



# Predictive Signatures for Lung Adenocarcinoma Prognostic Trajectory by Multiomics Data Integration and Ensemble Learning

Hayan Lee<sup>1</sup>(✉), Gilbert Feng<sup>2</sup>, Ed Esplin<sup>1</sup>, and Michael Snyder<sup>1</sup>(✉)

<sup>1</sup> Department of Genetics, Stanford University, Stanford, CA, USA  
{hayan.lee, explain, mpsnyder}@stanford.edu

<sup>2</sup> Electrical Engineering and Computer Sciences, University of California, Berkeley, Berkeley, CA, USA  
gilbertfeng@berkeley.edu

**Abstract.** Lung cancer is the most prevalent cancer worldwide. About 80% to 85% of lung cancers are non-small cell lung cancer (NSCLC). One of the major types of NSCLC is lung adenocarcinoma (LUAD), which solely accounts for approximately 40% of all cases. Although there has been a dramatic therapeutic improvement, the prognostic trajectory has relied on primarily clinical features such as tumor-nodal-metastasis (TNM) stage, age upon diagnosis, and smoking history for decades. It does not reflect molecular alterations on its pathway or heterogeneity of tumorigenesis. Here we propose an integrative multi-omics random forest model to predict survival time for LUAD patients. We identified multi-omics signatures with higher importance to better predict survival time than clinical annotations that physicians traditionally use. We confirmed that the integrative prediction model outperforms any single-omic-based model. We discovered that a methylation-based model performed best among any single-omic-based model for LUAD since it provides the most abundant signature candidates. Although methylation assay is costly in general, paradoxically, methylation offers the most economical pool as prognosis markers due to more abundant assay points.

**Keywords:** Survival time prediction · Integrative multi-omics model · Machine learning

## 1 Introduction

Lung cancer is the most prevalent cancer in many 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 [1, 2]. Lung adenocarcinoma (LUAD) is one of the major subtypes of non-small cell lung cancer, the only subtype in never smokers [3], along with lung squamous cell carcinoma (LUSC). LUAD accounts for approximately 40% of all lung cancer cases.

Prediction of its prognostic trajectory is important, especially to each patient. Traditionally prognostic trajectory has been estimated by clinical data, such as ages upon diagnosis, stage, and smoke history. This method is inaccurate because it does not consider

molecular characteristics in its pathway. Kaplan-Meier estimates provide the simplest and predictable way given the selected factor but cannot associate with related other factors and cannot be expressed as a functional form [4]. The Cox proportional hazards model can handle multiple variables and is realistic in that the hazard ratio changes over time, but it is not predictable to estimate survival time [5]. Recently there has been a dramatic improvement in treatment by molecule-targeting therapies. However, its prediction is not accurate enough because it still relies solely on clinical annotation and does not take molecular responses into account. Some data integration effort was made by incorporating gene expression data [6, 7]. Epigenetics data also have been exploited [8, 9]. There has been an effort to have even more omics data involved but most ended up with algorithmic integration rather than omics-level data integration [10].

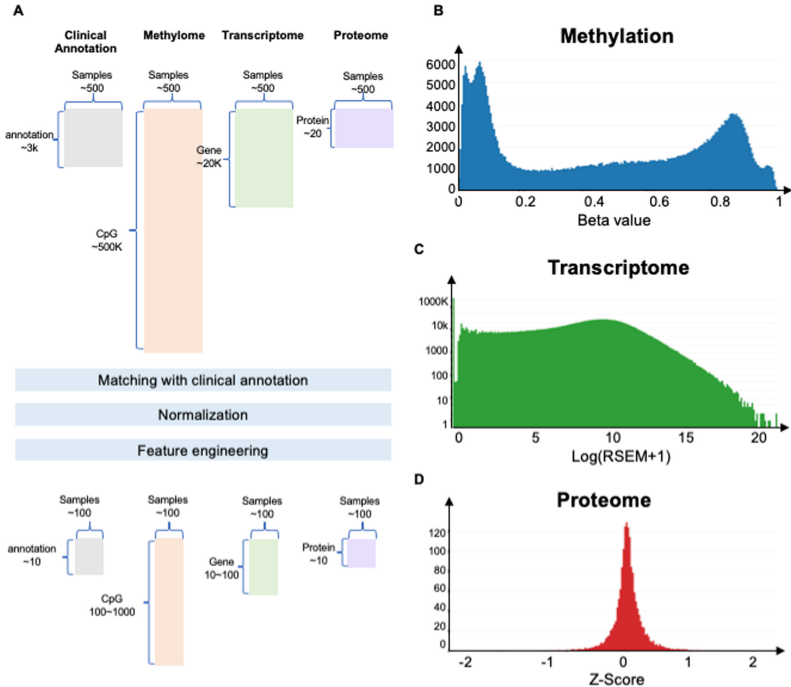
Thus, we propose an ensemble learning method on high-dimensional omics data with clinical annotation to predict survival time. We learned a random forest regression (RFR) model [11, 12] exploiting the Cancer Genome Atlas (TCGA) LUAD multi-omics data and clinical annotations [13, 14]. To build a survival time prediction model, ‘Days to death’ was selected as the prediction target. Random forest regression, an ensemble of multiple decision trees, was selected to learn a model across heterogeneous data types since it does not require normalizing features, while other machine learning algorithms usually do. Random forest regression 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.

