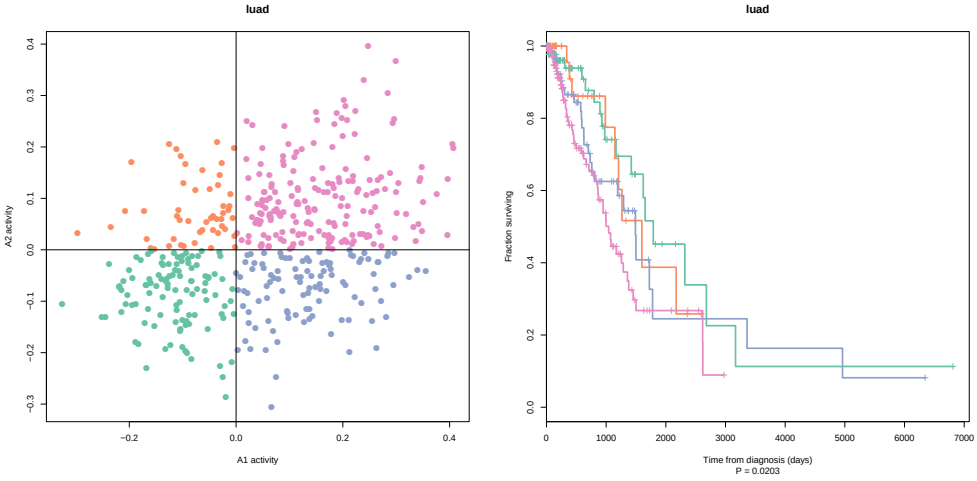


(b) TCGA kirc cohort

Figure 3.9: 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 3.9: (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) [50].

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 [47] [1] (Kendall's $\tau = 0.284$, $n = 110$, Table 3.3), 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 3.3).

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 [54] score (Kendall's $\tau = 0.663$, $n = 110$, Figure 3.10), a relationship supported by its close association to cell cycle-related MSigDB signatures (app:sigs-msigdb-corrs-axis1 on page 80).

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 3.3). 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 on which underlying axis is responsible. [2] 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 EMT (app:sigs-msigdb-corrs-axis2 on page 82). Prompted

[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.

[2] MP Fatal: Something for the discussion here – maybe that's why stroma is a slippery prognostic

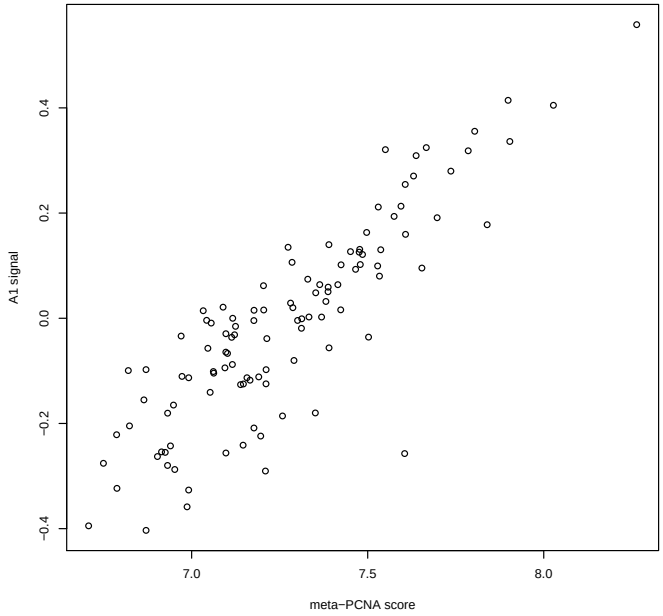


Figure 3.10: Axis A1 signal is closely associated with the meta-PCNA signature. A1 signal and meta-PCNA [54] scores were as evaluated on the APGI training set; Kendall's $\tau = 0.663$, $n = 110$, linear model $R^2 = 0.740$.

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 [20]. meta-EMT and A2 signals were strongly positively correlated (Kendall's $\tau = 0.568$, $n = 110$, linear model $R^2 = 0.557$, 3.11), 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 82), a proxy for expression of the acute phase response protein pentraxin 3.

3.3 Discussion

At the molecular level, the phenomenon of cancer has long been recognised as a composite of many processes [22], however the relative importance of each

Table 3.3: 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 signficant 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

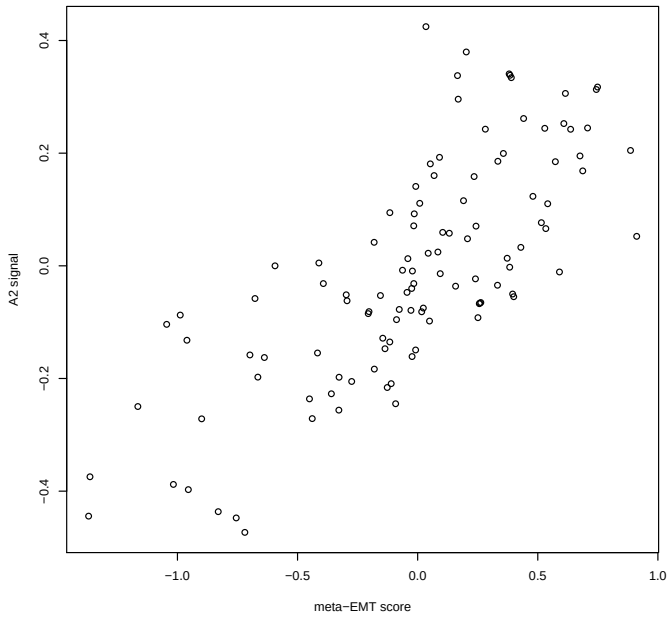


Figure 3.11: Axis A2 signal is closely associated with a signature of the EMT. A2 signal and meta-EMT [20] scores were as evaluated on the APGI training set; Kendall's $\tau = 0.568$, $n = 110$, linear model $R^2 = 0.557$.

process to a particular type of cancer has been largely uncertain. In pancreas cancer, a huge number of individual biomarkers are known [24], and attempts have been made to stratify cancers into empirical molecular subtypes [13], 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.

The thesis underlying this work was that fundamental disease processes control the survival of patients with resected PDAC, and that these processes can be identified from patterns of gene expression in patient tumours. To test this thesis, a global biology-agnostic approach was used to exhaustively identify expression modules strongly associated with outcome, and these modules were then linked to biological processes. Two independent outcome-associated modules were identified, and comparisons with published gene signatures indicated that one of these modules was likely measuring proliferative activity,

and the other the EMT, in the tumours.

Proliferation and the EMT are two of the ten major hallmarks of cancer [22], and so it is unsurprising that they are so closely associated with patient survival: the very definition of cancer involves uncontrolled and unceasing cell proliferation (TODO cite); and the EMT is a major enabling step in metastasis, the process by which most cancers are ultimately lethal (TODO cite). What *is* surprising is that the majority of hallmarks do *not* appear to be strongly associated with survival 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 [36].

Cancer is fundamentally a proliferative process: it is through inappropriate and continued proliferation, and the consequent destruction of normal tissues and disruption of homeostasis, that cancer progressively overwhelms the body. In keeping with this, the meta-PCNA signature of cell proliferation is prognostic in a wide range of cancers [54] – actually it isn't tested as such in Venet. TODO: should probably do what Venet suggests and pull some random genes as a test. * Something on prolif. Obviously a central aspect of cancer. Discuss meta-PCNA briefly and its pan-can applicability. A1/A2 also have pan-can prognostic ability, but interestingly it's subtly different from PCNA – eg kirc, where double pos is needed for poor prognosis.

* EMT. Huge topic, but essentially the idea is that this is a hallmark of metastasis. Fold in with the nomogram chapter, with A2/A4 as markers of occult met, which is a major mechanism of early recurrence for resected PDAC. How do A2/A4 compare with Ax2? I may need to put a bit of results in for this, but from memory they were poor proxies. Ax2 could be a superior indicator of met ability to A2/A4.

* Brief consideration of potential limitations – these are the big players, but can't exclude everything. Good examples are stroma and immune, which may also be playing a part.

* What about the stroma? The inverse corrs with the sigs is interesting. This links in with the current confusing lit on the influence of stroma on PC. Ie it's bad (http://www.ncbi.nlm.nih.gov/pubmed/22521638) vs it's good (http://www.ncbi.nlm.nih.gov/pubmed/24856585).

* Talk about validation. That bloody dataset. Link in with pan-cancer. I've talked about prolif. and EMT, but looks like they're also important in

other solid tumours. Definitely warrants follow-up.

* This technique is general – can be applied to anything. Might be a good way to assess prognostic sigs across all cancers. Another approach might be to take hallmarks and test, but the advantage of this is that it is totally unbiased. Disadvantage is complexity.

* Techniques. This approach was taken for its low bias, trying to capture as much biology as possible. Seems to indicate that Ax1/Ax2 are smooth axes, not the discrete groupings that are often sought. But this clustering is probably artefactual, when there are very many factors combining to yield a phenotype. It highlights the follow of clustering – look for clusters, and ye will find. The approach taken was global and inclusive, with the ability to link to bio., but unfortunately this led to very large sigs.

* More general application. Problem is the large cardinality of the sigs. These were generated more for understanding of biol. than ease of application. But it's possible that they could be cut down, see eg the approx. appendix. Similar approaches could maybe cut them down further, more work to do here. If trimmed down sufficiently, could even become replacements for A2/A4 etc.

