Interalgorithmic Consolidation for Pattern Recognition Applied to Melanoma Genomic Data*

Brody Kutt†
Department of Mathematics
Rochester Institute of Technology
Rochester, NY

Rachel Burdorf‡
Department of Biology
Colorado College
Colorado Springs, CO

Travaughn Bain§
Department of Mathematical Sciences
Florida Institute of Technology
Melbourne, FL, USA

Lisa Moore¶
Department of Biological Sciences
Florida Institute of Technology, Melbourne, FL

Munevver Mine Subasi‖**
Department of Mathematical Sciences
Florida Institute of Technology, Melbourne, FL

Abstract

Melanoma is a highly proliferative, chemo-resistant cancer without a durable response in most patients. Survival of patients with metastatic melanoma varies widely, and response rates to treatment range from 20-40% with combination therapy, and frequently there is no observed improvement in overall survival. Many of the available therapies have targeted the BRAFV600E mutation, which gives rise to increased cell proliferation through constitutive activation of the regulatory MAPK pathway. These drugs have not proven successful for all patients with tumors expressing these mutations, therefore identifying other key genes influencing response and survival is important.

In this paper, we utilize data from 62 skin tumor cell lines from the Cancer Cell Line Encyclopedia (CCLE) to examine features of gene expression ($> 19,000$) and DNA copy number variation ($> 20,000$) to assess the existence of clusters and, if so, the features which give rise to those clusters. In order to explore, analyze, and extract information from this large-scale dataset we adopt a ensemble feature selection approach that integrates a univariate multiclass Fisher ranking method with other well-known powerful machine learning techniques. In comparison to other stand-alone feature selection techniques, our proposed method provides a subset of features that can reliably distinguish between subtypes of melanoma cell lines across many different types of classifiers. Our methods reduced the 19,000 gene feature space into only the top 15 gene features that maintained the same initial clustering of the data. Of these 15 top genes, some were already known to be linked to melanoma prognosis, or linked to other cancers with novel relevance to melanoma, and some were never before linked to melanoma prior to this work.

*Supported by NSF REU Program – Award #: 1359341
Submitted to **2018 International Symposium on Artificial Intelligence and Mathematics**

†Email: bjk4704@rit.edu
‡Email: Rachel. Burdorf@ColoradoCollege.edu
§Email: tbain2013@my.fit.edu
¶Email: lmoore@fit.edu
‖Email: msubasi@fit.edu
**Corresponding Author.

1 Introduction

Melanoma is a devastating disease and incidence is on the rise while treatment options remain limited for the most aggressive forms of the disease. Several gene mutations are widely expressed across melanoma cases: for example, the BRAFV600E mutation occurs in approximately 60% of patients. This mutation results in constitutive activation of BRAF signaling in the ERK/MAP Kinase pathway with an end result of increased cell proliferation and survival (Ascierto and et al. 2012). Some of the most promising treatments have been developed as selective inhibitors of this pathway. However, 40% of patients do not carry mutations in this pathway, and of those that do, not all respond to these therapies. Even initial responsiveness may disappear with disease recurrence. Malignant melanoma has shown poor durability overall in response to available treatments. The question that arises from this information is what other factors drive treatment response, tumor aggressiveness, and the ability to metastasize from the original primary lesion site. A long-term goal of this project is to identify subset populations of melanoma as distinguished by gene expression, and subsequently assess drug response in these subset populations as a means to predict patterns that may indicate treatment response and prognosis.

