Machine Learning Toolbox for Alternative Splicing versus Differential Gene Expression Analysis in Lung Cancer Recurrence

An essay submitted in partial fulfillment of

the requirements for graduation from the

Honors College at the College of Charleston

with a Bachelor of Science in

Data Science

Tori McCaffrey

May 2014

Advisor: Paul Anderson

Introduction

Biomarker discovery is one of the most significant approaches for the effective diagnosis, prognosis, and treatment of many diseases including cancer. By identifying specific molecular events, physicians and scientists are able to determine a patient's medical phenotype more efficiently. The use of biomarkers provides powerful insight for the next steps of individualized treatment and personal therapy optimization [1]. Traditionally, biomarker discovery has been performed through differential gene expression (DGE) analysis, used to generate accurate predictive models for phenotypic classification [2]. However, the development of next generation massively parallel high throughput sequencing technology, such as RNA sequencing (RNA-seq), has resulted in supplementary data, including transcript start sites and splice variant (isoform) composition and expression [3]. This surplus of supplementary information allows for the generation of predictive models for genotype-to-phenotype investigations and for improved accuracy when compared to traditional DGE analysis. In particular, the addition of the splice variant (isoform) information is important to consider due to the fact that genes not only can be differentially expressed, due to cellular differentiation and environmental factors and disease, but also differentially alternatively spliced [3]. Recent studies on various mammalian tissues show that more than 90% of human genes exist as alternatively spliced variants [4]. In addition, it has been found that splicing occurs more frequently in neoplastic tissue, suggesting that alternative splicing events (ASE) may play a significant role to the malignant state of cancer [5]. Therefore, it is necessary to further the analysis of alternative splicing events to determine their non-redundant predictive capabilities, thus verifying the significance of alternative splicing events on phenotype.

The field of genomics is lacking an organized workflow or set of tools for the analysis of alternative splicing events, specifically in comparison to differential gene expression (DGE), with the power to create a predictive model. One of the few available open-source software toolboxes that are available for alternative splicing analysis, altAnalyze, is composed of several tools that perform different functions, such as principal component analysis, pathway analysis, clustering, visualizations, etc. It also provides the user with the option of using workflows to perform multiple different tests and analyses on the data. [6] Other current approaches for alternative splicing analysis exist as individual software tools that use splice variant databases as references. However, the capabilities of these tools, such as ALEXA-seq [4], MISO [7], and SpliceTrap [8], are lacking in several ways, often overlooking more complex alternative splicing patterns with more than two splice variants and unable to accommodate novel alternative splicing events that can be discovered by RNA-seq. This approach leads to a misinterpretation and incorrect quantification of the splicing events [9]. Another common method for alternative splicing analysis is the tool DiffSplice, which detects and visualizes differential alternative splicing through alternative splice modules or splice junctions [9]. However, all of the aforementioned tools and software are lacking in complete performance due to the fact that they are unable to create generate models from the data. Therefore, a more efficient approach for predictive model creation from both alternative splicing events and differential expression data produced by traditional next-generation bioinformatics software was sought.

The toolbox presented in this paper provides a package of tools and algorithms written in MATLAB programming language to be used for both gene expression and alternative splicing analyses of genomic data, ultimately focusing on the added value of alternative splicing events

when identifying putative biomarkers. A major feature of the toolbox is a novel multiple objective genetic algorithm based feature selection approach that combines alternative splicing events with gene expression for the generation of predictive models. Genomic sequence data is processed through a bioinformatic pipeline that outputs gene and isoform expression values for each sample. From these expression values, the relative abundances of splice variants are used to find alternative splicing events and are then combined with differential gene expression to create predictive models from orthogonal projections to latent structures discriminant analysis (OPLS-DA) [10][11]. The results of the predictive models are compared to the ranking of independent alternative splicing events (IASE) and are then cross-validated to determine predictive accuracy and possible confounding samples. The top features of models created from IASE, ASE, and DGE are to be compared in order to determine similarity using a Spearman's rank correlation test. Lastly, the genetic algorithm of the toolbox is used to determine the least amount of DGE and ASE features that would be able to generate an accurate prediction model. These significant features provide powerful information on the complementary and non-redundant predictive power of ASE. Overall the comprehensive set of tools improves the efficiency of alternative splicing and differential expression analysis provides a more efficient approach to experimental replication and can be used for similar biological studies involving DGE versus ASE analyses.

Methods

A. Data and Experimental Design

The experimental data used to test the accuracy of the predictive model came from RNA samples collected from twenty-one different lung adenocarcinoma tumors with known clinical outcomes from the American College of Surgery Oncology group. Adenocarcinoma is the most common type of non-small cell lung cancer (NSCLC), accounting for 50% of NSCLC cases. Out of the twenty-one RNA samples, ten of them were derived from patients who developed cancer recurrence within three years of their initial surgical resection (R, Relapse). The remaining eleven samples were derived from patients who remained disease free (DF) after three years.

RNA integrity was verified on an Agilent 2100 Bioanalyzer (Agilent Technologies, Palo Alto, CA). 100-200 ng of total RNA was used to prepare RNA-Seq libraries using the TruSeq RNA Sample Prep Kit following the protocol as described by the manufacturer (Illumina, San Diego, CA). Three samples per lane were clustered on a cBot as described by the manufacturer (Illumina, San Diego, CA). Clustered RNA-seq libraries were paired-end sequenced with 2X100 cycles on a HiScanSQ. Demultiplexing was performed utilizing CASAVA to generate Fastq files, which contain the whole transcriptome sequences.

