Assessing Histological Grades of Primary Breast Cancer Using Gene Signatures

Priyanka Arora, Lee Carraher,Ben Landis, Li Jing, Samuel Schmidt

December 13, 2013

Abstract

This paper is a re-Analysis of the Sotiriou et. al. paper on using tumor histological grade expressed gene signatures for breast cancer prognosis [1]. The principle assumptions of this process are that the classification of histological grade and tumor progression has a strong correlation with the length of survival time. Furthermore, we suggest grade 2 tumors are misclassified grade 1 or grade 3 tumors. The result of developing better gene signatures for classifying tumor grades can further assist in diagnosis and treatment options for breast cancer patients.

1 Background

Breast Cancer is the most common malignancy among women and is the second leading cause of cancer death.The American Cancer Society estimated 234,580 new cases of breast cancer in US, 2013 and 40,030 deaths related to breast cancer. The incident rate of new cases of breast cancer is highest for Whites and second highest for African Americans [2]. Breast cancer incidence is lowest for those of American Indian/Alaska Native, Asian American/Pacific Islander, and Hispanic/Latina descent. Similarly morality is highest for Whites and second highest for African Americans while the lowest mortality rate is for individuals of American Indian/Alaska Native, Asian American/Pacific Islander, and Hispanic/Latina descent [2].

1.1 Breast Cancer Heterogeneity and Treatment

The different outcomes of breast cancer related to age and race indicate that not all breast cancers behave similarly and that there are likely various environmental and genetic factors that influence outcomes. In fact, studies have shown that breast cancers are clinically and genetically heterogeneous [3]. Significant molecular heterogeneity was suggested by a comprehensive study on primary breast cancers by using multiple molecular information platforms covering genomic DNA copy number arrays, DNA methylation, exome sequencing, messenger RNA arrays, microRNA sequencing and reverse-phase protein arrays [4].

The analysis of molecular pathways has improved our understanding of the clinical behavior of breast cancer; however, the more we learn about the molecular characteristics of breast cancer, the more we appreciate the diversity of the disease [5] [6]. Breast cancer treatment decisions are made based on multiple variables, including genetic factors, disease burden, tumor markers, estrogen receptor status, and patient preference; however, treatments continue to evolve and the search for the optimal treatment protocol is ongoing [7].

Currently, breast cancer is treated with a multidisciplinary approach involving surgical oncology, radiation oncology, and medical oncology, which has been associated with a reduction in breast cancer mortality [7]. Generally, six types of standard treatment are used: Surgery, sentinel lymph node biopsy followed by surgery, radiation therapy, chemotherapy, hormone therapy and targeted therapy [8]. Most patients with breast cancer have surgery to remove the cancer from the breast. Some of the lymph nodes under the arm are usually taken out and looked at under a microscope to see if they contain cancer cells.

Radiation therapy may follow surgery in an effort to eradicate residual disease while reducing recurrence rates. Surgical resection with or without radiation is the standard treatment for ductal carcinoma in situ. Hormone therapy and chemotherapy are the 2 main interventions for treating metastatic breast cancer. Common chemotherapeutic regimens include Docetaxel, Cyclophosphamide, Doxorubicin, Trastuzumab, etc. Two selective estrogen receptor modulators (SERMs), tamoxifen and raloxifene, are approved for reduction of breast cancer risk in high-risk women [3]. In addition, new types of treatment such as high-dose chemotherapy with stem cell transplant is being tested in clinical trials.

1.2 Molecular Classification, Clinical Stages and Prognosis

Four main breast cancer classes were identified in a comprehensive study of human breast tumors based on mRNA expression profiles [4]: Luminal subtypes, HER2-enriched and Basal subtypes. The luminal subtypes are characterized as luminal A and luminal B. They are the most common subtypes of breast cancer and make up the majority of ER-positive breast cancers. The name "luminal" derives from similarity in gene expression between these tumors and the luminal epithelium of the breast. They typically express cytokeratins 8 and 18. The HER2-enriched subtype makes up about 10 to 15 percent of breast cancers and is characterized by high expression of HER2 and proliferation gene clusters and low expression of the luminal and basal gene clusters. These tumors are often negative for ER and PR. In basal-like tumors, most of these tumors fall under the category of triple-negative breast cancers because they are ER, PR and HER2 negative [4].

However, since breast cancers differ in many ways, such as in their cell of origin, the molecular alterations causing them and the susceptibility and defenses of the patient, and this makes it difficult to give the most appropriate treatment based on the molecular portraits [9]. Recent studies suggested using histologic grade can greatly reduce the heterogeneity of breast cancer outcome predictions [10]. The Bloom-Richardson breast cancer staging system has now basically been subsumed as the American Joint Committee on Cancer (AJCC) classfication system [11]. The paper Sotiriou et al used Elston grading system. The Elston-Ellis breast cancer grading system was a modification of the original Bloom-Richardson system, and is still in use in many places in Europe [12]. There are three factors that the pathologists take into consideration:

1. The amount of gland formation;
2. The nuclear features, i.e. the degree of nuclear pleomorphism;
3. The mitotic activity [12]

Based on Elston grading, histological grade 1 have the most favorable outcome and histological grade 3 having a poorer outcome. When histological grades are compared with survival time, histologic grade 1 tumor patients have a lower tumor recurrence rate or are a low risk group. In contrast histologic grade 3 tumor patients have a high tumor recurrence rate or are a high risk group. Grade 2 tumors have been difficult to classify based on their intermediate or unclear appearance and as a result grade 2 offers variable prognosis. Grade 2 tumors consist of a substantial number of tumors (30-60%) making classification of these tumors extremely important. In this paper Sotiriou et al. developed a 97-gene signature which resulted in the classification of grade 2 tumors into low or high recurrence risk based on GGI [9].

2 Purpose and Hypothesis

Since Grade 2 phenotypes are often associated with an intermediate prognosis and treatment recommendations for this phenotype are ambiguous, the goal of this paper is to determine if a gene signature can distinguish histologic grade 2 cancers into high and low risk phenotypes. We had 3 primary aims:

1. Validate the Sotiriou et al findings using the genomics portal data base

2. Determine a gene signature that will distinguish grade 1 and 3 tumors based on survival analysis

3. Identify biological pathways implicated by our gene signature

We hypothesize that our gene signature will distinguish grade 2 tumors with a good (grade 1 like) prognosis form grade 2 tumors with a poor (grade 3 like) prognosis and that biological pathways associated with our gene signature will primarily consist of cell cycle regulation processes. The study could improve our ability to predict breast cancer behavior , which is an essential step towards providing individualized treatment plan.

3 Methods and Materials

3.1 Data Acquisition and Sampling

Five gene expression datasets were used in this study which were obtained by microarray analysis (using Affymetrix U133A Genechips) of tumor samples from 661 patients with primary breast cancer.
They are enumerated below:

- The training set KJX64.
- The validation set KJ125.
- The National Cancer Institute (NCI) dataset from Sotiriou et al. [13]
- The Stanford/Norway (STNO) dataset from Sorlie et al. [14]
- The Nederlands Kanker Instituut (NKI) 2 dataset from Van de Vijver et al. [15]

Histologic tumor grade in these datasets was based on the Elston – Ellis grading system [12] and the standardized mean difference of Hedges and Olkin [16] was used to rank genes by their differential expression. The probe sets of the Affymetrix U133A Genechips were mapped to other microarray platforms by matching the Unigene identifiers (version 180), following the technique mentioned in Praz et al. [17].
For gene expression analysis, RNA was isolated by utilization of TRIzol, which is a monophasic solution of phenol and guanidinium isothiocyanate, used for solubilizing biological material and denaturation of proteins. Agilent Bioanalyzer was used for optimization and quality control of the RNA obtained from all those tumor samples and only the ones with good quality of RNA were acknowledged for further examination.

For the purpose of our analysis, while selecting the grade-associated genes, we only considered the ER-positive tumors and excluded the ones with ER-negative and NA status on account of the relationship between the ER status and histologic grade. As mentioned in the Sotiriou et al. paper [1], majority of the ER-negative tumors are grouped as either intermediate (Grade 2) or high (Grade 3) histologic grade, assuming that we had utilized all histologic grade 1 and 3 tumors despite their ER status during our analysis, we might have ended up choosing ER-identified genes that were falsely connected with grade.

