List of Corrections

Mah Dissertat'n

Mark Pinese

December 9, 2014

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

Introduction

Chapter 2

A Molecular Prognostic Nomogram for Pancreas Cancer

2.1 Introduction

2.2 Results

Cohort characteristics and subsetting

2.3 Discussion

2.4 Methods

Cohort recruitment and ethics

Sample collection, preparation, and gene expression microarrays

Chapter 3

Signatures of Survival Processes in Pancreas Cancer

[1]

3.1 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 2, 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 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 [8], 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, 11]. 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 [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 [12]. 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 [12], 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 [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 factoriza-

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

3.2 Results

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

Cohort characteristics and subsetting

[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

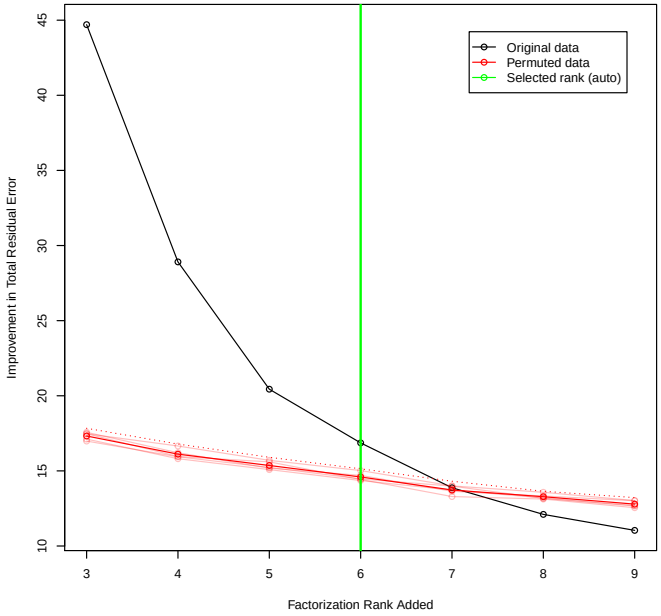


Figure 3.1: Automatic selection of factorization rank. SNMF/L was performed for varying ranks on either unpermuted data (black line) or data permuted within samples (red lines), and the improvement in total residual approximation error $\|A - WH\|_F$ calculated. The highest added rank for which the error 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 3.1).

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 3.2), and the basis matrix W was reasonably sparse (Figure 3.3). Small row L1 norm of the basis matrix is a desirable condition for this analysis, as it indicates that metagenes are

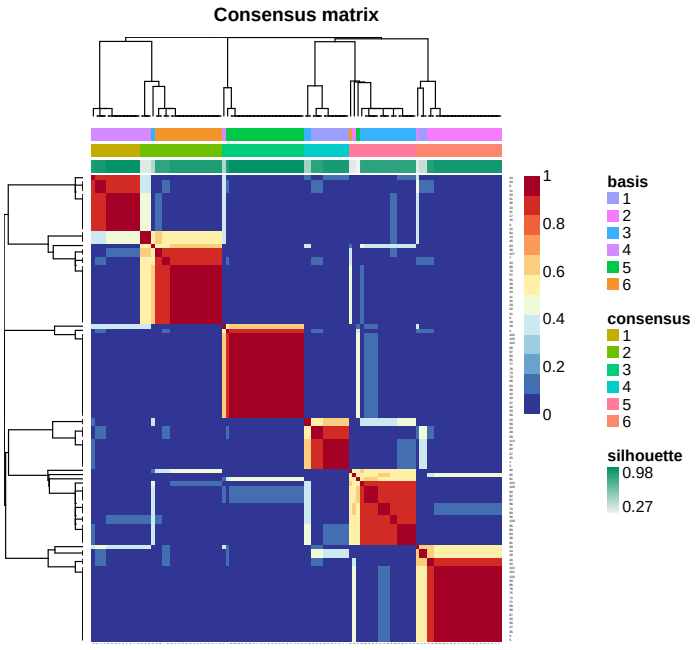


Figure 3.2: Clustering consensus matrix for the final rank-6 clustering. Colours indicate the stability of gene (in rows) and sample (in columns) clusters across random restarts of the factorization; at rank 6 this factorization was largely stable, with identical clusters assigned in all 500 random restarts 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 40.

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 APGI discovery cohort data, using non-negative least squares (NNLS)-estimated coefficients of each of the six metagenes as marginal predictors of outcome. The LASSO regularization parameter λ was chosen by 10-fold cross-validation to be the highest value for which the mean test set partial likelihood deviance was within one standard error of the lowest mean value. This resulted in a final model in which three metagenes, MG1, MG2, and MG5, were

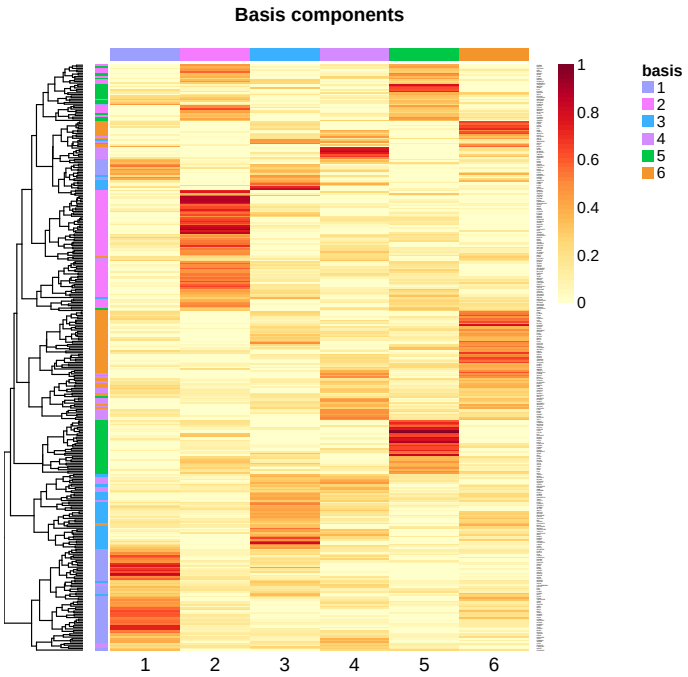


Figure 3.3: Basis matrix W of the final SNMF/L factorization. Rows represent genes, and columns metagenes, with cell colours proportional to the loading of a given gene on a given metagene. The loadings are sparse within rows, indicating that the metagenes are modular, each affecting the expression of largely distinct sets of target genes. A table of values of this basis matrix is available as app:sigs-w-matrix on page 40.

selected as prognostic (Figure 3.4).

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 3.5, 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 3.6, Kendall's τ test $P = 0.21$).

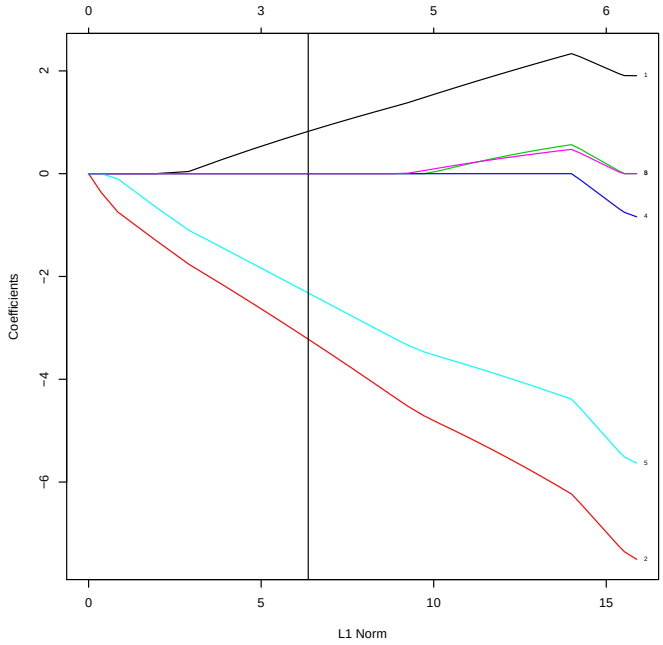


Figure 3.4: Coefficient vs penalty fit trajectories for the LASSO model predicting DSS from metagene expression. Each line represents the model coefficient for a metagene as the model is smoothly varied from a null model (L1 norm = 0), to a full unpenalised Cox fit (L1 norm ≈ 16). The vertical line indicates 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.

