Combining algorithms to find signatures that predict risk in early stage stomach cancer

J. B. Nation[1], Justin Cabot-Miller[2], Oren Segal[3], Robert Lucito[4], and Kira Adaricheva[5]*

[1]Department of Mathematics, University of Hawaii

jb@math.hawaii.edu

[2]Google Corporation

justinmil@google.com

[3]Department of Computer Science, Hofstra University

oren.segal@hofstra.edu

[4]Zucker School of Medicine at Hofstra-Northwell

robert.lucito@hofstra.edu

[5]Department of Mathematics, Hofstra University

kira.adaricheva@hofstra.edu

*Corresponding author

July 2, 2021

Abstract

This study applied two mathematical algorithms, LUST and D-basis, to the identification of prognostic signatures from cancer gene expression data. The LUST algorithm looks for metagenes, which are sets of genes that are either over-expressed or under-expressed in the same patients. While LUST runs unsupervised by clinical data, the D-basis algorithm uses implications and association rules to relate gene expression to clinical outcomes. The D-basis selects a small subset of the metagene (a signature) to predict survival.

The two algorithms, LUST and D-basis, were combined and applied to mRNA expression and clinical data from TCGA for 203 stage 1 and 2 stomach cancer patients. Two small (4-gene) signatures effectively predict survival in early-stage stomach cancer patients. These signatures could be used as a guide for treatment.

The first signature (DU4) consists of genes that are under-expressed on the long-survival/low-risk group: FLRT2, KCNB1, MYOC, TNXB. The second signature consists of genes that are over-expressed on the short-survival/high-risk group: ASB5, SFRP1, SMYD1, TACR2. Another 9-gene signature (REC9) predicts recurrence: BNC2, CCDC8, DPYSL3, MOXD1, MXRA8, PRELP, SCARF2, TAGLN, ZNF423. Each patient is assigned a score that is a linear combination of the expression levels for the genes in the signature. Scores below a selected threshold predict low risk/long survival, while high scores indicate a high risk of short survival.

The metagenes associate with TCGA cluster C1. Both our signatures and cluster C1 identify tumors that are genomically silent, and have a low mutation load or mutation count. Furthermore, our signatures identify tumors that are predominantly in the WHO classification of poorly cohesive and the Lauren class of diffuse samples, which have a poor prognosis.

Keywords: algorithms, cancer genomics, computational molecular biology, gene expression, pathways, stomach cancer, signature, survival

1 Introduction

The American Cancer Society estimates that 26,560 people in the United States will be diagnosed with stomach cancer in 2021 and an estimated 11,180 deaths will occur that year (American Cancer Soc. 2020). While the United States has one of the lowest rates for stomach cancer, other countries such as South Korea have close to 8 times the number of cases each year (population adjusted) (Ferlay et al. 2018; Prashanth and Barsouk 2018). Worldwide, stomach cancer is one of the most common cancer types and one of the top causes of cancer deaths.

As with many other cancers, prognosis for stomach cancer correlates well with the stage of the tumor. Five year survival for stage 1 cancer is approximately 65%, for stage 2 survival drops to approximately 35%, and for stage 3 the survival drops to approximately 25% (McLoughlin 2004). When the tumor reaches stage 4, there are no survival statistics, since virtually no one survives. The survival rate for stage 1 stomach cancer is surprisingly low compared to other cancers such as breast, ovarian, prostate, colon and melanoma stage 1 tumors (close to 90%).

The transcriptome can be used to identify tumor types as well as prognosis (Sørlie et al. 2001). A number of methods have been used to identify genes of interest including clustering, principal component analysis, and more recently machine learning methods (Alkhateeb et al. 2019). Genetic information can be used to understand the tumors at a molecular level and to predict clinical outcomes. A patient's prognosis can be a factor in determining treatment.

Our analysis used publicly available data comprising mRNA expression and clinical data from TCGA. Two innovative algorithms were combined to analyze the expression and outcome data, which identified two genetic signatures for predicting survival in stage 1 and stage 2 stomach cancer patients.

2 Methods

2.1 Dataset

mRNA gene expression and clinical data were downloaded for 415 stomach cancer patients from The Cancer Genome Atlas (TCGA). We then segregated the data for stage 1 and 2 stomach adenocarcinoma, resulting in 203 patients.

The mean expected survival time for this group based on the ecdf (empirical cumulative distribution function) is 842 days. Many of the patients (112) are censored at fewer than 842 days. Also, we want to leave a window around the mean: those dying or censored near the mean should perhaps not be classified as long- or short-term survivors (15 patients). Thus we restrict our data pool to those who died on or before 661 days from diagnosis (47 short-term survivors) and those who survived at least 1023 days (29 long-term survivors), for a total of 76 samples.

2.2 Analysis with LUST and D-basis

We use two totally different algorithms to identify candidate genes for signatures from the 20,531 genes in the TCGA data.

- The LUST algorithm looks for metagenes, which are sets of genes that either over-express or under-express in the same patients. A description of the LUST algorithm, and MATLAB code implementing it, can be found at the LUST github site[1]. The github site gives the metagenes found with each of the 33 types of cancer in the TCGA database.
- While LUST runs unsupervised by clinical data, the D-basis algorithm uses implications and association rules to relate gene expression to clinical outcomes, in this case survival and recurrence. A description of the D-basis algorithm, and C++ code

[1] https://github.com/tristanh314/lust-cancer-2019

implementing it, can be found at the D-basis github site[2].

The algorithms are run sequentially. LUST is run on expression for the entire TCGA gene set, and identifies the metagenes as clusters of candidates. D-basis takes the metagenes output by LUST and refines them to small prognostic signatures. The workflow is diagrammed in Figure 1.

Note that both algorithms look for signals that represent variation within cancer patients, not differences between tumors and normal tissue. A mathematical description of these algorithms is given in Appendix I.

2.3 Pathway analysis

Pathway analysis for metagenes R1, R2, and R3 was performed using Database for Annotation, Visualization, and Integrated Discovery (DAVID) (Huang da et al., 2009; version 6.7, https://david-d.ncifcrf.gov) based on Kyoto Encyclopedia of Genes and Genomes (KEGG) database and using the Panther Classification System (http://www.pantherdb.org/) based on reactome pathways. To conduct KEGG pathway analysis, the adjusted p-value < 0.05 was used as the threshold. To conduct Panther Classification System analysis, the highest enriched pathways were selected ($\sim$100 fold enrichment).

[2] https://gitlab.com/npar/dbasis

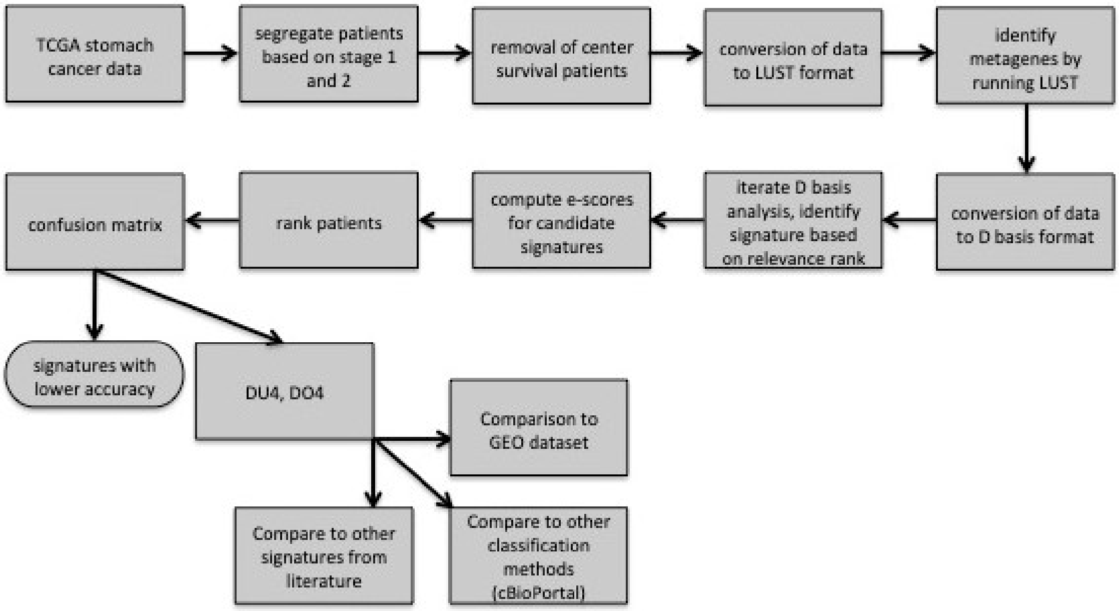

TCGA stomach cancer data
segregate patients based on stage 1 and 2
removal of center survival patients
conversion of data to LUST format
identify metagenes by running LUST
conversion of data to D basis format
iterate D basis analysis, identify signature based on relevance rank
compute e-scores for candidate signatures
rank patients
confusion matrix
signatures with lower accuracy
DU4, DO4
Comparison to GEO dataset
Compare to other signatures from literature
Compare to other classification methods (cBioPortal)

Figure 1: Flow chart of our method of combining algorithms to produce and test prognostic signatures.

3 Results

3.1 The signatures DU4 and DO4

The LUST github site gives the metagenes associated with each of the 33 types of cancer in the TCGA database. Some metagenes occur with many different cancers, while others are specific to only one or a few. LUST identifies 5 metagenes with stomach cancer. The most prominent is designated as the R metagene, which is unique to stomach cancer. Second is the A metagene, which is associated with immune response, and is found with all cancer types. Both split into 3 parts: R1, R2, and R3, and A1, A2, and A3, respectively. The genes in each submetagene are given in Appendix III. The remaining 3 occur in multiple cancers, but seem to have little predictive value for stomach cancer. The R3 submetagene is also found with esophageal cancer.

To extract from the metagene a smaller set of genes (signature) to predict survival, we used the D-basis algorithm, adding rows corresponding to clinical information to the matrix of discretized gene expression.

The D-basis algorithm, starting with the metagenes found by LUST, supplies two lists of genes that are candidates for predictive signatures. The first list consists of genes that are under-expressed on the long-survival/low-risk group. The top 4 genes on this list, all from metagene R, form an effective signature, which we label as DU4: FLRT2, KCNB1, MYOC, TNXB.

The second list of candidates consists of genes that are over-expressed on the short-survival/high-risk group. These genes are from metagenes R and A2. Those from metagene A2 give a reasonably good signature, but the ones from metagene R give more accurate predictions. Thus we get the signature DO4 consisting of the top 4 genes in the second list: ASB5, SFRP1, SMYD1, TACR2. We will formulate tests to predict survival based on these signatures.

3.2 Evaluating signatures: Kaplan-Meier curves and accuracy tests

Suppose we are given a signature consisting of k genes. The first task is to convert the expression data into a score for each patient. Consider the $k \times N$ matrix $\mathbf{M}$ of expression restricted to those k genes for N patients. (Expression has been preprocessed as usual, including log transform and quantile normalization.) The matrix $\mathbf{M}$ has a singular value decomposition $\mathbf{M} = \sum_{r=1}^{m} \sigma_r \mathbf{u}_r \mathbf{v}_r^T$ with $|\sigma_1| \geq |\sigma_2| \geq \cdots \geq |\sigma_m|$ and $\mathbf{u}_r$, $\mathbf{v}_r$ orthonormal sets of column vectors of length k, N respectively (Axler 2015; Dym 2014).[3] Note this decomposition is not very sensitive to small perturbations of the data. The matrix $\sigma_1 \mathbf{u}_1 \mathbf{v}_1^T$ is the rank-1 matrix $\mathbf{N}$ that minimizes the Euclidean (or Frobenius) distance $||\mathbf{M}–\mathbf{N}||$ from the original expression matrix $\mathbf{M}$. That means $\sigma_1 \mathbf{u}_1 \mathbf{v}_1^T$ is the best rank-1 approximation of $\mathbf{M}$, and it is very stable if the first singular value σ_1 is such that $|\sigma_1|$ is much greater than $|\sigma_2|$. For expression matrices of genes in a signature derived from LUST (or similar clustering algorithms), it is usually the case that $|\sigma_1|$ is much greater than $|\sigma_2|$, so that the

[3] While *singular value decomposition* is the standard mathematical terminology, in applied mathematics and statistics it is usually called the *principal component analysis.*

first component is by far the most important and stable. In particular, $\mathbf{u}_1$ does not change much if additional columns (patients) are added to, or removed from, the data.