The Sotiriou paper had 2 separate sets of previously unpublished tumor samples. The first set (KJX64) consisted of 64 tumors. These samples were all ER positive and nearly equally distributed between histologic grade 1 (n=33) vs. grade 3 (n=31). These patients all received tamoxifen therapy, but specimens were obtained prior to therapy initiation. The second set (KJ125) consisted of 125 tumor samples, which included both ER positive (n=86) and ER negative (n=37) tumors. The histologic grade was well distributed between grade 1 (n=34), grade 2 (n=46), or grade 3 (n=28), and 17 samples did not have histologic grade known. Importantly, none of these patients received systemic treatment.

3.2 Differential Expression Modeling Methods

Our method of assessing histological grade of primary breast cancer consists of three phases. The first phase is a semi-supervised ranking of genes and grade 1 and grade 3 tumors based on maximum differential expression. The second phase consists of selecting a set of genes that are up-regulated with grade 3 tumors and down regulated in grade 3. In the third phase, we generated two groups of tumor samples. Classification of the grade 2 tumors is then performed using these groups as models. Below we expand upon our process in detail.

The gene signature identification process consisted of assessing the Sotiriou breast cancer data set with Genomics Portals (genomicsportals.org). Genomics portals performs differential analysis on genes and samples using the CLEAN algorithm [18]. CLEAN co-clusters the genes and sample grades hierarchically, using data from the dataset and a database of known gene functional groups. Highly differentially expressed genes were identified with a p-value of .001, and fold change over expression level of 2. The results of this analysis were inspected in genomics portal's Treeview navigator [19], and the visually most cohesive up-regulated set of genes corresponding to grade 3 tumors were selected as our gene signature Figure 1. Inspection with Treeview failed to suggest a visually significant cluster of down-regulated genes for grade 3 tumors, therefore only up-regulated ones were included in our model. We acquired a set of 49 differentially expressed up-regulated grade 3 tumor genes from this analysis. As expected most of those genes turned out to be cell proliferation and cell cycle regulators, notably among them were:

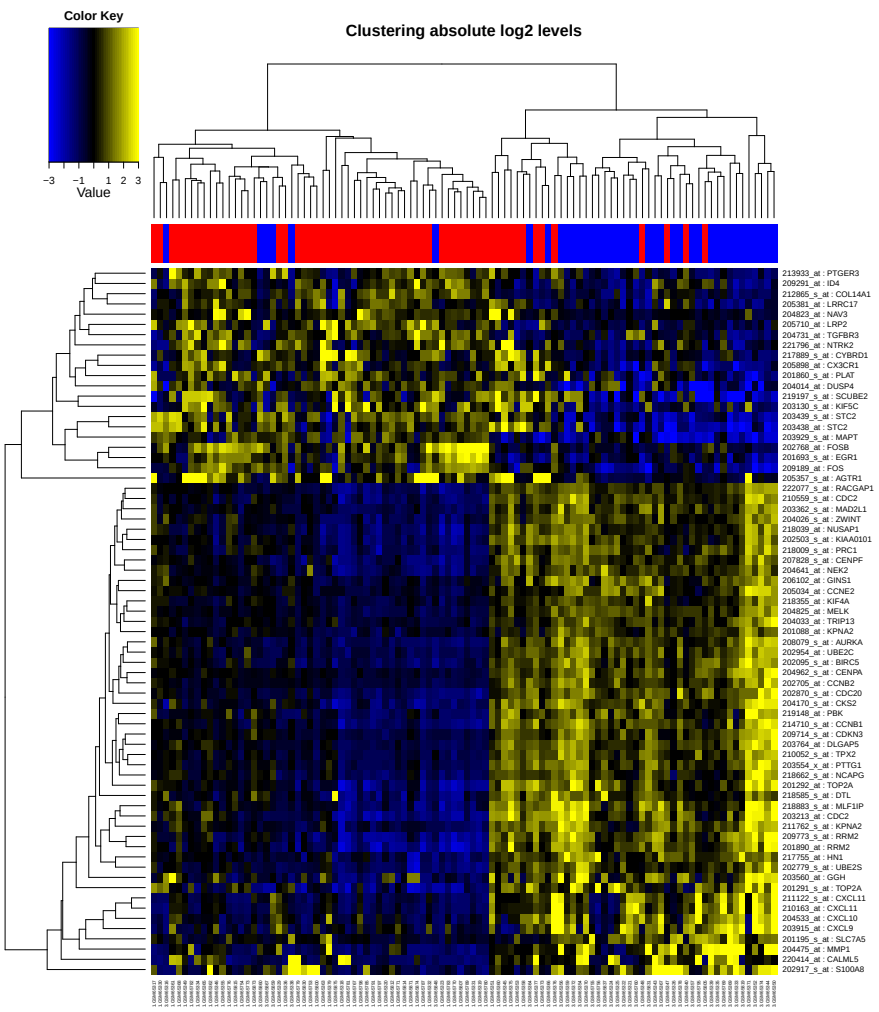

Figure 1: Grade 1 and Grade 3 Differentially Expressed.

- Chemokines (known to mediate cell migration and proliferation)
- Cyclins (control the progression of cells through the cell cycle by activating cyclin-dependent kinase-cdk enzymes)
- Cyclin-dependent kinase inhibitors (interact with, and dephosphorylate CDK kinase, thus preventing their activation; reported to be deleted, mutated, or over-expressed in several kinds of cancers)
- Aurora Kinase A/ Breast Tumor-Amplified Kinase (involved in microtubule formation and stabilization at the spindle pole during chromosome segregation, may play a role in tumor development and progression).
- Ubiquitin-conjugating enzyme E2S (required for cell cycle progression by degradation of mitotic cyclins, consequently, may be involved in cancer progression).

$$Centroid(Grade_X) = \left[\sum_{x \in X} \frac{x_1}{|X|}, \sum_{x \in X} \frac{x_2}{|X|}, \ldots, \sum_{x \in X} \frac{x_{|genesignature|}}{|X|} \right]$$

Figure 2: Grade Gene Expression Centroid Generation.

Using the selected set of 49 genes as a signature we supposed a |gene signature|−dimensional subspace to build our model. Withing the subspace, we used the clinically provided grade labels to generate centroids for the two , 1 and 3 - grade, groups following the expression (Figure 2) . Noise suppression refinement was performed by removing samples that deviated from the group cluster centers by more than 2 standard deviations of the cluster's mean variance.

For visual assessment of our analysis, we queried the grade 2 tumors against the grade 1 and 3 differentially expressed genes. The resulting chart shows a distinct clustering between the grade 1-like and grade 3-like tumors that follow approximately the grade 3 and grade 1 distributions. Assuming the dataset comprises the actual distribution of grade 1 to grade 3 tumors, approximately 55:64 for grade 1 and grade 3 respectively (Figure 3) we reassert our hypothesis of the lack of a grade 2 class distinction.

Classification of grade 2 tumors was then computed by nearest centroid under the euclidean norm following the method in Figure 4. Other distance norms could easily be substituted in place of euclidean, however in this analysis we assumed expression levels were orthogonal and linear. The new classifications were then used to recompute the KM-Survival analysis on tumor grades 2A (or grade 1-like) and 2b (or grade 3-like).