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 [15–17]. Yu et al. performed classification of long-term vs. short-term survival of non-small cell lung cancer patients but exploited mainly hematoxylin and eosin (H&E) histological image data with a few omics markers of interest [18]. Li et al. identified eight genes relating to survival in LUAD using only gene expression data [19]. Yu et al. learned a prediction model to classify short-term (<3yr) and long-term (>3yr) survival from LUAD using only somatic mutational features [20]. An integrative prediction model suggested RNA-seq should be more predictable on prognostics of survival time than other genomic data types but still failed to include the methylation data that eventually causes gene expression change. We extended omics data integration from conventional clinical history to methylation, gene expression, and protein abundance from TCGA LUAD patients.

## 2 Inherent Characteristics of Multi-omics Data

TCGA has generated a variety of omics data along with clinical annotations. TCGA detailed molecular levels 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 that 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 [20]. 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 from ~500 patients. About 20% of them have survival time annotations. Methylation data is already normalized by its assay design. We performed log normalization for gene expression data, and proteome data were also Z-score normalized. To further reduce model learning time, we performed feature engineering using Pearson correlation coefficients (PCC) [21, 22]. The preprocess reduced the original data matrix to 1/1000–1/2 (Fig. 1).



**Fig. 1.** Methylation, gene expression, and protein distribution from TCGA LUAD naturally vary. It presents that each omics data has its unique distribution signature. Accordingly, normalization should be adopted for prediction model learning. (A) For LUAD, TCGA collected data from ~500 patients and provided ~500K CpG methylation, ~20K gene expression, ~200 protein, and 100 clinical annotations, including ‘Days to death’, the target phenotype. (B) The distribution of the methylome (C) The distribution of transcriptome (D) The distribution of proteins.

### 3 Single-Omics Prediction Model

#### 3.1 Conventional Clinical History Based Model as a Baseline

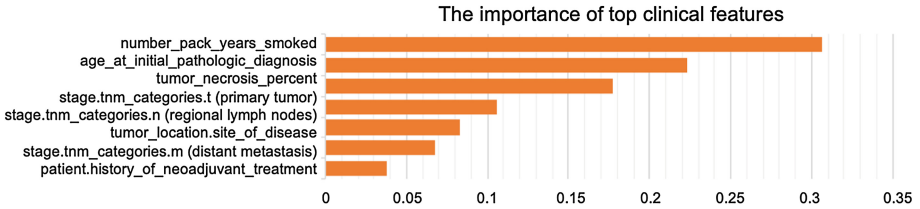
Clinical annotation data were retrieved from TCGA. We selected patients who had the ‘Days to death’ annotation and then selected other clinical history features that were recorded for all those patients such as ‘number\_pack\_years\_smoked’. Roughly ~50 clinical annotations were available for model learning. Since 50 features do not hurt

the learning efficiency, we trained a model with all 50 clinical features without further feature selection.

