Cancer Death Clock: Multi-omics Signatures for Lung Adenocarcinoma Survival Time Prediction

Hayan Lee[1], Gilbert Feng[2], Aaron Horning[1] and Michael Snyder[1a],

[1] *Department of Genetics, Stanford University. Stanford, CA, USA*

[2] *Electrical Engineering and Computer Sciences, University of California, Berkeley, Berkeley, CA, USA*

E-mail: hayan.lee@stanford.edu, gilbertfeng@berkeley.edu, ahorning@stanford.edu, mpsnyder@stanford.edu

Lung cancer is the most prevalent cancer in most countries worldwide. Lung adenocarcinoma (LUAD) solely accounts for approximately 40% of all cases. Although there has been a dramatic therapeutic improvement, the prognostic prediction has relied on mostly clinical features such as tumor-nodal-metastasis (TNM) stage, age upon diagnosis, and smoking history for decades. However, it is inaccurate and does not reflect molecular alterations on its pathway and heterogeneity of tumorigenesis. Here we propose an integrative random forest model to predict survival time exploiting multi-omics data. We identified multi-omics signatures with higher importance to better predict survival time than clinical annotations, traditionally used by physicians. We confirmed that the integrative prediction model outperforms any other model with a single type of omics data, while we found that methylation performed best among any single type omics-based model. Since methylation provides the most number of marker candidates, it has the highest chance to have one of the best predictors. We also performed cost analysis, finding that methylation is more costly than other types of omics technologies in general. Paradoxically, methylation is the most economical platform when the cost per marker is compared among methylome, transcriptome, and proteome technology since it provides the most abundant signature candidates.

Keywords: LUAD, Random forest regression, survival time prediction, nonlinear, non-parametric, marker cost analysis

1. Introduction

1.1. *Background of Lung Adenocarcinoma*

Lung cancer is the most prevalent cancer in most countries worldwide and has two subtypes: small cell lung cancer and non-small cell lung cancer, the latter of which comprises about 80-85% of lung cancers. Lung adenocarcinoma (LUAD) is one of the major subtypes of non-small cell lung cancer, along with lung squamous cell carcinoma (LUSC). LUAD accounts for approximately 40% of all lung cancer cases.

Prediction of its prognostic trajectory is critical, especially to each patient. Traditionally prognostic trajectory has been estimated by clinical data, such as ages upon diagnosis, stage, and

[a] Corresponding author

smoke history. This method is inaccurate because it does not consider molecular characteristics in its pathway. Recently, there has been a dramatic improvement on treatment by molecule-targeting therapies, but its outcome prediction is vague because it still relies on solely clinical annotation and does not take molecular responses into account. Thus, we propose an ensemble learning method on high dimensional omics data with clinical annotation to predict survival time. We learned a Random Forest regression model exploiting The Cancer Genome Atlas (TCGA)[1] LUAD omics data and clinical annotations.

1.2. *Random Forest Regression*

Days_to_death is very complex phenotype that involves many nonlinear factors including not only clinical factors such as age, gender, smoking history, alcohol intake, but also genetic factors such as genomic information, gene expression, protein abundance, etc. To build a prediction model, we chose Random Forest Regression (RFR) [2, 3], an ensemble of multiple decision trees. It is not required to normalize features for Random Forest as a decision tree does not. Since an integrative model should deal with a variety of data ranges, it is essential to choose a machine learning algorithm that does not require normalization or feature scaling. Random Forest also can deal with nonlinear solution space and a nonparametric model, which does not require any assumptions about the data distribution. Thus, it is ideal for our integrative nonlinear prediction model learning.

1.3. *Related work*

There are several works predicting survival time [4-6]. Long term vs. short term survival classification has been studied more preferably since two group classification is comparatively easier than multi group classification or regression. Yu et al. performed classification of long-term vs. short-term survival of non-small cell lung cancer patients but exploited mainly image data haematoxylin and eosin (H&E) stained histopathology whole-slide images with particular omics markers [7]. Li et al. identified eight genes relating to survival in LUAD using only gene expression data [8]. Yu et al. learned a prediction model to classify short-term (< 3yr) and long-term (> 3yr) survival from LUAD using only somatic mutational features [9]. I-Boost used an integrative prediction model that suggested RNA-seq is more prognostic of survival time than other genomic data types but still failed to include the methylation data that eventually causes gene expression change [10]. In our study we learned an integrated regression model of survival time from TCGA LUAD patients. We exploited methylation data as well as gene expression, protein abundance and clinical annotation.

2. Data

TCGA has generated a variety of omics data along with clinical annotations. TCGA detailed molecular level on various cancer types and collected methylation, gene expression, protein abundance along with genomic data such as copy number variation (CNV), somatic mutation, and microRNA expression. The previous studies show gene expression was the most predictable omics data type among clinical, gene expression, CNV, somatic mutation, microRNA expression, and

protein abundance, but it failed to include methylation data. Thus we integrated methylation, gene expression, and protein data along with clinical annotation to see if methylation data is more predictable than gene expression data.

