Prediction of clonal subtypes in breast invasive carcinoma

Hunter Boyce, Anna Shcherbina, Alice Yu

***Abstract*— Breast cancer is a highly heterogeneous disease that is classified into multiple clonal subtypes. Despite recent improvements in targeted treatment and clonal subtype detection methods, the complex genetics of tumor diversity remain poorly understood. This project examines genetic substructure in a population of breast cancer patients from The Cancer Genome Atlas (TCGA), and identifies networks of genes that correlate with specific clinical outcomes.**

I. Introduction

Breast cancer is the most commonly diagnosed cancer and the second leading cause of death among women[1][2]. The highly heterogeneous disease has many different clonal subtypes, characterized by tumor-specific sets of hormone receptors that result in different treatment responses. These receptors include Her2, Progesterone(PG), and Estrogen Receptors(ER), which all increase tumor growth speed and cause resistance to standard chemotherapy treatment[3][4][5]. Machine learning techniques can be applied to predict treatment response, disease severity, and survivability outcomes from a patient's genotype, and ultimately may elucidate ways to target treatment to better fit the patient.

The Cancer Genome Atlas (TCGA) is a rich database of genetic and electronic medical record information that can be mined to learn associations between clinical outcomes and patient genomes[6]. The breast invasive carcinoma dataset in TCGA is the largest and holds records for over 1000 cases of breast cancer. Patient clinical outcomes, such as menopause status and survival time post diagnosis (survivability) are also reported. However, the high dimensionality of the genomic data (input features) combined with the sparsity of the patient clinical and demographic data (outcomes) provides a challenge for data scientists to extract meaningful associations between the two. Supervised and unsupervised machine learning approaches make this problem more manageable. Researchers in the past have applied regression and clustering methods to genomic data to identify cancer driver mutations and therapeutic targets[7][8]. However, the causal agents of differing hormone receptor levels across patients remain poorly understood. This project applies unsupervised and supervised techniques to identify key features associated with distinct clonal cancer subtypes that can be used for prediction of clinical outcomes and provide leads for future genetics research.

II. Datasets

RNA expression level data (RNA-seq) and clinical metadata for 800 breast invasive carcinoma patients were downloaded (Fig. 1). The subjects were well-spread in age as well as clonal subtype and survivability. Most did not experience tumor metastasis and were postmenopausal, and the lack of positive training samples in these two categories led to their exclusion from analysis.

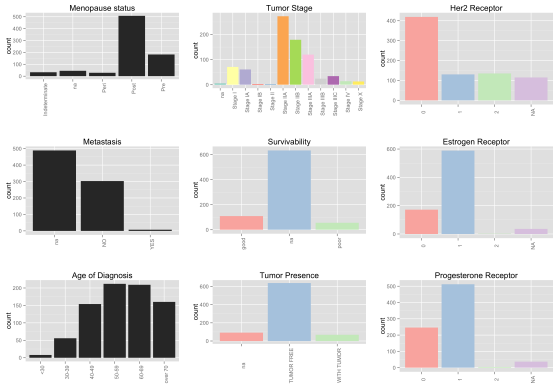


Fig. 1: TCGA subject clinical data (outcomes for analysis).

III. Feature Selection

A. L1 regularization and greedy forward selection approaches

The human genome contains an estimated 25,000 genes, but the dataset used in this project contains only 800 subjects, suggesting that the learning problem is over-constrained. Furthermore, the majority of human genes are not associated with breast cancer, which makes identifying the set of genes and pathways linked to the onset and outcome of breast invasive carcinoma more challenging. Consequently, supervised feature selection was performed to determine the genes associated with each outcome in Fig. 1.

