INTEGRATIVE ANALYSIS FOR LUNG ADENOCARCINOMA PREDICTS MORPHOLOGICAL FEATURES ASSOCIATED WITH GENETIC VARIATIONS*

CHAO WANG
Electrical and Computer Engineering, The Ohio State University
Columbus, Ohio, 43210, USA
Email: wang.2031@osu.edu

HAI SU
Biomedical Engineering, University of Florida
Gainsville, Florida, 32611, USA
Email: hai.su@bme.ufl.edu

LIN YANG
Biomedical Engineering, University of Florida
Gainsville, Florida, 32611, USA
Email: lin.yang@bme.ufl.edu

KUN HUANG
Biomedical Informatics, The Ohio State University
Columbus, Ohio, 43210, US
Email: kun.huang@osumc.edu

Lung cancer is one of the most deadly cancers and lung adenocarcinoma (LUAD) is the most common histological type of lung cancer. However, LUAD is highly heterogeneous due to genetic difference as well as phenotypic differences such as cellular and tissue morphology. In this paper, we systematically examine the relationships between histological features and gene transcription. Specifically, we calculated 283 morphological features from histology images for 201 LUAD patients from TCGA project and identified the morphological feature with strong correlation with patient outcome. We then modeled the morphology feature using multiple co-expressed gene clusters using Lasso-regression. Many of the gene clusters are highly associated with genetic variations, specifically DNA copy number variations, implying that genetic variations play important roles in the development cancer morphology. As far as we know, our finding is the first to directly link the genetic variations and functional genomics to LUAD histology. These observations will lead to new insight on lung cancer development and potential new integrative biomarkers for prediction patient prognosis and response to treatments.

1. Introduction

Lung cancer is one the most deadly cancers in the world. Among lung cancers, lung adenocarcinoma (LUAD) is a subtype of the non-small cell lung cancer (NSCLC) and is the most common histological type of lung cancers (*1*). However, despite the fact that it is a sub-classification of lung cancer, LUAD is a heterogeneous group of tumors with a highly variable prognosis and responses to treatment (*2*).

The high-throughput sequencing technologies are making targeted therapies possible for LUAD (*3*). The advance of these technologies allows molecular diagnostic biomarkers for the detection of lung cancer in addition to computed tomography (CT) screening (*4–7*). For example, the utility of epidermal growth factor receptor (EGFR) mutation testing is strongly recommended (*8*) in clinical practice. However, although EGFR-mutant lung cancers are

* This work is supported by UK-OSU Joint CCTS grant, NCI ITCR 1U01CA188547-01A1 grant, the OSU Pelotonia Fellowship, and the Ohio Supercomputer Center.

sensitive to EGFR tyrosine kinase inhibitors (TKIs), they develop resistance (*9*). Therefore, novel biomarkers for for LUAD are needed for enhanced personalized treatment.

Lung cancer diagnosis and classification have been traditionally based on imaging approaches, such as CT and histopathology (*10*, *11*). For instance, five distinct histologic subtypes and radiologic patterns have been reported recently. Traditionally, histopathology images serve as a golden standard for lung cancer diagnosis. Cellular and inter-cellular level morphology are usually used by the pathologists for making diagnosis decisions. However, the current pathology diagnosis is commonly based on individual pathologists' interpretations of the samples which are subject to large inter-observer variations and low throughput analysis. Unbiased quantitative pathology methods are showing promise by offering more cellular information (*12–14*). Recently, pathology informatics study on lung cancer has attracted more interests. In one study (*15*), the diagnostic significance of nuclear features in differentiating small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC) was investigated. Edwards et.al.(*16*) showed that adenocarcinoma diagnosis is more challenging compared to squamous carcinoma. An early automatic pathology analysis system was proposed in (*17*). In the study by Mijović et al. (*18*), diagnostic values of seven Karyometric variables are examined for diagnosis of major histological types of lung carcinoma. In Zhang et al's study (*19*), an image classification system is proposed to differentiate lung adenocarcinoma and squamous carcinoma. The work by Yao et al (*20*) developed topological features for lung cancer diagnosis. Compared to genomic biomarkers, advanced imaging may provide more clinically relevant information.

In order to take advantage of both the richness of histopathological information and molecular profiles, we aim to develop an integrative computational pipeline that exploits diagnostic images and mRNA expression. A related work on lung cancer was recently published on integrating histopathologcal images with genetic data for outcome prediction (*21*). The pipeline allowed us to discover the associations between cellular level and molecular level phenotypes, and thus novel biomarkers can be unveiled. In this paper, we extracted 283 histopathological features from LUAD tissue slides and initially attempted to identify co-expression gene clusters that have high correlation with these image features. Such approach in other cancers has led to new insight on cancer biology and new potential biomarkers (*22*). However, as shown in this paper, the morphology of LUAD is much more complicated and it turned out that the morphological features have low correlations with gene expression profiles. Figure 1 shows a 'highly-correlated' pairs between the imaging features and gene clusters. It is thus plausible that the LUAD morphology is regulated by any particular group of genes; instead a specific morphological characteristic is a manifestation of a combined effect from multiple groups of genes. Based on these quantitative experiments, we assert that a multivariate model is needed.

