# Integrated Gene Expression Probabilistic Models for Cancer Staging

Andrew H. Xia, Rivers School, Weston, MA 02493

Completed in 2012 as part of MIT PRIMES research for high school students

## Executive Summary

The purpose of this research project is to explore the vast amount of cancer data available for research and assess the usefulness of the data. This is a pilot project studying clinical cancer data as well as RNA sequencing data. The Cancer Genome Atlas (TCGA) data is downloaded and used for analysis. When the clinical cancer data is studied, it is found that a significant portion of data is missing for many patients. Cancer patients have their overall clinical cancer stages derived from their T, N, and M stages. However, often times the T, N, or M staging information are missing, indicating that the clinical cancer stage for each patient might be computed improperly. Therefore, it is necessary to verify the accuracy of cancer stages and deal with the problem of incomplete data. Analyzing the incomplete cancer data, recovering missing information, and determining each patient's cancer stage are the central goals of this project. We have demonstrated that cancer stages cannot be simply extrapolated using probability distributions. A new approach of using RNA sequencing data to determine cancer stages is explored. This method is more cost efficient and safer than surgery for staging cancer patients. Our work has established a statistical relationship between a list of genes and breast cancer stages.

# Integrated Gene Expression Probabilistic Models for Cancer Staging


## Abstract

The current system for classifying cancer patients' stages was introduced more than one hundred years ago. With the modern advance in technology, many parts of the system have been outdated. Because the current staging system emphasizes surgical procedures that could be harmful to patients, there has been a movement to develop a new Taxonomy, using molecular signatures to potentially avoid surgical testing. This project explores the issues of the current classification system and also looking for a potentially better way to classify cancer patients' stages. Computerization has made a vast amount of cancer data available online. However, a significant portion of the data is incomplete; some crucial information is missing. It is logical to attempt to develop a system of recovering missing cancer data. Successful completion of this research saves costs and increases efficiency in cancer research and curing. Using various methods, we have shown that cancer stages cannot be simply extrapolated with incomplete data. Furthermore, a new approach of using RNA Sequencing data is studied. RNA Sequencing can potentially become a cost-efficient way to determine a cancer patient's stage. We have obtained promising results of using RNA sequencing data in breast cancer staging.

## 1. Introduction

In the last two decades, there has been an explosion of data. With the help of computerization and Internet connection, it has become a lot easier to collect, store, and transfer data across the globe. With the convenience of accessing to an infinite amount of data, some issues have also arisen. In many areas, we seem to fall behind in analyzing the data and extracting meaningful information. The speed of collecting data often outpaced our ability of processing the information. Medical data is certainly no exception to this recent phenomenon.

In cancer research, there are many different sources of cancer data, offering similar measurements but in different formats, causing confusion for researchers. Also, the data itself is often incomplete and missing critical information.

The Cancer Genome Atlas (TCGA) [1] has been used as the primary data source of cancer information [2]. TCGA is supported by the National Cancer Institute. It is a growing website filled with information about more than two hundred types of cancer and millions of patients' cancer data. This website includes many forms of cancer data, including clinical data, SNPs, DNA Methylation, RNA sequencing, and more, across multiple levels of processed data.

In my research, the initial focus was on analyzing clinical cancer data. From TCGA I downloaded information about twenty types of cancer, and looked at the clinical data of the patients. Within the clinical cancer data, different types of statistics were given across the cancer types. It is often difficult to analyze the cancer types as a whole. After initial examination on the patient data and their cancer staging, it became rather clear that many patients' cancer stages were improperly classified. The current system of classifying cancer patients has been around for over one hundred years [3-4], which is known as the TNM staging system. This system describes the extent of cancer spread within the patients' body, focusing on the size of the tumor (T), the

type and amount of regional lymph nodes that are involved (N), and whether the cancer has metastasized and spread to different parts of the body (M). T, N, and M are all numerical values that combine to give a clinical cancer stage, commonly known to be stages I, II, III and IV. For example, if a patient has a T stage of 3, an N stage of 2, and an M stage of 0 in breast cancer, the patient is considered to have Stage III breast cancer.