In an effort to address this question, different studies have taken various approaches in a number of studies. One approach is to profile gene expression using techniques like cDNA microarray screening, or transcriptome analyses from patient tumor libraries. These techniques have the advantage of providing known information on tumor response to different therapies, tumor staging, prognosis and survival data. Downsides to this approach include limited access to tissues, tissue heterogeneity, and overall tissue availability (Ryu and et al. 2007). An alternate approach is the evaluation of melanoma cell line data, made possible by open access databases with patient and cell line datasets. Typical strategies for this approach are hierarchical clustering, similarity core analysis, and Elastic Net regression (Covell 2015; Garnett and et al. 2012; Rambov and et al. 2015; Ryu and et al. 2007). These methods have had some success in identification of potential genes that may be involved in upregulation of cell proliferation, drug response, and propensity for metastasis to distant sites. However, success has been limited due to the low correlation of results

from one approach to the next. General classes of gene products are identifiable, those being effectors of the cell cycle, its checkpoints, apoptosis, cell adhesion, tumor suppressors and DNA repair. The availability of data through publicly accessible large databases has paved the way for tailored approaches to data mining and machine learning. Many of these approaches in the literature have been used to assess drug sensitivity data with pathway and gene expression clustering (Brubaker and et al. 2014; Covell 2015; Garnett and et al. 2012; Jang and et al. 2014). Large datasets are a powerful tool for identifying pathways and gene expression patterns that are critical in determining overall sensitivity to treatments and prognosis. The Cancer Cell Line Encyclopedia (CCLE) is one such database, in which gene expression (GE), copy number (CN), and drug response data are available for over 1000 cancer cell lines (`http://www.broadinstitute.org/ccle/home`). In this study, we have selected the skin tumor cell lines from the CCLE as a starting point, and used a novel algorithmic approach to extract information from the large scale dataset to identify genes of interest that distinguish these 62 cell lines into 3 discrete clusters. To date, the use of biomarker data from these databases has provided limited success and will require confirmation in biologic systems to determine true correlative benefit. Hopefully these analyses will lead to development of molecular signatures for prognosis and staging, targeted therapies, and the ability to personalize treatment for the most durable response.

The rest of this paper is organized as follows. In Section 2, we present our methodology for determining the existence or non-existence of natural clusters in the data. In Section 3, we present our ensemble feature selection approach, the results obtained from it and discussion about the results. Section 4 summarizes our results and the main takeaways of the paper as well as proposes some possible avenues for future work.

2 Clustering of Melanoma Cell Lines

CCLE Gene expression and DNA copy number variation data for 62 skin tumor cell lines was used to systematically assess features of interest for distinguishing subtypes of melanoma cell lines. CCLE skin tumor cell line data that we used included $> 19,000$ mRNA gene expressions and $> 20,000$ DNA copy number variations (CNV). In order to identify the distinct clusters of CCLE skin tumor cell lines we applied the popular K-means++ method to data consisting of 62 cell lines with:

- Gene expression features alone.
- Copy number variation features alone.
- Gene expressions and copy number variation features together.

K-means is a widely used clustering technique that seeks to minimize the average squared distance between points in the same cluster. Simple euclidean distance was used as the distance metric since all of our feature data were scalar numeric values. Although it offers no accuracy guarantees, the simplicity and speed of K-means are very appealing in practice. Because of the sparsity of our data, we did not expect to find sufficiently complex, non-convex or dense clusters thus K-means sufficed. K-means++ was presented in (Arthur and Vassilvitskii 2007) where an algorithm that is $O(\log k)$-competitive with the optimal clustering was obtained by augmenting K-means with a simple, randomized seeding technique. Computational experiments have shown that the augmentation improves both the speed and the accuracy of K-means, often quite dramatically.

As shown in Figures 1-2, the initial fitting of the 19,000 gene expression features suggested approximately 5 true clusters as signified by the change in slope whereas there was no clear suggestion for the optimal number of clusters for the 20,000 copy number variations. Also, note that combining both the copy number variations and the gene expressions data did not affect the natural clustering of the CCLE skin tumor cell lines as can be seen in Figure 3.

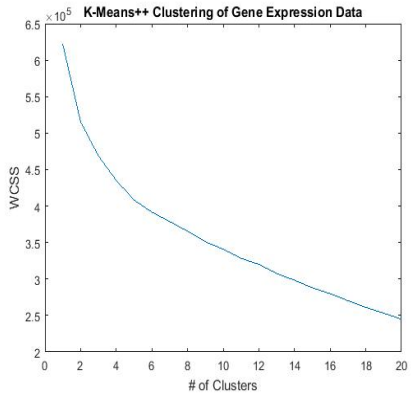

Figure 1: Within-cluster sum of squares (WCSS) error for 19,000 gene expression features over different values of K. Notice the elbow point at approximately 5 clusters where the slope changes.