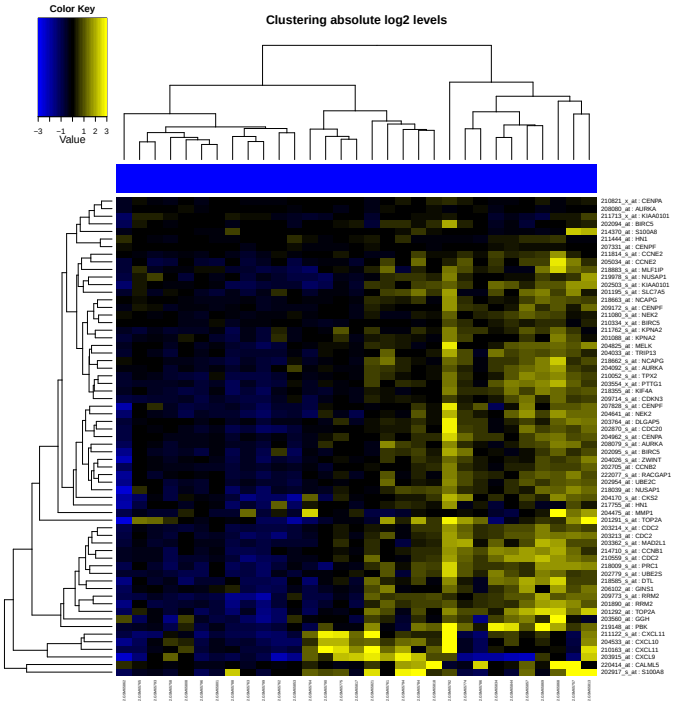

Figure 3: Grade 2 Differentially Expressed Genes from Grade 1 and Grade 3.

3.3 Unsupervised Machine Learning Methods

$$Grade(X) = \underset{C \in Centroids}{\operatorname{argmin}} \left(\sqrt{\sum_{g \in genelist} (X_g - C_g)^2} \right)$$

Figure 4: Nearest Centroid Classification.

In addition to the use of genomics portals to identify differentially expressed genes, two common unsupervised machine learning techniques were used to classify histological grades based on genetic expression. The first method used was the K-Means method. We will forgo an exhaustive explanation of K-Means here as it is a long known method with many descriptions and analysis available online. K-Means is an unsupervised clustering method that attempts to minimize the inter-cluster distance on a set of N points and K cluster assignments. The method is greedy and iterative, with few guarantees on optimality, however combined with random starts, the overall performance of K-Means is actually quite good for spherical clusters [20]. For our testing, following our aforementioned hypothesis of the existence of only two genetically identifiable tumor grades (low and high), we performed 100 tests using K-Means on 50:50 training:test data comprising grade 1 and grade 3 tumors. The reclassification accuracy results are given below.

Mean	0.8382
Mode	0.84
Min	0.58
Max	0.94
Var	0.00315

From the results it is clear there is significant error in the upper and lower bounds for accuracy. This though partially due to the K-Means algorithm itself, may have resulted from a badly conditioned training and test dataset split. As 50% grade 1 and 50% grade 3 would suggest the optimal split, we assert that the actual min accuracy performance is slightly lower than the real performance of K-Means, while the max is no better than could ideally be achieved from random 50:50 splitting.

The second unsupervised method, Projection to Latent Structures Regression (or Partial Least Squares Regression, *PLSR*) [21] proved to be slightly less accurate in its ability to correctly classify data. Despite its accuracy performance shortcomings, it did however perform well in its ability to identify highly differentially expressed genes. The explanation of these results is likely due to PLSR's maximization of the covariance between input data and a set of continuous output training vectors. Though the histological grade is a well ordered set, it is not continuous, and therefor fails to exploit some of the benefits of a covariance maximization algorithm. The top grade to expression covariance genes (loading vectors of PLSR) are listed below:

CALML5	51806	calmodulin-like 5
PBK	55872	PDZ binding kinase
RRM2	6241	ribonucleotide reductase M2
NEK2	4751	NIMA (never in mitosis gene a)-related kinase
CCNE2	9134	cyclin E2

4 Results and Discussion

In order to validate the aforementioned analysis, we created a set of Kaplan-Meier (KM) survival curves. These curves show that our gene sets and classification methods provide results that are at least as successful as those in the Sotiriou paper at separating the grade 2 tumors into two disjoint sets. For context, figure 5 shows three KM curves from the Sotiriou paper: Graph 5 shows the original data from their analysis. Grade 1, 2, and 3 tumor curves clearly separate from each other, clearly creating 3 separate and distinct classes. However, we would like to demonstrate that grade 2 tumors should actually be classified as either grade 1 or grade 3.

This is analysis is begun in graph B, as Sotiriou et al. used something they called the "Gene Expression

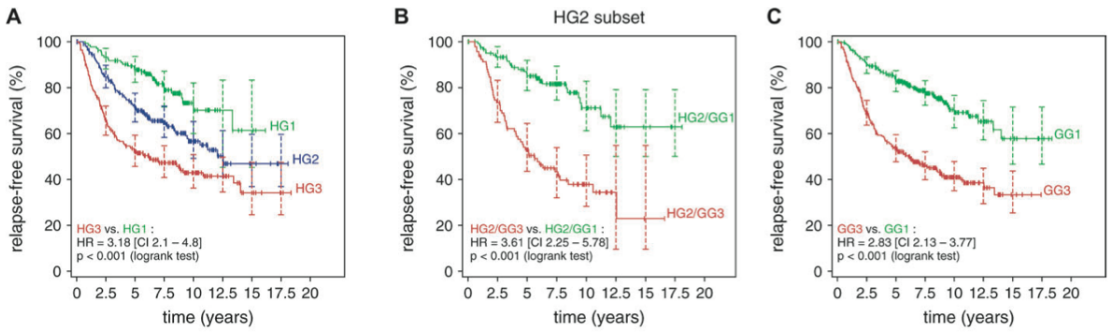

Figure 5: Kaplan-Meier curves from Sotiriou, et al.

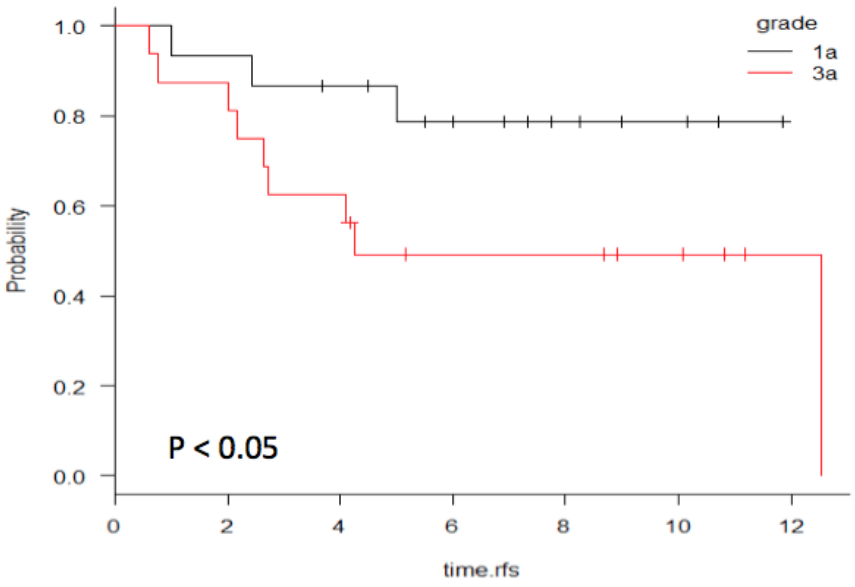

Figure 6: Reclassification of grade 2 samples using nearest-centroid method.

Grade Index" (GGI) to reclassify grade 2 tumors. This index is shown below:

$$GGI = scale\left(\sum_{j \in G_3} x_j - \sum_{j \in G_1} x_j - offset\right)$$

Gene Expression Grade Index, a tumor reclassification formula If the value of this formula was negative for a grade 2 sample, the sample would be reclassified as a grade 1 tumor, while a positive value would indicate a grade 3 tumor. Graph C shows all samples reclassified using the GGI. In our analysis, we chose a different re-classification technique. This was described above as the nearest-centroid method. We reclassified each grade 2 sample to either grade 1 or 3 to produce the graph in Figure 6.

Using Figure 6 we were able to see a difference between our re-graded curve and the original. Our reclassification actually intensified the gap between the grade 1 and 3 curves. Figure 7 shows the integration of grade 2 reclassifications in Figure 6 with the original grade 1 and 3 samples. That is, the curve labeled "grade 1" in Figure 7 includes all of the original grade 1 samples and grade 2 samples that were classified as "1-like", while the grade 3 curve contains the original grade 3 samples and the grade 2 samples that were classified as "3-like". The graph is analogous to the Sotiriou analysis shown in Figure 7-C.