The glmnet package in R was used to implement three approaches to feature selection[9] – elastic net regression[10], the group lasso[11], and forward subset selection via linear and logistic regression (for continuous and discrete outcomes, respectively). Greedy forward feature selection can be used to rank features (genes) in order of importance for determining an outcome of interest. However, multiple genes within a single pathway are likely to be correlated with certain outcomes. Because the expression levels of these genes are also correlated with each other, the forward selection technique will randomly select one feature(gene) from the group, since adding additional genes along the same pathway is unlikely to maximize the overall information content. The elastic net regression approach helps to avoid this pitfall, by using both ridge and lasso penalties to either join strongly correlated predictors or not at all:

$$f(x) + \lambda||x||_1 + (1-\lambda)||x||_2^2 \quad (1)$$

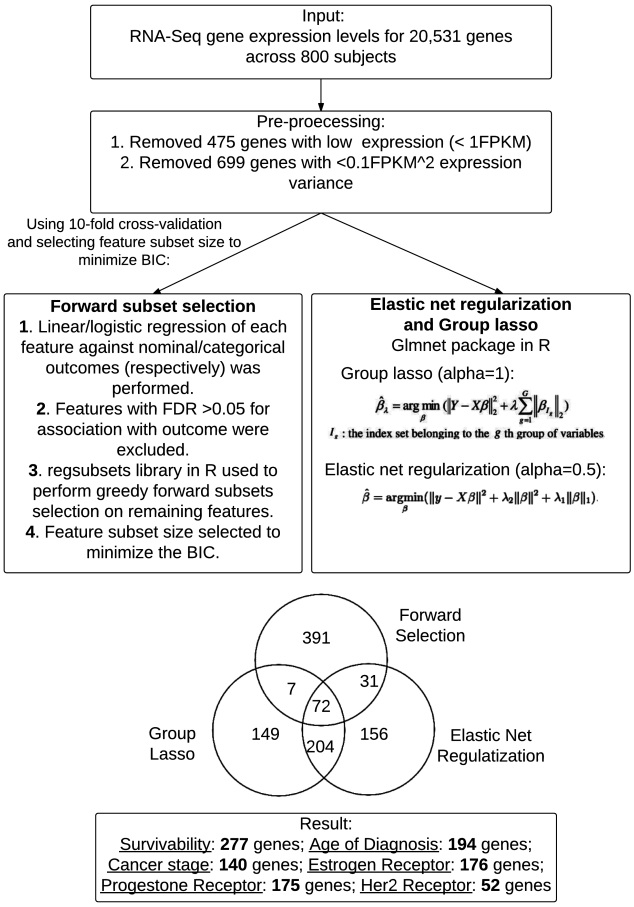


Fig. 2: Feature selection pipeline incorporating greedy forward selection, group lasso, and elastic net regression.

However, unlike forward selection, elastic net regression may not rank features accurately – weights for correlated features will be split by the number of features in the group, so that each of the features is weighted the equivalent of the group average. Similarly, the group lasso (ℓ_1, ℓ_2 mixed norm regularization), can be used to include or exclude entire groups of variables (i.e. gene pathways):

$$\text{minimize } f(x) + \lambda \sum_{i}^{N} ||x_i||_2 \tag{2}$$

However, the pathways must be specified *a priori* and the algorithm may fail to identify overlapping pathways, or pathways where only a small subset of genes are upregulated. Consequently, the union of all the feature selected from the three approaches were used (Fig. 2).

For each approach, selection was performed using 10-fold cross validation (training on 70% of the data and testing on 30% in each fold). Elastic net regression was performed with the alpha parameter set to 0.5. The glmnet package in R was used to perform a parameter sweep across lambda to identify the value that minimized the cross-validated mean error. For the group lasso, the alpha was set to 1, and a parameter sweep was again performed to determine lambda. For the survivability outcome, both the elastic net regression and the group lasso minimized all features to 0. To avoid this outcome, the chosen value of lambda was shifted by an amount that prevented 100 features from being reduced to 0.

