Classification and Regression Analysis of Lung Tumors from Multi-level Gene Expression Data

Pratheeba Jeyananthan
School of Electronics and Computer Science,
University of Southampton,
Southampton, UK
pj2u16@soton.ac.uk

Mahesan Niranjan
School of Electronics and Computer Science,
University of Southampton,
Southampton, UK
mn@ecs.soton.ac.uk

***Abstract*—We study classification and regression problems in lung tumours where high throughput gene expression is measured at multiple levels: epi-genetics, trancription and protein. We uncover the correlates of smoking and gender-specificity in lung tumors. Different genes are indicative of smoking levels, gender and survival rates at these different levels. We also carry out an integrative anaysis, by feature selection from the pool of all three levels of features. Our results show that the epi-genetic information in DNA methylation is a better marker for smoking status than gene expression either at the transcript or protein levels. Further, surprisingly, integrative anlysis using multi-level gene expression offers no significant advantage over the individual levels in the classification and survival prediction problems considered.**

***Index Terms*—Lung cancer, Survival prediction, Smoking mutated genes, Multi-omic, Gender specific genes, TCGA data repository**

I. Introduction

Cancer is a complex disease formed by heritable and environmental factors. Understanding of it causes and dynamics at the molecular level is crucial for early diagnostics and treatment planning because it is known that drug response, for example, widely varies across sub-populations of cancer sufferers. Hence molecular level stratification of patients is vital for knowing which drug will have good response from which group of patients. Advances in high throughput functional genomics, in which gene sequences, their epi-genetic modifications, expressions, and relative abundance of their products (i.e. proteins) are measured at a genomic level (i.e. throughout the genome), coupled with machine learning methods, is seen as an appealing avenue to pursue in the development of a better understanding of the disease and its effective treatment. Among the various cancers, lung cancer is a highly prevalent one with a specific environmental factor – smoking – being its main cause. Over 85% of incidence of lung cancer relate to smoking, and is the subject of this study.

High throughput measurement of gene expression at various levels (gene, epi-genetic, protein etc.) produce representations of a tumor sample in very high dimensions (methylation at $27,000$ sites, $20,000$ transcripts, $7,000$ proteins etc.). Two important issues arise when we attempt to apply statistical inference or machine learning methods on such data. Firstly, the number of patients on whom such measurements can be made is often much smaller, of the order of a few tens or hundreds. This leads to the problem often known a $n << p$ problem (*i.e.* number of samples far fewer than dimensionality). In this setting, any inference method, those that include density estimation in particular, suffer the effects of the *curse of dimensionality*. Techniques such as feature selection, feature reduction by subspace projections or regularisation have to be carefully applied to deal with this issue. Secondly, information contained in measurements taken at different levels of gene expression is often not the same, due to various types of regulatory mechanisms actin in cells. For example, genes that are transcribed need not all be translated into protein. Hence we will observe their expression at the level of the transcriptome, though they could have very different cellular function at the protein level. Thus, integrative analysis of measurements taken at different levels is of importance.

Several transcriptome-based studies have been reported in the literature, including the irreversible effect of tobacco [1], identifying differentially expressed genes between smokers and non-smokers [6], [7], [22] and differentiating smokers from non-smokers [20] or current smokers from others [24]. Methylation data has been used in studies to show how smoking could be identified using single [3], [5] or multiple methylation sites [35], and to infer the effects of maternal smoking on new-borns [27]. Methylation data was analysed between former and current smokers to see how methylation changes by smoking [36] and its residuals has also been suggested as a good marker of smoking in the past [30].

Relating lung cancer with smoking shows that smoking related methylations identified by [3], [5] have most association with lung cancer [14], [38]. This property of methylation was identified in transcriptome data as well [31].

Another area that has attracted attention is prediction of survival from high throughput genomic data. A wide range of algorithms including Bayesian ensemble method [4], PCA [10] and Wavelet based gene selection [13] for gene selection from trnscriptome data. Similarly survival prediction has been attempted from trancriptome and proteome data [8], [17], [25], [29], [32], [37].Interestingly, [33] report a study that links RNA degradation to survival in non-small cell lung cancer patients.

Gender specificity of gene expression in response to smok-

ing has also attracted attention in the literature [18], [26], [34].

In this paper, we have a comprehensive analysis on multiple levels of genes to find the connection between smoking and molecular data. up-regulated and down-regulated molecular data were analyzed between current smokers, lifelong non-smokers, reformed smokers ≤ 15 years and reformed smokers > 5 years and relationship between the molecular data and these four various smoking status were discussed. At the same time all those molecular levels were used to classify the patients into their corresponding smoking status.
Further, various gene levels of lung cancer patients were used to predict their individual survival rate. Performances by individual molecular data were compared to determine the best molecular data for survival prediction of lung cancer patients. Moreover, performance of the integrated molecular data also analyzed in the same problem.

II. Materials and methods

Transcriptome, proteome and methylation data of lung cancer patients were downloaded from The Cancer Genome Atlas (TCGA) data repository along with their survival rate, smoking status and patient gender.
For smoking related studies, we had data from 250 patients with transcriptome, 163 patients with proteome and 62 patients with methylation measurements. For survival prediction, the dataset consisted of 998 patients with transcriptome, 192 patients with proteome and 197 with methylation data available. For integrative analysis, however, the number of patients on whom all three measurements had been made is much smaller, 67.