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 52.

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

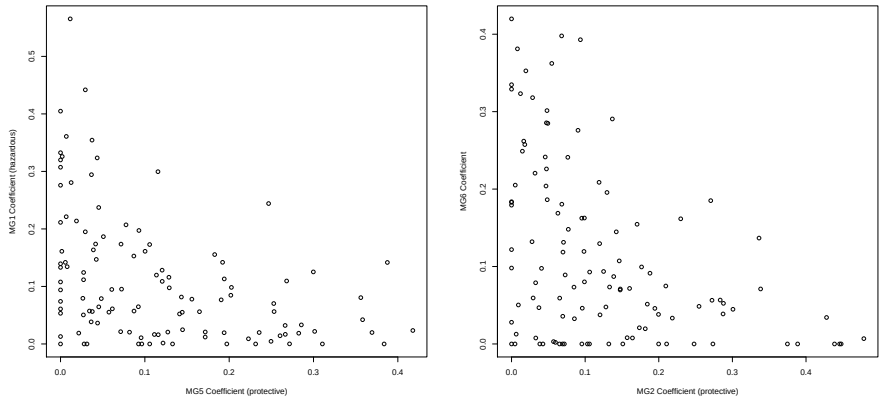


Figure 3.5: Prognostic metagenes form two axes of cell state. Metagene pairs MG1 and MG5, and MG2 and MG6, displayed mutually exclusive coefficient patterns in the APGI cohort, and could be combined to form just two axes of cell state.

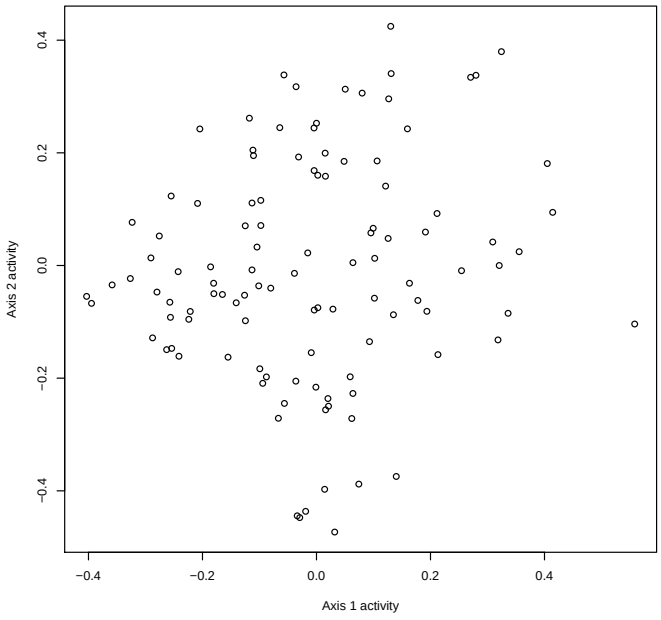


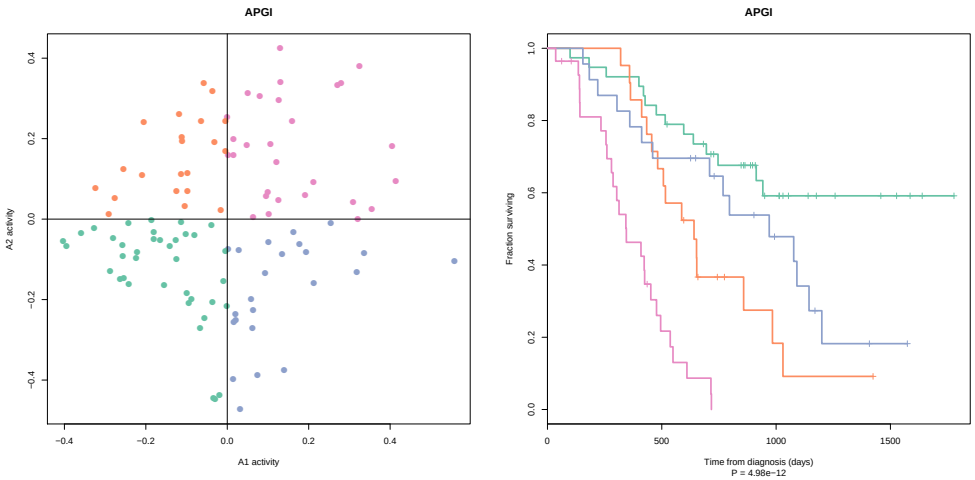
Figure 3.6: Prognostic axis signals are uncorrelated. Activity estimates of axes defined by highly correlated mutually exclusive metagene pairs (Axis A1 = MG1 - MG5, axis A2 = MG6 - MG2) were uncorrelated (Kendall τ test $P = 0.21$), indicating that these axis signals encoded orthogonal outcome-associated processes within tumours.

[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 3.1, 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 3.7.

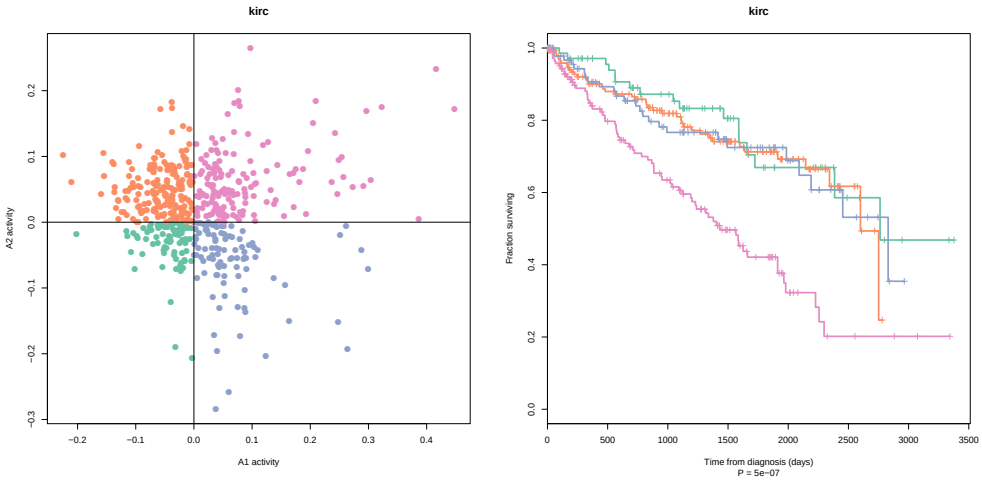
Meta-PCNA is a 130-gene signature of cell proliferation that has been found to be generally prognostic in a number of cancer cohorts [20]. To exclude the possibility that PARSE score simply recapitulated the known meta-PCNA signature, I examined whether PARSE contributed additional prognostic information 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 3.1, column b).