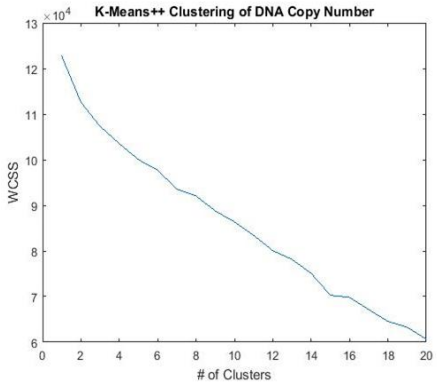

Figure 2: Within-cluster sum of squares (WCSS) error for 20,000 CNV features over different values of K. Notice the slope is relatively constant throughout.

For the sake of high-dimensional data visualization, we make use of popular Principle Components Analysis (PCA)

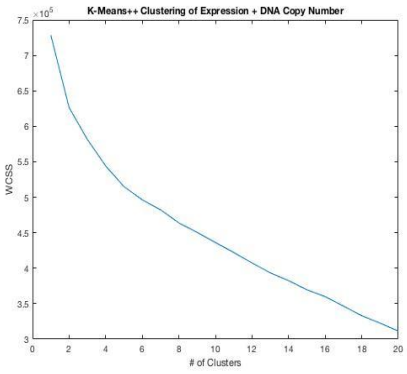

Figure 3: Within-cluster sum of squares (WCSS) error for 19,000 genes + 20,000 CNV features over different values of K. Notice the same elbow point as in Figure 1.

algorithm to project the data onto the two dimensions of greatest variation. More advanced visualization methods like t-Distributed Stochastic Neighbor Embedding (t-SNE) (van der Maaten and Hinton 2008), were considered but deemed not necessary for this project since our plots seemed to make sense for our data. Plots corresponding to all three 5-cluster datasets are presented in Figures 4-6.

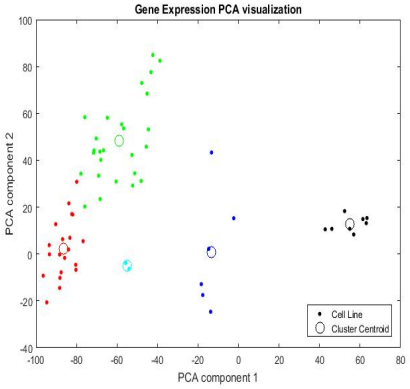

Figure 4: PCA plot of 5-clusters in 19,000 gene expression feature data

Information obtained on the cell lines from the American Type Culture Collection (ATCC) suggested that a single patient was the origin of the two cell lines clustered together in their own cluster. In addition, cell lines from one other cluster originated from Naval Biosciences Laboratory (NBL) cell lines, some of which have now been removed from the ATCC due to inability to confirm morphology. Due to the concern that these could not be confirmed as melanoma tumor lines, we removed these two clusters from further analysis. This resulted in three clusters for 49 skin tumor cell lines. Moreover, since the PCA plots in Figures 4-6 showed that the exclusion of copy number variations from the dataset provided a better clustering of the CCLE skin tumor cell