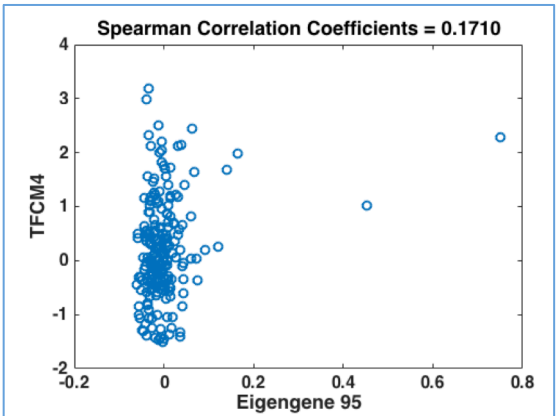

Figure 1: The scatter plot between the eigengene 95 and the *tfcm4*, which have the highest SCC between all eigengenes with tfcm4 (SCC = 0.170).

Therefore in this paper, we demonstrate that the morphological characteristics of LUAD can be explained by a combination of multiple gene clusters identified using sparse modeling based on the Lasso algorithm. In addition, we found that many of the gene clusters are associated with putative copy number variations, implying that genetic variations play important roles in the development cancer morphology. As far as

we know, our finding is the first to directly link the genetic variations and functional genomics to LUAD histology. These observations will lead to new insight on lung cancer development and potential new integrative biomarkers for prediction patient outcome and response to treatments.

2 Methods and Materials

Our analysis involve molecular and histological analysis based on data from The Cancer Genome Atlas (TCGA) LUAD project. The data we use include mRNA profiling, histological images and clinical data including survival information.

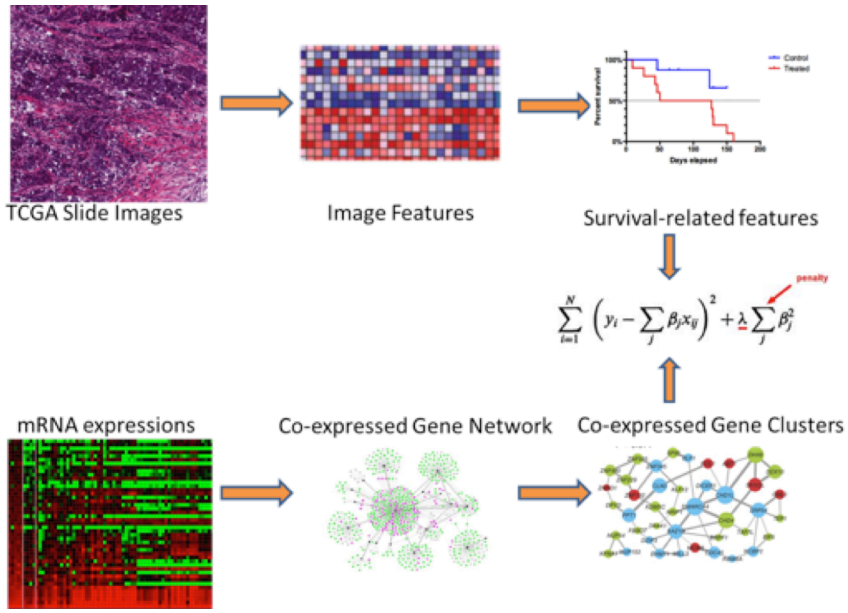

Figure 2: Schematic overview of constructing gene co-expression networks and analyzing the relationships between gene networks and morphological features.

2.1 Integrated Analysis Pipeline

We collected matched diagnostic images and gene expression data for a discovery dataset of 201 LUAD patients from the TCGA. The integrative analysis workflow is shown in Figure 2. Our automatic imaging processing pipeline detected cell nuclei and extracted predefined features evaluating staining variations. To select imaging features with clinical relevance, survival-related imaging features were identified. At molecular level, gene expression profiles (mRNA levels) were filtered and clustered using our co-expression network analysis algorithm. Strongly co-expressed gene clusters were represented by eigengenes. Then, we built a lasso regression model to select gene clusters that regulate the image feature that has the strongest association with survival times. By finding the co-expression patterns that are associated with the selected imaging feature, we can discover biological processes and genetic variations associated with cancer histology.

2.2 Image and Genomic Data Collection

We focus on LUAD patients with clinical information, genomic information, and histopathologic whole slide images. The data were downloaded from TCGA (The Cancer Genome Atlas) Data Portal. Data for 201 LUAD patients with all the three data types are downloaded for the experiments in 2014. For each patient, a representative image patch of size 1712 x 952 without damage or artifact is cropped from the tumor region. Expression profiles of 20,530 unique genes were investigated in the 201 patients (*23*).

