

# ASAH1 rGBM Manuscript ELISA Figure

Justin Sing, University of Toronto

Wed January 04 04:24:38 PM EST 2023

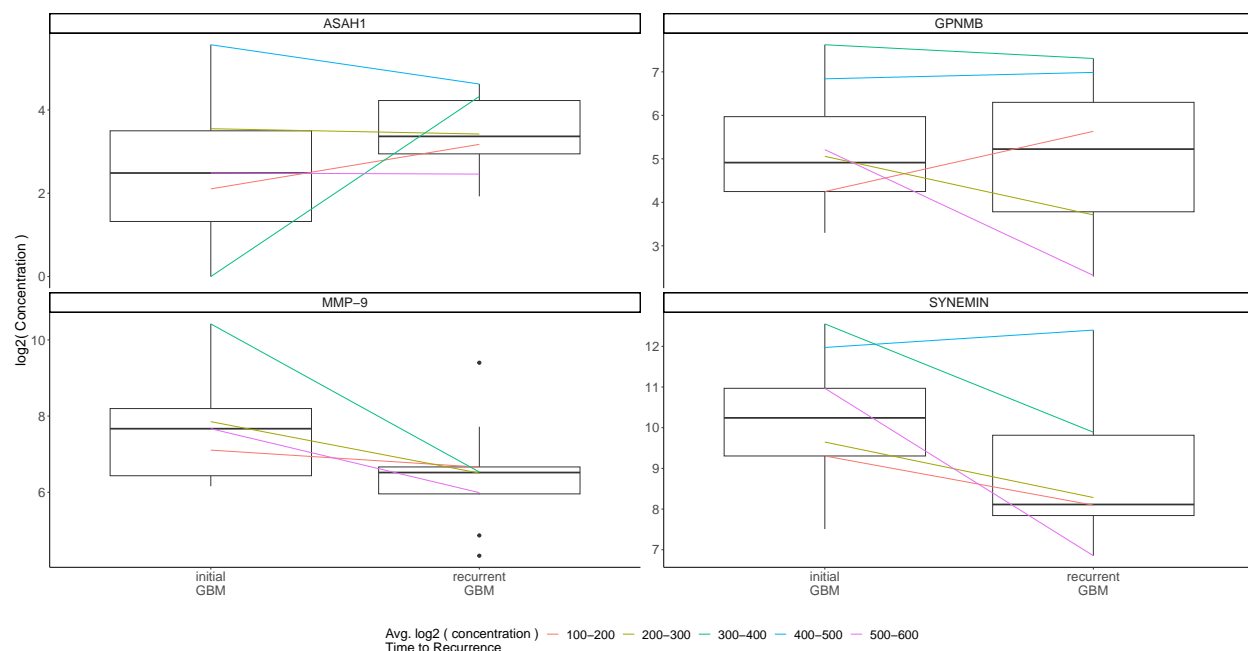
## Contents

<b>Subsetting and hold out patient for Validating Ratios</b>	<b>3</b>
<b>Comparison of Using All patients or Subsetting for Unique Patients per Class</b>	<b>4</b>
<b>Master Figure</b>	<b>6</b>
<b>Supplemental Figures and Tables</b>	<b>8</b>
<b>Methods</b>	<b>17</b>
Statistical Test for Comparison of Log2 Distributions . . . . .	17
Dimensionality Reduction . . . . .	17
Prediction Analysis . . . . .	17
<b>References</b>	<b>17</b>

```

##
## Call:
## lm(formula = Time.to.reccurence..days. ~ ASAH1_delta + SYNEMIN_delta +
##      GPNMB_delta + MMP.9_delta, data = ttr_delta)
##
## Residuals:
##      1      2      3      4      5      6      7      8
##  58.150 -60.848 -22.659  -8.495 -104.509 -52.616  55.801 -11.958
##      9
## 147.133
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    170.74     55.27   3.089  0.0366 *
## ASAH1_delta     10.28     29.43   0.349  0.7446
## SYNEMIN_delta  -12.01     26.90  -0.446  0.6785
## GPNMB_delta    -37.57     19.36  -1.941  0.1243
## MMP.9_delta    -26.99     38.94  -0.693  0.5263
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 107.5 on 4 degrees of freedom
## Multiple R-squared:  0.6284, Adjusted R-squared:  0.2569
## F-statistic: 1.691 on 4 and 4 DF, p-value: 0.3116
##
## Analysis of Variance Table
##
## Response: Time.to.reccurence..days.
##              Df Sum Sq Mean Sq F value Pr(>F)
## ASAH1_delta    1    172      172  0.0149 0.9088
## SYNEMIN_delta  1  33047   33047  2.8572 0.1662
## GPNMB_delta    1  39473   39473  3.4128 0.1384
## MMP.9_delta    1   5557    5557  0.4805 0.5263
## Residuals      4  46265   11566
##
## [1] "The estimated sample size needed for 80% power is: 29"

```



## Subsetting and hold out patient for Validating Ratios

I still think there is value in using the ratios, so I performed the sPLS-DA on the ratio data again, but instead of using all the ratios and inverse ratios from all patients, I do the following:

1. Randomly sample a single patient to use as a test patient, to make predictions on
2. Randomly sample  $n$  random patients from remaining patients and compute iGBM/rGBM ratios
3. For the rest of the patients I compute rGBM/iGBM ratios

