Electronic Supplementary Material (ESI) for Molecular BioSystems.

Hedgehog-mesenchyme gene signature identifies bi-modal prognosis in Luminal and Basal breast cancer sub-types

Wandaliz Torres-García[1] and Maribella Domenech[2]

[1]Department of Industrial Engineering and [2]Department of Chemical Engineering, University of Puerto Rico Mayagüez, Mayagüez, PR*

* **Conflict of Interest:** The authors have declared that no conflict of interest exists.

Supplemental Materials

Supp. A.

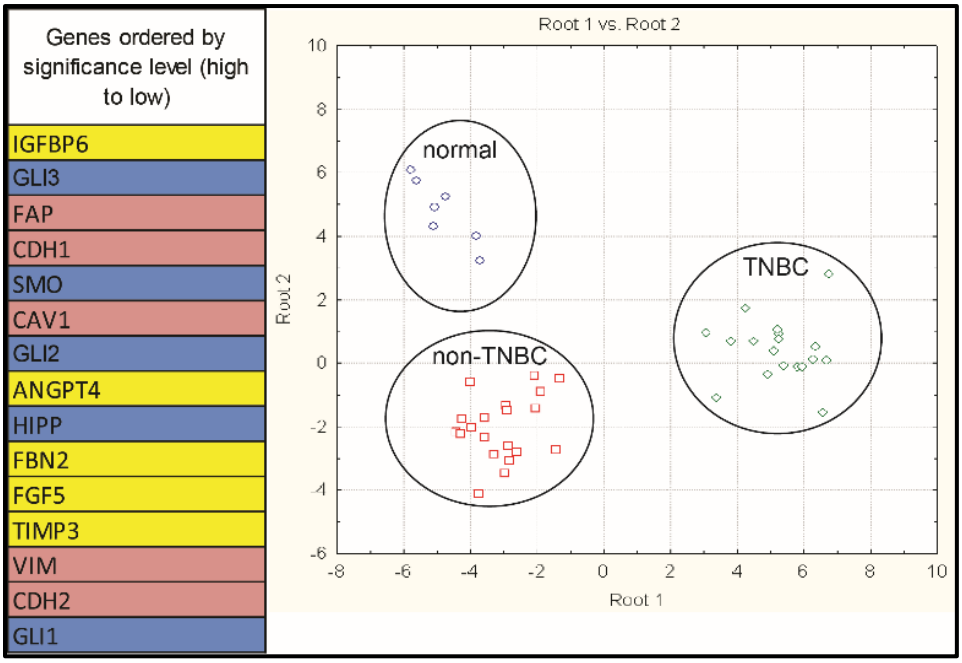


Figure A1. Two-dimensional representation of data from an LDA considering all samples (normal=7, non-TNBC=20 and TNBC=18) and using the 15-gene panel. The axes correspond to linear discriminant vectors and the points represent each sample.

Supp.B.

The RFs model outperformed the LDA model by a margin difference of up to 10% as seen in Table .

Table B1: Hh-mesenchymal gene signature performance using Random Forest and Linear Discriminant Analysis

Classification Model	Dataset	Number of Samples*	Error Rates (%)		AUCs	
			RFs	LDA	RFs	LDA
Subtype Signature Characterization	TCGA_Cell	587	32.030	39.520	0.800	0.767
	METABRIC	1974	42.650	42.710	0.682	0.689
	GSE 20685	327	41.590	42.810	0.770	0.796
	GSE 20711	90	50.000	45.450	0.657	0.702
	GSE 21653	266	42.110	34.960	0.720	0.773
	GSE 22226	129	44.190	44.190	0.696	0.760
	GSE 31448	294	37.070	32.650	0.721	0.778
Clinical Impact: Resistant vs Sensitive	GSE 58375	21	19.050	28.570	0.792	0.694
	GSE 77042	75	17.330	24.000	0.770	0.720

Tuning of parameters for Random Forest models measuring accuracy and kappa values.

We performed a 10-fold cross validation evaluation of accuracy (percentage of correctly classified instances) and kappa (similar to the accuracy evaluation but normalized to account for any imbalances in the classes) metrics. We tried different values within these two parameters using the largest gene expression dataset (METABRIC). As shown in Figures B1 and B2 from Supp.B, mtry equal to 3 showed best average accuracy and kappa metrics. Nonetheless their values across one standard error show insignificant changes. We expected this to happen since 15 predictors is not a large enough count compared to other common gene expression studies. Therefore, we validated that the default value of mtry in the performed R randomForest model is appropriate, in our case mtry=3. In terms of ntree, it is commonly known that with the larger the number of trees, the performance metrics estimation improves but the computational time increases. For this reason, we ran all our RFs models with 10,000 trees. In the tuning experiment, we did not observe any statistical differences in the accuracy nor kappa values across 9000, 10000 and 11000 trees (See Supp. B: Figure B1 & Figure B2). Thus, the choice of parameters (mtry=3 and ntree=10000) suit our purposes when using METABRIC dataset. We expect similar results with all other datasets since the set of predictors we are evaluating is relatively small and the construction of 10,000 individual decision trees per forest model should provide good metrics' precision.

Figure B1: Random forest parameter tuning based on accuracy using METABRIC dataset. (A) 10-Fold CV average accuracy values for different `mtry (1 through 15)` and `ntree (9000, 10000, 11000)` values. (B) 10-Fold CV average±1SD values.