B. Next Generation Processing

The Tuxedo pipeline [12] was used to map sequenced reads of mRNA to the human genome (hg19, UCSC) and deterministically quantify the amount of transcripts for each isoform and gene product. The tool Cuffdiff was then used to statistically analyze the differential expression of genes and isoforms between the two phenotypic groups: Disease Free (DF) and Relapse (R). The resulting data was investigated using CummeRbund [13], a RNA-seq data analysis package in R.

The overall procedure of processing the RNAseq data was organized into a bioinformatic pipeline on a local instance of a Galaxy Project server [14]–[16]. Two RNAseq data files, containing forward and reverse reads, were generated from each of the RNA samples using paired end (PE) sequencing. FastQC [17] was used to visualize the quality of the sequenced RNA for each dataset. Each set of PE read files was run through Trimmomatic [Lohse] to remove low quality base pairs and sequence adapters (synthetic sequences of DNA that are used to amplify and sequence the cDNA during RNA-seq) using these parameters [ILLUMINACLIP:TruSeq3-PE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:20 MINLEN:36]. The Trimmomatic website provides detailed information on the parameters used.

Trimmomatic filters and trims reads to generate high quality sequencing data from each PE reads file. The Galaxy tool, FastQ Groomer, was used to convert the format of datasets from Illumina to Sanger data format for downstream processing. TopHat 2 (v0.5) [12] was used to map the PE files to the human genome (hg19) with default parameters. TopHat outputs a binarily compressed sequence alignment/map (BAM) file containing the accepted transcript alignments. The aligned transcripts were then assembled into unique isoforms, characterized by gene, and quantified to ascertain gene and isoform expression values using Cufflinks (v2.1.1) [13], while performing bias correction, quartile normalization, and multi-read correction. hg19 was used for the reference genome (FASTA format) and annotation (GTF format). The assembled isoforms from each of the RNA samples were aggregated using CuffMerge (v1.0.0) [12] without reference annotation or sequences. Cuffdiff (v2.1.1) [12] was used to analyze the the samples as a whole, while performing bias correction, quartile normalization, and multi-read correction. The tool takes as input the aggregated isoforms GTF file from CuffMerge and the aligned transcript files from TopHat, which were specified as being one of two phenotypic groups: Disease Free (DF) and Relapse (R).

C. Exploratory Data Analysis

In order to understand the data more appropriately before training the predictive models, unsupervised exploratory data analysis was performed on the complete DGE and ASE datasets. These analyses included hierarchical clustering and principal component analyses (PCA). The CummeRbund package [13] was used to create expression value matrices for both DGE and ASE by using the output of Cuffdiff, which is represented as average linkage of exon per million fragments mapped (FPKM) values. The MATLAB bioinformatic toolbox used these DGE and ASE expression value matrices to perform both hierarchical clustering analysis and PCA analysis [20]. The results of these unsupervised analyses generated visualizations that aided in the understanding of the data. The PCA analyses visualized the gene expression and alternative splicing data by transforming the data into a new coordinate system, where each new dimension was computed as a linear combination of the original values in order to best explain the overall variance in the data [21]. Hierarchical clustering was achieved from heat map clustering using

the respective FPKM values for DGE and ASE.

1. Gene Expression Data: Using the FPKM values of the genes and isoforms from the output of CummeRbund, DGE and ASE expression value matrices were created by Cuffdiff. The tool did not incorporate genes that were deemed invalid from differential expression analysis into subsequent analyses (genes labeled "NOTEST"). The remaining genes and isoforms were normalized to unit variance and mean centered (i.e., auto-scaling). Genes that had only one single isoform (no alternative splicing events) were removed from the dataset to provide a fair comparison of the gene sets generated with predictive models of DGE versus ASE. The resulting FPKM expression dataset forms a normalized expression matrix (NEVM) of each gene per sample.

2. Alternative Splicing Data: In order to analyze differential alternative splicing events without being affected by the differential absolute expression of isoforms, the differences of the relative abundances of isoforms for each gene was calculated, thus removing the effect of differential expression between phenotypes. For example, consider a gene with three isoforms, where one of the isoforms accounts for 71% of the total amount of isoform expression and the other isoforms account for the remaining 29%. The relative abundances of isoforms must be different between phenotypes (e.g., isoform relative expression increases from 43% to 52%, in this case), in order for the gene to be considered a differentially alternative splicing event. This difference could result in an overall change in expression or it could be an independent event. A sample gene (FAM111B) that consists of three isoforms and shows a relative increase of expression of isoform NM_198947 (43% to 52% between DF and R) is shown in Figure 1. As expected, the other isoforms, NM_001142703 and NM_001142704, show a decrease of relative expression between DF and R from 55% to 45% and 58% to 42% respectively. Differentiating between the differential expression events and alternative splicing events creates complementary features for subsequent machine learning algorithms. The resulting dataset forms an isoform composition value matrix (ICVM) of each isoform per sample:

$$ICVM_{i,j} = \frac{FPKM_{i,j}}{\sum_{k=1}^{m} FPKM_{i,k}},$$

where $ICVM_{i,j}$ is the isoform composition value for the *jth* isoform and the *ith* gene and *m* is the number of isoforms for the gene *i.*