Greedy forward selection was performed using the regsubsets library in the R leaps package [12]. To reduce the $O(n^2)$ runtime of the algorithm, a regression step was performed first on each individual feature against the outcome of interest, using the Bonferroni method to correct for multiple testing. Any feature that did not have a regression P-value below 0.05 was excluded from the analysis. Only the surviving features were used as input to the regsubsets library. Features were added sequentially in the order in which testing error on the holdout set was minimized.

B. Biological Significance of Feature Sets

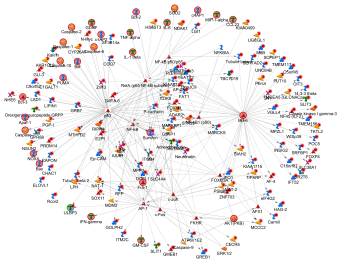
To ensure that the feature sets produced by feature selection were biologically relevant, they were analyzed with the Metacore software [13]. Given an input gene list, Metacore identifies associated gene networks and pathways. Three breast cancer-related networks from the Gene Ontology[14] were strongly associated with the down-selected feature sets (Fig. 3). Of the 107 genes in the apoptotic processes network, 87 were present in the survivability feature set. Similarly, 54 out of 61 genes in the reproductive system development network were present in the ER feature set, and 267 out of 321 genes in the growth factor receptor signaling network were present in several of the feature sets (survivability, ER, Her2, and age of diagnosis). These networks have strong associations with breast cancer in the literature. For example, the majority of the features(genes) within the apoptotic processes network interact with the p53 and Bcl-6 genes (Fig. 3a). Bcl-6 is an oncogene that is expressed in 68% of high-grade ductal breast carcinomas[15]. Similarly, the tumor suppressor gene p53 is the "most commonly altered gene in human cancer"[16]. The ESR1 gene at the center of the reproductive system development network (Fig. 4c) is the ER marker that defines the outcome of interest[17]. Though this discovery is circular, it serves as validation that the independent feature selection technique is choosing sets of biologically meaningful genes for data substructure identification. Finally, the HNF-1-alpha and NFK-beta genes that serve as hubs of the growth factor receptor signaling features are transcription factors that have been implicated in a number of different cancers[18][19].

IV. METHODS

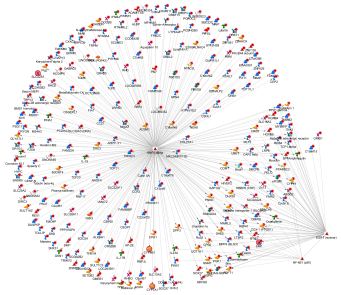
A. Unsupervised Methods

To identify genetic substructure within the data, unsupervised clustering was performed on each of the outcomes in Fig. 1. Each cluster was then correlated with clinical outcomes using Chi-Squared analysis for categorical outcomes and the ANOVA test for nominal outcomes. The highest-confidence clusters were obtained for the "survivability" and "ER" outcomes (Fig. 4).

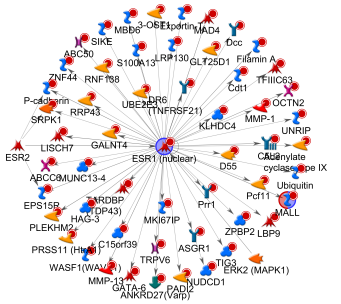
Hierarchical Clustering: Hierarchical clustering of subjects was performed with the Dynamic Tree Cut Package



(a) Regulation of apoptotic processes, FDR=1.7e-22,Z=9.87



(b) Growth factor receptor signaling, FDR=5.310e-9,Z=21.47



(c) Reproductive system development,FDR=1.76e-19, Z=9.10

Fig. 3: Feature selection identifies Gene Ontology networks and pathways of carcinoma-relevant genes. Genes that are members of the down-selected feature sets are marked by a red circle in the upper right corner.

from the R statistical software [20]. Rather than the fixed height branch cut of traditional hierarchical clustering, the Dynamic Tree Cut algorithm iteratively decomposes and combines nested clusters, stopping when the number of clusters becomes stable. Eight clusters (Fig. 4a) were found using the ER feature set, and two distinctive groups were identified using the survivability features (Fig. 4d). Individuals in the "blue" cluster survived an average of 304 days longer than those in the "red" cluster, as determined by ANOVA. These same individuals also overexpressed the SCUBE2 gene, a known breast tumor suppressor[21]. The correlation of this gene with the surviving patients supports its role in the suppression of tumor growth.