This results in the following data used in the sPLS-DA model

### Training Data

Patient ID	Class	ASAH1	SYNMEIN	GPNMB	MMP-9
6	iGBM	-4.3219281	2.6629410	0.3162515	3.8886097
10	rGBM	2.2103599	-3.8779037	2.5628539	-3.1538943
16	rGBM	1.6956066	0.8506559	0.1659252	-1.9814540
23	rGBM	-0.0257631	-3.6121574	1.6151451	-1.7166787
26	iGBM	1.7744403	0.4268049	4.2786146	-1.5475206
29	rGBM	3.3645724	0.4959143	-0.6652013	0.5074684
31	iGBM	-1.6599246	-2.8570919	-1.9497226	-1.3115642
34	iGBM	-0.1768778	0.7865737	2.1875247	1.1604031

### Test Data

Patient ID	Class	ASAH1	SYNMEIN	GPNMB	MMP-9
35	rGBM	-0.02599521	-4.115453	-2.892362	-1.679008

This way we do not duplicate patient IDs (i.e. having iGBM/rGBM and rGBM/iGBM for the same patient) and we have a single patient we can make a prediction on to see where in the PLS-DA dimension they would

land.

## Comparison of Using All patients or Subsetting for Unique Patients per Class

Lets compare to see if it makes a difference whether we use all patients for both iGBM/rGBM and rGBM/iGBM classes vs only using a unique set of patients for iGBM/rGBM and rGBM/iGBM (i.e. no duplicate patient in each class)

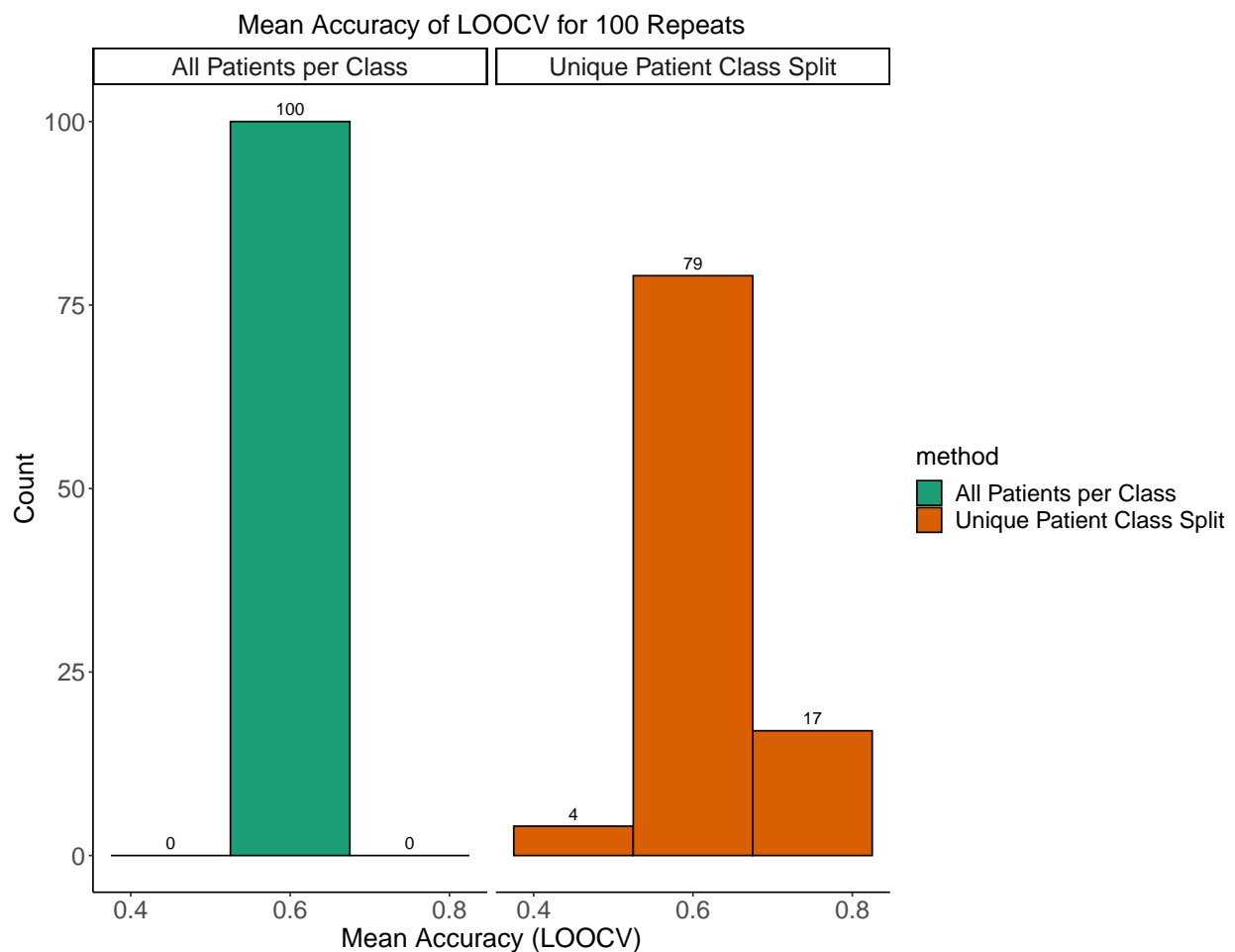
1. I perform a leave one out cross validation, leaving one patient out and using the rest of the data for training.
2. The left out patient gets using for assessing model performance accuracy.
3. The mean accuracy of each left out sample is computed
4. Steps 1-3 is repeated 100 times to assess the variation in mean accuracy of the LOOCV
5. 1-4 is performed for both cases of using all patients and using unique subset patients

```
##      2.5%      97.5%
```

```
## 0.4444444 0.7777778
```

