Application of Bioinformatics in the Revelation of NSCLC Biomarkers and Potential Targeted Drug Therapies

Mukund Subramanian1*, Jahangheer Shaik2, Rakesh Nagarajan2, and Theodore Waterson1

Student1, Teacher2: Ladue Horton Watkins High School, Saint Louis, MO 63124
Mentor2: Center for Biomedical Informatics, Department of Pathology and Immunology, Washington University, Saint Louis, MO 63124
*Corresponding author: mookiesub@gmail.com

Abstract
Lung Cancer is the leading cause of cancer-related death in the United States; yet despite this fact, common misconceptions hinder progress towards a better understanding of this fatal carcinoma. It is divided into two crudely differentiated sub-types which are distinguished by the appearance of their respective cells under a microscope: Small-Cell (SCLC) and Non-Small-Cell Lung Carcinomas (NSCLC). In order to find possible genetic perturbations causal for or associated with these diseases, scientists habitually utilize the voluminous capacity of microarray technology, which harnesses the process by which single-stranded DNA, derived from mRNA of differentially expressed genes, hybridized to their complementary probes immobilized on a chip. This genome-wide data is provided to bioinformatics laboratories for analysis, the focus of this investigation. We started with expression microarray data of Adenocarcinoma, Large Cell Carcinoma and Bronchioalveolar Carcinoma (sub-types of NSCLC) tissue from 149 patients. After performing Quality Control Checks to ensure that the data was entirely free of experimental biases, such as the Batch Effect, we proceeded to conduct differential analysis testing on 120 samples from patients with Adenocarcinoma, Large-Cell Carcinoma and Bronchioalveolar Carcinoma: sub-types of NSCLC. Data from these sub-types were put through the Significance Analysis of Microarrays (SAM) Test, to obtain a list of genes that were more significantly expressed in a certain NSCLC sub-types in comparison to its counterparts. The False Discovery Rate (FDR) of significant genes was set to 5% overall or average local if needed. We also conducted a 0% FDRSAM Test in order to identify potential NSCLC Biomarkers. We plugged the filtered 5% FDR differentially-expressed gene set into the GRANITE tool which draws Gene Relational Networks and identifies enriched pathways in a set of genes. After identifying these pathways, the most enriched genes in these pathways were plugged into the DrugBank Database in order to isolate specific drugs that would stall a gene, inhibit a rogue pathway or potentially thwart the cancer itself. The resulting genes in the 0% FDR data set could be used in the diagnosis of lung cancer as potential biomarkers of the disease. Hopefully, these beneficial results will pave the way for potential personalized targeted drug therapies to Lung Carcinomas.

Introduction
Lung Cancer is the leading cause of cancer-related death in the United States; in fact the vicious carcinoma claimed the lives of a stunning 158,683 out of 203,536 lung cancer patients in 2007. The scale of lung cancer death is so colossal, that in 2009, a greater number of lung-cancer-diagnosed patients died than those who died of breast, colon, pancreas and prostate cancers combined.9 It is projected that, by 2015, the worldwide number of Lung Cancer deaths alone will amount to a staggering 1,676,000.3 The survival rate for patients with lung cancer is a mere 15% in 5 years: the lowest survival rate of any other cancer.1 Yet current methods used to “treat” lung carcinomas, specifically chemotherapy, have been largely fruitless in their attempts to extend the lives of lung-cancer diagnosed patients. A recent study demonstrated that chemotherapy only contributes to the 5-year survival of 2.3% of all cancer cases.20 In skin cancer research, another study showed that vemurafenib,21 a drug that specifically targets a gene mutation rather than an entire pathway (chemotherapy’s ineffective method), has proven to be 8 times more effective than chemotherapy in shrinking tumors in patients. It seems that the identification of targeted therapies such as this one will lead to a new dawn in treatment. However, in the United States, lung cancer receives just $1,200 of federal funding per death, while breast cancer receives more than $27,000 per death, followed by $14,000 for prostate cancer and $6,500 for colon cancer.22 This lack of funding leads to a subsequent deprivation of research, which has starved the carcinoma of potential targeted therapies. These stunning figures provided the motivation for us to pursue research in this cutting-edge field of oncology.