2.3 Data Preprocessing and Imaging Feature Extraction

2.3.1 Imaging features

We adopt the cell detection and segmentation methods proposed in (*24*). In the cell

detection stage, a radial voting scheme with Gaussian pyramid is employed (*25*). For each image, a Gaussian pyramid is created. A single–pass voting is applied to each layer. The voting region receives scores weighted by a distance transform. Therefore, such weighted voting encourages the pixels closer to the cell center accumulates higher voting scores. The final voting score is obtained by summing up the voting scores from different layers. In the segmentation stage, a marker based active contour with a repulsive term is applied to the images using the detection results as the markers. An initial contour associated with each detected marker is created first. The contours evolve through an iterative procedure to reach the real boundaries of the cells. The repulsive term serves to prevent the contours from crossing and merging with each other.

Group 1: Geometry Features. Based on the segmentation results, five geometry features are calculated for each lung cancer cell to capture the cell shape information, including cell area, contour perimeter, circularity, major-minor axis ratio, and contour solidity. Contour solidity is defined as the ratio of the area of a cell region over the convex hull defined by the segmentation boundary.

Group 2: Pixel Intensity Statistics. Pixel intensity statistics features are used to capture the color of the segmented cells. This group of features are calculated based on the intensity of the pixels within the segmented cells, including intensity mean, standard deviation, skewness, kurtosis, entropy, and energy. *Lab* color space is used for a better color representation.

Group 3: Texture Features: Texture is an important feature found to be closely related to cancer diagnosis in radiomics. This is rooted in the fact that texture patterns are linked to difference in protein expressions (*26*). This group of features consists of co-occurrence matrix (*27*), center symmetric auto-correlation (CSAC) (*28*), local binary pattern (LBP) (*29*), texture feature coding method (TFCM) (*30*). The co-occurrence matrix (*27*) computes an estimation of the joint probability distribution of the intensity of two neighboring pixels. CSAC is a measure of the local patterns with symmetrical structure. These patterns are characterized by a series of local auto-correlation and covariance introduced in (*28*), including symmetric texture covariance (SCOV), variance (SVR), and within-pair variance (WVAR), and between-pair variance (BVAR). 3×3 pixel unit of each channel is considered. LBP (*29*) feature measures the local textures by assigning a binary code to a pixel with respect to its intensity and those of its neighboring pixels. A histogram of the generated binary codes reveals the distribution of the present repeated local patterns. Similar to LBP, in TFCM (*30*), a texture feature number (TFN) is assigned to each pixel by comparing this pixel with its neighbors in four directions: 0°, 45°, 90°, and 135°. A histogram is calculated based on the TFNs of one image patch.

2.3.2 Gene transcriptome data

The expression profiles of 201 samples with primary lung cancer adenocarcinoma from TCGA LUAD project were downloaded from TCGA data portal in January 2014. Specifically, RNA-seq data for the tumor samples were obtained using Illumina sequencing and processed as described in (*6*). The mapping results were converted to RPKM (read per kilobase per million reads) values for 20,530 genes. Genes with low expression levels (with no data in the top 15 percentile) and low variance (in the lowest 10 percentile) were removed resulting in 9,179 genes.

2.4 Gene co-expression network analysis and summarization

While our goal is to establish the relationships between gene expression levels and the imaging features, we first carry out gene co-expression network analysis (GCNA) to cluster

genes into co-expressed clusters. There are multiple reasons for carrying out GCNA before associating them with the imaging features. First, there is a large number of genes. If the association between every pair of gene and imaging feature is calculated and tested for significance, then more than half a million tests will be carried out which leads to low statistical power. In addition, since we will explore the association beyond univariate relationships using sparse analysis, the large number of genes (which are not always independent), pose serious computing challenges to the sparse modeling algorithms such as Lasso. Thus we first group genes with highly correlated expression profiles into co-expression clusters using GCNA then summarize the expression profiles within each cluster as an "eigengene" using the protocol described in (*31*). Essentially the expression profiles of each gene are first centralized (by subtracting the mean for each gene) and then standardized to have norm one. After the processing steps, singular value decomposition is applied to obtain the *eigengene* as the principal vector in the direction with the largest variance among the samples. Another advantage of the GCNA approach is that the highly co-expressed gene clusters are usually highly enriched in specific biological processes, regulatory factors or structural variations (e.g., copy number variations) (*32*), making the interpretation of the results straightforward.

While there are many algorithms for performing GCNA including the well known WGCNA package, we use a weighted network mining algorithm called local maximum quasi-clique merging (lmQCM) algorithm we recently developed (*32*). Unlike WGCNA which uses hierarchical clustering and does not allow overlaps between clusters, our algorithm is a greedy approach allowing genes to be shared among multiple clusters, in consistent with the fact the genes often participate in multiple biological processes. In addition, we have shown that lmQCM can find smaller co-expressed gene clusters which are often associated structual mutations such as copy number variations in cancers. The lmQCM algorithm has four parameters γ, α, t, and β. Among these parameters, γ is the most influencial, it decides if a new cluster can be initiated by setting the weight threshold for the first edge of the cluster as a subnetwork. In our GCN analysis, we directly use the absolute values of the Spearman correlation coefficients between expression profiles of genes as weights for which we have shown to be effective in previous studies.