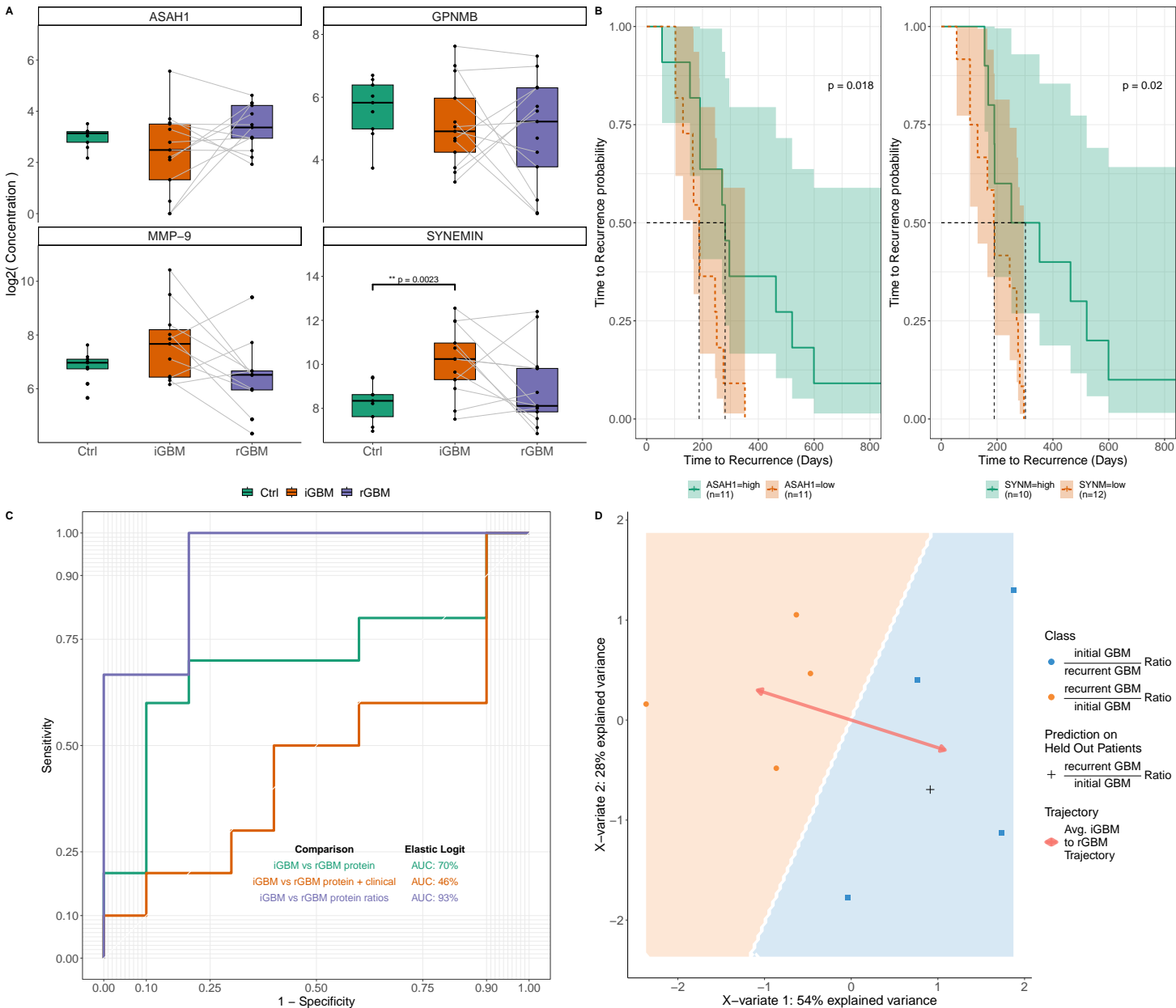
```
##      2.5%      97.5%
```

```
## 0.6666667 0.6666667
```



