DATA ANALYTICS FOR CANCER SURVIVAL AND TREATMENT

Sara He
Associate Consultant
Dell EMC
sara.he@dell.com

Diego Gallo
Data Scientist – Big Data
Dell EMC
diego.gallo@dell.com

William Schneider
Data Scientist
Dell EMC
william.schneider@dell.com

Wei Lin
Chief Data Scientist – Big Data
Dell EMC
w.lin@dell.com

Table of Contents

Disclaimer: The views, processes or methodologies published in this article are those of the authors. They do not necessarily reflect Dell EMC's views, processes or methodologies.

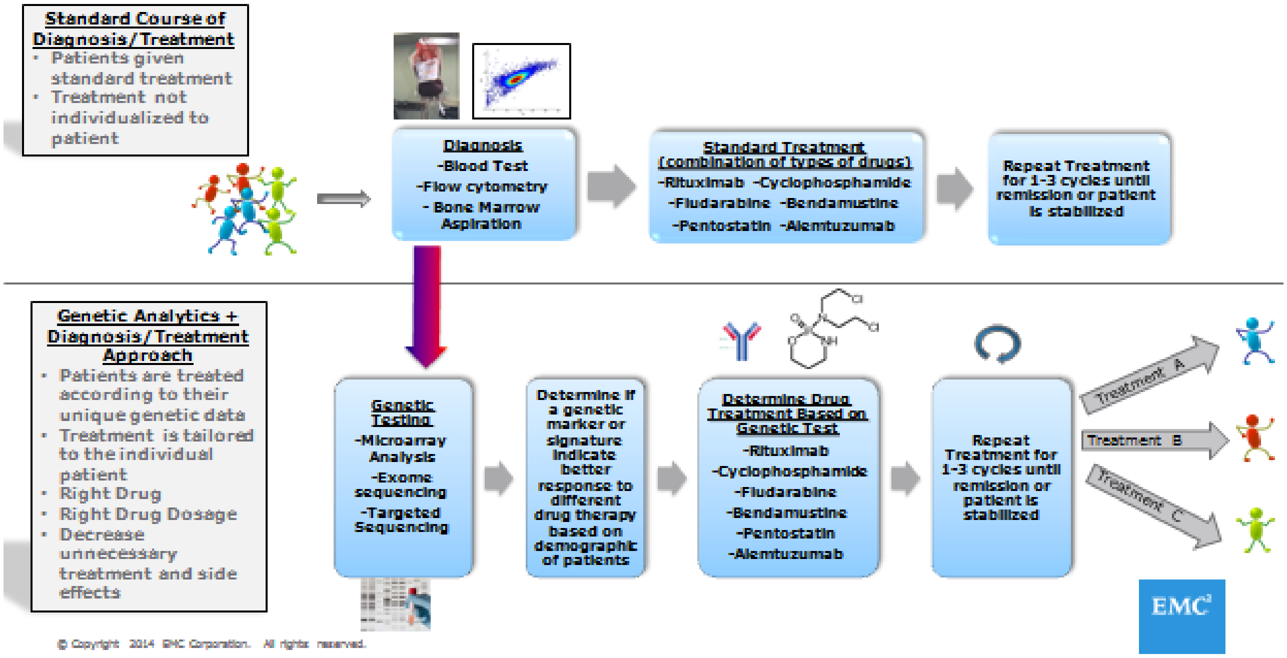
Abstract

Cancer is responsible for one in four deaths in the United States per year with an estimated average treatment cost of 102,395. Convergence of analytics, patient genetic information, and patient clinical information can offer a solution to decreases the cost of providing cancer treatment to patients by tailoring genetic information to clinical patient parameters such as drug therapies and outcome. By tailoring treatments to patients and screening potential patients with a higher risk for developing cancer based on genetic information, the healthcare system can save around 20% on treatment cost by avoiding unnecessary drug therapies [28]. Advancements have been made in improving cancer treatments and developing breakthrough drugs, however, many obstacles remain when it comes to treating this elusive disease. Cancer originates from a series of mutations unique to each individual which contributes to the complexity of disease treatment. It is no longer solely about determining the location of cancer but about deeper analyzing the inner workings of a tumor to extrapolate the specific genetic and molecular causes of the cancer. Genetic determinants in the formation of cancer have been the topic of research for decades and have led to the rise of personalized and precision medicine. Personalized medicine aim to treat a cancer based on the genetic makeup of the disease versus the current approach of using broad spectrum chemotherapy and radiation which can lead to unnecessary side effects and increase treatment cost. This approach requires genetic profiling in order to understand the genetic intricacies of a cancer cell and which genes are expressed at a higher or lower level than normal. The decreased cost and improved methods of genetic profiling offer a more thorough way of analyzing the genetic makeup of cancer to determine why patients respond differently to treatments and have a different prognosis. The explosive amount of genetic data generation creates the opportunity for big data analytics to translate huge amounts of raw genetic cancer data combined with clinical data into knowledge that can be applied to help future patients. By analyzing the genetic makeup of cancer patients, common genetic marker(s) can be analyzed to determine the likelihood of a certain outcome, help to predict a response to a given type of treatment, or identify potential therapeutic compounds.

Introduction

