Chapter 1

Signatures of Survival Processes in Pancreatic 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 **??**, 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 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

[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.

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

Despite extensive research, PDAC remains a poorly-understood disease. Recent genomic profiling has revealed the genetic alterations that accompany the disease [**?**], and a huge number of prognostic factors are known [**?**], 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 [**?**]. 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 [2]. A number of factors are likely to have contributed to this reluctance: deconvolution methods are numerically sensitive and require very large sets of high-quality measurements, 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 [**?**]) 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 [**?**]. The fraction of tu-

[2] MP Fatal: PCA, ICA cites

mour cells that are actually cancerous varies by more than 10-fold between tumours [**?**], 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 [**?**], 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 [**?**], 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) has reported a breakdown of PDAC GEX into a small number of biological modules [**?**]. 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 for this purpose. [3]

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 may 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 mea-

[3] MP Fatal: Tidy up this last sentence

surements 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 then tested for association with clinical course and other CPVs, as well as known general prognostic signatures, and their prognostic ability was validated by 10-fold cross-validation and testing in separate cohorts. Those metagenes that were found to be prognostic were then correlated with biological process signatures to associate the metagenes with biological processes.

Cohort characteristics and subsetting

The same homogeneous PDAC subset of the APGI cohort that was used in the work of **??** was employed for this analysis; see **??** on page **??** for case selection criteria and cohort characteristics.

Three metagenes 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

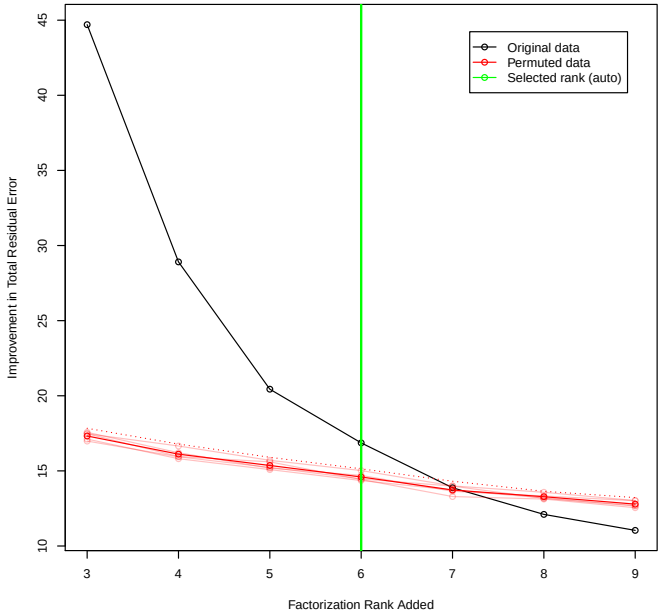

Figure 1.1: Automatic selection of factorization rank. SNMF/L was performed for varying ranks on either unpermuted data (black line) or data permuted within samples (red lines), and the improvement in total residual approximation error $\|A - WH\|_F$ calculated. The highest added rank for which the error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

were associated with time from diagnosis to disease-specific death 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 to ensure that even genes for which expression was only weakly prognostic, were included.

Factorization 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 (fig. 1.1).