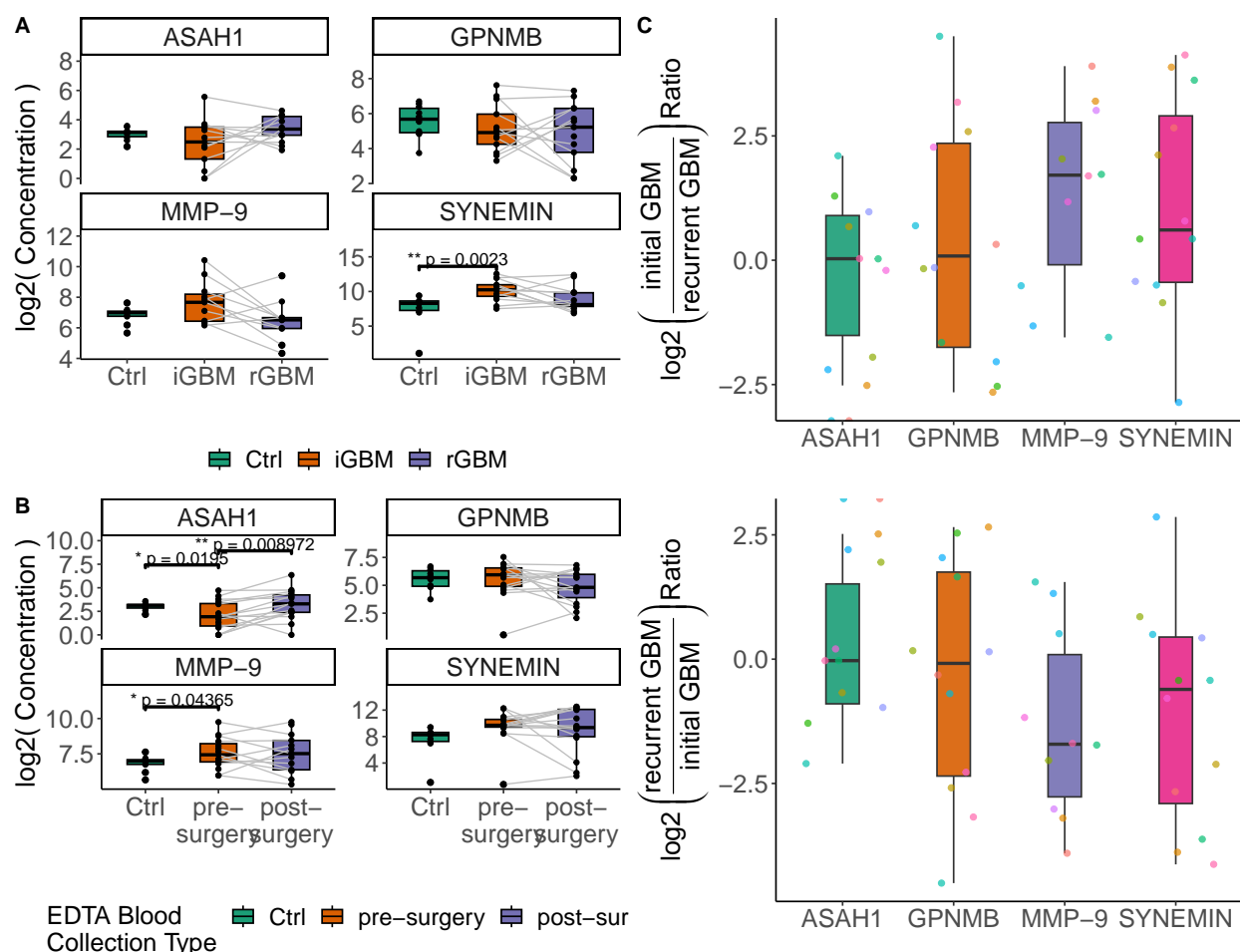
Justin Sing, University of Toronto

# Master Figure



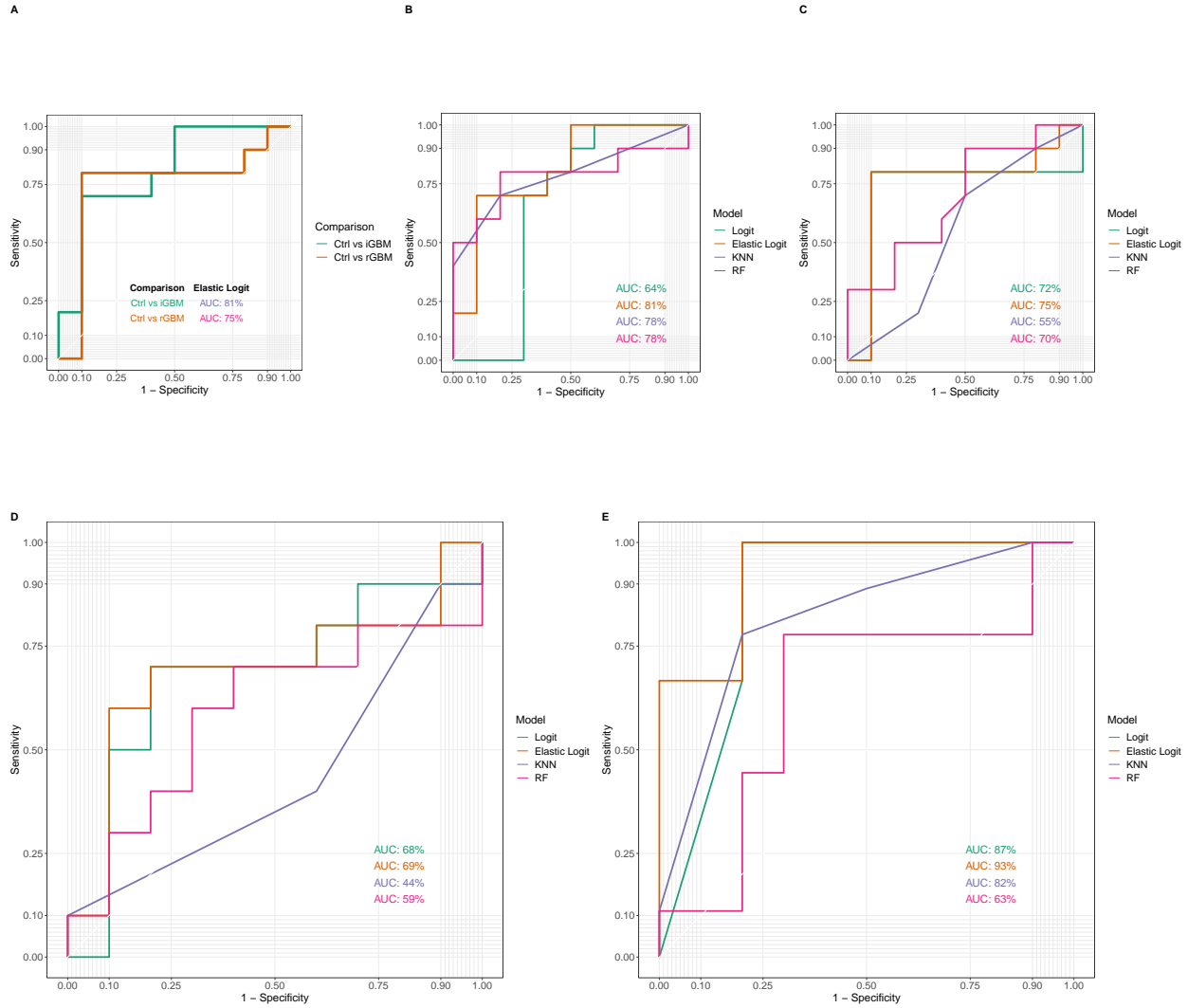
**Figure 3: A Promising Plasma Proteomics Signature for Stratification and Prediction of Recurrent Glioblastoma Multiforme.** Concentration of four proteins identified from TMT-MS tissue proteomics, were measured in patients' plasma samples using Enzyme-Linked Immunosorbent Assay (ELISA). **(a)** The log2 transformed protein concentration measurements demonstrates the difference in log2 concentration distributions between control, initial GBM and paired recurrent GBM populations. Note, a patient was omitted due to being an outlier, see Supp Figure X for full set of patients. **(b)** Kaplan-Meyer plots showing demonstrating high concentration of ASAH1 and high concentration of SYNEMIN contributing to a higher probability of tumor recurrence occurring at a later date. **(c)** The receiver operating characteristic (ROC) shows the potential of using plasma proteomics for separating each population, driven mostly from information contained in the proteins. **(d)** Reducing the data of the four proteins ratios demonstrates an average trajectory and separation of  $\frac{\text{initial GBM}}{\text{recurrent GBM}}$  vs  $\frac{\text{recurrent GBM}}{\text{initial GBM}}$ . One patient was randomly left out as a test patient (black + point), 4 samples were randomly sampled to represent  $\frac{\text{initial GBM}}{\text{recurrent GBM}}$  (blue points) and remaining 4 samples were used to represent  $\frac{\text{recurrent GBM}}{\text{initial GBM}}$ .