TCGA LUAD collected six types of omics data form ~500 patients. About 20% of them have survival time annotations. Methylation data is already normalized by its design. We performed log normalization for gene expression data, and proteome data were also z-score normalized. To further reduce model learning time we did feature engineering using Pearson correlation coefficients (PCC) [11] and ended up with data that was 1/1000~ to1/2 size of the original data matrix (Fig.1)

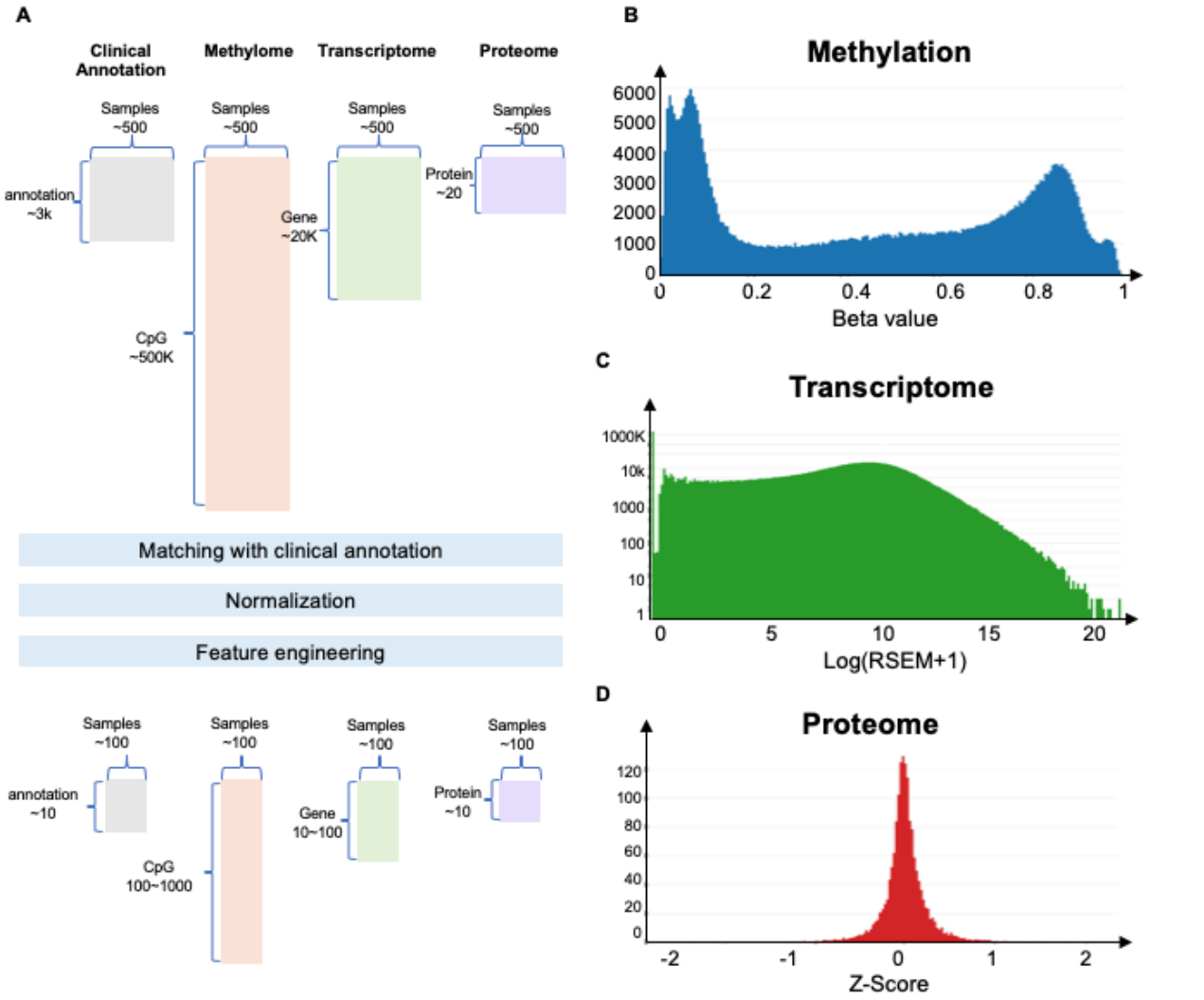

Fig. 1. Methylation, gene expression and protein abundance data of TCGA LUAD are adopted for prediction model learning. For LUAD, TCGA provided ~500 samples with ~500K CpG methylation, ~20K gene expression, ~200 protein abundance, and 100 clinical annotations including days_to_death, the phenotype (A). Distributions of each omics after appropriate normalization respectively are shown in B,C, and D.

3. Single-omics Prediction Model

3.1. *Methylation*

For methylation, we exploited the TCGA methylation data. TCGA adopted Illumina Infinium HumanMethylation450K BeadChip (HM450), where half a million CpGs were assayed to compute beta values, i.e., methylation ratio (which is the number of reads with methylated C divided by the total number of reads). After we extracted LUAD methylation data of the patients with days_to_death, methylation data of around 100 patients were used for prediction model training.