Feature importance in random forest is formulated in (1) and (2); that is each feature  $x$  can change prediction accuracy compared to random permutation of the feature  $x$ , tested in out-of-bag (OOB) data points. Feature importance of clinical annotation was displayed in Fig. 2, and overall performance was represented in Fig. 7. As expected, the most important feature is the smoking intensity which is represented in packs/years. The second most important feature was ‘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 importance by RFR were realistic and well-aligned with what has been known by the traditional method to estimate survival time by doctors.

$$Importance_x = \frac{1}{|all\ Tree|} \sum_{all\ Tree} \Delta Accuracy_{Tree,x}^{OOB} \quad (1)$$

$$\Delta Accuracy_{Tree,x}^{OOB} = \frac{1}{|OOB_T|} \sum_{i \in OOB_T} Accuracy\ with\ perm_x - Accuracy\ w/o\ perm_x \quad (2)$$



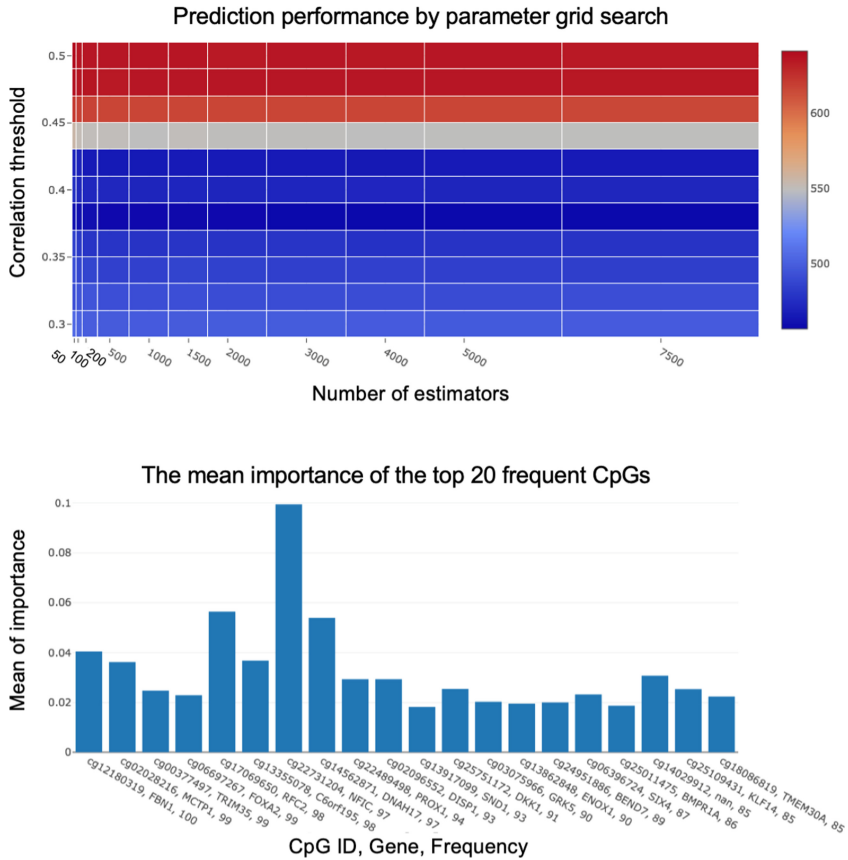
**Fig. 2.** Traditionally, survival time was estimated by doctors using well-known clinical history such as smoking intensity and time, age, and stage. Since PCC of clinical data is lower than other omics data, PCC thresholds were not applied. The RFR model learned from all clinical annotations and found that smoking intensity was the most important, followed by age, necrosis percentage, and tumor stage.

Clinical annotation-based RFR is selected as our baseline for the entire study. It is a strong baseline given that non-linear ensemble learning methods can address such complex heterogeneous data.

### 3.2 Methylation-Based Survival Time Prediction Model

For methylation, TCGA adopted Illumina Infinium HumanMethylation450K BeadChip (HM450) [23], where half a million CpGs were assayed to compute beta values, i.e., methylation ratio (which is the number of reads with methylated cytosine divided by the total number of reads). We extracted LUAD methylation data of the patients with ‘Days to death’. 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.

~100 patients were used for model training. The total number of assayed CpGs was ~485K. The beta value is already normalized as a ratio of 0 to 1. CpG loci that started with 'cg' were selected and the features with any missing data across the ~100 patients were excluded for model learning. The shape of the distribution of methylation data is bimodal (Fig. 1A), where CpGs are either highly methylated or unmethylated, which is very authentic compared to bell curve shares of transcriptome and proteome data.



**Fig. 3.** Methylome-base RFR model prediction performance was grid-explored along with the number of estimators and PCC threshold (top). Since RFR includes a randomization process, we ran RFR > 100 times. Thus the importance was averaged across 100 runs (bottom).