## Supplemental Figures and Tables

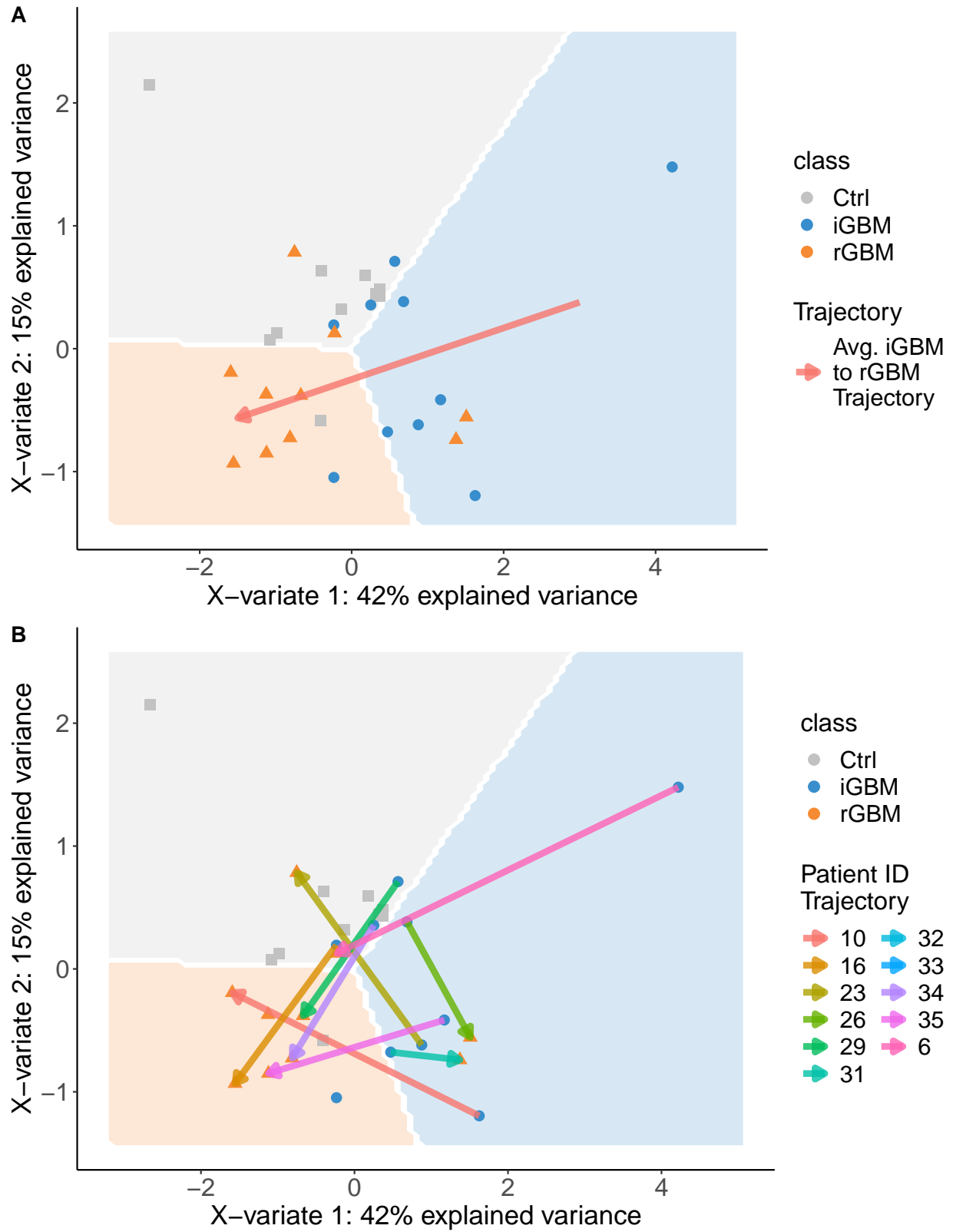


**Supplemental Figure 1 : Distribution of Pre- vs Post-Surgery Plasma Proteomics and Ratios of Initial and Recurrent GBM Plasma Proteomics.** (a) The  $\log_2$  transformed protein concentration measurements demonstrates the difference in  $\log_2$  concentration distributions between control, initial GBM and paired recurrent GBM population for all patients. (b) The  $\log_2$  transformed protein concentration measurements demonstrates the difference in  $\log_2$  concentration distributions between control, pre-surgery and post-surgery populations. (c) Demonstrates the  $\log$  transformed  $\frac{\text{initial GBM}}{\text{recurrent GBM}}$  ratio (top) as well as the inverse ratio (bottom) for the potential use of identifying a patients state to recurrence using these ratios.