In the United States, an estimated 1,658,370 new cases of patients diagnosed with cancer contributed to 589,430 deaths in 2015 and the national expenditure for treatment and care is predicted to reach $156 billion in 2020 [29]. Clearly, the personal and financial costs are large as cancer is a complex disease with no known cure. Though progress has been made in improving cancer treatments and developing breakthrough drugs, a lot of work remains. Previous strategies solely utilized standardized broad spectrum therapies such as chemotherapy, radiation, and stem cell transplant [1]. These strategies, however, can expose patients to harmful drug side effects and treatment failures. Advanced therapeutic developments include targeted therapies which use drugs that more precisely attack cancer cells without damaging normal cells and immunotherapy which utilizes the immune system to fight the disease. Recently, a rise in genetic profiling in cancer treatment is paving a new way of fighting the disease.

Cancer is a disease characterized by uncontrolled cell growth that spreads and invades other areas of the body. Normally cells are formed when the body needs them and die when they are damaged or grow old but somewhere within the cell cycle this normal process breaks down and cells divide uncontrollably creating cancer cells. The disruption of the normal cell cycle process in cancer formation is attributed to the mutation of genes, defined DNA sequences that act as a blueprint for cells. The development of cancer is linked to a series of genetic mutations unique to each individual, which contributes to the complexity of treating the disease. Patients with the same type of cancer can have completely different outcomes and treatment reactions due to specific gene mutations. This realization requires treatment recommendations which are unique to each patient, even with the same diagnosis. This leads to the concept of ‘personalized’ or ‘precision’ medicine.



Standard Course of Diagnosis/Treatment

- Patients given standard treatment
- Treatment not individualized to patient

Diagnosis
-Blood Test
-Flow cytometry
- Bone Marrow Aspiration

Standard Treatment (combination of types of drugs)
-Rituximab -Cyclophosphamide
-Fludarabine -Bendamustine
-Pentostatin -Alemtuzumab

Repeat Treatment for 1-3 cycles until remission or patient is stabilized

Genetic Analytics + Diagnosis/Treatment Approach

- Patients are treated according to their unique genetic data
- Treatment is tailored to the individual patient
- Right Drug
- Right Drug Dosage
- Decrease unnecessary treatment and side effects

Genetic Testing
-Microarray Analysis
-Exome sequencing
-Targeted Sequencing

Determine if a genetic marker or signature indicate better response to different drug therapy based on demographic of patients

Determine Drug Treatment Based on Genetic Test
-Rituximab
-Cyclophosphamide
-Fludarabine
-Bendamustine
-Pentostatin
-Alemtuzumab

Repeat Treatment for 1-3 cycles until remission or patient is stabilized

Figure 1: Illustration of the standard conventional method of cancer treatment versus the approach using genetic profiling for personalized treatment. With the treatment utilizing genetic profiling a, unique individualized treatment can be determined.