Analysis of TCGA data shows that the clinical data of many cancer patients lack T, N, and M stage information. This is problematic because anyone who uses the cancer data cannot be certain if the patient's cancer stage is really as printed. Many patients miss either one or two numerical values of T, N, and M, while still having a clinical cancer stage. This indicates that the cancer stage may not be determined correctly. For example, by definition one does not know what stage of cancer a patient has if the M stage data is left unavailable. Furthermore, when we conducted a test to compare the TCGA determined stage information with the definition of a clinical cancer stage based on T, N, and M stage data, we found that they were not always in agreement. Hence, issues exit in both completeness and accuracy of the clinical cancer data [5].

In our research, different data sources were used in order to find and fill in the missing information. Data entries beyond T, N, and M were studied in relation to cancer stages. For example, the 'days to death' column in the clinical cancer data table helps to determine cancer stages in retrospect, as there is a clear correlation for many cancer types between days to death and cancer stages. Six different methods were compared in statistics for accuracies in determining cancer stages using clinical cancer data.

A new approach was tested by using the RNA sequencing data. Specifically, RNA sequencing data in breast cancer was studied. Careful analysis between the clinical cancer data and the RNA sequencing data showed that a relationship can be established between a patient's

cancer stage and the genes in RNA sequencing data table. Our statistical study demonstrated high and low correlations of these factors.

In summary, the problem I attempted to solve here is to determine patients' cancer stages with incomplete cancer data. Cancer stage determination has many useful applications, not only it can help save costs, but more importantly it helps in cancer treatment and improves cancer patient's life quality. In this research project, several different methods were used and compared in accuracy. A new approach of using RNA sequencing data to determine cancer stages was experimented, which yielded in encouraging results.

## 2. Materials and Methods

The TCGA data portal is one of the main sources in this research. Computational analysis is performed on the data. At the initial stage, clinical cancer data is the primary focus. Raw clinical cancer data is formatted into standard forms and imported into the computer models written by me in R. R is a statistical programming language which is based on the 'S' language. In this interface there are many built in functions that allow the processing of data to be smoother and easier. Once the clinical data is imported into the model, statistical analysis is conducted. In TCGA, there are four levels of data, with Level 1 being unprocessed and Level 4 being highly processed. Level 3 data is the main focus in this project. There are many data entries for each patient. In addition, each cancer type has a large number of patients. The raw data is presented in multiple tables with patients in rows and parameters in columns. The parameter data includes background information such as how the patient's data is processed, and it also includes relevant statistics to the cancer. For example in the lung cancer section there is a column mentioning whether the patient is a smoker or not.

Of these parameters, our focus is mainly on four columns in the data table, which include the clinical overall cancer stage, the T stage, the N stage, and the M stage. Examination through these clinical cancer data shows that there is noticeable inconsistency in the overall clinical cancer stages and the relevant T, N, M parameters. For many patients, one of the T stage, or N stage, or M stage information is missing. For many other patients, a combination of the three stages is missing. This kind of data is considered to be Missing Not at Random (MNAR). The T, N, and M stages are critical in determining the patient's cancer stage in the current taxonomy. Several different tests are run on the incomplete data. It is found that for many cancer types there is a discrepancy between TCGA-published stages and the stages derived from the definition of the taxonomy.

More attention is then paid to RNA sequencing data. We are primary interested in looking at how RNA sequencing data can be related to clinical cancer data, as it is a relatively new technology. Not much research work has been done linking the usefulness of RNA sequencing data with clinical cancer staging.

RNA sequencing looks at the frequencies of each gene in a cell of a human body [6-8]. In RNA sequencing, a sample of a patient's body cells is extracted for analysis. The RNA strands of the cells float into millions of wells with an average diameter of 44 μm. Afterwards, the amount of light through each well is measured, and the amount of RNA present can be measured. After mapping the RNA snips to the DNA, one can tell how prevalent a gene may be. In a cancer cell, some genes are more prevalent than its normal value, meaning that the RNA level will be unnaturally high. This is relevant to cancer staging as patients will not have to go through surgery just to analyze a patient's cancer stage. Instead, a simpler test of extracting some cells in a human body reduces costs and also lowers the risks of complications from surgery. One can

also look at the amounts of RNA per strand across different stages, and look at which strands have the highest and lowest correlation between cancer stages.