2.4 Associations between Morphology and Transcriptomes

2.4.1 Correlation analysis

We first examined the correlation between the imaging features and the eigengenes for the gene clusters identified using lmQCM by calculating the Spearman correlation coefficients between them. However, as shown in the Results, the correlations between imaging features and eigengenes are not strong (none of them is significant if Bonferroni correction is applied for multiple test compensation). While this is different from the case in breast cancer, it suggests that the tissue morphology development is a complicated process involving in multiple processes and genetic factors. Thus in order to explain the morphology development, we need to resort to multi-variate modeling methods such as lasso regression.

2.4.2 Sparse modeling using Lasso regression

We model imaging features as manifestations of gene expression. Given the data availability, we focus on transcriptome data. Lasso regression model minimizes the residual sum of squares while at the same time enforcing sparsity of the model by adding a penalty term of the L_1-norm of the model coefficients.

Consider the linear regression model: we have (x_i, y_i), $i = 1,2, \dots, N$, where $x_i = (x_{i1}, \dots, x_{ip})^T$ and y_i are eigen-gene expression and image feature value for the i th

observation(patient sample), respectively. With regular regression model, the least square estimates are obtained by minimizing the residual squared error. However, in feature selection models to predict biomarkers, only imperative transcriptomes contribute to biological functions and processes, requiring more stringent and interpretable features. With large number of features, we would like to determine a small subset of them that can predict strong correlations. Let $\beta = (\beta_1, \dots, \beta_p)^T$ and β_0 to be a scalar. The lasso model estimate (β, β_0) by solving the following problem

$$\min_{\beta, \beta_0} \left(\frac{1}{2N} \sum_{i=1}^{N} (y_i - \beta_0 - x_i{}^T \beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j|\right), \quad (1)$$

where λ is nonnegative regularization parameter giving the weight for the model complexity term. As λ increases, the number of nonzero components of β decreases, leading to smaller numbers of predictors.

2.5 Identification of Survival-Related Image Features

Univariate Cox Proportional Hazard models are used to identify morphological features and genes that have expression related significantly to survival. Morphological features that have p-values less than 0.05 are recorded.

Table 1: Prognostic values of various image features in discover dataset. The features are listed by their significance in the survival model.

Feature Names	p-value		Feature Names	p-value
tfcm4	0.00456904		contrast1	0.01210092
tfcm9	0.00532429		tfcm12	0.01247155
tfcm3	0.00563955		tfcm11	0.01361604
tfcm1	0.0064998		csac23	0.01754474
tfcm2	0.00657692		tfcm7	0.0178572
tfcm10	0.00685436		fourier15	0.0178766
contrast2	0.0082282		csac5	0.01896244
tfcm8	0.0093341		entry4	0.01995154

2.6 Validating the Identified Genes in Other Data Set

For genes associated with the imaging features with the highest potential for predicting patient survival, we also test them on another publicly available dataset obtained from the NCBI Gene Expression Omnibus. The dataset GSExxxx contains transcriptome data of 149 non-small cell lung cancer patients, among which 90 are unique lung adenocarcinoma patients with clinical outcome (survival time and status). We use the genes to be tested as features to separate the 90 patients into two groups using K-means algorithms (K=2, Euclidean distance, average linkage, and 10 replicates). The survival times of the two groups are then visualized using Kaplan-Meier curves and compared using Cox Proportional Hazard regression.

2.7 Enrichment analysis of gene clusters

To interpret the biological meaning of the identified gene clusters, enrichment analysis tools such as TOPPGene (https://toppgene.cchmc.org/enrichment.jsp) are used. In addition, information about the genes are extracted from cBioPortal (http://www.cbioportal.org/).

3 Results

3.1 Image Feature Calculation

As shown in Figure 3 Left, the images reveal clear heterogeneity of the tumors among the patients. We calculated 283 image features from the images. As described in Section 2.3.1, there are multiple types of features and many features are strongly correlated (Figure 3 Right) such as part of the TCFM family (the block of 211 to 222). In this paper, we analyze

each feature individually, but some of the highly features can be combined in future analysis.

3.2 Survival-related Image Features and Gene Cluster

Using a univariate Cox proportional hazards regression model, we assessed the image features related risk score in the prediction of the LUAD patient survival. Significant