We employed mean absolute error (MAE) as our cost function [24]. We utilized 5-fold cross-validation to measure prediction model performance. To make model learning efficient and effective, we performed feature engineering by 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, and intervals set exponentially. We defined a parameter grid and ran random forest regression for each

cell to search the parameter combination space for the best-performed model. Since the algorithm relies on randomization, we ran random forest regression 100 times per cell to obtain more robust performance results (Fig. 3). The best performance was shown with a PCC threshold of 0.4, and the number of estimators does not seem to affect the performance significantly. 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 presented as an MAE heatmap. 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. 3). The genes related to the CpG and the actual frequency are shown with a CpG ID. The average importance value is on the Y-axis.

### 3.3 Transcriptome-Based Ensemble-Learning Model

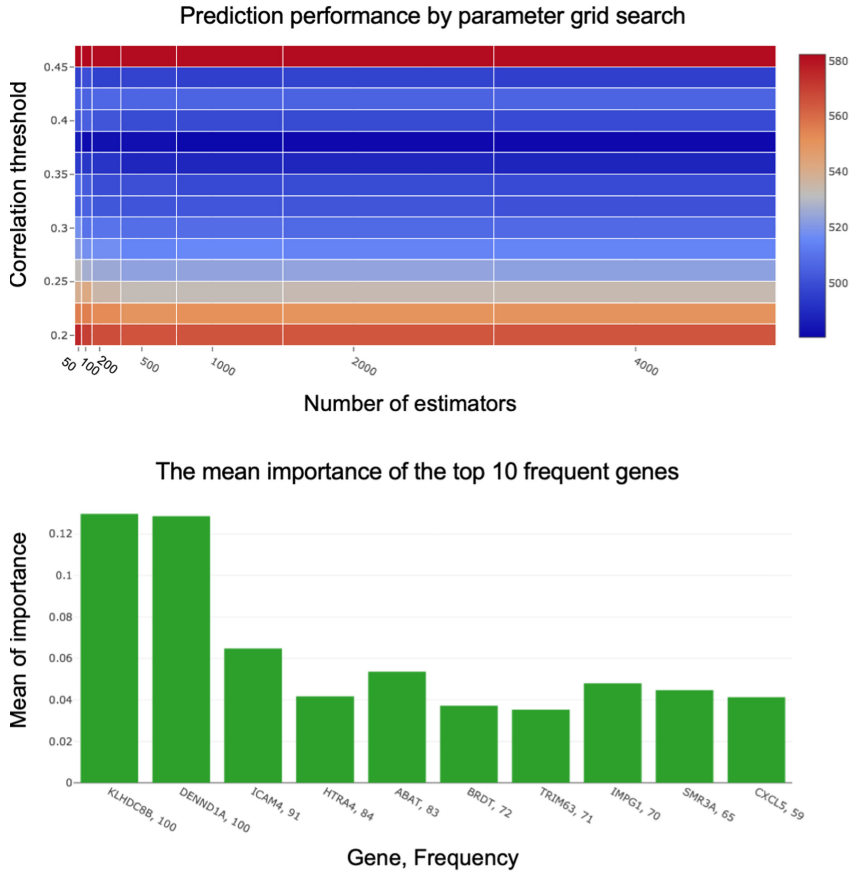
The gene expression data were also retrieved from TCGA. The data quantified over 20K genes for ~500 LUAD patients. The quantified gene expression levels were computed through RSEM [25, 26], which can deal with multiple isoforms fast by parallel computing the EM algorithm. The raw data was originally 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 some genes are not expressed at all (log 0 is not defined).

We compared patient barcodes, unique across the TCGA project, of the clinical annotation file with raw gene expression files. Then we selected patients who had ‘Days to death’ annotated along with gene expression. 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. 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. 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.

Ultimately, we found that features with a PCC threshold of ~0.34 running with 1000 estimators gave the lowest MAE. Overfitting degraded prediction performance when a few gene expression features with too high PCC were selected (Fig. 4). The average importance of the top 10 frequent genes is shown in Fig. 4 (bottom). KLHDC8B and DENND1A were shown in all 100 training and tests [27, 28].