For a control, we thought it might be interesting to run a similar analysis using the PAM50 geneset, a list of genes known to be differentially expressed across breast cancer tumor subtypes. The reclassification using the PAM50 genes (and nearest-centroid) is shown in the KM curve in Figure 8.

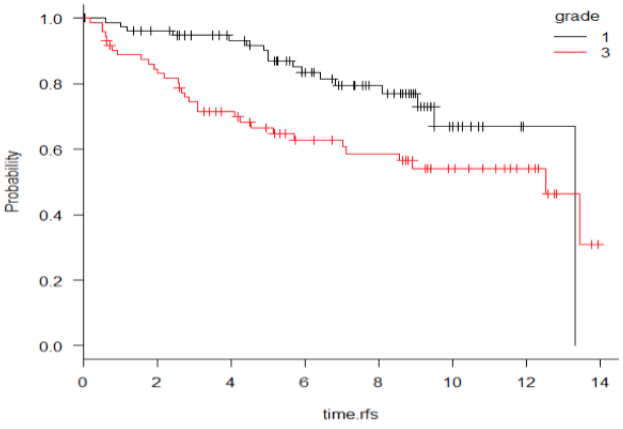

Figure 8: All samples reclassified using PAM50 genes.

4.1 Results

Baseline KM curves Based on the study's data without any reclassification, KM survival analyses

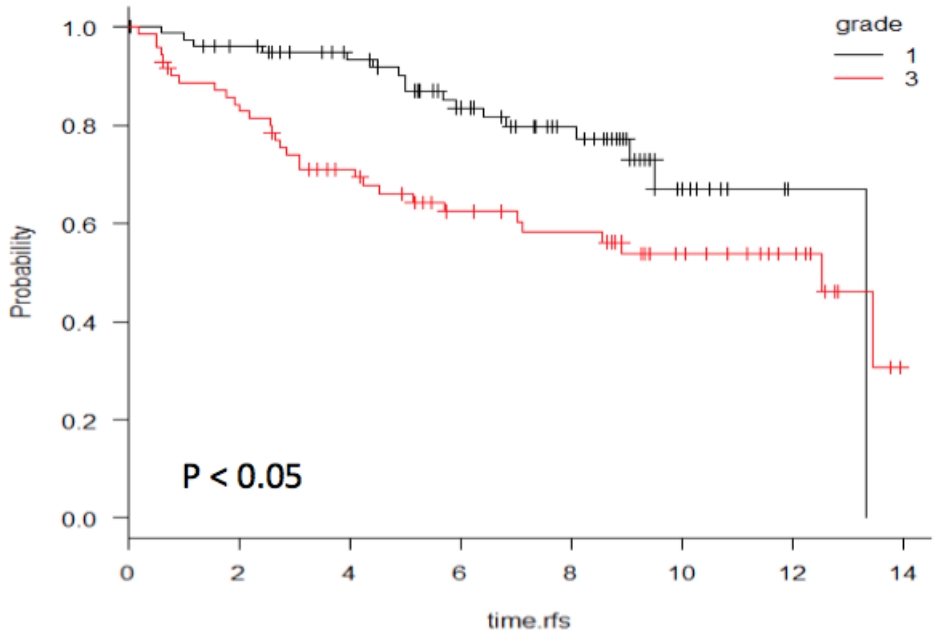

Figure 7: All samples reclassified using the nearest-centroid method.

with outcome defined as relapse free survival were performed for samples in the KJ125 set. KM analysis was performed for ER positive samples versus ER negative samples in the cohort (Figure 9). The ER negative samples tended to have worse prognosis than ER positive samples although the difference did not reach statistical significance. This trend is consistent with literature evidence that ER negative tumors are more malignant.

A KM curve was next generated for histologic grade among samples in the KJ125 cohort. Histologic grades 2 and 3 were associated with worse prognosis with regard to relapse free survival than grade 1 (Figure 10). Interestingly, there was no observed difference between grade 2 and 3 tumors.

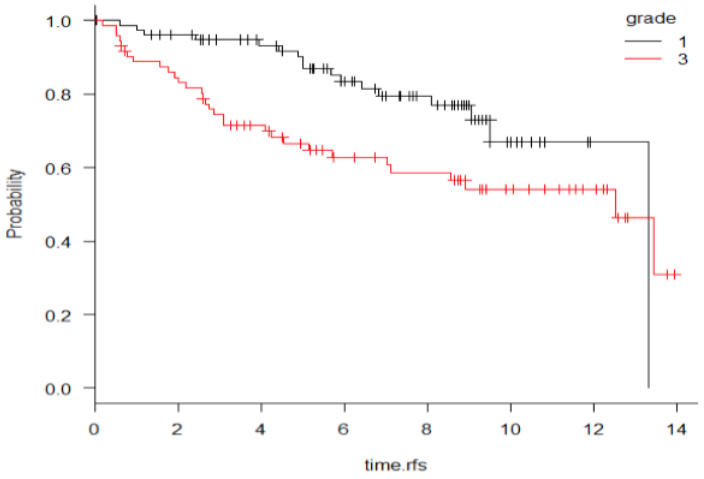

Figure 9: KM curve of relapse free survival over time in ER positive versus ER negative samples.

Unsupervised clustering between grade 1 and grade 3 tumors within only the KJX64 set was performed, which identified a list of 30 probe sets comprising 29 genes that are differentially expressed between grade 1 and grade 3 tumors (Table 1). Specifically, these genes were upregulated in histologic grade 3 tumors and downregulated in histologic grade 1 tumors.

The centroid-based approach utilizing calculated Euclidean distances was applied to the KJ125 samples. Figure 11 depicts a KM analysis for relapse free survival based on centroid-based reclassification of all KJ125 samples. This includes histologic grades 1, 2, 3, and those with histologic grade not available. Grade 1-like tumor samples were associated with significantly lower risk of disease relapse. Figure 12 depicts the KM analysis for centroid-based reclassification of only grade 2 tumor samples in the KJ125 set and also demonstrates better prognosis for those samples reclassified as grade 1-like.

202095_s_at	BIRC5	332	baculoviral IAP repeat-containing 5
214710_s_at	CCNB1	891	cyclin B1
210559_s_at	CDC2	983	cell division cycle 2, G1 to S and G2 to M
202870_s_at	CDC20	991	cell division cycle 20 homolog (S. cerevisiae)
209714_s_at	CDKN3	1033	cyclin-dependent kinase inhibitor 3
204962_s_at	CENPA	1058	centromere protein A
204170_s_at	CKS2	1164	CDC28 protein kinase regulatory subunit 2
203744_at	HMGB3	3149	high-mobility group box 3
211762_s_at	KPNA2	3838	karyopherin alpha 2 (RAG cohort 1, importin alpha 1)
203915_at	CXCL9	4283	chemokine (C-X-C motif) ligand 9
201890_at	RRM2	6241	ribonucleotide reductase M2
209773_s_at	RRM2	6241	ribonucleotide reductase M2
208079_s_at	AURKA	6790	aurora kinase A
201291_s_at	TOP2A	7153	topoisomerase (DNA) II alpha 170kDa
201195_s_at	SLC7A5	8140	solute carrier family 7 (cationic amino acid transporter, y+ system), 5
218009_s_at	PRC1	9055	protein regulator of cytokinesis 1
202705_at	CCNB2	9133	cyclin B2
205034_at	CNE2	9134	cyclin E2
203554_x_at	PTTG1	9232	pituitary tumor-transforming 1
204033_at	TRIP13	9319	thyroid hormone receptor interactor 13
204825_at	MELK	9833	maternal embryonic leucine zipper kinase
206102_at	GINS1	9837	GINS complex subunit 1 (Psf1 homolog)
202954_at	UBE2C	11065	ubiquitin-conjugating enzyme E2C
204026_s_at	ZWINT	1130	ZW10 interactor
210052_s_at	TPX2	22974	TPX2, microtubule-associated, homolog (Xenopus laevis)
204086_at	PRAME	23532	preferentially expressed antigen in melanoma
202779_s_at	UBE2S	27338	ubiquitin-conjugating enzyme E2S
222077_s_at	RACGAP1	29127	Rac GTPase activating protein 1
218039_at	NUSAP1	51203	nucleolar and spindle associated protein 1
218883_s_at	MLF1IP	79682	MLF1 interacting protein

Table 1: List of 30 probe sets comprising 29 genes which are upregulated in KJX64 histologic grade 3 samples and downregulated in grade 1 samples.