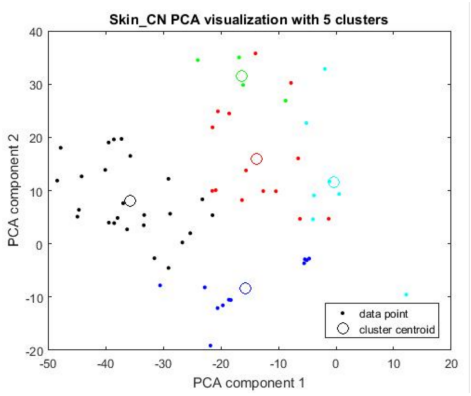

Figure 5: PCA plot of 5-clusters in 20,000 CNV feature data

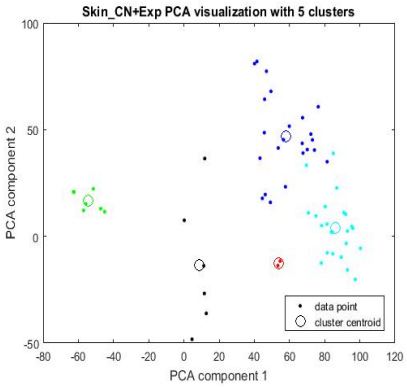

Figure 6: PCA plot of 5-clusters in 19,000 genes + 20,000 CNV feature data

lines, we continued our investigations on 49 skin tumor cell lines with the 19,000 gene expression features alone. Figures 7-8 show the PCA plot of the three clusters of 49 cell lines based on gene expressions only.

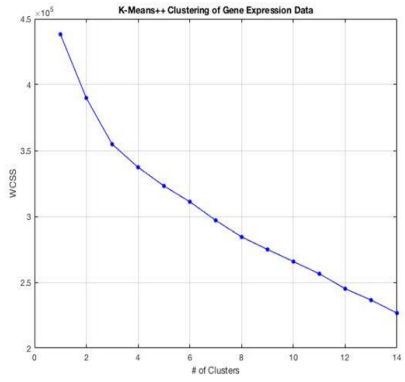

Figure 7: Within-cluster sum of squares (WCSS) error for 49 cell lines with 19,000 gene expression features over different values of K.

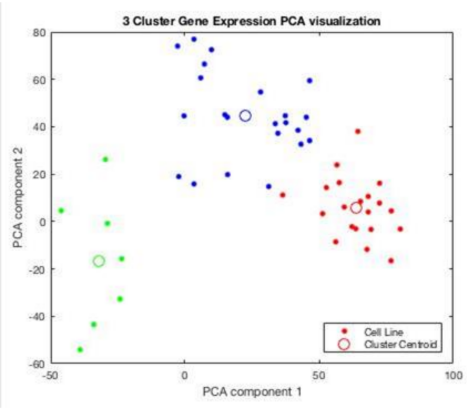

Figure 8: PCA plot of 3-clusters for 49 cell lines with 19,000 gene expression features.

3 Feature Selection and Classification

In order to identify the patterns of gene expression separating those clusters and to validate the three clusters obtained by the K-means++ algorithm we applied ensemble feature selection and classification methods. These analyses are outlined in the following sections.

3.1 Identification of Most Significant Gene Expressions

There exist many different algorithms for feature selection, but they are typically divided into three main groups: filter, wrapper and embedded methods. Filter methods rank each feature according to some univariate metric, and only the highest ranking features are used; the remaining features are eliminated. These methods tend to be computationally lightweight but do not consider the interactions between features. Wrapper algorithms explicitly search for the best subset of features. To assess the quality of a feature subset, wrapper methods rely on and interact with a classification algorithm and its ability to discriminate among the classes. The wrapper algorithm treats a classification algorithm as a black box, so any classification method can be combined with the wrapper. Standard optimization techniques (hill climbing, simulated annealing or genetic algorithms) can be used. Embedded methods search among different feature subsets, but unlike wrappers, the process is naturally within a certain classification algorithm itself instead of an outside process. An example of an embedded feature selection approach is the use of decision trees which have pruning mechanisms built within the algorithm itself that perform feature selection. For more information on feature selection methods and their applications to genomic and proteomic data the reader is referred to (Dubitz and et al. 2007) and the references therein.