A. Methods

Inference problems with high throughput genomic data are posed in very high dimensions, often the number of features far exceeding the number of samples that are available (the so called $n << p$ problem in statistics). Feature selection or other methods of dimensionality reduction (*e.g.* Principal Component Analysis) are required to address this. Though PCA type projections, forming a combination of features, can be statistically appealing from the point of view of high accuracies in inference, feature selection is usually preferable from the point of view of interpretability. Hence we used two methods of feature selection: ranking by Fisher ratio, which considers the performance of each feature individually, and greedy forward feature selection, which searches for combinations of features.

1) Fisher ratio: Fisher ratio is a measure of how discriminant a feature is, assuming their distributions are Gaussian. Let μ_1 and μ_2 be the means and σ_1 and σ_2 be the standard deviations of those distributions. Then the Fisher ratio is defined as $FR = \frac{(\mu_1-\mu_2)^2}{\sigma_1^2+\sigma_2^2}$.
The features were ranked by the Fisher ratio and the top few were selected to build classifiers. This has been a popular approach to feature selection in bioinformatics problems, starting from the early work of Golub *et al.* [15].

2) Greedy forward feature selection: Greedy forward selection is referred to as a *wrapper* [28] approach attempting to find a sub-optimal combination of features. It starts with a single feature, the one with the highest discrimination, and progressively includes features searching for the best combination of two, three... etc. feature classifiers. At each stage, the search is done linearly through the yet unselected features, drastically reducing the search required. Several high dimensional problems, including gene expression, whose discriminant subspaces are much smaller in dimension than the original feature space in which the problem is posed have been considered in Li *et al.* [21], justifying the choice of this method on the TCGA data.

3) Support Vector Machines (SVM): Empirically, support vector machines are generally considered suitable choice for high performance pattern classification. This is particularly true for problems that are posed in high dimensions because the focus of optimisation in SVMs is classification accuracy rather than the estimation of density as might be needed in a Bayesian classifier. Density estimation in high dimensions is notoriously difficult due to the *curse of dimensionality* and SVMs, by optimising a class boundary and its margin, offer a good way of circumventing them. The optimization problem solved in SVMs is,

$$\min_{\mathbf{w}\in\mathbb{R}^d,\xi_i\in\mathbb{R}^+} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i$$
$$\text{s.t.} \quad y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1-\xi_i,\ \xi \geq 0,\ i = 1,...n$$

where,
ξ_is are known as slack variables and C is a regularization parameter.

In our simulations, for simplicity, we used the convex optimisation toolbox `CVX`, within the `MATLAB` programming environment to implement SVMs. While there are far more computationally efficient implementations, the problems we consider are small enough that this choice was considered adequate. The weights of the SVM were computed on a training set, the C parameter was tuned using a validations set and the area under the receiver operating characteristic (AUC) was measured on an independent test set for reporting performances. Throughout, for classification, we used a linear SVM. When working in high dimensions, nonlinear kernels did not appear to yield significant improvements.

4) Linear regression: We used linear regression to model survival rates, which are continuous values in the range 0 to 1. Nonlinear models such as neural networks, though likely to produce better prediction accuracies, would require significantly more data than we have in TCGA, especially when we look at combinations of features. Let the input data matrix X consist of gene expression features as rows, and the outputs be contained in the vector $\boldsymbol{y}$, the linear

mapping being $y = \boldsymbol{w}^t \boldsymbol{x}$. It is assumed a bias term is included in the weight vector and the corresponding column of the matrix X is filled with ones. The weight vector $\boldsymbol{w}$ is then calculated as the pseudoinverse $w = (X^t X)^{-1} X^t \boldsymbol{y}$, implemented in MATLAB by the command `w = X \ y` which gives numerical stability. With regression, too, we used ten-fold cross validation to assess the uncertainty in inference.

5) Performance Metrics: We used area under the receiver operating characteristics curve (AUC) to quote classification results. This measure is far better than classification accuracies when the two classes are unbalanced and the cost of misclassification is unspecified [28]. For the regression problem of survival rate prediction, we used mean squared error (MSE), defined as: $\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$, where y_i and $\hat{y}_i$ are the i^{th} target and its prediction respectively.

B. Clinical data

Smoking status of the patients

The smoking status of each patient is provided in TCGA as one of four groups. These are (a) current smoker; (b) reformed smoker for $>$ 15 years; (c) reformed smoker for $\leq$ 15 years; and (d) lifelong non-smoker.

Survival rate or Survival probability

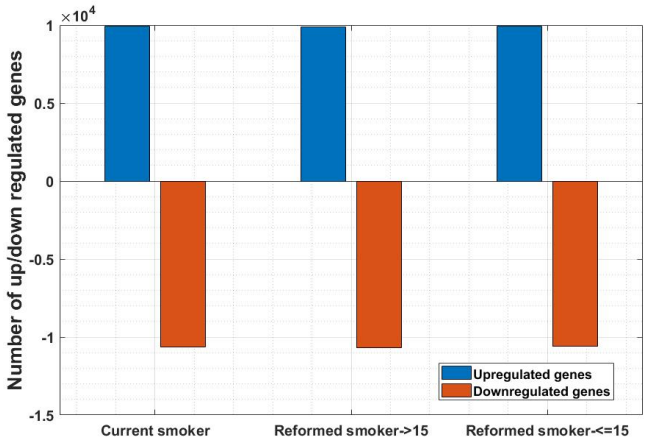
Survival rates, a continuous value between 0 and 1, are also provided by the TCGA repository. This is calculated based on the mortality of the patients (https://docs.gdc.cancer.gov/Data_Portal/PDF/Data_Portal_UG.pdf), based on their day to death or last follow up. The Kaplan-Meier estimator used to estimate the survival rate is given by $S(t_i) = \Pi_{t_i \leq t}\left(1 - \frac{d_i}{n_i}\right)$, where,