Next, the list of the 29 differentially expressed genes in the above table was queried against the RNAseq gene expression data from by PAM50 subtype in TCGA [4]. The clustering is demonstrated in Figure 13. There is a cluster of 108 samples with increased expression of the genes of interest, which can be considered grade 3-like. Interestingly, nearly half of the samples in this cluster were basal-like tumors based on PAM 50 signature (Table 14). In addition, 64% of all basal-like tumors in the analysis were found in this cluster. Furthermore, there were only 3 luminal A subtype tumors within the grade 3-like cluster which comprised only 1.4% of luminal A tumors in this dataset. These findings are consistent with observations that luminal A subtype tumors are less aggressive than basal-like tumors and serves as a validation of our generated gene list. The TCGA paper does not provide survival data, which would be another interesting validation to be explored in the future.

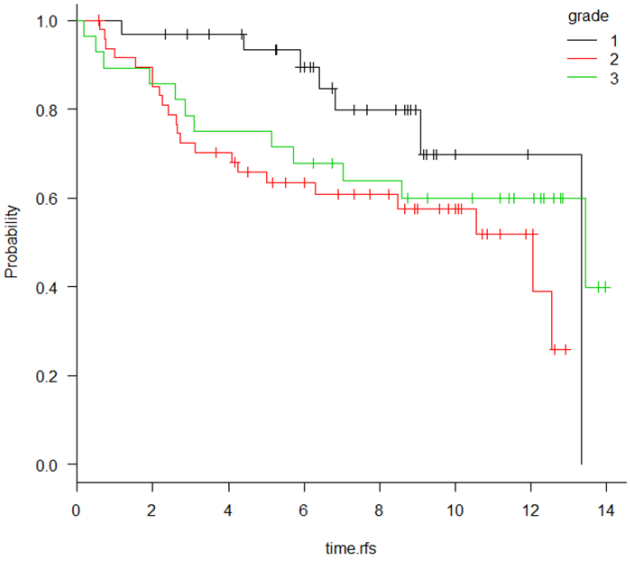

Figure 10: KM curve of relapse free survival over time in groups based on histologic grades.

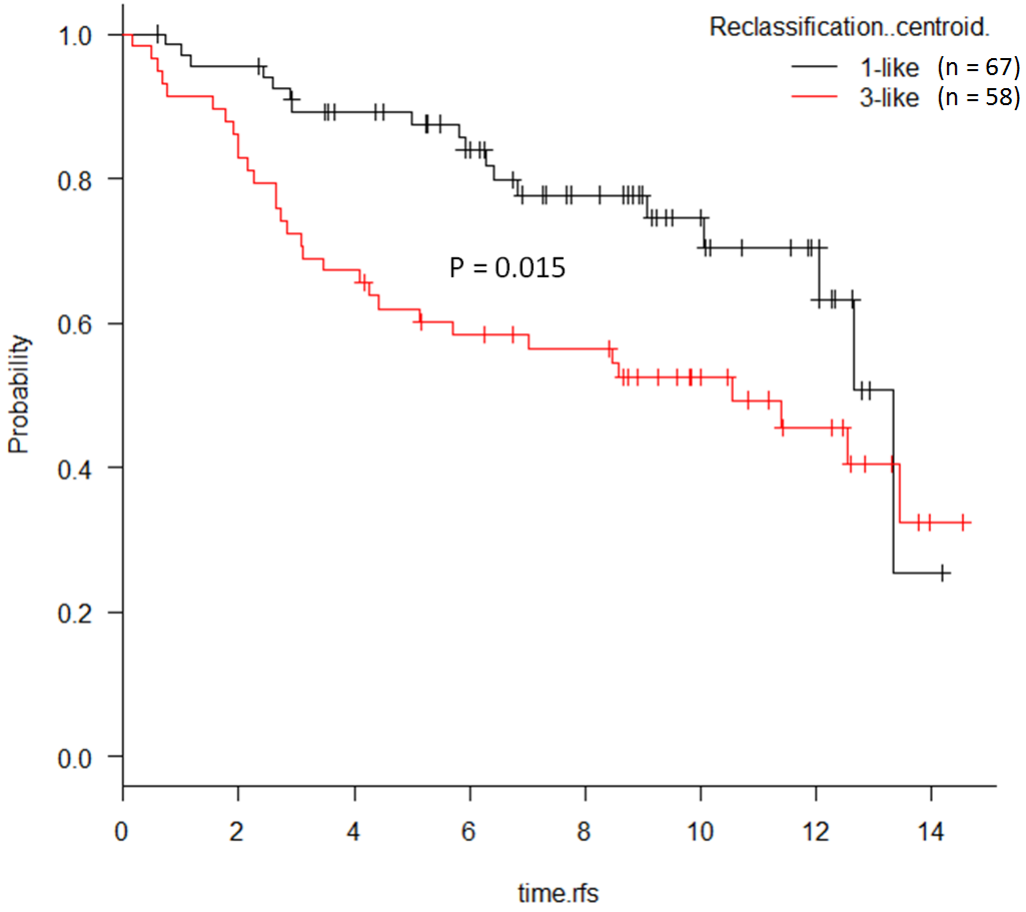

Figure 11: KM analysis of relapse free survival among all KJ125 tumors reclassified as 1-like or 3-like based on microarray gene expression profiles.