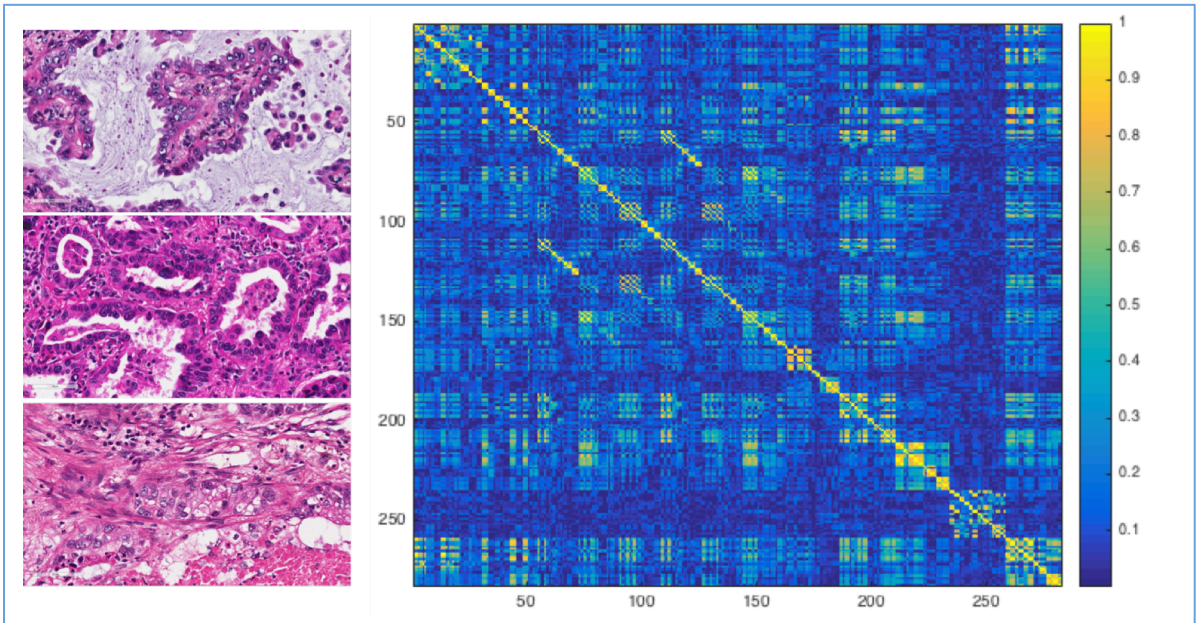

Figure 3. **Left**: Examples of the image patches from different patients. **Right**: Heatmap of the correlation (Spearman) matrix for the 283 image features.

morphological features are listed in Table 1. Among the six categories of imaging features, the *tfcm* category shows the most significant prognostic power, indicating texture features in lung adenocarcinoma have a strong potential for predicting patients' outcomes. In fact, all of the top six survival-related imaging features are in the *tfcm* category. Other features that capture prognosis are *contrast2*, *contrast*, *csac23*, *fourier15*, *csac5* and *entry4*.

3.2 Gene Co-Expression Network Analysis

As mentioned in Section 2.3.2, *9,179* genes were kept for analysis. The absolute value of the Spearman rank correlation coefficients were used for cluster detection using lmQCM algorithm. We allow the smallest gene clusters to have five genes. Then we found with γ = 0.75, t = 1, α = 1, and β = 0.4 the algoirthm yielded co-expressed gene clusters with balanced sizes. Specifically, it led to 95 clusters ranging from 5 to 120 genes. Many of the gene clusters are consistent to the ones frequently found in cancers. Most of these clusters involved in hallmark cancer biological processes such as cell cycle/genome stability (cluster 1), immune respones (cluster 2), translation / protein synthesis (cluster 3), and extracellular matrix development (cluster 7). However, some of them are more associated with specific cytobands (e.g., chr19p13), implying potential CNV sites.

3.3 Correlations between Image Features and Gene Clusters

The image analysis pipeline allowed us to quantify tumor characteristics on cellular level and associate these tumor characteristics with patient outcomes. In this study, we calculated 283 imaging features for the 201 patients and correlated with the 95 eigengenes. The correlation coefficient with the large absolute value is -0.2990 (p=1.7728e-05). In Table 2, we list the strongest correlation between eigengenes and the top five imaging features (in

Table 1) with the most significant power for predicting patient outcome. It is clear from the table that none of such correlations is statistically significant (after multiple test compensation), suggesting that complex phenomena such as cell and tissue morphology in lung cancers can only be explained by multiple molecular and genetic factors.

Table 2. Imaging features and the eigengenes with the strongest correlations with them.

Imaging feature	Eigengene (cluster)	SCC/p-value	Enrichment
tfcm4	95	0.1710/0.0153	18q12.1 (p=1.175e-9), all five genes on 18q12
tfcm9	59	0.1677/0.0174	16p11.2 (p=1.364e-10), all seven genes on 16p11
tfcm3	59	-0.1658/0.0188	
tfcm1	59	0.1704/0.0157	
tfcm2	59	0.1508/0.0327	

3.4 Lasso Regression Model for Imaging Features Using Eigengenes

Since the imaging features with prognostic power do not have strongly correlated gene clusters, we resort to multivariate models to explain the cell and tissue morphology using molecular data. Specifically, we built a lasso regression model. The lasso model selects a sparse set of eigengenes to explain the selected imaging feature. We rank the importance of image features by their significance in survival analysis. The top 10 image features in Table 1 belong to only two categories – TFCM and Contrast. Features within each category are highly correlated (for the eight TFCM features, the smallest of the absolute value of the SCC is 0.6840, the two SCC between the two Contrast features is 0.9923). Since eight out of 10 top image features are from the TFCM family, we chose one feature from for our modeling, namely *tfcm4*.