Table 3.1: 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

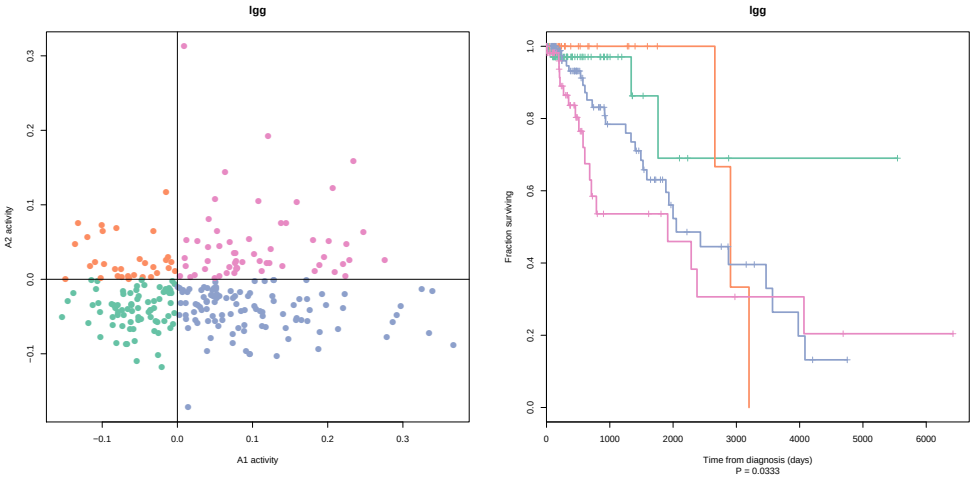


(a) APGI cohort

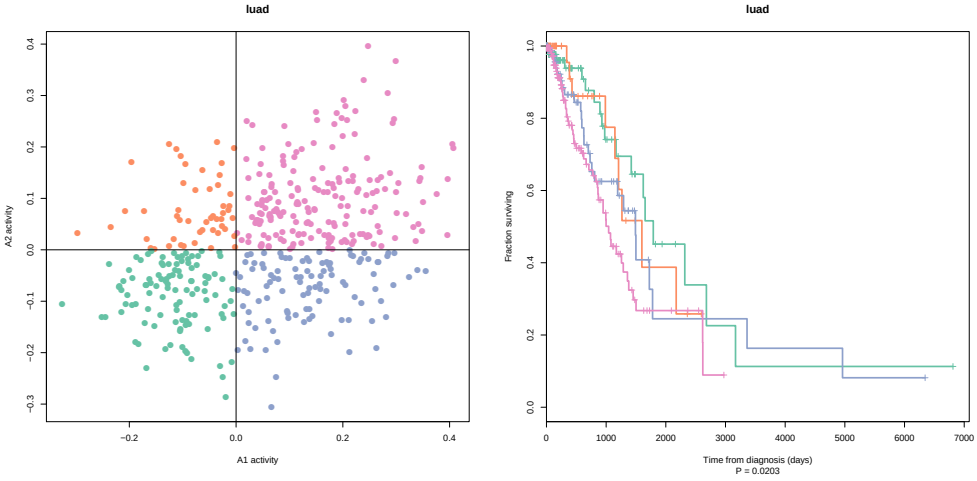


(b) TCGA kirc cohort

Figure 3.7: 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 3.7: (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 ECM remodelling 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

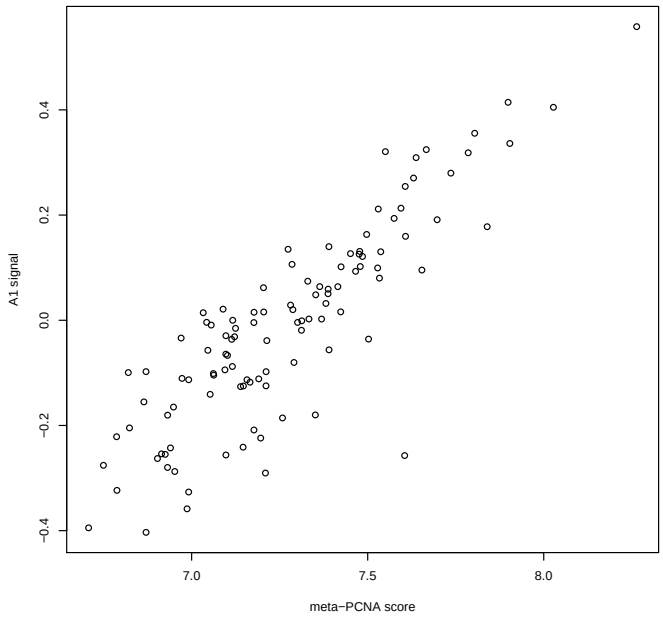


Figure 3.8: Axis A1 signal is closely associated with meta-PCNA signature score. A1 signal and meta-PCNA [20] scores were as evaluated on the APGI training set; Kendall's $\tau = 0.663$, $n = 110$, linear model $R^2 = 0.740$.

to clinical variates, meta-PCNA, 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 3.8), a relationship supported by its close association to cell cycle-related MSigDB signatures (app:sigs-msigdb-corrs-axis1 on page 49). 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 3.2), 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 3.2: 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.

MSigDB associations implicated extracellular matrix (ECM) processes, epithelial to mesenchymal transition (EMT), and possibly inflammation, in A2

activation. A2 signals were strongly positively correlated with gene set variation analysis (GSVA) scores for MSigDB signatures related to ECM interactions and remodelling, and a signature of LEF1-mediated EMT, strongly implicating matrix remodelling and invasion as the molecular basis of axis A2 (app:sigs-msigdb-corrs-axis2 on page 51). A potential link between A2 and inflammation was also indicated by a strong positive association between A2 signal and MSigDB GNF2_PTX3 GSVA score (Kendall's $\tau = 0.593$, app:sigs-msigdb-corrs-axis2 on page 51), a proxy for expression of the acute phase response protein pentraxin 3.

3.3 Discussion

3.4 Methods

Cohort recruitment and ethics

[3]

Sample collection, preparation, and gene expression microarrays

[4]

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` 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,

[3] MP Fatal: TODO: Cohort recruitment and ethics

[4] MP Fatal: TODO: Sample collection, preparation, and gene expression microarrays

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 [10], 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 [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 $\eta = \max(A)$.[5] 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

[5] Note that this parameter α is denoted β in the R `NMF` package; I use the symbol α here for consistency with [10]

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.

Calculation of the PARSE score on new 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.

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 valida-

tion. 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 [7]. 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.

Prognostic axis functional characterization

Clinical variate comparisons Prognostic axis activities calculated on the APGI data were tested for association with a restricted set of the available APGI CPVs, as outlined in Table 3.3. 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 [9]. Corrected P-values of less than 0.05 were considered significant.

Table 3.3: 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

MSigDB signature score comparisons Kendall correlation coefficients were calculated between axis activity estimates and GSVA scores for MSigDB gene sets, on the APGI expression dataset. A subset of the full MSigDB was used, as outlined in Table 3.4. Absolute correlations of greater than 0.5 were deemed substantive and reported for further characterisation.

Table 3.4: 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	*

Chapter 4

Comparative genomics

Outline ideas:

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

    - The patterns of mutations seen in common mouse models of pancreatic cancer match those consistently seen in human disease.
    - Different mouse models possess different mutation spectra, and models may be close fits to specific genetic subtypes of patients.

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

- Conclusion

4.1 Methods

Models

Sample Origin and Processing

Sequencing

QC

Mapping

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

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

Somatic SNV and Indel Detection

muTect and Strelka were used separately to detect somatic single nucleotide variants (SNVs) and insertion / deletion events (indels) in individual mouse

tumour and normal BAMs. muTect was supplied default parameters; Strelka used the parameter settings given in listing 4.1; these are the default parameters as recommended for use with the BWA mapper, with the exception that in this work isSkipDepthFilters was set to 1.

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

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

CNV and LOH Detection

Overview:

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

Loss of heterozygosity at individual loci

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

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

$$Hom = \{AA, CC, GG, TT\} \tag{4.1}$$
$$Het = \{AC, AG, AT, CG, CT, GT\} \tag{4.2}$$

The two models, HOM and HET, may be written

$$HOM : G \in Hom \tag{4.3}$$
$$HET : G \in Het \tag{4.4}$$

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

$$K = \frac{\mathcal{L}(HET)}{\mathcal{L}(HOM)} \tag{4.5}$$
$$= \frac{Pr(D|G \in Het)}{Pr(D|G \in Hom)} \tag{4.6}$$
$$= \frac{\sum_{g \in Het} Pr(D|G = g) Pr(G = g|G \in Het)}{\sum_{g \in Hom} Pr(D|G = g) Pr(G = g|G \in Hom)} \tag{4.7}$$

with D being the reads at the locus. We make the simplifying assumption that all genotypes in each of Hom and Het are equally likely, so that all

$Pr(G = g|G \in X) = \frac{1}{\|X\|}$ for $X \in \{Hom, Het\}$. Then

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

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

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

An implementation of the above heterozygous locus detection method is given in algorithm 1. The input posterior genotype likelihoods $\mathcal{L}(G = g|D)$ are supplied by `samtools mpileup -q 20 -Q 20 -v -u` operating on per-mouse normal sample BAMs, and the minimum value of K for a locus to be called as heterozygous is $\exp(minscore)$. Two additional filters are also employed in the algorithm: a locus is *not* reported as heterozygous if either the total read depth at the locus is less than *mindepth*, or if the difference in samtools-supplied log likelihood between the top two genotypes is less than *mindelta* nats. The latter filter is used to exclude any problem loci with an apparent triallelic state. [1]

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

In the general case with potential normal cell contamination of the tumour sample, it is not possible to use allelic imbalance as an indicator of LOH

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

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

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

begin

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

    return false;

  end

  `// Convert Phred-scaled likelihoods to nats`

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

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

  end

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

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

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

    return false;

  end

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

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

    case 2

      $L_{het} \longleftarrow L_2$;

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

    end

    case 3

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

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

    end

    case *default*

      return false;

    end

  endsw

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

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

    return false;

  end

  return true;

end

Algorithm 1: Determine if a locus is heterozygous

if the local copy number exceeds two. For example, in the triploid case, a LOH haplotype AAA, and a non-LOH haplotype AAB, both exhibit allelic imbalance. For this reason, allelic imbalance calls from the above test must be interpreted in the context of local copy number variation (CNV) estimates from the next procedure, and LOH calls only made if allelic imbalance is detected in regions of copy number $2n$ or less.