RNA sequencing data is obtained from the TCGA database as well. Experiments are carried out with a batch of breast cancer RNAseq data. As each patient in TCGA database has a barcode, it is possible to link the patients in the RNA sequencing data with the patients from the clinical stage. After RNAseq data is matched to a clinical cancer data, numerical studies are conducted between the two data sets. Some genes, such as Progesterone Receptor (PGR), Estrogen Receptor (ESR1, ESR2), Epidermal Growth Factor Receptor (EGFR), and Human Epidermal Growth Factor Receptor (ErbB-2, HER2), are already known to have some relation with breast cancer's development [9-10]. As a starting point, these genes were the primary targets in our study. Initially not much variation was observed across different stages. However, when we looked at the gene's location on the human genome in the RNA sequencing data, after comparing lower stage frequencies to higher stage frequencies, the results looked encouraging. We decided to look at all 25,000 or so genes in the human body.

In this batch of data, for every patient there are three data matrices: the 'exon', the 'junction', and the 'gene.' The rows consist of the genes or RNA snips, while the columns contain statistics such as raw counts, RKPM (Reads per Kilobase Per Million sequenced), and median length normalized. The exon and the junction are less processed compared to the gene data frame. Also, because the RNA sequencing technology breaks apart the RNA strands, in the exon and the junction data frames many rows are too short to match up to a particular gene. I used a t-test to compare the frequency of the raw counts of RNA, and also looked at the reads per kilobase of exon model per million mapped reads. In the study, I realized that seventy patients

was not enough for accurate analysis, as I only had less than ten Stage IV patients for the RNAseq data. Hence, I went back to TCGA looking for more data.

As the batches of data are tested in different conditions, a batch correction is often necessary. As each RNAseq patient is matched with the clinical data from TCGA by barcode, I use a function available in R to complete an ANOVA (analysis of variance) adjustment for the values. Different cancer stages are used as references. Next a t-test (statistical hypothesis test) is done between the earlier stages (Stage I and Stage II) and the later stages (Stage III and Stage IV) of breast cancer. In the test, the lowest correlation of raw counts across the stages is looked for. In order to control the error rate, A Bonferroni correction is carried out on the results. The Bonferroni correction is a simple method used to eliminate or lower error due to statistical chance. Furthermore, t-tests are conducted on the following parameters estimated of the genes for each patient: raw counts, raw counts normalized, and raw counts scaled. Detailed analysis of these results is presented in Results section.

Results obtained from analyzing the RNAseq data are encouraging. Although further research is needed, the techniques in our approach should have applications in furthering cancer staging research. RNA sequencing data is potentially a solution in solving the problem and filling in the gap of incomplete cancer data.

## 3. Results

For the clinical data, multiple data tests were run to determine the accuracy of TCGA-published clinical cancer stages versus the derived stages by TNM staging definition. Six different methods of finding cancer stages were applied and compared. In the first test, a data tree function was used. The cancer stages as defined by the T, N, and M stages were computed

and compared with the TCGA-published clinical cancer stages. Because the computed stages are strictly by TNM system definition, these computed stages are considered as the most accurate measurements and compared against the next five tests. In the second test, stages were computed by a random assignment of equally weighted stage probabilities to each patient. In the third test, unknown stages were simply given by a naïve replacement with the most common stage within the data set. In the fourth test, similar to the third test, unknown stages were assigned by a naïve replacement with the most common stage determined by a national source. In the fifth test, weighted distribution of cancer stages within the data set was used to randomly assign patients' new stages. In the sixth test, similar to the fifth test, weighted distribution of cancer stages calculated from a national source was used to randomly assign patients' new stages. All the stage results in these tests were compared with the computed cancer stages by the data tree function. Cohen's Kappa is used to measure the agreement between each tested result and the data tree function result. In statistics, Cohen's Kappa coefficient measures the agreement between two data sets and takes out so-called "chance agreement". For example, if the agreement appears high simply due to randomness, Cohen's Kappa is able to adjust that and reduces the contribution from chance.