500 random restarts of rank 6 SNMF/L were then performed on the

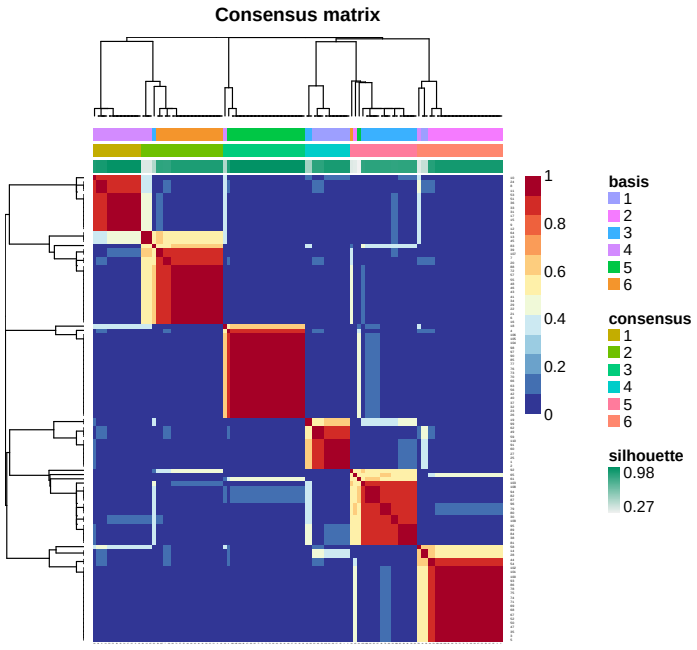

Figure 1.2: Clustering consensus matrix for the final rank-6 clustering. Colours indicate the stability of gene (in rows) and sample (in columns) clusters across random restarts of the factorization; at rank 6 this factorization was largely stable, with identical clusters assigned in all 500 random restarts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

survival-associated gene matrix to yield the final factorization. The resultant clustering consensus matrix was stable (fig. 1.2), and the basis matrix W was reasonably sparse (fig. 1.3). Small row L1 norm 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 for its design, which favours solutions with small W row L1 norm. A table of values of the basis matrix W is available as **??** on **??**–**??**.

Identifying prognostic metagenes The transcription patterns of genes associated with survival in the APGI cohort could be decomposed into just six largely distinct metagenes. Due to the presence of false positives in the 361 screened input genes, some of the metagenes will have no strong association with outcome. To identify which of the six metagenes were ultimately predictive of patient survival, I performed least absolute shrinkage and selection operator (LASSO) regression on the APGI discovery cohort data, using non-negative least squares (NNLS)-estimated coefficients of each of the six

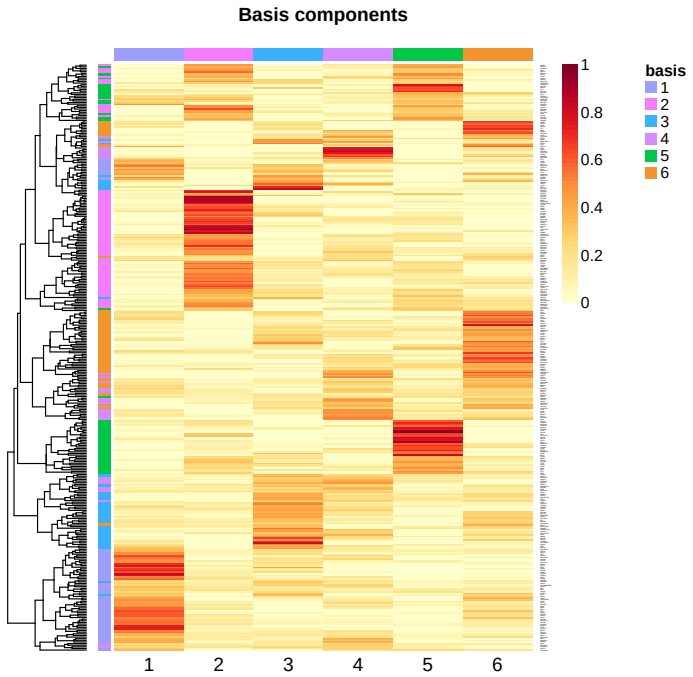

Figure 1.3: Basis matrix W of the final SNMF/L factorization. Rows represent genes, and columns metagenes, with cell colours proportional to the loading of a given gene on a given metagene. The loadings are sparse within rows, indicating that the metagenes are modular, each affecting the expression of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

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 (fig. 1.4).

The final signature developed for predicting time from diagnosis to death for patients PDAC therefore consists of three distinct metagenes: the protective MG2 and MG5, and the hazardous MG1. For external validation of the signature I combined observed coefficients for the three metagenes using their CV-selected APGI cohort LASSO fit coefficients, as Risk score $=$ $0.8238 \times$ MG1 coefficient $- 3.2195 \times$ MG2 coefficient $- 2.3208 \times$ MG5 coefficient. Full metagene scores for the 361 survival-associated genes (the W matrix) are available as **??** on **??**–**??**.