* Wrap-up. Thesis came up +ve, and two major players found. This has given new understanding of PDAC survival, and possibly surv in other solid tumours too. It's also illustrated a technique that might be applied to pan-cancer, to assess sigs on a truly global scale. These sigs are primarily tools for better understanding of the processes in the cancers, but immediately illustrate potential avenues for therapy.

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 factorized 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

Although it is impossible to conclusively exclude the effect of any other processes, the global unbiased nature of the analysis undertaken here establishes that proliferation and EMT are the predominant determinants of survival in PDAC

3.4 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 [6]. 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 [6]. 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 Illumina data (IDAT) scan files. All kit and microarray procedures were performed as per the manufacturer's instructions.

Data preprocessing

Microarray quality control and normalization 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 normalized as one batch using the `lumi` package. Normalization proceeded serially as: RMA-like background subtraction (`lumiB` method `"bgAdjust.affy"`), variance stabilizing transform (VST) (`lumiT` method `"vst"`), and quantile normalization (`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 normalized 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 [17], with a CPSS wrapper to reduce the false positive rate [46]. FAST statistics for time from diagnosis to DSD were calculated using R package `ahaz` on standardized 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 factorization

The gene × patient expression matrix of outcome-associated genes was decomposed into metagenes by the SNMF/L procedure of [29], 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 factorizations 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 factorization, the VST-transformed expression matrix A was approximately linearized by elementwise exponentiation, $a_{i,j} \leftarrow 2^{a_{i,j}}$. To reduce the influence of large variations in baseline expression on the factorization, 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.

Factorization rank was estimated following [16]: 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 factorizations 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 factorization 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 factorization 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 factorization 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 factorization 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 factorization.

The SNMF/L algorithm requires parameters α and η to control regularization; for all factorizations $\alpha = 0.01$, and $\eta = \max(A)$.[3] The default convergence criteria of the `NMF` package were used.

SNMF/L may not necessarily find a global optimum factorization; to address this, multiple random initializations 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 factorization; examination of approximation error distributions for these repeated runs indicated that these values were conservative, and factorizations 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 genes are listed in app:sigs-w-matrix on page 69), 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 factorization. Genes for which no expression data were avail-

[3] Note that this parameter α is denoted β in the R `NMF` package; I use the symbol α here for consistency with [29]

able (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 × sample matrix A' with rows matching the gene × metagene basis matrix W from SNMF/L. The metagene × 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 optimization 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 69.

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 × A1 activity + 1.548 × A2 activity.

Although not used in this work, a simplified procedure for the approximate calculation of PARSE scores was also developed; see app:sigs-parse-approx on page 83 for details.

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 [50] were estimated on the APGI cohort using a modification of the GSVA method [23]. GSVA with default settings was used to estimate expression scores for all MSigDB gene sets in the full 13,000 $\times$ 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 coordinated 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 $d_{i,j} = \frac{1}{2}(1 - \mathrm{cor}(s_i, s_j))$, 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 [54]. To estimate meta-ECM scores, log-scale GEX data were median centered, and then median values across samples were calculated for all genes in the two lists of [20] 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 characterization

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 3.4. Numeric variables were tested for assocation 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 [27]. 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 3.5. Absolute correlations of greater than 0.5 were deemed substantive and reported for further characterisation.

Attribution of work

Data for the APGI discovery cohort were generated as part of the APGI project, under the umbrella of the International Cancer Genome Consortium

Table 3.4: 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

(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 IDAT files through to analysis planning, signature development, testing, and interpretation, were performed solely by me.

Table 3.5: 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 4

Comparative genomics

Outline ideas:

- Introduction / overview:
  - The use of models in PC (very brief)
  - Specific models used in PC, with strong focus on the most common (KPC), and derivates. Cover ease-of-use briefly.
  - Current knowledge re: how appropriate the models are. Consider histology, genetic features, disease progress (incl. metastatic potential), response to therapy. Highlight gap in genetic information, and relevance to response to therapy.
  - Brief overview of known genetic features of human disease. Raise possibility of subtypes.
  - Wrap-up with overview of project:
    1. Collect matched tumour-normal DNA from a range of GEMMs.
    2. Sequence and determine conserved model-specific and general patterns of somatic mutation.
    3. Compare observed patterns to human disease.
       * Are genetic features of human disease recapitulated generally in the models?
       * Does a single model match the genetic features of human disease much better than the others?
       * Do specific models serve as simulations of certain subtypes of human disease?

- Overall thesis for this work:
  Matching patterns of genetic alterations in mouse models of pancreatic cancer to those seen in human disease can inform researchers as to which models are generally best, and which best match specific patient types.
  Sub-theses:
  * The patterns of mutations seen in common mouse models of pancreatic cancer match those consistently seen in human disease.
  * Different mouse models possess different mutation spectra, and models may be close fits to specific genetic subtypes of patients.

- Results
  1. Somatic SNV and indels
  2. CNV and LOH

- Conclusion

4.1 Methods

Models

Sample Origin and Processing

Sequencing

QC

Mapping

For initial mapping, all lanes were processed independently. SHRiMP was used to map colourspace reads to the mm10 genome using 'all-contigs' and 'single-best-mapping' options. Unpaired reads in the source fastq files were mapped as single reads; paired reads were mapped with pair mode 'opp-in', and a per-fastq insert size distribution estimated from a normal distribution fit to insert sizes of the first 10,000 reads. Likely duplicate reads were marked using Picard MarkDuplicates on each individual lane binary sequence alignment / map file (BAM), using an optical duplicate pixel distance parameter of 10.

Lane BAMs were progressively merged: first, duplicate lane BAMs for a given mouse and sample type (tumour or normal) were combined, then tumour and normal BAMs for a given mouse, and finally combined tumour-normal BAMs for all mice. Prior to each level of merging, the Genome analysis toolkit (GATK) was used to separately perform local alignment and base quality score recalibration (LA-BQSR) on each input BAM. Finally, the full experiment BAM file was recalibrated with LA-BQSR, and then split by mouse and sample type for analysis, yielding 62 paired tumour and normal final BAMs.

Somatic SNV and Indel Detection

muTect and Strelka were used separately to detect somatic single nucleotide variants (SNVs) and insertion / deletion events (indels) in individual mouse tumour and normal BAMs. muTect was supplied default parameters; Strelka used the parameter settings given in listing 4.1; these are the default parameters as recommended for use with the BWA mapper, with the exception that in this work isSkipDepthFilters was set to 1.

Listing 4.1: Strelka configuration file used for SNV / indel detection

```
[user]
isSkipDepthFilters = 1
maxInputDepth = 10000
depthFilterMultiple = 3.0
snvMaxFilteredBasecallFrac = 0.4
snvMaxSpanningDeletionFrac = 0.75
indelMaxRefRepeat = 8
indelMaxWindowFilteredBasecallFrac = 0.3
indelMaxIntHpolLength = 14
ssnvPrior = 0.000001
sindelPrior = 0.000001
ssnvNoise = 0.0000005
sindelNoise = 0.000001
ssnvNoiseStrandBiasFrac = 0.5
minTier1Mapq = 20
minTier2Mapq = 5
ssnvQuality_LowerBound = 15
sindelQuality_LowerBound = 30
isWriteRealignedBam = 0
binSize = 25000000
```

CNV and LOH Detection

Overview:

- Very brief background of CNV and LOH in tumours, and the possibility of detection from NGS data. Maybe pull in the hallmarks paper, or perhaps specific PC / GEMM examples.
- Brief overview of existing techniques and why unsuited?
  - CNV:
    - Exome pulldown complication
    - Ill-posed nature of problem
    - Human-specific methods
    - Outbred population-specific methods
  - LOH:
    - That Bayesian thing. Unfortunately affected by CNV, which is unknown.

Loss of heterozygosity at individual loci

This work took a simple approach to identify loci with significant evidence of loss of heterozygosity (LOH) in a tumour sample: locate high-confidence heterozygous loci in matched normal DNA, and then test only these heterozygous loci for a significant change in allelic fraction between matched tumour and normal samples. In regions of the genome with ploidy $2n$ and below, such allelic imbalance is indicative of LOH, even in the presence of unknown levels of diploid genome contamination.

Identifying heterozygous loci in normal DNA High-confidence heterozygous loci in normal DNA were identified by comparing posterior genotype likelihoods using a Bayesian model comparison (BMC) approach. BMC is a procedure for deciding which of two competing models is better favoured by the observed data; here the two models are, for a given locus: 'the locus is homozygous' (model HOM), and 'the locus is heterozygous' (model HET). The likelihoods of these two models (assessed on the reads observed at a locus) can be used to calculate a Bayes factor, which encodes which of the two models is better supported by the data at that locus, and how strongly. More

formally, we partition the ten possible diploid genotypes at a locus into two classes, Hom and Het:

$$Hom = \{AA, CC, GG, TT\} \tag{4.1}$$

$$Het = \{AC, AG, AT, CG, CT, GT\} \tag{4.2}$$

The two models, HOM and HET, may be written

$$HOM : G \in Hom \tag{4.3}$$

$$HET : G \in Het \tag{4.4}$$

where G is the true genotype at the locus. The Bayes factor K comparing HOM and HET is then

$$K = \frac{\mathcal{L}(HET)}{\mathcal{L}(HOM)} \tag{4.5}$$

$$= \frac{Pr(D|G \in Het)}{Pr(D|G \in Hom)} \tag{4.6}$$

$$= \frac{\sum_{g \in Het} Pr(D|G = g) Pr(G = g|G \in Het)}{\sum_{g \in Hom} Pr(D|G = g) Pr(G = g|G \in Hom)} \tag{4.7}$$

with D being the reads at the locus. We make the simplifying assumption that all genotypes in each of Hom and Het are equally likely, so that all $Pr(G = g|G \in X) = \frac{1}{\|X\|}$ for $X \in \{Hom, Het\}$. Then

$$K = \frac{\frac{1}{\|Het\|} \sum_{g \in Het} Pr(D|G = g)}{\frac{1}{\|Hom\|} \sum_{g \in Hom} Pr(D|G = g)} \tag{4.8}$$

$$= \frac{\frac{1}{\|Het\|} \sum_{g \in Het} \mathcal{L}(G = g|D)}{\frac{1}{\|Hom\|} \sum_{g \in Hom} \mathcal{L}(G = g|D)} \tag{4.9}$$

encodes the weight of evidence for the observed read data D favouring a locus being heterozygous over homozygous, and a value exceeding a given threshold is taken as significant evidence that the locus under consideration is heterozygous.