### 3.4 Proteome-Based Model

Processed Reverse Phase Protein Array (RPPA) data were retrieved from TCGA [29]. The data described the quantified protein abundance of 364 patients for 225 proteins. The data was already normalized, as displayed in Fig. 1D. The normalized data file had to be further engineered. We compared patient barcodes, unique across the TCGA

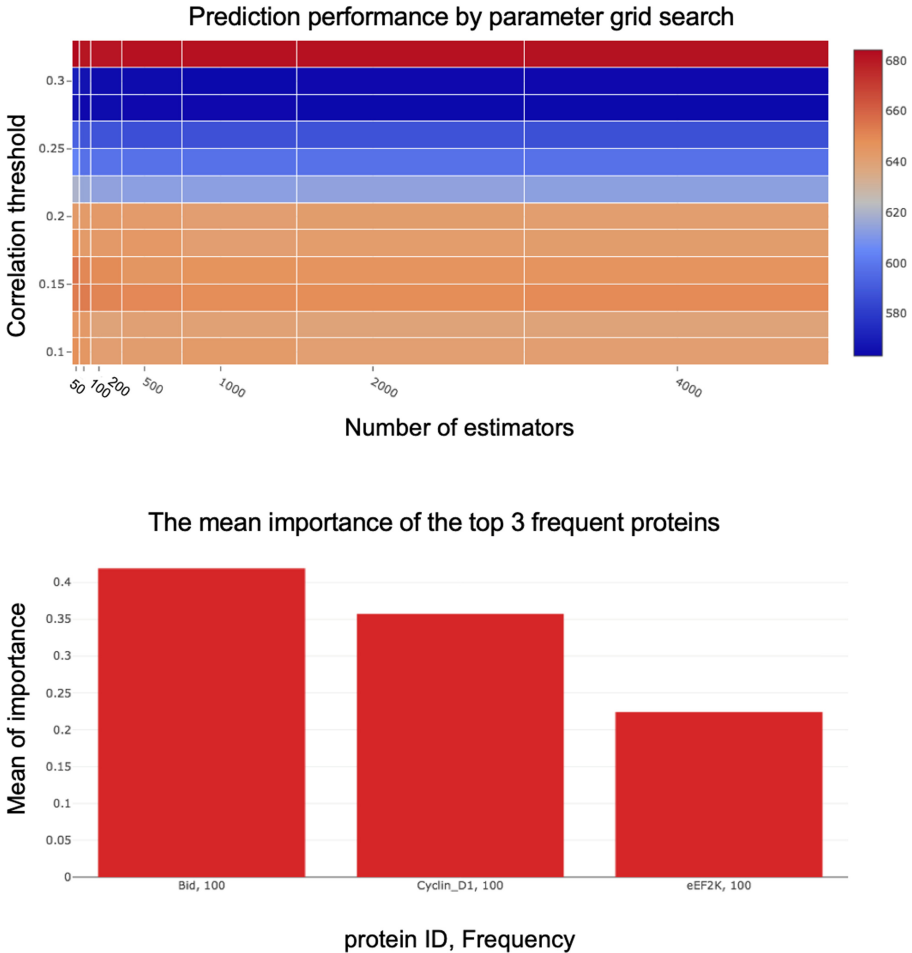


**Fig. 4.** The transcriptome-based RFR model prediction performance was grid-searched along with the number of estimators and PCC threshold (top). The top 10 genes by frequency are represented, and the mean importance was computed across 100 runs (bottom).

project, of the clinical annotation file with the normalized protein abundance file. Then we selected patients who had 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 was 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. 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 suppress randomization side effects. The top 3 proteins (BID, CCT5, EEF2K) by frequency were represented [30–32].

The results of the parameter tuning were demonstrated in Fig. 5. Ultimately, we found that features with a PCC threshold of  $\sim 0.28$  running with 4000 estimators gave the lowest MAE. Overfitting degenerated prediction performance when too few protein features due to extremely high PCC threshold were selected. Furthermore, too high a threshold may lose some of the informative features. The average importance values of the three most frequent proteins are shown in Fig. 5 (bottom).