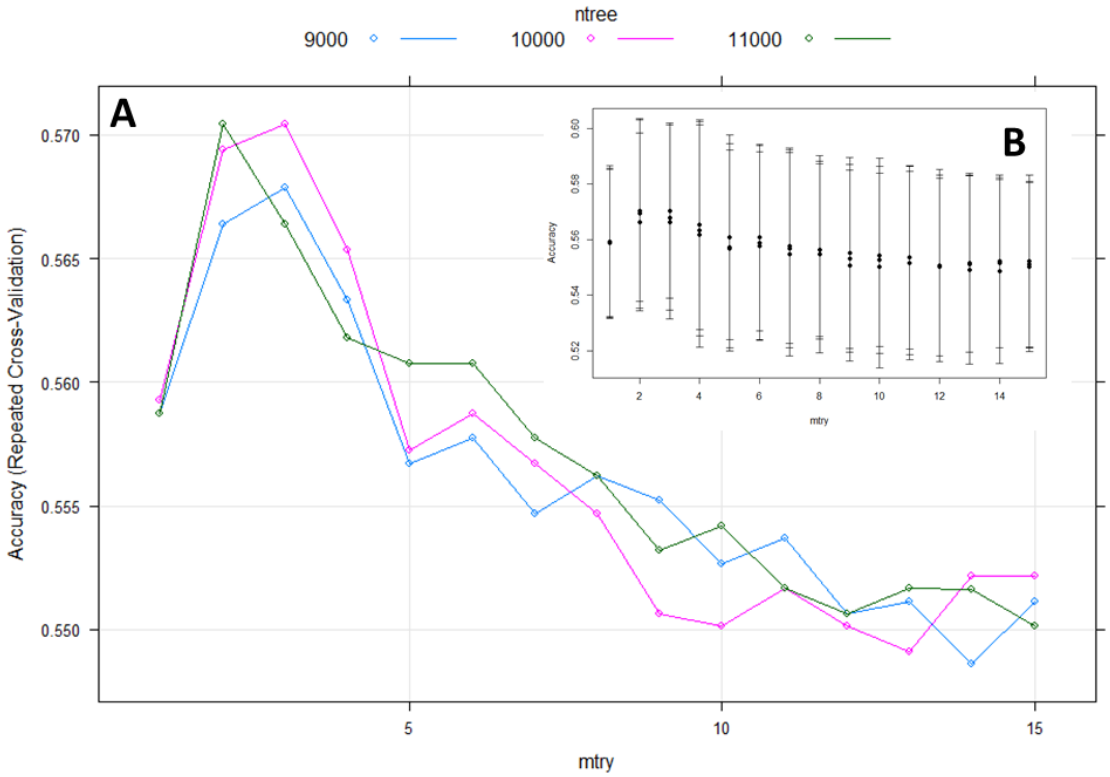
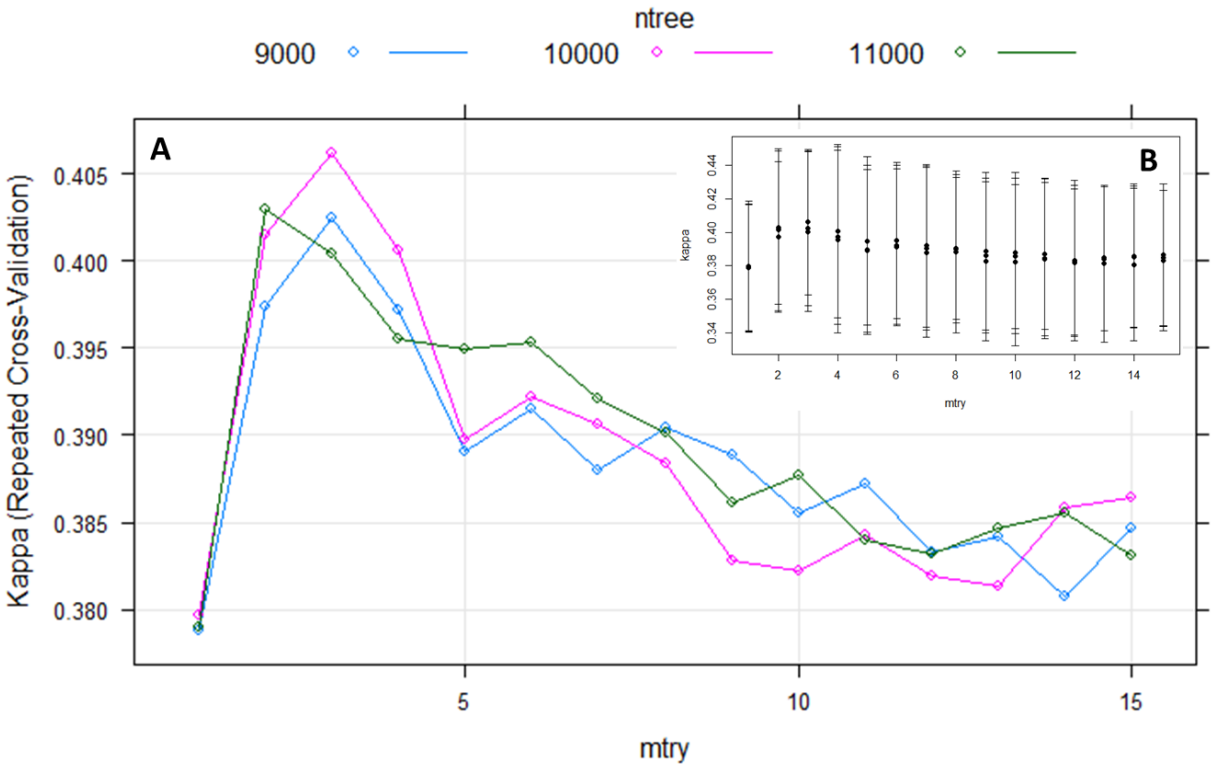


Figure B2: Random forest parameter tuning based on kappa using METABRIC dataset. (A) 10-Fold CV average kappa values for different `mtry (1 through 15)` and `ntree (9000, 10000, 11000)` values. (B) 10-Fold CV average±1SD kappa values.



Supp.C.

We only considered the following subtypes: Basal, Luminal A, Luminal B, HER2, Normal; as these are the most represented in literature and are used to classify breast cancer samples. A permutation test was performed to assess the predictive importance of the 15-gene signature in comparison with over one thousand random 15-gene sets. The Hh-mesenchyme signature ranked in the 85th percentile in terms of overall accuracy using the data from The Cancer Genome Atlas (TCGA).