An implementation of the above heterozygous locus detection method is given in algorithm 1. The input posterior genotype likelihoods $\mathcal{L}(G = g|D)$ are supplied by `samtools mpileup -q 20 -Q 20 -v -u` operating on per-mouse normal sample BAMs, and the minimum value of K for a locus to be called as heterozygous is $\exp(minscore)$. Two additional filters are also employed in the algorithm: a locus is *not* reported as heterozygous if either the total read

depth at the locus is less than *mindepth*, or if the difference in samtools-supplied log likelihood between the top two genotypes is less than *mindelta* nats. The latter filter is used to exclude any problem loci with an apparent triallelic state. [1]

Identifying tumour LOH at known normal heterozygous loci Given a set of loci that are known to be heterozygous with high confidence in the normal DNA of a given mouse, it is straightforward to test for LOH in the tumour DNA of the same mouse, provided the tumour ploidy at the locus is $2n$ or less. Considering only a single heterozygous locus, reads from a normal DNA sample will predominantly be for the two bases constituting the heterozygous genotype, possibly with a small number of reads from other bases due to sequencing or mapping errors. The number of reads for the two genotype bases may be quite different, as the exome capture processing step may favour one allele over the other, and lead to allelic bias in the observed read fractions. However, under the null hypothesis of no LOH and no mutation at the locus in the tumour DNA, if the tumour ploidy at the locus is $2n$ or less, then the relative proportions of reads for the two genotype bases should be the same in both the tumour and the normal samples. This null hypothesis can be tested using a contingency test comparing two binomial proportions; for this work I used the two sided Z-pooled test as implemented in R package `Exact`.

In the general case with potential normal cell contamination of the tumour sample, it is not possible to use allelic imbalance as an indicator of LOH if the local copy number exceeds two. For example, in the triploid case, a LOH haplotype AAA, and a non-LOH haplotype AAB, both exhibit allelic imbalance. For this reason, allelic imbalance calls from the above test must be interpreted in the context of local copy number variation (CNV) estimates from the next procedure, and LOH calls only made if allelic imbalance is detected in regions of copy number $2n$ or less.

Copy number variation at individual loci

Problem description Considering a single locus, either a single nucleotide or a contiguous stretch of DNA, the expected number of reads from a sequencing experiment that map to that locus is proportional to the copy number of

[1] MP Fatal: give instantiation values for the algo somewhere

Data: Total sequence depth at the locus D, minimum depth for call *mindepth*, list of alternate alleles A, list of Phred-scaled genotype likelihoods L, minimum likelihood difference in nats between top two genotypes *mindelta*, minimum Bayes factor in nats for heterozygous to be called over homozygous *minscore*.

Result: A boolean: true if the locus is called heterozygous, false if it is not.

begin

  if $D \leq mindepth$ **then**

    return false;

  end

  `// Convert Phred-scaled likelihoods to nats`

  for $i \leftarrow 1$ **to** $\|L\|$ **do**

    $L_i \longleftarrow -\frac{1}{10}\log(10)L_i$;

  end

  `// Ensure the likelihood difference between the two most likely genotypes is at least` *mindelta*`.`

  $L^* \longleftarrow L$ sorted in decreasing order;

  if $L_1^* - L_2^* \leq mindelta$ **then**

    return false;

  end

  `// Calculate combined likelihoods for heterozygous and homozygous genotypes`

  switch $\|A\|$ **do**

    case 2

      $L_{het} \longleftarrow L_2$;

      $L_{hom} \longleftarrow \log\left(\frac{1}{2}\sum_{i\in\{1,3\}}\exp(L_i)\right)$;

    end

    case 3

      $L_{het} \longleftarrow \log\left(\frac{1}{6}\sum_{i\in\{2,4,5,7,8,9\}}\exp(L_i)\right)$;

      $L_{hom} \longleftarrow \log\left(\frac{1}{4}\sum_{i\in\{1,3,6,10\}}\exp(L_i)\right)$;

    end

    case *default*

      return false;

    end

  endsw

  `// Compute the Bayes factor for heterozygous vs homozygous, and compare to the threshold`

  if $L_{het} - L_{hom} \leq minscore$ **then**

    return false;

  end

  return true;

end

Algorithm 1: Determine if a locus is heterozygous

the locus in the DNA input for sequencing. Based on this relationship it is – in principle – possible to estimate copy number from sequencing data, however a number of complicating factors are present, related to sequence 'mappability', exon capture affinity, sample contamination, and problem indeterminancy.

There are many regions in mammalian genomes for which it is challenging to map reads. These regions may be either poorly characterised themselves in the reference genome, or may be sufficiently like other parts of the genome for an unambiguous mapping to be impossible with the short and error-prone reads produced by next-generation sequencing (NGS) technologies. Most processing pipelines discard such ambiguous reads, with the net effect that difficult-to-map regions of the genome have much lower read depth than would be expected based on the quantity of DNA for those regions present as input to the sequencing procedure. Copy-number analysis techniques need to take this 'mappability' bias into account, or regions of reference DNA that are challenging to map may falsely be reported to undergo copy number loss.

A similar effect to 'mappability' bias is additionally present in datasets generated by exome sequencing. The process of exome enrichment necessarily favours certain regions of the reference genome (hopefully, the exome), over others. This enrichment is always imperfect: some non-target DNA will persist through the procedure, and not all target regions will be retained to the same degree. The ultimate effect of the exome enrichment procedure is to introduce an additional per-locus bias, 'exon capture affinity' that requires correction before copy number calls can be made. Unlike for 'mappability bias', ignoring exon capture affinity bias can lead to either false copy number loss or false copy number gain calls.

Contamination of tumour DNA is a universal problem in solid tumour sequencing. This contamination may be with non-cancerous diploid DNA, or alternate cancer genotypes present in the same sample, or both. In the case of CNV estimation based on read depth, the presence of contaminating diploid DNA causes a shrinkage of the observed CNV profile towards that of diploid cells, and reduces the signal-to-noise ratio (SNR) of the copy number estimates. CNV callers aware of this effect must take this effect into account in their calls, and may also be required to estimate the fraction of contaminating normal DNA. In tumour samples containing multiple tumour genotypes, with varying locus copy numbers, CNV estimates are for the mean copy number of the genotypes, weighted by their prevalence in the sample. In such cases,

deconvolution of the signal into its component genotypes based on a single sample of the tumour is impossible without the benefit of additional external information.

Ultimately, without knowledge of the number of cells input into the sequencing procedure, CNV estimation from NGS data is a fundamentally indeterminate problem. This is easily seen by considering the case of a hypothetical fully haploid tumour: the read counts of all loci will be completely consistent with those of a normal diploid sample. Without observing that the quantity of DNA present per input tumour cell is half that of a diploid cell, the haploid tumour and diploid normal samples would be completely indistinguishable. Information on the number of cells used for extraction is very seldom available, and so in almost all cases additional assumptions are required to assign absolute copy number to NGS read depth data.

Taking all the above complications into account, I developed an organism-agnostic CNV detection procedure for exome or whole genome sequencing (WGS) data that uses NGS read depths as input.

CNV model and test development The mathematical setup of the procedure is as follows. We reserve upper case variable symbols for random variables, and use lower case equivalents for observed values of these random variables. Consider m disjoint loci on the reference genome; these loci may be individual base pairs or contiguous regions. For a single matched tumour-normal sample pair, let the number of reads that were mapped to locus $i \in \{1 \dots m\}$ be n_i for the normal sample, and t_i for the tumour sample. Denote the total read depths at all examined loci as d_N and d_T, $d_N = \sum_{i=1}^{m} n_i$, $d_T = \sum_{i=1}^{m} t_i$. To consider normal DNA contamination effects, we suppose that the tumour sample is actually a mixture of normal cell diploid DNA, and cancer cell DNA, where the fraction of cancer cells in the sample is the unknown quantity $f \in (0, 1]$. Loci are subject to differential exome enrichment, locus size, and mapping biases, which are combined into the single per-locus quantity b_i, such that $\langle N_i \rangle \propto b_i$, $\langle T_i \rangle \propto b_i$, and $\sum_{i=1}^{m} b_i = 1$.

