plasma: Partial LeAst Squares for Multiomics Analysis

Kyoko Yamaguchi[1], Salma Abdelbaky[1], Lianbo Yu[2], Christopher C. Oakes[1], and Kevin R. Coombes[3]

[1]Division of Hematology, Department of Internal Medicine, Ohio State University, Columbus, OH USA
[2]Department of Biomedical Informatics, Ohio State University, Columbus, OH USA
[3]Division of Biostatistics and Data Science, Department of Population Health Sciences, Georgia Cancer Center at Augusta University, Augusta, GA USA

Abstract

Motivation

The rapid growth in the number and application of high-throughput "omics" technologies has created a need for better methods to integrate multiomics data sets. Much progress has been made in developing unsupervised methods, but supervised methods have lagged behind.

Results

We develop a novel algorithm, plasma, to train and validate models to predict time-to-event outcomes from multiomics data sets. The model is built on using two layers of the existing partial least squares algorithm to first select components that covary with the outcome in order to construct a joint Cox proportional hazards model. We apply plasma to the lung squamous cell carcinoma (LUSC) data from The Cancer Genome Atlas (TCGA). Our model successfully separates an independent test data set into high risk and low risk patients ($p = 0.0132$). The performance of the joint multiomics model is superior to that of the individual omics data sets. It is also superior to the performance of an approach that uses an unsupervised method (Multi Omics Factor Analysis; MOFA) to find factors that might work as predictors. Many of the factors that contribute strongly to the plasma model can be justified from the biological literature.

Availability and Implementation

The plasma R package can be obtained from The Comprehensive R Archive Network (CRAN) at https://cran.r-project.org/web/packages/plasma/index.html. The latest version of the package can be obtained from R-Forge at https://r-forge.r-project.org/R/?group_id=174. Source code and data for the analysis presented here can be obtained from GitLab, at https://gitlab.com/krcoombes/plasma.

Contact

Email: kcoombes@augusta.edu

Supplementary Information

Supplementary material is available from *Bioinformatics* online.

Introduction

Recent years have seen the development of numerous algorithms and computational packages for the analysis of multiomics data sets. At this point, one can find numerous review articles summarizing progress in the field (Adossa, Khan, Rytkonen, and Elo 2021; Graw et al. 2021; Heo, Hwa, Lee, Park, and An 2021; Picard, Scott-Boyer, Bodein, Perin, and Droit 2021; Reel, Reel, Pearson, Trucco, and Jefferson 2021; Subramanian, Verma, Kumar, Jere, and Anamika 2020; Vlachavas, Bohn, Uckert, and Nurnberg 2021). As with other applications of machine learning, the kinds of problems addressed by these algorithms are divided into two broad categories: unsupervised (e.g., clustering or class discovery) or supervised (including class comparison and class prediction) (Simon and Dobbin 2003). Advances in the area of unsupervised learning have been broader and deeper than advances in supervised learning.

One of the most effective unsupervised methods is Multi-Omic Factor Analysis (MOFA) (Argelaguet et al. 2018, 2020). A key property of MOFA is that it does not require all omics assays to have been performed on all samples under study. In particular, it can effectively discover class structure across omics data sets even when data for many patients have only been acquired on a subset of the omics technologies. As of this writing, we do not know of any supervised multiomics method that can effectively learn to predict outcomes when samples have only been assayed on a subset of the omics data sets.

MOFA starts with a standard method – Latent Factor Analysis – that is known to work well on a single omics data set. It then fits a coherent model that identifies latent factors that are common to, and able to explain the data well in, all the omics data sets under study. Our investigation (unpublished) of the factors found by MOFA suggests that, at least in some cases, it is approximately equivalent to a two-step process:

1. Use principal components analysis to identify initial latent factors in each individual omics data set.
2. For each pair of omics data sets, use overlapping samples to train and extend models of each factor to the union of assayed samples.

That re-interpretation of MOFA suggests that an analogous procedure might work for supervised analyses as well. In this article, we describe a novel two-step algorithm, which we call "*plasma*", to find models that can predict time-to-event outcomes on samples from multiomics data sets even in the presence of incomplete data. We use partial least squares (PLS) for both steps, using Cox regression to learn the single omics models and linear regression to learn how to extend models from one omics data set to another. To illustrate the method, we use the squamous cell lung cancer (LUSC) data set from The Cancer Genome Atlas (TCGA).

Methods

Our computational method is implemented and the data are available in version 1.0.0 of the `plasma` package (https://CRAN.R-project.org/package=plasma).

Data

The results included here are in whole or part based upon data generated by the TCGA Research Network. We downloaded the entire LUSC Level 3 data set (Cancer Genome Atlas Research Network 2017) from the FireBrowse web site (http://firebrowse.org/ (Deng, Bragelmann, Kryukov, Saraiva-Agostinho, and Perner 2017)) on 6 August 2018. We filtered the data sets so that only the most variable, and presumably the most informative, features were retained. To summarize:

1. From TCGA, we obtained 124 columns of clinical, demographic, and laboratory data on 504 patient samples. We removed six samples for whom no outcome data were available, leaving 498 patient samples. We removed any columns that always contained the same value, and any columns that were duplicates of other data in the set. We also removed any columns whose values were missing in more than 25% of the patients. We converted categorical variables into sets of binary variables using one-hot-encoding. We then separated the clinical data into three parts:
    1. Outcome (overall survival)
    2. Binary covariates (61 columns)

   3. Continuous covariates (2 columns)
2. Exome sequencing data for 484 patients with squamous cell lung cancer were obtained as mutation allele format (MAF) files. We removed any gene that was mutated in fewer than 4% of the samples. The resulting data set contained 897 mutated genes.
3. Methylation data for 364 LUSC patients was obtained as beta values computed by the TCGA from Illumina Methylation 450K microarrays. We removed any CpG site for which the standard deviation of the beta values was less than 0.25 or for which the mean was within 0.15 of the boundary values of 0 or 1. The resulting data set contained 1,940 highly variable CpG's.
4. Already normalized sequencing data on 2,588 microRNAs (miRs) was obtained for 338 patients. We removed any miR for which the standard deviation of normalized expression was less than 0.10, which left 872 miRs in the final data set.
5. Already normalized sequencing data on 20,531 mRNAs was obtained in 493 patients. We removed any mRNA whose mean normalized expression was less than 5 or whose standard deviation was less than 1.25. The final data set included 2,290 mRNAs.
6. Normalized expression data from reverse phase protein arrays (RPPA) was obtained from antibodies targeting 223 proteins in 322 patients. All data were retained for further analysis.

Imputation

We imputed missing data for any patient sample assayed in an input data set that yielded incomplete data. The underlying issue is that the PLS models for individual omics data sets will not make predictions on a sample if even one data point is missing. As a result, if a sample is missing at least one data point in every omics data set, then it will be impossible to use that sample at all.

For a range of available methods and R packages, see the CRAN Task View on Missing Data (https://CRAN.R-project.org/view=MissingData). We also recommend the R-miss-tastic web site on missing data (https://rmisstastic.netlify.app/). Their simulations suggest that, for purposes of producing predictive models from omics data, the imputation method is not particularly important. Because of the latter finding, we have implemented two simple imputation methods in the `plasma` package:

1. `meanModeImputer` will replace any missing data by the mean value of the observed data if there are more than five distinct values; otherwise, it will replace missing data by the mode. This approach works for both continuous data and for binary or small categorical data.
2. `samplingImputer` replaces missing values by sampling randomly from the empirical data distribution. For the LUSC data, we used the sampling imputer.