Copy number variation at individual loci

Problem description Considering a single locus, either a single nucleotide or a contiguous stretch of DNA, the expected number of reads from a sequencing experiment that map to that locus is proportional to the copy number of the locus in the DNA input for sequencing. Based on this relationship it is – in principle – possible to estimate copy number from sequencing data, however a number of complicating factors are present, related to sequence 'mappability', exon capture affinity, sample contamination, and problem indeterminancy.

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

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

Contamination of tumour DNA is a universal problem in solid tumour sequencing. This contamination may be with non-cancerous diploid DNA, or alternate cancer genotypes present in the same sample, or both. In the case of CNV estimation based on read depth, the presence of contaminating diploid DNA causes a shrinkage of the observed CNV profile towards that of diploid cells, and reduces the signal-to-noise ratio (SNR) of the copy number estimates. CNV callers aware of this effect must take this effect into account in

their calls, and may also be required to estimate the fraction of contaminating normal DNA. In tumour samples containing multiple tumour genotypes, with varying locus copy numbers, CNV estimates are for the mean copy number of the genotypes, weighted by their prevalence in the sample. In such cases, deconvolution of the signal into its component genotypes based on a single sample of the tumour is impossible without the benefit of additional external information.

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

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

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

We model the process of reads in NGS as a Bernoulli scheme, and use the weak dependence between read depths at different sites to derive a per-locus Poisson approximation. In this model the sequencer has a fixed s total physical sites available for sequencing; in the SOLiD 4 system these sites correspond to positions on the sequencing slide. Some of these sites yield observed sequence that is then mapped and used to estimate read depth, however many of them do not produce sequence reads, either because they are never populated with DNA, or because they fail low-level quality checks. We suppose that these

failed sites occur independent of the DNA sequence, and at a rate of r_F among all available sites. Then, a given physical sequencing site can either fail to yield sequence, with probability r_F, or it can produce observed sequence for one of m loci, each at probablity $(1 - r_F)b_i$, for $i \in \{1 \dots m\}$. This per-site categorical distribution, when sampled for each of s independent sites, results in a multinomial distribution on read depths,

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

where N_F is the number of failed sites (not observed), and N_i is the number of reads observed for locus i. The multinomial distribution induces a negative dependency on the number of reads observed at different loci, as the total read count s is fixed. However, for m large, or site failure rate r_F large[13], these negative dependencies are small, and

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

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

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

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

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

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

$$T_i \dot{\sim} Pois\,(k d_T b_i\,(1 + f\,(c_i - 1))) \tag{4.14}$$

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

The variable k is more than a convenient normalization constant: it encodes the signal expected of diploid loci in the tumour cells, and therefore

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

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

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

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

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

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

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

We test this hypothesis on R_i using the W_5 statistic of [6],

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

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

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

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

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

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

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

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

Data: An m-vector of normal locus read depths $\boldsymbol{n}$, an m-vector of tumour locus read depths $\boldsymbol{t}$, minimum normal sample depth $mindepth$.
Result: An m-vector of floats: for each locus, the one-sided P-value for CNV loss at that locus, $\boldsymbol{p_{loss}}$.

begin
  $d_N \longleftarrow \sum_{i=1}^{m} n_i$;
  $d_T \longleftarrow \sum_{i=1}^{m} t_i$;
  // Estimate k so that the modal ploidy signal will be called as diploid
  $s \longleftarrow \{(t_i/d_T) \div (n_i/d_N) : i \in \{1 \dots m\} \wedge n_i \geq mindepth\}$;
  $\hat{S} \longleftarrow KDE(s)$;
  $\hat{k} \longleftarrow mode(\hat{S})$;
  // Calculate P-values. W_5 and Φ are as defined in the text.
  $r_{H0} \longleftarrow \frac{d_T}{d_N}\hat{k}$;
  $\boldsymbol{p} \longleftarrow$ m-vector of NAs;
  for $i \leftarrow 1$ **to** m **do**
    if $n_i \geq mindepth$ **then**
      $p_i \longleftarrow \Phi\left(W_5(t_i, n_i)\right)$;
    end
  end
  return $\boldsymbol{p}$;
end

Algorithm 2: Calculate CNV loss P-values

Combining calls from adjacent loci

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

A number of approaches could be used to smooth LOH and CNV calls and share information between neighbouring loci; in this work I chose the hidden Markov model (HMM) formalism and extended the Pounds-Morris FDR

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

estimator[14] to the locality-sensitive case. The Pounds-Morris procedure fits the observed distribution of test P-values to a mixture of Uniform and Beta distributions. The Uniform distribution models the expected distribution of P-values under the null hypothesis, whereas the Beta distribution approximately fits the highly left-skewed distribution of P-values expected of tests for which the null hypothesis is false. After the observed distribution of P-values has been fit to the Beta-Uniform mixture model, the FDR associated with a given P-value can be estimated from the densities of the Beta and Uniform component distributions at that P-value.

The original Pounds-Morris procedure considers all tests as equivalent, and thus integrates no locality information, but for the LOH case combining the procedure with the locality-sensitive HMM is straightforward (figure 4.1). The HMM moves between two discrete states: *No LOH*, and *LOH*. The *No LOH* state emits a Uniform distribution of P-values, as expected under the null hypothesis of no LOH, whereas the *LOH* state emits a left-skewed Beta distribution of P-values, approximating the P-value distribution observed for loci at which the null hypothesis is false. Observed P-values at a chain of adjacent loci are fit to the HMM by standard algorithms implemented in R package depmixS4, and the posterior probability of a locus being in state *No LOH* directly gives the locality-adjusted FDR for that locus. In cases where too few extreme P-values are present to reliably estimate the parameters of the Beta distribution, the fit becomes unstable and FDR estimates potentially unreliable. To handle this situation gracefully, the method fits both the full *No LOH / LOH* model, and a restricted *No LOH* only model, and selects the model with the superior Bayesian information criterion (BIC).

Extension of the procedure to the CNV case requires three states: *Diploid*, *Loss*, and *Gain* (figure 4.2). We take advantage of the W_5 statistic's symmetry and fit the HMM to the one-sided p_{loss} CNV P-values; CNV loss is then indicated by P-values near zero, and CNV gain by P-values near one. The *Loss* and *Gain* states are modelled by Beta distributions, left-skewed in the *Loss* case, and right-skewed in the *Gain* case. The posterior probability of a locus being in state *Diploid* then gives the overall FDR for a CNV call at that locus. BIC model selection is performed as for the LOH case, except in this case four models are compared: *Diploid*, *Diploid / Loss*, *Diploid / Gain*, and *Loss / Diploid / Gain*.