We model the process of reads in NGS as a Bernoulli scheme, and use the weak dependence between read depths at different sites to derive a per-locus Poisson approximation. In this model the sequencer has a fixed s total physical sites available for sequencing; in the SOLiD 4 system these sites correspond to positions on the sequencing slide. Some of these sites yield observed sequence

that is then mapped and used to estimate read depth, however many of them do not produce sequence reads, either because they are never populated with DNA, or because they fail low-level quality checks. We suppose that these failed sites occur independent of the DNA sequence, and at a rate of r_F among all available sites. Then, a given physical sequencing site can either fail to yield sequence, with probability r_F, or it can produce observed sequence for one of m loci, each at probablity $(1 - r_F)b_i$, for $i \in \{1 \dots m\}$. This per-site categorical distribution, when sampled for each of s independent sites, results in a multinomial distribution on read depths,

$$(N_F, N_1, \dots, N_m) \sim Multi\,(s, (r_F, (1 - r_F)b_1, \dots, (1 - r_F)b_m)) \tag{4.10}$$

where N_F is the number of failed sites (not observed), and N_i is the number of reads observed for locus i. The multinomial distribution induces a negative dependency on the number of reads observed at different loci, as the total read count s is fixed. However, for m large, or site failure rate r_F large[39], these negative dependencies are small, and

$$N_i \dot{\sim} Pois\,(s(1 - r_F)b_i) \tag{4.11}$$

The quantity $s(1 - r_F) = \langle D_N \rangle$ is unknown, and we approximate it with the observed value d_N. Therefore, the final approximate model for read depth in the normal sample is

$$N_i \dot{\sim} Pois\,(d_N b_i) \tag{4.12}$$

For the tumour sample, the expression for the Poisson rate parameter is more complex than in the normal case, as locus copy number is no longer assumed constant. Ignoring for the moment the possibility of diploid DNA contamination in the tumour sample (i.e. let $f = 1$), and following the derivation used in the normal case, we find that the number of reads at locus i in pure tumour sample is distributed as

$$T_i \overset{f=1}{\dot{\sim}} Pois\,(d_T b_i c_i k_{pure}) \tag{4.13}$$

where c_i is the copy number of locus i in the tumour DNA, relative to diploid cells. $k_{pure} = 1/\sum_j b_j c_j$ is a normalization factor that ensures $\langle \sum T_i \rangle = d_T$. Now considering possible diploid DNA contamination, if tumour cells are

present at a fraction f, with the remainder diploid cells, the tumour locus read count is distributed as

$$T_i \dot{\sim} Pois\left(kd_T b_i \left(1 + f\left(c_i - 1\right)\right)\right) \tag{4.14}$$

Here k is no longer a simple normalization factor like k_{pure}, but is a value that involves sample purity and cancer cell DNA content. [2]

The variable k is more than a convenient normalization constant: it encodes the signal expected of diploid loci in the tumour cells, and therefore controls the absolute copy numbers called by the procedure. To see this, observe that the pure tumour ploidy signal is $c_i k$, and therefore that tumour ploidy relative to $2n$, c_i, is completely confounded with k. As noted earlier, without knowing the number of input cells in the tumour sample, it is impossible to determine absolute ploidy from NGS depth data, and so there is no way to conclusively determine the correct value for k. In this work I used the heuristic that the most common ploidy in a tumour cell should be diploid, and therefore selected values for k to ensure that the most common CNV call would be diploid (ie No CNV). This heuristic will almost certainly be wrong in cases, but is necessary given the fundamentally indeterminate nature of the CNV problem. Interpretation of the results of this CNV calling procedure must take into account the possibility that k is mis-specified, and that all CNV calls should be shifted appropriately.

Given the above approximate Poisson distributions for normal and tumour read depths as a function of locus ploidy, I developed a per-locus CNV test based on a ratio test for two Poisson-distributed random variables. Let R_i be the ratio of the read appearance rates at locus i in tumour and normal samples,

$$R_i = \frac{kD_T b_i \left(1 + f\left(c_i - 1\right)\right)}{D_N b_i} \tag{4.15}$$

$$= \frac{D_T}{D_N}\left(k\left(1 + f\left(c_i - 1\right)\right)\right) \tag{4.16}$$

Then, the null hypothesis of no CNV at locus i, $H_0 : c_i = 1$, is equivalent to a hypothesis on R_i,

$$H_0 : R_i = \frac{D_T}{D_N} k \tag{4.17}$$

[2] MP Fatal: Add the derivation in somewhere – perhaps an appendix. It's a pain in the arse so probs want to avoid the main text.

We test this hypothesis on R_i using the W_5 statistic of [21],

$$W_5(X_0, X_1) = \frac{2\left(\sqrt{X_0 + 3/8} - \sqrt{r_{H0}\left(X_1 + 3/8\right)}\right)}{\sqrt{1 + r_{H0}}} \tag{4.18}$$

where $r_{H0} = \frac{d_T}{d_N}\hat{k}$. This statistic is asymptotically normally distributed, so the one-sided copy number gain P-value ($H_1 : c_i > 1$) is

$$p_{gain} = 1 - \Phi\left(w_5(t_i, n_i)\right) \tag{4.19}$$

where w_5 is the observed value of the statistic W_5, and Φ is the cumulative distribution function of the standard normal distribution. W_5 is symmetric, so the one-sided P-value for copy number loss is

$$p_{loss} = \Phi\left(w_5(t_i, n_i)\right) \tag{4.20}$$

and the combined two-sided P-value for CNV at locus i is

$$p_{CNV} = \begin{cases} 2p_{loss} & \text{if } t_i/n_i < r_{H0} \\ 2p_{gain} & \text{if } t_i/n_i \geq r_{H0} \end{cases} \tag{4.21}$$

CNV detection procedure Pseudocode for the implementation of per-locus CNV detection is given in algorithm 2. [3]

Combining calls from adjacent loci

CNV and LOH are broad genomic events that typically affect many adjacent loci together, yet the methods presented in the preceding sections consider each locus in isolation. By examining loci separately, we disregard important information: that the CNV and LOH status of nearby loci is strongly correlated. Intuitively, by leveraging these local correlations and combining results from neighbouring loci, we can achieve more accurate CNV and LOH detection than if each locus were considered alone.

A number of approaches could be used to smooth LOH and CNV calls and share information between neighbouring loci; in this work I chose the hidden Markov model (HMM) formalism and extended the Pounds-Morris FDR estimator[41] to the locality-sensitive case. The Pounds-Morris procedure fits the observed distribution of test P-values to a mixture of Uniform and Beta distributions. The Uniform distribution models the expected distribution of

[3] MP Fatal: Add specific value of mindepth used

Data: An m-vector of normal locus read depths $\boldsymbol{n}$, an m-vector of tumour locus read depths $\boldsymbol{t}$, minimum normal sample depth $mindepth$.
Result: An m-vector of floats: for each locus, the one-sided P-value for CNV loss at that locus, $\boldsymbol{p_{loss}}$.
begin
  $d_N \longleftarrow \sum_{i=1}^{m} n_i$;
  $d_T \longleftarrow \sum_{i=1}^{m} t_i$;
  // Estimate k so that the modal ploidy signal will be called as diploid
  $s \longleftarrow \{(t_i/d_T) \div (n_i/d_N) : i \in \{1 \dots m\} \wedge n_i \geq mindepth\}$;
  $\hat{S} \longleftarrow KDE(s)$;
  $\hat{k} \longleftarrow mode(\hat{S})$;
  // Calculate P-values. W_5 and Φ are as defined in the text.
  $r_{H0} \longleftarrow \frac{d_T}{d_N}\hat{k}$;
  $\boldsymbol{p} \longleftarrow$ m-vector of NAs;
  for $i \leftarrow 1$ **to** m **do**
    if $n_i \geq mindepth$ **then**
      $p_i \longleftarrow \Phi\left(W_5(t_i, n_i)\right)$;
    end
  end
  return $\boldsymbol{p}$;
end

Algorithm 2: Calculate CNV loss P-values

P-values under the null hypothesis, whereas the Beta distribution approximately fits the highly left-skewed distribution of P-values expected of tests for which the null hypothesis is false. After the observed distribution of P-values has been fit to the Beta-Uniform mixture model, the FDR associated with a given P-value can be estimated from the densities of the Beta and Uniform component distributions at that P-value.