To begin research however, it was imperative that we truly understood Lung Cancer; contrary to common misinterpretations, lung cancer is actually a heterogeneous assortment of tumors of the lung, bronchus, alveoli and pleura.3 It is divided into two crudely differentiated sub-types known as Small-Cell and Non-Small-Cell Lung Carcinoma, indicative of what these respective cells look like under the scrutiny of a microscope. Of these Carcinomas, Non-Small Cell Lung Carcinomas, the focus of this study, represent the overwhelming majority of diagnoses at approximately 80.4%.4 Non-Small Cell Lung Carcinoma (NSCLC) is split into even more definitive sub-types: Adenocarcinoma, the most common form of all lung cancers, Squamous Cell Lung Carcinoma, Large-Cell Lung Carcinoma, and Bronchioalveolar Carcinoma, which, in this study, is considered to have its own explicit NSCLC Histology. These sub-types are widely accepted by the scientific community and provide a basis for oncologists to diagnose and treat with greater accuracy and specificity, due to differences in tumor appearance, location and formation.

The need for better analysis of tumors on a genome-wide
scale for specific genetic analysis has, as a result, instigated the increasingly habitual application of high-throughput screening assays by researchers, including DNA microarray technology. The voluminous capacity of this highly advanced technology enables it to analyze over 50,000 genes at once. The microarray is essentially a small chip with countless depressions upon it, each of which contains picomoles of various DNA sequences that encompass the entire human genome. Genes from samples are able to hybridize to their respective DNA counterpart sequences, permitting scientists to identify which genes are expressed more often in healthy patients and which genes are more active in cancer-diagnosed ones. Yet the true essence of microarray data is numbers. Each of these numbers defines levels of gene expression in cancer – the higher the number, the more the gene is expressed in cancer and vice-versa. Of course for a small data set, such tests could be easily done by hand. However the problem remains that one would have to painstakingly analyze 30,000 genes for each of the 149 patients in this study – no simple task by hand. This is where this project’s computer testing is absolutely imperative, for it effectively finds the genes that have the highest expression levels in a set of samples. After microarray testing, data is provided to bioinformatics laboratories for data analysis.

This study, an application of advanced biomedical informatics, attempts to determine significant differentially-expressed genes in microarray data in order to ameliorate cancer diagnosis and prognosis of three of the four NSCLC sub-types: Adenocarcinoma, Large-Cell Carcinoma (LCC), and Bronchioalveolar Carcinoma (BAC).