My findings are listed in tables in the Illustrations section. Of the fifteen cancer types that provided clinical cancer stages (some clinical cancer data didn't provide cancer staging information at all), three cancer types, ovarian, stomach, and uterus cancers, didn't provide T, N, or M staging information. Therefore, I could only experiment with twelve cancer types in making my data tree and cancer staging randomness tests. Breast cancer is used as an example to show what is computed in Table 1. The agreement is relatively high, at 91.7%, for the TCGA-published stages versus the computed stages by a data tree function. This indicates that the breast

cancer clinical data is reasonably complete and accurate. The weighted Kappa and unweighted Kappa were both high, at .868 and .781 respectively, meaning that the correlation is not simply due to randomness. In other cases, such as bladder cancer, the accuracy for the TCGA-published stages is relatively low. For many bladder cancer patients, the T, N, or M information is missing. This means that either the TNM staging was not tested, or tested but was not recorded, or was simply left out. In these cases when the agreement rate is low, it indicates that the data is incomplete and TCGA-published cancer stages cannot be treated with high accuracy.

In the other five tests, agreement is generally low to the computed data tree function. This shows that the cancer stages cannot be simply guessed based on probability, as different cancer stages have different characteristics.

For the RNA sequencing data, three tests were run on the data. For each patient, three statistics are used to analyze the RNA sequencing data: the raw counts per gene, the scaled estimates of raw counts per gene in each patient, and the raw counts normalized of each gene per patient. Three t-tests are conducted, one for each type of data, with Stage I and Stage II compared against Stage III and Stage IV. In the tests, the correlation of genes with the stages is calculated. A Bonferroni correction is done in order to eliminate errors due to randomness. In a Bonferroni correction, the p-values are adjusted by the number of comparisons, where p-value measures the statistical significance. This correction leads to a more focused group of genes to look at as compared to looking at the p-values of the results. For example, the raw counts normalized data column only have thirteen genes with Bonferroni values less than one, so the other genes may not be significant enough to be looked at for breast cancer research. The least correlated genes across the stages probably mean that that gene's expression volume either increases or decreases dramatically as the cancer develops. This implies that genes are related to

the cancer's development. These identified genes found in our three tests are displayed in Illustration sections below.

## 4. Illustrations

Table 1 shows the results from six tests on the TCGA clinical cancer data of breast cancer. Table 2 shows the results from the same tests for bladder cancer. Table 3 shows the results from four tests on the TCGA clinical cancer data of cervix cancer. Table 4 shows the same tests for colon cancer. Table 5 shows the same tests for head and neck squamous cell cancer. Table 6 shows the same tests for kidney renal clear cell cancer. Table 7 shows the same tests for kidney renal papillary cell cancer. Table 8 shows the same tests for liver cancer. Only data from TCGA is used for producing these results. No external source is used to improve the accuracy of the results.

| BRCA (849 Patients) | Accuracy | Weighted Kappa | Unweighted Kappa |
|---|---|---|---|
| 1. TCGA-published | 0.917 | 0.868 | 0.781 |
| 2. 25% all | 0.261 | 0.002 | 0.005 |
| 3. Most common national | 0.586 | 0.000 | 0.000 |
| 4. Most common internal | 0.180 | 0.000 | 0.000 |
| 5. By external distribution | 0.427 | 0.004 | -0.003 |
| 6. By internal distribution | 0.297 | -0.004 | -0.012 |

Table 1: The results of running multiple tests on the clinical cancer staging for Breast Cancer. All staging results are compared with the stages derived by Data Tree Function method from TNM definition system.