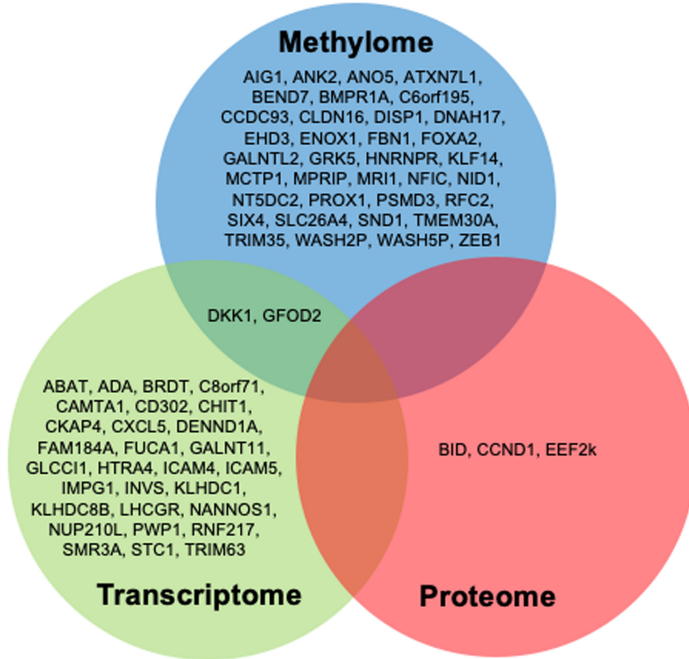
**Fig. 5.** Proteome-base RFR model prediction performance was grid-explored along with the number of estimators and PCC threshold (top). The top 3 proteins (BID, CCT5, EEF2K) by frequency were represented. The mean importance was computed across 100 runs (bottom).



## 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 GFOD2 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 two or more single omics-based models double confirmed. 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 to confirm genes 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 because it can handle non-linear solution space and 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 the gene expression-based model. These two single-omic-based models predict better than traditional survival time estimates based on smoke history, tumor stage, and age upon diagnosis. It is partly because having more features gives a higher chance to come across better predictor features. Note that the clinical annotation-based model was more robust than the protein-based model, even though protein provides more features than clinical annotations in general.

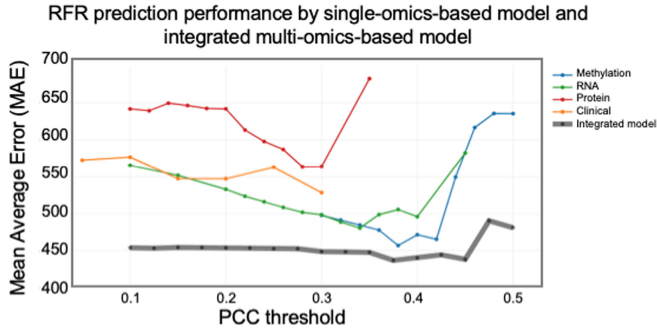
Prediction performance was measured in Table 1 and Fig. 7 (top). The best prediction performance, i.e., the lowest MAE, was recorded along with PCC threshold and a various number of estimators, meaning that decision trees. Each MAE data point was averaged after 5-fold cross-validation. Although the significance is hard to be proven, it is clear that the MAE of the integrative model is lower than any other single-omics-based model across all feature combinations. The improvement is more prominent when the MAE of single-omics-based models is higher in the far left and far right. It also can be interpreted that more noisy features when a lower PCC threshold is applied. At the same time, we lose more informative and predictable features when a higher PCC threshold is applied.

Running time was measured on a MacBook Pro with Intel(R) Core TM core i5 processor and 8 GB of RAM. The number of features significantly affected learning time. For example, methylation had the most marker candidate features and thus took the most extended runtime (Table 2).

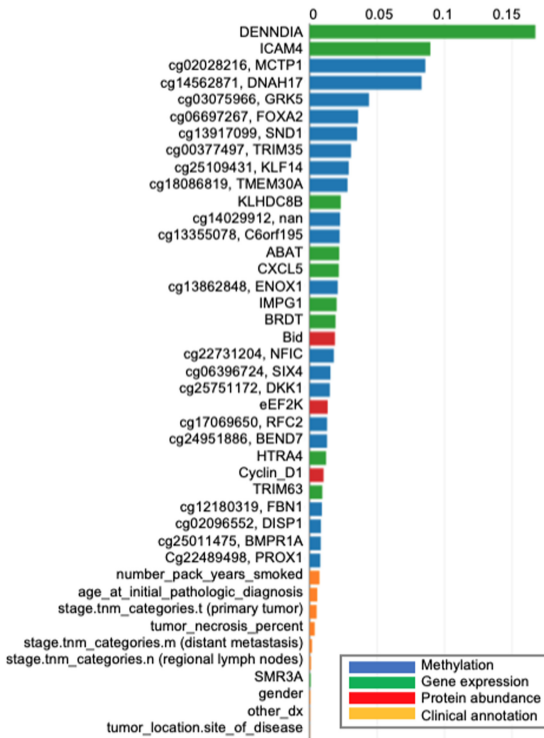
## 4.2 Omics-Marker Cost Analysis

We further investigated to find which omics data provide the most cost-effective markers (Table 3). 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 and highest predictive power per dollar [33]. The predictive power is formulated as a reverse or error rate, i.e., MAE and computed (3). Comparatively, only ~200 proteins are generated, thus offer the most costly marker.