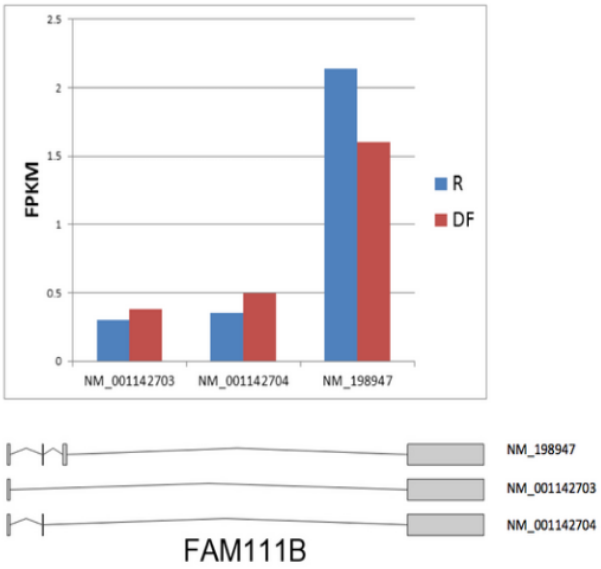


Fig. 1: The splice graph of FAM111B isoforms and their absolute expression values. The relative abundance of the NM_198947 splice variant is different between phenotypes, indicating an ASE.

3. Combined DGE and ASE: A third matrix was generated using the combined matrices of the NEVM and the ICVM in order to further analyze the role of DGE vs ASE when compared in the same analytical test. By combining the two expression matrices, there are 24,427 features of genes and isoforms. All of the tools discussed in this paper can take the combined data matrix as input to compare the resulting significance of DGE vs. ASE.

D. Analytical Tools

1. T-Test Tool to Detect Independent Alternative Splicing Events

The toolbox includes a two-sample t-test tool for samples with equal but unknown variances that determines the statistical difference of each feature between the two phenotypic groups (DF and R). It uses the respective expression value matrices (NEVM and ICVM) as input to identify significant differential expression events and more importantly, independent alternative splicing events (IASE). The two-sample t-test looks at each feature and determines how statistically different the two phenotypes are based on a false discovery rate of 0.05. The tool then generates a sorted list of features using the obtained p-value to rank the level of statistical difference between the two classes. The tool is also able to create a bar plot for a specific feature or set of features to visualize the statistical difference between the two phenotypes. Error bars were added to the bar plots using the standard error of the mean calculation.

2. *OPLS-DA*

The OPLS-DA tool was created to generate predictive models of lung cancer recurrence using orthogonal projections to latent structures discriminant analysis (OPLS-DA) [10], [11], which is a supervised, multivariate modeling technique used to determine the variation within X or the expression/ alternative splicing data that is correlated to Y or the class labels, in this case, phenotype. A second variation within X that has nothing to do with Y (e.g. noise) was filtered out resulting in a single latent vector (LV), analogous to a principal component in PCA.

Once generated, the models are cross-validated to find classification accuracy, where one sample from each phenotype is repeatedly left out at random for validation. The process can be iterated a variable number of times but in this experiment it was repeated 200 times. The individual results of this repeated process is recorded to identify the average prediction accuracy for each sample. This resulting data can then be used to find the effectiveness of the predictive models and to identify potential confounding samples.

3. Feature Masking Tool via Genetic Algorithm

Hybrid genetic algorithms (GA)/Bayesian classifiers have previously been applied to medical and biological datasets in order to obtain a reduced set of putative biomarkers for phenotype classification [22]. A feature selection tool was implemented via a multiple objective genetic algorithm that attempts to minimize the number of features included in the model while still maximizing the cross-validated coefficient of determination (R^2):

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}},$$

where SS_{Res} is the sum of squares of residuals and SS_{tot} is the total sum of squares, which is defined as the sum of the squared differences of each observation from the overall mean [23].

The chromosome of the GA consists of a bit vector, where each bit corresponds to a single feature. Specifically, in this experiment, there are 6,203 gene expression features and 18,224 alternative splicing event features derived from the same set of genes. Thus, the genetic algorithm has no bias towards selecting more expression or alternative splicing events. When the bit is set to 1, the classifier includes the gene expression feature or alternative splicing event feature, otherwise the feature is ignored and removed from the model. The GA population consists of 20 initially random bit vectors containing on average 100 features enabled. In each generation, the four most fit individuals survive without modification. The 1-point recombination operator accounts for the remaining 80% of the population in successive generations. The GA employs a stochastic universal sampling operator for parental selection prior to recombination. The mutation operator employs a Gaussian distribution to select bits for mutation. Selected bits are simply flipped. Evolution proceeds for a maximum of 200 generations, though it is halted if no improvement in the average spread of the Pareto solutions is obtained for 50 consecutive generations [23].