K-Means Clustering: The patients were clustered using k-means [22]. A Silhouette analysis was performed to determine the optimal value of k, and it was determined that k=2 provided the best data separation. Fig. 4 shows the two K-means clusters overlaid on the two most significant components from a PCA analysis of the data [23]. The ER features separated ER+ and ER- patients into two distinctive clusters. Similar results were seen in PG and Her2 Receptors. However, the survivability features only returned a small cluster of surviving patients and a mixture of patients in the other.

The overexpressed genes in the high-survivability clusters were associated with the cytokine-mediated signaling pathway (Fig. 4f). Similarly, the ER- cluster from both clustering methods had upregulated levels of genes associated with reproductive development (Fig. 4c).

B. Supervised Methods

A suite of supervised training models were implemented using the Python scikit-learn toolkit [24] to predict clinical outcomes from patient mRNA expression profiles.

Classification: There are multiple subtypes of breast cancer that are classified by the modification of expression levels of specific genes. However, these classifications can sometimes be indeterminate. Using the expression data allows for decision support where traditional means cannot make a confident assessment. Four classifiers were trained on the data and their performance is shown in Fig. 5. The only non-linear classifier, Random Forest, is an ensemble bootstrap method that builds decision trees on randomly sampled features and averages the predictions from all trees. The mRNA expression levels appear to be a good indicator for the Estrogen Receptor subtype of cancer (Fig. 5a), with linear classifiers averaging 96% accuracy (Table I). Similarly, for the Her2 Receptor subtype, linear classifiers average 85% accuracy. Random Forest performed worse than the linear classifiers in both cases. The same model was used to fit the training data instead of the testing data (Fig. 5c) and performed with 98% accuracy, leading to the conclusion that the Random Forest classifier is overfitting on the training data and is not generalizable. Other non-linear classifiers show this same trend. This suggests that the data is mostly linearly separable and the use of non-linear classifiers only hurts performance. Because of this assumption, the C parameter in the SVM model can be tuned low to avoid overfitting the training data and still perform well.

Regression: After diagnosis, the stage of cancer and prognosis is important for the patient to make informed decisions about their health. This is modeled as a regression problem to predict both the stage of cancer and the survivability of the patient. The regression models used are SVR with a linear kernel and a sigmoid kernel, and linear regression. All of the models tested have poor performance (Table II). Tumor Stage is misdiagnosed over an entire stage and survivability is predicted to be years off from the truth. This may be due to the sparseness of the outcome data – the majority of the patients have stage II or III, and stages I, IV are poorly represented. Additionally, this may be confounded by differing cancer subtypes. For survivability, breast cancer has a relatively good prognosis with 85% of patients surviving