The vector of *e-scores* is $\mathbf{s} = \mathbf{u}_1^T\mathbf{M}$. Thus the *e-score* for patient j is given by $s_j = \mathbf{u}_1^T\mathbf{m}_j$ where $\mathbf{m}_j$ is the j-th column of $\mathbf{M}$ (the expression levels for patient j). A related method is used in Okimoto et al. (2016).

It is useful to normalize the e-scores in some way; our normalization has -2 as the minimum value and 0 as the median, so that scores typically range from -2 to about 4.

Methods to evaluate the effectiveness of a signature include

- Kaplan-Meier survival curves,
- accuracy (percentage of true positive and true negative predictions),
- the F_1-score, the harmonic mean of precision and recall (sensitivity).

These criteria require some explanation.

To treat the e-score as a predictor of risk, we set a threshold θ on the e-score to divide the population into low-risk and high-risk groups. If a patient's score $s_j < \theta$, then patient j has low risk, while if $s_j \geq \theta$, then patient j is high-risk. (Of course, it could be the other way around, but for both our signatures low expression corresponded to low risk.) The threshold is usually chosen to maximize some measure. For this study, we chose θ to maximize the accuracy, as described below. We also tried choosing θ to maximize the mutual information (from information theory (Cover and Thomas 2006); see Appendix I) between the e-score and survival, but this gave the same thresholds as accuracy.

To apply the Kaplan-Meier method, survival analysis is done independently on the low-risk and high-risk groups, and the p-value for the log-rank test and/or Cox regression measures the probability that these two populations are the same. If the p-value is very small, then the score based on the signature has divided the patients into significantly different risk groups with respect to survival.

The *accuracy* of a predictor is based on the confusion matrix, as shown in Table 1.

	low score	high score
long survival	true neg	false pos
short survival	false neg	true pos

Table 1: Confusion matrix when the e-score is used as a predictor of risk. Thresholds must be set for both scores and survival.

The *accuracy* is

$$\mathrm{acc} = \frac{TP + TN}{N}$$

where TP is the number of true positives, TN is the number of true negatives, and N is the number of samples.

Accuracy is one of several measures of effectiveness that can be computed from the confusion matrix. While accuracy was used to determine the threshold, we also report the F_1-score for each signature, given by

$$F_1 = \frac{TP}{TP + \frac{1}{2}(FP + FN)}$$

Other alternative measures include weighted accuracy and the F_β score for a parameter β. The choice of the weights or β depends on medical considerations, so for this note we use only accuracy and the F_1-score. The article Hughes-Oliver (2019) discusses the various notions of accuracy, and compares accuracy to ROC curves.

We will visualize the confusion matrix with plots of survival against e-scores, as done in figures below. True positives will be in the lower right quadrant, false positives in the upper right quadrant, etc., mimicking Table 1.

3.3 Results: evaluating the signatures DU4 and DO4

The Kaplan-Meier survival curves for DU4 are given in Figure 2, along with the p-values for the logrank test and Cox regression.

To use accuracy, we must set a threshold θ for e-scores. Maximizing either the ac-

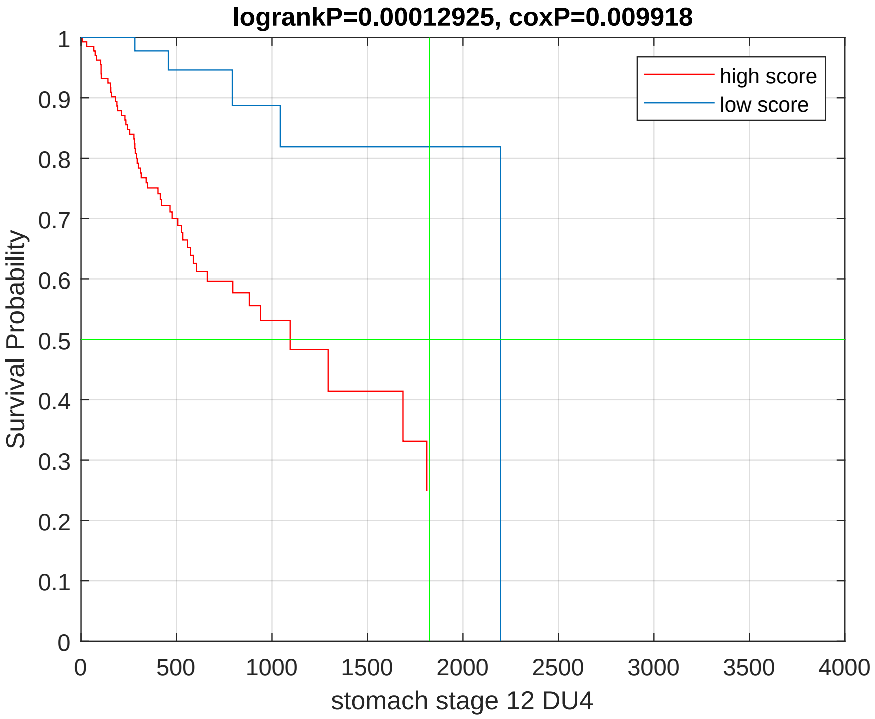


Figure 2: Kaplan-Meier survival curves for the 4-gene signature DU4. Survival times are in days. The top curve represents patients with low e-scores.

curacy or the mutual information between the e-score and survival gives the same result: $\theta = -0.9$. With this value, we get the values in Table 2, for an accuracy of $\frac{58}{76} = 76\%$. The F_1-score for DU4 is 0.83; note the sensitivity is an amazing $\frac{45}{47} = 96\%$.

	low score	high score
long survival	13	16
short survival	2	45

Table 2: Table for the signature DU4.

The confusion matrix for DU4 is illustrated in Figure 3. The patients are along the horizontal axis, arranged in order of increasing e-scores. Their e-scores are the ascending curve, with labels on the left vertical axis. There is a vertical line at $\theta = -0.9$. Survival times are on the right vertical axis, with a horizontal line at the mean expected survival time of 842 days. The + signs indicate the time of a patient's death, while circles represent uncensored patients that are still alive.

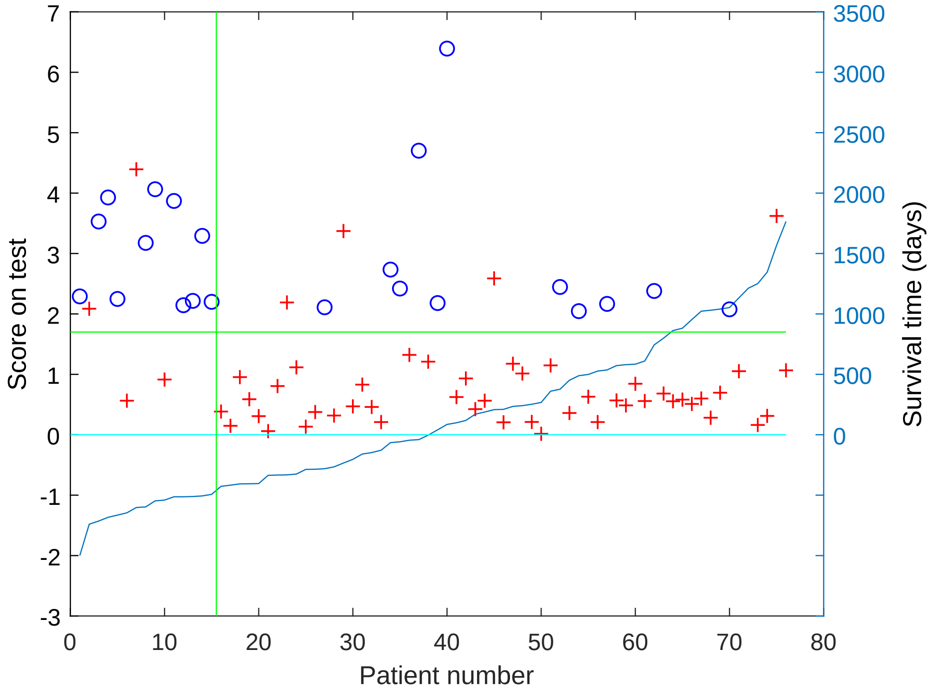


Figure 3: Predictions vs. survival for the DU4 signature.

Now we consider the same analyses for the signature DO4. The Kaplan-Meier survival curves and p-values based on the signature DO4 are given in Figure 4.

For the confusion matrix, using either method (maximizing accuracy or mutual information), we get a threshold of $\theta = -0.9$. With this value, the corresponding table is found in Table 3, for an accuracy of $\frac{56}{76} = 74\%$. The corresponding F_1-score for DO4 is 0.80. Finally, the results in Table 3 are illustrated in Figure 5.

	low score	high score
long survival	16	13
short survival	7	40

Table 3: Table for the signature DO4.

Thus the two signatures are comparable in terms of general effectiveness, with DU4 being more sensitive than DO4. Both signatures were tested for robustness by dividing

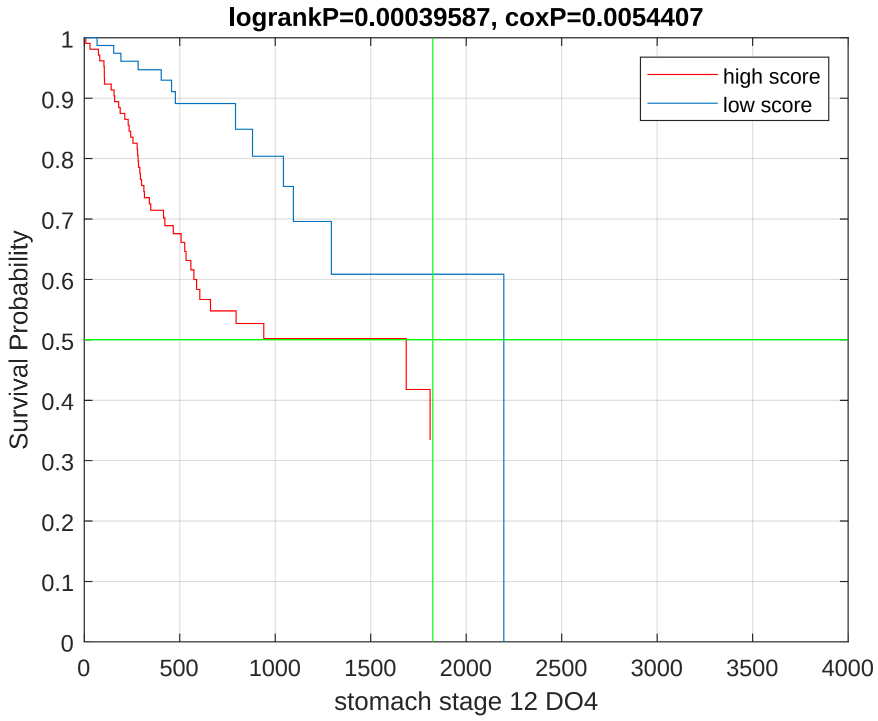


Figure 4: Kaplan-Meier survival curves for the 4-gene signature DO4. Survival times are in days.

the patient pool into two equal parts, and using one half to generate predictions for the other half. As noted earlier, the test vector $\mathbf{u}_1$ for the e-score should not change much when we use a large subset of the patient pool. For each half, the correlation (dot product) between $\mathbf{u}_1$ and the new first singular vector $\mathbf{u}_1'$ was > 0.999. Consequently, patients did not change risk status when only half the samples were used to generate the prediction, and there were no significant differences in the results: each half successfully predicted survival in the other half.