The total number of CpGs was ~485K, including both "cg" and "rf" prefix. Since the methylation level is already represented as a ratio with a range of 0 to 1, the data can be treated as normalized data. The shape of the distribution of methylation data is bimodal (Fig.1A), where CpGs are either highly methylated or unmethylated.

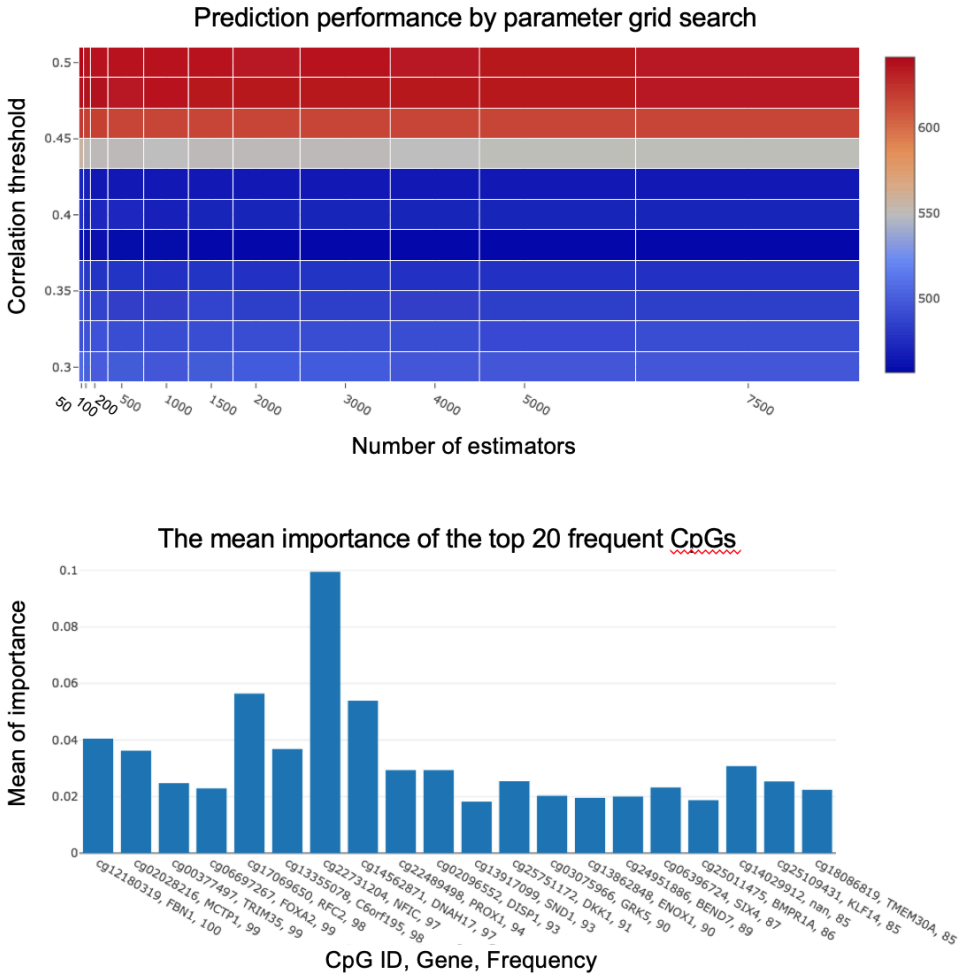

Fig. 2. Methylome-base RFR model prediction performance was grid-explored along with number of estimators and PCC threshold (top). The top 20 CpGs by frequency were represented. The mean importance was computed across 100 runs (bottom).

The beta value data file was processed and put through the Random Forest algorithm. Firstly, the barcodes in the clinical annotation and the barcodes of the methylation file were compared to select only patients who have a days_to_death clinical annotation. After set comparison, ~100 patients were used for model training. For features, CpG loci that started with "cg" were selected and the features with any missing data across the ~100 patients were excluded for model learning.

We employed mean absolute error (MAE) as our cost function. To make model learning efficient and effective, we performed feature engineering by Pearson Correlation Coefficient (PCC). We computed PCC and set various thresholds from 0.3 to 0.5 to select features. PCC allowed us to reduce the number of features down to tens of thousands from half a million. Along with PCC, we also experimented with a varied number of estimators from 50 to 7500, intervaled exponentially. To search the parameter combination space for the best-performed model, we defined a parameter grid and ran Random Forest regression for each cell. Since the algorithm relies on randomization, we ran Random Forest regression 100 times per cell to obtain more robust performance results (Fig. 2. top) The best performance was shown with a PCC threshold of 0.4 and the number of estimators does not seem to significantly effect the performance. Note that selecting features with only high correlation does not always guarantee better prediction, as it may cause an overfitting issue where the learned model fitted too much with the current data set, thus becoming unreliable for the future unknown data. The prediction performance is represented as Mean Absolute Error (MAE), whose values are expressed as colors in the parameter search grid. After running the Random Forest regression 100 times on each setting, a near-optimal setting was found at a PCC threshold of 0.38 with 1500 estimators.