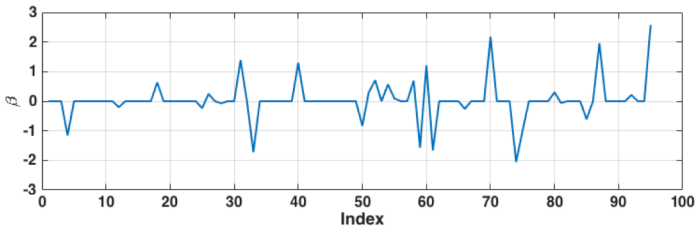

Figure 4. Coefficients (β values) of the lasso regression for *tfcm4*.

For *tfcm4*, it is found that the lowest MSE is found at λ = 0.0371 for the cost function in Eq.(1). Figure 4 shows the values of the coefficients β. Among the 95 eigengenes, 28 have non-zero coefficients among which 18 are larger than 0.5 and 12 are larger than 1. For the analysis of genes, we collected 185 genes from the 18 clusters with absolute value of coefficients larger than 0.5. In addition, Figure 5 shows the correlation between the combined eigengenes using the calculated β values with the *tfcm4* values in contrast to the correlation between the 95th eigengene (as listed in Table 2) and *tfcm4* (Figure 1).

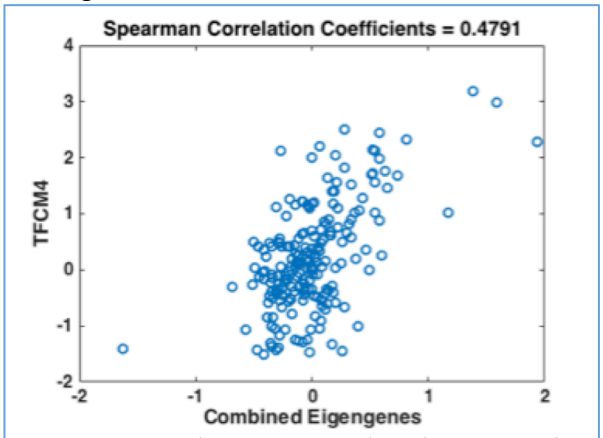

Figure 5. The scatter plot between the combined eigengenes using the lasso coefficients and the *tfcm4* (SCC = 0.4791).

3.5 Functional and Genetic Analysis of Gene Clusters Associated with Imaging Features

In order to understand the functional roles of the gene clusters associated with *tfcm4*, enrichment analysis was carried out

using TOPPGene and the results for the 18 gene clusters are shown in Table 2. Among the gene clusters whose eigengenes are associated with tfcm4, the largest cluster is the cluster #4, consisting of 59 genes and is highly enriched with ribosomal genes and thus protein translation function. Other related biological processes including immune response (response to virus, cluster #18), response to steroid hormone, negative regulation of epithelial cell proliferation, and mitochondrial ATP synthesis.

Interestingly, 14 out of the 18 gene clusters are highly enriched on specific cytobands. It has been previously noticed that many of the co-expressed clusters in cancers are associated with copy number variations (CNVs) in specific cytobands (*32*). CNVs are common genetic variations playing important roles cancer initiation and development. Functional CNVs usually lead to changes in expression levels of genes on that region due to the "dose effect", which also leads to co-expression of the transcribed genes. Figure 6 Left shows an example of the RPRD1A gene in cluster #95, whose mRNA level has a strong correlation with its copy number measurement and it shows a strong co-expression relationship with the ELP2 genes on the same cytoband.

Table 2: Gene clusters showing strong correlation with texture image feature tfcm, and their Gene Ontology terms and enriched cytobands.

Gene Cluster (size)	beta	GO Biological Process/p-values	Cytobands/p-values	Notes:
4 (59)	-1.1558	GO:0006614 SRP-dependent cotranslational protein targeting to membrane / 9.105E-98		
18 (14)	0.6328	GO:0009615response to virus/ 9.965E-15		
31 (10)	1.3894			Genes down-regulated in nsopharyngeal carcinoma relative to the normal tissue (p = 5.074e-19, all 10 genes)
33 (10)	-1.7213		19q13.42/5.525e-6	All 10 genes on 19q13.3-4
40 (8)	1.2977		8q24.13/3.263e-5	Seven genes on 8q21-24, one on 8q13
50 (8)	-0.8343	GO:0048545response to steroid hormone / 2.290E-8		
52 (8)	0.7075		7q33/4.800E-5	All eight genes on 7q21-36
54 (7)	0.5669	GO:0006413 translational initiation /1.096E-5	Yq11/2.305E-6, Xq13.2/2.856E-5	Four genes on Yq11, two on Xq13.2, one on Yp11.3
58 (7)	0.6952		8p21.1/ 6.631E-6	Five genes on 8p21, two on 8p12
59 (7)	-1.5729		16p11.2/1.364e-10	All seven genes on 16p11
60 (7)	1.2103		Xq28/1.982e-13	All seven genes on Xq27-28
61 (7)	-1.6639		6p21.1/4.436e-7	Six genes on 6p21-22, one on 6p12
70 (6)	2.1783		17q21.31/5.532e-	All six genes are on 17q21