Gene expression profiling is a way to measure thousands of genes simultaneously, in this case the genes of the tumor or cancer cells. Gene expression is the process where information from a gene is used to synthesize a functional gene product such as a protein or mRNA. Methods of gene expression profiling include microarray analysis which has the advantage of being able to analyze genes on a large scale. In microarray analysis, fluorescent probes bind to different gene products, mRNA in this case, and their fluorescent intensity is quantified in order to compare the expression levels of different genes [30]. In the case of cancer, genetic expression profiling is used to determine if certain genes are at a higher or lower activity level than a normal cell. This method has become more cost-effective and accurate over time. With the increased ability to gather this massive source of data for each patient diagnosed, the opportunity for big data analytics to discover unique genetic insights in cancer has increased significantly. With a larger pool of patients and their genetic profiles, common genetic markers can be determined with higher certainties, degrees of aggressiveness of the cancer can be understood, and the likelihood of favorable responses to commercially available drugs can be determined. Genetic profiling also has the potential to identify new genetic targets for scientists and the pharmaceutical industry.

The approach of utilizing genetics to fine-tune cancer treatment is based on identifying genetic biomarkers of a cancer cell. A biomarker is a genetic characteristic that is measured and used as an indicator of change in biological processes such as a mutation in a cancer cell. [2] These are used in cancer care to determine a person's predisposition to developing a type of cancer, for disease type identification, and to determine if a type of cancer will respond to a targeted therapy. This differs from standard chemotherapy because they act on specific molecules associated with cancer versus a broad attack on rapidly growing normal and cancerous cells with chemotherapy. An example of a target therapy is *Tratuzumab* (Herceptin) that targets the protein human epidermal growth factor receptor 2

(HER2), which is expressed at high levels on breast and ovarian cancer cells. Many melanomas express a mutant form of B-raf (*BRAF*) that drives disease progression and is treated by *Venmurafenib* (Zelboraf), which targets this mutant form of *BRAF* [3]. For these treatments to be effective these mutations have to be observed on the patient, otherwise the therapies would have no target and be ineffective. Genetic analysis to match patients to treatment has also been a subject of clinical trials. A phase 1 clinical trials program was initiated at MD Anderson Cancer Center in 2012 to match genetic aberrations to targeted drug agents. The patient's cancers were analyzed for specific biomarkers and mutations and matched to specific treatments accordingly. Patients with matched therapy were associated with a higher overall response rate (27% vs 5%) when compared to patients who did not have matched therapy, which suggests therapies matched to a patients' cancer signature has benefits[4]. The biological mechanisms for why cancer develops in a given patient can be determined by identifying specific genes that significantly differ between patient outcome groups, and available drugs that target those genes can be recommended. The genetic signature of a health outcome can also be used as a reference for future patients to determine whether or not a prognosis is favorable. From there an appropriate aggressive or mild treatment regimen can be applied, and patients with a poor prognosis can be treated more thoroughly while patients with a favorable prognosis can avoid unnecessary treatment.

The goal of this EMC Knowledge Sharing article is twofold:

1. To demonstrate the value of some of the current analytical approaches to improve treatment outcomes of a cancer based on clinical and genetic information
2. To illustrate a potential IT and business infrastructure which, when implemented, streamlines both the research development process and diagnosis / treatment decisions.

With this in place, the ROI of the organization is greatly affected. The integration of personal genetic information with clinical data through big data analytics offers an opportunity to revolutionize the healthcare paradigm by predicting patient outcome, improving drug efficacy, and establishing opportunities to introduce new healthcare economic models. A publicly available cancer data set is used in this work to illustrate two models: a prediction model using linear regression analysis, and a supervised classification model which is used for prediction and treatment recommendation.