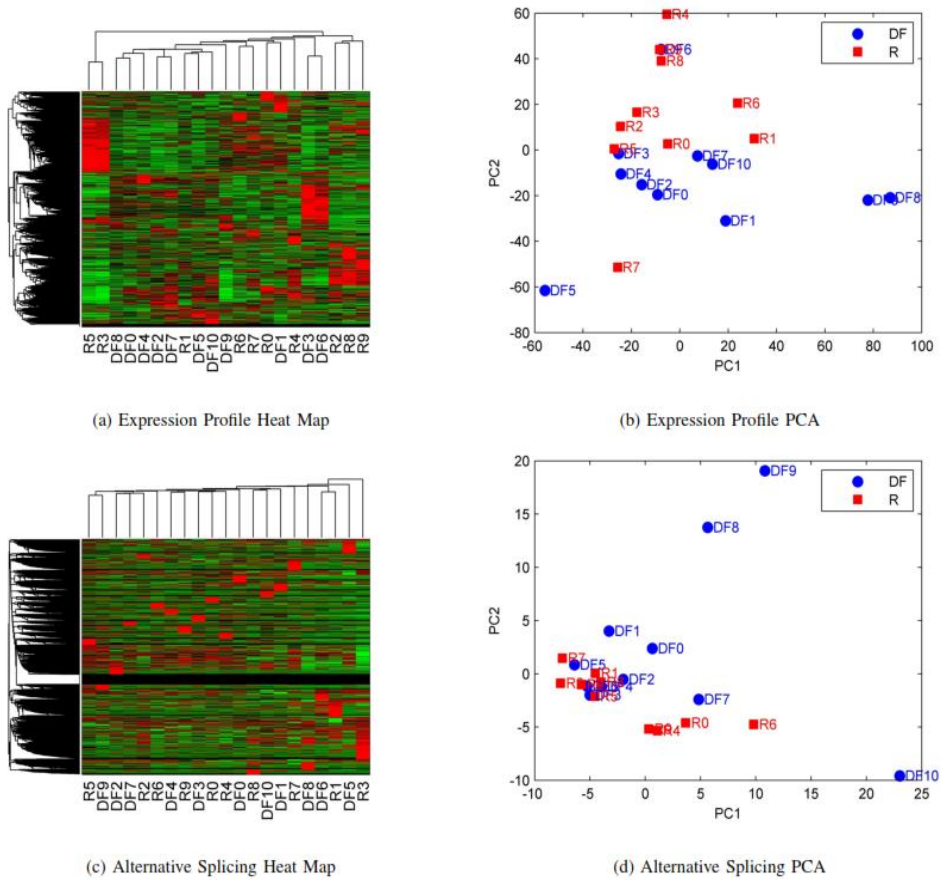
The two main objectives of the fitness function are (1) to maximize the cross-validated R^2 of a feature masked OPLS-DA model, and (2) to minimize the number of features included in the model. Additional benefits of the GA are that feature masking improves interpretability by

removing features with little or redundant impact on the classification accuracy. The average number of features selected over the validation procedure was measured, and the average percentage of those features that are ASE was also measured. These summary statistics were calculated using the solution on the Pareto front that resulted in the maximum R^2. The reduced set of genes generated by the GA serve as novel, putative biomarkers from both gene expression and alternative splicing events that can later be tested during subsequent validation studies.

Results

A. Unsupervised Analysis

After performing the cluster analysis and PCA, it was determined that these two unsupervised analyses did not show distinct phenotypic groups based on DGE or ASE prior to creating the predictive model. Heat maps were generated to show the hierarchical clustering of the biological samples (Figure 2 - DGE(a), ASE(c)). The PCA scores plot using DGE data (Figure 2(b)) shows a general clustering according to phenotype, but with several samples incorrectly clustering with the opposing class (e.g., DF with R group [DF6]; R with DF group [R7]). However, the PCA scores plot based on ASE (Figure (d)) does not show distinct clusters by phenotype. Further investigation of the analyses showed that several of the sample profiles are consistent with the other phenotype. This could be a result of confounding factors such as misclassification, tumor vs. stromal cell content, unique oncogene drivers or tumor suppressor gene los



(a) Expression Profile Heat Map

(b) Expression Profile PCA

(c) Alternative Splicing Heat Map

(d) Alternative Splicing PCA

Fig. 2: Heat maps from hierarchical clustering a) DGE and c) ASE or Principle Component Analysis (b), (d) derived from differential gene expression or alternative splicing data demonstrate that samples do not strongly correlate with clinical phenotype using unsupervised methods.

B. Predictive Models and Analysis

OPLS-DA predictive models were generated using the three inputs of alternative splicing events, differential expression, and the combined dataset of both ASE and DGE. The three models resulted in a cross-validated R^2 of 0.16, 0.07, and 0.09, respectively. The cross-validated accuracy of the three methods were 0.63, 0.90, and 0.84, respectively. The reason that the cross-validation summary statistics are low is most likely due to uniformly poor performance o a small subset of samples that exhibit profiles of the alternative phenotype.

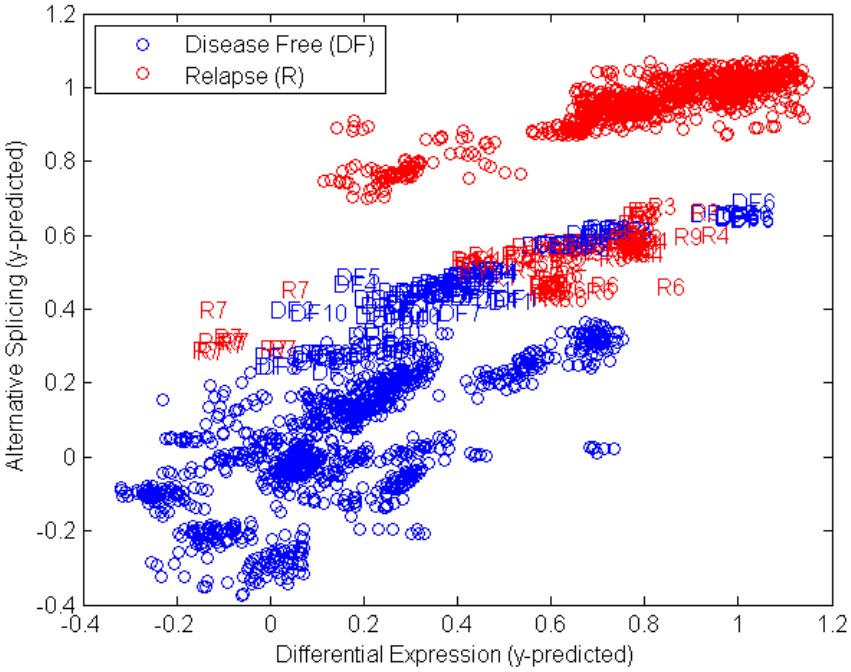


Figure 3: A combined scores plot visualizing the performance of models created from DGE and ASE during cross-validation.

