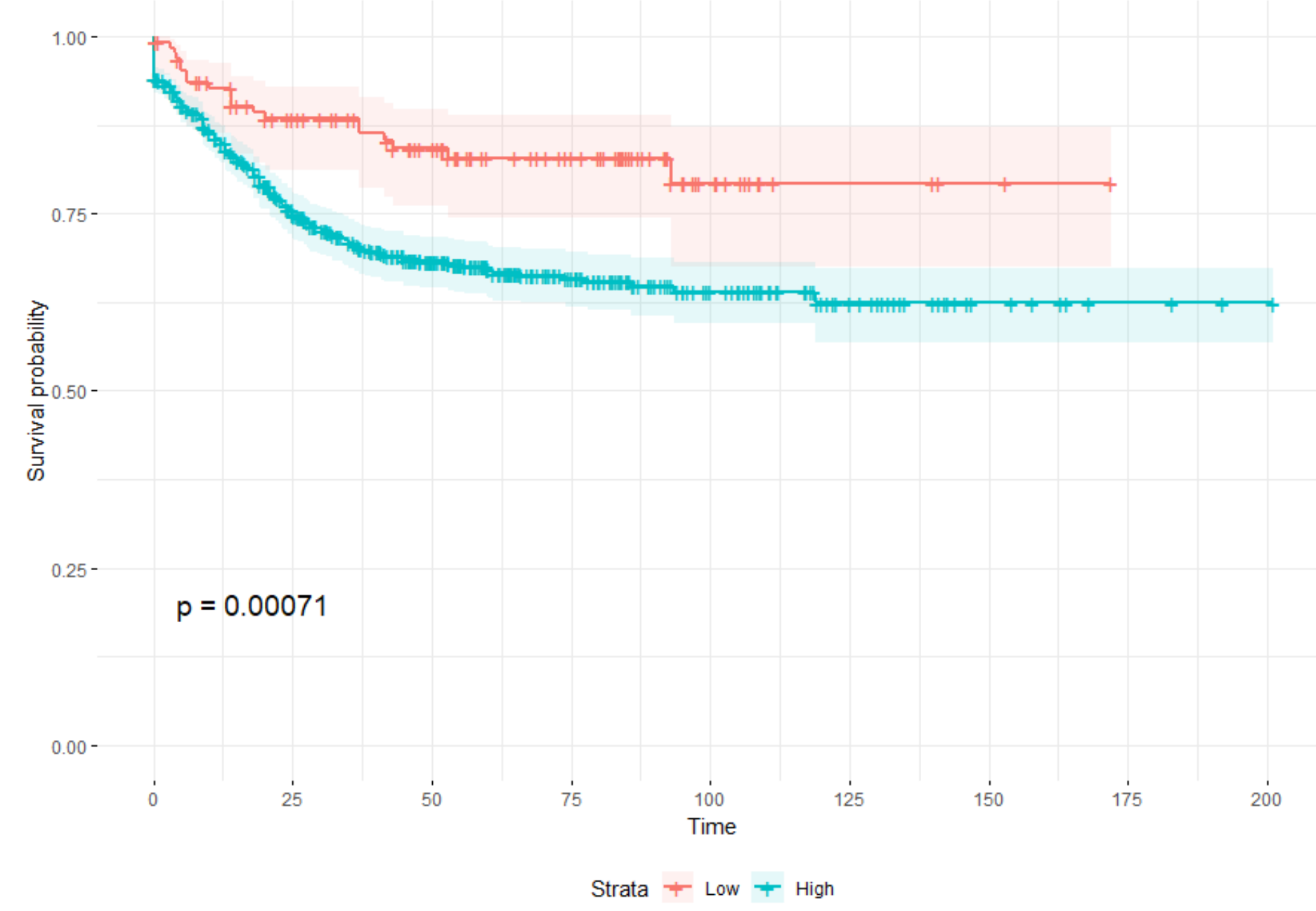
A Novel *In Silico* Approach to Identify Gene Signatures Associated with Recurrent Cancer

Kirsten Wohlars[1,2], Chris Yoo[2], Suwon Kim[3]
[1]Cornell University, Ithaca, NY
[2]Systems Imagination, Inc., Scottsdale, AZ
[3]Department of Basic Medical Sciences, University of Arizona College of Medicine, Phoenix, AZ

Introduction

Colorectal cancer is the fourth most prevalent cancer type in the United States by number of new cases in 2018 and the second leading cause of all cancer-related death [1]. Downregulation of inhibitor of growth factor 4 (ING4) has been correlated with increased metastasis and disease recurrence as well as many other indicators of cancer progression such as tumor size, histological grade, depth of serosa infiltration, and microvessel density in colon cancer [2]. Using a big data analysis approach, we determined that ING4 expression stratified recurrent patients (Figure 1).