| BLCA (65 Patients) | Accuracy | Weighted | Unweighted |
|---|---|---|---|
| 1. TCGA-published | 0.645 | 0.528 | 0.288 |
| 2. 25% all | 0.256 | -0.029 | -0.007 |
| 3. Most common national | 0.370 | 0.000 | 0.000 |
| 4. Most common internal | 0.000 | 0.000 | 0.000 |
| 5. By external distribution | 0.357 | -0.137 | -0.135 |
| 6. By internal distribution | 0.139 | 0.031 | 0.004 |

Table 2: Bladder Cancer results. All staging results are compared with the stages derived by Data Tree Function method from TNM definition system.

| CESC 12 Patients | Accuracy | Weighted | Unweighted |
|---|---|---|---|
| 1. TCGA-published | 0.750 | 0.660 | 0.576 |
| 2. 25% all | 0.182 | 0.074 | -0.010 |
| 3. Most common internal | 0.727 | 0.000 | 0.000 |
| 4. By internal distribution | 0.400 | -0.250 | -0.304 |

Table 3: Cervix Cancer Results. All staging results are compared with the stages derived by Data Tree Function method from TNM definition system.

| COAD 423 Patients | Accuracy | Weighted | Unweighted |
|---|---|---|---|
| 1. TCGA-published | 0.867 | 0.751 | 0.823 |
| 2. 25% all | 0.277 | -0.010 | 0.034 |
| 3. Most common internal | 0.417 | 0.000 | 0.000 |
| 4. By internal distribution | 0.272 | -0.019 | -0.032 |

Table 4: Colon Cancer Results. All staging results are compared with the stages derived by Data Tree Function method from TNM definition system.

| HNSC -502 Patients | Accuracy | Weighted | Unweighted |
|---|---|---|---|
| 1. TCGA-published | 0.737 | 0.617 | 0.591 |
| 2. 25% all | 0.262 | 0.027 | 0.055 |
| 4. Most common internal | 0.618 | 0.000 | 0.000 |
| 6. By internal distribution | 0.475 | -0.059 | -0.028 |

Table 5: Head and Neck Squamous Cell Cancer. All staging results are compared with the stages derived by Data Tree Function method from TNM definition system.

| KIRC -502 Patients | Accuracy | Weighted | Unweighted |
|---|---|---|---|
| 1. TCGA-published | 0.663 | 0.317 | 0.583 |
| 2. 25% all | 0.226 | -0.031 | -0.023 |
| 3. Most common internal | 0.368 | 0.000 | 0.000 |
| 4. By internal distribution | 0.244 | -0.043 | -0.035 |

Table 6: Kidney Renal Clear Cell Cancer Results. All staging results are compared with the stages derived by Data Tree Function method from TNM definition system.

| KIRP -84 Patients | Accuracy | Weighted | Unweighted |
|---|---|---|---|
| 1. TCGA-published | 0.512 | 0.187 | 0.408 |
| 2. 25% all | 0.206 | 0.810 | -0.020 |
| 3. Most common internal | 0.588 | 0.000 | 0.000 |
| 4. By internal distribution | 0.300 | -0.207 | -0.327 |

Table 7: Kidney Renal Papillary Cell Cancer Results. All staging results are compared with the stages derived by Data Tree Function method from TNM definition system.

| LIHC -53 Patients | Accuracy | Weighted | Unweighted |
|---|---|---|---|
| 1. TCGA-published | 0.727 | 0.560 | 0.637 |
| 2. 25% all | 0.273 | 0.147 | 0.065 |
| 3. Most common internal | 0.485 | 0.000 | 0.000 |
| 4. By internal distribution | 0.438 | -0.152 | 0.046 |

Table 8: Liver Cancer. All staging results are compared with the stages derived by Data Tree Function method from TNM definition system.