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

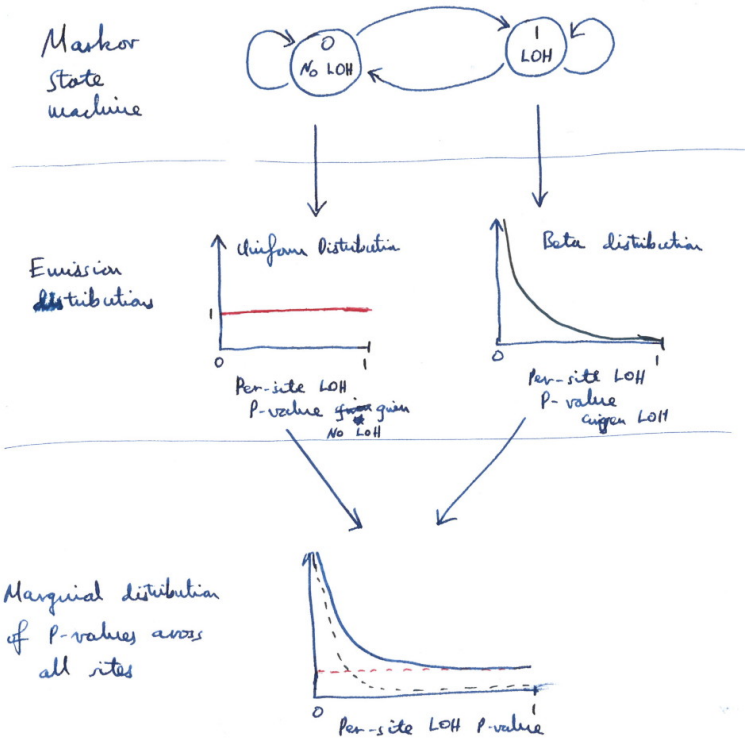


Figure 4.1: Locality-sensitive FDR estimation of LOH calls using a Markov chain Beta-Uniform mixture model.

Loss and *Gain* states (figure 4.2); although such transitions are biologically plausible, they were found to contribute to unstable fits in noisy data.

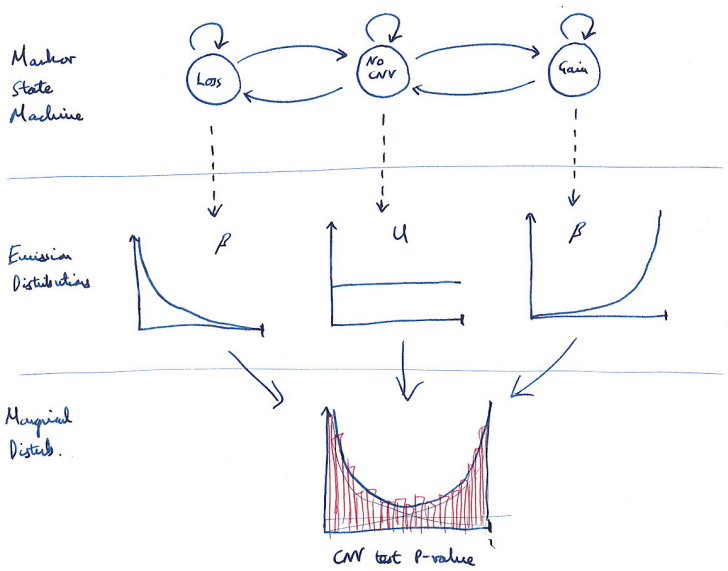


Figure 4.2: Locality-sensitive FDR estimation of CNV calls using a Markov chain double-Beta-Uniform mixture model.

Chapter 5

Conclusion

Appendices

Appendix A

Basis matrix W for the six survival-associated metagenes

	1	2	3	4	5	6
PPY	0.00	0.50	0.00	0.08	1.08	0.00
KRT6A	0.14	0.00	0.12	0.00	0.00	0.47
KRT17	0.29	0.00	0.39	0.16	0.12	0.51
DHRS9	0.00	0.00	1.00	0.34	0.00	0.17
SPP1	0.03	0.08	0.00	1.04	0.31	0.74
ADH1A	0.07	0.44	0.01	0.10	0.66	0.00
IGLL3P	0.17	0.15	0.00	0.00	0.76	0.00
DKK1	0.48	0.00	0.30	0.18	0.00	0.02
APCS	0.00	0.03	0.16	0.10	0.16	0.35
CST6	0.07	0.00	0.20	0.00	0.07	0.63
ANGPTL4	0.18	0.00	0.42	0.05	0.03	0.39
KRT7	0.46	0.00	0.56	0.00	0.14	0.44
PLAU	0.21	0.00	0.28	0.00	0.02	0.88
SCGB2A1	0.00	0.83	0.00	0.18	0.15	0.00
CCL19	0.00	0.00	0.00	0.00	0.95	0.00
CYP2S1	0.32	1.02	0.15	0.00	0.09	0.00
SLC2A1	0.18	0.12	1.00	0.41	0.00	0.70
ADM	0.00	0.00	0.52	0.51	0.00	0.36
FAM83A	0.25	0.00	0.12	0.00	0.00	0.22
FGG	0.05	0.04	0.00	0.14	0.01	0.22
KRT6C	0.12	0.00	0.00	0.00	0.00	0.16
PHACTR3	0.15	0.00	0.32	0.14	0.00	0.07
C9orf152	0.21	1.37	0.00	0.35	0.02	0.00
ALOX5AP	0.05	0.01	0.01	1.27	0.34	0.71
DCBLD2	0.40	0.00	0.12	0.00	0.14	0.84
CIDEC	0.03	0.00	0.43	0.28	0.00	0.00

FGB	0.00	0.00	0.02	0.32	0.00	0.08
SERPINB3	0.00	0.00	0.18	0.18	0.00	0.05
SLC16A3	0.13	0.38	1.10	0.42	0.00	1.00
FST	0.00	0.00	0.16	0.00	0.04	0.49
CAV1	0.42	0.00	0.19	0.08	0.27	0.84
TGFBI	0.19	0.00	0.15	0.19	0.05	1.00
COL12A1	0.00	0.13	0.03	0.53	0.19	1.65
SLC2A3	0.00	0.00	0.34	0.76	0.33	0.72
SUGCT	0.00	0.03	0.00	0.63	0.13	0.93
IL1R2	0.04	0.25	0.43	0.23	0.00	0.06
TCEA3	0.00	0.89	0.26	0.09	0.62	0.00
RAP1GAP	0.00	1.01	0.47	0.28	0.75	0.00
PXDN	0.00	0.00	0.38	0.59	0.31	1.19
FRZB	0.09	0.24	0.00	0.54	1.50	0.00
IL20RB	0.26	0.00	0.31	0.00	0.00	0.68
PLEKHS1	0.00	0.64	0.34	0.09	0.28	0.02
HSPB6	0.00	0.15	0.13	0.00	1.31	0.31
KANK4	0.00	0.00	0.20	0.47	0.00	1.23
COL7A1	0.00	0.00	0.59	0.00	0.00	0.59
C5orf46	0.00	0.00	0.00	1.06	0.13	1.04
VSTM2L	0.32	0.00	0.94	0.00	0.05	0.07
PTGES	0.57	0.02	0.57	0.07	0.00	0.56
FSCN1	0.37	0.07	1.06	0.13	0.14	0.74
CTSV	0.30	0.04	0.26	0.02	0.02	0.18
SPOCK1	0.12	0.00	0.03	0.52	0.34	1.27
RGS5	0.00	0.43	0.05	0.08	0.58	0.09
PHLDA1	0.08	0.14	0.72	0.13	0.62	1.50
IGFBP1	0.27	0.00	0.23	0.03	0.00	0.01
BAMBI	0.11	0.00	0.84	0.39	0.24	0.17
FLRT3	0.79	0.13	0.51	0.28	0.22	0.31
DSG3	0.18	0.00	0.21	0.00	0.00	0.54
ANGPTL2	0.00	0.00	0.37	0.87	0.18	0.92
ST6GAL1	0.17	0.84	0.00	0.23	0.67	0.09
SLC40A1	0.00	0.89	0.00	0.58	0.24	0.16
EMP3	0.25	0.00	0.46	0.16	0.22	0.56
RAB31	0.11	0.00	0.26	0.87	0.76	1.19
ST6GALNAC1	0.04	1.00	0.08	0.12	0.00	0.10
ACKR3	0.00	0.00	0.38	0.36	0.21	0.58
SLC12A2	0.04	0.91	0.34	0.10	0.49	0.18
ANKRD22	0.41	1.35	0.17	0.27	0.04	0.22
ENO2	0.36	0.34	0.79	0.03	0.00	0.94
EPHX2	0.00	0.59	0.11	0.17	0.68	0.00
MCEMP1	0.00	0.00	0.00	0.61	0.00	0.30
CDA	0.29	0.00	0.34	0.00	0.00	0.70