Figure 1: Kaplan-Meier Curve for Patient Stratification by ING4 ROC-value



We next explored the prognostic value of the ING4 downstream target genes previously reported [3]. We analyzed three independent gene expression profile data sets sourced from the NCBI GEO Database. The combined cohort consisted of 866 patients with colorectal cancer, 243 of whom experienced recurrence. All four cancer stages were represented in the sample cohort (Table I).

Table I: Summary of patient clinical stage, cohort size and disease free survival for all data sets

	GSE17538	GSE38832	GSE39582	Full Data Set
Cancer Stage				
Stage I	28	18	37	83
Stage II	70	35	267	372
Stage III	75	39	206	320
Stage IV	27	30	60	117
Cancer-Related Death				
Number of Deaths	55	9	179	243
Number of Patients	200	92	574	866
% Death	27.5%	9.78%	31.18%	28.06%

Methods

Figure 2: Boxplot of GAPDH normalized gene expression for recurrent vs. non-recurrent groups

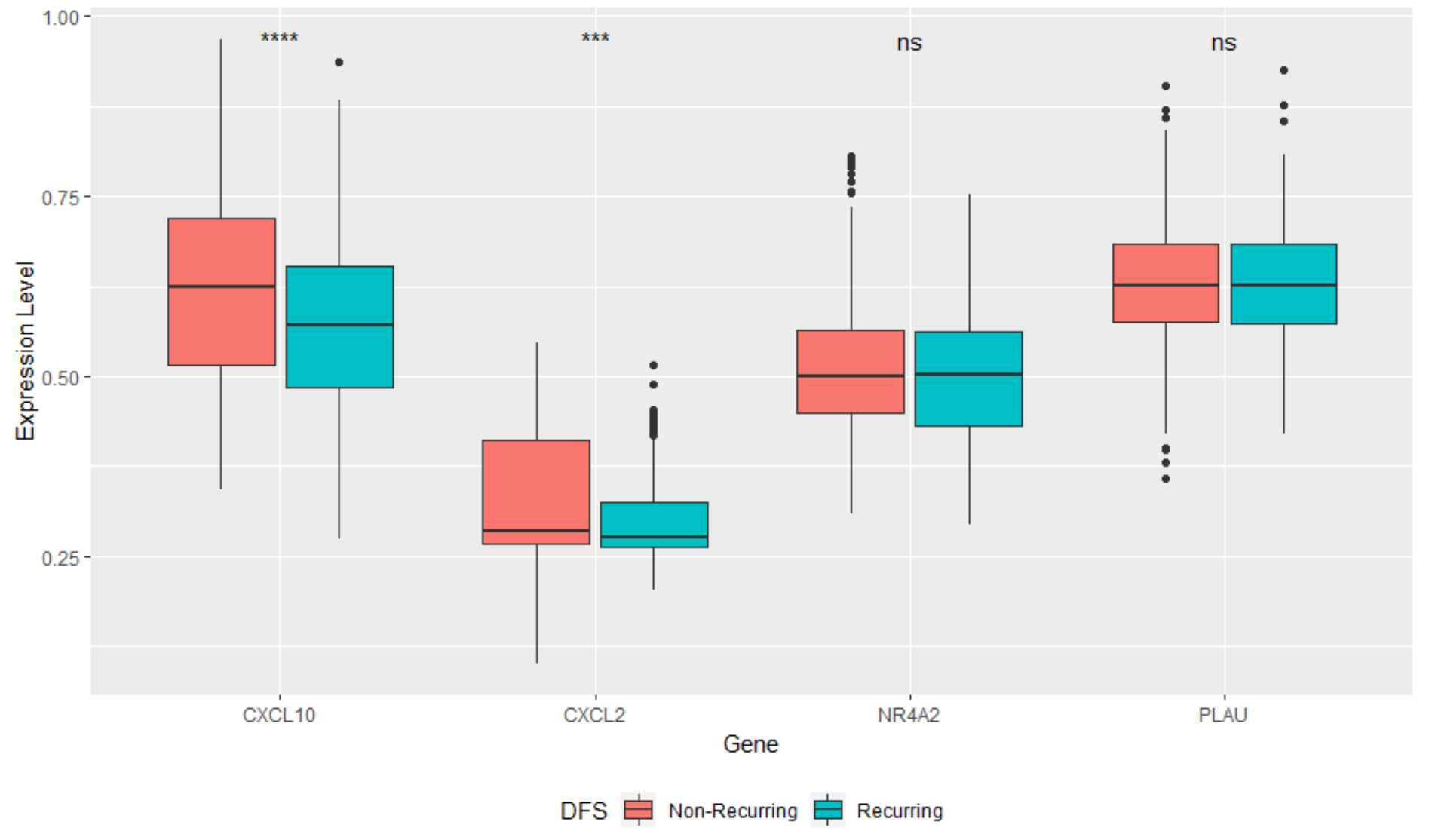
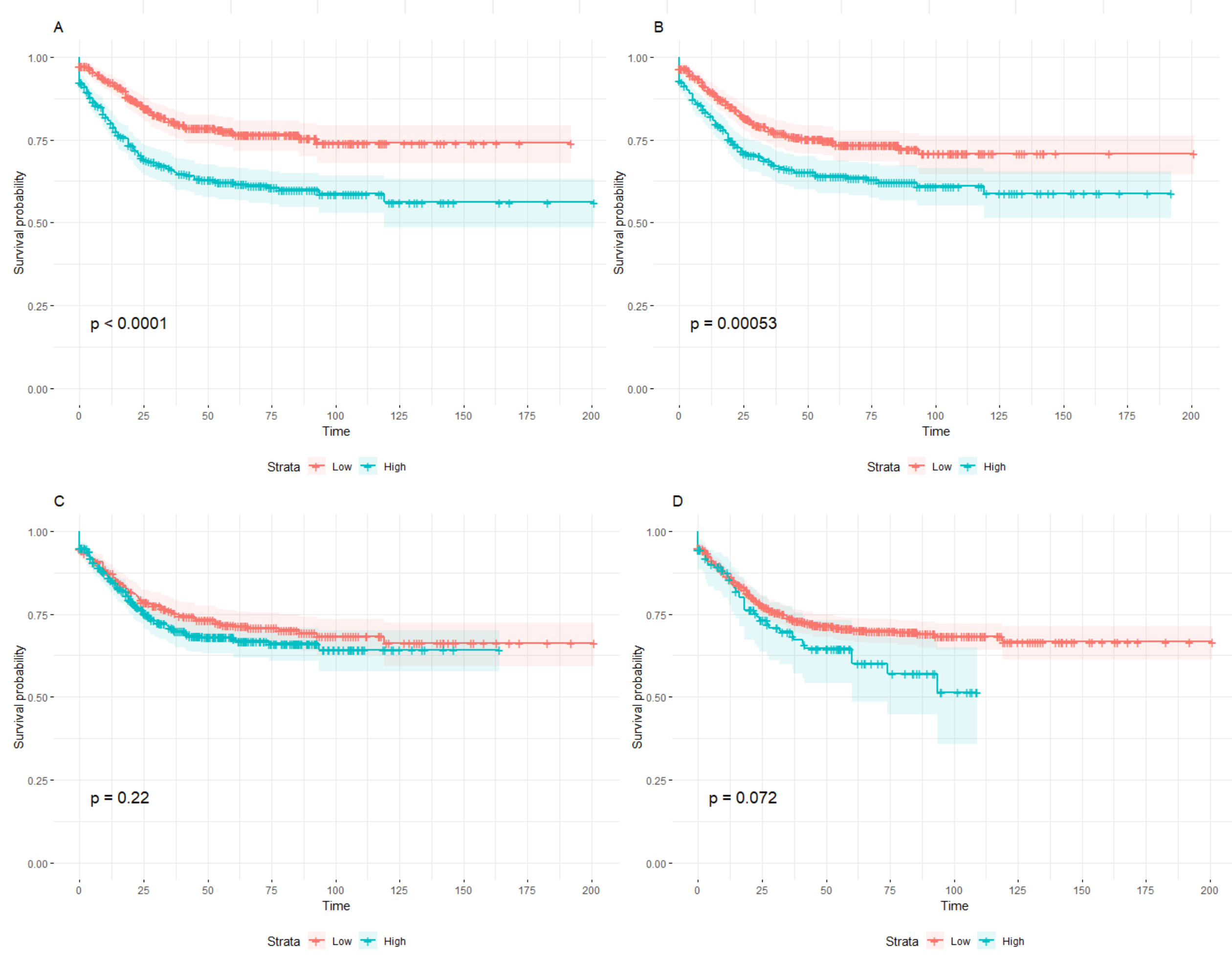