Computational Approach

The `plasma` algorithm is based on Partial Least Squares (PLS), which has been shown to be an effective method for finding components that can predict clinically interesting outcomes (Bastien, Bertrand, Meyer, and Maumy-Bertrand 2015). The workflow of the plasma algorithm is illustrated in **Figure 1** in the case of three omics data sets. First, for each of the omics data sets, we apply the PLS Cox regression algorithm (`plsRcox` Version 1.7.6 (Bertrand and Maumy-Bertrand 2021)) to the time-to-event outcome data to learn separate predictive models (indicated in red, green, and blue, respectively). Each of the individual omics models consists of three kinds of regression models:

- The `plsRcoxmodel` contains the coefficients of the components learned by PLS Cox regression. The number of components is determined automatically as a function of the logarithm of the number of features in the omics data set. The output of this model is a continuous prediction of "risk" for the time-to-event outcome of interest.
- Next, two separate models are constructed using the prediction of risk on the training data.
  - The `riskModel` is a `coxph` model using continuous predicted risk as a single predictor.
  - The `splitModel` is a `coxph` model using a binary split of the risk (at the median) as the predictor.

Each of these models may be incomplete, since they are not defined for patients who have not been assayed (shown in white) using that particular omics technology. Second, for each pair of omics data sets, we apply the

PLS linear regression algorithm (`pls` Version 2.8.1 (Mishra and Liland 2022)) to learn how to predict the Cox regression components from one data set using features from the other data set. This step extends (shown in pastel red, green, and blue, resp.) each of the original models, in different ways, from the intersection of samples assayed on both data sets to their union. Third, we average all of the different extended models (ignoring missing data) to get a single coherent model of components across all omics data sets. Assuming that this process has been applied to learn the model from a training data set, we can evaluate the final Cox regression model on both the training set and a test set of patient samples.

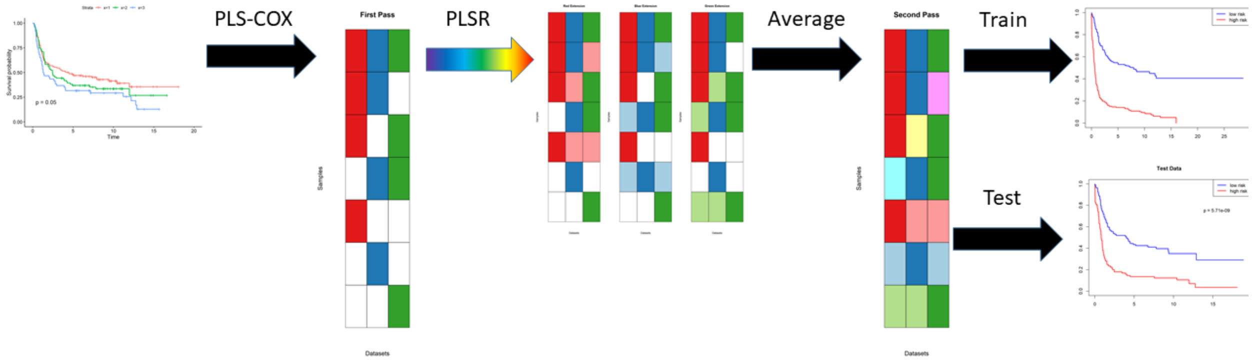

Figure 1: Workflow schematic for plasma algorithm with three omics data sets. See "Computational Approach" in the main text for an explanation.

All computations were performed in R version 4.2.2 (2022-10-31 ucrt) of the R Statistical Software Environment (R Core Team 2022). Cox proportional hazards models for survival analysis were fit using version 3.4.0 of the `survival` R package (Therneau and Grambsch 2000).

Gene Enrichment

In order to interpret the model, we converted feature identifiers into gene names. This task was straightforward for the mutation, mRNASeq, and RPPA data sets. We used annotation files from the Illumina web site to extract associated genes (possibly more than one per locus) from the methylation data set. Gene enrichment (pathway or annotation) analysis was performed by uploading gene lists to ToppGene (https://toppgene.cchmc.org/) (Chen, Bardes, Aronow, and Jegga 2009).

Terminology

Because of the layered nature of the plasma algorithm, we intend to use the following terminology to help clarify the later discussions.

1. The input data contains a list of *omics data sets*.
2. Each omics data set contains measurements of multiple *features*.
3. The first step in the algorithm uses PLS Cox regression to find a set of *components*. Each component is a linear combination of features. The components are used as predictors in a Cox proportional hazards model, which predicts the log hazard ratio as a linear combination of components.
4. The second step in the algorithm creates a secondary layer of components. We do not give these components a separate name. They are not an item of particular focus; we view them as a way to extend the first level components to more samples by “re-interpreting” them in other omics data sets.

Preparing the Data

To be consistent with the `MOFA2` R package (Argelaguet et al. 2020), all of the data sets are arranged so that patient samples are columns and assay features are rows. Each data set includes the same complete set of patients, with columns full of "NA's" to indicate samples that were not assayed.

Split Into Training and Test

We randomly assigned patients to separate training and test sets. We used 60% for training and 40% for testing. **Figure 2** presents a graphical overview of the number of samples (`N`) and the number of features (`D`) in each omics component of the training and test sets.