We further studied the CpGs that notably contributed to better prediction performance. The top 20 CpGs were selected by frequency (Fig. 2 (bottom)). Note that we ran 100 times per setting. The genes related to the CpG and the actual frequency are shown with a CpG ID. The average importance value is represented on the Y-axis.

3.2. *Transcriptome*

The gene expression data was also retrieved from TCGA. The data quantified over 20K genes for ~500 LUAD patients. The quantified gene expression levels were computed through RSEM [12,13], which can deal with multiple isoforms fast by parallel computing of the EM algorithm. The raw data was originally skewed with a long tail in the right. After log normalization, it appears more symmetric, with a mean of ~10 (Fig.1C). Note that we added one before taking logs because otherwise some genes are not expressed at all (as log 0 is not defined).

We compared patient barcodes, unique across the TCGA project, of the clinical annotation file with those of raw gene expression file. Then we selected only patients who had days_to_death annotated along with gene expression levels. The log-normalized gene expression data were further engineered. We also adopted PCC to narrow down the number of gene features for efficient model learning. PCC between days_to_death and log-normalized gene expression levels were computed. A variety of PCC thresholds from 0.2 to 0.45 resulted in hundreds to thousands of gene features since gene features with higher PCC than the thresholds were selected. Along with

PCC, the number of estimators was used for prediction performance grid search. Mean Absolute Error (MAE) was used as our cost function. We ran Random Forest regression 100 times per combination to learn robust prediction performance settings and to repress randomization side effects.

Fig. 3 (top) displays the results of the parameter tuning; ultimately, we found that features with a PCC threshold of ~0.34 running with 1000 estimators gave the lowest Mean Absolute Error. Overfitting degraded prediction performance when a few gene expression features with too high PCC were selected. The average importance of the top 10 frequent genes is shown in Fig. 3 (bottom). KLHDC8B and DENND1A were shown in all 100 training and tests.

