List of Corrections

Mah Dissertat'n

Mark Pinese

December 12, 2014 Build 0.0.36

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

List of Figures

List of Tables

List of Algorithms

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.

Chapter 1

Signatures of Survival Processes in Pancreas Cancer

[1]

1.1 Introduction

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 **??**, the wide 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 [3] 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 Australian Pancreatic Cancer Genome Initiative (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 fundamental biology of the one common cancer that has, to date, been almost entirely refractory to all the tools of modern molecular medicine.

[1] MP Fatal: TODO: put the thesis somewhere: That specific molecular processes control survival of resectable PC, and that these processes can be identified and detected using GEX data.

Despite extensive research, PDAC remains a poorly-understood disease. Recent genomic profiling has revealed the genetic alterations that accompany the cancer [2], and a huge number of prognostic factors are known [9] (refer to chap:intro 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 [15]. 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, 12]. A number of factors are likely to have contributed to this reluctance: deconvolution methods require relatively large sets of high-quality measurements [13], 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 [19]) have only recently reached maturity.

In addition to the general technical challenges of GEX deconvolution, issues particular to pancreas cancer significantly complicate attempts to identify molecular processes at work within the tumours. Pancreas cancer is challenging to sample, and mRNA in the tissue degrades rapidly once extracted, complicating sample collection. Additionally, a feature of PDAC is the presence of a dense desmoplastic stromal reaction throughout the tumour, that is formed by genetically normal patient stroma cells [14]. The fraction of tumour cells that are actually cancerous varies by more than 10-fold between tumours [2], 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 [3], 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 [14], 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 [3]. This study examined microdissected cancer cells only, and found that the transformed epithelial cells of PDAC could be placed into three major categories, based on their patterns of gene expression. Tumours from these three categories followed distinct clinical courses, and cell lines exhibited category-specific sensitivity to therapeutic drugs. As the first report to identify potential clinically relevant molecular subtypes within PDAC, the Collisson study was a significant advance in the understanding of the molecular processes at play within what was previously considered a homogeneous disease. However, it also possesses shortcomings that limit its clinical utility.

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

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

1.2 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

[2]

Two axes predict survival with resectable pancreatic cancer in multiple cohorts

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.

A complementary pair subset selection (CPSS) wrapper around the core sure independence screening (SIS)-feature aberration at survival times (FAST) variable selection method identified 361 genes (of 13,000 considered) that were associated with time from diagnosis to disease-specific death (DSD) in the APGI cohort. 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

[2] MP Fatal: TODO: Cohort characteristics and subsetting

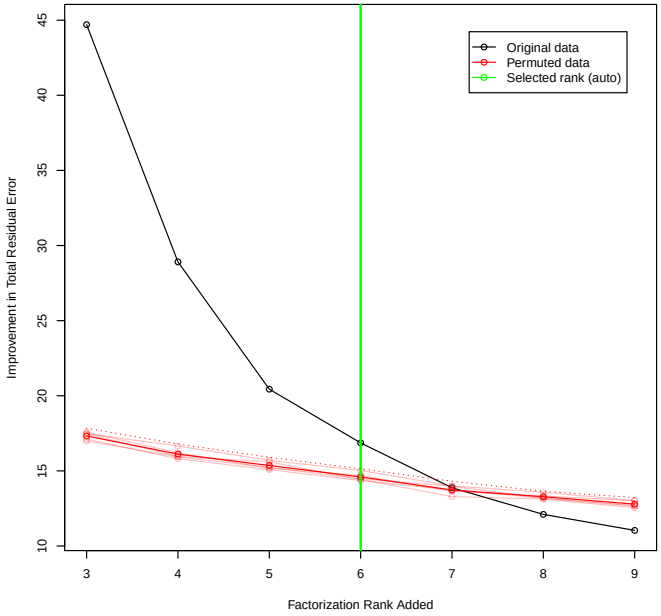


Figure 1.1: Automatic selection of factorization rank. SNMF/L was performed for varying ranks on either unpermuted data (black line) or data permuted within samples (red lines), and the improvement in total residual approximation error $\|A - WH\|_F$ calculated. The highest added rank for which the error 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).

to ensure that even genes for which expression was only weakly prognostic, were included.

Prognostic genes factorized into six metagenes The expression of the 361 survival-associated genes across 228 patients 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 1.1).