- $S(t_i)$ is the estimated survival probability or survival rate for the t time periods
- n_i is the number of subjects at risk at the beginning of time period t_i
- d_i is the number of subjects who die during time period t_i

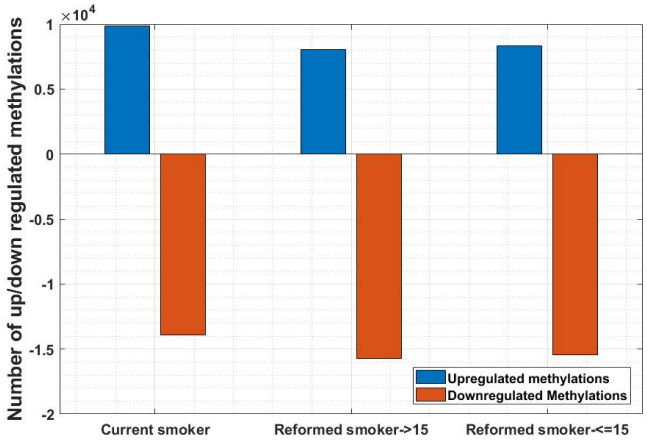
III. Results

A. DNA Methylation is significantly downreegulated in smokers

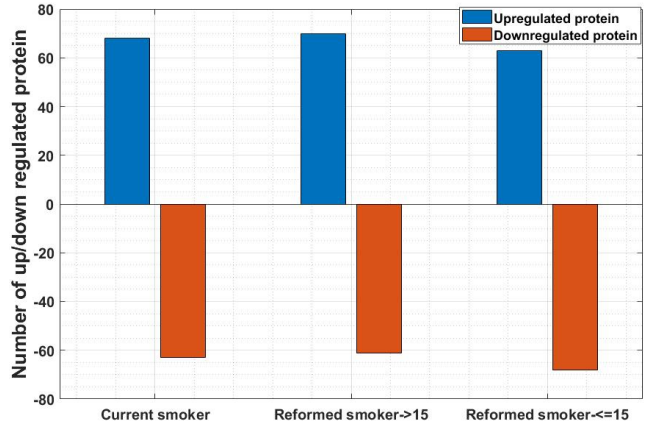
To assess the primary effect of smoking in gene expression across all the measurements, we looked at differential expression (up- or down-regulation) of genes in the three smoking groups with respect to the non-smokers. Figure I shows the numbers of up- or down-regulated genes at the transcript, DNA methylation and protein levels. We note that the primary observation one can make is that there is significantly higher amount of suppression of DNA methylation while the numbers of elevated and suppressed genes at the other levels is roughly the same. While a greater number of proteins show elevated expression than that show suppressed expression, the coverage of the proteome is not large enough to conclude this might be a general observation.



(A)



(B)



(C)

Fig. 1. Smoking effect on gene expression. A large number of genes are up- or down-regulated in the tumors of smokers in comparison to those in non-smokers. Very little difference exists between those who have given up smoking either in the recent past ($\leq$ 15 years) or significant time ago ($>$15 years). We note that the numbers of up- and down- regulated genes in transcript and protein levels is not significantly different. DNA methylation, however, significantly more suppression than elevation in the smoking population.

B. Gene expression separates tumors of lifelong non-smokers from current smokrs

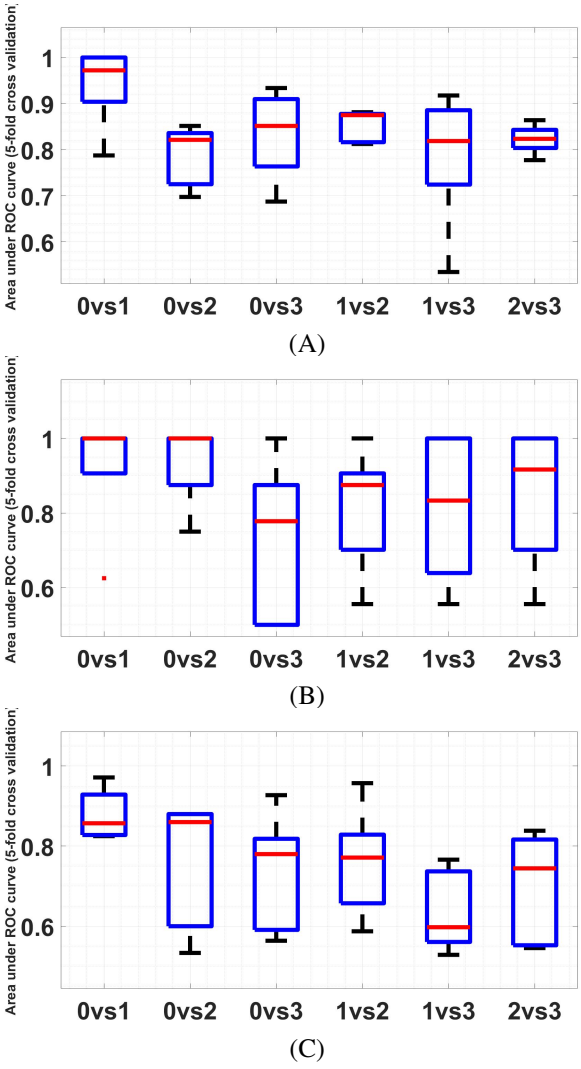


Fig. 2. Pairwise classification accuracies (quantified as area under the receiver operating characteristic curve) of the four different groups of smoking status among lung cancer patients. Top ten features selected by Fisher ratio were used to classify in each level of gene expression: A. transcriptome, B. DNA methylation and C. proteome. Labels: 0-Lifelong non-smoker, 1-Current smoker, 2-Current reformed smoker for $>$ 15 years, 3- Current reformed smoker $\leq$ 15 years