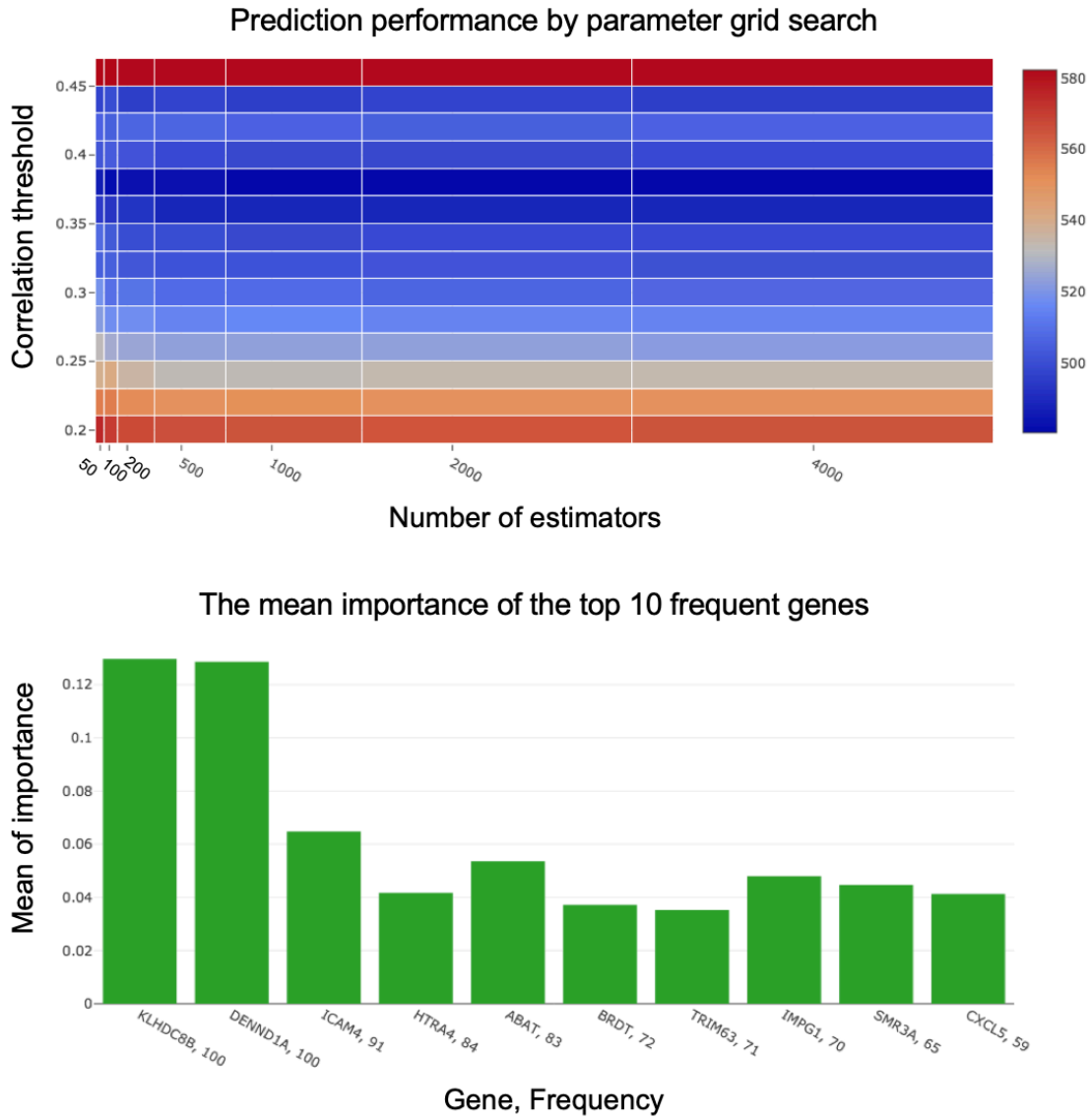

Fig. 3 The transcriptome-based RFR model prediction performance was grid-searched along with number of estimators and PCC threshold (top). The top 10 genes by frequency were represented. The mean importance was computed across 100 runs (bottom).

3.3. *Proteome*

Processed Reverse Phase Protein Array (RPPA) data were retrieved from TCGA. The data described the quantified protein abundance of 364 patients for 225 proteins. The data was already normalized, as displayed in Fig. 1D.

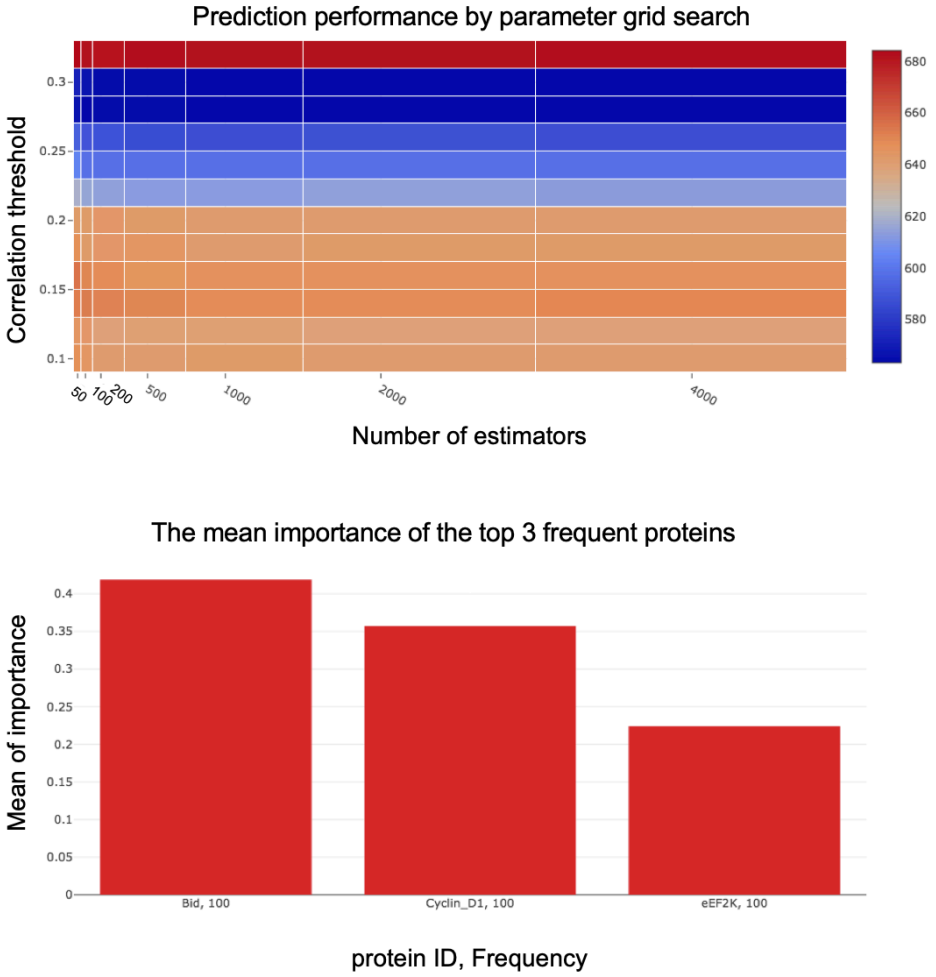

Fig. 4 Proteome-base RFR model prediction performance was grid-explored along with number of estimators and PCC threshold (top). The top 3 CpGs by frequency were represented. The mean importance was computed across 100 runs (bottom).

The normalized data file had to be further engineered. We compared patient barcodes, unique across the TCGA project, of the clinical annotation file with those of the normalized protein abundance file. Then we selected patients who had days_to_death annotated along with protein abundance levels. We also adopted PCC to narrow down the number of gene features for efficient model learning. PCC between days_to_death and normalized protein abundance levels were computed. A variety of PCC thresholds from 0.1 to 0.3 resulted in tens of protein abundance features since protein features with higher PCC than the thresholds were selected. Along with PCC, the number of estimators was used for prediction performance grid search. Mean Absolute Error (MAE) was used as our cost function. We ran 5-fold cross-validation to measure prediction

performance. We ran Random Forest regression 100 times per combination to learn robust prediction performance settings and to repress randomization side effects.