Validation of the three-metagene signature To ensure that the process for development of the three-metagene survival signature was reproducible,

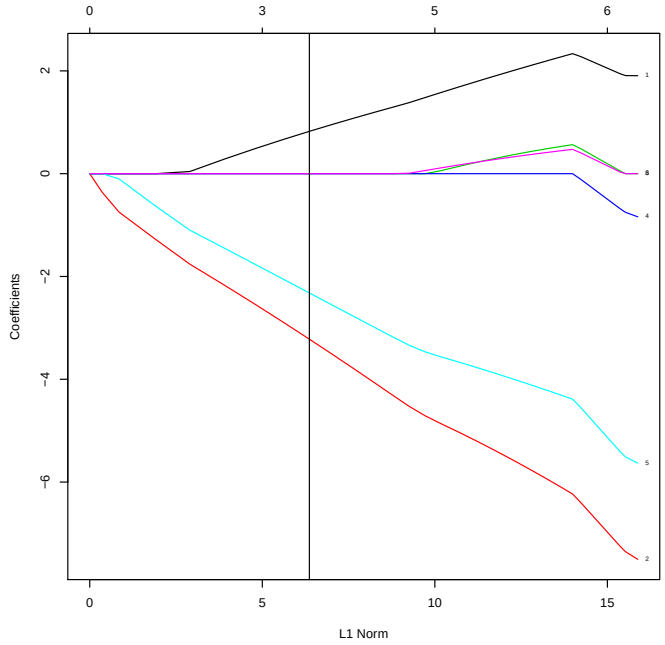

Figure 1.4: Coefficient vs penalty fit trajectories for the LASSO model predicting disease-specific survival from metagene expression. Each line represents the model coefficient for a metagene as the model is smoothly varied from a null model (L1 norm = 0), to a full unpenalised Cox fit (L1 norm ≈ 16). The vertical line indicates the optimal value of L1 norm as selected by 10-fold cross-validation; at this point in the trajectory only metagenes MG1, MG2, and MG5 contribute to prognosis estimates.

and that the signature itself was transferable to other cohorts, I performed cross- and external validation of the metagene survival signature. 10-fold cross-validation was performed on the full metagene discovery procedure, from supervised probe selection by CPSS-SIS-FAST to final LASSO coefficient estimation, including automatic NMF rank estimation. Cross-validated test set risk scores from the APGI discovery cohort were significantly predictive of time from diagnosis to disease-specific death by Cox regression (LRT $P =$[4]). The three-metagene survival signature described above also validated against the PDAC expression cohort GSE28735 [**?**] (LRT $P = 0.0139$), demonstrating the robustness of the metagene signature to cohort and platform effects. [5]

[4]MP Fatal: Update with new data
[5]MP Fatal: Talk about GSE21501 here

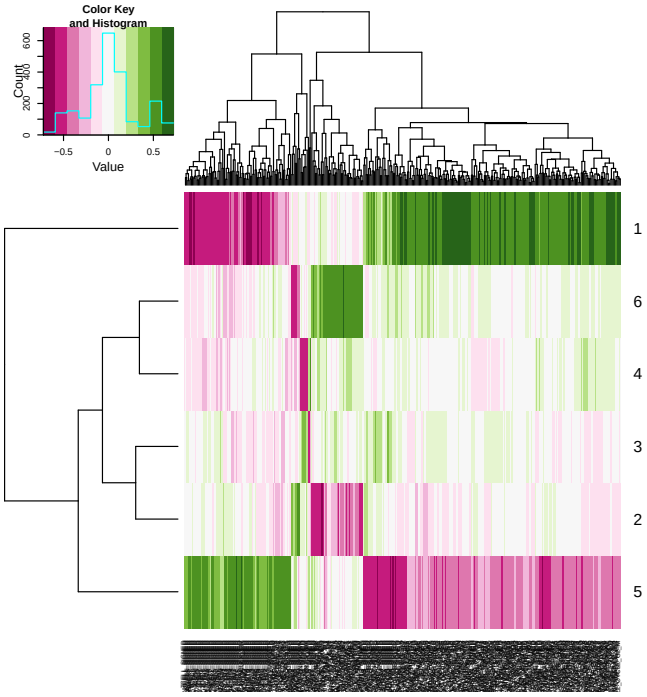

Figure 1.5: Correlations between metagenes and MSigDB signatures. Metagene coefficients, and GSVA MSigDB signature scores, were calculated on the APGI GEX data. Kendall's correlation coefficients were then calculated between the scores of each metagene and signature pair, and are plotted here. Metagenes MG1 and MG5, MG2 and MG6, and MG3 and MG4, form complementary pairs in terms of their correlation patterns with MSigDB signatures.