$$\text{Predictive power} = \frac{1}{MAE} \quad (3)$$



The mean importance of the top 40 frequent clinical features



**Fig. 7.** RFR prediction performance from 5-fold cross-validation shows that the clinical data-based prediction model, our baseline, 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 other models that rely solely on one type of omics data or clinical data (top). The mean importance values 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).

**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
# features with high importance	<10	~30	~10	<10	~10
# estimators	4000	7500	1000	4000	2000

**Table 2.** Running time analysis of RFR by PCC threshold

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

**Table 3.** Omics marker unit cost analysis

	Methylome	Transcriptome	Proteome
Total number of markers	~4.5M	~30K	~200
Library cost [34]	>\$300	~\$80	~\$320
Sequencing cost	~\$1,000	~\$1,000	0
Total cost/marker	~\$0.0029	~\$0.036	~\$1.60
Best MAE	456.8844	480.5176	563.304
Predictive power/USD(\$)	2.87	2.25	0.57

## 5 Discussion and Future Works

In this study, we used a random forest regression, non-parametric ensemble learning method, to predict the 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 DENND1A, ICAM4, cg02038216 (MCTP1), cg03075966 (GRK5), cg06697267 (FOXA2), etc. [35] It is observed that DENND1A is overexpressed in LUAD patients [36]. ICAM4 was identified as methylation markers by Wang et al. [37].

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 and (3) if the prediction power can be improved by adding image data, which is available to LUAD patients.

**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. Lung Carcinoma: Tumors of the Lungs. Merck Manual Professional Edition, Online Edition. Accessed 12 Aug 2021
2. What Is Lung Cancer? Types of Lung Cancer. <https://www.cancer.org/cancer/lung-cancer/about/what-is.html>. Accessed 12 Aug 2021
3. Subramanian, J., Govindan, R.: Lung cancer in never smokers: a review. *J. Clin. Oncol.* **25**(5), 561–570 (2007)
4. Goel, M.K., Khanna, P., Kishore, J.: Understanding survival analysis: Kaplan-Meier estimate. *Int. J. Ayurveda Res.* **1**, 274–278 (2010)
5. Katzman, J.L., et al.: DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Med. Res. Methodol.* **18** (2018). Article number: 24
6. West, M., et al.: Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc. Natl. Acad. Sci.* **98**, 11462–11467 (2001)
7. Van't Veer, L.J., et al.: Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**, 530–536 (2002)
8. Yu, X., et al.: Predicting lung adenocarcinoma disease progression using methylation-correlated blocks and ensemble machine learning classifiers. *PeerJ* **9**, e10884 (2021)
9. Cao, W., et al.: Multi-faceted epigenetic dysregulation of gene expression promotes esophageal squamous cell carcinoma. *Nat. Commun.* **11** (2020). Article number: 3675