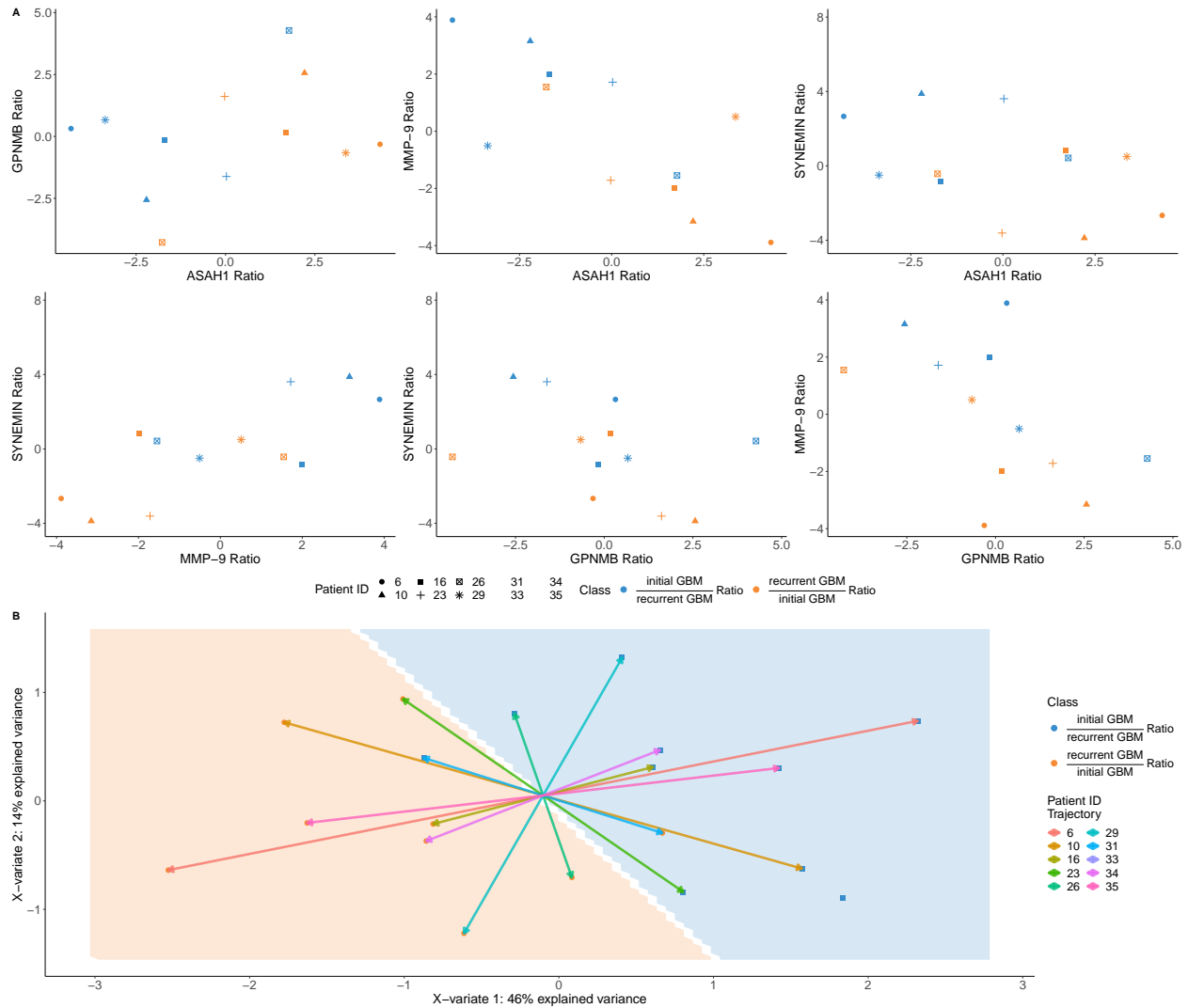


**Supplemental Figure 2 : Variation in Predictive Performance using Different Models.** Separation on Controls and testing parametric (logistic regression, logistic regression with elastic net regularization), non-parametric (k-nearest neighbours) and ensemble (random forest) supervised learning algorithms on different prediction tasks. (a) The receiver operating characteristic (ROC) shows the potential of using plasma proteomics for separating controls (health patients) from patients with GBM using an elastic logistic regression. (b) Most models have AUCs greater than 60% when predicting control vs initial GBM. (c) Most models have comparable AUCs, except for k-nearest neighbours when predicting for initial GBM vs recurrent GBM. (d) Including clinical data for *Tumor localization*, *Sex*, *Age at surgery (years)*, *Type of resection* reduces predictive performance for all models. (e) All models, except random forest, have AUCs above 80% when predicting initial GBM vs recurrent GBM using protein ratios.