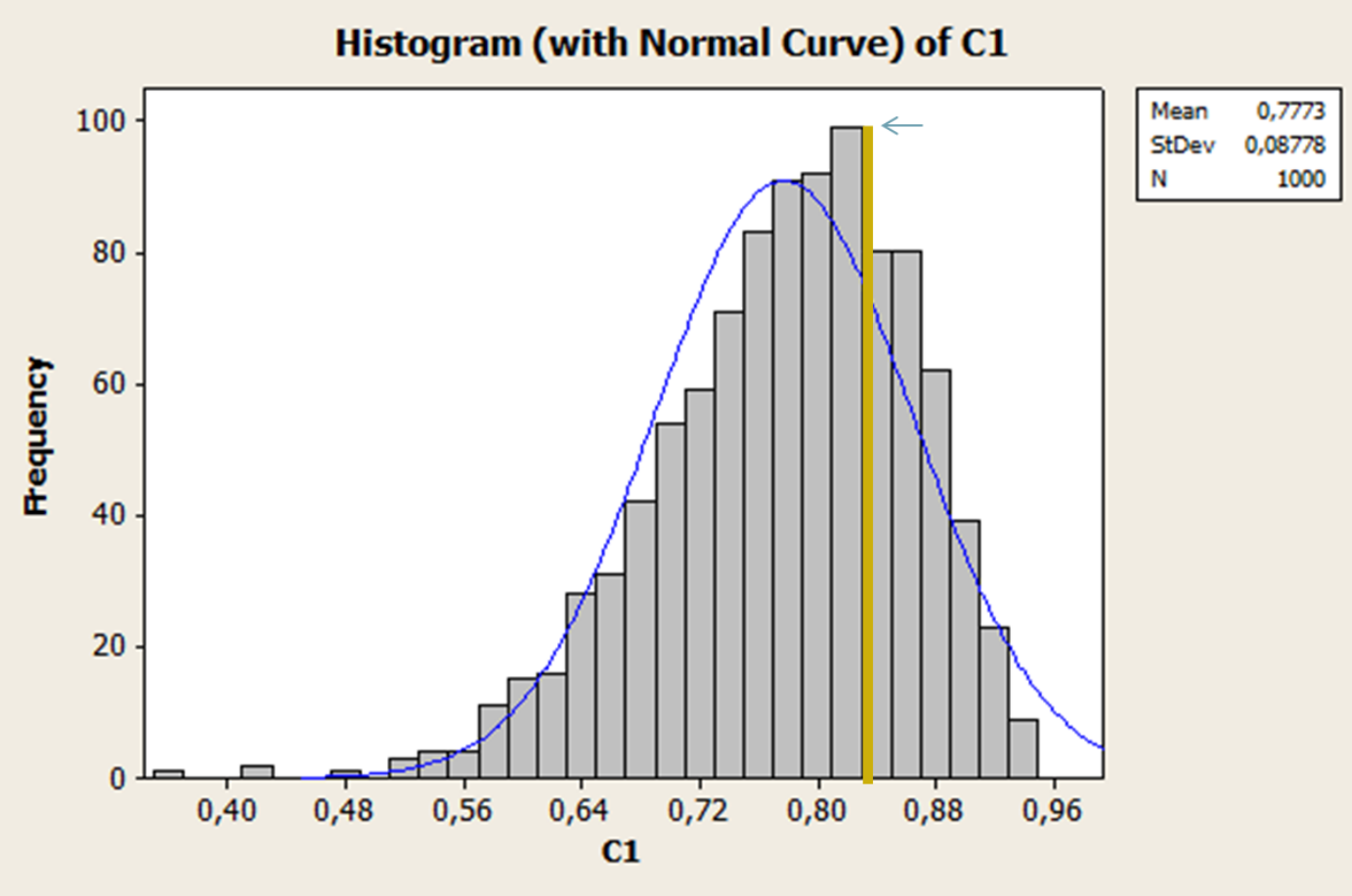


Figure C1: Histogram of accuracy values of 1000 random 15-gene sets compared to the Hh-mesenchyme signature towards breast cancer subtype discrimination using The Cancer Genome Atlas dataset.

Supp.D. Overall survival and Disease-Free Survival Analyses

The construction of hazard regression models is presented here. We incorporated several covariates including the 15-gene signature in both univariate and multivariate models using all samples from the METABRIC dataset (See Table D1 and Table D3). The hazard models for Basal samples (Table D2) did not produce significant expression patterns to discriminate survival. However, in Luminal A samples, further subgrouping of these samples, based on expression levels of genes such as IGFBP6, could improve clinical outcome and better therapeutic options assessments (see Table 2 and Table D4).

Table D1. Univariate Cox proportional hazard regression for overall survival using all METABRIC samples.

Univariate	beta	HR (95% CI for HR)	wald.test	p.value	
IGFBP6	-0.23	0.8 (0.74-0.86)	34	6.30E-09	***
HER2_STATUS	0.38	1.5 (1.2-1.7)	20	9.30E-06	***
PR_STATUS	-0.24	0.79 (0.7-0.89)	16	6.10E-05	***
CAV1	-0.13	0.88 (0.82-0.94)	16	6.50E-05	***
CorrectlyClassified	-0.23	0.8 (0.71-0.9)	14	0.00014	***
GLI2	-0.35	0.71 (0.56-0.89)	8.4	0.0038	**
SMO	-0.18	0.84 (0.73-0.96)	6.3	0.012	*
CDH2	0.072	1.1 (1-1.1)	5.3	0.021	*
FBN2	0.076	1.1 (1-1.2)	5.2	0.022	*
ER_STATUS	-0.15	0.86 (0.75-0.99)	4.5	0.033	*
VIM	-0.062	0.94 (0.88-1)	3.3	0.071	.
ANGPT4	-0.37	0.69 (0.45-1.1)	2.9	0.086	.
FAP	-0.037	0.96 (0.91-1)	1.4	0.24	
CDH1	0.028	1 (0.98-1.1)	1.3	0.25	
GLI1	-0.15	0.86 (0.65-1.1)	1.1	0.28	
FGF5	0.16	1.2 (0.84-1.6)	0.91	0.34	
HHIP	-0.18	0.84 (0.53-1.3)	0.55	0.46	
GLI3	0.026	1 (0.95-1.1)	0.44	0.51	
TIMP3	-0.0061	0.99 (0.94-1.1)	0.05	0.83	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table D2. Univariate Cox proportional hazard regression for overall survival for Basal samples using METABRIC dataset.

Basal Samples	beta	HR	(95% HR CI)	wald.test	p.value
GLI3	0.2	1.2	(0.94-1.6)	2.2	0.14
CDH2	0.086	1.1	(0.96-1.2)	1.9	0.17
ANGPT4	-0.7	0.5	(0.17-1.5)	1.6	0.21
GLI2	0.22	1.2	(0.82-1.9)	1.1	0.3
SMO	-0.1	0.9	(0.68-1.2)	0.55	0.46
FBN2	0.058	1.1	(0.91-1.2)	0.54	0.46
IGFBP6	-0.064	0.94	(0.79-1.1)	0.5	0.48
FGF5	-0.31	0.74	(0.31-1.8)	0.48	0.49
GLI1	-0.21	0.81	(0.39-1.7)	0.31	0.58
FAP	-0.021	0.98	(0.83-1.2)	0.07	0.79
HHIP	0.16	1.2	(0.33-4.2)	0.06	0.81
CDH1	0.011	1	(0.88-1.2)	0.02	0.88
CAV1	-0.0058	0.99	(0.84-1.2)	0	0.94
TIMP3	0.003	1	(0.87-1.2)	0	0.97
VIM	0.0013	1	(0.82-1.2)	0	0.99

Table D3. Multivariate Cox proportional hazard regression for overall survival using all METABRIC samples.

	coef	exp(coef)	se(coef)	z	Pr(>|z|)	
IGFBP6	-0.2068	0.81318	0.052669	-3.926	8.62E-05	***
CorrectlyClassifiedTRUE	-0.17617	0.838473	0.061164	-2.88	0.00397	**
PR_STATUS+	-0.18645	0.829904	0.071259	-2.616	0.00888	**
SMO	-0.17649	0.838205	0.075707	-2.331	0.01974	*
HER2_STATUS+	0.185466	1.203779	0.095689	1.938	0.0526	.
CDH2	0.048773	1.049982	0.033076	1.475	0.14033	
FBN2	0.044705	1.045719	0.034301	1.303	0.19247	
ER_STATUS+	0.031838	1.032351	0.089772	0.355	0.72285	
GLI2	0.00812	1.008153	0.136922	0.059	0.95271	
CAV1	-0.00147	0.998527	0.044975	-0.033	0.97386	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Rsquare= 0.044 (max possible= 1)
Likelihood ratio test (p= 2e-12); Wald test (p= 2.152e-12); Score (logrank) (p= 1.891e-12)