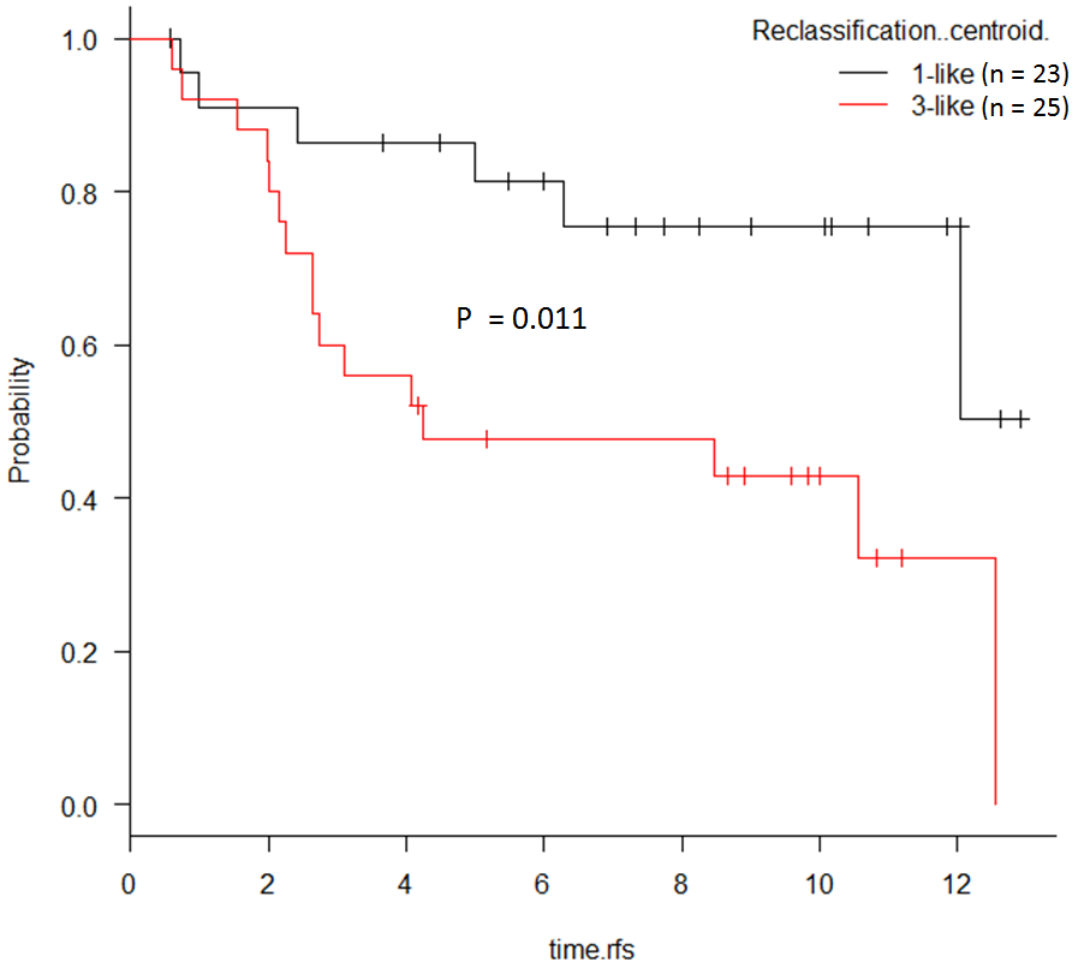

Figure 12: KM analysis of relapse free survival between histologic grade 2 tumors in KJ125 set reclassified as 1-like or 3-like based on gene expression profiles.

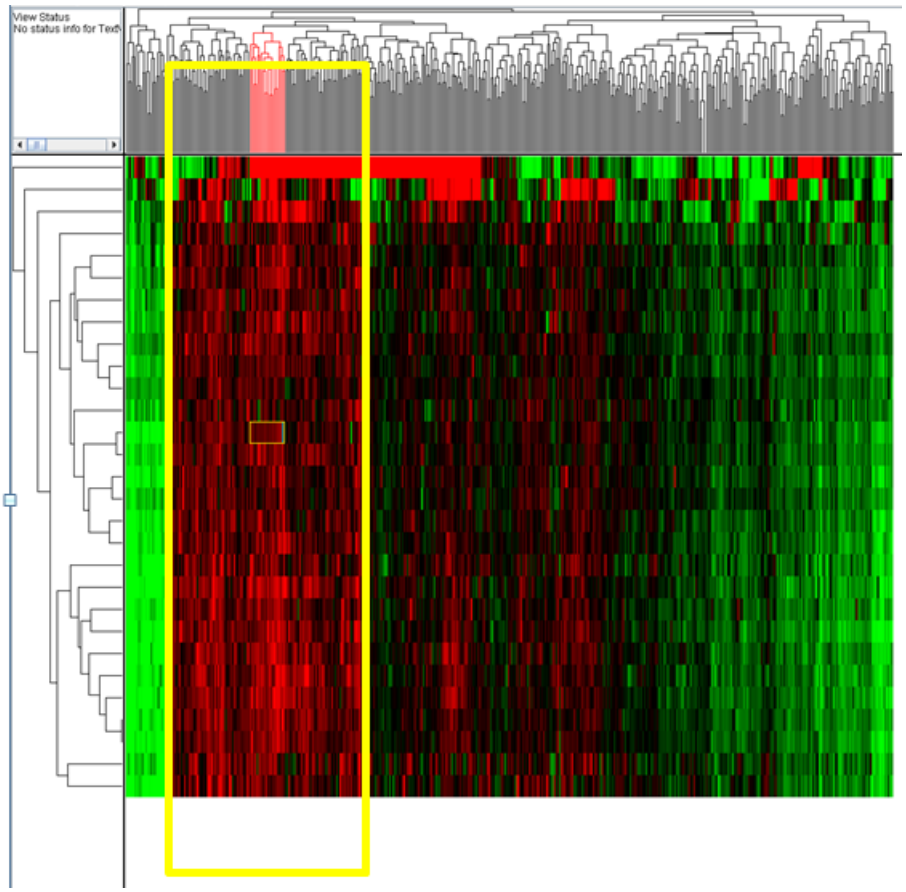

Figure 13: Query of TCGA RNAseq gene expression data with our generated list of genes based on Sotiriou microarray expression profiles. There is a cluster of increased gene expression (yellow box) containing 108 samples which can be considered grade 3-like.

We further explored the gene list of any functional annotation utilizing resource tools *DAVID* and *KEGG* [22] [23] [24] [25]. These analyses found that our gene list was functionally enriched for 3 pathways: cell cycle regulation, progesterone-mediated oocyte meiosis, and ubiquitin-mediated protein degradation in figures 15, 16, and 17. These pathways are consistent with the known features of advanced histologic grade such as mitosis and dedifferentiation. The majority of genes in our list overlapped with the list generated by Sotiriou et al. Several of the selected overlapping genes are shown in Table 2. Notably, these genes are known to be associated with oncologic processes in multiple other organs according to queries of the GeneCards web tool (www.genecards.org).

Subtype	**Number of tumors (N=434)**	**Number of tumors in cluster of increased expression (N=108)**	**Percent**
Normal	6	2	33%
LumA	201	3	1.4%
LumB	101	35	34.6%
Her2	47	17	36.2%
Basal-like	79	51	64.6%

Figure 14: Query of TCGA RNAseq gene expression data with our generated list of genes, clustered by PAM50 subtype. The cluster of increased expression (grade 3-like) has a preponderance of Basal-like tumors and paucity of luminal A tumors.

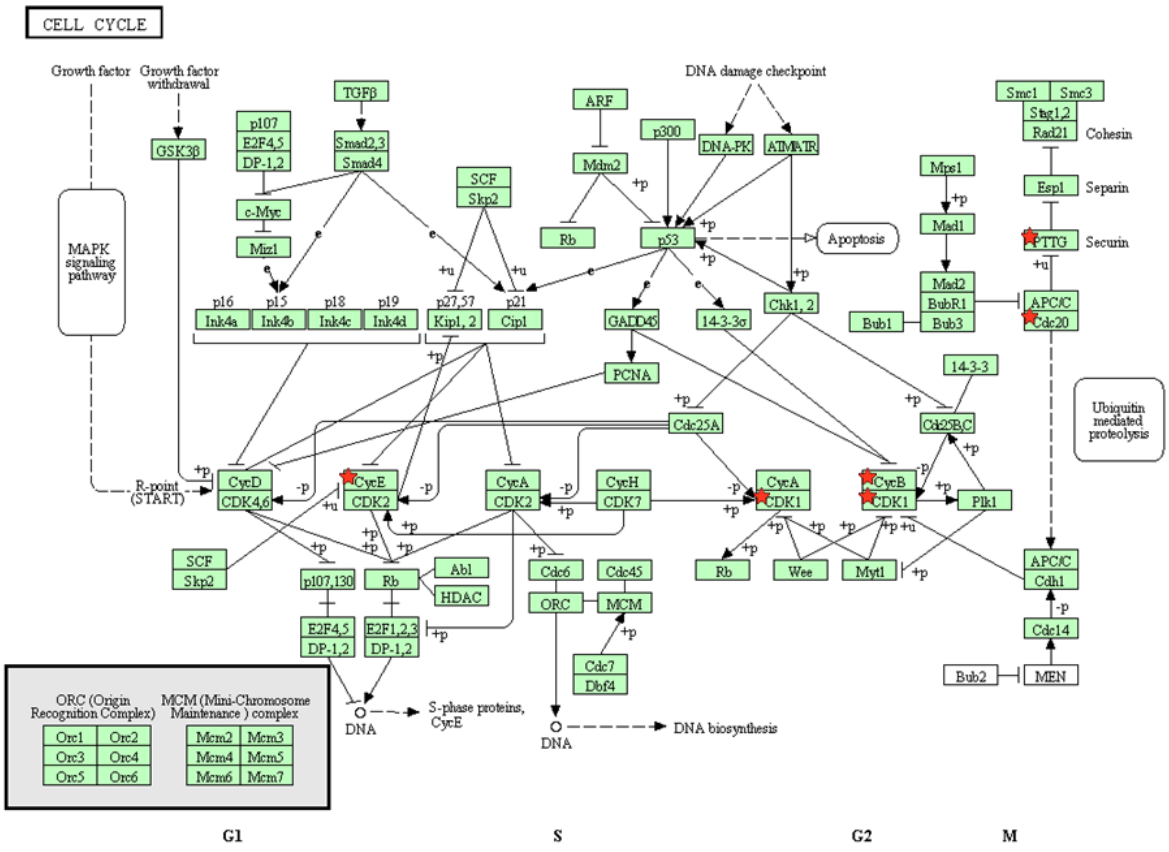

Figure 15: List of differentially upregulated genes in histologic grade 3 tumors is enriched for the pathway of cell cycle regulation.