**Materials and Methods**

To begin, we compiled microarray data from a wide variety of lung cancer patients, courtesy of the Washington University Department of Pathology and Immunology. Originally, we amassed a cohort of 149 lung cancer patients diagnosed with a variety of NSCLC Sub-Types and even further definitive classifications within these sub-types. This dataset was also a compilation of data from patients whose tests were conducted on two different Affymetrix Microarray Probe Platforms: Hu133+2 and Hu95AV2. Using an Integrated Genomics Suite called Partek, we endeavored to ensure that we used only the most accurate samples in our analysis. Thus, we also put the dataset through several statistical quality control checks. We used a Principal Component Analysis (PCA) Plot at this stage in the quality control checking – a useful statistical technique used to find patterns in datasets of high dimension, which we planned to use to find patterns within the microarray data of the various NSCLC sub-types to find relative similarity but some difference amongst the data obtained from patients diagnosed with the aforementioned types of carcinomas. Subsequently, in order to identify the significantly differentially-expressed genes distinguishing each NSCLC histological category, we surmised that the Significance Analysis of Microarrays (SAM) Test would be the best option in terms of the determination of these gene sets. We proceeded to put the data obtained after the Quality Control Checks into an Excel 10 File in order to utilize SAM. We used a one-versus-all testing method in which we compared Adenocarcinoma to the other types; LCC to the other sub-types and so on. The purpose of this was to identify genes that were more significantly expressed in these specific carcinomas in relative comparison to their counterparts. We proceeded to conduct the SAM Test using a set number of 100 algorithms in each instance. We first, however, desired to set a specific False Discovery Rate (FDR) – in essence, a statistical measure which accounts for the fact that some of the genes identified by the test could be falsely identified (false positives) – that we wanted the SAM Test to output. So, based on a study done by Harvard, Stanford and UNC-Chapel-Hill researchers on Adenocarcinoma using SAM, we concluded that a 5% FDR would suffice to eliminate error, but still obtain a large list of cancer-causing genes. If less than 200 genes were selected, the conditions were iteratively softened until a dataset of the desired size was acquired. Three SAM Tests were conducted in this manner with an overall 5% false discovery rate. Although the Adenocarcinoma vs. All and Large Cell Lung Carcinoma vs. All SAM Tests yielded reasonable results, the BAC vs. All Test did not dispense as desirable an outcome. With an output of a mere 5 genes, we were forced to relax the restrictions on the SAM parameters to a 20.84% FDR. However, this was only the overall FDR, and the average of the obtained local FDRs for the significant genes in BAC samples was roughly 5.69%. The average local FDR took into account only the genes of interest, rather than allowing the underexpressed genes to sway the FDR to greater levels. Thus, using the local FDR in this instance is permissible and would yield better results. For the Adenocarcinoma and LCC Samples’ Tests the FDR was 5.68% and 5.34% respectively. The size of the resulting significant gene sets for Adenocarcinoma, LCC and BAC respectively were: 3589, 2821, and 204 genes. We also conducted three other SAM Tests in the same manner, but with a 0% FDR, in order to identify an explicit set of significant genes that would be instrumental in cancer diagnosis. In addition to identifying specific sets of genes unique to/enriched in each NSCLC sub-type, we utilized a novel application created by Jahangheer Shaik, one of the major contributors to this project, called the Gene RelAtional Network of InTeracting Elements (GRANITE), which assists Bioinformaticians and Systems Biologists in depicting Gene Relational Networks (GRNs); GRNs depict functional relationships among genes, and are extremely complex, involving multiple genes and gene products operating at multiple levels. GRANITE provides a user-friendly platform on which data from several different databases can be integrated onto one site. The GRANITE tool allows one to assess which biological pathways in the human body are enriched in a certain set of genes, as well as the transcription factors inhibiting certain genes in any given dataset. Before conducting the GRANITE Tests we determined a list of approximately 36 pathways that we believed were important in the process of Lung Cancer, after researching online at the four premier Pathway Databases: KEGG, Reactome, PANTHER, and BioCarta. We conducted three different GRANITE tests by plugging in the enriched dataset obtained in the SAM Test for each respective NSCLC Histology. After conducting the tests we determined which pathways had the largest gene relational network and/or the largest number of genes for each NSCLC sub-class and determined the top three in each. If there were any ties in the number of GRNs in the pathways given, all of the tied pathways were listed. The final test we conducted was the
Results

Quality Control Check Results: When, at first, we viewed the PCA Plot of the data on Partek, as shown in Figure 1a, we were convinced that something was amiss. The data, despite the fact that all of it concerned Lung Cancer, was extremely definitive as it was split into two explicit groupings within the cohort of patients with the same disease. We then proceeded to label all of our data with proper probes mentioned earlier: Hu133+2 and Hu95AV2.

The results of this Probe Mapping are shown in Figure 1b. As is clearly visible the notorious Batch Effect is taking place in our data. The Batch Effect, one of the great drawbacks to microarray technology, occurs when microarray tests are conducted at different times or in different conditions, leading to a lack of comparability between the datasets. As our data contained the older Hu95AV2 and the newer Hu133+2 probes, the Batch Effect clearly grouped the two probe types together in a non-biologically-related, biased method. In order to rid the data of the batch effect, we normalized the data using the Quantile Normalization Process, a statistical technique for making two distributions identical in statistical properties, and obtained a Box-and-Whiskers Plot of the Data shown in Figure 1c. Satisfied that the normalization had gone smoothly, we then proceeded to get rid of the outliers.

However, the removal of the Batch Effect using Partek left no clear method by which to identify patterns in a PCA Plot, as shown in Figure 1d, we at last reasoned that we ought to remove the older Probe Type: the Hu95AV2 Probe and conduct our testing on the newer Hu133+2 Probe Type Data only. In the figure, it is impossible to find any subtle patterns which will help us slightly distinguish the various sample types. This “distinguishability” is an essential facet to any good dataset. Since there were only 25 samples conducted on the Hu95AV2 microarray probe-type, we chose to eliminate these probes to allow for comparability between the remaining samples without the Batch Effect or other errors, and since this older probe type is more prone to error anyway. After eliminating the Hu95AV2 Probe Data from our file and obtaining the PCA plot for the remaining Hu133+2 data, we eliminated the outliers and obtained the final PCA plot shown in Figure 1e. This excellent plot is not too distinct, yet it is explicit enough to be able to differentiate...
between the sub-types, showing that our data is largely accurate and sufficient to test on. After completing Quality Control Checking there were a total of 120 microarray patient samples remaining (all Hu133+2); in total there were, 24 LCC samples, 13 BAC samples and 83 Adenocarcinoma samples. The successful completion of these cursory tests allowed the study to proceed using the final dataset from the Quality Control Checks.