The original Pounds-Morris procedure considers all tests as equivalent, and thus integrates no locality information, but for the LOH case combining the procedure with the locality-sensitive HMM is straightforward (figure 4.1). The HMM moves between two discrete states: *No LOH*, and *LOH*. The *No LOH* state emits a Uniform distribution of P-values, as expected under the null hypothesis of no LOH, whereas the *LOH* state emits a left-skewed Beta distribution of P-values, approximating the P-value distribution observed for

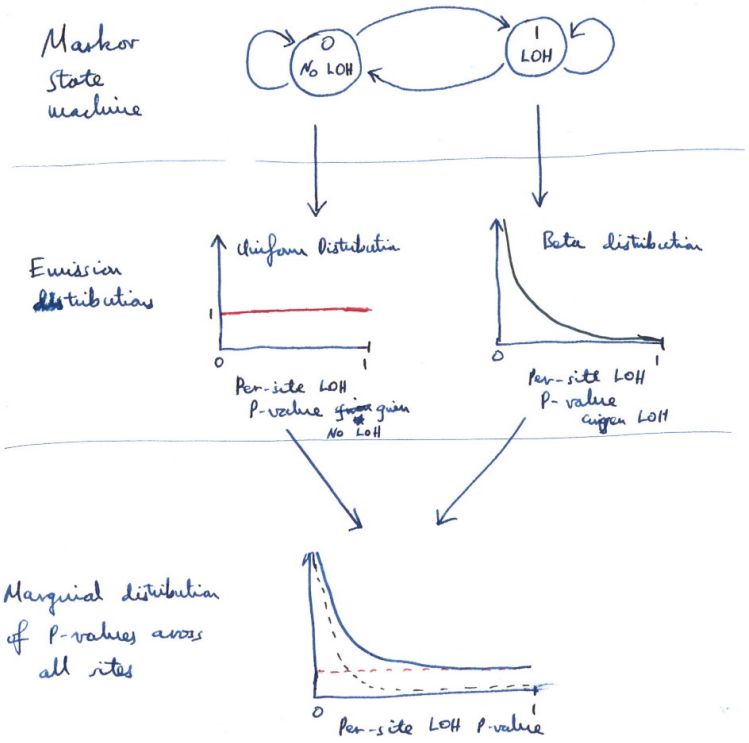


Figure 4.1: Locality-sensitive FDR estimation of LOH calls using a Markov chain Beta-Uniform mixture model.

loci at which the null hypothesis is false. Observed P-values at a chain of adjacent loci are fit to the HMM by standard algorithms implemented in R package depmixS4, and the posterior probability of a locus being in state *No LOH* directly gives the locality-adjusted FDR for that locus. In cases where too few extreme P-values are present to reliably estimate the parameters of the Beta distribution, the fit becomes unstable and FDR estimates potentially unreliable. To handle this situation gracefully, the method fits both the full *No LOH / LOH* model, and a restricted *No LOH* only model, and selects the model with the superior BIC.

Extension of the procedure to the CNV case requires three states: *Diploid*, *Loss*, and *Gain* (figure 4.2). We take advantage of the W_5 statistic's symmetry and fit the HMM to the one-sided p_{loss} CNV P-values; CNV loss is then indicated by P-values near zero, and CNV gain by P-values near one. The

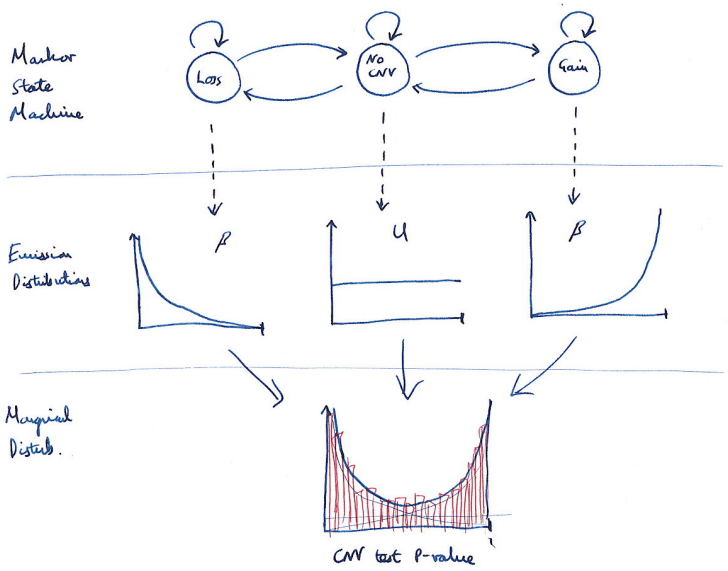


Figure 4.2: Locality-sensitive FDR estimation of CNV calls using a Markov chain double-Beta-Uniform mixture model.

Loss and *Gain* states are modelled by Beta distributions, left-skewed in the *Loss* case, and right-skewed in the *Gain* case. The posterior probability of a locus being in state *Diploid* then gives the overall FDR for a CNV call at that locus. BIC model selection is performed as for the LOH case, except in this case four models are compared: *Diploid*, *Diploid / Loss*, *Diploid / Gain*, and *Loss / Diploid / Gain*.

Although the given procedure is simple in formulation, some additional complexities were required for a practical implementation, all related to the high degree of flexibility of the Beta distribution. The Uniform distribution is a special case of the Beta distribution, and therefore in cases where the distribution of P-values is near Uniform (ie. all sites appear to satisfy the null hypothesis), the fitting problem is ill-posed. This issue was resolved by enforcing Beta parameters $\alpha \leq 0.95$ for LOH and CNV loss detection, and $\beta \leq 0.95$ for CNV gain detection. For FDR correction of CNV P-values, structural zeros were placed on the probabilities of direct transitions between *Loss* and *Gain* states (figure 4.2); although such transitions are biologically plausible, they were found to contribute to unstable fits in noisy data.

Chapter 5

Conclusion

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
c5.M_PHASE/c5.MITOSIS/c5.M_PHASE_OF_MITOTIC_CELL_CYCLE
c5.REGULATION_OF_MITOSIS
c4.GNF2_RFC3/c4.GNF2_RFC4/c4.GNF2_SMC2L1/c4.GNF2_CKS1B/c4.GNF2_CKS2/c4.GNF2_TT
c5.CELL_CYCLE_PROCESS/c5.MITOTIC_CELL_CYCLE/c5.CELL_CYCLE_PHASE
c5.SPINDLE
c4.MORF_BUB1B
c6.CSR_LATE_UP.V1_SIGNED
c5.SPINDLE_POLE
c2.PID_PLK1_PATHWAY
c5.ORGANELLE_PART/c5.INTRACELLULAR_ORGANELLE_PART
c2.REACTOME_CELL_CYCLE/c2.REACTOME_CELL_CYCLE_MITOTIC
c2.REACTOME_CYCLIN_A_B1_ASSOCIATED_EVENTS_DURING_G2_M_TRANSITION
c2.REACTOME_MITOTIC_PROMETAPHASE
c2.KEGG_CELL_CYCLE
c5.CHROMOSOME_SEGREGATION
c4.MORF_FEN1
c2.REACTOME_G1_S_SPECIFIC_TRANSCRIPTION
c2.REACTOME_ACTIVATION_OF_THE_PRE_REPLICATIVE_COMPLEX/c2.REACTOME_ACTI
c2.REACTOME_E2F_ENABLED_INHIBITION_OF_PRE_REPLICATION_COMPLEX_FORMATIO
c2.REACTOME_E2F_MEDIATED_REGULATION_OF_DNA_REPLICATION
c5.CELL_CYCLE_GO_0007049
c2.REACTOME_KINESINS
c3.V$ELK1_02
c5.SPINDLE_MICROTUBULE
c5.MITOTIC_CELL_CYCLE_CHECKPOINT
c2.REACTOME_CELL_CYCLE_CHECKPOINTS/c2.REACTOME_G1_S_TRANSITION/c2.REACT
c4.MORF_ESPL1
c4.MORF_BUB1
c4.MORF_BUB3/c4.MORF_RAD23A
c5.CONDENSED_CHROMOSOME
c4.MORF_RFC4/c4.MORF_RRM1
c2.BIOCARTA_G2_PATHWAY
c3.SCGGAAGY_V$ELK1_02
c2.PID_AURORA_A_PATHWAY
c5.MITOTIC_SISTER_CHROMATID_SEGREGATION/c5.SISTER_CHROMATID_SEGREGATION
c4.MORF_UNG
c2.PID_FOXM1PATHWAY
c4.MORF_GSPT1
c2.REACTOME_METABOLISM_OF_NUCLEOTIDES
c2.PID_ATR_PATHWAY
c2.BIOCARTA_MCM_PATHWAY
c4.MORF_CCNF
c5.CELL_CYCLE_CHECKPOINT_GO_0000075
c5.MITOTIC_SPINDLE_ORGANIZATION_AND_BIOGENESIS/c5.SPINDLE_ORGANIZATION_AN
c4.MORF_EI24
c5.DOUBLE_STRAND_BREAK_REPAIR
c4.GNF2_PA2G4/c4.GNF2_RAN
c2.REACTOME_G2_M_DNA_DAMAGE_CHECKPOINT
c2.KEGG_PYRIMIDINE_METABOLISM

Appendix C

MSigDB signatures correlated with axis A2

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

GeneSet
c2.PID_INTEGRIN1_PATHWAY
c2.PID_INTEGRIN3_PATHWAY
c2.PID_UPA_UPAR_PATHWAY
c4.GNF2_PTX3
c2.KEGG_ECM_RECEPTOR_INTERACTION
c2.PID_INTEGRIN5_PATHWAY
c4.GNF2_MMP1
c2.REACTOME_EXTRACELLULAR_MATRIX_ORGANIZATION/c2.REACTOME_COLLAGEN_F
c5.AXON_GUIDANCE
c2.KEGG_FOCAL_ADHESION
c2.PID_SYNDECAN_1_PATHWAY
c2.REACTOME_CELL_EXTRACELLULAR_MATRIX_INTERACTIONS
c2.PID_INTEGRIN_CS_PATHWAY
c5.TISSUE_DEVELOPMENT
c5.COLLAGEN
c6.CORDENONSI_YAP_CONSERVED_SIGNATURE
c6.LEF1_UP.V1_SIGNED
c2.REACTOME_INTEGRIN_CELL_SURFACE_INTERACTIONS
c5.AXONOGENESIS/c5.CELLULAR_MORPHOGENESIS_DURING_DIFFERENTIATION
c6.STK33_NOMO_SIGNED
c7.GSE17721_CTRL_VS_CPG_12H_BMDM_SIGNED
c7.GSE1460_INTRATHYMIC_T_PROGENITOR_VS_THYMIC_STROMAL_CELL_SIGNED