Gene	Protein	Function	Associated diseases
TPX2	TPX2, microtubule -associated, homolog (Xenopus laevis)	Required for assembly of mitotic spindles	Hepatocellular carcinoma Colorectal cancer
CCNB1	Cyclin B1	Initiates mitosis	Many cancers
CDC20	cell division cycle 20 homolog (S. cerevisiae)	Cell cycle regulation	Ataxia-telangiectasia (↑ cancer risk)
PTTG1	pituitary tumor-transforming 1	Chromosome stability, negatively regulates TP53	Thymic hyperplasia Nelson syndrome (pituitary adenoma)
UBE2C	ubiquitin-conjugating enzyme E2C	Targeting proteins for degradation	Esophageal carcinoma Papillary thyroid carcinoma
KPNA2	karyopherin alpha 2 (RAG cohort 1, importin alpha 1)	Nuclear protein import	Nijmegen breakage syndrome (↑ cancer risk)
BIRC5	baculoviral IAP repeat-containing 5	Inhibits apoptosis and promotes cell proliferation	Cervical cancer Laryngeal carcinoma

Table 2: Selected genes of interest which were identified to be upregulated in Grade 3 tumors in our analyses as well as the analyses by Sotiriou. These genes are recognized to be associated with other oncologic processes and systemic disorders associated with cancer risk.

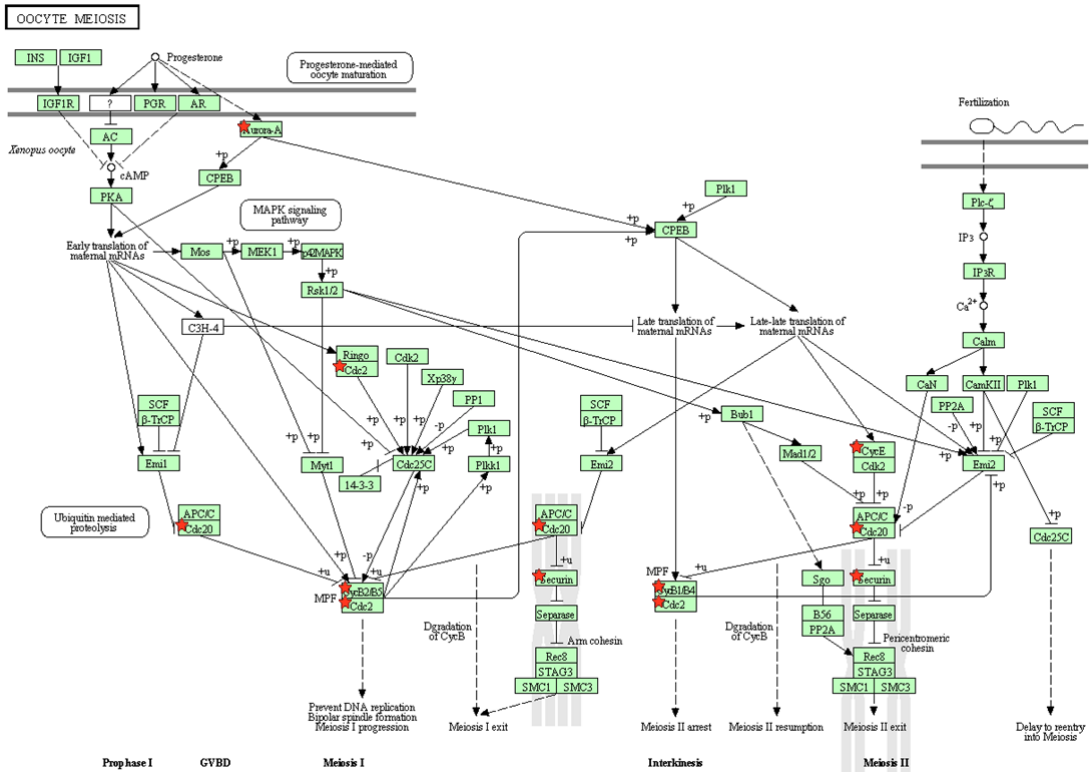

Figure 16: List of differentially upregulated genes in histologic grade 3 tumors is enriched for the pathway of progesterone-mediated oocyte maturation.

4.2 Discussion

As shown above, our discovered geneset and reclassification method were able to split grade 2 tumors into two disjoint sets. These sets may be more helpful in tumor diagnosis and treatment, as grade 2 diagnoses are currently not considered "not informative for clinical decision making" [13]. Our methods were able to produce a result that was slightly better than the Sotirou paper at a significant p-value (p=0.05). In addition, using the PAM50 geneset, we were able to reproduce these results and achieve an almost identical graph. This was unexpected, as PAM50 genes are known to separate samples based on their subtype rather than their grade. However, it is notable to mention that grade 1 and 3 tumors are more likely to be Basal and Luminal tumors, respectively. This explains the similarity between our and the PAM50 results.

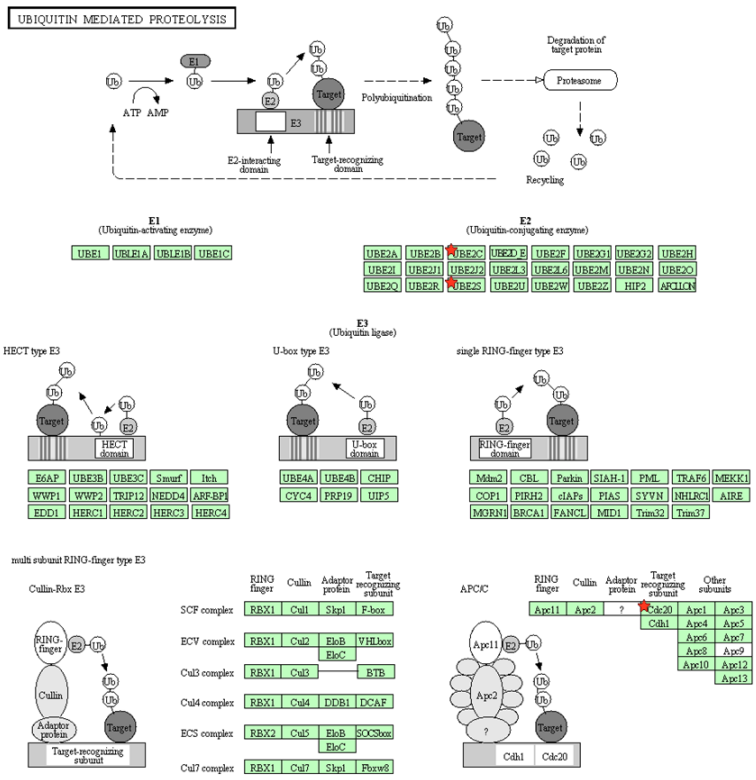

Figure 17: List of differentially upregulated genes in histologic grade 3 tumors is enriched for the pathway of ubiquitin-mediated protein degradation.

Limitations to the study include pathologist variation and possible misclassification, region or race specific differences since two populations are presented, and complications to analysis with ER status.

In our study, we only used estrogen receptor positive tumors to look at the differentially expressed genes. Thus the conclusions are limited to only the ER positive tumors and inferences cannot be made as to ER negative tumors. Future work would include looking into the differentially expressed genes in both estrogen receptors positive and negative tumors and then compare the gene list between each other. This can shed light on which specific genes are dependent on the ER status of tumors.