3.4 Validation

While our emphasis is on the method of obtaining signatures, we need to show that combining algorithms yields accurate prognostic signatures. The paper Xie et al. 2020 discusses different methods of finding predictive signatures, and analyzes 39 prognostic signatures for gastric cancer from the literature. We ran a test to compare those 39 signatures,

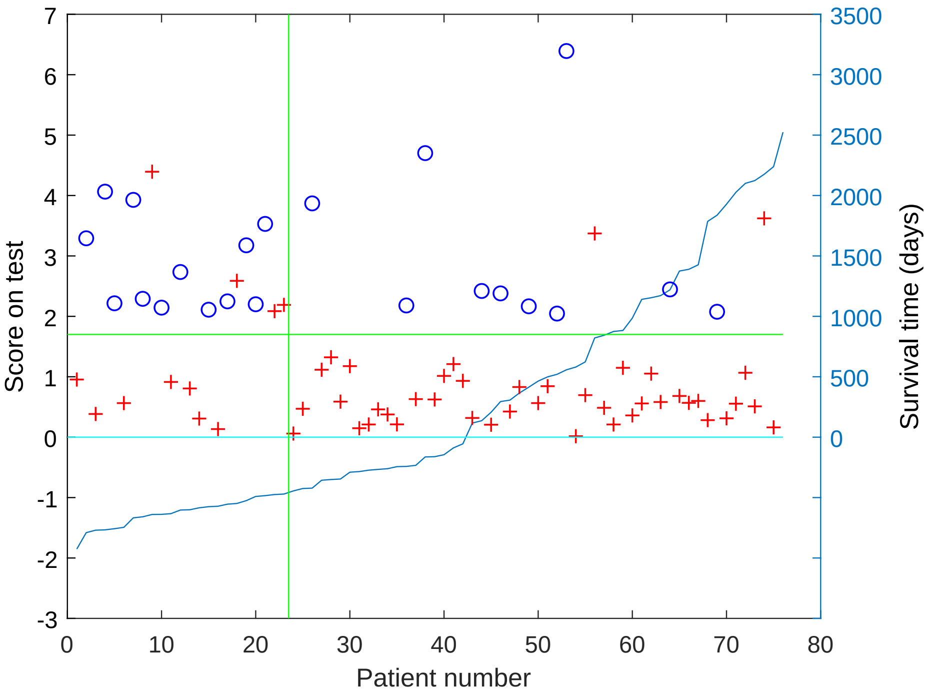
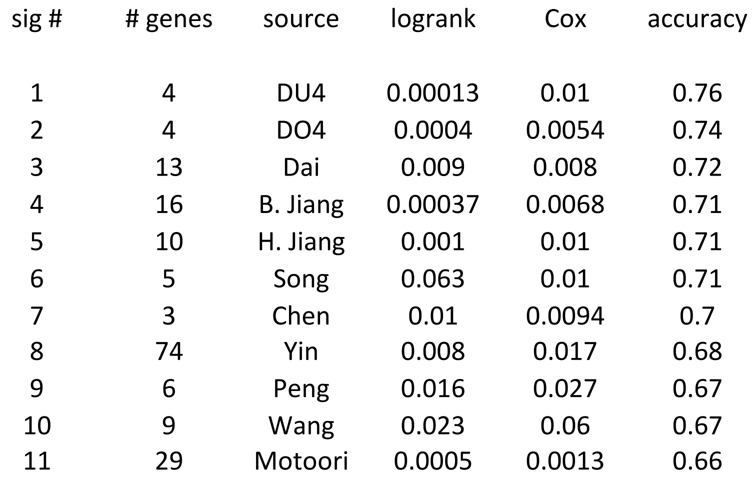


Figure 5: Predictions vs. survival for the DO4 signature.

plus one from Meng et al. 2020, with DU4 and DO4. We tested each signature on the TCGA stage 1 and 2 stomach cancer cohort, exactly as described above. For each signature, we found the threshold e-score that produced the most accurate prediction, and also recorded the logrank and Cox regression p-values for that classification. The results for DU4, DO4, and the top 9 signatures from the literature are shown in Table 4. The signatures DU4 and DO4 were clearly the most accurate, with a few close contenders. The remaining 31 signatures, not shown, had less accuracy.

It is interesting to note that the 3rd place signature, from Dai et al. 2019, contains 2 genes from metagene R, 7 genes whose expression is strongly correlated with DU4, and 4 more genes. The 5th place signature, from H. Jiang et al. 2020, has 1 gene from R, 6 genes whose expression is strongly correlated with DU4, and 3 more genes. The 8th place gene, from Yin et al. 2013, contains 20 genes from R (mostly R2), plus 53 more. The remaining top signatures, however, did not intersect metagene R.

(In the preceding discussion, we say that a gene is strongly correlated with DU4 if its



sig #	# genes	source	logrank	Cox	accuracy
1	4	DU4	0.00013	0.01	0.76
2	4	DO4	0.0004	0.0054	0.74
3	13	Dai	0.009	0.008	0.72
4	16	B. Jiang	0.00037	0.0068	0.71
5	10	H. Jiang	0.001	0.01	0.71
6	5	Song	0.063	0.01	0.71
7	3	Chen	0.01	0.0094	0.7
8	74	Yin	0.008	0.017	0.68
9	6	Peng	0.016	0.027	0.67
10	9	Wang	0.023	0.06	0.67
11	29	Motoori	0.0005	0.0013	0.66

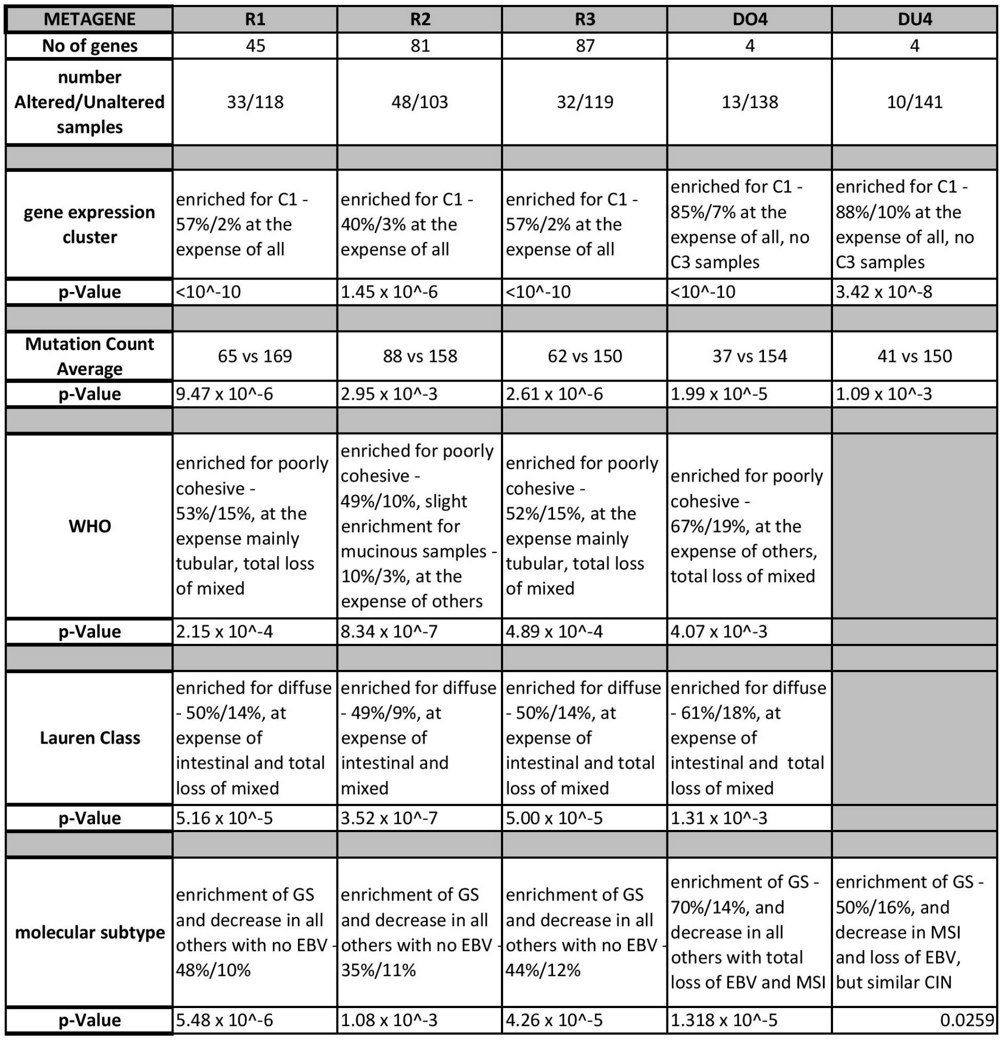
Table 4: The most accurate signatures for predicting survival in stomach cancer: DU4, DO4, and the top 9 from the literature.

differential expression between the low-risk and high-risk groups, as determined by DU4, has a p-value less than 10^{-8}.)

We also tested the signatures on the GEO (Gene Expression Omnibus) stomach cancer dataset GSE84433 (Yoon et al. 2020); see Appendix II.

3.5 Comparison in WHO and Lauren

To determine how well the signatures segregated the tumors analyzed along with other clinical parameters, we used TCGA clinical and expression data easily queried on cBioPortal. Although we used the TCGA legacy tumor set for our analyses, we wished to compare our signatures to the results from the 2014 TCGA stomach cancer manuscript (TCGA Research Network 2014). To do so, from our set of 203 samples analysed, we removed 52 samples that the TCGA had removed from the legacy set during the 2014 publication. We then compared the remaining 151 tumors, for gene expression of the genes in our metagenes or signatures using a z-score of ± 2.0 different from diploid samples. Using the gene expression we were able to segregate the tumor sample set, and the number of tumors for this segregation is shown in Table 5 as a ratio of affected over unaffected. For example, tumors with high expression of genes in the signature DO4 are affected and tumors with low



METAGENE	R1	R2	R3	DO4	DU4
No of genes	45	81	87	4	4
number Altered/Unaltered samples	33/118	48/103	32/119	13/138	10/141
gene expression cluster	enriched for C1 - 57%/2% at the expense of all	enriched for C1 - 40%/3% at the expense of all	enriched for C1 - 57%/2% at the expense of all	enriched for C1 - 85%/7% at the expense of all, no C3 samples	enriched for C1 - 88%/10% at the expense of all, no C3 samples
p-Value	<10^-10	1.45 x 10^-6	<10^-10	<10^-10	3.42 x 10^-8
Mutation Count Average	65 vs 169	88 vs 158	62 vs 150	37 vs 154	41 vs 150
p-Value	9.47 x 10^-6	2.95 x 10^-3	2.61 x 10^-6	1.99 x 10^-5	1.09 x 10^-3
WHO	enriched for poorly cohesive - 53%/15%, at the expense mainly tubular, total loss of mixed	enriched for poorly cohesive - 49%/10%, slight enrichment for mucinous samples - 10%/3%, at the expense of others	enriched for poorly cohesive - 52%/15%, at the expense mainly tubular, total loss of mixed	enriched for poorly cohesive - 67%/19%, at the expense of others, total loss of mixed	
p-Value	2.15 x 10^-4	8.34 x 10^-7	4.89 x 10^-4	4.07 x 10^-3	
Lauren Class	enriched for diffuse - 50%/14%, at expense of intestinal and total loss of mixed	enriched for diffuse - 49%/9%, at expense of intestinal and mixed	enriched for diffuse - 50%/14%, at expense of intestinal and total loss of mixed	enriched for diffuse - 61%/18%, at expense of intestinal and total loss of mixed	
p-Value	5.16 x 10^-5	3.52 x 10^-7	5.00 x 10^-5	1.31 x 10^-3	
molecular subtype	enrichment of GS and decrease in all others with no EBV - 48%/10%	enrichment of GS and decrease in all others with no EBV - 35%/11%	enrichment of GS and decrease in all others with no EBV - 44%/12%	enrichment of GS - 70%/14%, and decrease in all others with total loss of EBV and MSI	enrichment of GS - 50%/16%, and decrease in MSI and loss of EBV, but similar CIN
p-Value	5.48 x 10^-6	1.08 x 10^-3	4.26 x 10^-5	1.318 x 10^-5	0.0259