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 factorizations 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

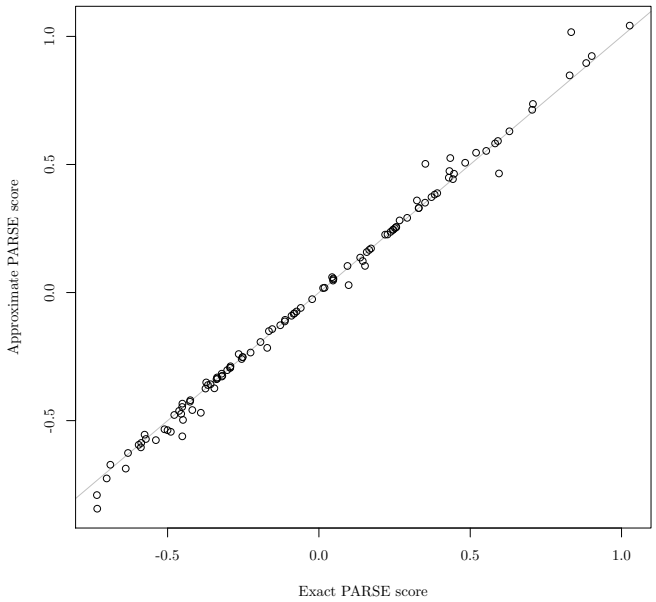


Figure D.1: The linear PARSE score approximation $P \approx kA$ closely matches the exact version calculated using NNLS, when evaluated on APGI GEX data.

(Figure D.1).

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 table (below), and arrange rows of A so that they exactly match the order of rows 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, follows.

	Value
A4GALT	0.00418
A4GNT	-0.01632
ABHD16A	0.00143
ABHD5	0.01227
ABLIM1	-0.01392
ACE	-0.00556
ACKR3	0.00802
ACYP2	-0.01298
ADH1A	-0.01845
ADM	0.00122
AGRP	-0.00509
AKIP1	0.00545
AKR1A1	-0.01321
ALDH5A1	-0.02452
ALOX5AP	-0.00179
AMOT	-0.00825
ANGPTL2	0.01178
ANGPTL4	0.01365
ANKLE2	0.01205
ANKRD22	-0.00941
ANKRD37	0.00474
ANLN	0.04364
APCDD1	0.01244
APCS	0.00602
ARFGAP3	-0.01070
ARHGAP24	-0.02524
ARHGEF19	-0.00476
ARL4C	0.02609
ARSD	-0.01466
ASPM	0.01593
ATAD2	0.02602
ATF7IP2	-0.00405
ATL3	0.00972
AURKB	0.01869

AXIN2	-0.01658
B3GALTL	0.01113
BAMBI	-0.00680
BBS2	0.00587
BCKDK	-0.02452
BCL11B	-0.02161
BIRC5	0.02419
BOC	-0.03047
BTN3A1	-0.00868
C1orf56	-0.00865
C1QTNF6	0.01572
C2orf70	-0.01360
C5orf46	0.01559
C9orf152	-0.02152
CA8	-0.01129
CACHD1	-0.01313
CADPS2	-0.02136
CAMK1G	-0.01790
CAPN6	-0.02615
CARHSP1	-0.01515
CATSPER1	0.00163
CAV1	0.02989
CCDC88A	0.01480
CCL19	-0.01715
CCNB1	0.03071
CCR7	-0.01775
CD70	0.00954
CDA	0.02792
CDC45	0.01256
CDK12	-0.01624
CDK2	0.01546
CEBPB	0.00404
CEP55	0.03755
CFDP1	-0.00617
CHAF1B	0.00920
CHEK1	0.03669

CHN2	-0.02051
CIDEC	-0.00596
CIDECP	-0.00684
CKAP2L	0.03545
CLEC3B	-0.01500
CNIH3	0.01413
CNNM1	-0.01611
COL12A1	0.04098
COL5A3	0.03177
COL7A1	0.01688
COLGALT1	0.02272
COLGALT2	-0.00903
COX4I2	-0.00943
CSNK1D	-0.01128
CST6	0.02032
CTSL	-0.01263
CTSV	0.00987
CYP2S1	-0.01044
DCAF8	-0.02374
DCBLD2	0.03351
DCUN1D5	0.02056
DENND1A	0.01898
DERA	0.01568
DHRS9	-0.00454
DKK1	0.00649
DNAJC9	0.01385
DPY19L1	0.00749
DSG2	0.01463
DSG3	0.02070
DYNC2H1	-0.01537
E2F7	0.03923
EDIL3	0.01326
EIF2AK3	-0.02073
ELMOD3	-0.03300
EMP3	0.01550
ENO2	0.02998

EPHX2	-0.02392
ERRFI1	0.01597
EXOSC8	-0.00850
EYA3	0.02671
FAH	0.01035
FAM120AOS	-0.00980
FAM134B	-0.01945
FAM189A2	-0.01692
FAM83A	0.01202
FAM91A1	0.01341
FBXO22	0.00649
FBXW8	-0.00891
FEM1B	0.04785
FER	0.02675
FGB	-0.00252
FGD6	0.02545
FGG	0.00548
FHDC1	-0.01380
FLRT3	0.01416
FRZB	-0.03715
FSCN1	0.02159
FST	0.01504
FYN	-0.01133
GAB2	-0.03742
GABPB1	0.01929
GAPDH	0.02073
GATA6	-0.01780
GATC	0.02661
GIMAP2	-0.03176
GINS2	0.01713
GNPAT	-0.01458
GOLM1	-0.01171
GPC3	-0.02419
GPR176	0.00563
HIPK2	-0.02620
HJURP	0.02296

HRASLS2	0.00196
HSP90B1	-0.00641
HSPB6	-0.01586
ICAM2	-0.00232
IDH2	0.00528
IFT140	-0.02068
IGFBP1	0.00427
IGLL3P	-0.01241
IKBIP	-0.00033
IL1R2	-0.00660
IL20RB	0.02671
IL33	-0.00991
ITGA5	0.01407
ITPKB	-0.01390
KANK4	0.03261
KCNQ3	0.00040
KCTD10	0.01501
KCTD5	-0.01440
KIAA0513	-0.02989
KIAA1549L	0.01354
KIF14	0.01477
KIF20A	0.02967
KIF2C	0.01417
KLHL5	0.02641
KNTC1	0.02375
KRT17	0.01644
KRT6A	0.01795
KRT6C	0.00798
KRT7	0.01916
KYNU	0.01181
LAMA5	0.00174
LCNL1	-0.01571
LDHA	0.04004
LETM2	0.01687
LGALS9B	-0.00232
LINC01184	-0.01837

LMO3	-0.02246
LMTK2	0.00804
LOC100506562	-0.00290
LOX	0.02695
LYNX1	0.00001
MAP3K8	0.00338
MARCKSL1	-0.00884
MARS2	-0.01442
MC1R	-0.02281
MCEMP1	0.00025
MCM10	0.02451
MCM4	0.02708
MCOLN2	-0.01684
MELK	0.02067
MEOX1	-0.01961
MIF	0.01560
MIR99AHG	-0.03712
MME	0.01102
MRAP2	-0.01810
MRPL24	-0.01395
MTRNR2L1	-0.01563
NACC2	0.00733
NAMPT	0.00071
NCAPD2	0.02756
NCAPG	0.04487
NELFE	-0.00390
NEURL2	0.01012
NFIA	-0.03387
NFIX	-0.01186
NMB	-0.00205
NPM1	-0.01520
NR0B2	-0.01468
NRP2	0.00250
NUP155	0.02330
OAZ1	-0.00134
ORC1	-0.00199

P2RY2	0.01288
P2RY8	-0.03043
P4HA1	0.00225
P4HA2	0.01770
PAX8	0.01350
PAX8-AS1	0.00830
PBXIP1	-0.01174
PCDH20	-0.00861
PCF11	-0.01710
PCOLCE2	-0.00752
PDLIM7	0.01678
PEX11B	-0.02280
PFKFB4	0.00525
PGAM5	0.00973
PGBD3	0.01700
PHACTR3	0.00172
PHLDA1	0.03330
PHOSPHO2	-0.02129
PIGL	0.00833
PLAC9	-0.02093
PLAU	0.03213
PLEKHS1	-0.01672
PLIN2	-0.01174
PLIN3	-0.00506
PLOD1	0.00369
PLOD2	0.02261
POC1A	0.01507
POLA2	0.00692
POP5	-0.00224
POU2AF1	-0.02222
PP7080	-0.01242
PPAPDC1A	0.02867
PPM1H	-0.02311
PPP1R12B	0.00096
PPP1R14B	0.01352
PPP1R3C	0.00125

PPY	-0.02787
PRC1	0.02492
PRDM16	-0.02289
PREP	-0.01799
PRKCDBP	0.00755
PRMT7	-0.01665
PROSER2	0.01761
PRR11	0.01859
PTGES	0.02681
PTPN21	0.01723
PXDN	0.02281
PYGL	0.01714
RAB31	0.01316
RACGAP1	0.02957
RALGAPB	0.02214
RAP1GAP	-0.03483
RASL11B	-0.01808
RAVER2	-0.01352
RBMS2	0.02834
RERE	-0.01635
RERGL	-0.01801
RFC5	0.01848
RFK	-0.01090
RFX2	-0.00264
RGS3	-0.00319
RGS5	-0.01505
RHOF	0.02828
RMND5A	-0.00614
RNF103	-0.03019
RPA2	-0.02756
RPIA	-0.02226
SAMD5	-0.00655
SCGB2A1	-0.01773
SCYL2	0.01826
SDIM1	-0.01083
SEC23IP	-0.01125