			10	
74 (6)	-2.0544		8p11.2/1.048e-9	All six genes are on 8p11.2
75 (6)	-1.0093	GO:0050680 negative regulation of epithelial cell proliferation / 3.290E-6	17q11.2/6.880e-7	All six genes are on 17q11-12
85 (5)	-0.6095	GO:0042776 mitochondrial ATP synthesis coupled proton transport / 6.311E-9	21q22.11/3.344E-5	Four genes on 21q21-22
87 (5)	1.9569		19q13.2/1.131e-6	All five genes on 19q13
95 (5)	2.5783		18q12.1/1.175e-9	All five genes on 18q12

3.4 Prognostic Validation

Validation on heterogeneous external data sets allows for evaluation of the generalizability. To test the importance of cilium-related genes, we further performed survival analysis on a publicly available dataset with 90 LUAD patients. Among the 185 genes correlated with the image feature category tfcm, 118 of the gene symbols can be matched exactly to the external dataset. In the validation dataset, lung adenocarcinoma patients were stratified into two groups using K-means based on their expression levels of the 118 genes. In both datasets, a statistically significant group of patients with worse outcomes were differentiated (n = 44 and n = 46, respectively). The difference between the two groups is significant (Cox hazard proportional model p-value 2.1285e-6). Figure 7 shows the Kaplan-Meier curves of the two patient cohorts.

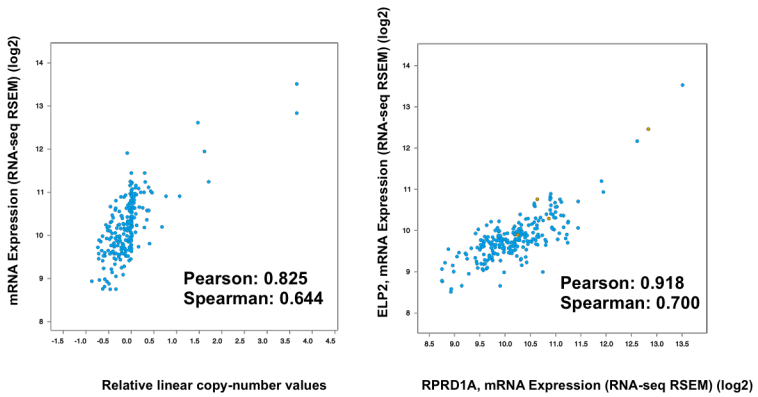

Figure 6. **Left**: Gene RPRD1A shows strong correlation between mRNA level and copy number values. **Right**: RPRD1A and ELP2 (both in cluster #95) are strongly co-expressed.

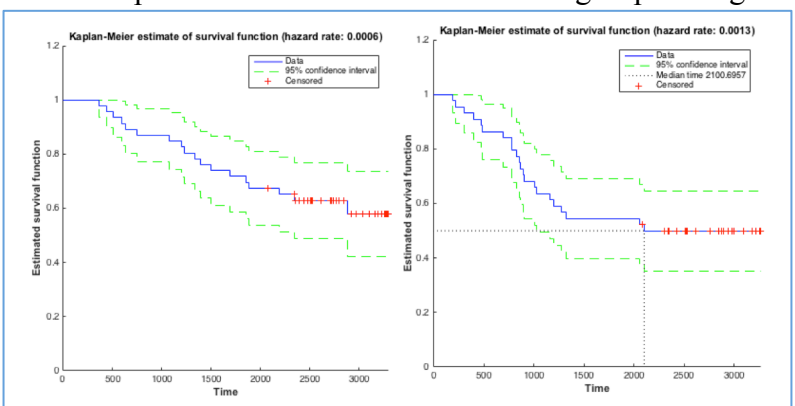

Figure 7: Kaplan–Meier survival curves of prognostic model on two clusters from the lung adenocarcinoma dataset. Survival curves of two groups (group 1 with 44 patients on the left and group 2 with 46 patients on the right). The difference of survival time between the two groups is significant (Cox hazard proportional model p-value 2.1285e-6).

4 Discussion and Conclusion

Our integrative analysis

pipeline allows us to find survival related textural features of lung adenocarcinoma. In addition to the image features, we also demonstrated that modeling of the histology at cellular and tissue levels using "omics" data may involve multiple groups of genes. Interestingly, our results showed that that the histological phenotype may be manifestations of multiple genetic variations, especially copy number variations. Specifically, many of the enriched cytobands we identified have been previously associated with lung cancer development including 19q13 (*33*, *34*) , 8q24 (*33*), 7q21-36 (*35*), 8p21 (*33*), 16p11 (*36*), Xq27-28, 6p21 (*34*), 17q21 (*34*), 21q22 (*35*), and 18q12 (*33*). While there is no report on the association of Xq27-28 with lung cancer, Xq26 has been shown to be associated with lung cancers (*36*), suggesting that the genetic variations should be further explored to identify potential "driver" genes for lung cancer. We also showed that the genes in the clusters can indeed predict patient prognosis, which leads to discovery of potential biomarkers. While our study is focused on patient prognosis, the process can be repeated for patient treatment response prediction with appropriate data. Overall we demonstrated that the morphology is a complex phenomenon and its development may involve multiple groups of genes. In cancers, this process is even more complex as the genetic variations also contribute significantly to this process. Our findings indeed support this notion.