In part because of the high dimensionality and sparseness of the CCLE melanoma data, feature selection methods often yield vastly different results in terms of the returned utility of one feature versus another. These varying results depend entirely on what algorithm is used and not the structure of the data itself. Also, because of the power of the well-known classification algorithms, they can easily learn patterns in the data that might not be biologically relevant. Using a large enough subset of features, even with randomly selected features, can yield good classification accuracy for any one algorithm despite the high probability that the features are not actually relevant to the outcome variable. This lends the task of biological interpretation to be extremely difficult if not impossible. We propose that the features which most believably distinguish the data will show relevance across many different kinds of feature selection algorithms and will also yield high classification performances across many different types of classifiers with a relatively small subset size. To reach this end, we employ a feature selection pipeline which can be thought of as an ensemble of algorithms. A univariate statistical technique called the Fisher Score and a multivariate SVM-based technique (SVM-RFE+CBR) (Yan and Zhang 2015) are first used to reduce the dataset to a size that is more manageable for running a third technique that includes 500 randomized trials of a sequential floating forward search (Pudil, Novoviov, and Kittler 1994) wrapped around the popular K-nearest neighbors algorithm. The randomization is captured by randomizing the training and test sets on each trial. The goal of our approach is not simply to train one learner or fit one model to get good classification performance. Instead, we wish for overall knowledge gain about the feature set itself through the use of multiple models. This is difficult with stand-alone methods since the CCLE melanoma dataset contains a small number of observations yet several thousands of features as is common in other genomic or proteomic data. In order to avoid the "curse of dimensionality" (Duda and et al. 2001) our feature selection procedure, described below, combines the results and strengths of different algorithms.

Interalgorithmic Consolidation for Feature Selection:

S0. Input Data: 49 skin tumor cell lines with 19,000 gene expressions.

S1. Apply the Fisher ranking method (univariate correlation based method extended to multiple features) on "Input Data" and select top 500 gene expressions based on Fisher ranking.

S2. Perform Support Vector Machines - Recursive Feature Elimination with Correlation Bias Reduction (SVM-RFE+CBR) method (Yan and Zhang 2015) on "Input Data" and select top 500 gene expressions based on SVM-RFE+CBR ranking.

S3. Reduced Data: Combine the top 500 gene expressions obtained in (S1) with top 500 genes expressions obtained in (S2) for 49 cell lines. We eliminate the presence of duplicates. The resulting reduced data contains 49 cell lines with 928 gene expressions.

S4. Perform randomized trials of a sequential floating forward search wrapped around the K-Nearest Neighbor algorithm on "Reduced Data" and record the genes that occur the most often in the results.

S5. Output Data: 49 cell lines with the top 15 genes obtained

from the randomized sequential trials of k-Nearest Neighbors in (S4).

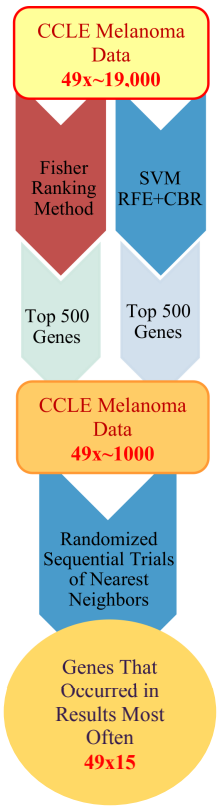

Figure 9: Pipeline for feature selection in CCLE melanoma data

Our method can be visualized graphically in Figure 9. The features which occur most often in our randomized trials are the features we regard as the best, as they are seen to emerge repeatedly as good features for classification across randomization, as opposed to the results of one execution in isolation. The cut off for feature frequency is in fact very steep as shown in Figure 10.

3.2 Top 15 Most Significant Genes

Our analyses showed that only the top 15 genes obtained from the randomized trials in (S4) were sufficient for optimal classification performance. Only under 15 features did classification performance seem to drop across various classification algorithms. These genes are given in Table 1.

TBC1D16	SEMA6A	AVPI1
TRIM9	ARHGEF6	GST01
DYNC1I1	GPR127B	AHR
YPEL2	PIK3CD	C16ORF52
CD274	SPATA13	SMTN

Table 1: Top 15 genes for CCLE skin cell line cluster differentiation.