Prognostic metagenes define two axes of cell state and survival

To link metagene expression with potential underlying biology, metagene coefficients on the APGI discovery cohort were compared to clinical variates, scores of a general prognostic signature, and scores for signatures from the molecular signatures database (MSigDB) [**?**].

Correlations between metagene coefficients and MSigDB scores revealed a parity in the metagenes: MG1 and MG5, MG2 and MG6, and MG3 and MG4, formed complementary pairs in terms of their correlation with MSigDB signatures (fig. 1.6). This parity reflects mutual exclusivity interactions between metagene coefficients, and suggests that the six metagenes encode only three underlying biological states. The three prognostic metagenes were therefore combined into two prognostic axes: MG2-MG6, and MG5-MG1.

MSigDB correlations, as well as comparisons to a general survival signature, revealed that the MG5-MG1 axis primarily reflected the proliferative state of cells. Coefficients of the hazardous MG1 were strongly correlated with

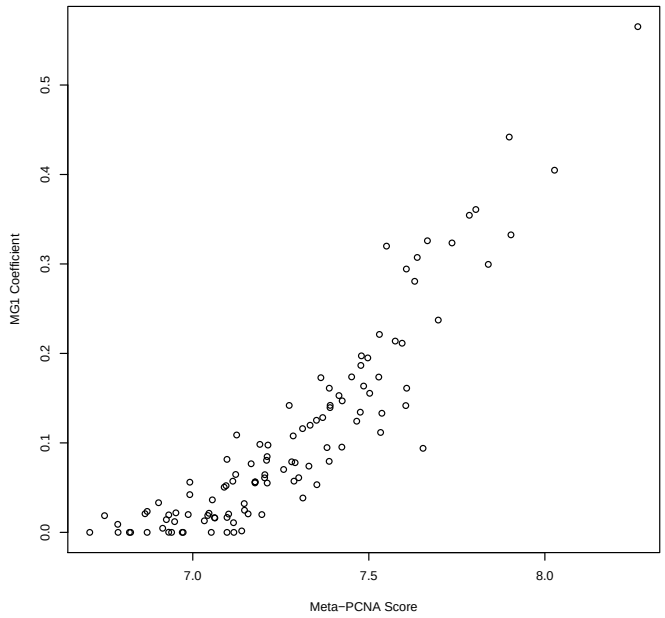

Figure 1.6: MG1 coefficients are closely associated with meta-PCNA signature score. MG1 coefficients and meta-PCNA scores [**?**] evaluated on the APGI training set were highly correlated, with Kendall's $\tau = 0.757$, $n = 110$.

scores of the general outcome predictor meta-PCNA (Kendall's $\tau = 0.757$, $n = 110$), a 130 gene signature of cell proliferation [**?**] (**??**). Supporting this finding, MG1 coefficients were also strongly associated with gene set variation analysis (GSVA) estimates of the E2F1 response, and of a number of cell-cycle related MSigDB sets (table 1.1). Coefficients for the protective MG5 were significantly negatively correlated with qPure estimates of cancer cell fraction [**?**] (table 1.2), and positively correlated with the GSVA score of genes with promoter regions within ±2 kb of the STAT5A target motif `AWTTCYNGGAANY`. Taken together, these associations suggest that a high MG1 coefficient (equivalently, a low MG5 coefficient) is a marker of aggressive tumours with high proliferative activity, relatively low stromal content, and poor outcome. Conversely, high a MG5 coefficient (low MG1) indicates a less aggressive tumour with higher stromal content, and possibly increased immune involvement.