Pairwise classification accuracies of tumors from patients with different smoking status are shown in Figure 2. These classifications carried out with the top ten features in each level of gene expression, filtered based on Fisher ratios. Among the different combinations, separating life-long non-smokers from those currently smoking yields near-perfect classification. Across the three levels of gene expression, methylation data gives the highest accuracy than the other two.

Further, for most of the pairs, methylation data gave median accuracy $\geq$ 0.89. For lifelong non-smokers versus reformed smoker for $\leq$ 15 years (0.79) and current smoker versus reformed smoker for $\leq$ 15 years (0.82), methylation data gave comparably low median accuracy than transcriptome. Transcriptome data gave comparably low accuracy than methylation in all the other pairs and their median accuracies range between 0.81 and 0.88.

Accuracy of current smoker versus reformed smoker for $\leq$ 15 years is comparably low in all three molecular levels, which shows that the molecular expression of current smoker is nearly same to reformed smoker for $\leq$ 15 years.

C. Gender Specificity of Gene Expression

For each smoking status, ten genes with high difference between male and female were discussed in this section. TTTY15, CYorf15B, RPS4Y1, USP9Y, UTY, DDX3Y, CYorf15A, ZYF, KDM5D and E1F1AY were differently expressed between male and female of non-smokers. Most of the genes are same for current smokers, reformed smoker for $\leq$ 15 years and reformed smoker for $>$ 15 years, except CYorf15B of non-smokers was replaced by XIST for current smokers and reformed smoker for $>$ 15 years, and by NLGN4Y for reformed smoker for $\leq$ 15 years, Table I.

We compared our results on high throughput data with microarray data and found Y-chromosome related genes were over expressed in male and X chromosome related gene XIST was over expressed in female. Here, we found some new genes related to Y-chromosome which are TTTY15, CYorf15B, CYorf15A, ZYF, NLGN4Y and KDM5D.

D. Smoking mutated genes are gender specific

It has been shown that smoking mutated genes are gender specific [26]. They concluded that female smokers are more vulnerable to cigarette smoking induced diseases than male smokers. They identified 175 differently expressed genes between male smokers and non-smokers, 237 differently expressed genes between female smokers and non-smokers. Four up-regulated genes (RGS6, ELL3, TBXA2R and GRM5) and two down-regulated genes (RAB6B and GPR15) were differently expressed between female smokers and male smokers.

In this study, we analyzed gender specific smoking mutated genes between each pair of smoking status. For each pair, top ten differently expressed genes were identified in this study. Table II illustrates that almost every smoking mutated genes are gender specific, except few of them.

E. Selected features on survival prediction shows their connection with tumor formation

We have three distinct molecular data in this study. Each of these data is separately used in feature selection and different number of features were selected from each molecular data. 506 transcriptome, 35 proteins and 86 methylation were selected from 20530 transcriptome, 191 proteins and 23924 methylation correspondingly.

Transcriptome features: GO analysis of selected set of genes [11] & [12] shows that they are related to DNA damage related activities. These are the functions of selected features of gene data:

TABLE I

TOP TEN GENES WITH DIFFERENTIAL GENDER SPECIFIC EXPRESSION IN EACH OF THE FOUR GROUPS CONSIDERED: (A) LIFELONG NON-SMOKERS, (B) CURRENT SMOKERS, (C) REFORMED SMOKER FOR $>$ 15 YEARS AND (D) REFORMED SMOKER FOR $\leq$ 15 YEARS.

Lifelong non-smokers	TTTY15	CYorf15B	RPS4Y1	USP9Y	UTY	DDX3Y	CYorf15A	ZFY	KDM5D	EIF1AY
Current smokers	XIST	ZFY	CYorf15A	TTTY15	USP9Y	EIF1AY	RPS4Y1	UTY	KDM5D	DDX3Y
Reformed smokers $>$ 15 years	TTTY15	XIST	CYorf15A	USP9Y	EIF1AY	UTY	ZFY	KDM5D	RPS4Y1	DDX3Y
Reformed smokers $\leq$ 15 years	CYorf15A	EIF1AY	NLGN4Y	TTTY15	UTY	RPS4Y1	ZFY	KDM5D	USP9Y	DDX3Y

TABLE II

TOP TEN GENES WHICH, IN A GENDER SPECIFIC WAY, DISCRIMINATE BETWEEN THE DIFFERENT STATUS OF SMOKING IN LUNG CANCER. PAIRWISE CLASSIFICATIONS OF THE DIFFERENT GROUPS WERE CARRIED OUT SEPARATELY FOR MALE AND FEMALE PATIENTS. LABELS: 1-CURRENT SMOKERS, 2-REFORMED SMOKER FOR $>$ 15 YEARS AND 3-REFORMED SMOKER FOR $\leq$ 15 YEARS