Table D4. Multivariate Cox proportional hazard regression for overall survival for Luminal A samples using METABRIC dataset.

	coef	exp(coef)	se(coef)	z	Pr(>|z|)	
PR_STATUS+	-0.46513	0.62806	0.11608	-4.007	6.15E-05	***
IGFBP6	-0.37683	0.68603	0.10048	-3.75	0.000177	***
CorrectlyClassifiedTRUE	-0.10358	0.9016	0.12591	-0.823	0.41069	
FAP	-0.0518	0.94952	0.06843	-0.757	0.449107	
CAV1	0.01872	1.0189	0.0918	0.204	0.838386	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Likelihood ratio test (p=1.747e-08); Wald test(p=1.469e-08); Score (logrank) (p=1.161e-08)

Overall Survival by Receptor Status

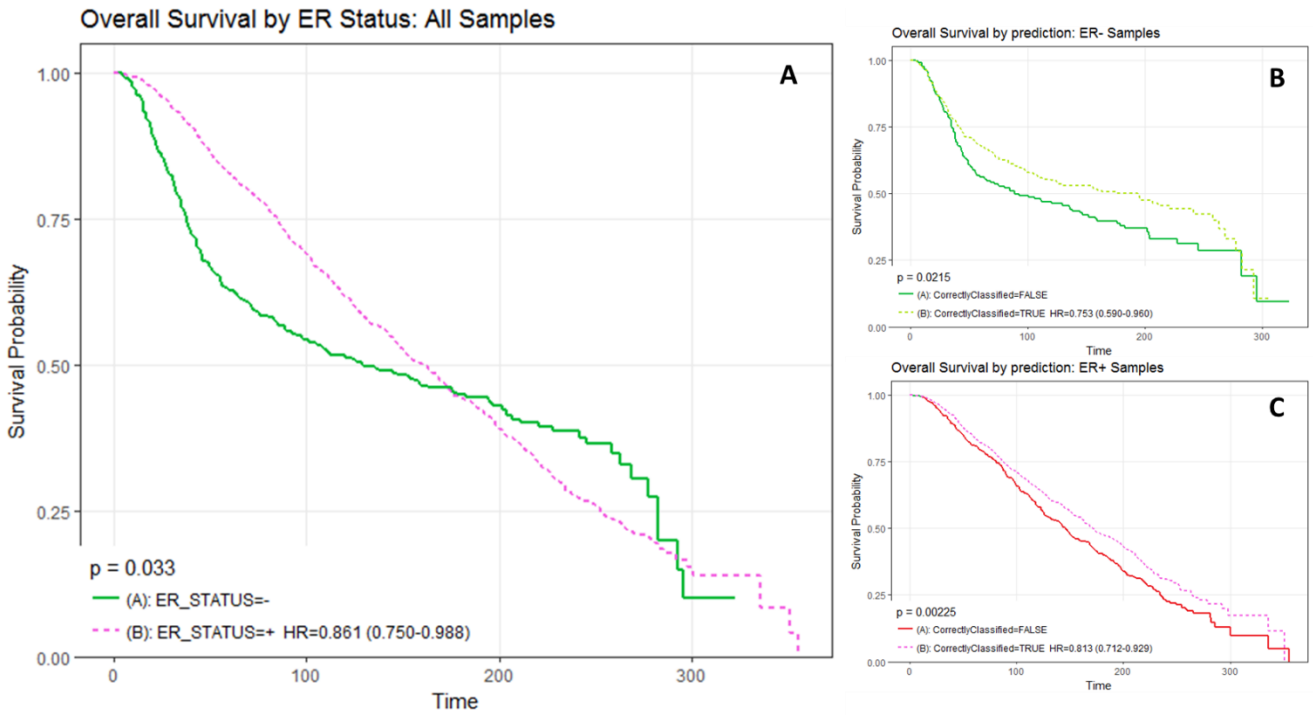


Figure D1: Kaplan Meier overall survival curves and log-rank tests by ER receptor status using all samples from METABRIC dataset. (A) ER- versus ER+ samples. (B) ER- samples across Hh15 model prediction. (C) ER+ samples across Hh15 model prediction.

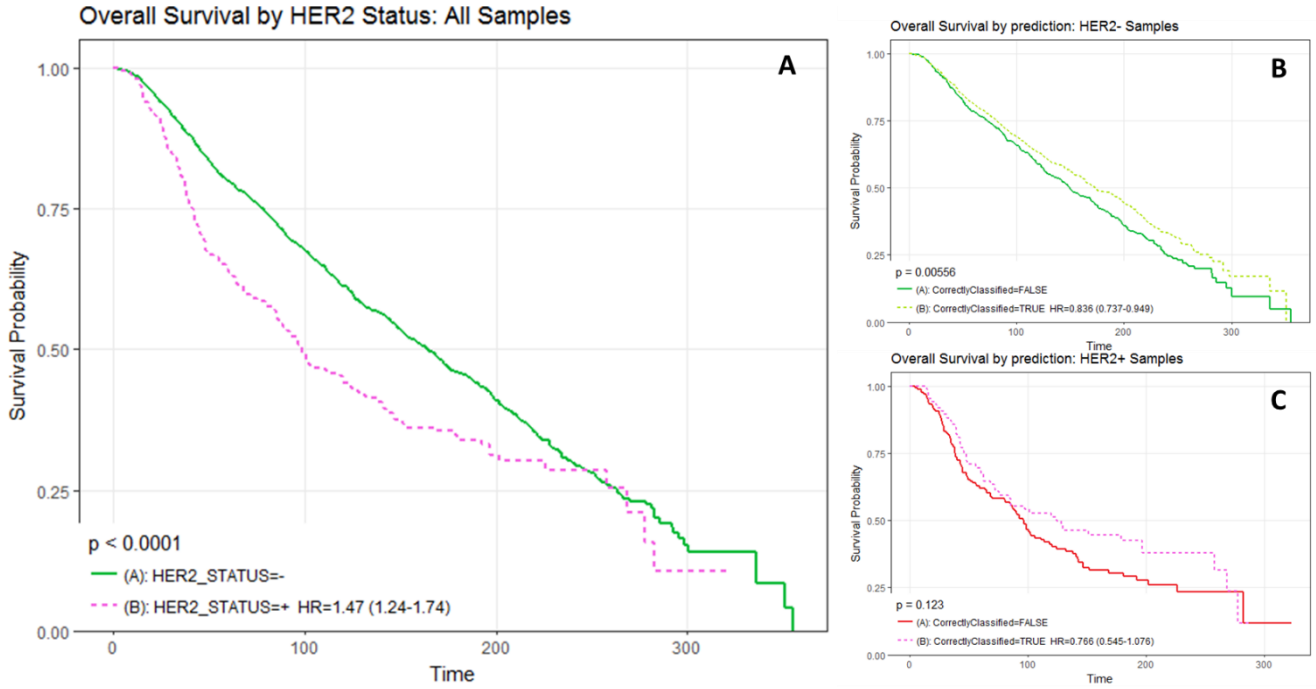


Figure D2: Kaplan Meier overall survival curves and log-rank tests by HER receptor status using all samples from METABRIC dataset. (A) HER2- versus ER+ samples. (B) HER2- samples across Hh15 model prediction. (C) HER2+ samples across Hh15 model prediction.