The MG2-MG6 axis correlated with stromal content and tumour grade: conditions of high MG2 and low MG6 were associated with lower stromal content, lower grade, and longer survival. The MG6 coefficient was associated with tumour microscopic pathological grade (Holm-corrected $P = 0.0288$, 150 tests performed), although this dependence was weak: on average, MG6 score was 0.0623 higher in grade 3 or 4 tumours over grade 1 or 2, with $R^2 = 0.122$.

Table 1.1: MSigDB signatures substantially correlated to the MG5-MG1 axis.

MSigDB Signature	Correlation MG1	MG5
c2 REACTOME $G_S\alpha$ signalling events	-0.36	0.55
c5 Phosphoric diester hydrolase activity	-0.35	0.52
c3 V$STAT5A 01	-0.31	0.52
...	...	...
c5 M phase, c5 Mitosis, c5 M phase of mitotic cell cycle	0.68	-0.50

MG6 coefficients were strongly correlated positively correlated with GSVA scores for MSigDB signatures related to extracellular matrix (ECM) interactions and remodelling, and MG2 anticorrelated with a signature of LEF1-mediated epithelial to mesenchymal transition (EMT) (table 1.3), strongly implicating matrix remodelling and invasion as the molecular basis of MG2-MG6 axis score. The protective MG2 metagene was also positively associated with tumour cancer cell fraction, the opposite of the situation observed for the MG5-MG1 axis, where high cancer cell fraction was associated with poorer outcome, through the hazardous MG5 metagene. 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 metagene is responsible.

The Collisson subgroups are the result of junk science and can't be properly validated, though I certainly tried

1.3 Discussion

1.4 Methods

Cohort recruitment and ethics

The APGI cohort was used for metagene discovery; see **??** on page **??** for recruitment and ethics details of this cohort.

Sample collection, preparation, and gene expression microarrays

See **??** on **??** for details on sample processing for the APGI discovery cohort.

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 for categorical variates. Only three associatons were signficant at a 5% FWER level by Holm's correction; these are highlighted. Tests were also performed for the non-survival associated metagenes MG3 and MG4, and were taken into account for multiple testing correction, but their results are not reported here.

Variable	MG5	MG1	MG2	MG6
Age at diagnosis	0.7512	0.8816	0.6816	0.7516
Ethnicity	0.9939	0.6228	0.7244	0.0192
Gender	0.3733	0.1384	0.0207	0.0388
Histological subtype	0.5660	0.8934	0.5620	0.0595
Pack years smoked	0.5061	0.3481	0.2604	0.4738
Pathological grade	0.0147	0.0064	0.0057	1.87E-4
Perineural invasion	0.0642	0.3263	0.0645	0.8348
Cancer cell fraction	1.38E-5	0.0473	1.51E-5	0.1064
Recurrence site				
Bone	0.4605	0.7822	0.8047	0.0953
Brain	0.2810	0.7745	0.0299	0.3089
Liver	0.3725	0.1294	0.0469	0.4266
Lung	0.0285	0.4887	0.2979	0.0897
Lymph nodes	0.3963	0.3318	0.9865	0.7819
Mesentery	0.4057	0.5023	0.2771	0.1203
Omentum	0.3392	0.1208	0.2790	0.0596
Other	0.0692	0.6825	0.2960	0.1711
Pancreatic bed	0.9389	0.8693	0.3251	0.9235
Pancreas remnant	0.6621	0.5516	0.2399	0.2693
Peritoneum	0.8085	0.9581	0.0528	0.0276
Staging: M	0.7803	0.3965	0.3624	0.2488
Staging: N	0.1124	0.7312	0.1811	0.5892
Staging: T	0.6051	0.1377	0.5530	0.4779
Staging: Overall stage	0.1922	0.0382	0.2195	0.4592
Tumour location	0.1092	0.4698	0.2301	0.1814
Tumour longest axis length	0.8154	0.9408	0.1352	0.4037
Vascular invasion	0.7806	0.6266	0.1746	0.0862

Data preprocessing

6

[6]MP Fatal: Move this to the nomogram chapter

Table 1.3: MSigDB signatures substantially correlated to the MG2-MG6 axis.