0 VS 1	**Male**	TSPAN2	TXNIP	BMX	PACSIN3	C12orf65	CFH	NPR3	ITIH3	CACNA2D3	MRPS5
	Female	ZNF564	SLC40A1	RBM17	EHMT2	SERGEF	C12orf69	CNNM1	GTPBP4	LUZP2	GPR15
0 VS 2	**Male**	NPR3	C5orf23	ADAMTS6	KIAA0776	DACT1	C3orf30	CSDC2	CNTNAP3	ACPP	ELP2P
	Female	PPP1R3G	CNNM1	C14orf181	CTSD	CCDC144B	LGALS9	LOC653786	RAB42	TMEM86A	TREM2
0 VS 3	**Male**	RNGTT	AVIL	CNTNAP3	PAAF1	HDAC7	KHK	LOC100130522	PDSS2	CACNA2D3	CDS2
	Female	PITPNC1	ENPP3	PRDM1	MPI	LHX9	CNNM1	ZNF702P	OLIG1	SERGEF	GPR15
1 VS 2	**Male**	MTPAP	PSMC3IP	HYLS1	CD82	CST5	ALDH1A1	SESN1	CLEC9A	GPR15	TNFSF13
	Female	JSRP1	RRAS2	SND1	C12orf69	AHRR	FAM65C	LUZP2	TMEM56	CRYGN	GPR15
1 VS 3	**Male**	KIAA0895L	CCR9	FAM102B	LIPA	KIAA1409	SNX2	BREA2	IQGAP2	CLEC9A	GM2A
	Female	FAM5C	PLCXD3	CCDC138	PSPH	BAT3	GTF2H4	TMEM56	NRG2	NEGR1	SKIV2L
2 VS 3	**Male**	ACOX3	RAPGEF3	CASR	ARID3C	MRPL19	PRICKLE4	PLAA	IMMT	IL17RC	GPR15
	Female	ZNF681	ZSCAN18	HMGN3	LEP	LIN7C	HNRPDL	LOC728855	ZNF675	AHRR	GPR15

- regulation of intrinsic apoptotic signaling pathway in response to DNA damage
- steroid biosynthetic process
- regulation of response to DNA damage stimulus
- trans membrane receptor protein tyrosine kinase signaling pathway
- enzyme linked receptor protein signaling pathway

Since DNA damage has been proved to be the root cause of cancer, the selection of these features from transcriptome of lung cancer patients is reasonable.

Proteome & Methylation features:

35 features were selected out of 191 total protein data and 86 methylation were selected out of 23924 methylation and used separately in the survival prediction. Go annotation of these genes reports that they are related to cell cycle regulation activities. Go terms related to these genes are:

- cell cycle arrest
- negative/positive regulation of cell cycle
- regulation of gene expression
- cell death
- cellular response to hypoxia
- negative regulation of cell differentiation
- regulation of organ growth
- cell proliferation or regulation of cell proliferation
- cell aging
- regulation of cell death or regulation of programmed cell death
- cell development

Cell proliferation and cell death have close relationship with carcinogens. The above mentioned GO terms justify the selected features. Comparing the selected features on various molecular levels shows that there are no common features between transcriptome and proteome or transcriptome and methylation while we have a single common feature PEA15 between transcriptome and proteome.

F. Integration of multi-omic data might not increase the accuracy of survival prediction

Figure 3 illustrates that almost all three of those molecular data predict the individual survival rate of lung cancer patients with high accuracy. Variations of training error between molecular data are very low. Comparing test error shows that transcriptome data gave highest accuracy among them.

Generally, transcriptome data was used in the survival prediction. Using the expression of a single gene in survival prediction is a technique used in a few studies. Proteome data and methylation data is rarely used in survival prediction. [32] revealed that protein data can significantly differentiate high survival risk and low survival risk of glioma patients. Here, we showed that transcriptome data might predict the individual survival better than other two molecular data on lung cancer patients, with minor differences in accuracy of others.

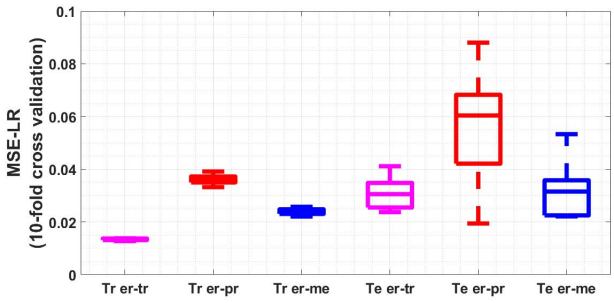


Fig. 3. Training and test errors of a linear regression model on predicting survival from the three levels of gene expression: transcriptome, proteome and DNA methylation. Variation across ten-fold cross validation are shown as box plots. 300 transcript, 20 methylation and 35 protein features were selected by the greedy forward feature selection algorithm.

All three gene level data were available along with survival data only for 68 patients. These 68 patients were used to check the performance of integrated molecular data on survival prediction of lung cancer patients. Here greedy forward feature selection algorithm selected 32 transcriptome features, 22 proteome features, 35 methylation features and 37 features from integrated data. Figure 4 compares the performance by single level with the performance by integration.