**Supplemental Figure 3 : Individual Patient Trajectory in Low Dimensional Space.** (a) A sparse partial least squares discriminant analysis shows separation between the different population groups, as well as an overall trajectory of a initial GBM patient going towards a recurrent GBM patient in a low

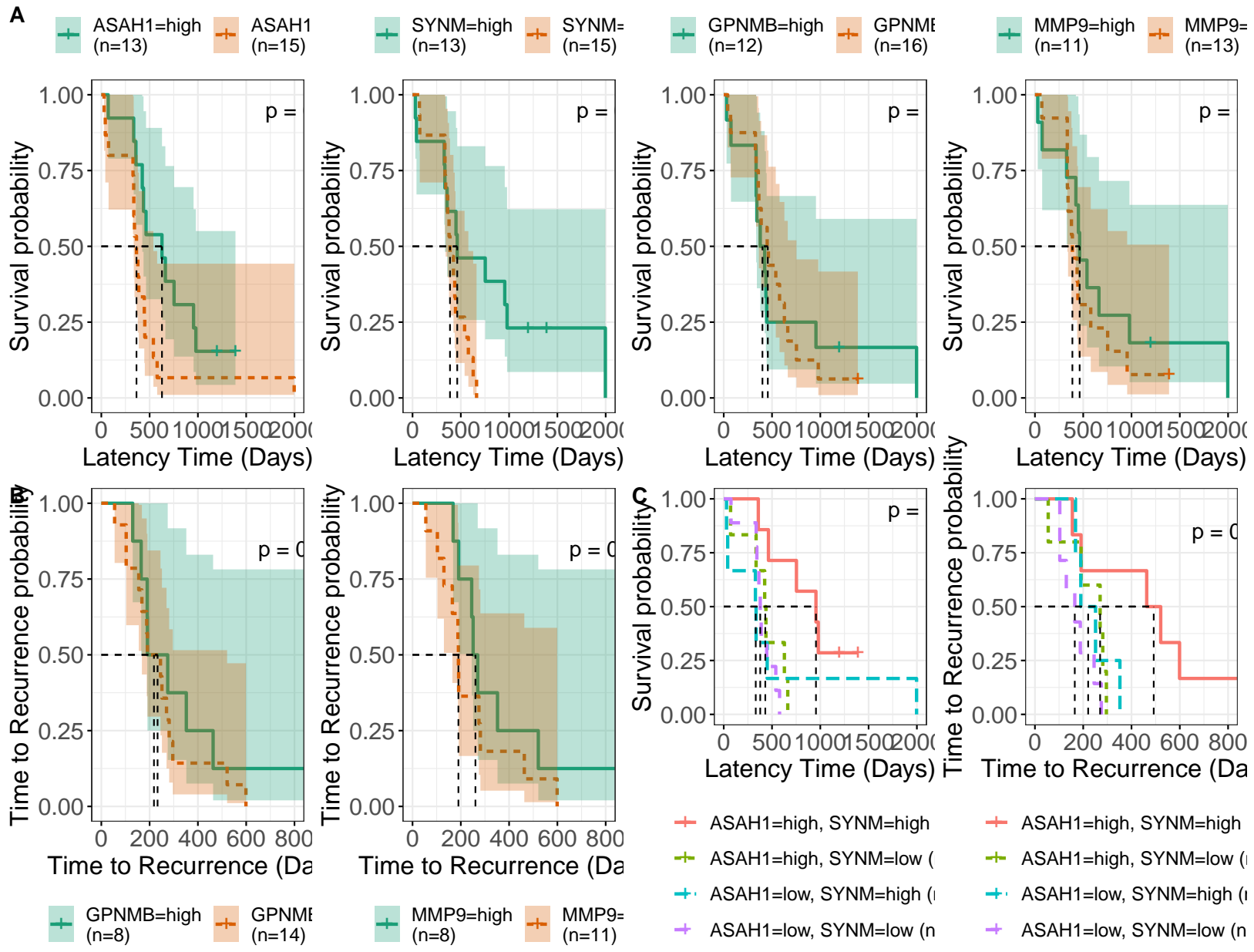
dimensional space. **(b)** Individual trajectories demonstrate the individual variance of patients path to recurrence in a low-dimensional space.



**Supplemental Figure 4 : Information Contained in Plasma Proteomic Ratios for Initial GBM vs Recurrent GBM Stratification.** (a) Pairwise protein ratios shows promise of individual protein-pair ratios being able to stratify initial vs recurrent plasma GBM samples. (b) Individual trajectories demonstrate the individual variance of patients path to recurrence in a low-dimensional space using ratios.