Methodology

The data is collected from patients suffering from Chronic Lymphocytic Leukemia, a slow progressing blood cancer that contributes to 25% of leukemia cases. Genetic and clinical data for 267 patients were obtained from the International Cancer Gene Consortium (ICGC). The goal is to determine if a specific group of genes, called a gene signature, contributes to the clinical outcome of the patient (partial remission, complete remission, progression, stable, and death) and how likely a favorable outcome results for a patient given a type of treatment. One limitation is that this data set does not contain detail on the type of treatment experienced by the patients.

We first present some introductory data profiling results for the patients in this data set, and perform some exploration using cluster analyses and Principal Component Analysis (PCA). While clinical data is straightforward to understand, genetic data consists of a massive amount of variables, of which only a few may contribute to the response to treatment. The first demonstration of modeling techniques consists of a classification model derived from [24] which predicts likelihood of a positive response from the treatment as well as identification of specific genes which are important in the clinical outcome. The second model is a regression model of the genetic data performed in order to see if the patient outcome can be predicted based on genetic expression. One limitation in this analysis is the lack of specific treatment data. A future study would be to incorporate treatment information into either of the models presented.

Data Profiling

For Chronic Lymphocytic Leukemia, patients are diagnosed at stage 1 for first symptomatic state, stage 2 for intermediate disease state, and stage 3 for advance disease state. The 267 patients were usually diagnosed at Stage 1 of the disease (84% of patients), since that is the first symptomatic stage. After treatment, the majority of patients became stable, with 5% experiencing partial remission and 14% complete remission, totaling 75% of the patients with positive outcomes. Of the negative outcomes, 7% underwent relapse and 18% experienced progression of the disease. The breakout of outcome by diagnosis stage is shown below, where stage 1 patient's exhibit similar outcomes to the whole group.

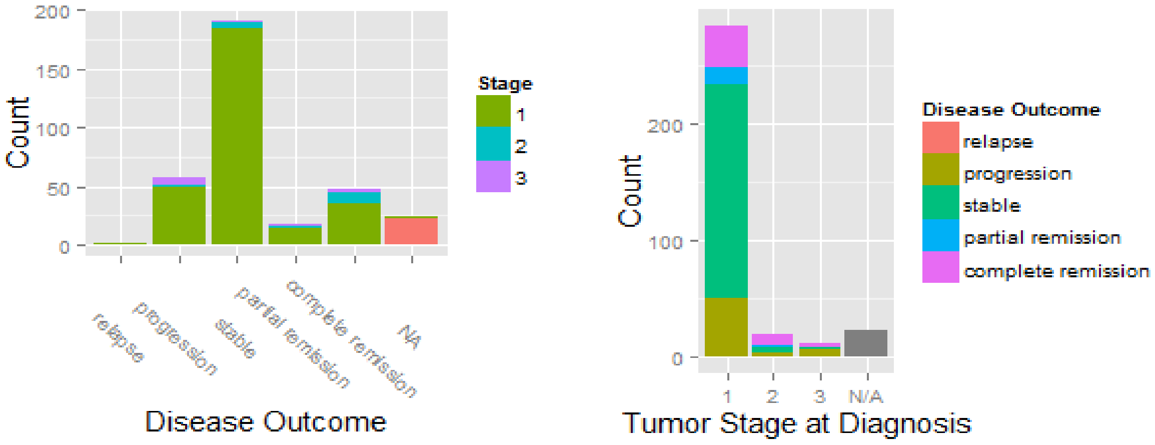


Figure 2: Disease outcomes in the data set. Left: outcomes broken up by stage (1, early; 2, middle; 3, late) at diagnosis. Right: Tumor stage broken up by disease outcome.