An additional feature of the toolbox is a tool to generate a combined scores plot, which shows the compared performance of two methods, i.e., ASE vs. DGE, ASE vs. Combined, etc. The plot shows the predicted values for differential expression and alternative splicing OPLS-DA models (Figure 3), where the unlabelled circles are training samples that have been predicted via the iterative cross-validation described above. The label data points on the plot represent the validation set that was withheld during each iteration. It is apparent from the diminished separation between the groups that over-fitting occurred. The over-fitting would be due to the high-dimensionality of the data set relative to the sample size and the influence of confounding samples. The separation on the y-axis is entirely a result of the ASE. Conversely, the separation on the y-axis is entirely due to the DGE profiles. While both methods provide discriminant information between the two phenotypes, the information is not redundant, since the Pearson correlation, ρ, is between the alternative splicing y-predicted and differential expression y-predicted is 0.896. The average performance per sample for each model was investigated to measure the influence of these and other samples on the model. As mentioned above, the cause for misclassification would most likely be a result of confounding factors such as misclassification, tumor vs. stromal cell content, unique oncogene drivers or tumor suppressor gene loss.

C. Comparing Alternative Splicing Event Sets

Using the results produced by OPLS-DA to obtain the average rank of genes, isoforms, and

combined genes and isoforms, as well as the resulting p-values obtained from the two-sample t-tests on DE and ASE, the top 100 genes and isoforms for each test were determined. There were 49 genes and 66 isoforms that overlap in the list of the top 100 genes and isoforms, respectively, for OPLS-DA and the t-test. A Wilcoxon signed rank test on the distribution of absolute difference in rank order indicates that the overall rank order is significantly different between results from predictive modeling versus independent alternative splicing event identification ($p < 0.001$). This indicates that putative biomarker identification for our predictive model that uses a combination of multiple ASE to predict recurrence of lung cancer differs from the markers selected independently.

The top 100 genes and/or isoforms generated from the combined dataset of NEVM and ICVM resulted in a combination of both genes and isoforms. This indicates that the information contained in ASE is representative of a non-redundant set of genes correlated to phenotype and that ASE provides independent, supplementary information to DGE.

A heat map of the samples using the top 100 ranked genes from DGE and ASE predictive models during cross-validation is shown in Figure 4**.** As shown in the heat maps, there is a stronger cluster to phenotype association for ASE. Though DF samples do not form a distinct cluster based on ASE, R samples form a definitive phenotype cluster. This suggests that clustering by ASE features removes some of the contra-clustering noise from analysis. For these reasons, ASE features, such as those shown in Figure 4, provide useful, new information that can be used to create predictive models and find putative biomarkers that are non-redundant with those found with traditional DGE analysis.

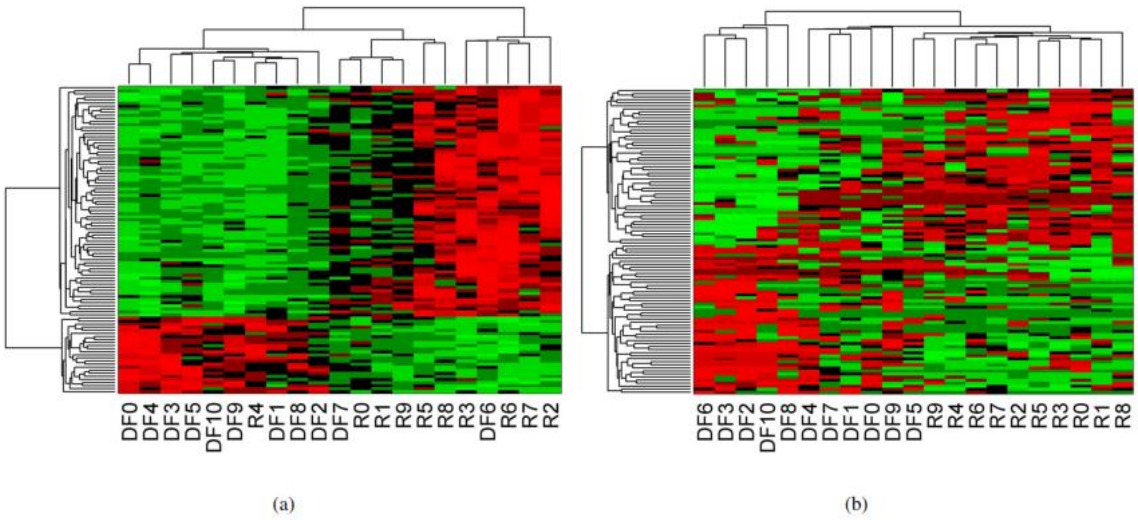


Fig. 4: Heatmap samples using top 100 ranked genes from (a) expression profiles and (b) alternative splicing event based predictive models.