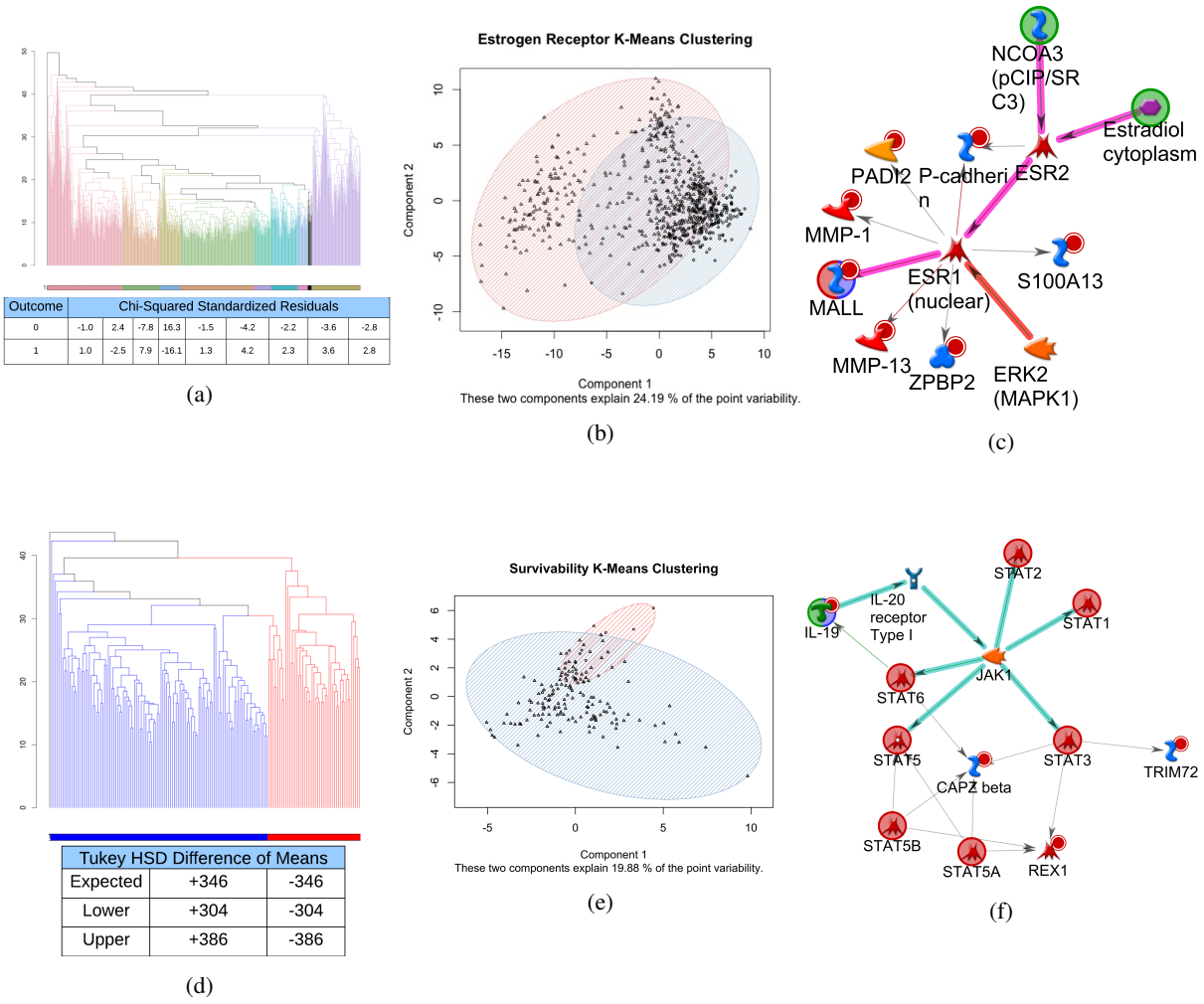


Fig. 4: Unsupervised clustering of estrogen receptor and survivability feature sets. a) Dynamic cut hierarchical clustering of subjects on the ER feature set. Table illustrates standardized Chi-Squared residuals for each cluster for the outcomes "0" (no ER) and "1" (ER present). b) K-Means clustering of subjects on the ER feature set, followed by PCA analysis. c) GO process: developmental process involved in reproduction (FDR=1.86e-24, Z=152) d) Hierarchical clustering of subjects on the survivability feature set. ANOVA results of survivability (days) against cluster membership are illustrated in the table. e) K-means clustering of subjects on the survivability feature set. f)GO process:cytokine-mediated signaling pathway (FDR=2.13e-14, Z=119).

after 5 years [2] leaving few data points corresponding to death events on which to train.