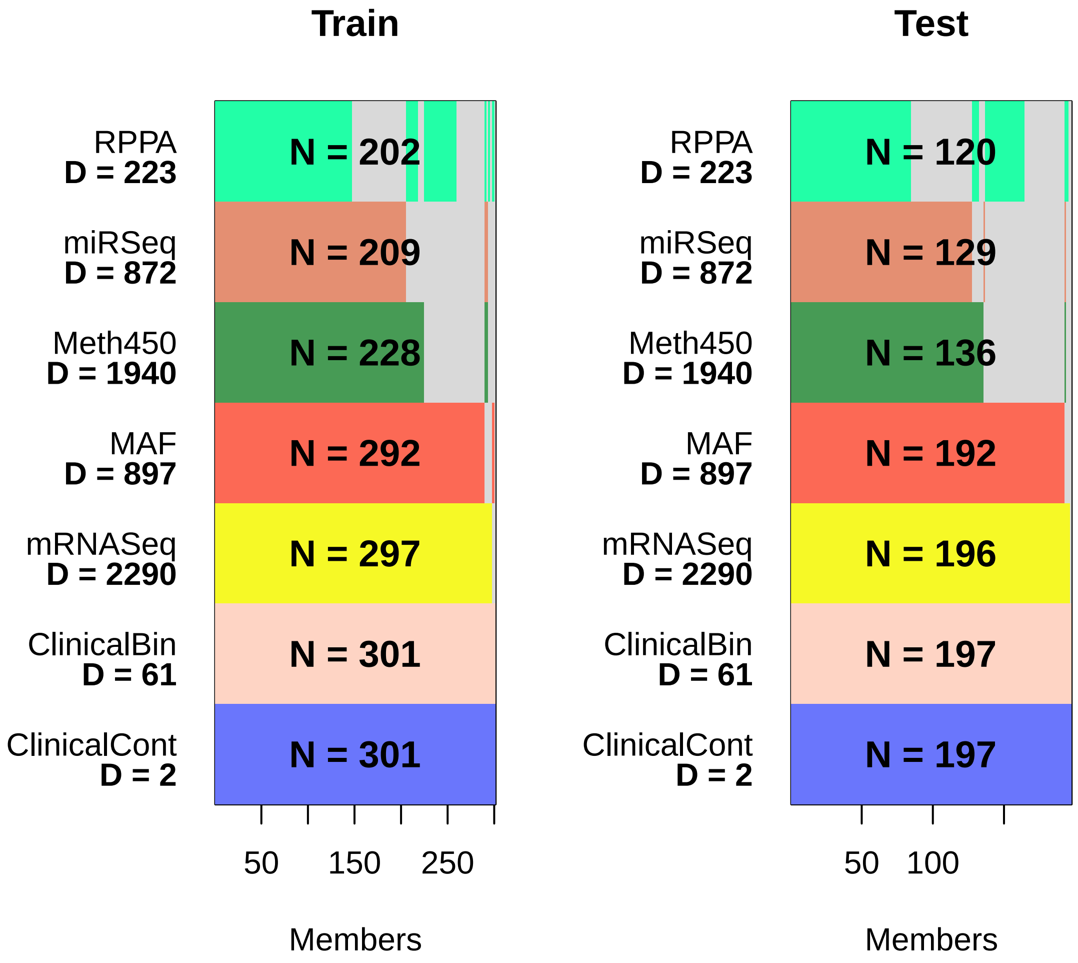

Figure 2: Overview of training and test data. (N is the number of samples in a data set; D is the number of features. RPPA = reverse phase protein arrays; Meth450 = Illumina 450K methylation arrays; MAF = mutation allele format; ClinicalBin = binary clinical features; ClinicalCont = continuous clinical features.)

Results

Individual PLS Cox Regression Models

In the first step of the `plasma` algorithm, we fit PLS Cox models on each omics data set. On the training set, six of the seven contributing data sets are able to find a PLS model that can successfully separate high risk from low risk patients (**Figure 3**). Only the continuous clinical data (which contains just two variables, age at diagnosis and the number of pack-years that the patient smoked) fails to find a useful prognostic model.

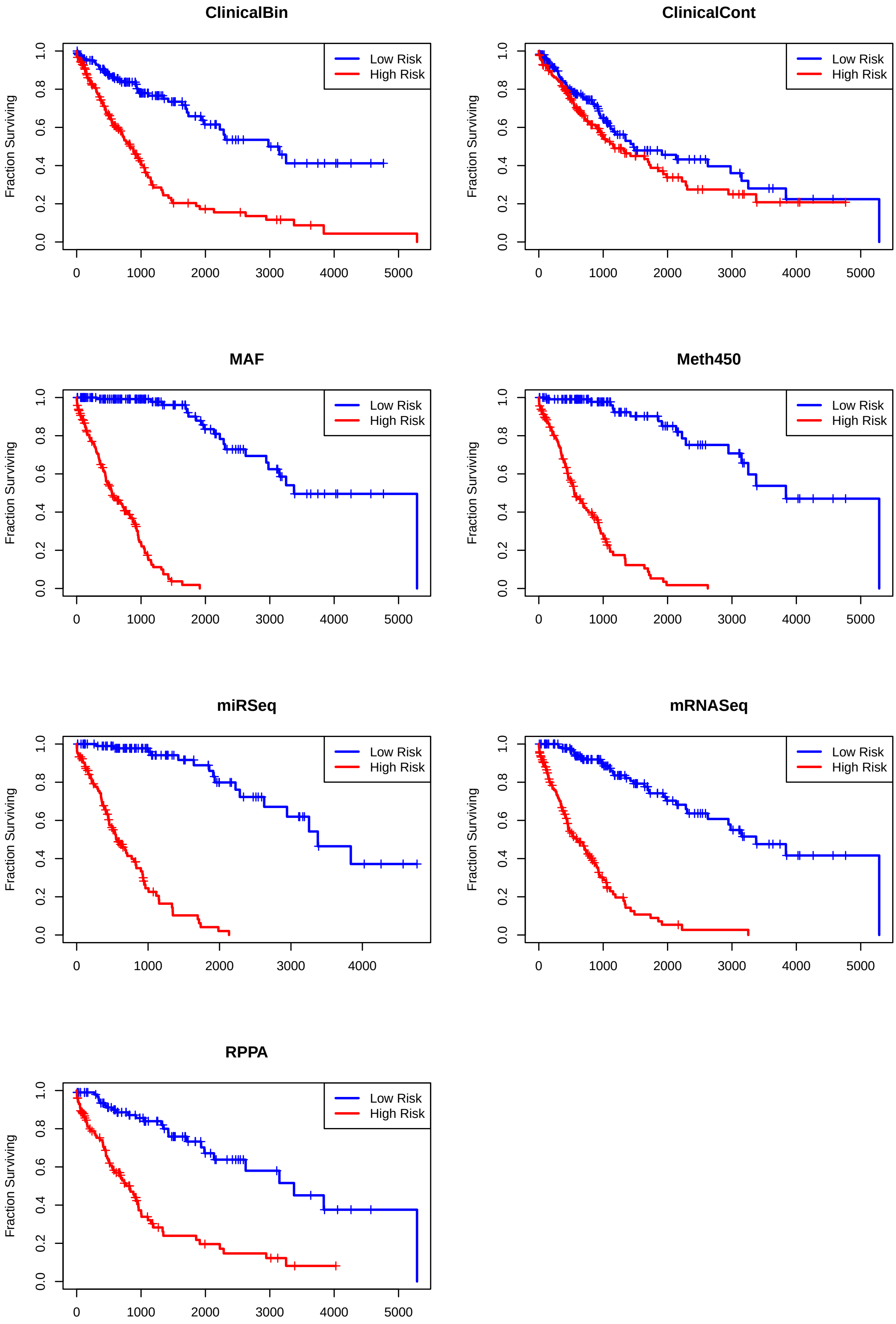

Figure 3: Kaplan-Meier plots of overall survival on the training set from separate PLS Cox omics models.

Unified Model of Overall Survival

The second step of the algorithm is to extend the individual omics-based models across other omics data sets. Since this step is performed for all pairs of data sets, in our case there are seven different sets of predictions of each PLS components. These different predictions are averaged. The structure of complete models created is the same as for the separate, individual omics models. **Figure 4A** shows the final composite Kaplan-Meier plot using the predicted risk, split at the median value, on the training data set.

Validation on an Independent Test Set

We then applied the final composite model to the test set. **Figure 4B** uses the predicted risk, split at the median of the training data, to construct a Kaplan-Meier plot on the test data. The model yields a statistically significant (p = 0.0132) separation of outcomes between the high and low risk patients.

Single Omics Predictions on the Test Set

Next, we compared the test results of the joint omics model to the predictions made on the test data from the separate singleomics models. The Kaplan-Meier plots in **Figure 5** show that most of the individual models have poor performance on the test data, especially when compared to the results of the combined model learned jointly from all the omics data sets.

Interpreting the Model

At this point, our model appears to be a fairly complex black box. We have constructed a matrix of components, based on linear combinations of actual features in different omics data sets. These components are then combined in yet another linear model that predicts the time-to-event outcome through Cox regression. In this section, we want to explore how the individual features from different omics data sets contribute to the models.

Our first step in opening the black box is to realize that not all of the components discovered from the individual omics data sets survived into the final composite model. Some components were eliminated (by stepwise feature selection based on the Akaike Information Criterion) because they appeared to be nearly linearly related to components found in other omics data sets. So, we can examine the final composite model more closely.