D. Putative Combined Biomarkers via Feature Masking

The multiobjective genetic algorithm was used to generate subsets of putative genes and

alternative splicing events that could be identified as potential biomarkers, by maximizing the cross-validated R^2, while minimizing the number of genes and alternative splicing events included in the model. During the cross-validation iterations, the genetic algorithm terminated before the maximum number of generations was reached as the average change in the spread of the Pareto solutions was less than the specified tolerance. Feature selection was carried out simultaneously on an equal number of DGE and ASE features. If the gene expression data was consistently more predictive of the two phenotypes, the algorithm would show a distinct bias towards those features; however, after aggregating the results from the cross-validation iterations, the average percent of ASE features included was 46%. Further, the average number of features selected was 213 (i.e., features not masked). The feature selection resulted in a slight decrease in accuracy from 76% to 68%, but reduced the number of putative biomarkers from >24,000 to approximately 200 genes. The near equal selection of alternative splicing events and gene expression profiles indicates that alternative splicing analysis provides a complementary and non-redundant set of features for lung adenocarcinoma diagnosis and phenotype classification.

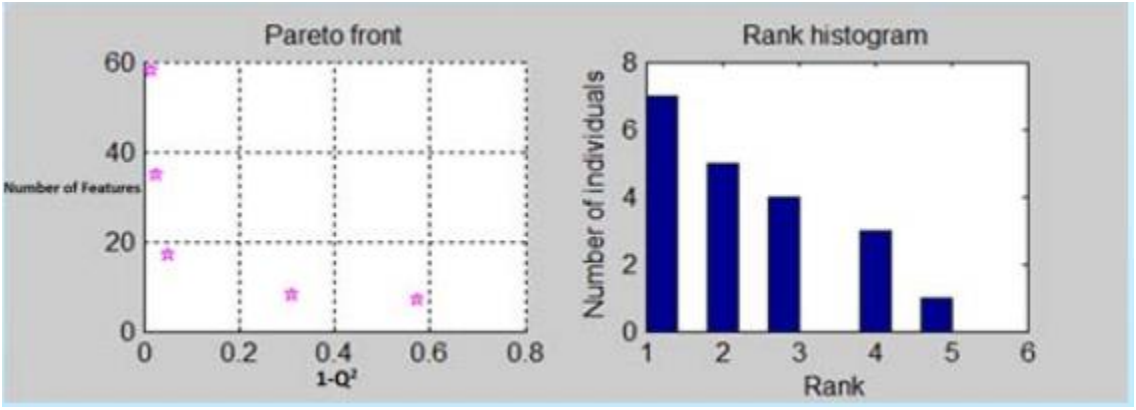


Fig. 5: Sample visualization of multiobjective genetic algorithm output, including the plot of Pareto front, and Rank of Individuals.

E. ROC Curve to Visualize Performance

In order to visualize performance of the implemented methods, a Receiver Operating Characteristic (ROC) curve feature was added to the toolbox. An ROC curve is a graphical plot that shows the performance of a binary classifier system as its discrimination is varied, by plotting the fraction of true positives versus the fraction of false positives [24]. As shown in Figure 5, ROC curves were generated using the results of the OPLS-DA models for (a) DGE, (b) ASE, and (c) Combined data. The figures were created using the leave-one-out cross-validated raw test data. The general shape of the resulting plots is what is expected to be generated from a relatively accurate predictive model.