SAM Results: After conducting the following 5% FDR SAM Tests – Adenocarcinoma vs. All, LCC vs. All and BAC vs. All — we obtained the plots shown in Figures 2a, 2b, and 2c in the same respective order. The significant genes obtained after the SAM Test were subsequently plugged in to the GRANITE tool in order to identify genes involved in cancerous pathways.

GRANITE Results: In Adenocarcinoma we discovered that the Janus Kinase-Signal Transducer and Activator of Transcription or Jak-STAT Signaling Pathway had the highest number of GRNs, followed by a three-way tie for second between Angiogenesis, mTOR Signaling, and Regulation of the Actin Cytoskeleton, and succeeded by Tight Junction in 3rd Place. However, in an odd turn of events the Regulation of the Actin Cytoskeleton Pathway possessed the largest number of genes, with Jak-STAT in tow and the remainder of the pathways following in the very same sequence. In Large-Cell Carcinoma, Jak-STAT was once again found as the pathway with the greatest amount of GRNs; however the remaining pathways did not at all resemble the results obtained in the Adenocarcinoma GRANITE test. After Jak-STAT, Wnt Signaling was found as the second largest pathway in terms of enriched genes followed by the EGF-Receptor (EGFR) Interaction Pathway in third. Lastly, for Bronchioalveolar Carcinoma, none of the pathways we ascertained to be significant in tumorigenesis and the progression of Lung Cancer were present, except one pathway: Wnt Signaling, which had quite an impressively large GRN for such a comparatively small dataset. Observers of this study must note that overly general and ambiguous pathways that could be linked to a plethora of different diseases were not used or analyzed as a part of this research. Two notable examples would be the MAP-Kinase (MAPK) Signaling Pathway and the well-known pathway calling for programmed Cell Death, Apoptosis.

The Results of GRANITE in terms of the enriched genes are displayed in the tables shown in Figures 3a, 3b and 3c, in the following respective order: BAC, Adenocarcinoma and LCC.

The GRANITE results of this study are consistent with some of the most highly-respected scientific literature currently in the field. We noticed that many of the genes involved in Adenocarcinoma were identified in both this study and others by Tongji University\(^{19}\), and the University of Michigan\(^{24}\). Significant genes involved in LCC match many genes found in a study presented at the famous Seminar in Cell Developmental Biology\(^{17}\). Lastly, many significant genes corresponding to BAC in this
study are consistent with those found in a North Shore University Hospital investigation\textsuperscript{25} and a University of California study\textsuperscript{18}. In addition, the aforementioned investigation conducted by Harvard University, Stanford University, and the University of North Carolina Chapel Hill\textsuperscript{7} also confirms several of our GRANITE-identified genes. This consistency with top literature clearly establishes the benefits of the methods utilized in our study and promises much.

Using the resulting Gene List obtained from the GRANITE Test, we proceeded to identify drugs which would halt the carcinogenic impacts of the significant genes involved on cancer-related pathways.

\textbf{DrugBank Results:} Overall, we obtained 21 drugs that inhibited some of the rogue pathways which aided lung tumor cells in tumorigenesis and tumor maintenance.

For Adenocarcinoma, we discovered that the following drugs would inhibit a certain mutant gene in their respective corresponding pathway: the drugs Adenovite, Adephos and Azucaps all inhibited the PRKAA1 Gene involved in the mTOR Signaling Pathway, Sprycel impeded the STAT5B Gene involved in the Jak-STAT Signaling Pathway, Certican and Torisel each repressed the FRAP1 Gene in the mTOR Signaling Pathway, and lastly, Hydroxyfasudil suppressed the ROCK1 Gene in the Regulation of the Actin Cytoskeleton Pathway. These results are shown in Figure 4b.