Table II: Summary of cox proportional hazard statistics for individual gene survival analysis with patient stratification by mean, median and ROC value.

	ING4	CXCL10	CXCL2	NR4A2	PLAU
Mean					
Cutoff Value	0.537	0.611	0.319	0.506	0.629
p-value	0.8819	<0.0001	0.0206	0.4361	0.7594
Hazard Ratio	1.019	1.909	1.404	1.105	1.04
95% Confidence Interval	(0.791, 1.314)	(1.468, 2.483)	(1.052, 1.874)	(0.859, 1.422)	(0.809, 1.338)
Median					
Cutoff Value	0.531	0.605	0.282	0.501	0.626
p-value	0.9394	<0.0001	0.0035	0.3381	0.7257
Hazard Ratio	1.01	1.884	1.459	1.131	1.046
95% Confidence Interval	(0.785, 1.299)	(1.451, 2.446)	(1.13, 1.883)	(0.879, 1.455)	(0.813, 1.346)
ROC					
Cutoff Value	0.497	0.608	0.28	0.495	0.717
p-value	7e-04	<0.0001	5e-04	0.2246	0.0736
Hazard Ratio	2.165	1.979	1.564	1.17	1.357
95% Confidence Interval	(1.37, 3.422)	(1.521, 2.574)	(1.214, 2.015)	(0.908, 1.506)	(0.97, 1.9)

* Cutoff values represent the mean, median and ROC cutoff after GAPDH normalization

Figure 3: Kaplan-Meier curves for patients stratified by ROC value for A) CXCL10, B) CXCL2, C) NR4A2, and D) PLAU



All gene expression data was normalized ~~with~~ to GAPDH. To determine if a particular gene had a positive or negative effect on cancer recurrence, we compared the gene expressions of the recurrent and non-recurrent groups (Figure 2). After doing so, we assigned a binary code for expression values per gene: for, we used CXCL10 and CXCL2, 0 for expression values lower than the threshold and 1 for expression values higher than the threshold. For NR4A2 and PLAU, we used 0 for expression values higher than the threshold and 1 for expression values lower than the threshold. This method was repeated for three times for thresholds equivalent to the mean, median and best ROC value.