Patient survival time, in the next figures, shows a median survival time around 7 years, exhibited by a mix of patients of various ages. Longer survival times over 10 years are experienced by patients between 50 and 65. This could potentially be due to disease progression or due to the patient's age. Chronic lymphocytic leukemia is diagnosed mostly during stage 1 which is the least advanced stage.

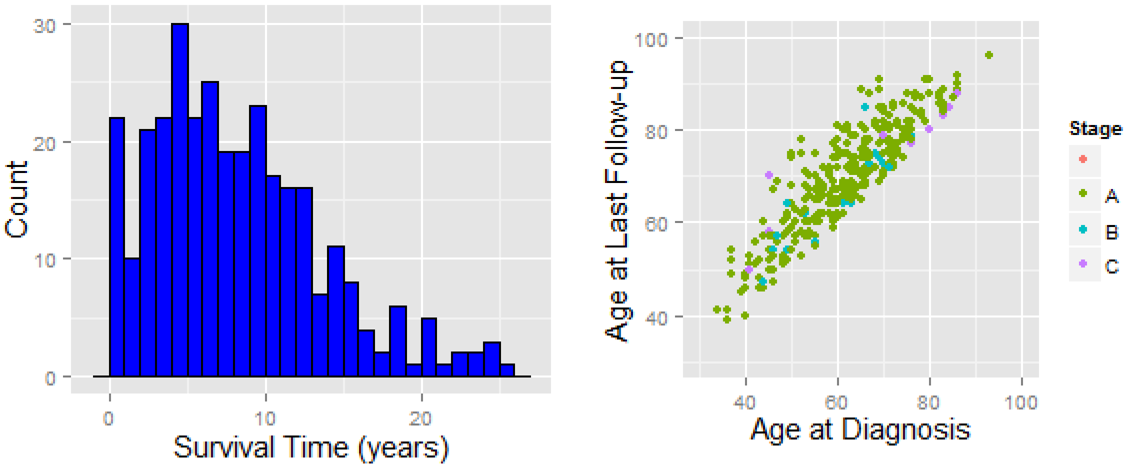


Figure 3: Patient survival time related to age. Left: histogram of the survival times of the data set. Right: age at diagnosis combined with age at last follow-up. The difference between these two ages is within a year less than the survival time.

The patient base is 40% female and disproportionally more men under complete remission as well as progression. When diagnosed at stage 1 it is observed the majority of the patients are stable due to the fact that patients who are stable generally are not treated unless they exhibit negative symptoms of the disease. Among the patients diagnosed at stage 2, we can observe a relatively higher number of patients under complete remission, which might be explained due to more extreme treatment options since the cancer is already at a later stage. Finally, among the patients diagnosed at stage 3, we can observe a relatively higher number of patients with progression, which might be explained by the fact that the cancer was discovered too late to treat it effectively. Chronic lymphocytic leukemia is diagnosed more frequently between the ages of 50 and 70.

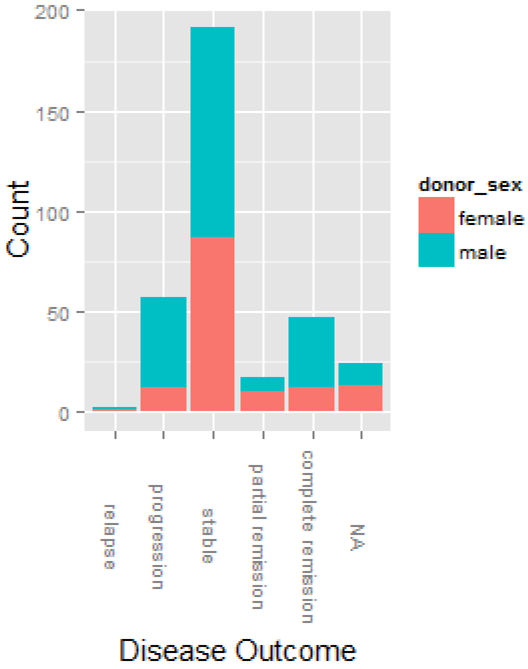


Figure 4: Gender of patients by disease outcome.