Justin Sing, University of Toronto

13



Supple-

**mental Figure 5 : Survival and Time to Recurrence Analysis.** (a) Kaplan-Meyer plots showing demonstrating high concentration of ASAH1 and high concentration of SYNEMIN contributing to a higher probability of tumor recurrence occurring at a later date. (b) Kaplan-Meyer plots demonstrating the combination of both ASAH1 and SYNEMIN and their contribution of survival probability and time to recurrence.

### Logistic Regression with Elastic Net Regularization Performance Metrics

Comparison	Sensitivity	Specificity	Pos Pred Value	Neg Pred Value	Balanced Accuracy
Ctrl vs iGBM	70%	70%	70%	70%	70%
Ctrl vs rGBM	80%	80%	80%	80%	80%
iGBM vs rGBM protein	70%	70%	70%	70%	70%
iGBM vs rGBM protein + clinical	50%	50%	20%	80%	50%
iGBM vs rGBM protein ratios	80%	88.89%	88.89%	80%	84.44%

Supplemental Table I : Logistic Regression with Elastic Net Regularization Performance Metrics

### Statistical Comparisons Between Population Groups per Genes

Comparison	Statistical Test <sup>1</sup>	ASAHI <sup>2</sup>	GPNMB <sup>2</sup>	MMP-9 <sup>2</sup>	SYNEMIN <sup>2</sup>
<b>Control vs Pre-Surgery vs Post-Surgery Plasma Samples</b>					
Ctrl vs post- surgery	Permutation Test	0.7476000	0.1693500	0.1244000	0.2567000
Ctrl vs pre- surgery	Permutation Test	0.0195000	0.9946500	0.0436500	0.0625000
pre- surgery vs post- surgery	Paired Permutation Test	0.0089720	0.2421880	0.4755860	0.5427860
<b>Control vs Initial GBM vs Recurrent GBM Plasma Samples</b>					
Ctrl vs iGBM	Permutation Test	0.1880500	0.4017500	0.0797000	0.0023000
Ctrl vs rGBM	Permutation Test	0.2551500	0.2481500	0.4577500	0.1374500
iGBM vs rGBM	Paired Permutation Test	0.1992188	0.6123047	0.1171875	0.1054688

<sup>1</sup>Permutation test using t-test statistics for non-paired samples, or paired permutation test using pair t-test statistics for paired samples.

<sup>2</sup>Values in cells represent the non-adjusted p-value from the corresponding statistical test.

Supplemental Table II : Statistical Comparisons Between Population Groups per Genes



## Methods

### Statistical Test for Comparison of Log2 Distributions

Log2 transformed distributions of protein concentrations were compared pairwise for control, initial GBM, and recurrent GBM samples for each protein, ASAH1, GNMPB, MMP-9 and SYNEMIN. Statistical p-values were computed using a permutation test, that utilizes t-test statistics but performs random permutations of the data for computing a p-value. Note, for paired sample data, initial GBM vs recurrent GBM, a paired permutation test was performed instead. Visualization and statistical analyses were performed in R (v4.1.2) and RStudio (v2021.09.2+382), permutation tests were performed using broman (v0.80).

### Dimensionality Reduction

To visualize the samples plasma proteomics on a different dimension, we applied a sparse Partial Least Squares Discriminant Analysis (sPLS-DA), which maximizes the covariance between the explanatory and response variable(s). This was performed on the log2 transformed protein concentrations, as well as the log2 transformed ratios ( $\frac{\text{initial GBM}}{\text{recurrent GBM}}$  vs  $\frac{\text{recurrent GBM}}{\text{initial GBM}}$ ) of each protein. Individual trajectories are displayed as vector arrows connecting initial GBM patient samples to their corresponding recurrent GBM patient samples, in the sPLS-DA space using the first two components that contain the most amount of explained variance. An average trajectory direction can be computed by taking the average of the individual trajectory vectors, to visualize an overall trajectory of initial GBM state to recurrent GBM state. To visualize decision boundaries for class separation (control, initial GBM and recurrent GBM samples), we applied a maximum distance approach. [1] Visualization and statistical analyses were performed in R (v4.1.2) and RStudio (v2021.09.2+382), sPLS-DA analysis was performed using mixOmics (v6.18.1).

### Prediction Analysis

Prediction tasks were performed using parametric (logistic regression, logistic regression with elastic net regularization), non-parametric (k-nearest neighbours) and ensemble (random forest) supervised learning algorithms. General predictive tasks were performed on: controls vs initial GBM samples, controls vs recurrent GBM samples, initial GBM samples vs recurrent GBM samples. Data used in prediction tasks are log2 transformed protein concentrations, log2 transformed protein concentrations supplemented with clinical data (*Tumor localization, Sex, Age at surgery (years), Type of resection*), and log2 transformed ratios of  $\frac{\text{initial GBM}}{\text{recurrent GBM}}$  vs  $\frac{\text{recurrent GBM}}{\text{initial GBM}}$  protein measurements. Model training and prediction performances are performed on the entire data set using leave-one-out cross-validation due to the small sample size. Visualization and statistical analyses were performed in R (v4.1.2) and RStudio (v2021.09.2+382), prediction analyses were performed using caret (v6.0-93).

## References

1. Rohart, F, B Gautier, A Singh, and K-A Lê Cao. 2017. "MixOmics: An R Package for 'Omics Feature Selection and Multiple Data Integration." *PLoS Computational Biology* 13 (11). Cold Spring Harbor Labs Journals.