For Large Cell Carcinoma, we found that the following drugs – Pegasys, Wellferon, Intron A, Avaferon, Betaseron, Roferon A, Alferon and PEG-Intron – all inhibited the IFNAR2 Gene involved in the Jak-STAT Signaling Pathway, Actimmune impeded the IFNGR1 Gene involved in the Jak-STAT Signaling Pathway, Adagen repressed the GRB2 Gene in the Jak-STAT Signaling Pathway, Aerolone checked the PIK3R1 Gene involved in the Jak-STAT Signaling Pathway, Sprycel constrained the STAT5B Gene in the Jak-STAT Signaling Pathway and lastly, Eskalith suppressed the CTNNB1 Gene in the Wnt Signaling Pathway. These results are shown in Figure 4c.

Finally, for Bronchioalveolar Carcinoma, we ascertained the following drug as a proper inhibitor: the drug Velcade stalled the PSMB2 Gene in the Wnt Signaling Pathway. These results are shown in Figure 4a.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{fig1c.pdf}
\caption{Final PCA Plot of Data after elimination of the Hu95AV2 Probe Type samples and the elimination of the Batch Effect. This plot shows slight subleties between the sub-types yet not overwhelming discrepancy in sample expression in patients of the same general disease.}
\end{figure}

\begin{figure}[h]
\centering
\includegraphics[width=0.4\textwidth]{fig2a.pdf}
\caption{SAM Plot of Adenocarcinoma samples indicating a much larger number of samples overexpressed in comparison to other sub-types and an acceptable FDR.}
\end{figure}

\begin{figure}[h]
\centering
\includegraphics[width=0.4\textwidth]{fig2b.pdf}
\caption{SAM Plot of LCC samples indicating a relatively large number of samples overexpressed in comparison to other sub-types and an even lower FDR than that of the Adenocarcinoma samples.}
\end{figure}
Discussion

According to the 2010 Report of the American Thoracic Society, early diagnosis plays a crucial role in prognosis in lung cancer and would be greatly supported by the discovery of a simple marker of lung cancer. Pinpointing genetic predisposition to lung cancer could augment the efficacy of tumor screening. This may be related to deviations within genes that are associated with constraining cell growth and other similar biological pathways involved in the potential creation and maintenance of lung tumors. Gene expression profiling experiments have indicated that biomarkers of this nature may, in the future, be valuable in the diagnosis of lung cancer.

For the 0% FDR SAM Test, we obtained a list of 135 Genes in Adenocarcinoma, and 45 Genes at the same FDR in LCC. However at the 0% FDR for BAC we only obtained one significant gene, so we suggest that those interested in this study use the original BAC 5% Average Local FDR Dataset in Diagnosis. This list of significant genes could be useful in the advancement and amelioration of early Lung Cancer identification, as these genes are the very same biomarkers which are imperative to the betterment of diagnosis. The clinical relevance of these genes could be promising to future oncologists, as they could identify whether a patient has a form of NSCLC in a faster and more reliable way, enhancing the chances of prevention in the early stages.

This study not only improves diagnosis, but lengthens prognosis as well. The drugs targeting the rogue pathways that we have identified through the GRANITE and DrugBank Tests could be future therapeutic substances taken by patients in order to ease the burden of the disease and potentially rid them of it. If a patient is diagnosed with Adenocarcinoma, for example, he or she could be put on one of the identified drugs in the following list: Adenovite, Adephos, Azucaps, Sprycl, Certican, Torisel or Hydroxysufadil. These drugs would essentially inhibit certain genes in a rogue pathway and hopefully stop the pathway from aiding the cancerous tumor. For example, Azucaps would stall the functioning of the Jak-STAT Signaling Pathway, a crucial participant in Non-Small-Cell Lung Cancer due to the fact that it is responsible for antiapoptosis.

Admittedly there could be flaws in the research. Our data did not include the Squamous Cell Carcinoma sub-type of NSCLC and it used only 100 algorithms in the SAM Test due to computer capabilities as opposed to the maximum of 2000 to ensure near-perfect accuracy. Or, perhaps, the Quality Control Checks did not go as planned. Whatever the issue, this study still proposes a promising idea for the future, and is accurate enough despite these setbacks.

In conclusion, the results of this investigation should be looked into by drug testers to determine the effectiveness of the genes we identified on lung-cancer-diagnosed patients. Hopefully this will aid doctors in making headway in the translation of the identified genes and drugs into proper cancer treatment.
References


identified by bioinformatics analysis of public expression data. Cancer Research, 67(15), 7431-7438. AMER ASSOC CANCER RESEARCH.