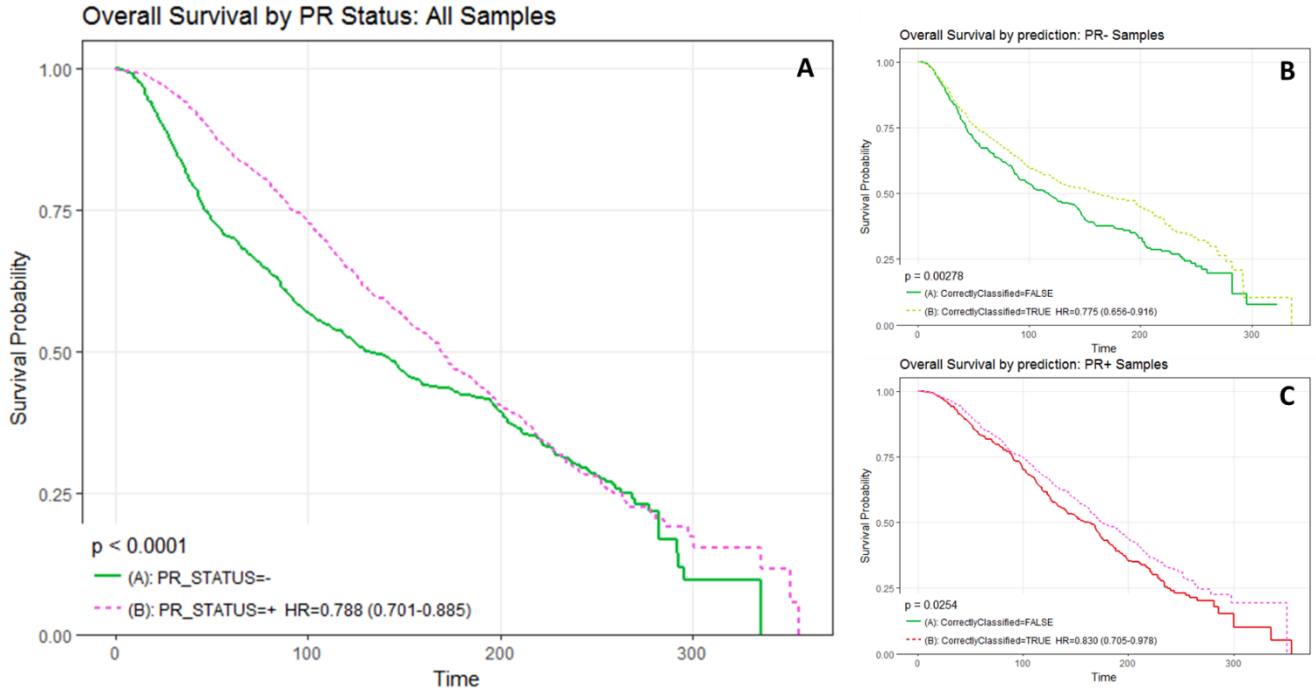


Figure D3: Kaplan Meier overall survival curves and log-rank tests by PR receptor status using all samples from METABRIC dataset. (A) PR- versus ER+ samples. (B) PR- samples across Hh15 model prediction. (C) PR+ samples across Hh15 model prediction.

Study of overall and disease-free survival across Luminal A and basal samples from the METABRIC dataset across different therapeutic approaches.

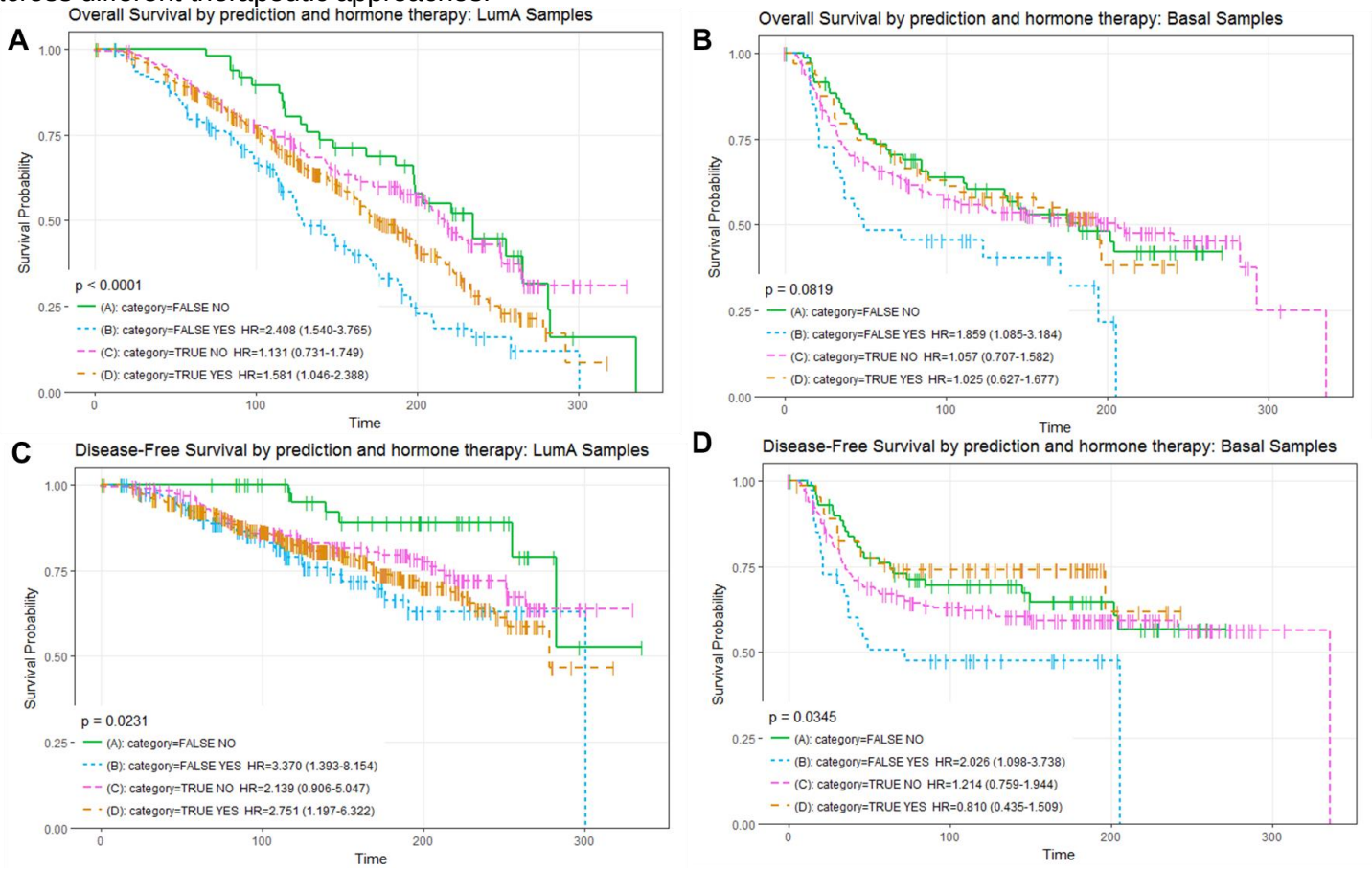


Figure D4: Overall and DFS curves by prediction/hormone therapy. (A-B) shows overall survival for Luminal A and basal samples from the METABRIC datasets respectively; inspected for differences on whether the 15-gene signature correctly classifies them (TRUE:Correctly Classified, FALSE:Incorrectly Classified) and if hormone therapy was applied(Yes/NO). (C-D) shows DFS curves for Luminal A and basal respectively.