Of these 15 genes, some are already known to be linked

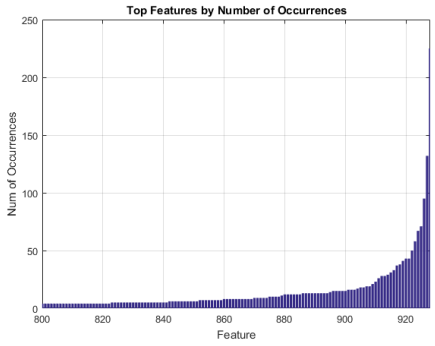

Figure 10: Number of occurrences of the gene expression features during randomized trials. Feature index is on the X axis while frequency in the results set across 500 randomized trials are shown on the Y axis. We find that the frequency of gene expression feature occurrence is largely dominated by only a few features.

to melanoma prognosis or known to be linked to other cancers, but their relationship to melanoma was not previously shown. Some of the top 15 genes seemingly most relevant to cancer:

- **TBC1D16** is suggested to regulate EGFR in melanoma as a result of a hypomethylation event conferring poor survival, exacerbated growth and also may increase BRAF and MEK inhibitor sensitivity (Vizoso and et al. 2015).
- **TRIM9** Tripartate Motif Containing 9 is expressed in many cancer cell lines, the TRIM proteins have been identified in other hierarchical analyses for melanoma. TRIM9 is thought to be a modifier of disease incidence and progression in lung cancer. TRIM9 has been identified as a ubiquitin ligase (E3) (Wang and et al. 2016).
- **DYNC1I1** Cytoplasmic Dynein 1 intermediate chain 1 protein is thought to regulate dynein function important for vesicle motility and trafficking of organelles. It is regulated by microphthalmia-associated transcription factor (MITF) which is important to melanocyte development. Frequent somatic mutations of MITF have been reported with cutaneous melanoma.
- **CD274** This gene codes for Programmed death-ligand 1 (PD-L1) which is currently the target of several large trials showing substantial benefit with anti-PD-L1 for late stage melanoma (Ascierto and Marincola 2015).
- **GST01** Glutathione S-transferase omega-1 polymorphisms associated with the increased risk of developing breast and liver cancer and has not been previously implicated in melanoma.

3.3 Relevance of TBC1D16 in Melanoma

The top performing gene expression in terms of the frequency in the results set over the randomized trials was a gene known as TBC1D16. As shown in Figure 10, TBC1D16 vastly outperforms every other gene. It has been shown in a recent publication that TBC1D16 has significant

relevance to the metastatic potential of melanoma and response to BRAF and MEK inhibitors. TBC1D16 is also proposed to strongly regulate vesicle trafficking. It is also key to regulation of the EGFR pathway which exerts control on survival signaling in melanoma. Expression profiling of tumor cell lines can provide an avenue for identifying novel candidates in developing a molecular signature for melanoma. A molecular signature could aid in targeting patient centric treatment approach as well as identifying aggressiveness of the disease. Large-scale sequencing efforts are being spearheaded by groups such as the International Cancer Genome Consortium and the Cancer Genome Atlas to identify new drug targets and to confirm genes that may predict sensitivity to drugs. Evidence that TBC1D16 may have a role in regulating the EGRF pathway by activation of a GTP-ase activating protein and influence response to MEK, BRAF and EGRF inhibitors, all drugs targeting metastatic melanoma, supports our finding that this is an important gene in the disease profile.

3.4 Validation of the Top 15 Genes

We used K-means++ clustering on the reduced CCLE skin cancer tumors with the top 15 genes only. We found that the optimal number of clusters for the reduced data as signified by the elbow point was much more pronounced to be 3 as shown in Figure 11.

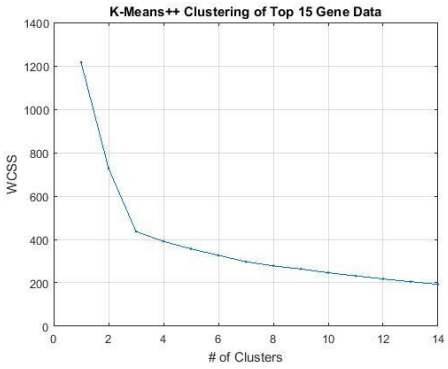

Figure 11: Within-cluster sum of squares (WCSS) error for the top 15 gene expression data across different values of K. Note the very pronounced elbow point at exactly 3 clusters.