PLIN2	0.31	0.00	0.08	1.02	0.47	0.21
SERPINH1	0.00	0.01	0.39	0.22	0.43	1.02
FAM134B	0.00	0.82	0.00	0.23	0.21	0.00
NFIX	0.00	0.88	0.14	0.00	1.39	0.80
LYNX1	0.03	0.00	0.26	0.17	0.00	0.10
LDHA	0.65	0.47	0.00	0.32	0.05	1.17
SOD2	0.58	0.12	0.00	0.47	0.40	0.17
PCDH20	0.00	0.43	0.00	0.15	0.00	0.00
ITGA5	0.00	0.00	0.48	0.27	0.12	0.68
ZNF185	0.25	0.17	1.02	0.48	0.00	0.72
PLOD2	0.15	0.09	0.24	0.29	0.17	0.89
TNFRSF6B	0.63	0.00	0.07	0.18	0.00	0.39
MME	0.00	0.00	0.06	0.45	0.04	0.58
MRAP2	0.04	0.78	0.00	0.22	0.23	0.00
PLAC9	0.07	0.00	0.00	0.11	1.29	0.08
ERRFI1	0.16	0.03	0.55	0.35	0.29	0.79
PP7080	0.10	0.97	0.00	0.04	0.00	0.00
DSG2	0.43	0.57	0.18	0.51	0.04	0.71
APCDD1	0.00	0.14	0.15	0.13	0.60	0.84
PRKCDBP	0.26	0.00	1.02	0.51	0.26	0.59
SULF2	0.17	0.15	0.46	0.19	0.39	0.77
TUBA1C	1.31	0.55	0.54	0.53	0.27	0.50
PCOLCE2	0.00	0.01	0.12	0.54	0.00	0.05
LAMA5	0.37	0.08	1.02	0.00	0.34	0.18
P4HA1	0.04	0.10	0.41	0.84	0.00	0.55
RASL11B	0.00	0.19	0.07	0.22	1.21	0.31
KYNU	0.61	0.09	0.07	0.54	0.00	0.28
CTSL	0.39	0.00	0.20	1.18	0.47	0.22
MARCKSL1	0.15	1.34	0.30	0.00	0.00	0.26
PRC1	0.96	0.35	0.04	0.04	0.00	0.32
C1QTNF6	0.00	0.00	0.59	0.62	0.22	0.97
CCR7	0.06	0.00	0.00	0.00	1.05	0.00
HRASLS2	0.33	0.00	0.30	0.22	0.00	0.00
CHN2	0.00	0.50	0.00	0.34	0.44	0.00
PYGL	0.08	0.00	0.31	0.34	0.14	0.74
MELK	1.02	0.29	0.00	0.23	0.01	0.22
LOX	0.21	0.00	0.08	0.39	0.09	0.92
CDC45	0.96	0.08	0.11	0.34	0.03	0.00
AXIN2	0.00	0.52	0.44	0.13	0.81	0.29
ATL3	0.64	0.03	0.16	0.49	0.25	0.29
CAMK1G	0.09	0.24	0.00	0.03	0.88	0.00
ABLIM1	0.01	0.91	0.32	0.00	0.61	0.34
TRIM2	0.13	1.15	0.31	0.31	0.36	0.00
TWIST1	0.00	0.00	0.20	0.91	0.12	1.20

ARSD	0.15	1.24	0.19	0.00	0.22	0.14
CEBPB	0.07	0.07	1.29	0.53	0.51	0.81
CEP55	1.42	0.33	0.00	0.17	0.00	0.46
GINS2	1.08	0.18	0.39	0.07	0.00	0.00
MCM4	1.28	0.14	0.31	0.03	0.01	0.13
PPP1R3C	0.00	0.02	0.13	0.37	0.03	0.26
MTRNR2L1	0.28	0.56	0.49	0.07	0.55	0.00
CDK12	0.19	0.28	0.00	0.08	0.83	0.00
BIRC5	1.38	0.17	0.37	0.55	0.00	0.24
SPHK1	0.26	0.00	0.27	0.09	0.62	1.41
A4GALT	0.03	0.00	1.30	0.08	0.36	0.52
ICAM2	0.50	0.20	0.48	0.31	0.40	0.13
ANKRD37	0.06	0.18	0.22	0.72	0.01	0.57
STK39	0.15	1.00	0.24	0.14	0.08	0.12
ASPM	1.17	0.39	0.20	0.17	0.04	0.04
PFKFB4	0.55	0.22	0.68	0.43	0.14	0.29
IDH2	0.71	0.43	0.40	0.21	0.33	0.23
SGSM1	0.00	0.93	0.08	0.02	0.84	0.00
SELENBP1	0.00	1.20	0.36	0.20	0.26	0.00
P4HA2	0.32	0.17	0.12	0.54	0.11	0.74
LMO3	0.00	0.11	0.00	0.01	1.18	0.01
KLHL5	0.42	0.16	0.00	0.35	0.70	1.14
HIPK2	0.26	1.25	0.07	0.24	0.52	0.00
NAMPT	0.34	0.00	0.05	0.75	0.32	0.35
NCAPG	1.61	0.44	0.00	0.00	0.00	0.52
PLOD1	0.06	0.00	1.21	0.75	0.37	0.80
C2orf70	0.11	1.09	0.02	0.00	0.00	0.00
RERGL	0.24	0.00	0.00	0.11	1.18	0.00
CFDP1	0.35	0.55	0.74	0.67	0.00	0.26
RACGAP1	1.37	0.37	0.14	0.19	0.07	0.33
SNRPB	0.99	0.08	0.41	0.90	0.02	0.00
CLEC3B	0.06	0.07	0.12	0.01	0.81	0.00
ANLN	1.17	0.24	0.08	0.08	0.00	0.72
ZFPM1	0.00	1.22	0.29	0.00	0.43	0.15
UPP1	0.55	0.00	0.79	0.43	0.16	0.11
AURKB	1.00	0.11	0.14	0.00	0.01	0.00
SYNE2	0.00	0.88	0.24	0.00	0.28	0.28
SOBP	0.00	0.20	0.81	0.10	1.36	0.00
GAPDH	0.48	0.39	0.83	0.24	0.00	0.72
SERTAD2	0.29	0.14	0.90	0.99	0.49	0.44
TPX2	1.32	0.15	0.04	0.15	0.04	0.11
POC1A	1.38	0.33	0.32	0.47	0.00	0.00
PDLIM7	0.20	0.00	0.41	0.37	0.11	0.68
TSTD1	0.17	1.22	0.48	0.07	0.45	0.02