In addition to the clinical data, we explore the gene expression microarray data for these patients, about 18,000 genes. Due to the number of attributes, it's useful to focus on the genes with the greatest amount of variation in the data set. It turns out that the genes are well represented, as seen in the next figure. The scale of the expression levels ranges from 0 to 12, with only non-zero standard deviations. The peak in mean occurs around 4 and 0.1 for standard deviation for 0.1. Plotted to the right is the ratio of the standard deviation to the mean, showing a good representation of the spread of each variable.

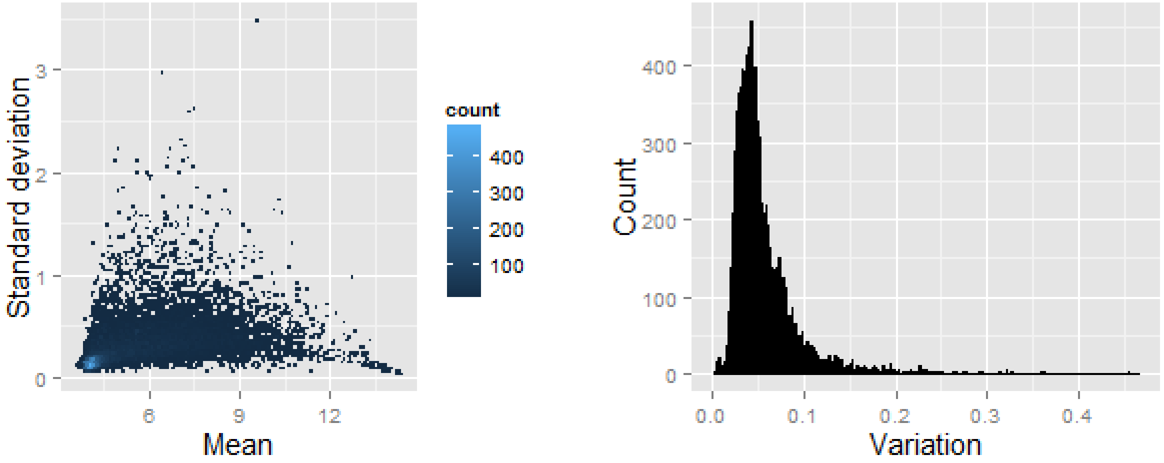


Figure 5: Sample variable variability. Left: Color plot of the density of genes with a given mean and standard deviation. Right: Histogram of the sample coefficient of variation.

While the spread of each gene expression individually may be significant, the data set as a whole is explored by a principal component analysis next. The first principal component on the left appears to have significant skewed weight toward the zero, indicating that a small percentage of the genes are significant to understanding the data set. The set of genes which have a weight larger than 0.01 is about 2,600, or about 14% of the genes, and similarly with the second component.