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 1.2), and the basis matrix W was reasonably sparse (Figure 1.3). Small row L1 norm of the basis matrix is a desirable condition for this analysis, as it indicates that metagenes are

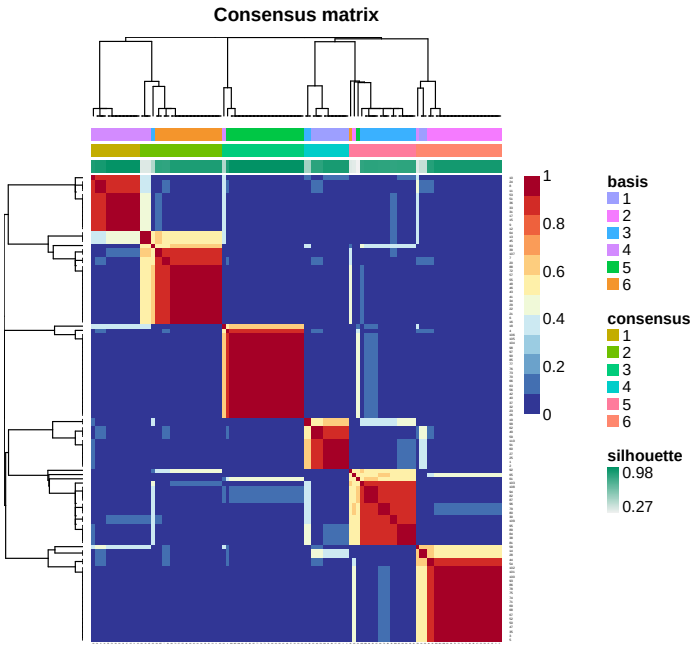


Figure 1.2: Clustering consensus matrix for the final rank-6 clustering. Colours indicate the stability of gene (in rows) and sample (in columns) clusters across random restarts of the factorization; at rank 6 this factorization was largely stable, with identical clusters assigned in all 500 random restarts to the majority of genes and samples.

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 small W row L1 norm. A table of values of the basis matrix W is available as app:sigs-w-matrix on page 24.

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 228-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,

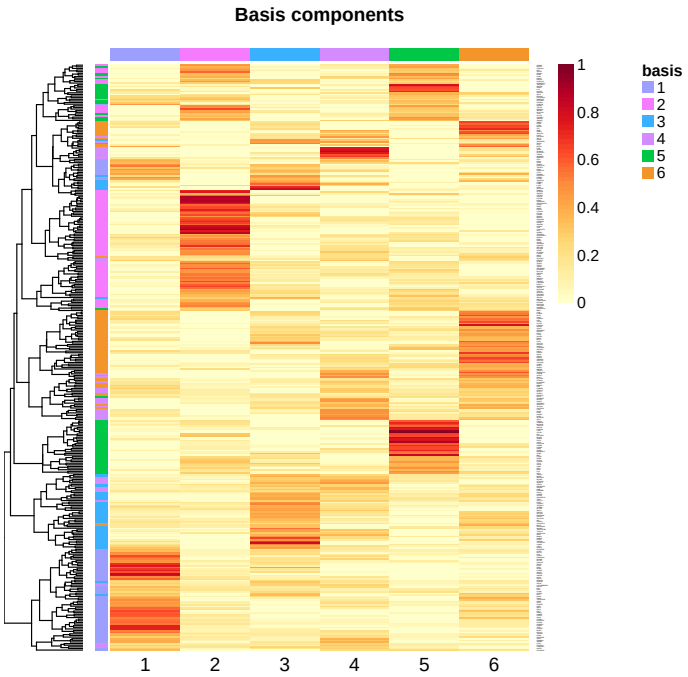


Figure 1.3: Basis matrix W of the final SNMF/L factorization. Rows represent genes, and columns metagenes, with cell colours proportional to the loading of a given gene on a given metagene. The loadings are sparse within rows, indicating that the metagenes are modular, each affecting the expression of largely distinct sets of target genes. A table of values of this basis matrix is available as app:sigs-w-matrix on page 24.

MG2, and MG5, were selected as prognostic (Figure 1.4).