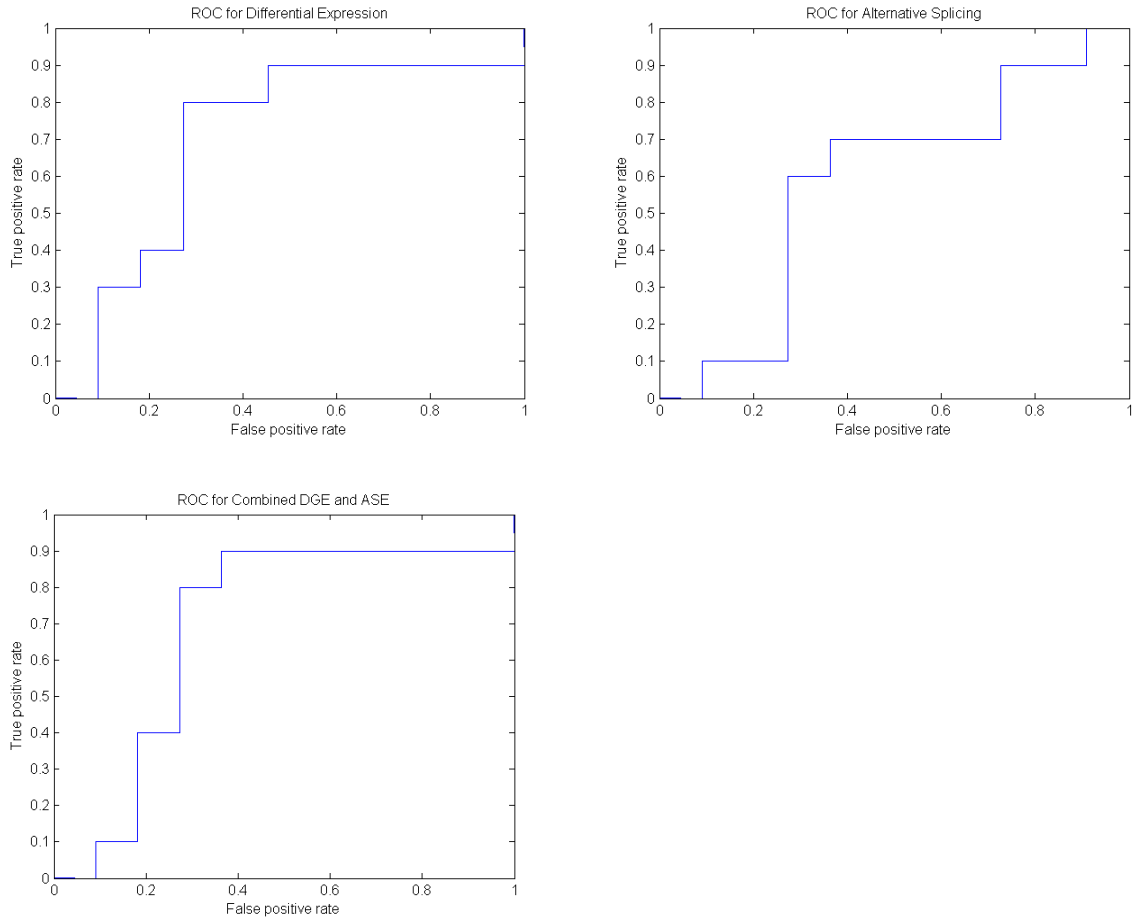


Fig. 6: The ROC curves illustrate the performance of classification by OPLS-DA for ASE, DGE and the Combined data, respectively. The graphs plot the fraction of true positives vs. the fraction of false positives, where Relapse represents the positive class.

Conclusion

The toolbox presented in this paper would offer a new analytical approach to bioinformatic analyses, while improving the efficiency of experimental design by offering a comprehensive set of tools. The novel method of this study combines alternative splicing events and differential gene expression data to determine whether specific patterns exist that are representative of lung cancer and correlate with tumor aggressiveness and patient prognosis. By identifying splice variants associated with either indolent or aggressive cancers for the selection of patients that are more likely to have their cancer relapse, physicians and researchers can more readily understand the disease and can offer more aggressive treatment to those affected. This approach to treatment has the potential to significantly improve response rates and patient survival. In addition, the identification of specific alternatively spliced variants offers a new resource for novel biomarker and therapeutic targets. Previous research in the field has only involved the functional role of individual splice variants in cancer progression, but the possible association of splice variant signatures with patient outcome had not yet been examined [22]. This study clearly addresses a area in cancer research that needs to be further researched and understood. Thus, because the

tools and features of this toolbox can also be applied to any similar biological data involving two phenotypes, it has the potential for significant clinical care improvement.

By comparing predictive models created from two types of alternative splicing analysis to a predictive model created from differential gene expression, it can be shown that ASE analysis provides non-redundant and complementary features which can be used for predictive models. Our genetic algorithm showed that ASE features contribute approximately the same to a feature masked predictive model as DGE features do. Cross-validation of OPLS-DA models from DGE and ASE showed that some samples consistently failed to be predicted, whereas for other samples predictive accuracy was model dependent; thus indicating that the features from DGE and ASE have complementary predictive power. A Wilcoxon signed-rank test showed that the top 100 significant features from ASE and DGE predictive models were significantly different, further supporting the claim that predictive features of lung adenocarcinoma are driven by ASE.

Future investigations will include validation of ASE and DGE phenotype predictive features, an analysis comparing the enriched gene sets and pathways affected by ASE and DGE features, performance of ASE predictive models on other disease phenotypes and measuring the effect of sequencing coverage on predictive model accuracy. Combining the rank orders of the ASE and IASE when selecting genes for validation is advised. Real time PCR or NanoString nCounter Analysis System (NanoString Technologies, Inc., Seattle, WA) analyses will be used for validation using additional DF or R patient RNA samples.

The results obtained from the analytical toolbox used in this study have shown that ASE is an important tool to be considered when creating predictive models for disease phenotype prediction. Not only were models created, but driving features were also identified. These specific features can be deemed as biomarkers and will be further investigated and used for future pathological and molecular investigations. The methods described can be used to predict other disease phenotypes as well, adding to the bioinformatic toolbox used by clinical researchers in pathology, functional genomics, and systems biology. A future goal would be to have the machine learning toolbox made publicly available to aid in the generation of predictive models and biomarker identification.

Acknowledgements

I acknowledge the support from Dr. Anderson in the College of Charleston Computer Science Department and the MUSC Oncology Department. The authors acknowledge funding from the SC EPSCoR/IDeA GEAR:Cyberinfrastructure Program (P. Anderson) and South Carolina State Appropriation (D. Watson). We also acknowledge support from the Genomics Shared Resource, Hollings Cancer Center, and Medical University of South Carolina. This shared resource is supported in part by the Hollings Cancer Center, Medical University of South Carolina Support Grant (P30 CA 138313).

References

[1] J. Botling, K. Edlund, and M. Lohr, “Biomarker discovery in non-small cell lung cancer : integrating gene expression profiling , meta-analysis and tissue microarray validation,” *Cell*, no. 150, pp. 1107–1120, 2012.

[2] E. O’Brien, J. Lerman, R. Chang, D. Hyduke, and B. Palsson, “Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction.” *Molecular Systems Biology*, vol. 9, p. 693, Oct. 2013.

[3] M. Griffith, O. L. Griffith, J. Mwenifumbo, R. Goya, A. S. Morrissy, R. D. Morin, R. Corbett, M. J. Tang, Y. Hou, T. J. Pugh, G. Robertson, S. Chittaranjan, A. Ally, J. K. Asano, S. Y. Chan, H. I. Li, H. Mcdonald, K. Teague, Y. Zhao, T. Zeng, A. Delaney, M. Hirst, G. B. Morin, S. J. M. Jones, I. T. Tai, and M. A. Marra, “Alternative expression analysis by RNA sequencing,” *Nature Methods,* vol. 7, no. 10, 2010.