Table 1: Table 1: Coefficients of the omics components retained in the final model. (coef = log hazard ratio, exp(coef) = hazard ratio)

	coef	exp(coef)	se(coef)	z	p
ClinicalBin1	-1.846	0.158	0.610	-3.029	0.00245
ClinicalCont1	-1.184	0.306	0.311	-3.810	0.00014
MAF1	1.664	5.282	0.248	6.698	0.00000
MAF2	1.660	5.258	0.266	6.237	0.00000
MAF3	3.153	23.417	0.598	5.277	0.00000
Meth4502	0.789	2.202	0.160	4.935	0.00000
Meth4504	-0.834	0.434	0.183	-4.571	0.00000
miRSeq1	2.676	14.522	0.915	2.924	0.00345
miRSeq3	4.119	61.502	1.089	3.782	0.00016
RPPA1	2.817	16.734	0.635	4.436	0.00001
RPPA2	1.030	2.802	0.258	3.993	0.00007
RPPA3	0.461	1.586	0.215	2.141	0.03229

We see that at least one component discovered from four of the five "true" omics data sets survived in the final model; only the mRNA components failed to make the cut. In addition, one component each from the binary and continous clinical data was retained in the final model.

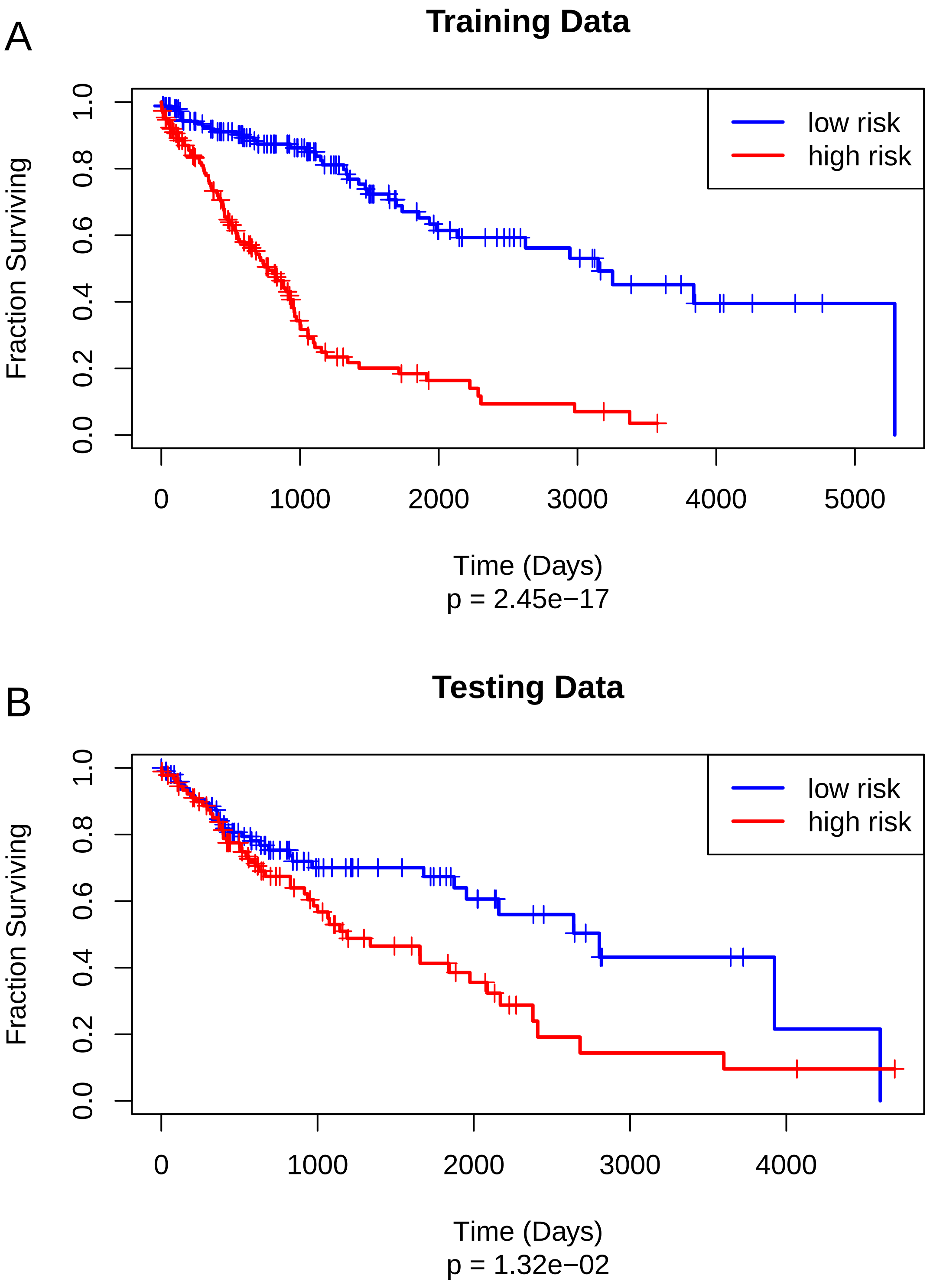

Figure 4: Kaplan-Meier plot of overall survival on (A) the training set and (B) the test set using the unified `plasma` Cox model.