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 1.5, 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 1.6, 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

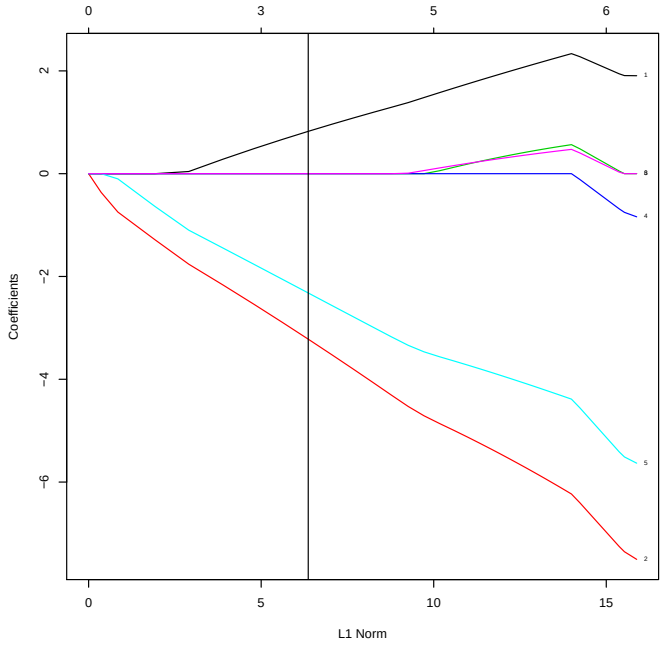


Figure 1.4: Coefficient vs penalty fit trajectories for the LASSO model predicting DSS from metagene expression. Each line represents the model coefficient for a metagene as the model is smoothly varied from a null model (L1 norm = 0), to a full unpenalised Cox fit (L1 norm ≈ 16). The vertical line indicates 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.

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 linear combination of gene expression estimates, and is detailed in app:sigs-parse-approx on page 36.

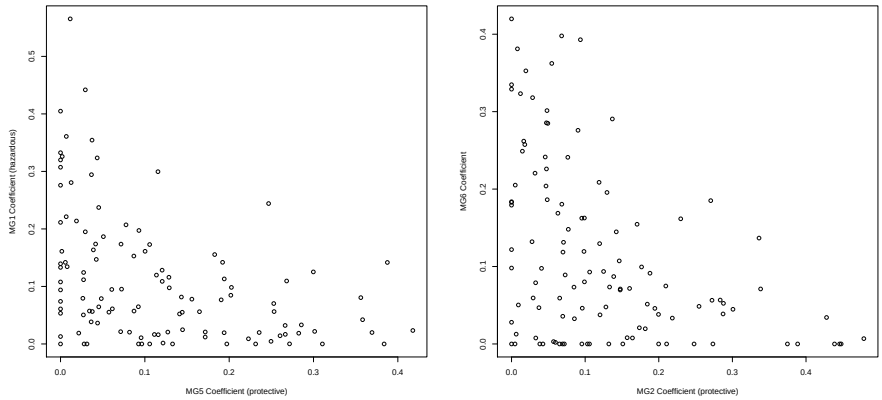


Figure 1.5: Prognostic metagenes form two axes of cell state. Metagene pairs MG1 and MG5, and MG2 and MG6, displayed mutually exclusive coefficient patterns in the APGI cohort, and could be combined to form just two axes of cell state.