[4] L. H. LeGault and C. N. Dewey, “Inference of alternative splicing from RNA-Seq data with probabilistic splice graphs.,” *Bioinformatics*, vol. 29, no. 18, pp. 2300–10, Sep. 2013.

[5] C. J. David and J. L. Manley, “Alternative pre-mRNA splicing regulation in cancer: pathways and programs unhinged,” *Genes & Development,* vol. 24, pp. 2343–2364, 2010.

[6] D. Emig, N. Salomonis, J. Baumbach, T. Lengauer, B. R. Conklin, and M. Albrecht, “AltAnalyze and DomainGraph: analyzing and visualizing exon expression data.,” Nucleic Acids Res., vol. 38, pp. W755–62, Jul. 2010.

[7] Y. Katz, E. T. Wang, E. M. Airoldi, and C. B. Burge, “Analysis and design of RNA sequencing experiments for identifying isoform regulation.,” *Nat. Methods,* vol. 7, no. 12, pp. 1009–1015, 2010.

[8] J. Wu, S. Sun, W. R. Mccombie, A. R. Krainer, and M. Q. Zhang, “SpliceTrap : a method to quantify alternative splicing under single cellular conditions,” *Bioinformatics,* vol. 27, no. 21, pp. 3010-6, 2011.

[9] Y. Hu, Y. Huang, Y. Du, C. F. Orellana, D. Singh, A. R. Johnson, A. Monroy, P.-F. Kuan, S. M. Hammond, L. Makowski, S. H. Randell, D. Y. Chiang, D. N. Hayes, C. Jones, Y. Liu, J. F. Prins, and J. Liu, “DiffSplice: the genome-wide detection of differential splicing events with RNA-seq.,” *Nucleic Acids Res.*, vol. 41, no. 2, p. e39, Jan. 2013.

[10] J. Trygg and S. Wold, “Orthogonal projections to latent structures (O-PLS),” Journal of Chemometrics, vol. 16, no. 3, pp. 119–128, 2002.

[11] M. Bylesj, M. Rantalainen, O. Cloarec, J. Nicholson, E. Holmes, and J. Trygg, “Opls discriminant analysis: combining the strengths of pls-da and simca classification.” Journal of Chemometrics, vol. 20(8-10), pp. 341–351, 2006.

[12] C. Trapnell, A. Roberts, L. Goff, G. Pertea, D. Kim, D. R. Kelley, H. Pimentel, S. L. Salzberg, J. L. Rinn, and L. Pachter, "Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks." Nature protocols, vol. 7(3), pp. 562–78, Mar. 2012.

[13] L. Goff, C. Trapnell, and D. Kelley, CummeRbund: Analysis, exploration, manipulation, and visualization of Cufflinks high-throughput sequencing data., 2012, r package version 2.2.0.

[14] J. Goecks, A. Nekrutenko, J. Taylor, and T. G. Team, "Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences." Genome Biology, vol. 11(8):R86, Aug. 2010.

[15] D. Blankenberg, G. Von Kuster, N. Coraor, G. Ananda, R. Lazarus, M. Mangan, A. Nekrutenko, and J. Taylor, "Galaxy: a web-based genome analysis tool for experimentalists." Current Protocols in Molecular Biology, vol. Chapter 19:Unit 19.10.1-21, Jan. 2010.

[16] B. Giardine, C. Riemer, R. Hardison, R. Burhans, L. Elnitski, P. Shah, Y. Zhang, D. Blankenberg, I. Albert, J. Taylor, W. Miller, W. Kent, and A. Nekrutenko, "Galaxy: a platform for interactive large-scale genome analysis." Current Protocols in Molecular Biology, vol. 15(10), pp. 1451–5, Oct. 2005.

[17] S. Andrews, "Fastqc a quality control tool for high throughput sequence data." [Online]. Available: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

[18] M. Lohse, A. Bolger, A. Nagel, A. Fernie, J. Lunn, M. Stitt, and B. Usadel, "Robina: a user-friendly, integrated software solution for rna-seq-based transcriptomics." Nucleic Acids Research.

[19] C. Trapnell, B. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. van Baren, S. Salzberg, B. Wold, and L. Pachter, "Transcript assembly and abundance estimation from rna-seq reveals thousands of new transcripts and switching among isoforms." Nature Biotechnology, vol. 28(5), pp. 511–515, Jul. 2010.

[20] MATLAB, version 7.10.0 (R2010a). Natick, Massachusetts: The MathWorks Inc., 2010.

[21] I. Jolliffe, Principal Component Analysis. New York, New York: Springer-Verlag, 1986.

[22] M. Raymer, T. Doom, L. Kuhn, and W. Punch, "Knowledge discovery in medical and biological datasets using a hybrid bayes classifier/evolutionary algorithm," IEEE Transactions on Evolutionary
Computation, pp. 802–813, 2003.

[23] P. E. Anderson, M. R. Paul, E. S. Hazard, V. A. Mccaffrey, D. K. Watson, and P. M. Watson, "Predictive Modeling of Lung Cancer Recurrence using Alternative Splicing Events versus Differential Expression Data," 2014.

[24] A. P. Bradley, “The Use of the Area Under the ROC Curve in the Evaluation of Machine Learning Algorithms,” Pattern Recognition, vol. 30, no. 7, pp. 1145–1159, 1997.