The top 15 genes and their expression levels provide the same exact clustering with the same cell lines clustering together (see Figure 4) as when clustering in the original 19,000 gene expression feature space. The PCA plot of the three clusters using the top 15 genes is presented in Figure 12 which shows the increased density and separability of clusters as when compared to Figure 8.

In order to further validate the discriminating power of the top 15 genes among the three clusters we applied various classification techniques on the 49 cell lines with the 15 gene expression features. The cluster index was used as the class variable. Table 2 shows the 10-fold cross validated per-cluster classification accuracy of various classification algorithms as implemented in WEKA (Hall and et al. 2009).

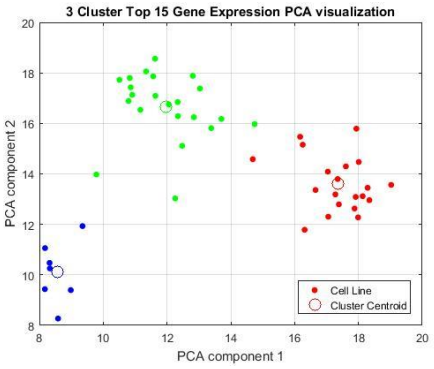

Figure 12: PCA plot of 3-clusters using the top 15 gene expression features only. Note the relative density and separability of the clusters.

Cluster	Multilayer Perceptron	Logic Regression	Naive Bayes Multinomial
1	100%	100%	100%
2	100%	100%	100%
3	100%	85.70%	100%
Average	100%	95.23%	100%
Cluster	**k-Nearest Neighbor**	**Logic Model Tree**	**Random Forest**
1	100%	85%	100%
2	100%	100%	100%
3	100%	100%	100%
Average	100%	95%	100%

Table 2: Classification accuracy for top 15 genes

3.5 Decision Tree Classification Model

It is instructive to develop a human interpretable model for the sake of understanding what combinations of gene expressions lead to cluster indices. We applied a decision tree classification method on data consisting of the 49 cell lines with the top 15 gene expression features to obtain a classification model consisting of combinatorial patterns of gene expressions. A Decision Tree is a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features (Quinlan 1993). The decision tree classification model for CCLE skin cancer tumors as shown in Table 3 consists of one pattern for cluster 1, one for cluster 2, and two patterns for cluster 3, where decision rules are based on three genes including GSTO1, TBC1D16, and TRIM9.

Cluster	**Decision Tree Rules**
1	$GSTO1 \geq 12.7765$ & $TBC1D16 \geq 7.28245$
2	$GSTO1 < 12.7765$ & $TRIM9 \geq 5.03635$
3	$GSTO1 \geq 12.7765$ & $TBC1D16 < 7.28245$
3	$GSTO1 < 12.7765$& $TRIM9 < 5.03635$

Table 3: Decision tree classification model

The performance of the final classification model was evaluated with several executions of a 10-fold cross validation technique. The per-cluster cross-validation accuracy of the decision tree classification model are provided in Table 4. The high classification performance provides evidence for the reliability of the rules generated by the algorithm.

Average Accuracy	Cluster 1	Cluster 2	Cluster 3
98.33%	95%	100%	100%

Table 4: Average accuracy of the decision tree model

4 Conclusions & Future Work

This paper presents the usage of an off-the-shelf clustering technique as well as a novel ensemble feature selection technique on genomic data that has identified 3 distinct clusters of melanoma cell lines. Despite making use of everything that is available, further biological experimentation is likely needed to increase the confidence of the cluster's reliable presence across all different instances of melanoma. The identification of TBC1D16 as the primary gene providing cluster separability is a significant result given its known relationship to melanoma. This suggests its potential relevance for melanoma prognosis and anti-cancer drug selection given its marked variable expression across different melanoma clusters. Other genes in our results that have not been previously linked to melanoma may offer themselves as points of interest for further biological exploration. Similar testing will be applied to other public access genomic databases for melanoma to compare results.