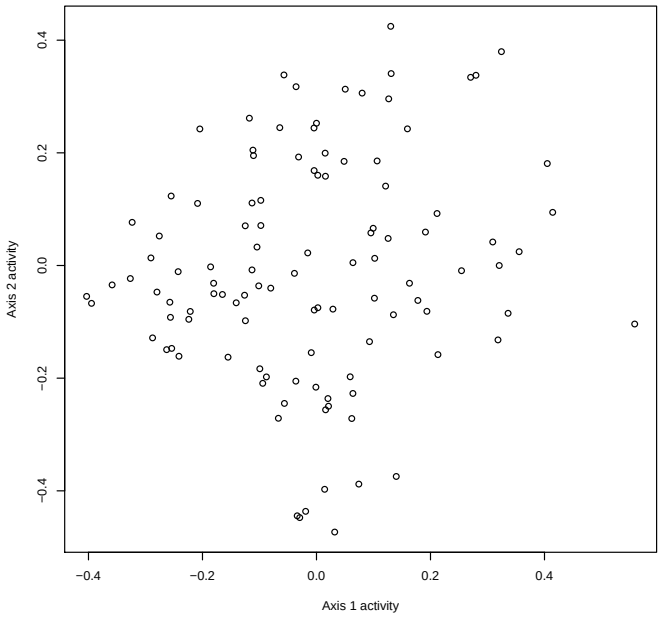


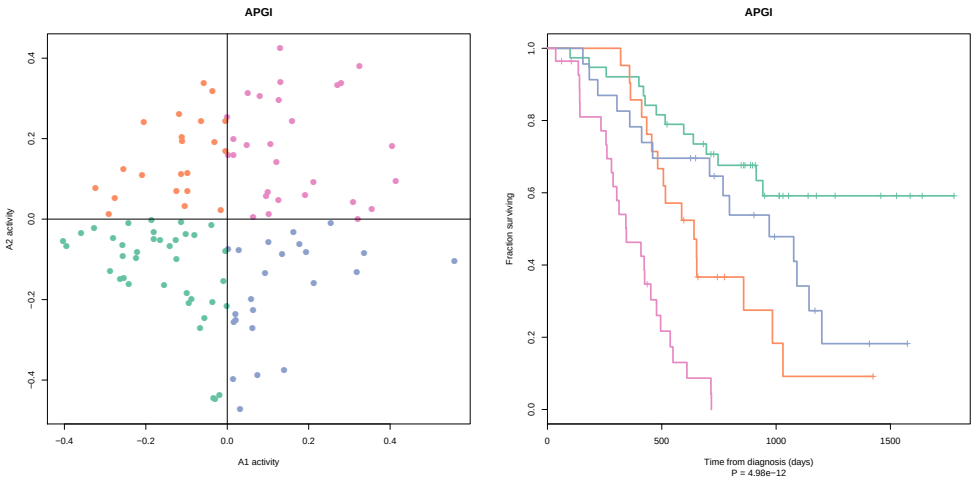
Figure 1.6: Prognostic axis signals are uncorrelated. Activity estimates of axes defined by highly correlated mutually exclusive metagene pairs (Axis A1 = MG1 - MG5, axis A2 = MG6 - MG2) were uncorrelated (Kendall τ test $P = 0.21$), indicating that these axis signals encoded orthogonal outcome-associated processes within tumours.

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 [21] (LRT $P = 0.0149$) and The Cancer Genome Atlas (TCGA) paad (LRT $P = 0.0156$), but not in GSE21501 [18] (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 1.1, 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 1.7.

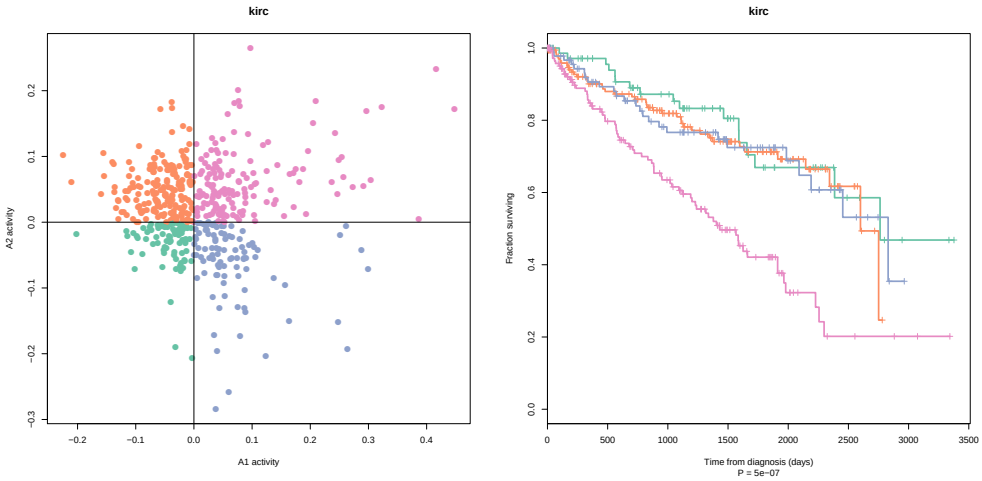
Meta-PCNA is a 130-gene signature of cell proliferation that has been found to be generally prognostic in a number of cancer cohorts [20]. To exclude the possibility that PARSE score simply recapitulated the known meta-PCNA signature, I examined whether PARSE contributed additional prognostic information to meta-PCNA in the large TCGA cohorts. In TCGA kidney renal clear cell carcinoma, lower grade glioma, and lung adenocarcinoma, there was significant evidence that the PARSE score provided prognostic information beyond that given by meta-PCNA, at a 5% FWER (Table 1.1, column b).