References

1. S. Couraud, G. Zalcman, B. Milleron, F. Morin, P.-J. Souquet, *Eur. J. Cancer*. **48**, 1299–311 (2012).
2. P. A. Russell *et al.*, *J. Thorac. Oncol.* **6**, 1496–504 (2011).
3. E. Conde *et al.*, *Clin. Transl. Oncol.* **15**, 503–8 (2013).
4. D. Hokka *et al.*, *Lung Cancer*. **79**, 77–82 (2013).
5. X. Li *et al.*, *Neoplasma*. **59**, 500–7 (2012).
6. E. A. Collisson *et al.*, *Nature*. **511**, 543–50 (2014).
7. C. Camps, Jantus-Lewintre, Usó, Sanmartin, *Lung Cancer Targets Ther.*, 21 (2012).
8. P. F. Robert T. Adamson, *Am. J. Manag. Care*. **19** (2013).
9. J. Chmielecki *et al.*, *J. Thorac. Oncol.* **7**, 434–42 (2012).
10. J. H. M. Austin *et al.*, *Radiology*. **266**, 62–71 (2013).
11. L. M. Solis *et al.*, *Cancer*. **118**, 2889–99 (2012).
12. Y. Yuan *et al.*, *Sci. Transl. Med.* **4**, 157ra143–157ra143 (2012).
13. a. H. Beck *et al.*, *Sci. Transl. Med.* **3**, 108ra113–108ra113 (2011).
14. H. Wang, F. Xing, H. Su, A. Stromberg, L. Yang, *BMC Bioinformatics*. **15**, 1–12 (2014).
15. F. B. Thunnissen *et al.*, *Pathol. Res. Pract.* **188**, 531–5 (1992).
16. S. L. Edwards *et al.*, *J. Clin. Pathol.* **53**, 537–40 (2000).
17. K. Kayser, D. Radziszowski, P. Bzdyl, R. Sommer, G. Kayser, *Rom. J. Morphol. Embryol.* **47**, 21–8 (2006).
18. M. Mijovic, Zaklina; Mihailovic, Dragan; Kostov, *Med. Biol.* **15**, 28 – 32 (2008).
19. X. Zhang, L. Yang, W. Liu, H. Su, S. Zhang, in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2014* (2014), pp. 479–486.
20. J. Yao *et al.*, in *Proceedings of the 6th International Workshop on Machine Learning in Medical Imaging - Volume 9352* (Springer-Verlag New York, Inc., 2015; http://link.springer.com/10.1007/978-3-319-24888-2_35), pp. 288–295.
21. X. Zhu *et al.*, in *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)* (IEEE, 2016; http://ieeexplore.ieee.org/document/7493475/), pp. 1173–1176.
22. C. Wang *et al.*, *J. Am. Med. Inform. Assoc.* **20**, 680–7.

23. TCGA, The Cancer Genome Atlas - Data Portal.
24. F. Xing, L. Yang, in *2013 IEEE 10th International Symposium on Biomedical Imaging* (IEEE, 2013), pp. 386–389.
25. X. Qi, F. Xing, D. J. Foran, L. Yang, *IEEE Trans. Biomed. Eng.* **59**, 754–65 (2012).
26. P. Lambin *et al.*, *Eur. J. Cancer*. **48**, 441–6 (2012).
27. R. M. Haralick, K. Shanmugam, I. Dinstein, *IEEE Trans. Syst. Man. Cybern.* **3**, 610–621 (1973).
28. K. Laws, in *24th Annual Technical Symposium*, T. F. Wiener, Ed. (International Society for Optics and Photonics, 1980), pp. 376–381.
29. T. Ojala, M. Pietikainen, T. Maenpaa, *IEEE Trans. Pattern Anal. Mach. Intell.* **24**, 971–987 (2002).
30. M.-H. Horng, Y.-N. Sun, X.-Z. Lin, *Comput. Med. Imaging Graph.* **26**, 33–42 (2002).
31. P. Langfelder, S. Horvath, *BMC Bioinformatics*. **9**, 559 (2008).
32. J. Zhang, K. Huang, *Cancer Inform.* **1**, 1 (2016).
33. B. R. Balsara *et al.*, *Cancer Res.* **57**, 2116–20 (1997).
34. P. P. Medina *et al.*, *Hum. Mol. Genet.* **18**, 1343–52 (2009).
35. F. Li, L. Sun, S. Zhang, *Oncol. Rep.* **34**, 1701–7 (2015).
36. N. A. Levin *et al.*, *Cancer Res.* **54**, 5086–91 (1994).