In addition, looking at expression signatures based on AJCC staging classification systems may decease some data set variation. The main criteria of the AJCC staging systems include tumor size, lymph node status, and distant metastasis. Future studies should also examine age of onset, races and geographic regions to look for variability and to improve treatment for different groups.

4.2.1 Caveats

While we were able to analyze the given dataset in an unbiased manner, there was no way for us to control the data given to us. In that vein, we were unable to control the research population for tumor size, age, margin, ER status, or lymph node status.

On a similar note, there was no way for us to alter our survival analysis based on what treatments were being offered to each patient. The possibility that different patients were participating in different treatments throughout their lives is enough to alter the confidence intervals on our KM curves.

4.2.2 Future Work

Our work with the Sotirou paper was only able to reproduce their analysis and improve upon it slightly. There is still a large amount of work that could be done to further validate our methods, as well as to delve deeper in to the work done by both our teams.

Firstly, in order to prove our work we would like to perform validation in some other independent datasets. In addition, we could perform a prospective study at some point in the future. Unfortunately, this does seem unlikely.

Moreover, an easier way to validate our results would be to perform the same analysis using the Distant Metastasis Free Survival Rate (DMFS). Our work solely included Relapse Free Survival (DFS) data.

References

[1] C. Sotiriou, P. Wirapati, S. Loi, A. Harris, S. Fox, J. Smeds, H. Nordgren, P. Farmer, V. Praz, B. Haibe-Kains, C. Desmedt, D. Larsimont, F. Cardoso, H. Peterse, D. Nuyten, M. Buyse, M. J. Van de Vijver, J. Bergh, M. Piccart, and M. Delorenzi, "Gene expression profiling in breast cancer: Understanding the molecular basis of histologic grade to improve prognosis," *Journal of the National Cancer Institute*, vol. 98, no. 4, pp. 262–272, 2006.

[2] I. C. Henderson and A. J. Patek, "The relationship between prognostic and predictive factors in the management of breast cancer," *Breast Cancer Res. Treat.*, vol. 52, no. 1-3, pp. 261–288, 1998.

[3] C. DeSantis, R. Siegel, P. Bandi, and A. Jemal, "Breast cancer statistics, 2011," *CA: a cancer journal for clinicians*, vol. 61, no. 6, pp. 408–418, 2011.

[4] D. C. e. a. Koboldt, "Comprehensive molecular portraits of human breast tumours," *Nature*, vol. 490, pp. 61–70, Oct 2012.

[5] P. T. Simpson, J. S. Reis-Filho, T. Gale, and S. R. Lakhani, "Molecular evolution of breast cancer," *J. Pathol.*, vol. 205, pp. 248–254, Jan 2005.

[6] K. Komaki, N. Sano, and A. Tangoku, "Problems in histological grading of malignancy and its clinical significance in patients with operable breast cancer," *Breast Cancer*, vol. 13, no. 3, pp. 249–253, 2006.

[7] E. M. Kesson, G. M. Allardice, W. D. George, H. J. Burns, and D. S. Morrison, "Effects of multidisciplinary team working on breast cancer survival: retrospective, comparative, interventional cohort study of 13 722 women," *BMJ*, vol. 344, p. e2718, 2012.

[8] J. H. Howard and K. I. Bland, "Current management and treatment strategies for breast cancer," *Curr. Opin. Obstet. Gynecol.*, vol. 24, pp. 44–48, Feb 2012.

[9] F. Bertucci and D. Birnbaum, "Reasons for breast cancer heterogeneity," *J. Biol.*, vol. 7, no. 2, p. 6, 2008.

[10] L. W. Dalton, S. E. Pinder, C. E. Elston, I. O. Ellis, D. L. Page, W. D. Dupont, and R. W. Blamey, "Histologic grading of breast cancer: linkage of patient outcome with level of pathologist agreement," *Mod. Pathol.*, vol. 13, pp. 730–735, Jul 2000.

[11] H. J. BLOOM and W. W. RICHARDSON, "Histological grading and prognosis in breast cancer; a study of 1409 cases of which 359 have been followed for 15 years," *Br. J. Cancer*, vol. 11, pp. 359–377, Sep 1957.

[12] C. Elston and I. Ellis, "Pathological prognostic factors in breast cancer. i. the value of histological grade in breast cancer: experience from a large study with long-term follow-up. cw elston & io ellis. histopathology 1991; 19; 403–410," *Histopathology*, vol. 41, no. 3a, pp. 151–151, 2002.

[13] C. Sotiriou, S.-Y. Neo, L. M. McShane, E. L. Korn, P. M. Long, A. Jazaeri, P. Martiat, S. B. Fox, A. L. Harris, and E. T. Liu, "Breast cancer classification and prognosis based on gene expression profiles from a population-based study," *Proceedings of the National Academy of Sciences*, vol. 100, no. 18, pp. 10393–10398, 2003.

[14] T. Sørlie, C. M. Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen, T. Hastie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, T. Thorsen, H. Quist, J. C. Matese, P. O. Brown, D. Botstein, P. E. Lønning, and A.-L. Børresen-Dale, "Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications," *Proceedings of the National Academy of Sciences*, vol. 98, no. 19, pp. 10869–10874, 2001.

[15] M. J. van de Vijver, Y. D. He, L. J. van't Veer, H. Dai, A. A. Hart, D. W. Voskuil, G. J. Schreiber, J. L. Peterse, C. Roberts, M. J. Marton, M. Parrish, D. Atsma, A. Witteveen, A. Glas, L. Delahaye, T. van der Velde, H. Bartelink, S. Rodenhuis, E. T. Rutgers, S. H. Friend, and R. Bernards, "A gene-expression signature as a predictor of survival in breast cancer," *N. Engl. J. Med.*, vol. 347, pp. 1999–2009, Dec 2002.

[16] L. V. Hedges, I. Olkin, M. Statistiker, I. Olkin, and I. Olkin, "Statistical methods for meta-analysis," 1985.

[17] V. Praz, V. Jagannathan, and P. Bucher, "CleanEx: a database of heterogeneous gene expression data based on a consistent gene nomenclature," *Nucleic Acids Res.*, vol. 32, pp. D542–547, Jan 2004.

[18] J. Freudenberg, V. Joshi, Z. Hu, and M. Medvedovic, "Clean: Clustering enrichment analysis," *BMC Bioinformatics*, vol. 10, no. 1, pp. 1–15, 2009.

[19] K. Shinde, M. Phatak, F. Johannes, J. Chen, Q. Li, J. Vineet, Z. Hu, K. Ghosh, J. Meller, and M. Medvedovic, "Genomics portals: integrative web-platform for mining genomics data," *BMC Genomics*, vol. 11, no. 1, pp. 1–10, 2010.

[20] A. Jain, "Data clustering: 50 years beyond k-means," in *Machine Learning and Knowledge Discovery in Databases* (W. Daelemans, B. Goethals, and K. Morik, eds.), vol. 5211 of *Lecture Notes in Computer Science*, pp. 3–4, Springer Berlin Heidelberg, 2008.

[21] S. Wold, M. Sjöström, and L. Eriksson, "Pls-regression: a basic tool of chemometrics," *Chemometrics and Intelligent Laboratory Systems*, vol. 58, no. 2, pp. 109 – 130, 2001. ¡ce:title¿PLS Methods¡/ce:title¿.

[22] d. a. W. Huang, B. T. Sherman, and R. A. Lempicki, "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources," *Nat Protoc*, vol. 4, no. 1, pp. 44–57, 2009.

[23] M. Kanehisa, S. Goto, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe, "Data, information, knowledge and principle: back to metabolism in KEGG," *Nucleic Acids Res.*, Nov 2013.

[24] M. Kanehisa, "Molecular network analysis of diseases and drugs in KEGG," *Methods Mol. Biol.*, vol. 939, pp. 263–275, 2013.

[25] H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono, and M. Kanehisa, "KEGG: Kyoto Encyclopedia of Genes and Genomes," *Nucleic Acids Res.*, vol. 27, pp. 29–34, Jan 1999.