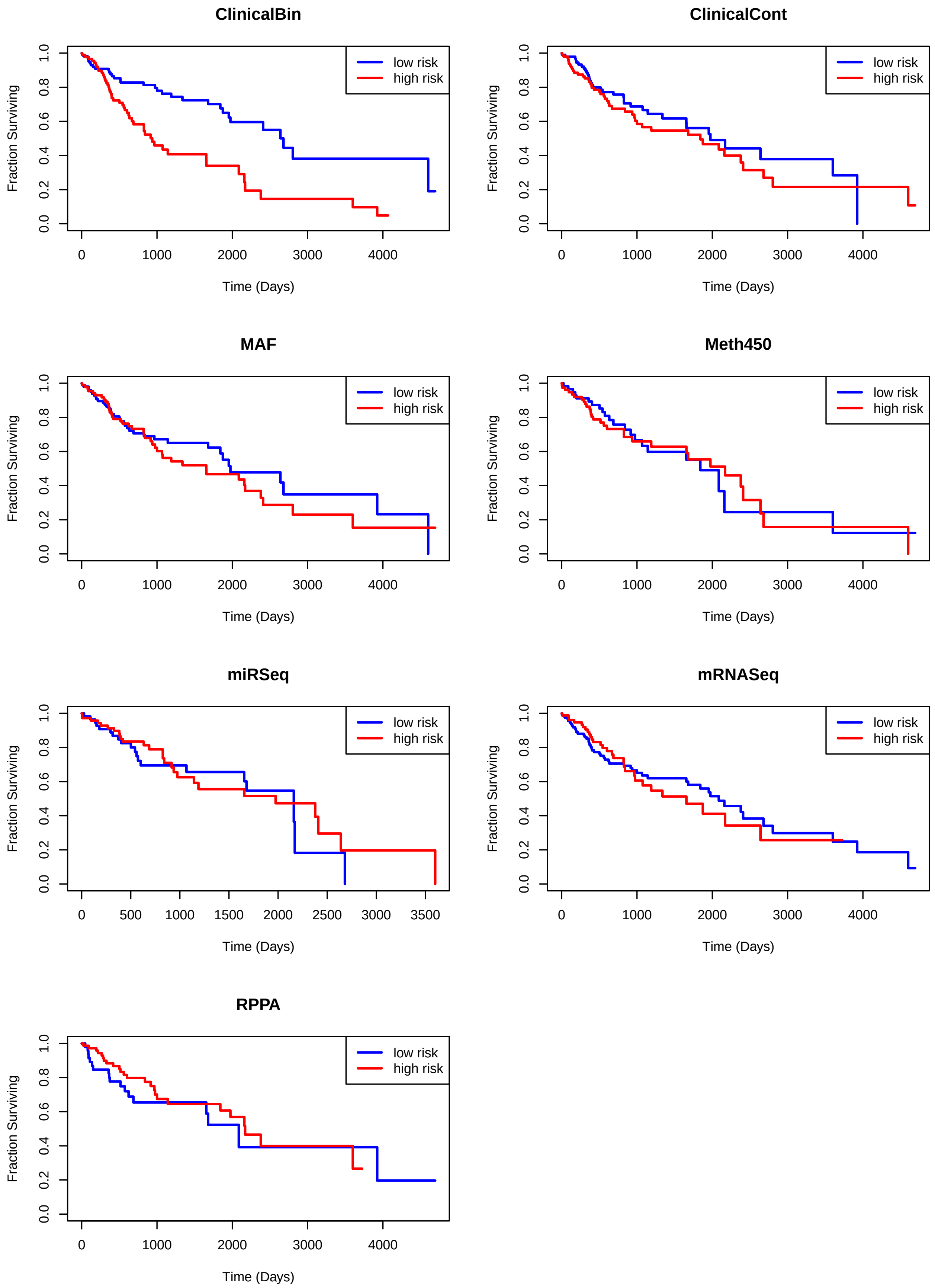

Figure 5: Kaplan-Meier plots of overall survival on the test set from separate PLS Cox omics models.

Final (Composite) Weights

A key point to note is that the core of the plasma model consists of a composition of two levels of linear models. One model starts with individual omics data sets and predicts the "*components*" that we have learned across all data sets. The second model uses the components as predictors in a Cox proportional hazards model of time-to-event (i.e., overall survival) data. The composite of these two steps is yet another linear model. In particular, we can compute the matrix that links the features from the omics-level data sets directly to the final survival outcome. Features from the individual omics data sets with the highest weight in the final model are shown in **Figure 6**.

Gene Enrichment Using ToppGene

For each component, we determined the list of significant genes (with the highest weights) contributing to that component. We uploaded those gene lists to the ToppGene web site (Chen et al. 2009) to perform gene enrichment analyses. We performed a separate gene enrichment analysis using the union of all significant genes from all components. Finally, we performed a similar analysis using the genes with the largest contribution to the final composite model. The negative log10 p-values of the enrichment scores from each ToppGene category were used for two-way clustering of the annotation categories and the components. Results for pathway annotations are presented in **Figure 7**. Results for other ToppGene database categories are shown in **Supplemental Figures S1-S4**.

Discussion

One of the most interesting aspects of this study is that the plasma algorithm successfully discovers a model on the training set that generalizes to produce statistically significant results on the test set (**Figure 4**). However, it accomplishes this goal even though most of the individual omics models do not generalize well to the data set. We can think of three possible reasons for this performance.

1. "*Wisdom of the crowd.*" There has long been an idea in the machine learning field that combining ensembles of weak models can give rise to a strong model. Well-established examples of this idea are bagging and boosting (Kotsiantis 2011; Maclin and Opitz 1997; Sutton 2005)
2. "*Out of phase.*' ' Each omics data set may overfit the model in a different way. Instead of reinforcing each other, the extent of overfitting may cancel out.
3. "*Feature Elimination.*" The combined method successfully identifies useful predictive factors. So, we are still able to fit a generalizable Cox model on the final components.

More research will need to be performed on a variety of data sets to determine whether this phenomenon is more general.

The interpretation of the model (**Figure 6**) in terms of which features contribute to the final predictive model supports the idea that plasma has found a biologically sensible model. By examining the clinical binary features, we see that a continued history of smoking and tumor presence at the time of sample collection are associated with greater hazard ratio, as expected. Interestingly, we also find that tumors in the lower lobes of the lung (ICD-10 C34.3) are less deadly than tumors in the upper lobes (ICD 10 C34.1). Previous literature on the effects of location have been variable and inconsistent, but a large-scale meta-analysis came to the opposite conclusion, and attributed the differences to a differential response to surgery or radiotherapy (Lee, Lee, and Park 2018).

Several of the mutated genes supporting the model also have expected hazard ratios. For example, mutations in *KMT2D* are positively associated with a higher hazard ratio. *KMT2D* is one of the more commonly mutated genes in lung squamous cell carcinoma; according to a recent study (Pan et al. 2023), the loss of *KMT2D* is thought to upregulate RTK-RAS signaling. The second highest-weighted gene, *DNAH5*, codes for a protein that is found in the microtubule-associated protein complex responsible for creating cilia found in airways. Mutations in this gene are associated with primary ciliary dyskinesia (Omran et al. 2000), which predisposes affected individuals to chronic respiratory infections. Notably, there have been no statistically significant associations found between primary ciliary dyskinesia and lung cancer (Tilley, Walters, Shaykhiev,