Tables 9-11 show the results obtained from conducting t-tests on the RNA sequencing data. Table 9 shows the raw counts normalized for a gene compared between lower stages (Stage I and Stage II) and higher stages (Stage III and Stage IV). Table 10 shows the raw counts scaled estimate for each gene compared between lower stages (Stage I and Stage II) and higher stages (Stage III and Stage IV). Table 11 shows the raw counts for each gene compared between lower stages (Stage I and Stage II) and higher stages (Stage III and Stage IV). The genes listed in the tables achieved the lowest p-values or the highest significance in the test. Following the t-test, the identified fifteen most relevant genes are shown in the order of significance in each table.

| | row.names | gene_id raw counts norm | t.test | bonferroni |
|---|---|---|---|---|
| 1 | 14703 | RBBP8\|5932 | 1.953229e-08 | 0.000394142 |
| 2 | 15758 | SEPT10\|151011 | 3.631884e-08 | 0.0007328778 |
| 3 | 20326 | ZNF660\|285349 | 1.351671e-06 | 0.02727537 |
| 4 | 5444 | EHHADH\|1962 | 2.254037e-06 | 0.04548422 |
| 5 | 18107 | TMC3\|342125 | 6.473242e-06 | 0.1306235 |
| 6 | 17321 | ST6GALNAC5\|81849 | 1.792886e-05 | 0.3617865 |
| 7 | 4240 | CRIM1\|51232 | 1.89824e-05 | 0.3830459 |
| 8 | 18340 | TMEM84\|283673 | 3.126003e-05 | 0.6307961 |
| 9 | 20327 | ZNF662\|389114 | 3.213864e-05 | 0.6485256 |
| 10 | 18970 | TUSC3\|7991 | 3.328041e-05 | 0.6715655 |
| 11 | 19513 | WDR35\|57539 | 3.572338e-05 | 0.7208621 |
| 12 | 19985 | ZNF167\|55888 | 3.811759e-05 | 0.7691749 |
| 13 | 12171 | NTRK3\|4916 | 4.087733e-05 | 0.8248636 |
| 14 | 13692 | PNMAL1\|55228 | 5.195462e-05 | 1 |
| 15 | 11885 | NIPSNAP3B\|55335 | 6.283171e-05 | 1 |

Table 9: Raw Counts Normalized of Genes –least correlated. There are 404 genes with p-values below 0.05.

| | row.names | gene id scaled estimate | t.test | bonferroni |
|---|---|---|---|---|
| 1 | 146 | ACAD11\|84129 | 3.738221e-06 | 0.07672699 |
| 2 | 2681 | C7orf41\|222166 | 7.344661e-06 | 0.1507492 |
| 3 | 734 | ANKRD57\|65124 | 3.724748e-05 | 0.7645046 |
| 4 | 1464 | BCO2\|83875 | 4.608242e-05 | 0.9458416 |
| 5 | 1179 | ATL1\|51062 | 5.672853e-05 | 1 |
| 6 | 37 | A4GNT\|51146 | 6.11229e-05 | 1 |
| 7 | 2365 | C22orf23\|84645 | 7.469377e-05 | 1 |
| 8 | 3395 | CDC14B\|8555 | 8.176817e-05 | 1 |
| 9 | 2628 | C6orf192\|116843 | 0.0001010976 | 1 |
| 10 | 3132 | CCDC27\|148870 | 0.0001064962 | 1 |
| 11 | 309 | ADAMTS5\|11096 | 0.0001484724 | 1 |
| 12 | 366 | ADORA1\|134 | 0.000185699 | 1 |
| 13 | 3234 | CCL28\|56477 | 0.0001954212 | 1 |
| 14 | 3072 | CCDC109B\|55013 | 0.0001984308 | 1 |
| 15 | 819 | APBB1\|322 | 0.000199244 | 1 |