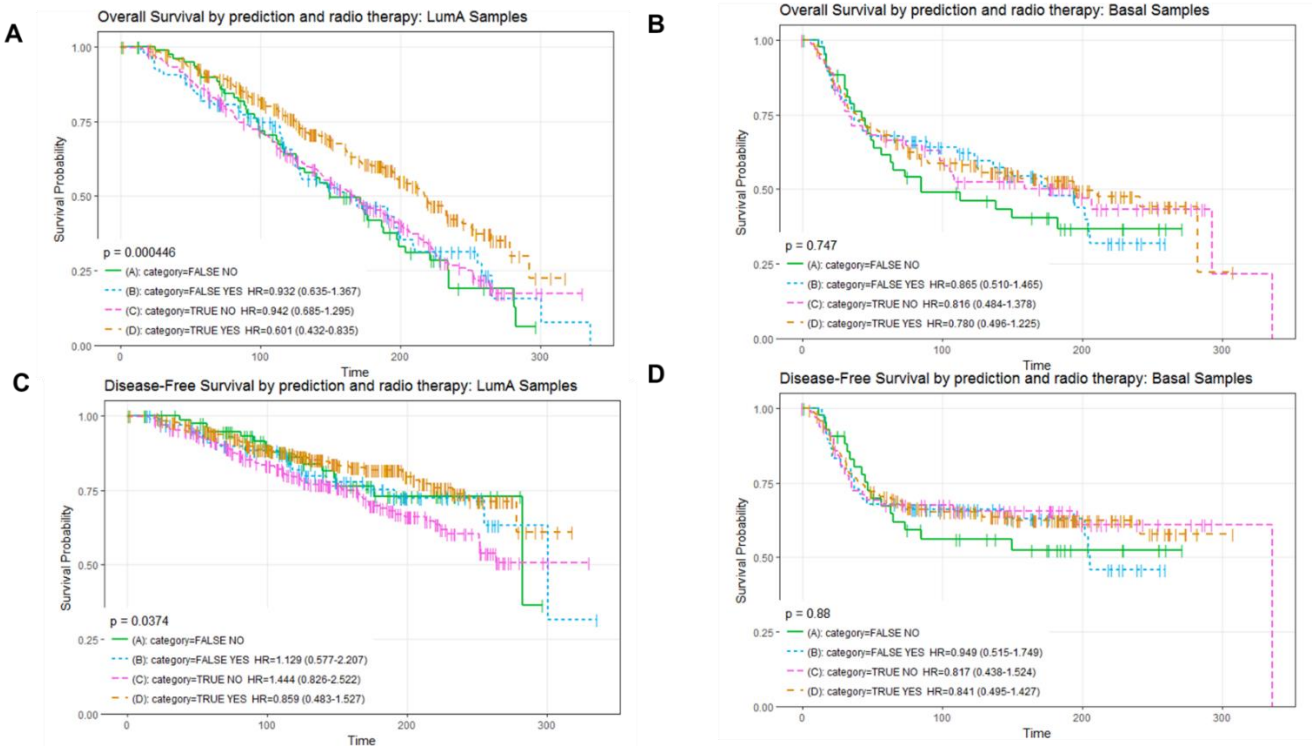


Figure D5: Overall and DFS curves by prediction/radio therapy. (A-B) shows overall survival for Luminal A and basal samples from the METABRIC datasets respectively; inspected for differences on whether the 15-gene signature correctly classifies them (TRUE:Correctly Classified, FALSE:Incorrectly Classified) and if radio therapy was received(Yes/NO). (C-D) shows DFS curves for Luminal A and basal respectively.