Fig. 4 (top) displays the results of the parameter tuning; ultimately, we found that features with a PCC threshold of ~0.28 running with 4000 estimators gave the lowest Mean Absolute Error. Overfitting degenerated prediction performance when too few protein features due to extremely high PCC threshold were selected. The average importances of the three most frequent proteins are shown in Fig. 4 (bottom).

3.4. *Clinical annotation*

Clinical annotation data was retrieved from TCGA. We selected patients who had the days_to_death annotation, and then selected clinical annotation features that were recorded for all those patients. Roughly ~50 clinical annotations were available for model learning. Since PCC was low across available features and 50 features does not hurt the learning efficiency, we learned a model with all 50 clinical features without further feature selection.

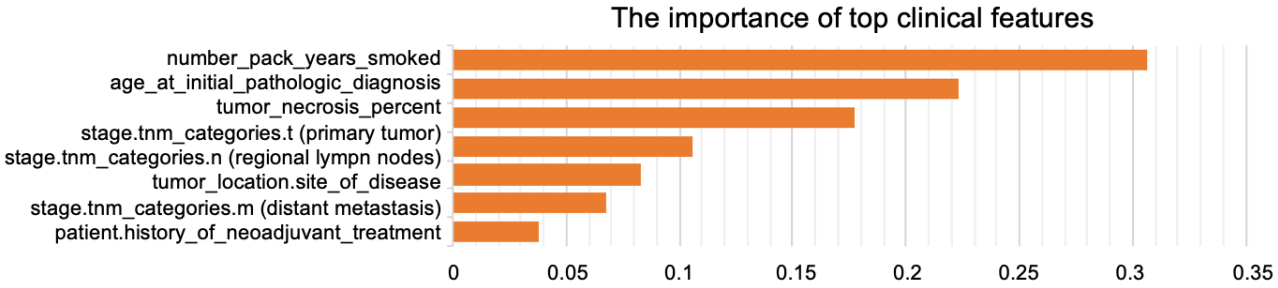

Fig. 5 Traditionally, survival time was estimated by doctors using clinical data such as smoking history, age, and stage. Since PCC of clinical data is lower than other omics data, PCC thresholds were not applied. The RFR model learned from all clinical annotations and found that smoking history was the most important, followed by age, necrosis percentage, and tumor stage.

Feature importance was displayed in Fig. 5 and overall performance was recorded in Fig. 7. As expected, the most important feature is smoke intensity which is represented in packs/years, followed by age upon diagnosis, then tumor stage information; primary tumor (T) was the most significant factor, followed by regional lymph node (N) and distant metastasis (M). The revealed importances by RFR was well-aligned with what has been the traditional method to estimate survival time by doctors.

4. Integrative multi-omics prediction model

4.1. *Integrative modeling*

We tried to find the intersection among methylation, gene expression, and protein abundance (Fig. 6). DKK1 and CFOD2 are confirmed by both methylation and gene expression data. Since there was a low amount of protein data available, none of the genes in methylation or gene expression data could be cross-confirmed by protein data.

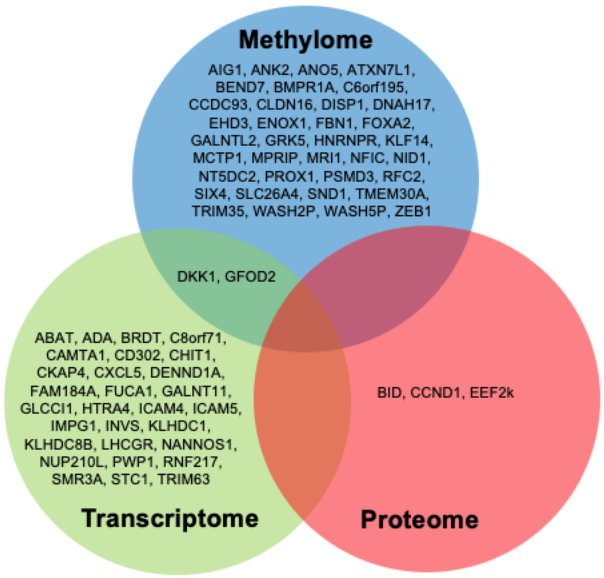

Fig. 6 We further investigated if there are any genes that were double confirmed by two or more single omics-based models. DKK1 and GFOD2 displayed significance in both methylation and gene expression data. Since TCGA generated only a handful of proteome data (~200), it was unlikely for genes to be confirmed by proteomic data.

This inspired us to further develop an integrative model with all the heterogeneous omics data and the clinical annotations. We again chose Random Forest Regression (RFR) because it can handle non-linear solution space and it does not require intense normalization. We selected features across the three omics data and the clinical annotations by PCC thresholds. We learned a model, measured prediction performance after 5-fold cross-validation, and plotted the prediction performance as MAE.

Firstly, the integrative model outperformed any single omics-based prediction model across all PCC thresholds (Fig.7). For single-omics based models, the methylation-based model performed best followed by gene expression based model. These two omics-based models predict better than traditional survival time estimates based on smoke history, tumor stage, and age upon diagnosis. This can be partly attributed to the fact that having more features gives a higher chance to come across better predictor features. Note that the clinical annotation-based model was stronger than the protein-based model despite the fact that protein provided more features than clinical annotations.