MSigDB Signature	Correlation MG2	MG6
c2 PID integrin 1 pathway	-0.50	0.54
c2 PID integrin 3 pathway	-0.48	0.55
c2 PID UPA UPAR pathway	-0.43	0.56
c2 PID integrin5 pathway	-0.37	0.60
c2 KEGG ECM receptor interaction	-0.45	0.51
c4 GNF2 MMP1	-0.43	0.53
c2 KEGG focal adhesion	-0.43	0.52
c2 REACTOME extracellular matrix organization, c2 REACTOME collagen formation	-0.42	0.53
c2 Farmer breast cancer cluster 5, c2 Anastassiou cancer mesenchymal transition signature, c4 GNF2 CDH11	-0.43	0.51
c5 Axon guidance	-0.42	0.51
c2 PID syndecan 1 pathway	-0.37	0.55
c5 Collagen	-0.37	0.51
c6 Cordenonsi YAP conserved signature	-0.53	0.34
c3 TGANTCA v$AP1 C	-0.28	0.59
c6 LEF1 up v1 signed	-0.53	0.33
c2 PID AVB3 integrin pathway	-0.32	0.52
c3 V$AP1 Q6 01	-0.23	0.50
c3 V$AP1 Q6	-0.16	0.54
c3 V$AP1 Q4	-0.14	0.54

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` 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 disease-specific survival were identified by SIS-FAST [**?**], with a CPSS wrapper to reduce the false positive rate [**?**]. FAST statistics for time from diagnosis to disease-specific death 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 [**?**], 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.

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 [**?**]: 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 $\eta = \max(A)$.[7] The default convergence criteria of by 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 data

The following procedure was used to estimate metagene expression scores (coefficients) from gene expression measurements of a cohort. Measurements were

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

subset to the 361 outcome-associated genes identified by CPSS-SIS-FAST, 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'.

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

Associating metagene expression with clinical variables

Metagene coefficients evaluated on the APGI data were tested for association with a restricted set of the available APGI CPVs, as outlined in table 1.4. Numeric variables were tested for assocation with each metagene by Kendall's τ test; factor and boolean variables using ANOVA with the CPV as the explanatory variable. 150 tests in total were performed (25 variables, 6 metagenes), and P-values were corrected together using the Holm-Bonferroni procedure [**?**]. Corrected P-values of less than 0.05 were considered significant.

Cross-validation of the metagene discovery process

The full metagene discovery pipeline, from survival-associated gene selection by CPSS-SIS-FAST, through automatic rank estimation, NMF factorization, and LASSO model fitting, was applied to training sets in a 10-fold cross-validation loop. Test set metagene coefficients for each fold were calculated from that fold's fitted NMF W matrix as described in section 1.4, and test set coefficients were applied to the fitted LASSO model for that fold, to yield test set linear predictors. Linear predictors from each test set were combined to form a test vector containing a cross-validated prediction for each patient in the APGI cohort, each patient's prediction having been made from data that did not include that patient. Overall prognostic ability was assessed by comparing a Cox proportional hazards model predicting time from diagnosis to disease specific death using the cross-validated test vector as the sole covariate, to an intercept-only model, using a likelihood ratio test.

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,

Table 1.4: CPVs tested for association with each metagene's coefficients.

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

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.

For each validation data set, metagene coefficients were calculated as described in section 1.4. Metagene coefficients were then combined to form per-sample risk predictions using the LASSO fit produced on the APGI discovery data. The APGI survival signature was tested within each data set using likelihood ratio tests comparing a Cox model using signature risk predictions as

the sole covariate, with an intercept-only Cox model.

GSVA scoring

The expression of gene sets from the MSigDB [**?**] were estimated on the APGI cohort using a modification of the GSVA method [**?**]. GSVA with default settings was used to estimate expression scores for all MSigDB gene sets in the full 13,000 × 228 VST-scaled APGI GEX data matrix. MSigDB contains both undirected gene sets such as metabolic pathways, in which members of the set are not expected a-priori to move in concert, and directional signatures, with paired `*_UP` and `*_DN` components that would be expected to change in 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.

Metagene functional characterization

Kendall correlation coefficients were calculated between metagene coefficients and GSVA gene set scores, on the APGI expression dataset. Absolute correlations of greater than 0.5 were deemed substantive and reported for further characterisation.