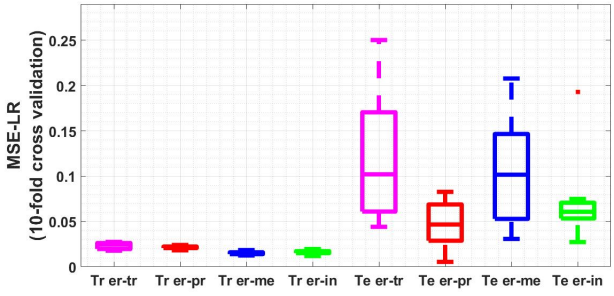


Fig. 4. Comparing feature selection on individual levels of expression with integrated analysis (right-most columns for training and testing). The first three columns differ slightly from those in Fig. 3 because the number of samples for which all three levels of expression are available is different and the models were re-estimated on the reduced set. With the reduced dataset 32 transcriptomic, 22 proteomic and 35 methylation features were selected. The integrative analysis (greedy search for combination of features using all three feature sets) selects 37 features (23 methylation and 14 transcriptomic with none of the protein expression levels contributing to the combination).

Comparing accuracies shows that, training error of methylation data and integrated data have no significance difference between them and higher than other two training error. And, there is no significance difference between the test error of proteome data and integrated data. As we have few number of patients in this integrated dataset, this result might need further analysis. [19] already exposed that integration of multi-omic data may improve the accuracy of prediction of their clinical variables. Here, we suggest that integrating the omic-data might not increase or decrease the accuracy of survival prediction of lung cancer patients.

IV. Discussion

Features with high differentiation power were identified between each pair of smoking status. Analyzing the selected features between pairs shows that there are some common features between pairs. For instance, gene GPR15 is identified in four different pairs, except lifelong non-smokers versus reformed smokers $>$ 15 years and current smokers versus reformed smokers $\leq$ 15 years. We compared our features with other studies and [2] identified some transcriptomes to differentiate people with different smoking habits and their features on microarray data of whole blood with our features ends up with empty set. PLA2G1B of [22] is common with our features where they used microarray data of lung adenocarcinoma patients.

Methylation 450K and 27K data were used in smoking related studies. Methylation cg05575921 of gene AHRR and methylation cg03636183 of gene F2RL3 were widely discussed in smoking related studies [3], [5], [14] and [38]. However, our study does not find these features as a super methylation to differentiate various smoking status. [3] used methylation 450K data in their study and [5] used blood of general population to get the methylation 27K data and their criteria for selecting the subjects to experiment is too stringent. Despite, in our study, we used methylation 27K data of lung cancer patients from the TCGA without any conditions. This might be the reason, why our study could not find those markers. And it reveals that these features are dependent on the subject's health condition and the technique we used to measure the data.

Besides, [6] and [22] studied on up-regulated and down-regulated transcriptome data between smokers and non-smokers. [6] identified 88 up-regulated and 106 down-regulated genes between active smokers and non-smokers from transcriptome of buccal mucosa tissue. At the same time, [22] identified 83 up-regulated and 213 down-regulated genes between smokers and non-smokers who are lung adenocarcinoma patients.

Further, selected features from transcriptome data on survival prediction was compared with published features of [33]. They identified 997 genes on survival prediction of lung cancer patients. They used microarray transcriptome of non-small cell lung cancer patients and analyzed the variance to select the features. They mentioned few selected features where NLRC4 and MAGEA3 were common with our transcriptome features.

Moreover, [34] and [18] studied on gender specific genes. Microarray expression of brain was used in [34] and identified DBY, SMCY, UTY, RPS4Y1 and USP9Y (genes from Y chromosome) and XIST (gene from X chromosome) are differently expressed between male and female. [18] studied on microarray transcriptome data of idiopathic dilated cardiomyopathy (IDCM) patients. 35 genes were over expressed and 16 genes were under expressed between male and female. Genes related to Y-chromosome such as USP9Y, DDX3Y, RPS4Y1 and E1F1AY were over expressed in male. X-chromosome related genes such as XIST was over expressed in female. Our study on high throughput sequencing data of lung cancer

patients has some common genes such as UTY, RPS4Y1, USP9Y, DDX3Y, RPS4Y1, E1F1AY and XIST and some new biomarkers such as TTTY15, CYorf15A, CYorf15B, ZFY, KDM5D and NLGN4Y.

V. Conclusion

In this work, we have carried out classification and regression analysis on multi-level gene expression data on samples from a highly prevalent cancer in which environmental influences are significant. We have extracted genes that are relevant for predicting smoking status, gender specificity and survival at all three levels of epi-genetic, transcript and protein abundances. On the raw data, we note that the predominant effect of smoking is the relative suppression of DNA methylation, not seen in the other two levels. We also see that those currently smoking ar easily differentiated from other subgroups. Somewhat surprisingly, integrated analysis, in which we seek to identify combinations of features across all three levels of gene expression did not give significant improvement over taking the features individually. This could be because the size of the dataset for integrative analysis (the patients in which all three measurements were made) was substantially smaller than those with any one measurement. Referring the properties of genes selected using gene ontology analysis confirms some of the observations we make.

Acknowledgment

Pratheeba Jeyananthan is funded by a scholarship from the Institute for Life Sciences (IfLS) University of Southampton. Mahesan Niranjan acknowledges support from the project Joining the Dots: From Data to Inference (EP/N014189/1), funded by the Engineering and Physical Sciences Research Council, UK.

References

[1] Beane, J., Sebastiani, P., Liu, G., Brody, J., Lenburg, M. and Spira, A. (2007). Reversible and permanent effects of tobacco smoke exposure on airway epithelial gene expression. *Genome Biology*, **8(9)**, p.R201.