Table 1.1: 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

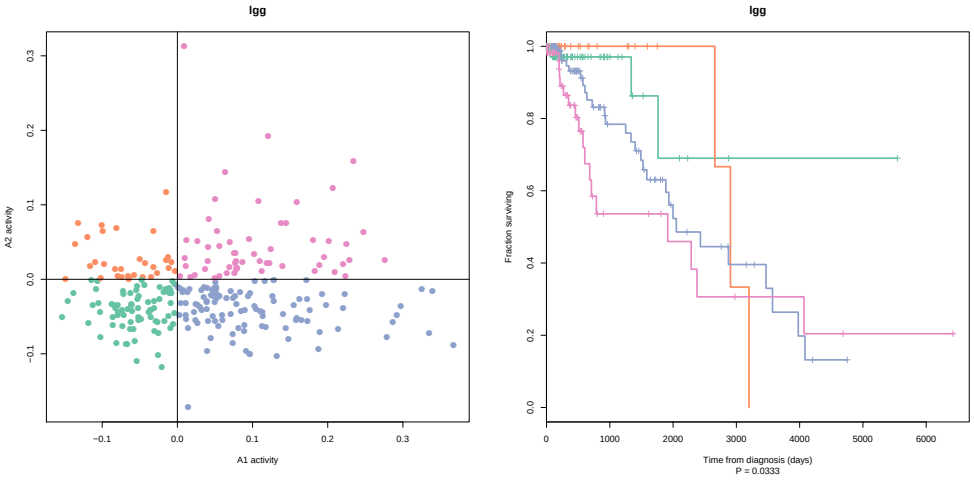


(a) APGI cohort

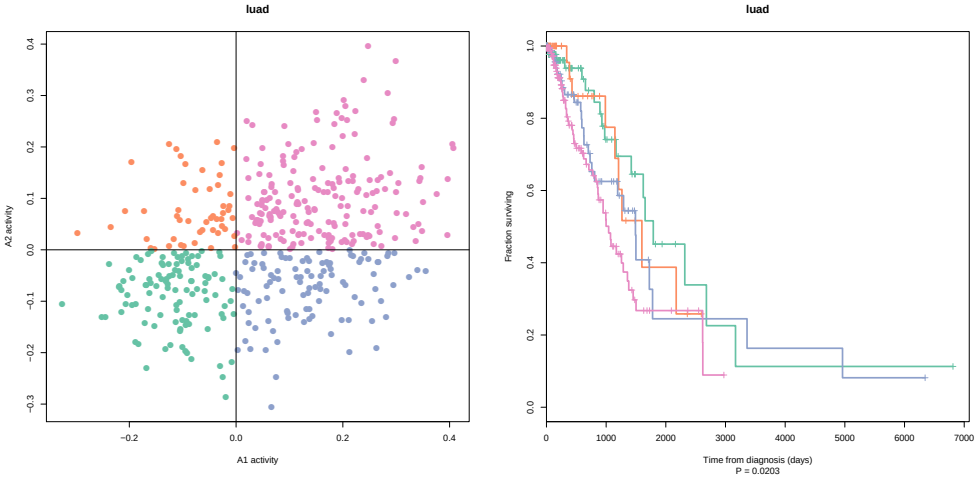


(b) TCGA kirc cohort

Figure 1.7: 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 1.7: (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