PLIN3	0.34	0.26	0.97	0.93	0.14	0.41
IL33	0.24	0.04	0.00	0.13	0.68	0.00
CA8	0.00	0.69	0.05	0.01	0.05	0.00
SAMD5	0.13	0.54	0.00	0.00	0.09	0.00
NFIA	0.12	0.84	0.00	0.39	1.50	0.27
KCTD5	0.38	0.51	1.13	0.61	0.00	0.00
CCNB1	1.43	0.46	0.13	0.25	0.02	0.36
TM9SF3	0.00	1.08	0.22	0.00	0.16	0.21
KIF20A	1.37	0.29	0.21	0.23	0.00	0.29
PROSER2	0.93	0.18	0.40	0.37	0.27	0.40
COLGALT1	0.40	0.16	0.62	0.43	0.16	0.88
PPM1H	0.00	0.85	0.46	0.27	0.24	0.00
NCAPD2	1.38	0.41	0.16	0.12	0.20	0.32
PREP	0.06	0.98	0.30	0.20	0.02	0.00
DPY19L1	0.34	0.36	0.30	0.54	0.08	0.51
CKAP2L	1.78	0.22	0.27	0.03	0.00	0.09
ZBED2	0.16	0.00	0.18	0.00	0.00	0.64
MIR99AHG	0.04	0.28	0.39	0.45	1.79	0.22
P2RY2	0.18	0.03	0.77	0.22	0.00	0.50
KIF2C	0.80	0.13	0.11	0.01	0.00	0.00
PPP1R14B	0.37	0.26	0.78	0.00	0.37	0.59
GPC3	0.10	0.23	0.00	0.00	1.27	0.00
MAP3K8	0.20	0.00	0.07	0.31	0.56	0.43
NMB	0.21	0.19	0.66	0.79	0.00	0.36
RAVER2	0.20	0.91	0.05	0.09	0.27	0.06
SPIN4	0.85	0.32	0.80	0.39	0.22	0.40
AMOT	0.07	0.82	0.14	0.52	0.43	0.57
POP5	0.56	0.51	1.52	0.23	0.11	0.18
COLGALT2	0.00	0.60	0.00	0.02	0.00	0.00
DCUN1D5	1.36	0.08	0.00	0.86	0.96	0.72
DNAJC9	0.78	0.11	0.37	0.12	0.13	0.15
KCTD10	0.38	0.13	0.29	0.44	0.51	0.79
MIF	0.43	0.33	0.96	0.44	0.00	0.68
SLAMF9	0.04	0.00	0.00	0.67	0.00	0.07
MCOLN2	0.20	0.28	0.00	0.00	0.94	0.00
CSNK1D	0.21	0.38	1.56	0.48	0.16	0.23
TMED1	0.26	0.34	1.15	0.83	0.49	0.28
CADPS2	0.26	1.29	0.00	0.55	1.02	0.57
MEOX1	0.00	0.05	0.16	0.04	0.96	0.00
GIMAP2	0.15	0.72	0.00	0.66	0.77	0.00
RFC5	1.08	0.24	0.00	0.52	0.16	0.31
CARHSP1	0.75	0.53	0.87	0.90	0.26	0.00
SLC15A1	0.00	0.00	0.48	0.00	0.06	0.06
BCL11B	0.20	0.92	0.23	0.24	0.42	0.00

CDK2	1.06	0.25	0.01	0.52	0.33	0.33
KIAA1549L	0.38	0.08	0.26	0.66	0.15	0.64
HJURP	1.33	0.24	0.23	0.02	0.00	0.00
FYN	0.01	0.52	0.12	0.13	1.69	0.87
RNF103	0.03	1.25	0.17	0.55	0.29	0.06
ACYP2	0.25	0.89	0.00	0.23	0.85	0.41
CD70	0.09	0.00	0.21	0.36	0.00	0.43
PPAPDC1A	0.00	0.00	0.00	0.76	0.00	1.22
TPD52L2	0.63	0.16	1.31	0.65	0.44	0.23
TOM1	0.00	0.10	1.49	0.81	0.68	0.52
DERA	1.18	0.20	0.46	0.60	0.29	0.32
TREM1	0.05	0.00	0.09	0.71	0.00	0.30
UFC1	0.00	1.19	0.25	0.47	0.30	0.00
TCTA	0.00	0.75	0.82	0.09	0.98	0.02
ALDH5A1	0.10	0.99	0.55	0.06	0.90	0.22
KNTC1	1.07	0.14	0.44	0.08	0.15	0.28
XXYLT1	0.24	0.00	1.05	1.08	0.46	0.87
SMOX	0.37	0.29	1.43	1.00	0.18	0.00
ARFGAP3	0.03	0.30	0.54	0.84	0.49	0.54
SEPW1	0.03	0.95	1.24	0.00	0.63	0.56
ANKLE2	0.75	0.14	0.62	0.51	0.19	0.38
TLE4	0.05	0.88	0.07	0.33	0.90	0.47
RBMS2	0.61	0.15	0.00	0.40	0.32	0.89
AKR1A1	0.25	1.08	0.26	0.29	0.66	0.45
RERE	0.05	0.74	0.62	0.00	0.99	0.42
ATAD2	0.94	0.07	0.11	0.03	0.11	0.31
SPOCD1	0.00	0.00	0.18	0.21	0.00	0.76
DYNC2H1	0.00	1.61	0.15	0.00	0.76	0.67
CAPN6	0.00	0.75	0.00	0.23	0.64	0.00
RPIA	0.46	1.35	0.22	0.19	0.46	0.00
P2RY8	0.23	0.07	0.00	0.28	1.66	0.00
ARHGEF19	0.08	0.08	1.20	0.52	0.45	0.51
ARL4C	0.00	0.02	0.30	0.49	0.30	1.23
CHAF1B	0.99	0.30	0.20	0.02	0.52	0.10
FHDC1	0.18	1.24	0.22	0.02	0.00	0.05
POLA2	0.84	0.22	0.33	0.13	0.21	0.00
AGRP	0.00	0.00	0.00	0.68	0.00	0.17
RPA2	0.47	0.70	0.70	0.41	1.42	0.24
MRPL24	0.16	1.13	0.22	0.12	0.22	0.18
PRDM16	0.00	1.12	0.00	0.00	0.53	0.09
POU2AF1	0.06	0.47	0.00	0.00	0.92	0.00
MC1R	0.10	0.13	1.08	0.87	0.47	0.13
TNFRSF17	0.03	0.05	0.00	0.08	0.58	0.00
FAH	0.68	0.42	0.36	0.21	0.32	0.39

HSP90B1	0.53	0.46	0.78	0.90	0.30	0.38
TRAPPC2	0.51	1.08	0.00	0.49	0.62	0.14
ARHGAP24	0.06	1.06	0.02	0.75	1.10	0.62
ABHD16A	0.66	0.72	0.00	0.00	0.52	0.22
TMTC4	0.00	1.29	0.33	0.21	0.20	0.28
SCYL2	0.70	0.39	0.00	0.98	0.41	0.96
TOR2A	0.00	0.99	0.48	0.20	0.53	0.00
IKBIP	0.29	0.00	0.30	1.12	0.15	0.47
DENND1A	0.82	0.00	0.25	0.19	0.00	0.18
BCKDK	0.22	0.29	0.87	1.07	0.40	0.11
KIAA0513	0.08	1.04	0.17	0.32	0.59	0.00
CNNM1	0.00	0.87	0.41	0.00	0.09	0.00
VPS35	0.39	1.39	0.00	0.53	0.00	0.25
ZPLD1	0.00	0.00	0.19	0.03	0.03	0.11
CHEK1	1.52	0.16	0.00	0.00	0.11	0.27
PEX11B	0.11	1.35	0.00	0.53	0.29	0.25
BTN3A1	0.66	0.71	0.07	0.25	0.99	0.30
FBXO22	0.50	0.36	0.00	0.58	0.00	0.31
BBS2	0.25	1.14	0.00	0.22	1.00	1.16
DCAF8	0.00	1.14	0.48	0.11	0.53	0.19
ITPKB	0.00	0.83	0.61	0.00	1.19	0.67
SH3GL1	0.12	0.11	1.01	1.25	0.22	0.00
PBXIP1	0.00	0.51	0.41	0.00	0.44	0.17
GAB2	0.04	0.74	0.38	0.64	1.36	0.27
NACC2	0.53	0.00	0.72	0.25	0.00	0.11
EXOSC8	0.93	0.60	0.28	1.02	0.37	0.15
ATF7IP2	0.00	0.20	0.12	0.00	0.03	0.00
MCM10	1.14	0.14	0.00	0.01	0.00	0.08
PGAM5	0.92	0.00	0.39	0.49	0.00	0.00
AKIP1	0.64	0.24	0.60	0.71	0.78	0.72
STAT5B	0.00	0.91	0.32	0.06	1.31	0.22
KIF14	1.12	0.36	0.20	0.43	0.00	0.13
FAM189A2	0.00	1.00	0.00	0.02	0.11	0.00
GNPAT	0.17	0.95	0.14	0.44	0.18	0.19
PAX8	0.77	0.00	0.56	0.00	0.00	0.00
GABPB1	0.74	0.20	0.00	0.74	0.22	0.67
TARBP2	0.68	0.38	1.22	0.61	0.18	0.00
ABHD5	0.15	0.75	0.00	0.75	0.40	1.17
NUP155	1.13	0.41	0.06	0.33	0.23	0.46
FAM120AOS	0.18	1.05	0.00	0.28	0.71	0.57
CATSPER1	0.12	0.00	0.92	0.00	0.00	0.10
RFK	0.00	0.66	0.12	0.00	0.43	0.21
CIDECP	0.11	0.02	0.52	0.28	0.11	0.00
CACHD1	0.00	0.69	0.02	0.00	1.08	0.49