SELENBP1	-0.02707
SEPW1	-0.01161
SERPINB3	-0.00201
SERPINH1	0.02086
SERTAD2	-0.00995
SGSM1	-0.02933
SH3GL1	-0.02784
SLAMF9	-0.00761
SLC12A2	-0.01821
SLC15A1	-0.00139
SLC16A3	0.01842
SLC2A1	0.01424
SLC2A3	0.00438
SLC30A3	-0.01126
SLC40A1	-0.02146
SMOX	-0.02258
SNORA11D	-0.00256
SNRPB	0.00276
SOBP	-0.03269
SOD2	0.00120
SPHK1	0.03861
SPIN4	0.01254
SPOCD1	0.02117
SPOCK1	0.03046
SPP1	0.00175
ST3GAL2	-0.02187
ST6GAL1	-0.02118
ST6GALNAC1	-0.01232
STAT5B	-0.03172
STK39	-0.01196
SUGCT	0.01833
SULF2	0.01494
SYNE2	-0.00968
TAF5L	-0.01213
TARBP2	-0.01019
TCEA3	-0.02679

TCTA	-0.03326
TGFBI	0.03259
THSD7B	-0.01931
TLE4	-0.01794
TM9SF3	-0.01255
TMED1	-0.01796
TMEM26	0.03659
TMTC4	-0.01797
TNFRSF10D	-0.00315
TNFRSF17	-0.01180
TNFRSF6B	0.02308
TOM1	-0.01640
TOM1L2	0.00266
TOR2A	-0.02926
TPD52L2	-0.00579
TPX2	0.02590
TRAPPC2	-0.01920
TREM1	-0.00073
TRERF1	0.00581
TRIM2	-0.02689
TSTD1	-0.02503
TUBA1C	0.02053
TWIST1	0.02246
UFC1	-0.03123
UHRF2	0.01445
UPP1	0.00182
USP30	0.00629
VPS35	-0.01219
VSTM2L	0.00352
WNT2B	-0.00812
XXYLT1	0.00341
ZBED2	0.02396
ZFPM1	-0.02180
ZNF185	0.01435
ZNF565	-0.00565
ZNF658	-0.01988

ZPLD1	0.00165
ZSCAN16	-0.00720
ZSCAN32	-0.02184

Glossary

APGI Australian Pancreatic Cancer Genome Initiative. iv, 5, 17, 22–26, 28, 33, 35, 36, 38, 40, 43, 44, 46–49, 83, 84

AUC area under the curve. 13

BAF B allele frequency. 37

BAM binary sequence alignment / map file. 53, 54, 56

BIC Bayesian information criterion. 10, 65, 66

BMC Bayesian model comparison. 55

CA-19-9 carbohydrate antigen 19-9. 4, 7

CNV copy number variation. v, 57, 59, 60, 62–66

CPH Cox proportional hazard. 9–11

CPSS complementary pair subset selection. 24, 44–46

CPV clinico-pathological variable. iv, 4, 5, 7, 22, 23, 37, 39, 44, 49, 50

CT computed tomography. 1, 7

CV cross-validation. 26

DSD disease-specific death. 24, 44

DSS disease-specific survival. iii, 32, 44

ECM extracellular matrix. 37

EMT epithelial to mesenchymal transition. 5, 36–38, 40–42

EUS endoscopic ultrasound. 1, 5–7

FAST feature aberration at survival times. 24, 44–46

FDR false-discovery rate. 24, 45, 64–66

FNA fine needle aspirate. 5–7

FWER familywise error rate. 26, 27, 34

GATK Genome analysis toolkit. 54

GEO Gene Expression Omnibus. 47

GEX gene expression. iii, 17–24, 44, 46, 48, 49, 84

GG generalised gamma. 11

GSVA gene set variation analysis. 38, 48, 49

HMM hidden Markov model. 63–65

IBS integrated Brier score. iii, 13–15

ICA independent component analysis. 19–21

ICGC International Cancer Genome Consortium. 49

IDAT Illumina data. 43, 50

IHC immunohistochemical. 5, 7

IHC immunohistochemistry. 5, 6

indel insertion / deletion event. 54

KM Kaplan-Meier. iii, 11–13

LA-BQSR local alignment and base quality score recalibration. 54

LASSO least absolute shrinkage and selection operator. iii, 25, 26, 32

LOESS local regression. 9

LOH loss of heterozygosity. 55, 57, 63, 64, 66

MDS multidimensional scaling. 44

MSigDB molecular signatures database. i, ii, iv, 18, 37, 38, 48, 49, 51, 80–82

MSKCC Memorial Sloan-Kettering Cancer Center. 4

NCBI National Center for Biotechnology Information. 47

NGS next-generation sequencing. 59, 60, 62

NMF non-negative matrix factorization. iii, 19–21, 23–25, 45, 83

NNLS non-negative least squares. 25, 26, 47, 83, 84

NSWPCN New South Wales Pancreatic Cancer Network. 6–9, 11–15

PARSE prognostic axis risk stratification estimate. ii–iv, 26, 27, 34–37, 47, 48, 83, 84

PCA principal component analysis. 19–21

PCOP the Pancreas Cancer Outcome Predictor. 3, 5–7

PDAC pancreatic ductal adenocarcinoma. 17, 18, 21–24, 26, 28, 36, 40–44, 48

PH proportional hazard. 9–11

PPV positive predictive value. 4

SIS sure independence screening. 24, 44–46

SNMF/L sparse non-negative matrix factorization, long variant. iii, 24, 25, 29, 31, 45–47

SNP single nucleotide polymorphism. 37

SNR signal-to-noise ratio. 59

SNV single nucleotide variant. 54

TCGA The Cancer Genome Atlas. iv, 26, 27, 34–36, 47, 48

TD-ROC time-dependent receiver operating characteristic. 13

VST variance stabilizing transform. 43, 45, 48

WGS whole genome sequencing. 60

References

[1] 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, August 2000.

[2] Nils D Arvold, Andrzej Niemierko, Harvey J Mamon, Carlos Fernandez-del Castillo, and Theodore 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):1383–90, August 2011.

[3] Umashankar K Ballehaninna and Ronald 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):105–19, June 2012.

[4] Giuliano Barugola, Massimo Falconi, Rossella Bettini, Letizia Boninsegna, Andrea Casarotto, Roberto Salvia, Claudio Bassi, and Paolo Pederzoli. The Determinant Factors of Recurrence Following Resection for Ductal Pancreatic Cancer. *Journal of the Pancreas (Online)*, 8(1):132–140, 2007.

[5] 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. *Pancreas*, 37(4):462, 2008.

[6] Andrew V Biankin, Nicola Waddell, Karin S Kassahn, Marie-Claude Gingras, Lakshmi B Muthuswamy, Amber L Johns, David K Miller, Peter J Wilson, Ann-Marie Patch, Jianmin Wu, and Others. Pancreatic cancer

genomes reveal aberrations in axon guidance pathway genes. *Nature*, 491(7424):399–405, 2012.

[7] Ahmet Bilici. Prognostic factors related with survival in patients with pancreatic adenocarcinoma. *World journal of gastroenterology : WJG*, 20(31):10802–12, August 2014.

[8] Paul Blanche, Jean-François Dartigues, and Hélène Jacqmin-Gadda. Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. *Statistics in medicine*, 32(30):5381–97, December 2013.

[9] Paul Blanche, Jean-François Dartigues, and Hélène Jacqmin-Gadda. Review and comparison of ROC curve estimators for a time-dependent outcome with marker-dependent censoring. *Biometrical journal. Biometrische Zeitschrift*, 55(5):687–704, September 2013.

[10] L Breiman. Random forests. *Machine learning*, pages 5–32, 2001.

[11] Murray F. Brennan, Michael W. Kattan, David Klimstra, and Kevin Conlon. Prognostic Nomogram for Patients Undergoing Resection for Adenocarcinoma of the Pancreas. *Annals of Surgery*, 240(2):293–298, August 2004.

[12] Jean-Philippe Brunet, Pablo Tamayo, Todd R Golub, and Jill P Mesirov. Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the National Academy of Sciences of the United States of America*, 101(12):4164–9, March 2004.

[13] Eric A Collisson, Anguraj Sadanandam, Peter Olson, William J Gibb, Morgan Truitt, Shenda Gu, Janine Cooc, Jennifer Weinkle, Grace E Kim, Lakshmi Jakkula, Heidi S Feiler, Andrew H Ko, Adam B Olshen, Kathleen L Danenberg, Margaret A Tempero, Paul T Spellman, Douglas Hanahan, and Joe W Gray. Subtypes of pancreatic ductal adenocarcinoma and their differing responses to therapy. *Nature medicine*, 17(4):500–3, April 2011.

[14] Christopher Cox, Haitao Chu, MF Schneider, and A Muñoz. Parametric survival analysis and taxonomy of hazard functions for the generalized gamma distribution. *Statistics in medicine*, 26:4352–4374, 2007.

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

[16] Attila Frigyesi and Mattias Höglund. Non-negative matrix factorization for the analysis of complex gene expression data: identification of clinically relevant tumor subtypes. *Cancer informatics*, 6, 2008.