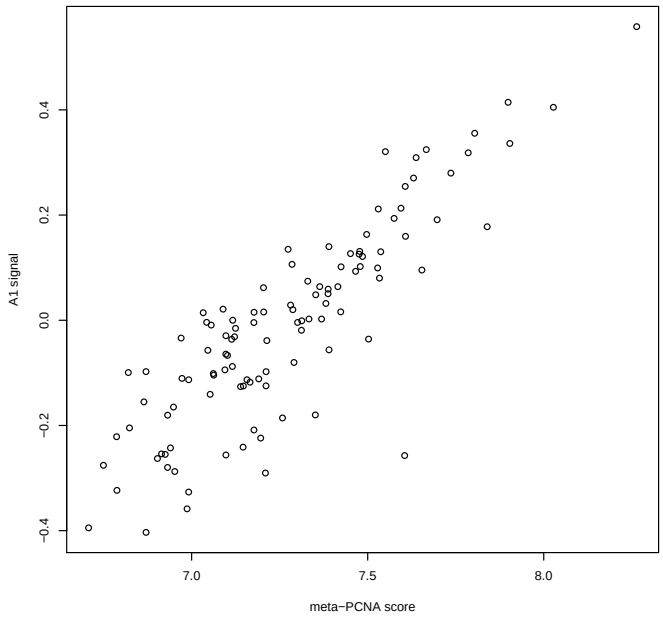


Figure 1.8: Axis A1 signal is closely associated with meta-PCNA signature score. A1 signal and meta-PCNA [20] scores were as evaluated on the APGI training set; Kendall's $\tau = 0.663$, $n = 110$, linear model $R^2 = 0.740$.

to clinical variates, known survival signatures, and scores for signatures from the molecular signatures database (MSigDB) [19].

MSigDB correlations, as well as comparisons to a general proliferative signature, revealed that the PARSE axis A1 (MG1 − MG5) primarily reflected the proliferative state of cells. A1 signal was very strongly correlated with meta-PCNA [20] score (Kendall's $\tau = 0.663$, $n = 110$, Figure 1.8), a relationship supported by its close association to cell cycle-related MSigDB signatures (app:sigs-msigdb-corrs-axis1 on page 33). A1 signal was also significantly positively correlated with qPure [17] estimates of cancer cell fraction in the tumour (Kendall's $\tau = 0.284$, $n = 110$, Table 1.2), although this association was weak (linear model $R^2 = 0.155$).

Among the clinical variables tested, PARSE axis A2 (MG6 − MG2) correlated with stromal content and tumour grade: conditions of high A2 signal were associated with higher stromal content, higher grade, and shorter survival. A2 signal was positively correlated with tumour microscopic pathological grade (Holm-corrected $P = 0.0067$, 50 tests performed), although this dependence was weak: on average, A2 signal was 0.1103 higher in grade 3 or 4 tumours over grade 1 or 2, with $R^2 = 0.119$. A2 signal was also negatively associated with tumour cancer cell fraction, the opposite of the positive re-

Table 1.2: 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.

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

lationship observed for axis A1, despite signal in both axes being positively 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.

A number of MSigDB signatures were associated with A2 signals, among them integrins, extracellular matrix (ECM) processes, and a signature for

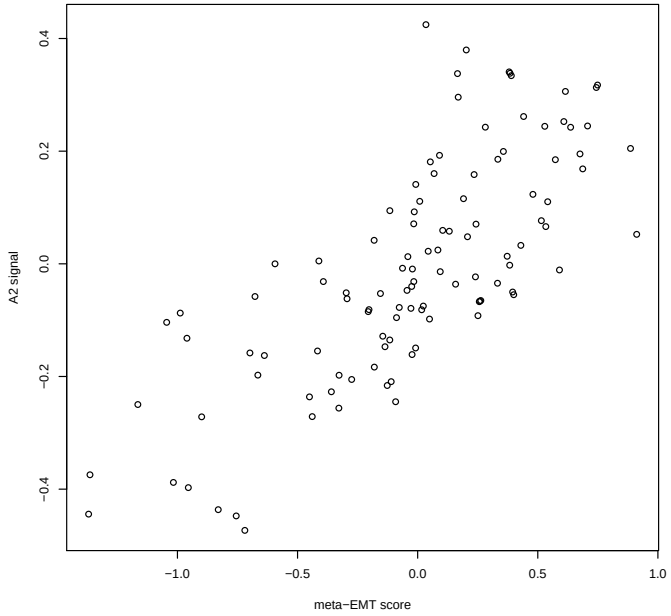


Figure 1.9: Axis A2 signal is closely associated with meta-ECM signature score. A2 signal and meta-ECM [6] scores were as evaluated on the APGI training set; Kendall's $\tau = 0.568$, $n = 110$, linear model $R^2 = 0.557$.