NR0B2	0.00	0.84	0.00	0.00	0.14	0.00
TMEM26	0.04	0.02	0.10	0.49	0.22	1.45
NELFE	0.94	0.23	0.59	0.86	0.36	0.08
ZSCAN16	0.30	1.45	0.00	0.02	0.51	0.51
FAM91A1	0.98	0.20	0.16	0.79	0.00	0.27
PHOSPHO2	0.34	1.07	0.00	0.47	0.41	0.05
KCNQ3	0.00	0.13	0.17	0.78	0.09	0.52
RHOF	0.75	0.17	0.48	0.14	0.00	0.59
COX4I2	0.00	0.17	0.07	0.00	0.99	0.33
MARS2	0.75	1.02	0.00	0.40	0.50	0.00
BOC	0.00	0.00	0.32	0.00	1.61	0.00
ZSCAN32	0.35	1.16	0.50	0.30	0.73	0.24
PCF11	0.26	0.94	0.25	0.10	1.11	0.41
SEC23IP	0.34	1.30	0.00	0.53	0.36	0.46
E2F7	1.04	0.00	0.03	0.02	0.00	0.54
COL5A3	0.00	0.00	0.18	0.04	0.07	1.03
SNORA11D	0.08	0.27	0.48	0.44	0.00	0.27
OAZ1	0.86	0.59	0.66	1.12	0.52	0.59
LETM2	0.44	0.00	0.39	0.00	0.00	0.28
EIF2AK3	0.18	1.27	0.00	0.38	0.61	0.33
ST3GAL2	0.34	0.00	0.80	1.07	0.44	0.00
PRR11	0.82	0.05	0.23	0.00	0.00	0.09
ELMOD3	0.00	1.16	0.69	0.39	0.53	0.09
CNIH3	0.00	0.06	0.00	0.32	0.00	0.60
B3GALTL	0.36	0.33	0.56	0.38	0.49	0.77
PRMT7	0.14	1.50	0.44	0.00	0.18	0.22
FGD6	0.55	0.00	0.13	0.14	0.00	0.50
TRERF1	0.49	0.29	0.38	0.13	0.05	0.13
RALGAPB	1.00	0.50	0.29	0.76	0.26	0.80
UHRF2	0.15	0.29	0.33	0.50	0.66	1.10
GOLM1	0.00	0.71	0.12	0.05	0.00	0.00
PAX8-AS1	0.57	0.04	0.34	0.07	0.01	0.00
THSD7B	0.09	0.20	0.00	0.29	0.96	0.11
TAF5L	0.22	1.06	0.18	0.24	0.23	0.22
PPP1R12B	0.17	0.32	0.78	0.63	0.03	0.49
LINC01184	0.63	0.80	0.00	0.34	0.81	0.00
RFX2	0.00	0.22	0.24	0.00	0.46	0.30
WNT2B	0.09	0.11	0.00	0.01	0.45	0.00
TOM1L2	0.19	0.00	0.63	0.33	0.05	0.23
TNFRSF10D	0.15	0.11	0.66	0.46	0.00	0.18
GATA6	0.05	0.88	0.09	0.14	0.19	0.00
SDIM1	0.00	0.05	0.24	0.00	0.50	0.00
ZNF658	0.00	0.88	0.00	0.00	0.91	0.28
IFT140	0.00	1.09	0.52	0.00	0.26	0.07

LGALS9B	0.11	1.02	0.00	0.00	0.35	0.49
LMTK2	0.74	0.36	0.31	0.53	0.02	0.24
FER	0.50	0.10	0.18	0.44	0.18	0.87
NRP2	0.15	0.00	0.50	0.00	0.00	0.05
EYA3	0.00	0.09	0.53	0.00	0.00	0.91
ZNF565	0.07	0.29	0.07	0.06	0.24	0.08
GATC	1.02	0.11	0.00	0.48	0.07	0.47
CCDC88A	0.00	0.17	0.47	0.01	0.80	1.02
USP30	0.54	0.14	0.39	0.00	0.08	0.00
LOC100506562	0.58	0.29	0.60	0.60	0.11	0.11
RMND5A	0.27	0.12	0.26	0.71	0.00	0.07
FBXW8	0.25	0.26	0.66	0.93	0.18	0.33
EDIL3	0.00	0.00	0.00	0.86	0.01	0.82
A4GNT	0.00	0.74	0.05	0.05	0.37	0.07
ORC1	0.98	0.32	0.16	0.95	0.12	0.01
FEM1B	0.30	0.30	0.00	0.00	0.08	1.42
SLC30A3	0.45	0.50	0.08	0.21	0.66	0.07
C1orf56	0.00	0.87	0.00	0.37	0.11	0.36
NEURL2	0.69	0.12	0.00	0.26	0.72	0.43
PGBD3	0.62	0.36	0.43	0.20	0.56	0.74
PTPN21	0.27	0.17	0.32	0.49	0.27	0.84
LCNL1	0.11	0.28	0.01	0.27	0.53	0.00
ACE	0.03	0.83	0.05	0.00	0.00	0.18
NPM1	0.00	1.05	0.00	0.00	0.08	0.04
RGS3	0.24	0.12	0.00	0.81	0.23	0.32
PIGL	1.06	0.15	0.56	0.30	0.24	0.00
GPR176	0.43	0.31	0.00	0.74	0.37	0.59

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, non-negative. In the reverse problem of PARSE calculation, A and $\hat{W}$ are supplied, and H is to be estimated. I propose an approximation that removes the requirement that H be non-negative, $\hat{H} = \hat{W}^{+}A$, where $\hat{W}^{+}$ is the Moore-Penrose pseudoinverse of $\hat{W}$.
[1]

[1]MP Fatal: TODO: finish calculation of this approximation. Also consider a low-cardinality version.

Glossary

APGI Australian Pancreatic Cancer Genome Initiative. 3, 5, 6, 8, 9, 11, 13–15, 20–22

BAM binary sequence alignment / map file. 25, 26, 28

BIC Bayesian information criterion. 35

BMC Bayesian model comparison. 27

CNV copy number variation. v, 30, 31, 33–35

CPSS complementary pair subset selection. 6, 18, 20

CPV clinico-pathological variable. iv, 5, 6, 16, 18, 21, 22

CV cross-validation. 10

DSD disease-specific death. 6, 18

DSS disease-specific survival. iii, 10, 18

ECM extracellular matrix. 14, 16, 17

EMT epithelial to mesenchymal transition. 16, 17

FAST feature aberration at survival times. 6, 18, 20

FDR false-discovery rate. 6, 18, 35

FWER familywise error rate. 12

GATK Genome analysis toolkit. 25

GEO Gene Expression Omnibus. 20

GEX gene expression. 3–6, 18, 21

GSVA gene set variation analysis. 17, 21, 22

HMM hidden Markov model. 34, 35

IDAT Illumina data. 17

indel insertion / deletion event. 25

LA-BQSR local alignment and base quality score recalibration. 25

LASSO least absolute shrinkage and selection operator. iii, 8–10

LOH loss of heterozygosity. 27, 28, 30, 34, 35

MDS multidimensional scaling. 18

MSigDB molecular signatures database. i, iv, 4, 15–17, 21–23, 49–51

NCBI National Center for Biotechnology Information. 20

NGS next-generation sequencing. 30, 31, 33

NMF non-negative matrix factorization. iii, 5–8, 18, 52

NNLS non-negative least squares. 8, 10, 20, 52

PARSE prognostic axis risk stratification estimate. i, iii, iv, 10, 12–15, 20, 21, 52

PDAC pancreatic ductal adenocarcinoma. 3–6, 10, 14, 21

SIS sure independence screening. 6, 18, 20

SNMF/L sparse non-negative matrix factorization, long variant. iii, 7–9, 18–20

SNR signal-to-noise ratio. 30

SNV single nucleotide variant. 25

TCGA The Cancer Genome Atlas. iv, 12–14, 20

VST variance stabilizing transform. 17, 18, 21

WGS whole genome sequencing. 31

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

[7] 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.

[8] 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.

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

[10] 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.

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

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

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

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

[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.