10. Kwek, S., Nguyen, C.: iBoost: boosting using an instance-based exponential weighting scheme. In: Elomaa, T., Mannila, H., Toivonen, H. (eds.) *Machine Learning: ECML 2002*, pp. 245–257. Springer, Heidelberg (2002). [https://doi.org/10.1007/3-540-36755-1\\_21](https://doi.org/10.1007/3-540-36755-1_21)
11. Ho, T.K.: Random decision forests. In: *Proceedings of 3rd International Conference on Document Analysis and Recognition (1995)*
12. Ho, T.K.: The random subspace method for constructing decision forests. *IEEE Trans. Pattern Anal. Mach. Intell.* **20**, 832–844 (1998)
13. Cancer Genome Atlas Research Network, et al.: The cancer genome atlas pan-cancer analysis project. *Nat. Genet.* **45**, 1113–1120 (2013)
14. The Cancer Genome Atlas Research Network: Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511**, 543–550 (2014)
15. Campbell, K.R., Yau, C.: A descriptive marker gene approach to single-cell pseudotime inference. *Bioinformatics* **35**, 28–35 (2019)
16. Liu, L., et al.: Favorable outcome of patients with lung adenocarcinoma harboring POLE mutations and expressing high PD-L1. *Mol. Cancer* **17** (2018). Article number: 81
17. Nie, D., et al.: Multi-channel 3D deep feature learning for survival time prediction of brain tumor patients using multi-modal neuroimages. *Sci. Rep.* (2019). **9** Article number: 1103
18. Yu, K.-H., et al.: Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. *Nat. Commun.* **7** (2016). Article number: 12474
19. Li, S., et al.: Identification of an eight-gene prognostic signature for lung adenocarcinoma. *Cancer Manag. Res.* **10**, 3383–3392 (2018)
20. Yu, J., et al.: LUADpp: an effective prediction model on prognosis of lung adenocarcinomas based on somatic mutational features. *BMC Cancer* **19** (2019). Article number: 263. <https://doi.org/10.1186/s12885-019-5433-7>
21. National Cancer Institute: Pearson Correlation Coefficient. Definitions (2020)
22. Gooch, J.W.: Pearson correlation coefficient. In: Gooch, J.W. (ed.) *Encyclopedic Dictionary of Polymers*, pp. 990–990. Springer, New York (2011). [https://doi.org/10.1007/978-1-4419-6247-8\\_15317](https://doi.org/10.1007/978-1-4419-6247-8_15317)
23. Infinium HumanMethylation450K BeadChip Product Files. [https://support.illumina.com/array/array\\_kits/infinium\\_humanmethylation450\\_beadchip\\_kit/downloads.html](https://support.illumina.com/array/array_kits/infinium_humanmethylation450_beadchip_kit/downloads.html)
24. Chai, T., Draxler, R.R.: Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. *Geosci. Model Dev.* **7**, 1247–1250 (2014)
25. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *Bioinformatics* 69–102 (2014). <https://doi.org/10.1201/b16589-9>
26. Teng, M., et al.: A benchmark for RNA-seq quantification pipelines. *Genome Biol.* **17**, 203 (2016)
27. Expression of KLHDC8B in lung cancer - The Human Protein Atlas. <http://www.proteinatlas.org/ENSG00000185909-KLHDC8B/pathology/lung+cancer/LUAD>. Accessed 12 Aug 2021
28. Zengin, T., Önal-Süzek, T.: Analysis of genomic and transcriptomic variations as prognostic signature for lung adenocarcinoma. *BMC Bioinformatics* **21**, 368 (2020)
29. Reverse phase protein microarray (reverse phase protein array, RPPA, ‘reverse phase array’). *The Dictionary of Genomics, Transcriptomics and Proteomics 1* (2015)
30. Billen, L.P., Shamas-Din, A., Andrews, D.W.: Bid: a Bax-like BH3 protein. *Oncogene* **27**, S93–S104 (2008)
31. Meng, Y., et al.: CCT5 interacts with cyclin D1 promoting lung adenocarcinoma cell migration and invasion. *Biochem. Biophys. Res. Commun.* **567**, 222–229 (2021)
32. Expression of EEF2K in lung cancer. The Human Protein Atlas. <https://www.proteinatlas.org/ENSG00000103319-EEF2K/pathology/lung+cancer/LUAD>. Accessed 12 Aug 2021
33. Epigenomics Core @ WCMC. <http://epicore.med.cornell.edu/pricelist.ph>. Accessed 12 Aug 2021

34. Infinium MethylationEPIC Kit. <https://www.illumina.com/products/by-type/microarray-kits/infinium-methylation-epic.html>
35. Li, C., Long, Q., Zhang, D., Li, J., Zhang, X.: Identification of a four-gene panel predicting overall survival for lung adenocarcinoma. *BMC Cancer* **20**, 1198 (2020)
36. McAllister, J.M., et al.: Overexpression of a DENND1A isoform produces a polycystic ovary syndrome theca phenotype. *Proc. Natl. Acad. Sci. U. S. A.* **111**, E1519–E1527 (2014)
37. Wang, R., Zhu, H., Yang, M., Zhu, C.: DNA methylation profiling analysis identifies a DNA methylation signature for predicting prognosis and recurrence of lung adenocarcinoma. *Oncol. Lett.* **18**, 5831–5842 (2019)