LEF1-mediated epithelial to mesenchymal transition (EMT) (app:sigs-msigdb-corrs-axis2 on page 35). Prompted by the strong positive correlation between A2 and the LEF1 overexpression signature, I investigated the association between A2 signal and score for a general signature of EMT, meta-EMT [6]. meta-EMT and A2 signals were strongly positively correlated (Kendall's $\tau = 0.568$, $n = 110$, linear model $R^2 = 0.557$, 1.9), 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 35), a proxy for expression of the acute phase response protein pentraxin 3.

3

[3]MP Fatal: TODO: Consider comparing A1 and A2 vs meta-PCNA and meta-ECM in TCGA – are A1/A2 better than the metas? Model complexity is the same so therefore can just compare partials – woo

1.3 Discussion

At the molecular level, the phenomenon of cancer has long been recognised as being a composite of many processes [7], however the relative importance of each process to a particular disease has been largely uncertain. In pancreas cancer, a huge number of individual biomarkers are known [9], and some attempts have been made to stratify cancers into molecular subtypes [3], but no studies have provided a comprehensive analysis of which basic hallmarks of cancer are actually important in controlling patient outcome. This work fills that gap in knowledge, and is the first to exhaustively identify proliferation and the EMT as the major molecular processes that determine survival of patients with pancreas cancer.

Discussion points: PARSE = meta-PCNA + meta-ECM (+ immune + stroma?). Context-dependency of stroma signal GSE21501 – why didn't it validate? Relevance to future work.

1.4 Methods

Cohort recruitment and ethics

[4]

Sample collection, preparation, and gene expression microarrays

[5]

Data preprocessing

Microarray quality control and normalization Illumina data (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`

[4] MP Fatal: TODO: Cohort recruitment and ethics
[5] MP Fatal: TODO: Sample collection, preparation, and gene expression microarrays

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.

Summary The above preprocessing steps yielded matched CPV and resected tumour GEX data for 13,000 genes across 228 patients.

Outcome-associated gene selection

Genes that were associated with DSS were identified by SIS-FAST [5], with a CPSS wrapper to reduce the false positive rate [16]. 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 $\times$ patient expression matrix of outcome-associated genes was decomposed into metagenes by the SNMF/L procedure of [11], 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 $\times$ sample matrix A into smaller matrices, the gene $\times$ metagene basis matrix W, and metagene $\times$

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 [4]: 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 $\eta = \max(A)$.[6] 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

[6]Note that this parameter α is denoted β in the R `NMF` package; I use the symbol α here for consistency with [11]

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 24), 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 available (the genes being either filtered out in preprocessing or not measured at all) were assigned scaled expression values of zero. These manipulations yielded a gene $\times$ sample matrix A' with rows matching the gene $\times$ metagene basis matrix W from SNMF/L. The metagene $\times$ sample coefficient matrix H' for the new cohort was then estimated by NNLS implemented in R package `nnls`, solving for each column of $a'_{*,i}$ of A' the 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 24.

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

Calculation of the PARSE score on new cohort data

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

Although not used in this work, a simplified procedure for the approximate calculation of PARSE scores was also developed; see app:sigs-parse-approx on page 36 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 [19] were estimated on the APGI cohort using a modification of the GSVA method [8]. 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 $d_{i,j} = \frac{1}{2}(1 - \text{cor}(s_i, s_j))$, and were used to cluster gene sets using R `hclust` and complete linkage. R `cutree` identified

clusters of highly similar gene sets, using a distance threshold of 0.02; gene set activities within each cluster were merged by taking median values across all samples, to form a new merged gene set activity estimate. Following merging, 7,633 single and compound gene set activity estimates remained across 228 samples.

meta-PCNA and meta-ECM score calculation

Scores for the meta-PCNA signature were calculated from GEX data as described in [20]. 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 [6] 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 1.3. 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 [10]. 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 1.4. Absolute correlations of greater than 0.5 were deemed substantive and reported for further characterisation.

Table 1.3: CPVs tested for association with prognostic axis signals.

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

Table 1.4: 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	*

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, $\widehat{H} = \widehat{W}^{+}A$, 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:

$$P = LH \tag{D.1}$$

$$\approx L\widehat{H} \tag{D.2}$$

$$= L\widehat{W}^{+}A \tag{D.3}$$

$$= \widehat{LW}A \tag{D.4}$$

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 $\widehat{LW}$ is an approximate gene loading vector for calculation of an approximate PARSE score. Approximation of P by $\widehat{LW}A$ appears excellent; when tested on APGI gene expression measurements, the approximation closely matched the more laborious exact NNLS solution (Figure D.1).