Results

Figure 4: Histogram of risk score for recurrent vs. non-recurrent patients

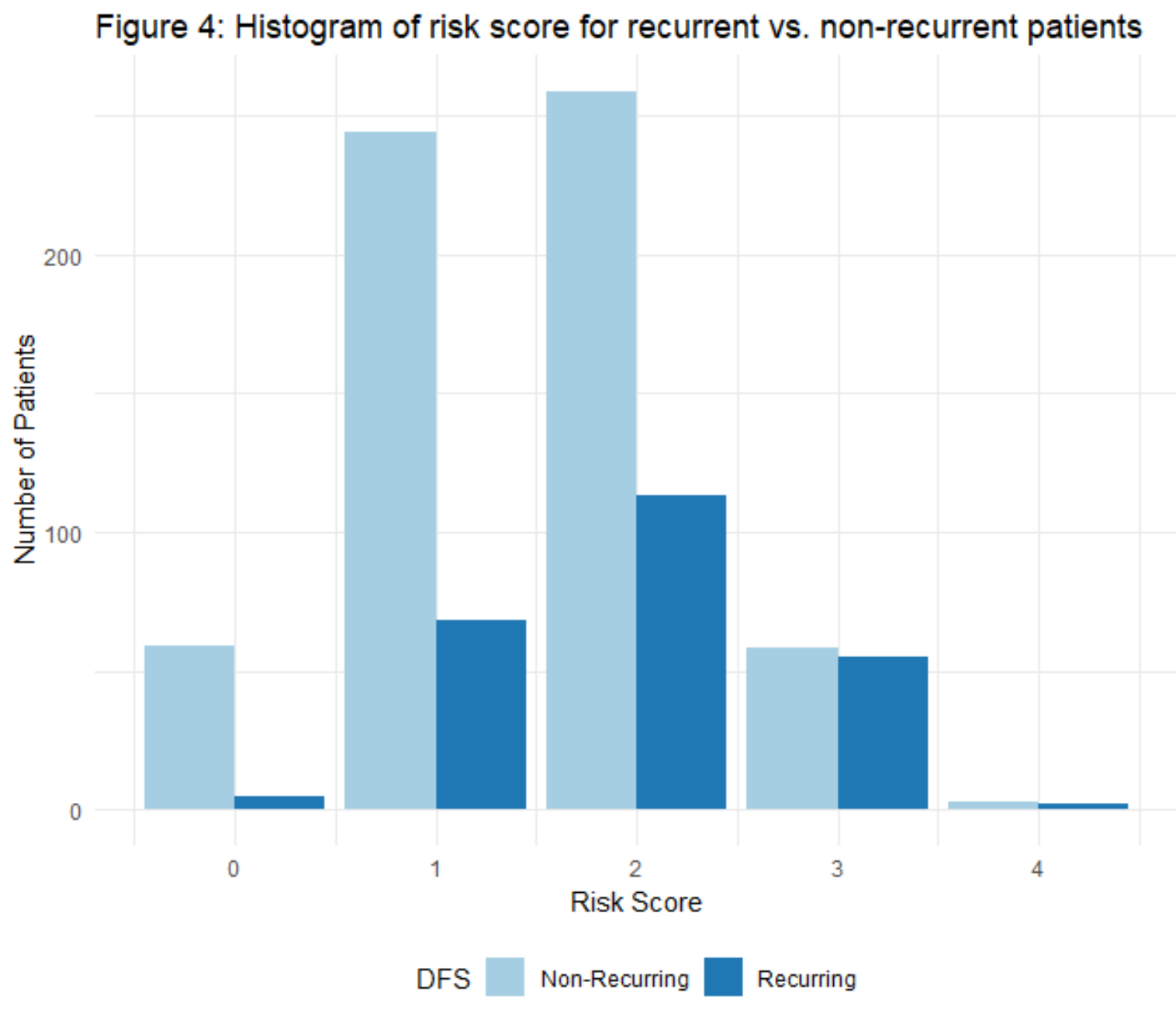