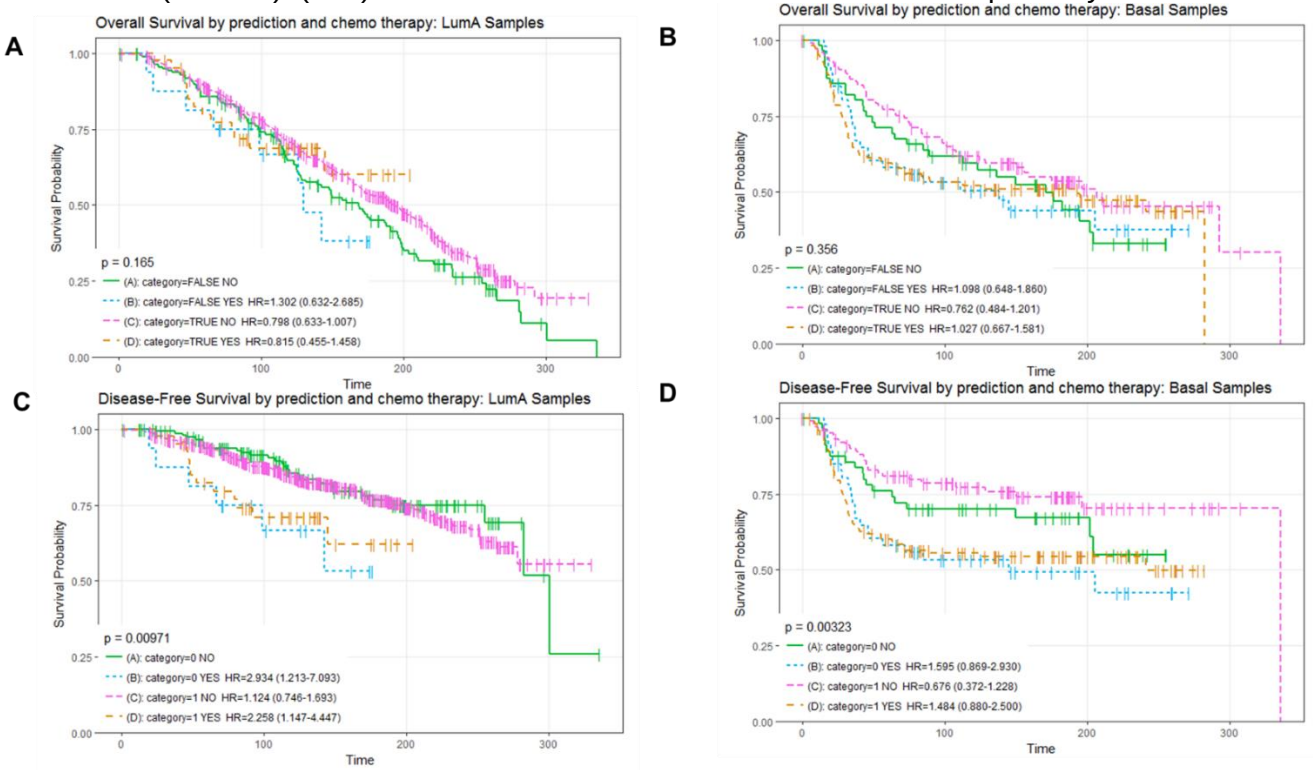
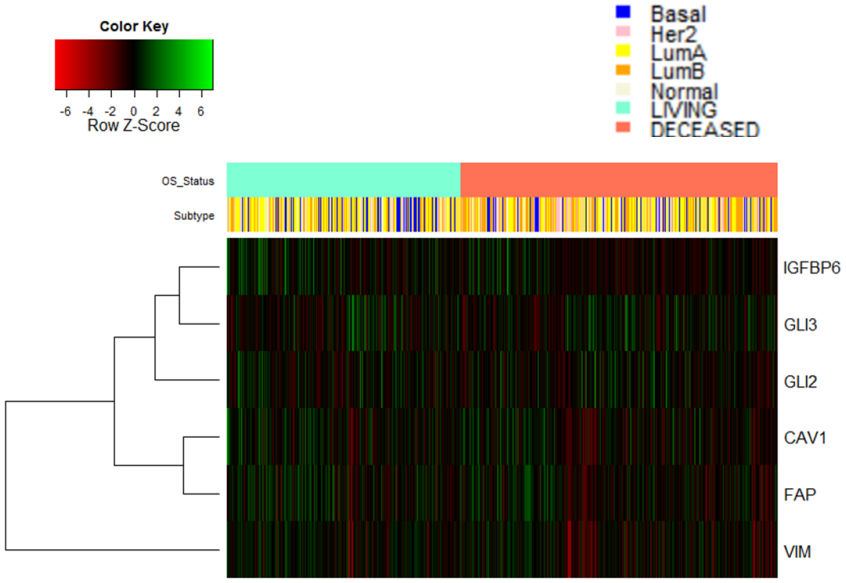


Figure D6: Overall and DFS curves by prediction/chemotherapy. (A-B) shows overall survival for Luminal A and basal samples from the METABRIC datasets respectively; inspected for differences on whether the 15-gene signature correctly classifies them (TRUE:Correctly Classified, FALSE:Incorrectly Classified) and if chemotherapy was applied(Yes/NO). (C-D) shows DFS curves for Luminal A and basal respectively.

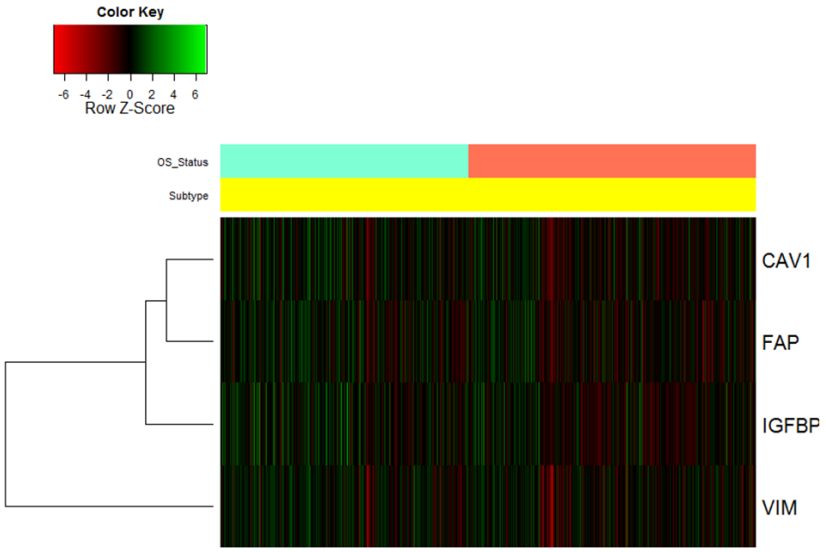
Differential Expression in Overall Survival using a non-parametric approach

Differential Expression in Overall Survival and Disease-Free Survival outcome using non-parametric multiple hypothesis testing, Wilcoxon Permuted Significance Test with p-value adjustment by Westfall & Young (1993).



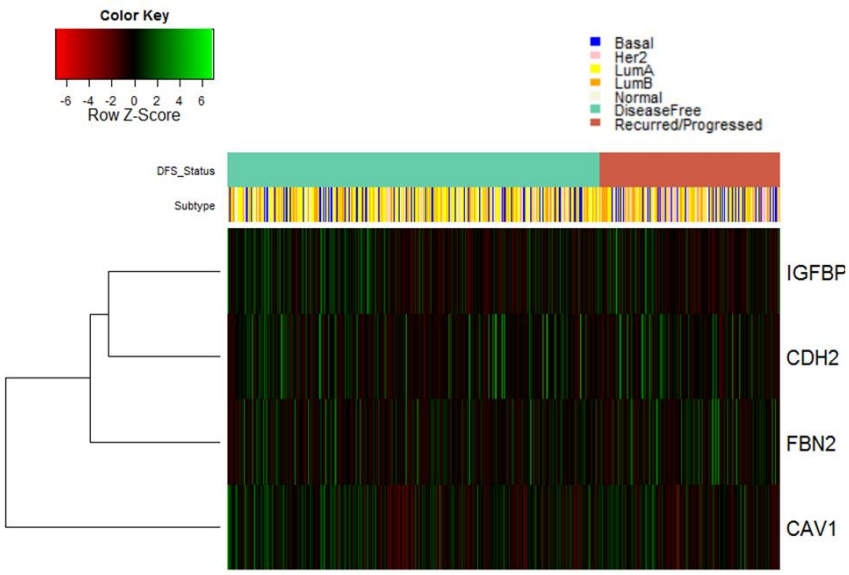
All Samples	index	teststat	rawp	adjp
IGFBP6	1	-6.7648	0.0001	0.0001
CAV1	6	-6.7548	0.0001	0.0001
VIM	13	-6.5818	0.0001	0.0001
GLI2	7	-3.3421	0.0008	0.0085
FAP	3	-3.3049	0.0011	0.0096
GLI3	2	3.1534	0.0014	0.0153
ANGPT4	8	-2.1561	0.0303	0.24
TIMP3	12	-1.614	0.1059	0.5827
FBN2	10	1.44048	0.1505	0.6733
SMO	5	-1.3594	0.176	0.6798
FGF5	11	0.88767	0.3656	0.9042
CDH2	14	0.43982	0.6687	0.9891
GLI1	15	0.34292	0.7389	0.9891
CDH1	4	0.30216	0.768	0.9891
HHIP	9	-0.1969	0.845	0.9891

Figure D7: Wilcoxon Permuted Significance Test with adjustment for all samples in the METABRIC dataset. Heatmap is sorted by overall survival.