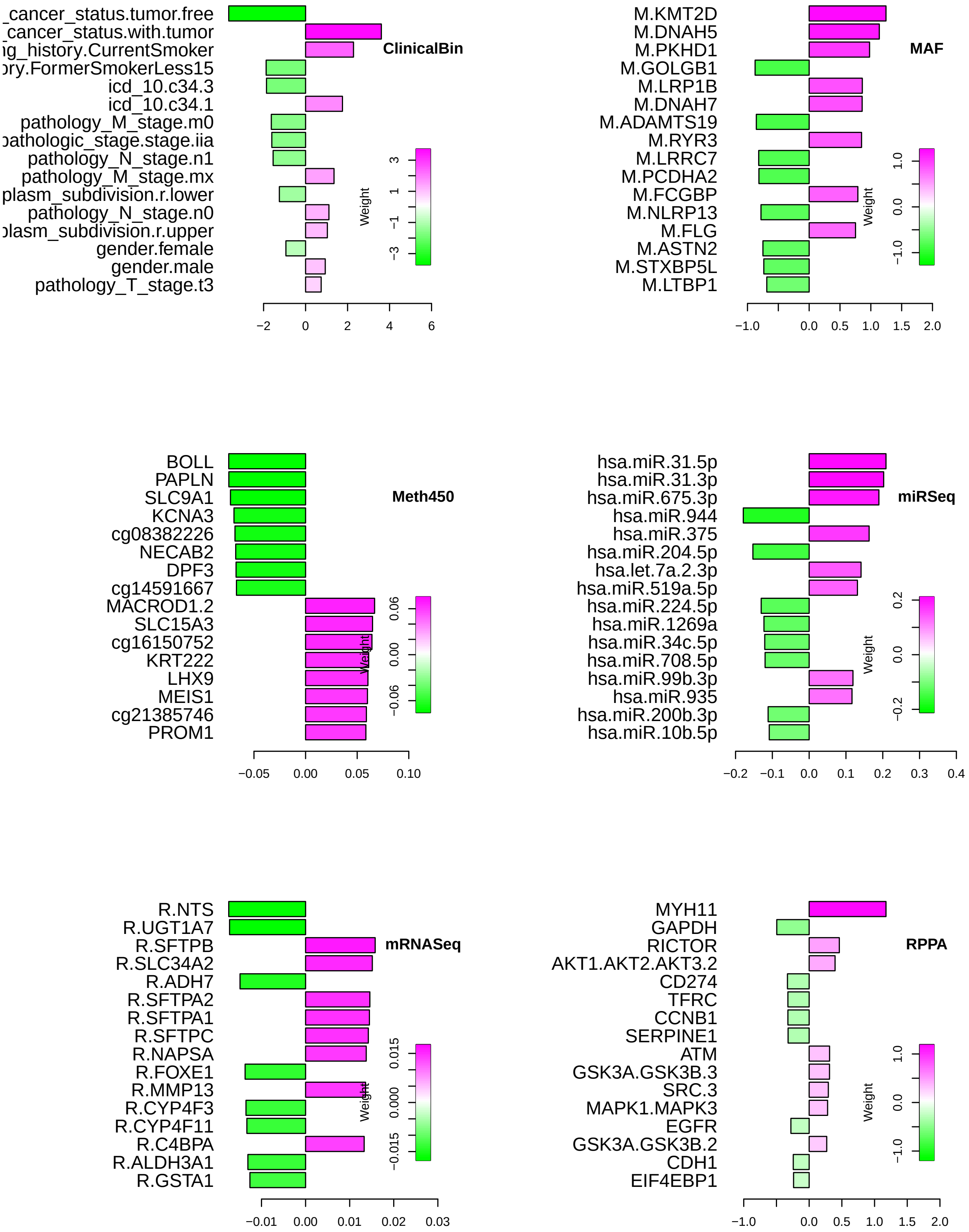

Figure 6: Highly weighted features in the final Cox model. Positive weights (magenta) indicate increaded hazard ratio; negative weights (green) indicate decreased hazard ratio. Prefix M = MAF; prefix R = mRNASeq.

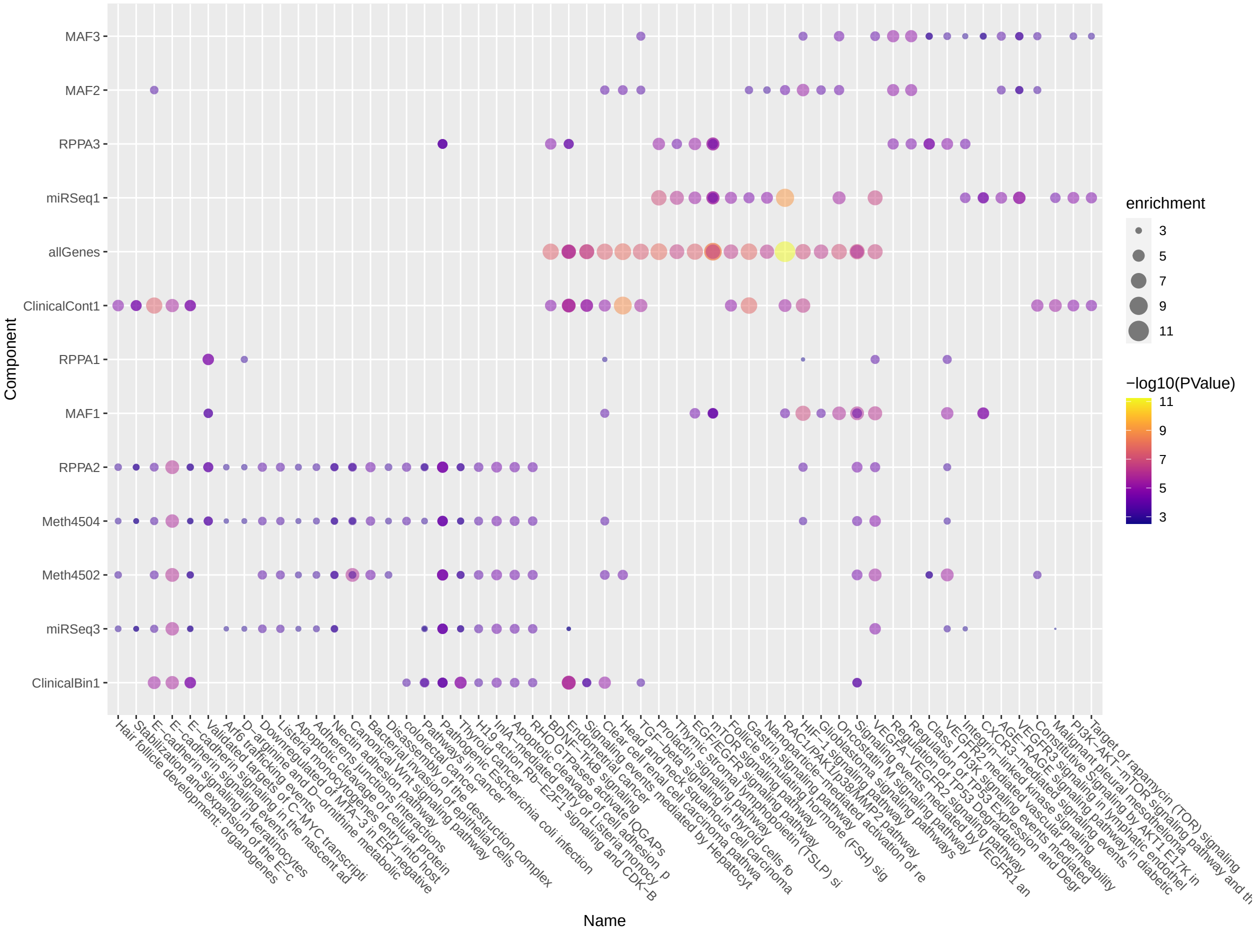

Figure 7: Clustering components based on pathway annotation p-values.

and Crystal 2015). Smoking, which is known to significantly raise lung cancer risk, can also damage and shorten cilia (Cao, Chen, Dong, Xie, and Liu 2020). Because our model is trained on lung cancer survival and not lung cancer onset, the role of *DNAH5* mutations may be that they are insufficient to significantly increase the onset of cancer, but may facilitate cancer progression. It is also worth noting that a TRA2B-DNAH5 fusion was found in 3% of squamous cell lung cancer patients, and this fusion protein was found to promote cancer progression by an ERK1/2-matrix metalloproteinase 1 (MMP1) signaling axis (Li et al. 2016). Further studies would need to be done in order to elucidate how *DNAH5* could contribute to the lung cancer progression through loss or gain of function mutations.

The highest-weight microRNAs in the model correspond to hsa-miR-31, which is associated in our data with shorter survival. This miR is known to promote tumor growth and correlates with low survival in previous studies (Davenport et al. 2021; Edmonds et al. 2016; Tan et al. 2011). By contrast, hsa-miR-34 is associated in our model with longer survival. It is known to be a tumor suppressor in lung cancer (Zhang, Liao, and Tang 2019).

In the reverse-phase protein array data set, we observe a large weight from MYH11, which has been studied in acute myeloid leukemia (AML) in the context of a CBFB-MYH11 fusion gene that is associated with good survival. It is largely unstudied in the realm of LUSC and may present a novel avenue for further exploration. Both RICTOR and AKT have relatively high weights, and are associated with increased hazards ratio. RICTOR/mTORC2 and AKT are known to be upregulated in certain cancers. Overexpression of RICTOR in a non-small cell lung cancer patient-derived xenograft mouse model has been shown to promote tumor growth (Kim, Rhee, and Chen 2020). The AKT/MTOR signaling pathway is dysregulated in many kinds of cancers, and is associated with both metastasis and poor prognosis in lung cancer (Lu et al. 2020). RICTOR/mTORC2 and AKT are also known to be upregulated in insulin-resistant patients, and an association was found between increased fasting serum insulin levels and incidence of lung cancer (Argirion, Weinstein, Männistö, Albanes, and Mondul 2017).

The results of the pathway annotation using ToppGene (**Figure 7**) gives us insight into the biological characteristics of the PLS components that were found by the final `plasm` model to be significantly related to OS. In figure 7, we see that some components from the same dataset are similar in enrichment pattern; for example MAF2 and MAF3 and RPPA1 and RPPA2, which may indicate that these components are highly correlated with each other. There are a few interesting proteins mentioned in some of these terms, such as NRF2, ABCA3, CSF2RA/B. In healthy cells, NRF2 is a transcription factor that combats oxidative stress; ABCA3 is expressed in alveolar type II pneumocytes, and is required for surfactant synthesis. The receptors of CSF2 are essential for proper surfactant metabolism.