Table 5: shows the comparison of the metagenes R1, R2, and R3 and the signatures DO4 and DU4 to analysis of the TCGA stomach cancer dataset by cBioPortal. The first column identifies the entries in the 5 columns. The five columns to the right correspond to the metagenes R1, R2, R3 and the signatures DO4, DU4. The information underneath the row identifying the metagenes and signatures have a corresponding title in the first column. If there is a p-value given it corresponds to the row above it. If there is a row grayed out it is because the corresponding p-value was above 0.05. The row altered/unaltered is the number of samples out of the total of 151 that the metagene or signature identify. As an example the first column for R1 which encompasses 45 genes, out of the 151 tumor samples, 33 had altered expression of R1 and 118 did not. All entries below in that column refer to this segregation of tumor samples. As an example for gene expression clusters where there are 4 clusters identified in the TCGA dataset we found that for the 33 tumors, 57% were from the C1 cluster and for the 118 tumors only 2% were from the C1 cluster. The rest follows this nomenclature. The mutation count average is the average number of mutations found for the altered vs the unaltered samples. The Lauren and WHO classifiers are histologic classifiers for stomach cancer. Lauren uses two general categories and WHO break down into 4 general categories, each with the addition of a mixed category. The molecular subtype is a measure of the genome of the tumor samples (if the sample is EBV positive, has microsatellite instability, has many large chromosomal mutations, or has very few mutations).The p-value for mutation count was calculated with Wilcoxon test and all others with Chi squared test.

expression of genes in DO4 are unaffected. The signatures segregated the tumors along several important clinical parameters: gene expression clusters (enriched for cBioPortal cluster C1), mutation count (low mutation load), WHO classification (enriched for poorly cohesive), molecular subtype (enriched for genetically silent), and Lauren class (enriched for diffuse). All these factors are associated with a poor prognosis. We also analyzed the signatures with the entire tumor set, with similar results; see Table 7 in Appendix III.

A slightly different analysis shows these distinctions in another light. We ask: *what clinical properties are differentiated by the low-risk and high-risk groups?* In this case we used all 203 stage 1 and 2 patients from TCGA, with no censoring, and the signature DU4 with a threshold score $\theta = -0.9$ to divide low/high risk.

For the Lauren classification, 37 patients had *diffuse* tumors. Of these, 2 were in the low-risk group (based on the signatures) and 35 were in the high-risk group. For the WHO classification, 31 tumors were deemed *poorly cohesive*, all of which were assigned to the high-risk group. A similar distinction occurred with the molecular subtype classification: 30 tumors were deemed *genetically silent*, with 2 in the low-risk group and 28 in the high-risk group. Other classes had some differences between low and high risk, but nothing as marked as the ones above.

The results are identical if we divide low/high risk using the signature DO4 with a threshold score $\theta = -1.0$.

3.6 Pathway analysis

The results of KEGG pathway analysis using DAVID on metagenes R1, R2, and R3 are given in Appendix III, Tables 8, 9, 10. For metagene R1 many cancer-related pathways were identified, including PI3K-AKT, RAS, G-protein pathways, and PTK2 signaling. Analysis of metagene R2 identified pathways involved in collagen/extra-cellular matrix homeostasis and cellular interaction with the ECM. Pathway analysis of metagene R3 identified many pathways that are specific to the stomach, such as smooth muscle contraction, but also cancer-related pathways that were found with R1.

Using Panther Classification Systems for Reactome, pathway analysis of the metagenes found similar results as the KEGG analysis. Signal transduction pathways were identified for R1 (such as over 100-fold enrichment were FRFR, SHH and PI3K-AKT pathways). The most enriched pathways for R2 involved extra-cellular matrix homeostasis (including collagen degradation, collagen chain trimerization, dermatin sulfate biosynthesis, and ECM proteoglycan synthesis), and for R3 a mix of smooth muscle contraction and signal transduction pathways.

3.7 Recurrence

Finding signatures that predict recurrence turns out to be more difficult. Here we ran the tests on all stages of stomach cancer, censoring all patients who have survived below the empirical mean of 895 days without recurrence. That gives a pool of 139 patients, 75 of whom had recurrence before 895 days. Thus "failure" in this scenario is recurrence, which is not the same as "disease-free survival", since deaths before 895 days due to the original tumor are censored. (We are interested in when a patient who is at one point disease-free has a recurrence.)

Again the D-basis provided a list of candidates from the R and A metagenes, and the list of top candidates from metagene R provided the most accurate signature for predicting recurrence: BNC2, CCDC8, DPYSL3, MOXD1, MXRA8, PRELP, SCARF2, TAGLN, ZNF423. This signature is designated as REC9. The table for recurrence prediction is Table 6, with an accuracy of $\frac{92}{139} = 66\%$. The F_1-score for REC9 is 0.70. The Kaplan-Meier curves and prediction figures for recurrence are given in Figures 6 and 7.

	low score	high score
long non-recurrence	36	28
short recurrence	19	56

Table 6: Table for the 9-gene recurrence signature REC9.

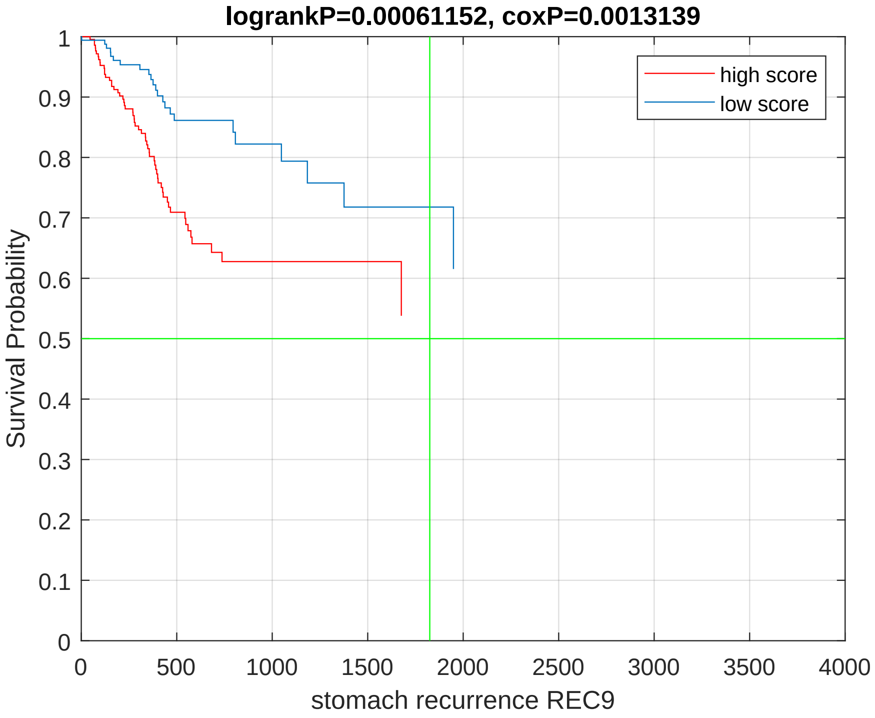


Figure 6: Kaplan-Meier curves for non-recurrence with the 9-gene signature.

4 Discussion

We have combined two algorithms to analyze tumor gene expression along with clinical data to identify gene expression that associated with survival. Using the first algorithm (LUST) we identified metagenes, which were refined by D-basis to signatures DU4 and DO4 that accurately predicted survival.

The metagenes identified as R1, R2, and R3 associate with cBioPortal gene expression cluster C1. The metagenes identify tumor sets that are enriched for mutation of the CDH1 gene. The TCGA analysis found that tumors that were in the genomically silent grouping, were enriched for CDH1 and RHOA gene mutation (TCGA Research Network 2014).

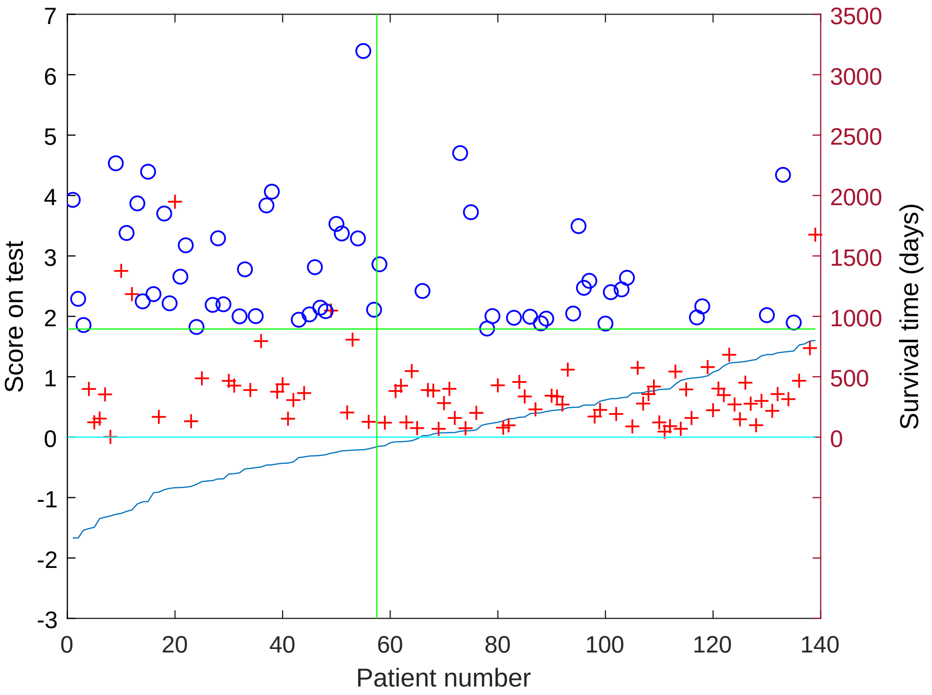


Figure 7: Predictions vs. non-recurrence for the 9-gene signature REC9. The + signs indicate the time of recurrence, either distant or local, while circles represent patients who survive at least 895 days without recurrence.

The metagenes identify tumor sets that are also in the genomically silent range and have low mutation loads. Interestingly, the tumor sets identified by the metagenes are enriched for mutation of the CDH1 gene, but not the RHOA gene. The gene signatures refined from the metagenes recapitulate the above findings from the metagenes. Furthermore, the signatures DU4 and DO4 identify tumors that are predominantly in the WHO classification of poorly cohesive and the Lauren class of diffuse samples. The WHO classification of poorly cohesive identifies tumors (including those tumors with signet cells present) that have a poor prognosis (Jouini et al. 2020). Tumors of the Lauren classification that are classified as diffuse are tumors that are less well differentiated, and thus also have a poorer prognosis. The presence of a CDH1 mutation overlaps with diffuse tumors identified by Lauren classification (TCGA Research Network 2014; X. Li et al. 2016; Nemtsova et al. 2020). While mutation of CDH1 can contribute to a poor outcome, it is only one of several such factors. As a prognostic marker for survival, the presence of a CDH1 mutation has an accuracy of only 53%. That is, with the same TCGA sample set, if patients with a CDH1 mutation are declared *high-risk* and those without *low-risk*, only 40 of the 76 patients are labelled correctly.