LumA	index	teststat	rawp	adjp
IGFBP6	1	-6.1404	0.0001	0.0001
CAV1	6	-5.0606	0.0001	0.0001
VIM	13	-4.5198	0.0001	0.0003
FAP	3	-4.4581	0.0001	0.0003
GLI2	7	-2.439	0.0153	0.1459
ANGPT4	8	-2.4168	0.0151	0.1459
FBN2	10	-2.2223	0.0262	0.2104
TIMP3	12	-1.9939	0.048	0.3082
CDH1	4	-1.8232	0.0658	0.3818
SMO	5	1.80657	0.0757	0.3818
CDH2	14	-1.3755	0.1623	0.5944
FGF5	11	0.83706	0.3997	0.8761
HHIP	9	0.47095	0.6407	0.9526
GLI1	15	-0.4396	0.6622	0.9526
GLI3	2	0.39083	0.6919	0.9526

Figure D8: Wilcoxon Permuted Significance Test with adjustment for Luminal A samples in the METABRIC dataset. Heatmap is sorted by overall survival.



All Sampl	index	teststat	rawp	adjp
IGFBP6	1	-4.3841	0.0001	0.0002
FBN2	10	3.97177	0.0002	0.0009
CDH2	14	3.47856	0.0012	0.0067
CAV1	6	-3.2826	0.0014	0.0115
TIMP3	12	-1.8593	0.0649	0.4817
VIM	13	-1.6441	0.1002	0.6258
ANGPT4	8	-1.386	0.1628	0.7907
GLI3	2	-1.1955	0.2387	0.8702
FGF5	11	1.15004	0.2474	0.8702
HHIP	9	1.11058	0.2669	0.8702
GLI1	15	0.9949	0.3124	0.8702
FAP	3	0.65266	0.514	0.934
SMO	5	-0.5775	0.5597	0.934
GLI2	7	-0.5519	0.5811	0.934
CDH1	4	0.16634	0.8724	0.934

Figure D9: Wilcoxon Permuted Significance Test with adjustment for all samples in the METABRIC dataset. Heatmap is sorted by disease-free survival.

Supp.E.

The gene signature did a better job characterizing resistant versus sensitive cases than subtypes. GLI1, GLI2 and SMO were found as the most relevant genes to discriminate these cases using MDG score from the RFs. All three genes tend to be less expressed in most of the sensitive instances and highly expressed in the resistant ones as shown in the heatmaps here.

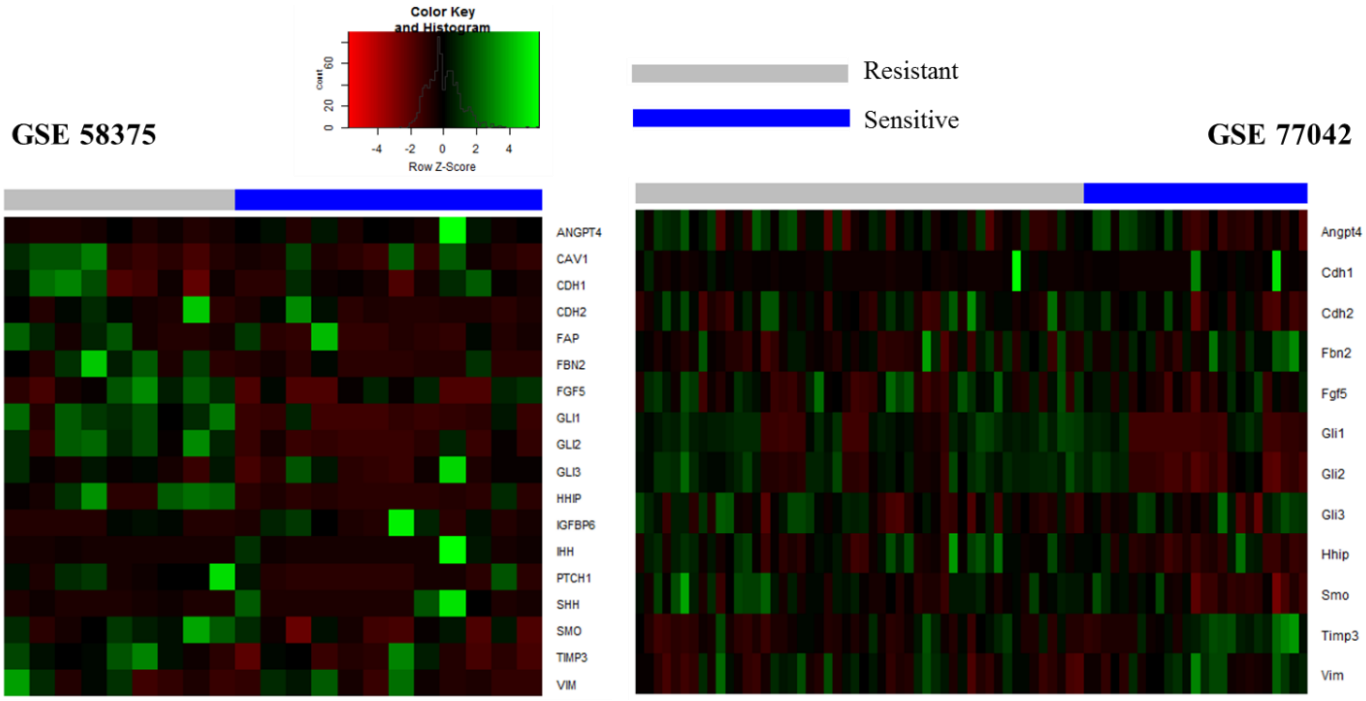


Figure E1: Heatmaps of resistant and sensitive samples across the Hedgehog-mesenchyme 15-gene expression for GSE58375 and GSE77042.

Supplemental Materials References

1. Lopez-Otin C, Diamandis EP. Breast and prostate cancer: an analysis of common epidemiological, genetic, and biochemical features. *Endocr Rev.* 1998;19(4):365-396.
2. Coffey DS. Similarities of prostate and breast cancer: Evolution, diet, and estrogens. *Urology.* 2001;57(4 Suppl 1):31-38.
3. Brabender J, Marjoram P, Salonga D, et al. A multigene expression panel for the molecular diagnosis of Barrett's esophagus and Barrett's adenocarcinoma of the esophagus. *Oncogene.* 2004;23(27):4780-4788.
4. Nebozhyn M, Loboda A, Kari L, et al. Quantitative PCR on 5 genes reliably identifies CTCL patients with 5% to 99% circulating tumor cells with 90% accuracy. *Blood.* 2006;107(8):3189-3196.