Prediction performance was measured in Table 1 and Fig. 7 (top). The best performance (with the lowest MAE) was recorded along with PCC threshold and number of estimators (decision trees).

Table 1. Prediction performance

	Clinical annotation	Omics data			Integrative model
		Methylation	RNA	Protein	
Best Performance (the lowest MAE)	547.0829	456.8844	480.5176	563.304	436.8226
# of features before feature engineering	~3K	~500K	~30K	~2K	~100
PCC threshold	NA	0.38	0.34	0.28	0.375
# features used for training models (at optimal Pearson threshold)	~50	<100	<100	<10	Methylome : ~20 Transcriptome ~10 Proteome <10 Clinical Annotation ~10
# eatures with high importance	<10	~30	~10	<10	~10
# estimators	4000	7500	1000	4000	2000

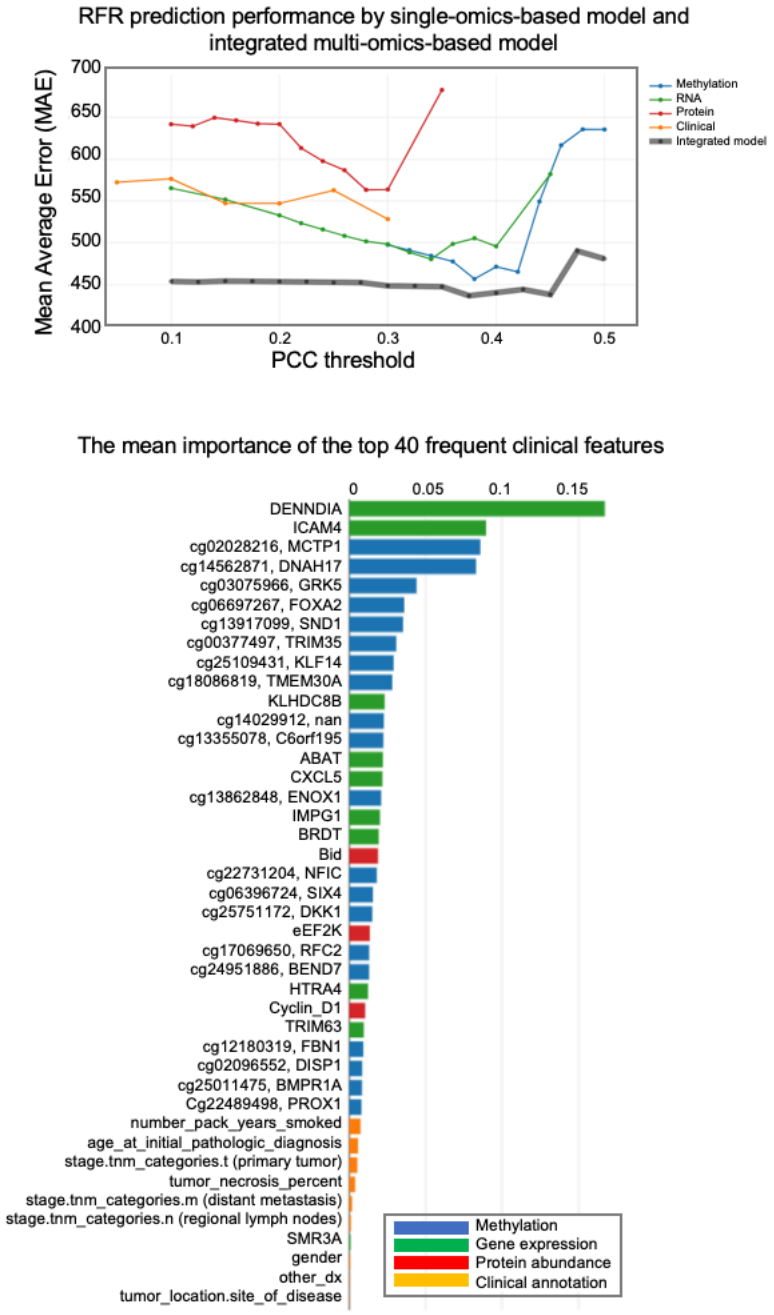

Fig. 7. RFR prediction performance from 5-fold cross-validation shows that the clinical data-based prediction model performs better than the protein-based model. However, it also reveals that the gene expression or methylation-based model can outperform traditional survival time estimates. The integrated multi-omics data prediction model outperforms any model that relies on solely one type of omics data or clinical data (top). The mean importances of the top 40 features by frequency are shown after 100 runs and 5-fold cross-validation. Interestingly, the top two features are from gene expression, followed by methylation features. We identified ~20 omics features more significantly predictable than traditional clinical features (bottom).

Running time was measured on a MacBook Pro with Intel(R) Core TM core i5 processor and 8GB of RAM. The number of features significantly affected learning time. For example, methylation has the most number of marker candidate features, and thus took the longest runtime.