TABLE I: Classifier Performance

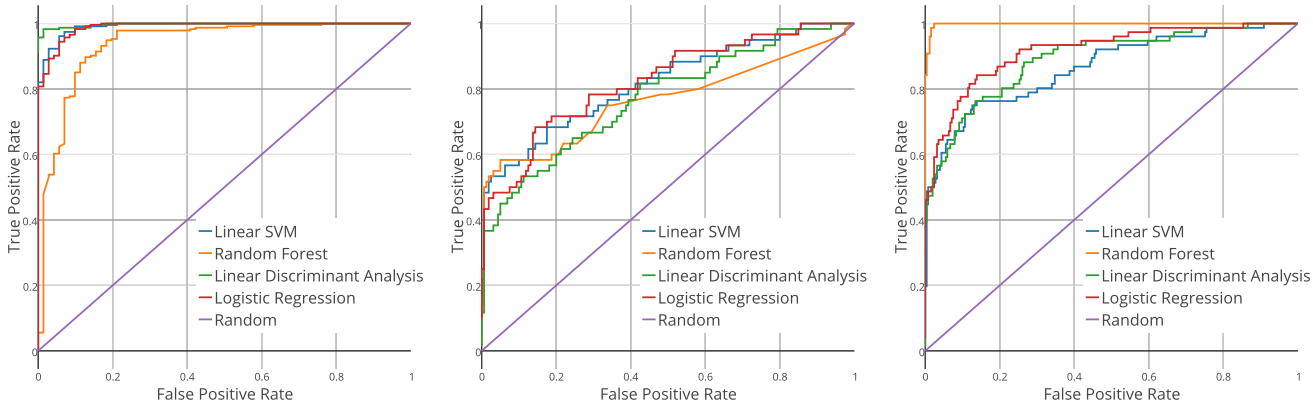
Classifier	Precision	Recall	Accuracy	AUC
Estrogen Receptor				
Linear SVM	0.97	0.97	0.97	0.98
Random Forest	0.92	0.92	0.92	0.95
LDA	0.94	0.94	0.94	0.96
Logistic Regression	0.97	0.97	0.97	0.98
Her2 Receptor				
Linear SVM	0.87	0.86	0.86	0.78
Random Forest	0.84	0.85	0.85	0.72
LDA	0.83	0.83	0.84	0.76
Logistic Regression	0.83	0.84	0.84	0.79

TABLE II: Regression Performance

Regressor	RMSE	Median Absolute Error	R^2
Tumor Stage (Range: 1-10)			
SVR linear	2.31	1.36	0.10
SVR sigmoid	2.54	0.9	-0.08
Linear Regression	2.05	1.35	0.30
Survivability (Range: 158-4456)			
SVR linear	764.29	497.39	0.60
SVR sigmoid	1212.04	909.00	0.00
Linear Regression	479.35	381.36	0.84