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

[18] 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):2529–45, 1999.

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

[20] Christian J Gröger, Markus Grubinger, Thomas Waldhör, Klemens Vierlinger, and Wolfgang Mikulits. Meta-analysis of gene expression signatures defining the epithelial to mesenchymal transition during cancer progression. *PloS one*, 7(12):e51136, January 2012.

[21] Kangxia Gu, Hon Keung Tony Ng, Man Lai Tang, and William R Schucany. Testing the ratio of two poisson rates. *Biometrical journal*, 50(2):283–98, April 2008.

[22] Douglas Hanahan and Robert a Weinberg. Hallmarks of cancer: the next generation. *Cell*, 144(5):646–74, March 2011.

[23] Sonja Hänzelmann, Robert Castelo, and Justin Guinney. GSVA: gene set variation analysis for microarray and RNA-Seq data. *BMC bioinformatics*, 14(1):7, January 2013.

[24] H C Harsha, Kumaran Kandasamy, Prathibha Ranganathan, Sandhya Rani, Subhashri Ramabadran, Sashikanth Gollapudi, Lavanya Balakrishnan, Sutopa B Dwivedi, Deepthi Telikicherla, Lakshmi Dhevi N Selvan, Renu Goel, Suresh Mathivanan, Arivusudar Marimuthu, Manoj Kashyap, Robert F Vizza, Robert J Mayer, James a Decaprio, Sudhir Srivastava,

Samir M Hanash, Ralph H Hruban, and Akhilesh Pandey. A compendium of potential biomarkers of pancreatic cancer. *PLoS medicine*, 6(4):e1000046, April 2009.

[25] Patrick J Heagerty and Yingye Zheng. Survival model predictive accuracy and ROC curves. *Biometrics*, 61(1):92–105, March 2005.

[26] Choon-Kiat Ho, Jörg Kleeff, Helmut Friess, and Markus W Büchler. Complications of pancreatic surgery. *HPB : the official journal of the International Hepato Pancreato Biliary Association*, 7(2):99–108, January 2005.

[27] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, 6(2):65–70, 1979.

[28] Hemant Ishwaran, Udaya B. Kogalur, Eugene H. Blackstone, and Michael S. Lauer. Random survival forests. *The Annals of Applied Statistics*, 2(3):841–860, September 2008.

[29] Hyunsoo Kim and Haesun Park. Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. *Bioinformatics*, 23(12):1495–502, June 2007.

[30] Tae Hyun Kim, Sung-Sik Han, Sang-Jae Park, Woo Jin Lee, Sang Myung Woo, Tae Yoo, Sung Ho Moon, Seong Hoon Kim, Eun Kyung Hong, Dae Yong Kim, and Joong-Won 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):e743–8, December 2011.

[31] Sang Hyub Lee, Haeryoung Kim, Jin-Hyeok Hwang, Eun Shin, Hye Seung Lee, Dae Wook Hwang, Jai Young Cho, Yoo-Seok Yoon, Ho-Seong Han, and Byung Hyo Cha. CD24 and S100A4 expression in resectable pancreatic cancers with earlier disease recurrence and poor survival. *Pancreas*, 43(3):380–8, April 2014.

[32] Su-In Lee and Serafim Batzoglou. Application of independent component analysis to microarrays. *Genome biology*, 4(11):R76, January 2003.

[33] Su-In Lee and Serafim Batzoglou. Application of independent component analysis to microarrays. *Genome biology*, 4(11):R76, January 2003.

[34] Michael S. Lewicki and Terrence J. Sejnowski. Learning Overcomplete Representations. *Neural Computation*, 12(2):337–365, February 2000.

[35] 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.

[36] Guopei Luo, Jiang Long, Bo Zhang, Chen Liu, Jin Xu, Quanxing Ni, and Xianjun Yu. Stroma and pancreatic ductal adenocarcinoma: An interaction loop. *Biochimica et Biophysica Acta - Reviews on Cancer*, 1826(1):170–178, 2012.

[37] Robert C MacCallum, Keith F Widaman, Shaobo Zhang, and Sehee Hong. Sample size in factor analysis. *Psychological Methods*, 4(1):84–99, 1999.

[38] Daruka Mahadevan and Daniel D Von Hoff. Tumor-stroma interactions in pancreatic ductal adenocarcinoma. *Molecular cancer therapeutics*, 6(4):1186–97, April 2007.

[39] David R McDonald. On the Poisson approximation to the multinomial distribution. *Canadian Journal of Statistics*, 8(I):115–118, 1980.

[40] A Popescu, Adriana-Mihaela Ciocâlteu, D I Gheonea, Sevastia Iordache, Carmen Florina Popescu, A Sftoiu, and T Ciurea. Utility of Endoscopic Ultrasound Multimodal Examination with Fine Needle Aspiration for the Diagnosis of Pancreatic InsulinomaA Case Report. *Current Health Sciences Journal*, 38(1), 2012.

[41] S. Pounds and S. W. Morris. Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values. *Bioinformatics*, 19(10):1236–1242, July 2003.

[42] Bisakha Ray, Mikael Henaff, Sisi Ma, Efstratios Efstathiadis, Eric R Peskin, Marco Picone, Tito Poli, Constantin F Aliferis, and Alexander Statnikov. Information content and analysis methods for multi-modal high-throughput biomedical data. *Scientific reports*, 4:4411, January 2014.

[43] Patrick Royston, Douglas G Altman, and Willi Sauerbrei. Dichotomizing continuous predictors in multiple regression: a bad idea. *Statistics in medicine*, 25(1):127–41, January 2006.

[44] Samir a Saidi, Cathrine M Holland, David P Kreil, David J C MacKay, D Stephen Charnock-Jones, Cristin G Print, and Stephen K Smith. Independent component analysis of microarray data in the study of endometrial cancer. *Oncogene*, 23(39):6677–83, August 2004.

[45] Charitini Salla, Panagiotis Konstantinou, and Paschalis Chatzipantelis. CK19 and CD10 expression in pancreatic neuroendocrine tumors diagnosed by endoscopic ultrasound-guided fine-needle aspiration cytology. *Cancer*, 117(6):516–21, December 2009.

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

[47] Sarah Song, Katia Nones, David Miller, Ivon Harliwong, Karin S Kassahn, Mark Pinese, Marina Pajic, Anthony J Gill, Amber L Johns, Matthew Anderson, and Others. qpure: A tool to estimate tumor cellularity from genome-wide single-nucleotide polymorphism profiles. *PloS one*, 7(9):e45835, 2012.

[48] Edward B Stelow, Carolyn Woon, Stefan E Pambuccian, Michael Thrall, Michael W Stanley, Rebecca Lai, Shawn Mallery, and H Evin Gulbahce. Fine-needle aspiration cytology of pancreatic somatostatinoma: the importance of immunohistochemistry for the cytologic diagnosis of pancreatic endocrine neoplasms. *Diagnostic cytopathology*, 33(2):100–5, August 2005.

[49] Jeran K Stratford, David J Bentrem, Judy M Anderson, Cheng Fan, Keith a Volmar, J S Marron, Elizabeth D Routh, Laura S Caskey, Jonathan C Samuel, Channing J Der, Leigh B Thorne, Benjamin F Calvo, Hong Jin Kim, Mark S Talamonti, Christine a Iacobuzio-Donahue, Michael a Hollingsworth, Charles M Perou, and Jen Jen Yeh. A six-gene signature predicts survival of patients with localized pancreatic ductal adenocarcinoma. *PLoS medicine*, 7(7):e1000307, July 2010.

[50] Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, Scott L. Pomeroy, Todd R. Golub, Eric S. Lander, and Jill 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.

[51] Andrew E Teschendorff, Michel Journée, Pierre a Absil, Rodolphe Sepulchre, and Carlos Caldas. Elucidating the altered transcriptional programs in breast cancer using independent component analysis. *PLoS computational biology*, 3(8):e161, August 2007.

[52] B Y Terry M Therneau, Patricia M Grambsch, Division Biostatistics, Mayo Clinic, and Thomas R Fleming. Martingale-based residuals for survival models. *Biometrika*, 77(1):147–160, 1990.

[53] Nobukazu Tsukamoto, Shinichi Egawa, Masanori Akada, Keiko Abe, Yuriko Saiki, Naoyuki Kaneko, Satoru Yokoyama, Kentaro Shima, Akihiro Yamamura, Fuyuhiko Motoi, Hisashi Abe, Hiroki Hayashi, Kazuyuki Ishida, Takuya Moriya, Takahiro Tabata, Emiko Kondo, Naomi Kanai, Zhaodi Gu, Makoto Sunamura, Michiaki Unno, and Akira Horii. The expression of S100A4 in human pancreatic cancer is associated with invasion. *Pancreas*, 42(6):1027–33, August 2013.

[54] David Venet, Jacques E Dumont, and Vincent Detours. Most random gene expression signatures are significantly associated with breast cancer outcome. *PLoS computational biology*, 7(10):e1002240, October 2011.

[55] Geng Zhang, Peijun He, Hanson Tan, Anuradha Budhu, Jochen Gaedcke, B Michael Ghadimi, Thomas Ried, Harris G Yfantis, Dong H Lee, Anirban Maitra, Nader Hanna, H Richard Alexander, and S Perwez Hussain. Integration of metabolomics and transcriptomics revealed a fatty acid network exerting growth inhibitory effects in human pancreatic cancer. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 19(18):4983–93, September 2013.