Pathway analysis for metagenes R1, R2, and R3 found pathways that could be segregated into roughly two categories, pathways of normal stomach genes and pathways that were more cancer related.

The metagenes were used to identify the signatures DO4 and DU4. Many of the genes in DO4 and DU4 have been indicated as factors in cancer by other studies: FLRT2 in Dai et al. (2019), TNXB in Yan et al. (2019), SFRP1 in S. Li et al. (2018), and SMYD1 in J. Song et al. (2019). More details are in Appendix II.

A recent study (Zhou et al. 2020) found 9 genes whose over-expression indicated a poor prognosis in stomach cancer. These genes were found using the ESTIMATE algorithm on TCGA data, and validated on GEO dataset GSE84433 (Yoon et al. 2020). Of the 9 genes, 7 were from R3 (CNN1, FLNC, HAND2, MYL9, PLN, SPARCL1, SYNC), 1 was from R2 (SFRP2), and only 1 not associated with the R metagene (CARTPT). That study also identified 31 genes more generally associated with stomach cancer survival: 1 was from R1, 2 from R2, 14 from R3, 4 from the R metagene additional genes, and 10 were not part of the R metagene.

Thus our combination of algorithms has found the large set R of related genes that seem to be a factor in the progression of stomach cancer, and two small signatures within it that predict survival.

Acknowledgements Hofstra University supported participation and poster presentation of these results at the Gordons Research Conference *Human Genetics and Genomics* in NH, July 2019.

Disclosures No competing financial interests exist.

References

Adaricheva, K. and J. Nation (2017). Discovery of the D-basis in binary tables based on hypergraph dualization. *Theor. Comp. Sci., Part B* 658, 307–315.

Adaricheva, K., J. Nation, et al. (2015). Measuring the Implications of the D-basis in Analysis of Data in Biomedical Studies. *Proceedings of ICFCA-15, Nerja, Spain*. Springer, 39–57.

Alkhateeb, A. et al. (2019). Transcriptomics Signature from Next-Generation Sequencing Data Reveals New Transcriptomic Biomarkers Related to Prostate Cancer. *Cancer Informatics* 18, 1–12.

American Cancer Soc. (2020). *Cancer Facts and Statistics*. URL: https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2021/cancer-facts-and-figures-2021.pdf.

Axler, S. (2015). Linear Algebra Done Right, 3rd ed. New York: Springer.

Chen, C.N. et al. (2005). Gene expression profile predicts patient survival of gastric cancer after surgical resection. *J. Clin. Oncol.* 23, 7286–7295. URL: https://doi.org/10.1200/JCO.2004.00.2.

Cover, T. and J. Thomas (2006). Elements of Information Theory, 2nd. ed. Hoboken: Wiley.

Dai, J. et al. (2019). Whole genome messenger RNA profiling identifies a novel signature to predict gastric cancer survival. *Clin. Transl. Gastroenterol.* 10(1), 1–9. URL: https://doi.org/10.14309/ctg.0000000000000004.

Dym, H. (2014). Linear Algebra in Action, 2nd ed. Providence: Amer. Math. Soc., Graduate Studies in Mathematics.

Ferlay, J. et al. (2018). *Global Cancer Observatory: Cancer Today. Lyon, France: International Agency for Research on Cancer*. URL: https://gco.iarc.fr/today.

Hughes-Oliver, J.M. (2019). Assessment of prediction algorithms for ranking objects. *Notices Amer. Math. Soc.* Feb. 182–190.

Jiang, B. et al. (2019). An immune related gene signature predicts prognosis of gastric cancer. *Medicine* 98. URL: https://doi.org/10.1097/MD.0000000000016273.

Jiang, H. et al. (2020). A 21-gene support vector machine classifier and a 10-gene risk score system constructed for patients with gastric cancer. *Mol. Med. Rep.* 21, 347–359. URL: https://doi.org/10.3892/mmr.2019.10841.

Jouini, R. et al. (2020). Prognostic significance of poorly cohesive gastric carcinoma in Tunisian patients. *Heliyon* 6(3), e03460.

Li, S. et al. (2018). Inhibition of DNMT suppresses the stemness of colorectal cancer cells through down-regulating Wnt signaling pathway. *Cell Signal* 47, 79–87.

Li, X. et al. (2016). Distinct Subtypes of Gastric Cancer Defined by Molecular Characterization Include Novel Mutational Signatures with Prognostic Capability. *Cancer Research* 7, 1724–32.

McLoughlin, J.M. (2004). Adenocarcinoma of the stomach: a review. *Proc. Baylor Univ. Med. Cent.* 17(4), 391–399.

Meng, C. et al. (2020). Discovery of prognostic signature genes for overall survival in gastric cancer. *Comp. Math. Methods Med.* URL: https://doi.org/10.1155/2020/5479279.

Motoori, M. et al. (2005). Prediction of recurrence in advanced gastric cancer patients after curative resection by gene expression profiling. *Int. J. Cancer* 114, 963–968. URL: https://doi.org/10.1002/ijc.20808.

Nation, J. et al. (2017). A comparative analysis of mRNA expression for 33 different cancers. URL: https://github.com/tristanh314/lust-cancer-2019.

Nemtsova, M.V. et al. (2020). Clinical relevance of somatic mutations in main driver genes detected in gastric cancer patients by next-generation DNA sequencing. *Science Reports* 1, 504.

Okimoto, G. et al. (2016). The joint analysis of multiple, high-dimensional data types using sparse matrix factorizations of rank-1 with applications to ovarian and liver cancer. *BioData Mining* 9:24. DOI: `10.1186/s13040-016-0103-7`. URL: `http://www.biodatamining.org/content/9/1/24`.

Peng, P.L. et al. (2018). Identification of a novel gene pairs signature in the prognosis of gastric cancer. *Cancer Med.* 7, 344–350. URL: `https://doi.org/10.1002/cam4.1303`.

Prashanth, R. and A. Barsouk (2018). Epidemiology of gastric cancer: global trends, risk factors and prevention. *Prz. Gastroenterol.* 14, 26–38.

Segal, O. et al. (2018). The D-basis algorithm for association rules of high confidence. *IT in Industry* 6 (N3).

Shannon, C. (1948). A mathematical theory of communication. *Bell System Tech. Jour.* 27, 379–423 & 623–656.

Song, J. et al. (2019). Expression patterns and the prognostic value of the SMYD family members in human breast carcinoma using integrative bioinformatics analysis. *Oncol. Lett.* 17(4), 3851–3861.

Song, L.X. et al. (2019). A 5-gene prognostic combination for predicting survival of patients with gastric cancer. *Med. Sci. Monitor* 25, 6313–6320. URL: `https://doi.org/10.12659/MSM.914815`.

Sørlie, T. et al. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl. Acad. Sci. USA* 98(19), 10869–10874.

TCGA Research Network (2014). Comprehensive molecular characterization of gastric adenocarcinoma. *Nature* 513, 202–209.

Wang, Z. et al. (2017). Identification and validation of a prognostic 9-genes expression signature for gastric cancer. *Oncotarget* 8, 73826–73836. URL: https://doi.org/10.18632/oncotarget.17764.

Xie, L. et al. (2020). Systematic review of prognostic gene signature in gastric cancer patients. *Front. Bioeng. Biotechnol.* URL: https://doi.org/10.3389/fbioe.2020.00805.

Yan, S.P. et al. (2019). LncRNA LINC01305 silencing inhibits cell epithelial-mesenchymal transition in cervical cancer by inhibiting TNXB-mediated PI3K/Akt signalling pathway. *J. Cell Mol. Med.* 23(4), 2656–2666.

Yin, Y. et al. (2013). Converting a microarray signature into a diagnostic test: a trial of custom 74 gene array for clarification and prediction the prognosis of gastric cancer. *PLoS ONE*. URL: https://doi.org/10.1371/journal.pone.0081561.

Yoon, S.J. et al. (2020). Deconvolution of diffuse gastric cancer and the suppression of CD34 on the BALB/c nude mice model. *BMC Cancer* 20, 314.

Zhou, L. et al. (2020). Exploring TCGA database for identification of potential prognostic genes in stomach adenocarcinoma. *Cancer Cell Int.* 20, 264.

5 Appendix I: Mathematical description of the algorithms

5.1 Data conversion

The original expression of 20,531 mRNA data from tumor tissue of 415 stomach cancer patients from TCGA was log transformed and quantile normalized before discretizing. To obtain the entry table for LUST, each entry of original table was converted to 0, +1 or -1. These entries indicate the over-expressed (+1), under-expressed (-1), or normally expressed (0) genes for each patient. The conversion is done using a density parameter D, i.e., the proportion of non-zero entries. In the current study, the value of $D = 0.5$ was used so that roughly a quarter of values were converted to +1 and a quarter to -1. Other tests were done for values $0.4 \leq D \leq 0.6$, which did not change the results significantly.

5.2 The LUST algorithm

The LUST (Lattice Up-Stream Targeting) algorithm is a discrete method that uses a variation of association rules to find clusters of genes with similar expression patterns, and within those clusters selects groups of genes that maximize some given objective function. The objective function might represent for example the degree of interaction between the genes (as reflected in the expression data), or some clinical outcome such as survival.

It should be emphasized that we are comparing gene expression for tumors, not tumor vs. normal. We seek genes whose expression varies significantly in cancer patients, with the intention of using this information to tailor treatment based on genetic signatures.

In its function used for this paper, LUST is applied to the TCGA mRNA expression data, with an objective function that measures the degree of interaction between the genes. This allows us to identify groups of genes that are part of the same biological process, e.g., immunity or cell division or metabolism, that have sufficient variation across the samples. These we will refer to as *metagenes*. Knowing the metagenes that occur in the gene expression for a particular type of cancer allows us to determine the biological factors that affect the progress of the disease, using gene functional annotation.

The LUST program can also run in a second mode, with an objective function that depends on clinical outcomes. By restricting the input to gene expression from a metagene, and maximizing the objective function, LUST produces signatures predictive of survival. This has been done effectively for many different types of cancer Nation et al. 2017. But for stomach cancer, we found that the signatures produced by using LUST for the first step, and the D-basis for the second step, were more accurate than those produced by LUST alone.

The LUST programs are written in MATLAB (MathWorks, Natick, MA, USA). The programs, a manual, and the different metagenes found in 33 different cancer data sets in TCGA are presented in our GitHub site: https://github.com/tristanh314/lust-cancer-2019.

The input of the LUST algorithm consists of:

- Data which can be a combination of gene expression, microRNA expression, methylation, and possibly other variables.
- The values of parameters *density* and *conftol* (short for *confidence tolerance*).

The output of the algorithm is:

- Groups of genes (metagenes), ranked using an objective function.

Not all the groups obtained need be related to the disease and/or clinical outcomes, and some will be more relevant than others, but the results can be analyzed to identify metagenes of interest.

Assume the density D has been fixed. The parameter *conftol* adjusts the sensitivity of the algorithm; usually we take *conftol* $\doteq 0.75$. The optimum value of *conftol* was between 0.7 and 0.8 for every cancer considered.