[2] Beineke, P., Fitch, K., Tao, H., Elashoff, M., Rosenberg, S., Kraus, W. and Wingrove, J. (2012). A whole blood gene expression-based signature for smoking status. *BMC Medical Genomics*, **5(1)**.

[3] Bojesen, S., Timpson, N., Relton, C., Davey Smith, G. and Nordestgaard, B. (2017). AHRR (cg05575921) hypomethylation marks smoking behaviour, morbidity and mortality. *Thorax*, **72(7)**, 646-653.

[4] Bonato, V., Baladandayuthapani, V., Broom, B., Sulman, E., Aldape, K. and Do, K. (2010). Bayesian ensemble methods for survival prediction in gene expression data., *Bioinformatics*, **27(3)**, 359-367.

[5] Breitling, L., Yang, R., Korn, B., Burwinkel, B. and Brenner, H. (2011). Tobacco-Smoking-Related Differential DNA Methylation: 27K Discovery and Replication. *The American Journal of Human Genetics*, **88(4)**, 450-457.

[6] Cao, C., Chen, J., Lyu, C., Yu, J., Zhao, W., Wang, Y. and Zou, D. (2016). Correction: Bioinformatics Analysis of the Effects of Tobacco Smoke on Gene Expression. *PLOS ONE*, **11(3)**, p.e0150778.

[7] Charlesworth, J., Curran, J., Johnson, M., Gring, H., Dyer, T., Diego, V., Kent, J., Mahaney, M., Almasy, L., MacCluer, J., Moses, E. and Blangero, J. (2010). Transcriptomic epidemiology of smoking: the effect of smoking on gene expression in lymphocytes. *BMC Medical Genomics*, **3(1)**.

[8] Chaudhary, K., Poirion, O., Lu, L. and Garmire, L. (2017). Deep LearningBased Multi-Omics Integration Robustly Predicts Survival in Liver Cancer,. *Cancer Research*

[9] Chauhan, S., Kaur, J., Kumar, M., Matta, A., Srivastava, G., Alyass, A., Assi, J., Leong, I., MacMillan, C., Witterick, I., Colgan, T., Shukla, N., Thakar, A., Sharma, M., Siu, K., Walfish, P. and Ralhan, R. (2015). Prediction of recurrence-free survival using a protein expression-based risk classifier for head and neck cancer. *Oncogenesis*, **4(4)**, e147-e147.

[10] Chen, X. and Wang, L. (2009). Integrating Biological Knowledge with Gene Expression Profiles for Survival Prediction of Cancer.,*Journal of Computational Biology*, **16(2)**, 265-278.

[11] Eden, E., Navon, R., Steinfeld, I., Lipson, D. and Yakhini, Z. (2009). GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists.*BMC Bioinformatics*, **10(1)**, 48.

[12] Eden, E., Lipson, D., Yogev, S. and Yakhini, Z. (2007). Discovering Motifs in Ranked Lists of DNA Sequences. *PLoS Computational Biology*, **3(3)**, e39.

[13] Farhadian, M., Mahjub, H., Moghimbeigi, A., Lisboa, P., Poorolajal, J. and Mansoorizadeh, M. (2015). Wavelet-based gene selection method for survival prediction in diffuse large B-cell lymphomas patients.*International Journal of Data Mining and Bioinformatics*, **13(2)**, 197.

[14] Fasanelli, F., Baglietto, L., Ponzi, E., Guida, F., Campanella, G., Johansson, M., Grankvist, K., Johansson, M., Assumma, M., Naccarati, A., Chadeau-Hyam, M., Ala, U., Faltus, C., Kaaks, R., Risch, A., De Stavola, B., Hodge, A., Giles, G., Southey, M., Relton, C., Haycock, P., Lund, E., Polidoro, S., Sandanger, T., Severi, G. and Vineis, P. (2015). Hypomethylation of smoking-related genes is associated with future lung cancer in four prospective cohorts. *Nature Communications*, **6(1)**.

[15] Golub, T. (1999). Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science*, **286(5439)**, 531-537.

[16] Grant, M. and Boyd, S. (2013). CVX: Matlab software for disciplined convex programming, version 2.0 beta. http://cvxr.com/cvx.

[17] Han, G., Zhao, W., Song, X., Kwok-Shing Ng, P., Karam, J., Jonasch, E., Mills, G., Zhao, Z., Ding, Z. and Jia, P. (2017). Unique protein expression signatures of survival time in kidney renal clear cell carcinoma through a pan-cancer screening. *BMC Genomics*, **18(S6)**.

[18] Heidecker, B., Lamirault, G., Kasper, E., Wittstein, I., Champion, H., Breton, E., Russell, S., Hall, J., Kittleson, M., Baughman, K. and Hare, J. (2009). The gene expression profile of patients with new-onset heart failure reveals important gender-specific differences. *European Heart Journal*, **31(10)**, 1188-1196.

[19] Huang, S., Chaudhary, K. and Garmire, L. (2017). More Is Better: Recent Progress in Multi-Omics Data Integration Methods. *Frontiers in Genetics*, **8**.

[20] Landi, M., Dracheva, T., Rotunno, M., Figueroa, J., Liu, H., Dasgupta, A., Mann, F., Fukuoka, J., Hames, M., Bergen, A., Murphy, S., Yang, P., Pesatori, A., Consonni, D., Bertazzi, P., Wacholder, S., Shih, J., Caporaso, N. and Jen, J. (2008). Gene Expression Signature of Cigarette Smoking and Its Role in Lung Adenocarcinoma Development and Survival. *PLoS ONE*, **3(2)**, p.e1651.