Figure 5: Kaplan-Meier curve of patient stratification by risk score

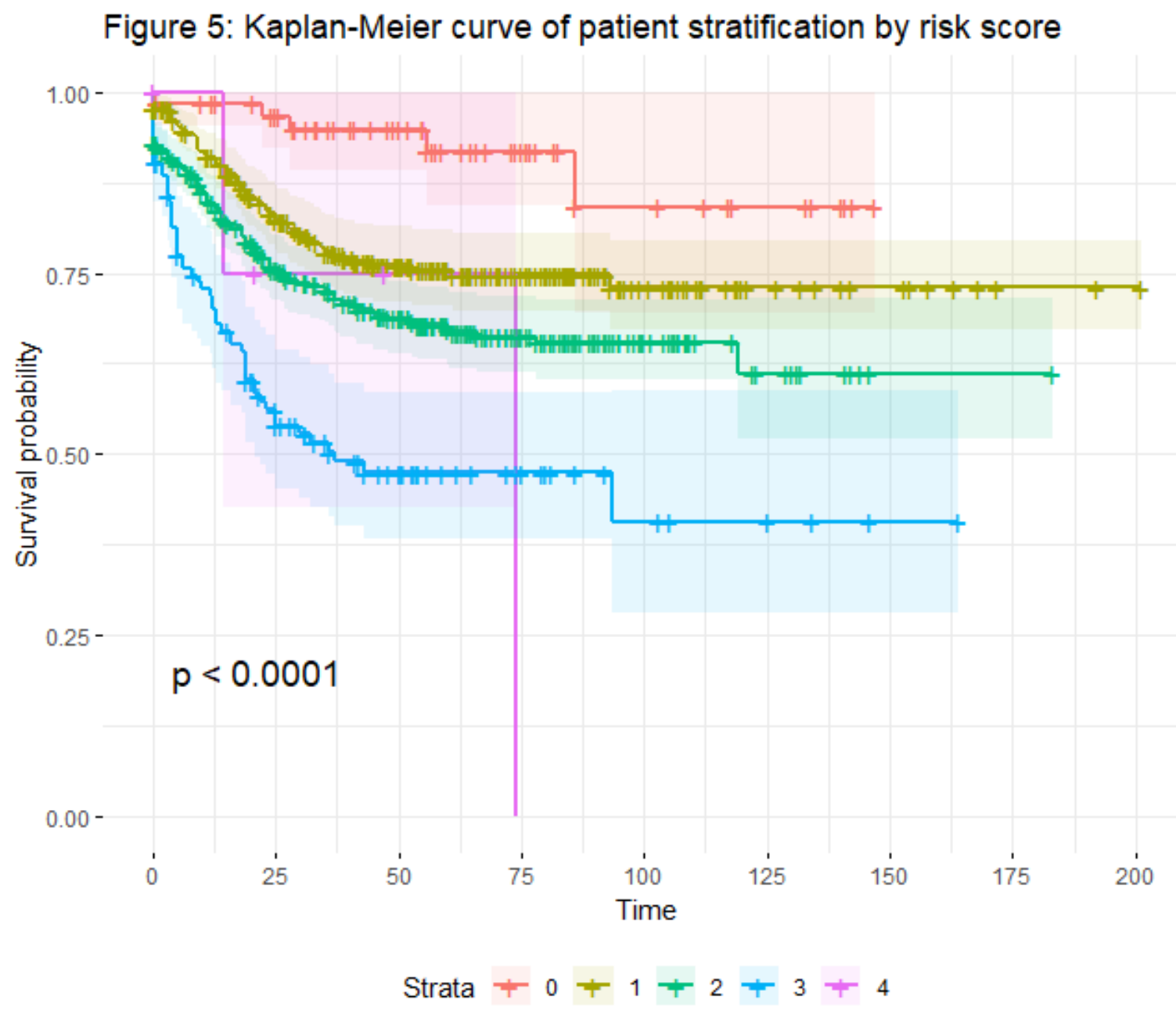
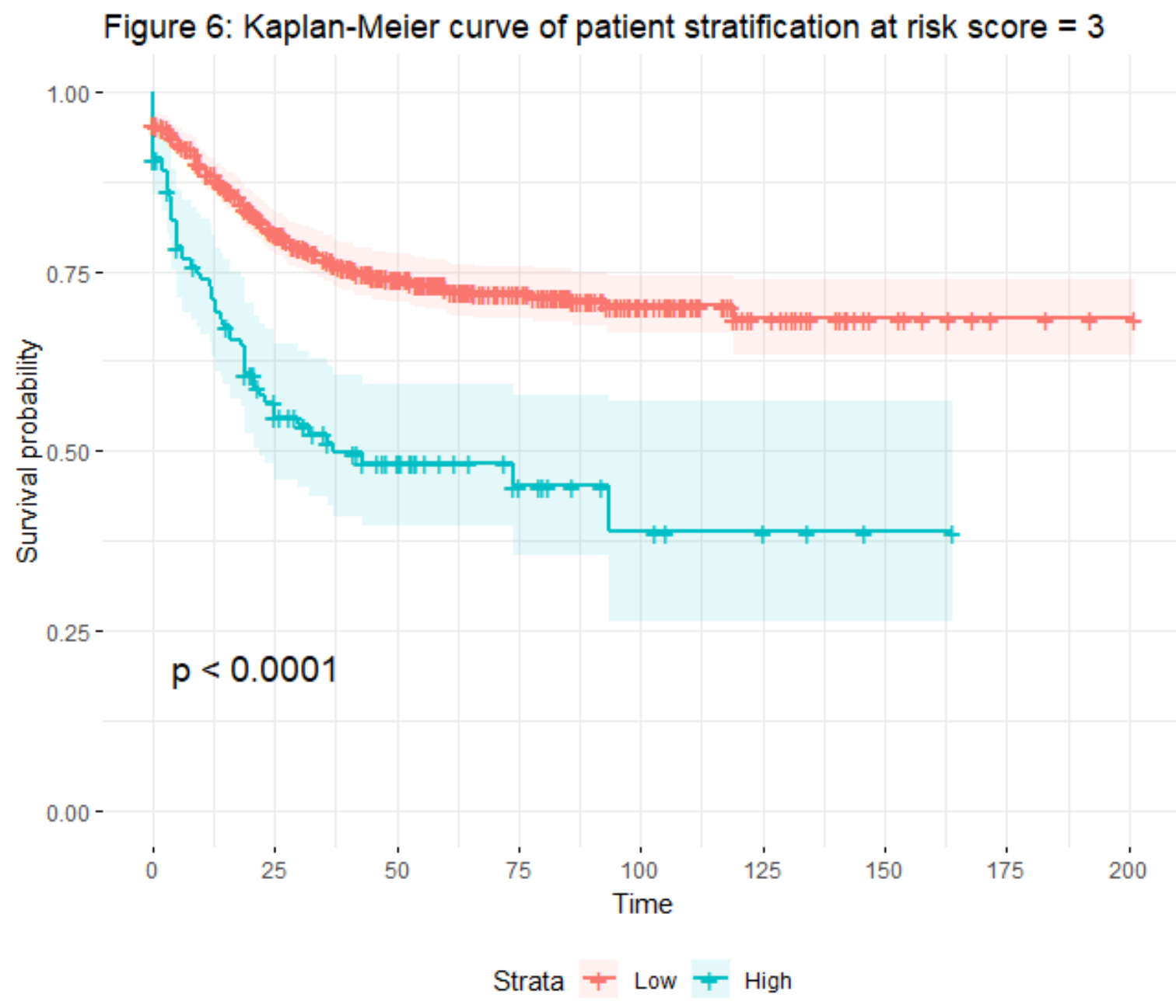