For a row (gene) X, let X^+ denote the set of columns (samples) that are marked $+1$, and let X^- denote the set of columns that are marked -1. We say that X *regulates* Y, and write $X \to Y$, if the following hold:

1. $\dfrac{|X^+ \cap Y^+|}{|X^+|} \geq$ *conftol*

2. $\dfrac{|X^- \cap Y^-|}{|X^-|} \geq$ *conftol*

In words, X regulates Y if the conditional probability that a patient over-expresses Y, given that the patient over-expresses X, is at least *conftol*, and likewise for under-expression. The premise is that if X *does* strongly regulate Y biologically, then it will be reflected as regulation in the data.

Now we say that gene X is *equivalent* to gene Y, and write $X \approx Y$, if $X \to Y$ and $Y \to X$ both hold. This means that X and Y are acting in concert, and heuristically are part of some common process.

The algorithm begins by calculating:

1. for each gene X, a list of all genes Y such that $X \to Y$,

2. for each gene X, a list of all genes Y such that $X \approx Y$.

Now we form groups of genes as follows. Initially, the groups consist of a gene X and all the genes equivalent to it,

$$F_X = \{Y : Y \approx X\}.$$

These groups may overlap substantially. Overlapping groups are *merged* according to the following scheme: if a larger group contains at least *overlappercent* (a parameter with default value 0.6) of the genes of a smaller group, then the groups are combined.

The LUST algorithm will generate many groups that are candidates for metagenes. These candidates are ranked using an *objective function* that we seek to maximize. We regard a metagene as a directed graph, with edges determined by the relation $X \to Y$. The objective function should be a graph-theoretic measure of the probability of obtaining a set of vertices of that size and density of edges. *The objective function for metagenes does not depend on clinical outcomes*, but only on the expression data.

Suppose we are given a metagene M with n genes. Regarding M as a directed graph, let $|E|$ be the number of arrow relations (edges) $X \to Y$ between genes (vertices) of M. (Here $X \approx Y$ counts as two directed edges.) A complete directed graph on n vertices would have $n(n-1)$ edges. Hence the *edge density* of M is given by

$$\delta(M) = \frac{|E|}{n(n-1)} \, .$$

The objective function should be increasing in both the size n of the metagene and its edge density. A simple objective function with this property is

$$f(M) = n \cdot \frac{|E|}{n(n-1)} = \frac{|E|}{n-1}$$

and this is what we used.

As with density, for any particular study one should try several values of *conftol* and compare the results. Lowering *conftol* increases the sensitivity of the algorithm, and as long as values of *conftol* below 0.5 are avoided, false discoveries are not an issue. A typical real expression data matrix of size $20,000 \times 100$ with *conftol* $= 0.7$ will generate 150,000 to 200,000 arrows. The expected number of arrow relations in a random matrix of the same size and density is 0.01. When there are more columns (samples), the expected number of random arrows is less.

Now the LUST program is run on the expression matrix $\mathbf{E}$ for the type of cancer being considered. The output consists of the following:

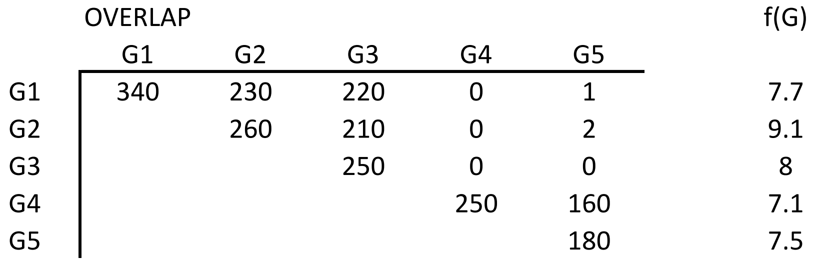
1. a list of the genes in each of the largest groups $G_1, \ldots, G_m$, where the default value of m is 32,

2. a table giving the size of the groups $|G_i|$ and their intersections $|G_i \cap G_j|$,

3. the value $f(G_i)$ of the objective function on each group.

Next we form a *pseudo-equivalence* by setting $G_i \equiv G_j$ if $G_i \cap G_j$ is large. This is not a transitive relation. From each pseudo-equivalence class, choose a representative G_k that maximizes the objective function $f(G)$. These representatives G_k will be the *metagenes* for this cancer. Figure 8 illustrates this process for a collection of 5 groups.

Currently, the metagenes are identified by inspection of the output. We obtain from 4 to 8 metagenes in this way for each type of cancer. These can be ranked according to their values with the objective function $f(G)$.



	OVERLAP					f(G)
	G1	G2	G3	G4	G5	
G1	340	230	220	0	1	7.7
G2		260	210	0	2	9.1
G3			250	0	0	8
G4				250	160	7.1
G5					180	7.5

Figure 8: Schematic output of the LUST algorithm. The overlaps show that group G1 contains 340 genes, G2 contains 260 genes, and their intersection $G1 \cap G2$ contains 230 genes, etc. The overlaps indicate that groups G1, G2 and G3 represent the same biological process or pathway; we would choose G2 as the representative metagene since it has the highest score on the objective function $f(G)$. Groups G4 and G5 are from a different process; for this set of candidate metagenes we would choose G5 as the representative based on its score.

Let us consider how one should interpret metagenes. Lists of the genes contained in the metagenes found by the LUST algorithm are given on the LUST Github site.

A metagene is collection of genes acting in concert. The probability of this happening for a large number of samples without their being part of some common process (or interacting pathways) is low, as evidenced by the false discovery rate. Moreover, these genes are expressed differentially in the samples. Thus the metagene points to some feature or factor that varies in cancer patients. (It will not in general distinguish tumor from normal.) This may reflect aggressiveness of the tumor, proliferation, patient response to the disease, some factor unique to the organ in question (e.g., smooth muscle function in stomach cancer, lipid metabolism in HCC, digestive function or insulin regulation in pancreatic cancer), or possibly some differences not related to the disease.

The significance of the metagene can be measured by the objective function $f(M)$, or how well the signatures obtained from the metagene separate the survival curves (or some other clinical response). These are rather different measures, which are rather loosely correlated. One purpose of signatures is to identify the patients that have the factor in question, e.g., impaired immune response or metabolic dysfunction.

If signatures from the metagene separate the Kaplan-Meier survival curves, then there is *something* there that needs to be understood. *Caveat:* None of our current analysis addresses mutations or methylation or microRNA regulation of gene expression. These factors are relevant and should be included in later studies. The LUST algorithm supports multiple data types.

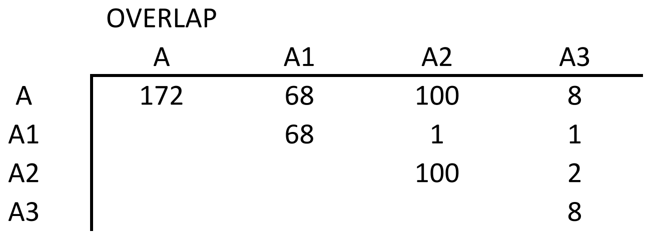
Sometimes a collection of genes that is a single group for some cancers can split into two or more metagenes for others, with the parts being disjoint or nearly so. This is particularly true for metagene A, related to immune regulation. Let us call those parts A1 , A2 and A3. When we look at the groups that are candidates for metagene A, several options occur.

- Some groups are a mixture of A1, A2 and A3.
- Some groups contain A2 and A3 combined, but separate from A1.
- Sometimes all three parts form distinct groups.

Whether we regard this situation as one metagene with three parts, or three related and slightly overlapping metagenes, is a matter of convenience.

In stomach cancer, there is a large combined A metagene, along with smaller split versions of A1, A2, and A3. But for example, in colon, rectum, and some kidney cancerts, A1 is barely present.

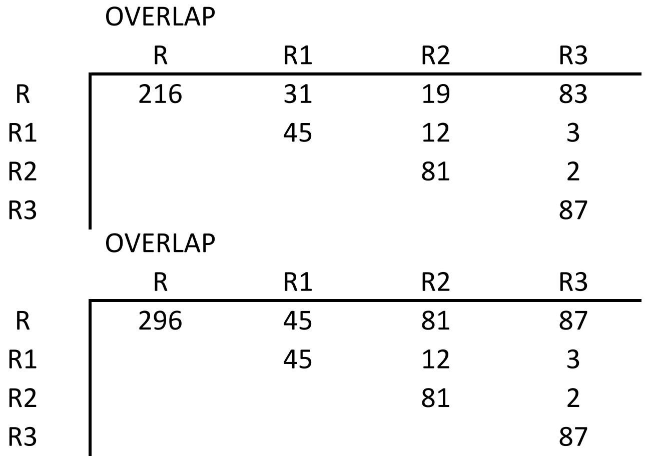
Metagene R for stomach cancer also splits into three parts. In this case, metagene R does not exactly contain the smaller metagenes, but the overlap is too large to consider



OVERLAP

	A	A1	A2	A3
A	172	68	100	8
A1		68	1	1
A2			100	2
A3				8

Figure 9: LUST output for the A metagene. The large metagene A contains 3 smaller metagenes: A1, A2, A3 which barely overlap. This is an idealized version, compiled from the output for many cancers, and these entries would be embedded (not necessarily consecutively) in a much larger table of output.



OVERLAP

	R	R1	R2	R3
R	216	31	19	83
R1		45	12	3
R2			81	2
R3				87

OVERLAP

	R	R1	R2	R3
R	296	45	81	87
R1		45	12	3
R2			81	2
R3				87

Figure 10: LUST output for the R metagene from stomach cancer. The metagene R has a large overlap with 3 smaller metagenes: R1, R2, R3 and some additional genes which are in none of the smaller ones. The first table is as it actually appears in the stomach cancer output; the second is an idealized version with R containing the parts R1, R2, R3. Again these entries would be contained in a much larger table of output.

them distinct. Figure 10 indicates how these metagenes appear in the LUST output. A version of metagene R also occurs with esophageal cancer.

5.3 The D-basis algorithm

D-basis is a new algorithm described in Adaricheva and Nation (2017) that discovers the rules $S \to d$ in a table with entries 0 and 1. Here S is a set of rows (genes) and d is another row (say, a clinical parameter, or another gene). The rule $S \to d$ is found in the table, if all entries of 1 for set S imply an entry of 1 in row d, for each column (patient) of the matrix. In practice, the algorithm computes rules that hold in almost all columns, that is, rules such that the probability of $d = 1$, when the entries in S are all 1, is above a given threshold.

The *support* of a rule $S \to d$ is the number of columns, where all entries in rows $S \cup d$ are 1. The D-basis provides flexibility of several parameters built into the testing, which may affect the ranking of the genes with respect to the parameter of *relevance*. The top genes were identified through testing with variation of the *minimal support*, which refers to percentage of patients validating the rules connecting genes and survival/recurrence. In our testing, we varied minimal support between 5% and 15% of patients in the testing set.

For a fixed row d and any other row x, one can compute the total support of all rules $S \to d$ such that x is in S. This parameter shows the frequency that x appears in implications targeting d. The algorithm can also compute a similar frequency of x, when targeting $\neg d$, i.e., an additional row where all entries in d are switched. The ratio of the two frequencies gives the *relevance* of row x to d. Thus, all rows of the table can be ranked in their relevance to a fixed row d.

The D-basis can be applied to the entry table formed by metagenes found by LUST, and choose d as a marker for the low-risk patients (longer survival), or high-risk patients, or recurrence parameter, among many other options. It was used in ovarian cancer analysis in Adaricheva, Nation, et al. (2015). Additional functionality of the D-basis was developed in Segal et al. (2018).

We tested the algorithm on random tables of size and density equivalent to real data for stomach cancer. In the process of finding the first signature above, the D-basis algorithm produced 8041 implications $S \to d$ of support ≥ 8 with $|S| \leq 4$ from a 460×76 expression matrix. The estimated probability of obtaining at least 8041 implications $S \to d$ of support ≥ 8 and $|S| \leq 4$ in a 460×76 random binary table with the same density as our data is 1.64×10^{-7}. We conclude that the signature represents a real signal in the data, and not a random artifact.