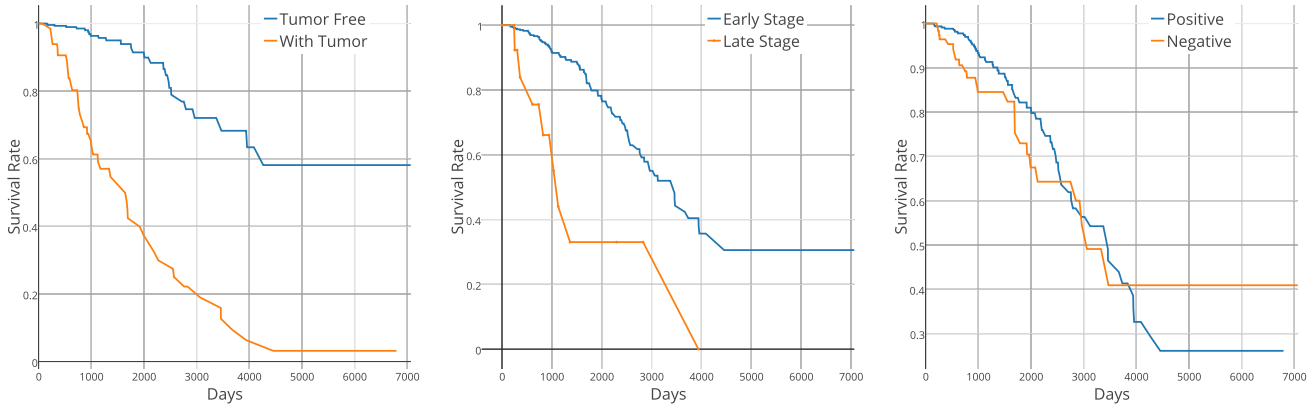
C. *Survival Analysis*

A Kaplan-Meier estimate can determine which outcomes correspond to good and poor survival. The survival function



(a) (b) (c)

Fig. 5: Receiver Operating Characteristic (ROC) curves for subtypes of Breast Cancer. Four classifiers are shown (SVM with linear kernel, Random Forrest, Linear Discriminant Analysis, and Logistic Regression) with a random classifier included for reference. (a) ROC curve for Estrogen Receptor. (b) ROC curve for Her2 Receptor. (c) ROC curve for Her2 receptor on the training data.



(a) (b) (c)

Fig. 6: Kaplan-Meier survival curves. (a) Survival for patients that present with a tumor and those that are tumor free. (b) Survival for patients that are early stage (Stage I, II, III) and late stage (Stage IV, V). (c) Survival for ER positive and negative patients

is defined as:

$$\hat{S}(t) = \prod_{t_i < t} \frac{n_i - d_i}{d_i} \tag{3}$$

where n_i is the number of patients at risk of death at time t and d_i is the number of patient deaths at time t. Kaplan-Meier estimates were obtained using the lifelines Python module [25]. Figures 6a and 6b validate the survival data and show curves that are expected; early stage, tumor free patients survive longer than late stage, patients with tumors. However, most other outcomes are not correlated with survival rates (Fig. 6c).

V. CONCLUSION

Feature selection via a union of the group lasso, elastic net regression, and greedy forward selection identified a number of genes, networks, and biological pathways associated with survivability and receptor status. Most significantly, apoptosis, growth factor receptor, and reproductive development networks from the literature were strongly represented in the down-selected gene sets. Hierarchical and K-means clustering approaches revealed substructure within the population, highlighting individual differences in survivability and estrogen receptor status. The two techniques produced strongly overlapping clusters. Of the supervised learning techniques used to predict cancer subtype, linear classifiers such as LDA, and linear SVM produced over 95 percent accuracy, while more complex clusters such as boosting and the random forest led to overfitting on the training data. Regression analysis revealed that expression levels of mRNA are not good predictors of survival time or tumor stage.

REFERENCES

[1] Cancer Statistics Review, 1975-2012 - SEER Statistics. Web. 14 Nov. 2015. http://seer.cancer.gov/csr/1975_2012/.

[2] Breast Cancer - Statistics, Cancer.Net, 25-Jun-2012. [Online]. Available: Breast Cancer Facts :: The National Breast Cancer Foundation. www.nationalbreastcancer.org.Web. 15 Oct. 2015.

[3] Puglisi et al. Current challenges in HER2-positive breast cancer. Crit Rev Oncol Hematol. 2015 Oct 31. pii: S1040-8428(15)30065-2. doi: 10.1016/j.critrevonc.2015.10.016.

[4] Stellato C, Porreca I, Cuomo D, Tarallo R, Nassa G, Ambrosino C. The busy life of unliganded estrogen receptors. Proteomics. 2015 Oct 28. doi: 10.1002/pmic.201500261.

[5] Matsumoto et al. Biological markers of invasive breast cancer. Jpn J Clin Oncol. 2015 Oct 20. pii: hyv153.