[21] Li, H. and Niranjan, M. (2007). Discriminant subspaces of some high dimensional pattern classification problems. In: *Machine Learning for Signal Processing*,. 27-32.

[22] Liu, Y., Ni, R., Zhang, H., Miao, L., Wang, J., Jia, W. and Wang, Y. (2016). Identification of feature genes for smoking-related lung adenocarcinoma based on gene expression profile data. *OncoTargets and Therapy*, **Volume 9**, 7397-7407.

[23] Lynch, C., Abdollahi, B., Fuqua, J., de Carlo, A., Bartholomai, J., Balgemann, R., van Berkel, V. and Frieboes, H. (2017). Prediction of lung cancer patient survival via supervised machine learning classification techniques. *International Journal of Medical Informatics*, **108**, 1-8.

[24] Martin, F., Talikka, M., Hoeng, J. and Peitsch, M. (2015). Identification of gene expression signature for cigarette smoke exposure responsefrom man to mouse. *Human & Experimental Toxicology*, **34(12)**, 1200-1211.

[25] Oberthuer, A., Kaderali, L., Kahlert, Y., Hero, B., Westermann, F., Berthold, F., Brors, B., Eils, R. and Fischer, M. (2008). Subclassification and Individual Survival Time Prediction from Gene Expression Data of Neuroblastoma Patients by Using CASPAR. *Clinical Cancer Research*, **14(20)**, 6590-6601.

[26] Paul, S. (2014). Differential Effect of Active Smoking on Gene Expression in Male and Female Smokers. *Journal of Carcinogenesis & Mutagenesis*, **05(06)**.

[27] Reese, S., Zhao, S., Wu, M., Joubert, B., Parr, C., Hberg, S., Ueland, P., Nilsen, R., Midttun, ., Vollset, S., Peddada, S., Nystad, W. and London, S. (2016). DNA Methylation Score as a Biomarker in Newborns for

Sustained Maternal Smoking during Pregnancy. *Environmental Health Perspectives*, **125(4)**.

[28] Scott, M.J.J. and Niranjan, M. and Melvin, D.G. and Prager, R.W. (1998). Maximum realisable performance: a principled method for enhancing performance by using multiple classifiers. *Proceedings of the British Machine Vision Conference*, Southampton, England.

[29] Seok, J., Davis, R. and Xiao, W. (2015). A Hybrid Approach of Gene Sets and Single Genes for the Prediction of Survival Risks with Gene Expression Data. PLOS ONE, 10(5), p.e0122103.

[30] Shenker, N., Ueland, P., Polidoro, S., van Veldhoven, K., Ricceri, F., Brown, R., Flanagan, J. and Vineis, P. (2013). DNA Methylation as a Long-term Biomarker of Exposure to Tobacco Smoke. *Epidemiology*, **24(5)**, 712-716.

[31] Spira, A., Beane, J., Shah, V., Steiling, K., Liu, G., Schembri, F., Gilman, S., Dumas, Y., Calner, P., Sebastiani, P., Sridhar, S., Beamis, J., Lamb, C., Anderson, T., Gerry, N., Keane, J., Lenburg, M. and Brody, J. (2007). Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer. *Nature Medicine*, **13(3)**, 361-366.

[32] Stetson, L., Dazard, J. and Barnholtz-Sloan, J. (2016). Protein Markers Predict Survival in Glioma Patients. *Molecular & Cellular Proteomics*, **15(7)**, 2356-2365.

[33] Valk, K., Vooder, T., Kolde, R., Reintam, M., Petzold, C., Vilo, J. and Metspalu, A. (2010). Gene Expression Profiles of Non-Small Cell Lung Cancer: Survival Prediction and New Biomarkers. *Oncology*, **79(3-4)**, 283-292.

[34] Vawter, M., Evans, S., Choudary, P., Tomita, H., Meador-Woodruff, J., Molnar, M., Li, J., Lopez, J., Myers, R., Cox, D., Watson, S., Akil, H., Jones, E. and Bunney, W. (2003). Gender-Specific Gene Expression in Post-Mortem Human Brain: Localization to Sex Chromosomes. *Neuropsychopharmacology*, **29(2)**, 373-384.

[35] Wan, E., Qiu, W., Carey, V., Morrow, J., Bacherman, H., Foreman, M., Hokanson, J., Bowler, R., Crapo, J. and DeMeo, D. (2015). Smoking-Associated Site-Specific Differential Methylation in Buccal Mucosa in the COPDGene Study. *American Journal of Respiratory Cell and Molecular Biology*, **53(2)**, 246-254.

[36] Wilson, R., Wahl, S., Pfeiffer, L., Ward-Caviness, C., Kunze, S., Kretschmer, A., Reischl, E., Peters, A., Gieger, C. and Waldenberger, M. (2017). The dynamics of smoking-related disturbed methylation: a two time-point study of methylation change in smokers, non-smokers and former smokers. *BMC Genomics*, **18(1)**.

[37] Yasrebi, H., Sperisen, P., Praz, V. and Bucher, P. (2009). Can Survival Prediction Be Improved By Merging Gene Expression Data Sets?. *PLoS ONE*, **4(10)**, e7431.

[38] Zhang, Y., Elgizouli, M., Schttker, B., Holleczek, B., Nieters, A. and Brenner, H. (2016). Smoking-associated DNA methylation markers predict lung cancer incidence. *Clinical Epigenetics*, **8(1)**.