5.4 Mutual information

In general, given discrete random variables X and Y, the *mutual information* between X and Y is

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)}$$

where $p(x,y)$ is the probability of the pair (x,y), and $p(x)$, $p(y)$ are the marginal probabilities. The concept goes back to Shannon (1948); see Cover and Thomas (2006) for a modern treatment. The mutual information measures (in bits) the reduction in the uncertainty of X given that you know Y: in terms of the entropy H and conditional probability,

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$$

In our application, X is the discretized survival (0 for short survival, 1 for long survival) and Y is the discretized e-score (0 if $s_j < \theta$, 1 if $s_j \geq \theta$). The probabilities are determined emperically, by their frequencies. By adjusting θ, we can maximize the mutual information between survival and the e-scores.

6 Appendix II: Additional tables, figures and analysis

.

The pathway analysis reflected in Table 5 was done on patients diagnosed at stage 1 or 2. Table 7 gives the analysis for stages 1–4, which shows no significant differences.

Tables 8, 9, and 10 give the pathway analysis performed on metagenes R1, R2, and R3 using DAVID. The tables show the pathways identified, with other information including the corrected p-value.

As indicated in the text, some of the genes in DU4 and DO4 have been specifically indicated as factors in cancer by other studies. FLRT2 and TNXB were found with decreased gene expression in the long survival group. Decrease in FLRT2 gene expression was found by Dai et al. (2019) to correlate with better survival in gastric cancer. In cervical cancer, TNXB protein is involved in the activation of the PI3K pathway in the transition of EMT (Yan et al. 2019). Other genes, including SFRP1 and SMYD1 were found with higher gene expression in the long survival group. SFRP1 is a soluble inhibitor of the WNT pathway and is frequently found with promoter methylation in cancer cells (S. Li et al. 2018). We propose that overexpression of SFRP1 would act in an autocrine and paracrine manner decreasing cell proliferation.

We tested the signatures DU4 and DO4 on the GEO (Gene Expression Omnibus) dataset GSE84433 (Yoon et al. 2020) consisting of 357 stomach adenocarcinoma patients. This dataset is somewhat different from the TCGA data, including (1) there is no staging in the clinical data, and (2) a different Illumina array was used, with multiple probes for some genes. So we used all the samples, not trying to separate out the early stage cases, and initially used all the probes from the signatures. It turned out that using only some of the probes from a signature gave better predictions.

More significantly, the mean expected survival time for the GEO cohort, based on the ecdf, is 3450 days, compared to 842 days for the TCGA cohort. Whatever the reasons for this disparity, the results should be interpreted with some caution. This discrepancy also makes it impractical to use accuracy as a measure of a signature's prognostic effectiveness, as it is not clear what constitutes a long survival time.

On the GSE84433 data, the DU4 signature was not very predictive of survival, with a p-score of just $p = 0.064$ on the log-rank test. However, if we use only 2 probes from the signature, 1 for TNXB and 1 for FLRT2, we get $p = 0.0049$. The entire DO4 signature performed noticeably better, with $p = 0.00093$. The Kaplan-Meier curves for DO4 on the GEO dataset are shown in Figure 11. Using only 3 probes from the DO4 signature, 2 for ASB5 and 1 for SMYD1, this improves to $p = 0.00052$.

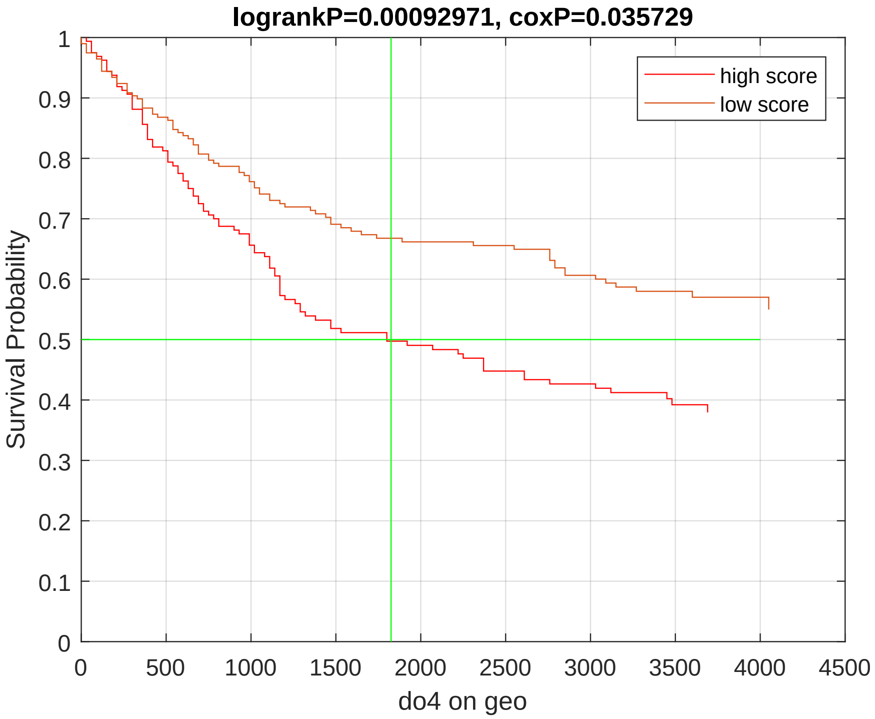


Figure 11: Kaplan-Meier survival curves for the 4-gene signature DO4 on the GEO dataset GSE84433.

7 Appendix III: The genes in metagenes A and R

Recall that the LUST algorithm identifies *metagenes* as sets of genes that tend to over-express on the same patients, and to under-express on the same patients. The assumption is that this signal in the expression data means that these genes are part of some common biological process, which may or may not be related to the disease. Our concern, of course, is to find metagenes that are related to disease progression. For stomach cancer, this is the A and R metagenes, which are listed below.

We should mention that the metagenes found by LUST for esophageal cancer were somewhat different (Nation et al. 2017), except for a small version of R3, and our signatures do not predict survival for esophageal cancer very well. This distinction is similar to that found for hepatocellular carcinoma and cholangiocarcinoma for liver cancer (Nation et al. 2017).

The algorithm allows one metagene to be contained in another. This can be interpreted to mean that the submetagene represents a more tightly connected cluster within the larger one. Of particular interest is the case when a large metagene (with a few hundred genes) splits into parts that are disjoint or nearly disjoint. Both the A and R metagenes split into 3 parts.

Metagene A, consisting of genes related to immune response, splits into A1, A2, and A3 in almost every type of cancer, including stomach cancer.

A1: ACAP1, ARHGAP9, BTLA, CCL5, CCR5, CD247, CD27, CD2, CD3D, CD3E, CD3G, CD48, CD52, CD5, CD6, CD74, CD8A, CD8B, CD96, CRTAM, CTSW, CXCR3, CXCR6, FASLG, GPR171, GRAP2, GZMA, GZMH, GZMK, ICOS, IL12RB1, IL21R, IL2RB, IL2RG, ITGAL, ITK, KIAA0748, KLRK1, LCK, LTA, MAP4K1, NKG7, P2RY10, PDCD1, PRF1, PTPN7, PTPRCAP, PYHIN1, RASAL3, SASH3, SCML4, SEPT1, SH2D1A, SIRPG, SIT1, SLA2, SLAMF1, SLAMF6, SPN, TBC1D10C, TBX21, THEMIS, TIGIT, TRAF3IP3, TRAT1, UBASH3A, ZAP70, ZNF831.

A2: AIF1, AOAH, ARHGAP30, ARHGAP9, BIN2, BTK, C17orf87, C1QA, C1QB, C1QC, C1orf162, C3AR1, CD14, CD163, CD300A, CD300C, CD300LF, CD33, CD37, CD4, CD53, CD74, CD86, CLEC4A, CORO1A, CSF1R, CSF2RB, CYBB, CYTH4, DOCK2, DOCK8, DOK2, EVI2B, FCER1G, FCGR1A, FCGR1B, FCGR1C, FCGR3A, FERMT3, FGD2, FPR3, FYB, GIMAP1, GIMAP4, GPR65, GPSM3, HAVCR2, HCK, HCST, HK3, HLA-DMB, IGSF6, IKZF1, IL10RA, IRF8, ITGAM, ITGB2, KLHL6, LAIR1, LAPTM5, LCP2, LILRB1, LILRB2, LILRB4, LRRC25, LST1, LY86, MNDA, MPEG1, MRC1, MS4A4A, MS4A6A, MS4A7, MSR1, MYO1F, NCF2, NCKAP1L, PIK3AP1, PILRA, PLEK, PSTPIP1, PTPRC, RNASE6, SASH3, SELPLG, SIGLEC1, SIGLEC7, SIGLEC9, SLAMF8, SLA, SLC7A7, SLCO2B1, SPI1, TBXAS1, TFEC, TLR8, TNFAIP8L2, TYROBP, VSIG4, WAS.

A3: CD74, HLA-DMA, HLA-DMB, HLA-DOA, HLA-DPA1, HLA-DPB1, HLA-DRA, HLA-DRB1

Metagene R, which is unique to stomach cancer, also contains 3 nearly disjoint parts (though R1 and R2 do have 12 genes in common). The R3 submetagene is also found in esophageal cancer.

R1: AKT3, ASAM, BNC2, C10orf72, CALD1, CDH11, CRISPLD2, DCN, DDR2, ECM2, EDNRA, FBN1, FBXL7, FGFR1, FSTL1, GLI3, GLT8D2, GPC6, GUCY1A3, GUCY1B3, JAM3, LAMA4, LHFP, MAP1A, MPDZ, MSRB3, NAP1L3, NDN, OLFML1, PDGFRB, PDLIM3, RECK, RUNX1T1, SDC2, SGCD, SSC5D, STON1, SYDE1, TCF4, TSHZ3, VGLL3, ZCCHC24, ZFPM2, ZNF423, ZNF521.

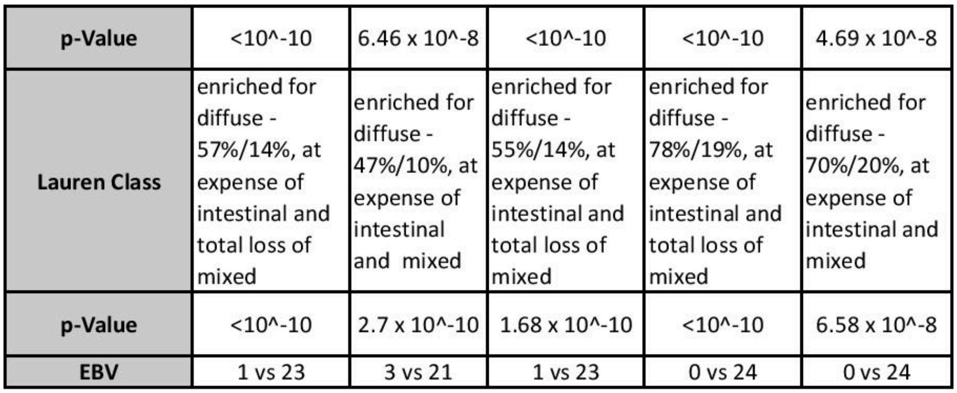
R2: ADAMTS12, ADAMTS2, AEBP1, ANTXR1, BGN, BICC1, BNC2, C10orf72, C1S, CCDC80, CCDC8, CDH11, COL10A1, COL12A1, COL1A1, COL1A2, COL3A1, COL5A1, COL5A2, COL6A1, COL6A2, COL6A3, COL8A1, COL8A2, CPZ, CSDC2, CTSK, DCN, DDR2, EFEMP2, EMILIN1, FAM180A, FAP, FBLN2, FBN1, FIBIN, FNDC1, FRMD6, FSTL1, GFPT2, GGT5, GLT8D2, GNB4, GPC6, GREM1, HMCN1, ISLR, ITGA11, ITGBL1, LAMA2, LUM, MFAP5, MMP2, MXRA8, NAP1L3, OLFML2B, PCOLCE, PLXDC2, PODN, PRKD1, PRRX1, PTGER3, RAB31, RSPO3, SCARF2, SERPINF1, SFRP2, SFRP4, SPARC, SPOCK1, SSC5D, SULF1, THBS1, THBS2, THY1, TIMP2, TIMP3, TMEM119, VCAN, VGLL3, ZEB2.