[6] The Cancer Genome Atlas - Data Portal. Web. 15 Oct. 2015.https://tcga-data.nci.nih.gov/tcga/.

[7] Cruz J, Wishart D. Applications of machine learning in cancer prediction and prognosis. Cancer Inform. 2006; 2: 5977. Published online 2007 February 11.

[8] Bastani M et al. A machine learned classifier that uses gene expression data to accurately predict estrogen receptor status. PLoS One. 2013; 8(12): e82144. Published online 2013 December 2. doi: 10.1371/journal.pone.0082144.

[9] Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software. 2010, 33(1) 1-22. http://www.jstatsoft.org/v33/i01/.

[10] Zou, Hui, Zou Hui, and Hastie Trevor. Regularization and Variable Selection via the Elastic Net. Journal of the Royal Statistical Society. Series B, Statistical methodology 67.2 (2005): 301320. Print.

[11] Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. J. R. Statist. Soc. B (2006) 68, Part 1, pp. 49-67.

[12] LEAPS R package. https://cran.r project.org/web/packages/leaps/leaps.pdf.

[13] http://thomsonreuters.com/en/products-services/pharma-life-sciences/pharmaceutical-research/metacore.html.

[14] Ashburner et al. Gene ontology: tool for the unification of biology (2000) Nat Genet 25(1):25-9. Online at Nature Genetics.

[15] Logarajah S, Hunter P, Kraman M, Steele D, Lakhani S, Bobrow L, Venkitaraman A, Wagner S. BCL-6 is expressed in breast cancer and prevents mammary epithelial differentiation. Oncogene (2003) 22, 55725578. doi:10.1038/sj.onc.1206689.

[16] Vogelstein B, Sur S. p53: The most frequently altered gene in human cancers. Cell Communication.http://www.nature.com/scitable/topicpage/p53-the-most-frequently-altered-gene-in-14192717.

[17] Yamamoto-Ibusuki, Mutsuko et al. C6ORF97-ESR1 Breast Cancer Susceptibility Locus: Influence on Progression and Survival in Breast Cancer Patients. European journal of human genetics: EJHG 23.7 (2015): 949956. Print.

[18] Belanger A, Tojcic J, Harvey M, Guillemette C. Regulation of UGT1A1 and HNF1 transcription factor gene expression by DNA methylation in colon cancer cells. BMC Molecular Biology. 2010, 11-9. doi:10.1186/1471-2199-11-9.

[19] Biswas DK, Shi Q, Baily S, Strickland I, Ghosh S, Pardee AB, Iglehart JD. NF-kappa B activation in human breast cancer specimens and its role in cell proliferation and apoptosis. Proc Natl Acad Sci U S A. 2004 Jul 6;101(27):10137-42. Epub 2004 Jun 25.

[20] Langfelder, P., B. Zhang, and S. Horvath. Defining Clusters from a Hierarchical Cluster Tree: The Dynamic Tree Cut Package for R. Bioinformatics 24.5 (2007): 719720. Print.

[21] Lin YC, Lee YC, Li LH, Cheng CJ, Yang RB. Tumor suppressor SCUBE2 inhibits breast-cancer cell migration and invasion through the reversal of epithelial-mesenchymal transition.J Cell Sci. 2014 Jan 1;127(Pt 1):85-100. doi: 10.1242/jcs.132779. Epub 2013 Nov 8.

[22] Hartigan, J. A. and Wong, M. A. (1979). A K-means clustering algorithm. Applied Statistics 28, 100108.

[23] Pison, G., Struyf, A. and Rousseeuw, P.J. (1999) Displaying a Clustering with CLUSPLOT, Computational Statistics and Data Analysis, 30, 381392.

[24] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

[25] Davidson-Pilon, C., Lifelines, (2015), Github repository. https://github.com/CamDavidsonPilon/lifelines.