The horizontal axis of Figure 7 can be roughly partitioned into extracellular matrix (ECM), metabolism, and surfactant-associated ToppGene terms. Using this rough partioning scheme, we see that MAF1, RPPA1, and RPPA2 forms a group enriched for metabolism, specifically dysfunction of surfactant metabolism. We see that ClinicalBin1 and Meth4502 group together in being enriched for metabolism-related terms but not surfactant-related terms. We see that miRSeq is quite unique in being highly enriched for ECM-related terms like collagen. Above this, the components miRSeq1, miRSeq3, MAF2, MAF3, ClinicalCont1, and Meth4504 are defined by a combination of the aforementioned terms with a high enrichment for terms related to fibrosis and keratinization.

In addition to grouping the components together in terms of patterns of ToppGene enrichments, we thought it may be beneficial to perform enrichment analysis on the union of highly weighted genes from all PLASMA components (allGenes) and on the set of genes that are highly weighted in the final composite Cox regresson model to predict OS (finalGenes). The enrichment pattern from allGenes, which appears to be similar to that of miRSeq1, is enriched for pathway terms related to the ECM (collagen, wound healing, etc.). The enrichment pattern from finalGenes, which appears similar to that of MAF1, RPPA1, and RPPA2, is enriched for pathway terms that show some contextual overlap with allGenes in terms related to fibrosis, but have additional enrichments in surfactant metabolism dysfunction. It has been shown that smoking causes keratinization of the lung, and that lung cancers with a keratinizing phenotype have poor prognosis compared to cancers with a non-keratinizing phenotype (Park et al. 2017). Considering that cytokeratin is considered a negative marker for fibroblasts, the keratinization signal may be coming from the lung-tumor itself (Shiga et al. 2015) whereas the ECM signal may be coming from activated fibroblasts subtypes that were extracted alongside the tumor during the lung biopsy and are known to promote cancer progression and metastasis through fibrosis and ECM remodeling (Kalluri 2016). Overall, our ToppGene results likely point to the contribution of the tumor microenvironment with the tumor itself in affecting patient survival.

As a comparison, we have also performed `MOFA` (Argelaguet et al. 2018, 2020) using the same training and testing data (**Supplementary Material**). Although the factors from `MOFA` are defined such that the first factor, Factor 1, accounts for the greatest variance in the model, the factors may or may not be significantly associated with the outcome, and a post-hoc survival analysis would need was done to assess this. It may be the case that some factors, although they are significantly associated with outcome, account for very small variance in the final `MOFA` model, which hinders interpretability. This was the case with the TCGA-LUSC dataset, in which, when 10 factors were learned from the `MOFA` model, only Factor 8 was significantly associated with OS in the multivariate Cox Proportional Hazards model with and without stepwise feature selection (**Tables S3** and **S4**).

In summary, we have identified a method analogous to that of `MOFA` that allows us to combine different omics data. A major difference between `MOFA` and `plasma` is that while the `MOFA` model learns "factors" that are composites of the variables in an unsupervised fashion, the `plasma` model learns "components" that are composites of the variables in a supervised fashion using the outcome response variables. Because the `plasma` components are learned in a way that maximizes the covariance in the predictors and the response, these components will be automatically associated with the outcome. This could be advantageous in that dissecting the weights associated with the components would yield a list of variables from different omics datasets that contribute the most to defining the outcome, and any additional analyses could be refined by looking at these high-weighted variables most closely.

Funding

This work was supported by a collaboration grant from the Department of Biomedical Informatics at The Ohio State University and by startup funds from the Ohio State University Comprehensive Cancer Center and from the Georgia Cancer Center at Augusta University.

Conflict of Interest: None declared.

References

Adossa N, Khan S, Rytkonen KT, Elo LL (2021). "Computational Strategies for Single-Cell Multi-Omics Integration." *Comput Struct Biotechnol J*, **19**, 2588–2596. https://doi.org/10.1016/j.csbj.2021.04.060.

Argelaguet R, Arnol D, Bredikhin D, Deloro Y, Velten B, Marioni JC, Stegle O (2020). "MOFA+: A Statistical Framework for Comprehensive Integration of Multi-Modal Single-Cell Data." *Genome Biol*, **21**(1), 111. https://doi.org/10.1186/s13059-020-02015-1.

Argelaguet R, Velten B, Arnol D, Dietrich S, Zenz T, Marioni JC, Buettner F, Huber W, Stegle O (2018). "Multi-Omics Factor Analysis-a Framework for Unsupervised Integration of Multi-Omics Data Sets." *Mol Syst Biol*, **14**(6), e8124. https://doi.org/10.15252/msb.20178124.

Argirion I, Weinstein SJ, Männistö S, Albanes D, Mondul AM (2017). "Serum Insulin, Glucose, Indices of Insulin Resistance, and Risk of Lung Cancer." *Cancer Epidemiology, Biomarkers & Prevention: A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology*, **26**(10), 1519–1524. https://doi.org/10.1158/1055-9965.EPI-17-0293.

Bastien P, Bertrand F, Meyer N, Maumy-Bertrand M (2015). "Deviance Residuals-Based Sparse PLS and Sparse Kernel PLS Regression for Censored Data." *Bioinformatics*, **31**(3), 397–404. https://doi.org/10.1093/bioinformatics/btu660.

Bertrand F, Maumy-Bertrand M (2021). "Fitting and Cross-Validating Cox Models to Censored Big Data With Missing Values Using Extensions of Partial Least Squares Regression Models." *Front Big Data*, **4**, 684794. https://doi.org/10.3389/fdata.2021.684794.

Cancer Genome Atlas Research Network (2017). "Integrated Genomic Characterization of Oesophageal Carcinoma." *Nature*, **541**(7636), 169–175. https://doi.org/10.1038/nature20805.

Cao Y, Chen M, Dong D, Xie S, Liu M (2020). "Environmental Pollutants Damage Airway Epithelial Cell Cilia: Implications for the Prevention of Obstructive Lung Diseases." *Thoracic Cancer*, **11**(3), 505–510. https://doi.org/10.1111/1759-7714.13323.

Chen J, Bardes EE, Aronow BJ, Jegga AG (2009). "ToppGene Suite for Gene List Enrichment Analysis and Candidate Gene Prioritization." *Nucleic Acids Res*, **37**(Web Server Issue), W305–11. https://doi.org/10.1093/nar/gkp427.

Davenport ML, Echols JB, Silva AD, Anderson JC, Owens P, Yates C, Wei Q, Harada S, Hurst DR, Edmonds MD (2021). "miR-31 Displays Subtype Specificity in Lung Cancer." *Cancer Research*, **81**(8), 1942–1953. https://doi.org/10.1158/0008-5472.CAN-20-2769.

Deng M, Bragelmann J, Kryukov I, Saraiva-Agostinho N, Perner S (2017). "FirebrowseR: An R Client to the Broad Institute's Firehose Pipeline." *Database (Oxford)*, **2017**. https://doi.org/10.1093/database/baw160.

Edmonds MD, Boyd KL, Moyo T, Mitra R, Duszynski R, Arrate MP, Chen X, Zhao Z, Blackwell TS, Andl T, et al. (2016). "MicroRNA-31 Initiates Lung Tumorigenesis and Promotes Mutant KRAS-driven Lung Cancer." *The Journal of Clinical Investigation*, **126**(1), 349–364. https://doi.org/10.1172/JCI82720.