In terms of future work that can extend from this, it is of interest to see if the genes listed in this work can be used in a predictive manner. Similar approaches have been taken with melanoma gene expression assays in clinical trials which hope to provide better diagnostics for staging of the disease (Castle DecisionDx Melanoma and Myriad myPath Melanoma). Both of these assays utilize a similar gene profile that includes a mixture of housekeeping genes, immune modulators and those involved in pathogenesis (Ferris and et al. 2016), (Clarke and et al. 2017). Approaches such as presented here could help tailor treatment options for patients based on the predicted outcome that is based on the genetic profile of their individual tumor.

5 Acknowledgements

First two authors were supported by National Science Foundation (NSF) Research Experience for Undergraduates (REU) Grant – Award #: 1359341

References

Arthur, D., and Vassilvitskii, S. 2007. k-means++: the advantages of careful seeding. *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms,* 1027–1035.

Ascierto, P., and et al. 2012. The role of BRAF V600 mutation in melanoma. *Journal of Translational Medicine* 10:85.

Ascierto, P., and Marincola, F. 2015. The year of Anti-PD-1/PD-L1s against melanoma and beyond. *EBIOMedicine* 2:92–93.

Brubaker, D., and et al. 2014. Drug intervention response predictions with PARADIGM (DIRPP) identifies drug resistant cancer cell lines and pathway mechanisms of resistance. *Biocomputing* 125–135.

Clarke, L., and et al. 2017. An independent validation of a gene expression signature to differentiate malignant melanoma from benign melanocytic nevi. *Cancer* 123(4):617–628.

Covell, D. 2015. Data mining approaches for genomic biomarker development: Applications using drug screening data from the cancer genome project and the cancer cell line encyclopedia. *PLoS ONE* 10(7):e0127433.

Dubitz, W., and et al. 2007. Fundamentals of data mining in genomics and proteomics. *Springer*.

Duda, R., and et al. 2001. Pattern classification. *Wiley-Interscience Publication*.

Ferris, L., and et al. 2016. Identification of high-risk cutaneous melanoma tumors is improved when combining the online American Joint Committee on Cancer Individualized Melanoma Patient Outcome Prediction Tool with a 31-gene expression profile-based classification. *J Am Acad Dermatol* 76(5):818–825.

Garnett, M., and et al. 2012. Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature* 483:570–575.

Hall, M., and et al. 2009. The WEKA Data Mining Software: An Update. *SIGKDD Explorations* 11(1).

Jang, I., and et al. 2014. Systematic assessment of analytical methods for drug sensitivity prediction from cancer cell line data. *Biocomputing* 63–74.

Pudil, P.; Novoviov, J.; and Kittler, J. 1994. Floating search methods in feature selection. *Pattern Recognition Letters* 15(11):1119 – 1125.

Quinlan, J. 1993. C4.5: Programs for Machine Learning. *Morgan Kaufmann Publishers*.

Rambov, and et al. 2015. New functional signatures for understanding melanoma biology from tumor cell lineage-specific analysis cell reports. *Cell Rep* 13(4):840–853.

Ryu, B., and et al. 2007. Comprehensive expression profiling of tumor cell lines identifies molecular signatures of melanoma progression. *PLoS ONE* 2(7):e594.

van der Maaten, L., and Hinton, G. 2008. Visualizing high-dimensional data using t-SNE. *Journal of Machine Learning Research* 9:2579–2605.

Vizoso, M., and et al. 2015. Epigenetic activation of a cryptic tbc1d16 transcript enhances melanoma progression by targeting EGFR. *Nat Medicine* 21(7):741–750.

Wang, X., and et al. 2016. TRIM9 is up-regulated in human lung cancer and involved in cell proliferation and apoptosis. *Int J Clin Exp Med* 9(6):10461–10469.

Yan, K., and Zhang, D. 2015. Feature selection and analysis

on correlated gas sensor data with recursive feature elimination. *Sensors and Actuators B: Chemical* 353–363.