Figure 6: Kaplan-Meier curve of patient stratification at risk score = 3



We characterized each patient by a risk score of 0-4 which indicated the number of genes that correlated with recurrence, thus a higher score represented a greater risk of recurrence. The risk score was used to stratify patients into different survival groups (Figure 5). Risk score was then used to further stratify patients into those who had a score greater than or less than a particular score. Survival analysis was conducted at each stratification level. Ultimately, a risk score of at least three produced the most significant results (Figure 6).

Conclusions

Patients who had at least 3 genes they below, NR4A2 and PLAU, or above, CXCL10 and CXCL2 the best ROC value experienced recurrence at significantly more than and faster than patients who did not. However, using the receiver operating characteristic curve (ROC) to set the risk threshold potentially overfits the results. To account for this, we ran the same model using mean expression values as the risk threshold. This model was found to be significant as well, and performed similarly to the ROC Model for hazard ratio, percent of patients correctly classified and Akaike Information Criteria (AIC) (Table III).

Table III: Comparison of final mean and ROC models

	p-value	Hazard Ratio	% Correcty Classified	AIC
Mean	4.50e-7	1.913	65.13%	3108.5
ROC	3.35e-10	2.513	71.48%	3100.9

* % Correctly classified is given by the number of individuals who experience recurrent disease and were classified as high risk and those who did not experience recurrent disease and were characterized as low risk.

The ROC model performed slightly better in all aspects. However, the results of the mean model are comparable suggesting that the expression levels of these genes likely contribute to cancer recurrence

Conclusion/Future plans

- We devised a binary score system for four genes and were able to assess a prognostic value of the genes collectively.
- Expression of CXCL10 and CXCL10 higher than a ROC or mean value correlated with faster recurrence, suggesting their roles in aggressive disease.
- Expression of NR4A2 and PLAU lower than a ROC or mean value correlated with faster recurrence suggesting their functions in the suppression of aggressive disease.
- Are ROC or mean values an appropriate threshold for patient stratification?

References

[1] "Cancer Stat Facts: Common Cancer Sites." Surveillance, Epidemiology, and End Results Program, seer.cancer.gov/statfacts/html/common.html.
[2] You, Qi, et al. "Downregulated Expression of Inhibitor of Growth 4 (ING4) in Advanced Colorectal Cancers: A Non-Randomized Experimental Study." Pathology & Oncology Research, vol. 17, no. 3, 2011, pp. 473–477., doi:10.1007/s12253-010-9301-7.
[3] Byron S.A., Min E., Thal T.S., Hostetter G., Watanabe A.T., Azorsa D.O., Little T.H., Coya, T., Kim S. (2012). "Negative Regulation of NF-κB by the ING4 Tumor Suppressor in Breast Cancer." PLoS ONE, 7 (10) , art. no. e46823