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

1. Prepare a linear gene × sample matrix A, in which values for each row (gene) have been scaled to encompass the range 0 to 1.

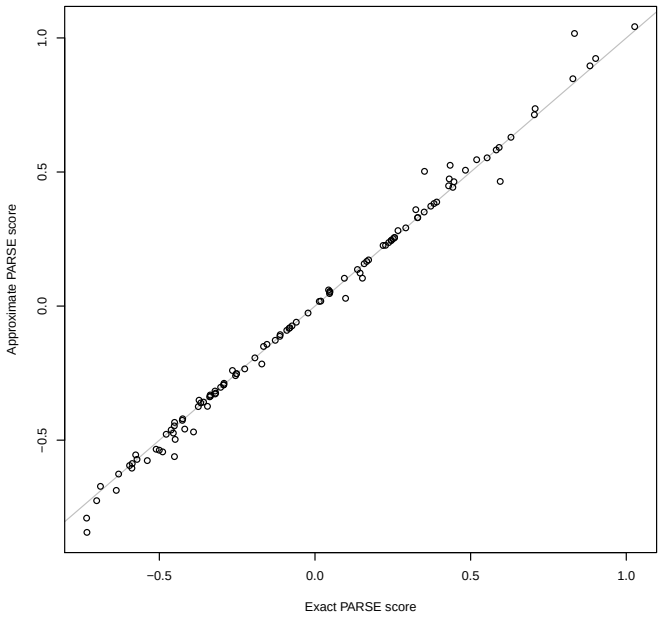


Figure D.1: The linear PARSE score approximation $P \approx \widehat{LW}A$ closely matches the exact version calculated using NNLS, when evaluated on APGI GEX data.

2. Subset A to only the genes present in the $\widehat{LW}$ table (below), and arrange rows of A so that they exactly match the order of rows of $\widehat{LW}$. If genes present in $\widehat{LW}$ are missing from A, insert all-zero rows for these genes into A.

3. Calculate approximate PARSE scores P as $P = \left(\widehat{LW}\right)^T A$. This is equivalent to, for each column (sample) of A, multiplying each entry of the column of A with the corresponding entry of $\widehat{LW}$, and summing the results.

The $\widehat{LW}$ vector 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. 1, 3, 4, 6, 7, 9, 11–13, 15, 19–21, 36, 37

CPSS complementary pair subset selection. 4, 17, 19

CPV clinico-pathological variable. iii, 3, 4, 14, 17, 21, 22

CV cross-validation. 8

DSD disease-specific death. 4, 17

DSS disease-specific survival. ii, 8, 17

ECM extracellular matrix. 14

EMT epithelial to mesenchymal transition. 12, 15, 16

FAST feature aberration at survival times. 4, 17, 19

FDR false-discovery rate. 4, 17

FWER familywise error rate. 10

GEO Gene Expression Omnibus. 19

GEX gene expression. 1–4, 17, 19–21, 37

GSVA gene set variation analysis. 15, 20, 21

IDAT Illumina data. 16

LASSO least absolute shrinkage and selection operator. ii, 6–8

MDS multidimensional scaling. 17

MSigDB molecular signatures database. i, iii, 2, 13–15, 20–22, 33–35

NCBI National Center for Biotechnology Information. 19

NMF non-negative matrix factorization. ii, 4–6, 17, 18, 36

NNLS non-negative least squares. 6, 8, 19, 36, 37

PARSE prognostic axis risk stratification estimate. i–iii, 8, 10–13, 19, 20, 36, 37

PDAC pancreatic ductal adenocarcinoma. 1–4, 10, 12, 20

SIS sure independence screening. 4, 17, 19

SNMF/L sparse non-negative matrix factorization, long variant. ii, 5–7, 17–19

TCGA The Cancer Genome Atlas. iii, 10–12, 20

VST variance stabilizing transform. 16, 18, 20

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] 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.

[3] 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.

[4] 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*, (2003), 2008.

[5] 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.

[6] 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.

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

[8] 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.

[9] 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.

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

[11] 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.

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

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

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

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.