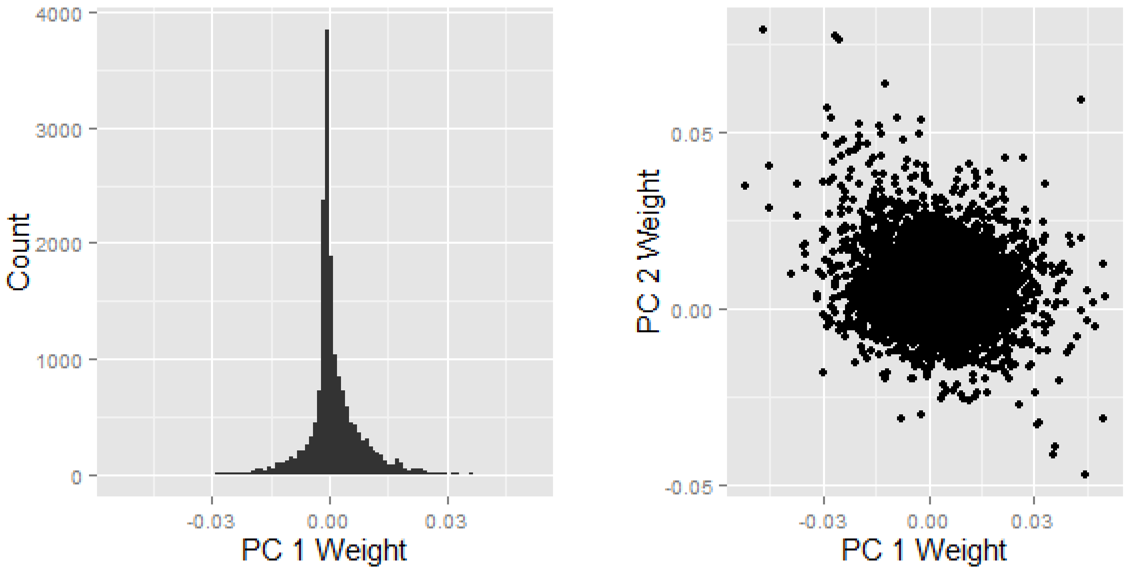


Figure 6: Weights of the first two principal components in the data set. Left: histogram of the first PC. Right: Scatter plot of the first two PCs.

Another common visualization in the field is a two-way clustered heat map of a subset of the genes. Chosen here is the set which is significant with the outcome. About 2,100 genes were selected, shown in Figure 7.

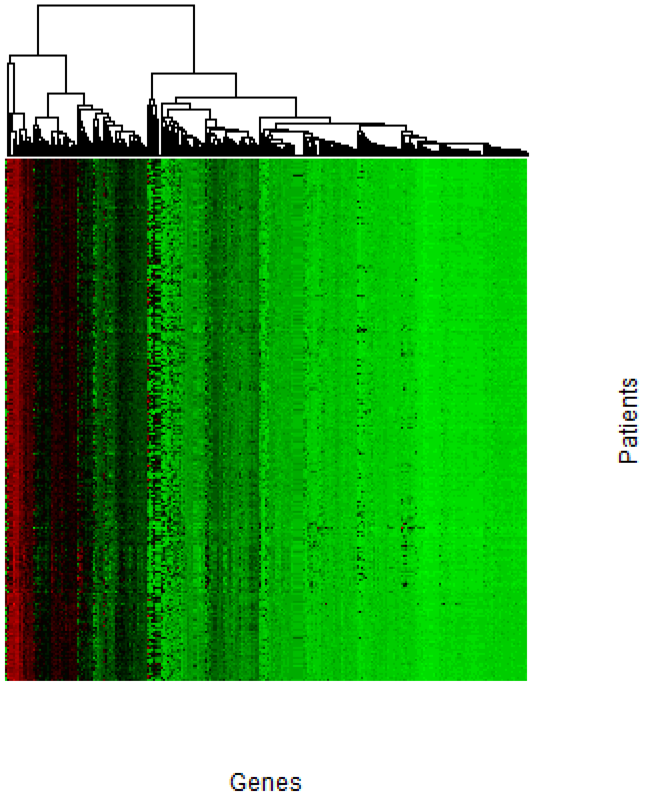


Figure 7: Clustered heat map of the genes which have a 0.05 significance when fit against the outcome. Green is a higher expression level, red is lower.

Classification Model

In this section, an approach is performed closely related to [24] which uses a classification mechanism based solely on genetic data, specifically, a set of the most significant genes. The motivation for the model is that a small subset of gene expression levels correlates significantly with the outcome. The authors define a binary outcome, so in this analysis a positive outcome is defined as stable or better. In Figure 7, the genes which are significant individually compared with the outcome are retained from the original set and clustered. To correct for multiple-testing errors, the Bonferroni method results in 38 significant genes. The cluster result is shown in the next figure, where, visually, a few gene pairs can be seen.