Table 10: Raw counts scaled estimates of each gene. There are 462 genes that have p-values below 0.05

| | row.names | gene id raw counts | t.test | bonferroni |
|---|---|---|---|---|
| 1 | 14703 | RBBP8\|5932 | 4.048424e-08 | 0.0008169314 |
| 2 | 20326 | ZNF660\|285349 | 1.187114e-06 | 0.02395478 |
| 3 | 18970 | TUSC3\|7991 | 2.952124e-06 | 0.0595709 |
| 4 | 15758 | SEPT10\|151011 | 3.282039e-06 | 0.06622827 |
| 5 | 20327 | ZNF662\|389114 | 1.037953e-05 | 0.2094484 |
| 6 | 18107 | TMC3\|342125 | 1.419508e-05 | 0.2864425 |
| 7 | 5021 | DMD\|1756 | 1.794848e-05 | 0.3621823 |
| 8 | 2681 | C7orf41\|222166 | 2.901215e-05 | 0.5854362 |
| 9 | 11885 | NIPSNAP3B\|55335 | 3.308038e-05 | 0.6675289 |
| 10 | 17321 | ST6GALNAC5\|81849 | 3.551831e-05 | 0.7167241 |
| 11 | 19985 | ZNF167\|55888 | 3.662499e-05 | 0.7390557 |
| 12 | 14695 | RAX\|30062 | 4.862599e-05 | 0.9812239 |
| 13 | 4989 | DLEC1\|9940 | 5.563888e-05 | 1 |
| 14 | 5444 | EHHADH\|1962 | 5.815263e-05 | 1 |
| 15 | 4240 | CRIM1\|51232 | 6.655326e-05 | 1 |

Table 11: Raw counts of genes, after the t-test was completed. There are 333 genes with p values below 0.05.

## 5. Discussion

A comparison between the clinical cancer stages as determined by TCGA data base and the actual cancer stages as determined by the taxonomy definition shows a noticeable discrepancy. One cannot simply guess the patient's cancer stage. As Cohen's Kappa is essentially zero in many tested cases, it is shown that the agreement is low between the guessed values and actual stages, after factoring out chance as a cause of correlation.

RNA sequencing can provide an alternative way to determine cancer stages if the T, N, and M stages of cancer patients cannot be determined. After comparing the RNA sequencing data of the lower stage patients with the higher stage patients in a statistical t-test across three different measurements for breast cancer, we found that certain genes have significant relationships to cancer stages. For example, in our study, gene RBBP8 has the highest change

between stages of breast cancer. Researchers have proven that gene RBBP8 is linked to the development of breast cancer [9-10]. Our study confirms such linkage.

The important genes can be found by the Bonferroni values. There are thirteen genes that have Bonferroni values below one for the raw counts normalized in Table 9. There are twelve genes that have Bonferroni values below one for the raw counts in Table 11. There are four genes that have Bonferroni values below one for the scaled estimate in Table 10. More specifically, for the scaled estimate section the four genes are ACAD11, C7orf41, ANKRD57, and BCO2. In biomedical research, these genes can be explored to see if there is indeed any relation between the prevalence of these genes and the development of cancer. The scaled estimate section is probably the best place to look for those genes, as the number of candidates appears to be the fewest for breast cancer. The four genes listed above can be the ones to study first.

This project takes a unique statistical approach to the large world of cancer research. We first looked at the current taxonomy of determining cancer stages. We then examined the clinical data from TCGA to quantify the accuracy of the published cancer stages. Six different methods are tested to demonstrate that cancer stages cannot simply be guessed based on probability distribution. We concluded that the missing data problem cannot be solved using statistical extrapolation approaches.

The real significance of this research is that we found a way to link genes to cancer stages by using RNA sequencing data. Using breast cancer as an example, we found that the RNA sequencing data can be analyzed statistically to identify a list of genes which have the strongest correlation to cancer stages. In our view, this is an effective way to solve the problem of missing clinical cancer stage information, without going through invasive surgical procedures.

RNA sequencing data is usually complete as the process of collecting RNA sequencing data is fairly secure. However, as the method is relatively new, accuracy of the RNA sequencing data depends on the sequencing depth. Sometimes there are improper associations of lab results with genes. This poses challenges for our research. In addition, statistical methods have other limitations. These are topics for further studies.