R3: ACTA2, ACTG2, ADAM33, ADCY5, ANGPTL1, ANK2, AOC3, ASB5, ATP1A2, BOC, C20orf200, C2orf40, C7, CALD1, CASQ2, CCDC80, CDH19, CHRDL1, CHRDL2, CHRM2, CNN1, CPXM2, DES, FAM19A4, FGF7, FHL1, FLNC, GNAO1, GRIK5, HAND2, HSPB6, HSPB7, HSPB8, IGF1, INMT, JPH2, KCNB1, KCNMA1, KCNMB1, KIAA0408, KIAA2022, LDB3, LMO3, LMOD1, LOC399959, LOC728264, MAMDC2, MGP, MRGPRF, MSRB3, MYH11, MYL9, MYLK, MYOCD, MYOC, NBLA00301, NEXN, NFASC, NGFR, NRXN1, NXPH3, OGN, PCDH10, PGM5, PLIN4, PLN, PLP1, PRELP, PRUNE2, RBPMS2, RGMA, SCN7A, SCRG1, SFRP1, SLITRK5, SMYD1, SOX10, SPARCL1, SSC5D, SYNC, SYNM, SYNPO2, TAGLN, TCEAL2, THBS4, TMEM35, TNS1.

However, the genes in R1, R2, and R3 do not exhaust all of metagene R. There are many genes in metagene R that do not show up in one of the parts, and in fact some of these are in our signatures: ABCA8, ABCC9, ABI3BP, ANGPTL2, ASPN, ATP8B2, BARX1, BHMT2, C14orf132, C6orf186, C7orf58, C9orf4, CADM3, CCL19, CDO1, CILP, CLIP3, COL14A1, CRTAC1, CRYAB, CYBRD1, CYP1B1, DAAM2, DACT3, DCLK1, DLG2, DNAJB5, DPYSL3, DZIP1, EFEMP1, EFS, EML1, EPHA7, FAM107A, FERMT2, FILIP1, FLNA, FLRT2, FOXP2, FXYD6, GEFT, GHR, GPRASP1, JAM2, KANK2, LDOC1, LGI2, LIMS2, LOC572558, MAP1B, MAP6, MAPK10, MEF2C, MEOX2, MN1, MOXD1, MRVI1, MXRA8, NDN, NEGR1, NRK, OMD, PDE1A, PDE3A, PDZD4, PDZRN4, PEG3, PLXNA4, PODN, PPP1R12B, PRICKLE2, PTGIS, PYGM, RCAN2, RGMA, RNF150, SCN2B, SDPR, SETBP1, SGCA, SLIT2, SLIT3, SMOC2, SORBS1, SVEP1, SYNE1, SYT11, TACR2, TGFB1I1, TMEM47, TMTC1, TNXB, TPM2, TSPAN2, VIPR2, ZCCHC24, ZDHHC15, ZEB1, ZFHX4.

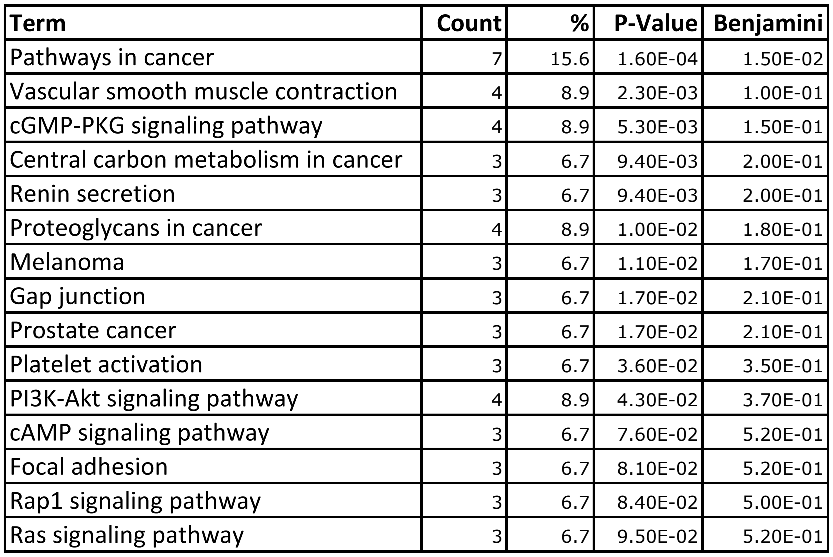
Because metagene A occurs with every type of cancer, we can refine the splitting into submetagenes by comparing multiple data sets. Metagene R is a major signal in only stomach cancer, with R3 also found with esophageal cancer. This leaves the splitting into submetagenes somewhat more coarse than the splitting for metagene A.

METAGENE	R1	R2	R3	DO4	DU4
No of genes	45	81	87	4	4
number Altered/Unaltered samples	70/195	107/158	71/194	30/235	27/238
gene expression cluster	enriched for C1 - 58%/5% at the expense of all	enriched for C1 - 44%/3% at the expense of all	enriched for C1 - 58%/5% at the expense of all	enriched for C1 - 90%/10% at the expense of all, but no C3 samples	enriched for C1 - 84%/12% at the expense of all, but no C3 samples
p-Value	<10^-10	<10^-10	<10^-10	<10^-10	<10^-10
Mutation Count Average	61 vs 116	77 vs 134	58 vs 134	39 vs 137	39 vs 116
p-Value	2.17 x 10^-10	2.12 x 10^-6	<10^-10	2.36 x 10^-10	3.36 x 10^-10
WHO	enriched for poorly cohesive - 59%/15%, at the expense tubular 30%/58%, total loss of mixed	enriched for poorly cohesive - 48%/12%, at the expense of all but mainly tubular 35%/62%	enriched for poorly cohesive - 56%/15%, at the expense tubular 27%/60%, total loss of mixed	enriched for poorly cohesive - 79%/20%, at the expense of tubular 17%/55%, with total loss of mucinous and mixed	enriched for poorly cohesive - 70%/21%, at the expense of tubular 22%/54%, with total loss of mucinous
p-Value	3.35 x 10^-10	1.39 x 10^-8	6.68 x 10^-9	2.02 x 10^-9	3.21 x 10^-6
CIMP					only 1 sample compared to 92, 3%/39%
p-Value					1.46 x 10^-3
molecular subtype	enrichment of GS and decrease in all others with 1 EBV - 48%/10%	enrichment of GS and decrease in all others with 3 EBV - 37%/9%	enrichment of GS and decrease in all others with 1 EBV - 48%/10%	enrichment of GS and decrease in all others with total loss of EBV - 70%/14%	enrichment of GS and decrease in all others with total loss of EBV and MSI - 63%/16%



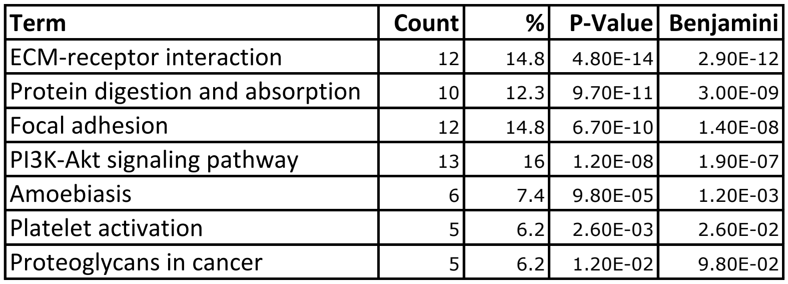
p-Value	<10^-10	6.46 x 10^-8	<10^-10	<10^-10	4.69 x 10^-8
Lauren Class	enriched for diffuse - 57%/14%, at expense of intestinal and total loss of mixed	enriched for diffuse - 47%/10%, at expense of intestinal and mixed	enriched for diffuse - 55%/14%, at expense of intestinal and total loss of mixed	enriched for diffuse - 78%/19%, at expense of intestinal and total loss of mixed	enriched for diffuse - 70%/20%, at expense of intestinal and mixed
p-Value	<10^-10	2.7 x 10^-10	1.68 x 10^-10	<10^-10	6.58 x 10^-8
EBV	1 vs 23	3 vs 21	1 vs 23	0 vs 24	0 vs 24

Table 7: Analysis on all 4 stages



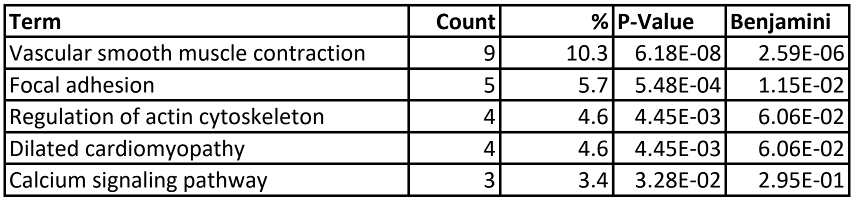
Term	**Count**	**%**	**P-Value**	**Benjamini**
Pathways in cancer	7	15.6	1.60E-04	1.50E-02
Vascular smooth muscle contraction	4	8.9	2.30E-03	1.00E-01
cGMP-PKG signaling pathway	4	8.9	5.30E-03	1.50E-01
Central carbon metabolism in cancer	3	6.7	9.40E-03	2.00E-01
Renin secretion	3	6.7	9.40E-03	2.00E-01
Proteoglycans in cancer	4	8.9	1.00E-02	1.80E-01
Melanoma	3	6.7	1.10E-02	1.70E-01
Gap junction	3	6.7	1.70E-02	2.10E-01
Prostate cancer	3	6.7	1.70E-02	2.10E-01
Platelet activation	3	6.7	3.60E-02	3.50E-01
PI3K-Akt signaling pathway	4	8.9	4.30E-02	3.70E-01
cAMP signaling pathway	3	6.7	7.60E-02	5.20E-01
Focal adhesion	3	6.7	8.10E-02	5.20E-01
Rap1 signaling pathway	3	6.7	8.40E-02	5.00E-01
Ras signaling pathway	3	6.7	9.50E-02	5.20E-01

Table 8: R1 Pathway analysis using DAVID



Term	**Count**	**%**	**P-Value**	**Benjamini**
ECM-receptor interaction	12	14.8	4.80E-14	2.90E-12
Protein digestion and absorption	10	12.3	9.70E-11	3.00E-09
Focal adhesion	12	14.8	6.70E-10	1.40E-08
PI3K-Akt signaling pathway	13	16	1.20E-08	1.90E-07
Amoebiasis	6	7.4	9.80E-05	1.20E-03
Platelet activation	5	6.2	2.60E-03	2.60E-02
Proteoglycans in cancer	5	6.2	1.20E-02	9.80E-02

Table 9: R2 Pathway analysis using DAVID



Term	Count	%	P-Value	Benjamini
Vascular smooth muscle contraction	9	10.3	6.18E-08	2.59E-06
Focal adhesion	5	5.7	5.48E-04	1.15E-02
Regulation of actin cytoskeleton	4	4.6	4.45E-03	6.06E-02
Dilated cardiomyopathy	4	4.6	4.45E-03	6.06E-02
Calcium signaling pathway	3	3.4	3.28E-02	2.95E-01

Table 10: R3 Pathway analysis using DAVID