Graw S, Chappell K, Washam CL, Gies A, Bird J, Robeson MS, Byrum SD (2021). "Multi-Omics Data Integration Considerations and Study Design for Biological Systems and Disease." *Mol Omics*, **17**(2), 170–185. https://doi.org/10.1039/d0mo00041h.

Heo YJ, Hwa C, Lee GH, Park JM, An JY (2021). "Integrative Multi-Omics Approaches in Cancer Research: From Biological Networks to Clinical Subtypes." *Mol Cells*, **44**(7), 433–443. https://doi.org/10.14348/molcells.2021.0042.

Kalluri R (2016). "The Biology and Function of Fibroblasts in Cancer." *Nature Reviews. Cancer*, **16**(9), 582–598. https://doi.org/10.1038/nrc.2016.73.

Kim LC, Rhee CH, Chen J (2020). "RICTOR Amplification Promotes NSCLC Cell Proliferation Through Formation and Activation of mTORC2 at the Expense of mTORC1." *Molecular Cancer Research: MCR*, **18**(11), 1675–1684. https://doi.org/10.1158/1541-7786.MCR-20-0262.

Kotsiantis S (2011). "Combining Bagging, Boosting, Rotation Forest and Random Subspace Methods." *Artificial Intelligence Review*, **35**(3), 223–240. https://doi.org/10.1007/s10462-010-9192-8.

Lee HW, Lee C-H, Park YS (2018). "Location of Stage I–III Non-Small Cell Lung Cancer and Survival Rate: Systematic Review and Meta-Analysis." *Thoracic Cancer*, **9**(12), 1614–1622. https://doi.org/10.1111/1759-7714.12869.

Li F, Fang Z, Zhang J, Li C, Liu H, Xia J, Zhu H, Guo C, Qin Z, Li F, et al. (2016). "Identification of TRA2B-DNAH5 Fusion as a Novel Oncogenic Driver in Human Lung Squamous Cell Carcinoma." *Cell Research*, **26**(10), 1149–1164. https://doi.org/10.1038/cr.2016.111.

Lu J, Zang H, Zheng H, Zhan Y, Yang Y, Zhang Y, Liu S, Feng J, Wen Q, Long M, et al. (2020). "Overexpression of p-Akt, p-mTOR and p-eIF4E Proteins Associates with Metastasis and Unfavorable Prognosis in Non-Small Cell Lung Cancer." *PLOS ONE*, **15**(2), e0227768. https://doi.org/10.1371/journal.pone.0227768.

Maclin R, Opitz D (1997). "An Empirical Evaluation of Bagging and Boosting." In *Proceedings of the fourteenth national conference on artificial intelligence and ninth conference on Innovative applications of artificial intelligence* 546–551. AAAI Press, Providence, Rhode Island.

Mishra P, Liland KH (2022). "Swiss Knife Partial Least Squares (SKPLS): One Tool for Modelling Single Block, Multiblock, Multiway, Multiway Multiblock Including Multi-Responses and Meta Information Under the ROSA Framework." *Anal Chim Acta*, **1206**, 339786. https://doi.org/10.1016/j.aca.2022.339786.

Omran H, Häffner K, Völkel A, Kuehr J, Ketelsen UP, Ross UH, Konietzko N, Wienker T, Brandis M, Hildebrandt F (2000). "Homozygosity Mapping of a Gene Locus for Primary Ciliary Dyskinesia on Chromosome 5p and Identification of the Heavy Dynein Chain DNAH5 as a Candidate Gene." *American Journal of Respiratory Cell and Molecular Biology*, **23**(5), 696–702. https://doi.org/10.1165/ajrcmb.23.5.4257.

Pan Y, Han H, Hu H, Wang H, Song Y, Hao Y, Tong X, Patel AS, Misirlioglu S, Tang S, et al. (2023). "KMT2D Deficiency Drives Lung Squamous Cell Carcinoma and Hypersensitivity to RTK-RAS Inhibition." *Cancer Cell*, **41**(1), 88–105.e8. https://doi.org/10.1016/j.ccell.2022.11.015.

Park HJ, Cha Y-J, Kim SH, Kim A, Kim EY, Chang YS (2017). "Keratinization of Lung Squamous Cell Carcinoma Is Associated with Poor Clinical Outcome." *Tuberculosis and Respiratory Diseases*, **80**(2), 179–186. https://doi.org/10.4046/trd.2017.80.2.179.

Picard M, Scott-Boyer MP, Bodein A, Perin O, Droit A (2021). "Integration Strategies of Multi-Omics Data for Machine Learning Analysis." *Comput Struct Biotechnol J*, **19**, 3735–3746.

https://doi.org/10.1016/j.csbj.2021.06.030.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Reel PS, Reel S, Pearson E, Trucco E, Jefferson E (2021). "Using Machine Learning Approaches for Multi-Omics Data Analysis: A Review." *Biotechnol Adv*, **49**, 107739. https://doi.org/10.1016/j.biotechadv.2021.107739.

Shiga K, Hara M, Nagasaki T, Sato T, Takahashi H, Takeyama H (2015). "Cancer-Associated Fibroblasts: Their Characteristics and Their Roles in Tumor Growth." *Cancers*, **7**(4), 2443–2458. https://doi.org/10.3390/cancers7040902.

Simon RM, Dobbin K (2003). "Experimental Design of DNA Microarray Experiments." *Biotechniques*, **Suppl**, 16–21.

Subramanian I, Verma S, Kumar S, Jere A, Anamika K (2020). "Multi-Omics Data Integration, Interpretation, and Its Application." *Bioinform Biol Insights*, **14**, 1177932219899051. https://doi.org/10.1177/1177932219899051.

Sutton CD (2005). "Classification and Regression Trees, Bagging, and Boosting." *Handbook of Statistics*, **24**, 303–329. https://doi.org/10.1016/S0169-7161(04)24011-1.

Tan X, Qin W, Zhang L, Hang J, Li B, Zhang C, Wan J, Zhou F, Shao K, Sun Y, et al. (2011). "A 5-MicroRNA Signature for Lung Squamous Cell Carcinoma Diagnosis and Hsa-miR-31 for Prognosis." *Clinical Cancer Research*, **17**(21), 6802–6811. https://doi.org/10.1158/1078-0432.CCR-11-0419.

Therneau TM, Grambsch PM (2000). *Modeling Survival Data: Extending the Cox Model*. (K. Dietz, M. Gail, K. Krickeberg, J. Samet, and A. Tsiatis, Eds.). Springer, New York, NY. https://doi.org/10.1007/978-1-4757-3294-8.

Tilley AE, Walters MS, Shaykhiev R, Crystal RG (2015). "Cilia Dysfunction in Lung Disease." *Annual Review of Physiology*, **77**, 379–406. https://doi.org/10.1146/annurev-physiol-021014-071931.

Vlachavas EI, Bohn J, Uckert F, Nurnberg S (2021). "A Detailed Catalogue of Multi-Omics Methodologies for Identification of Putative Biomarkers and Causal Molecular Networks in Translational Cancer Research." *Int J Mol Sci*, **22**(6). https://doi.org/10.3390/ijms22062822.

Zhang L, Liao Y, Tang L (2019). "MicroRNA-34 Family: A Potential Tumor Suppressor and Therapeutic Candidate in Cancer." *Journal of Experimental & Clinical Cancer Research*, **38**(1), 53. https://doi.org/10.1186/s13046-019-1059-5.