## 6. Conclusions and Further Work

When the TCGA data is more complete, data tree function shows that the TCGA-published cancer stages have 91.7% accuracy, e.g. in the breast cancer section. When the clinical data is incomplete, e.g. in ovarian cancer, stomach cancer, and uterus cancer, the TCGA-published cancer stages are less accurate. By using six different methods, we have shown that cancer stages cannot be simply assigned with statistical extrapolation when data is incomplete.

We have explored a new approach to use the RNA sequencing data to determine cancer stages. In this research work, we are able to find certain genes linked with cancer stage development. These genes display the largest changes across the stages. For example, gene RBBP8, has been associated with the development of breast cancer. Our statistical method not only identifies a single gene, but a list of genes ranked by their correlation to cancer stages. This opens up possible research directions of other genes linked to breast cancer. To understand the fundamental relationship between the genes and the cancer stages, further bio-medical research is required.

In conclusion, when clinical cancer data is incomplete, cancer staging cannot be solved simply using probability distributions. Some published clinical cancer stages are derived from such probability distributions and therefore not accurate. One method we plan to explore in the

future is to reconstruct the data using low-rank matrix completion method [11]. Applying t-test on RNA sequencing data has led to promising results in solving the incomplete data problem. The list of genes identified through this research can potentially open new areas for research of the relationship between genes and development of cancer. The results obtained are practically useful. We have demonstrated that using RNA sequencing to determine cancer stages is a viable approach.

## 7. Acknowledgement

This research is part of the MIT PRIMES (Program for Research in Mathematics, Engineering and Science) for high school students. I have learned a great deal through my first scientific research project. I'd like to thank Prof. Pavel Etingoff, Dr. Slava Gerovitch, and Dr. Tanya Khovanova of MIT for organizing the PRIMES program. I am especially grateful to my mentors Dr. Gil Alterovitz and Dr. Jeremy Warner for guiding me through my research. I also would like to thank my parents for driving me to MIT and always being supportive.

## 8. References

1. *The Cancer Genome Atlas Data Portal*. 2012. 20 July 2012. <https://tcga-data.nci.nih.gov/tcga/tcgaHome2.jsp>.

2. *CLC Bio: User Manual*. n.d. 20 July 2012.

3. National Research Council of the National Academies. *Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease*. Washington, DC: The National Academies Press, 2011.

4. Edge, S.B., et al. *AJCC Cancer Staging Manual 7th Edition*. New York: Springer, 2009.

5. Joseph G. Ibrahim, Haitao Chu, and Ming-Hui Chen. "Missing Data in Clinical Studies: Issues and Methods." *Journal of Clinical Oncology* (2012): 3297-3303.

6. Mardis, Elaine R. "Next-Generation DNA Sequencing Methods." *Annual Reviews* 9 (2008): 387-402.

7. *How-to/RNASeq analysis*. 18 August 2012. Website. 20 August 2012. <http://seqanswers.com/wiki/How-to/RNASeq_analysishttp://seqanswers.com/wiki/How-to/RNASeq_analysis>.

8. Cai G, Li H, Lu Y, Huang X, Lee J, Müller P, Ji Y, Liang S. "Accuracy of RNA-seq and its dependence on Sequencing depth." *Bioinformatics* (2012). <http://www.rna-seqblog.com/data-analysis/expression-tools/accuracy-of-rna-seq-and-its-dependence-on-sequencing-depth/>.

9. Kate D. Sutherland1, Jane E. Visvader1, David Y.H. Choong2, Eleanor Y.M. Sum1, Geoffrey J. Lindeman1, Ian G. Campbell2,†,*. "Mutational analysis of the LMO4 gene, encoding a BRCA1-interacting protein, in breast carcinomas." *International Journal on Cancer* (2003, Volume 1, Issue 107): 155-158.

10. Matthew Meyerson, Stacey Gabriel and Gad Getz. "Advances in understanding cancer genomes through second-generation sequencing." *Nature Reviews* 11 (October 2010): 685-696.

11. Candes, Emmanuel J and Benjamin Recht. "Exact Matrix Completion via Convex Optimization." *Communications of the ACM* 55.6 (2012): 111-119.