Table 2. Running time analysis of RFR

	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5
Protein	<1hr	<1hr	<1hr	<1hr	<1hr	<1hr	NA	NA	NA
Clinical	<1hr	<1hr	<1hr	<1hr	<1hr	NA	NA	NA	NA
RNA	>48 hrs	~20hrs	~6hrs	~2hrs	<1hr	<1hr	<0.5hr	<0.5hr	NA
Methylation	>2 weeks	>2 weeks	>36hrs	>24hrs	~15hrs	~5hrs	~1.5hrs	<1hr	<1hr
Integrative model	<1hr	<1hr	<1hr	<1hr	<1hr	<1hr	<1hr	<.5hr	<0.5hr

4.2. *Omics-marker cost analysis*

We further investigated to find which omic provides the most cost-effective markers. Though the methylation and protein cost more than RNA-seq, the methylation platform is the most economical because methylation generates millions of CpG markers, resulting in the lowest total cost/marker. Comparatively, only ~200 protein abundance are generated for about the same price in the protein platform, and the RNA-seq also ultimately provides a higher total cost/marker due to providing only 20-30K gene expression markers compared to the 5M-30M CpG markers provided by the methylation platform.

Table 3 Omics marker unit cost analysis [14]

	Methylome	Transcriptome	Proteome
Total number of markers	5M	20K	200
Library cost	$300	$80	$320
Sequencing cost	$1000	$1000	NA
Library cost/marker	$0.00006	$0.004	$1.6
Sequencing cost/marker	$0.0002	$0.05	NA
Total cost/marker	$0.00026	$0.054	$1.6

5. Discussion and future works

In this study, we used a nonparametric ensemble learning method, Random Forest Regression, to predict survival time of lung adenocarcinoma (LUAD) patients from heterogeneous omics data and clinical annotations. We specifically chose LUAD because it is widely accepted that smoking history is one of the most important factors to estimate survival time, along with other clinical factors such as age and tumor stage. Our goal was to identify omics markers that outperform such clinical markers, which have previously been the most reasonable factors in predicting survival

time, and we successfully found such better-predicting omics markers, such as DENNDIA, ICAM4, cg02038216 (MCTP1), cg03075966 (GRK5), cg06697267(FOXA2), etc. In the future, it would be interesting to apply RFR to other types of cancer data from TCGA to see (1) if methylation markers consistently outperform gene expression markers and (2) if there are any common methylation/gene expression markers to predict survival time.

Acknowledgments and Funding

The authors thank Akshay Sanghi for discussing the significance of smoking history in lung cancer patients. This work used the Genome Sequencing Service Center by Stanford Center for Genomics and Personalized Medicine Sequencing Center, supported by the grant award NIH S10OD020141.

References

1. Cancer Genome Atlas Research Network, "Comprehensive molecular profiling of lung adenocarcinoma," Nature, vol. 511, no. 7511, pp. 543–550, Jul. 2014.
2. T. K. Ho, "Random decision forests," Proceedings of 3rd International Conference on Document Analysis and Recognition.
3. T. K. Ho, "The random subspace method for constructing decision forests," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, no. 8. pp. 832–844, 1998.
4. K. R. Campbell and C. Yau, "A descriptive marker gene approach to single-cell pseudotime inference," Bioinformatics, vol. 35, no. 1, pp. 28–35, Jan. 2019.
5. L. Liu et al., "Favorable outcome of patients with lung adenocarcinoma harboring POLE mutations and expressing high PD-L1," Mol. Cancer, vol. 17, no. 1, p. 81, Apr. 2018.
6. D. Nie et al., "Multi-Channel 3D Deep Feature Learning for Survival Time Prediction of Brain Tumor Patients Using Multi-Modal Neuroimages," Sci. Rep., vol. 9, no. 1, p. 1103, Jan. 2019.
7. K.-H. Yu et al., "Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features," Nat. Commun., vol. 7, p. 12474, Aug. 2016.
8. S. Li et al., "Identification of an eight-gene prognostic signature for lung adenocarcinoma," Cancer Manag. Res., vol. 10, pp. 3383–3392, Sep. 2018
9. J. Yu *et al.*, "LUADpp: an effective prediction model on prognosis of lung adenocarcinomas based on somatic mutational features," *BMC Cancer*, vol. 19, no. 1, p. 263, Mar. 2019.
10. K. Y. Wong *et al.*, "I-Boost: an integrative boosting approach for predicting survival time with multiple genomics platforms," *Genome Biol.*, vol. 20, no. 1, p. 52, Mar. 2019.
11. R. D. Yockey, "The Pearson r Correlation Coefficient," SPSS® Demystified. pp. 156–163, 2018.
12. B. Li and C. N. Dewey, "RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome," BMC Bioinformatics, vol. 12, no. 1. 2011.
13. M. Teng et al., "A benchmark for RNA-seq quantification pipelines," Genome Biol., vol. 17, p. 74, Apr. 2016.
14. "Epigenomics Core @ WCMC." [Online]. Available: http://epicore.med.cornell.edu/pricelist.php. [Accessed: 06-Aug-2019].