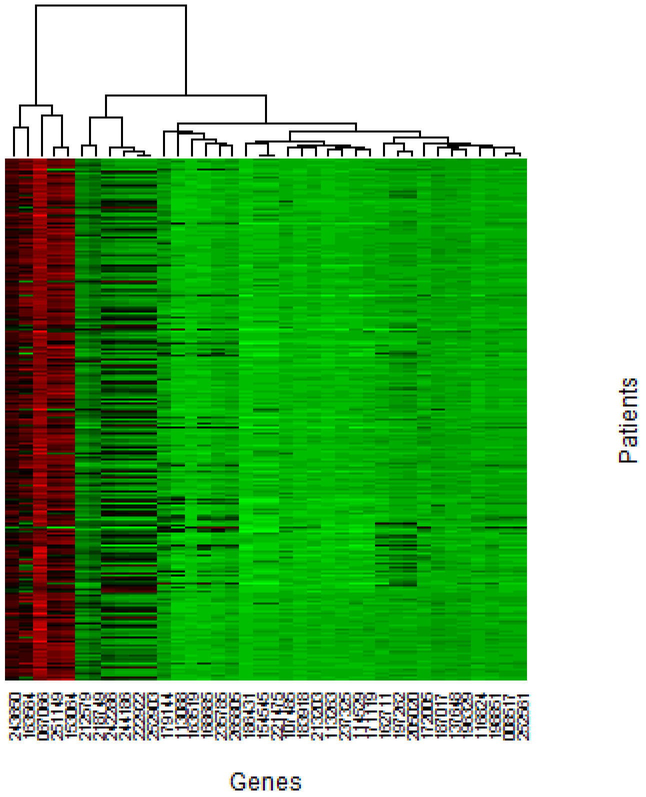


Figure 8: Heat map of the genes which are significant to the 0.05 level with the Bonferoni correction.

In order to train the model on the outcome, k=1 cross-validation is performed by correlating each patient's genes with the average of the genes for patients who were positive and negative for the outcome. The larger correlation determines the prediction, and the accuracy is taken to be the percent of correct predictions over the 267 patients.

The gene set which is used for the above trained model is successively increased until the accuracy degrades. The order of the increase is given by the correlation of that gene with the outcome. In this data set, 79% accuracy was obtained using 25 genes. The following figure illustrates the progression of the accuracy as more genes were added to the variable set.

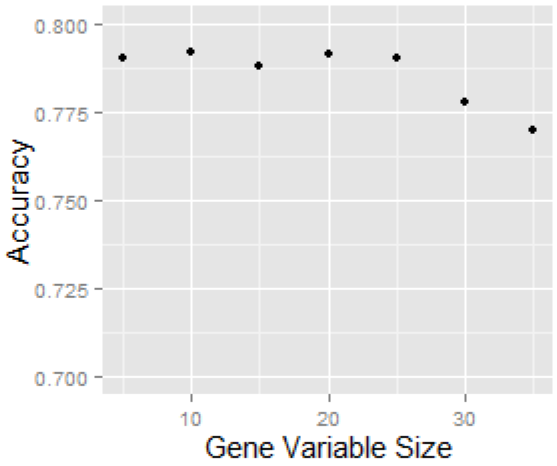


Figure 9: Accuracy of the training method using an increasing number of gene variables.

To gain a sense of the expression levels of the resulting model, a heat map is shown in the next figure ordered by the leave-one-out sample's correlation with the average positive outcome. To the right, the correlation value is plotted with the colors indicating negative or positive outcome. A few observations: correlations are generally high for the patient base, indicating that the expression levels don't vary too much across the patients. However, there is a sharp decrease at the bottom with the lower correlation patients. Furthermore, there are two genes which cluster especially well on the left, indicating they have similar predictive power.

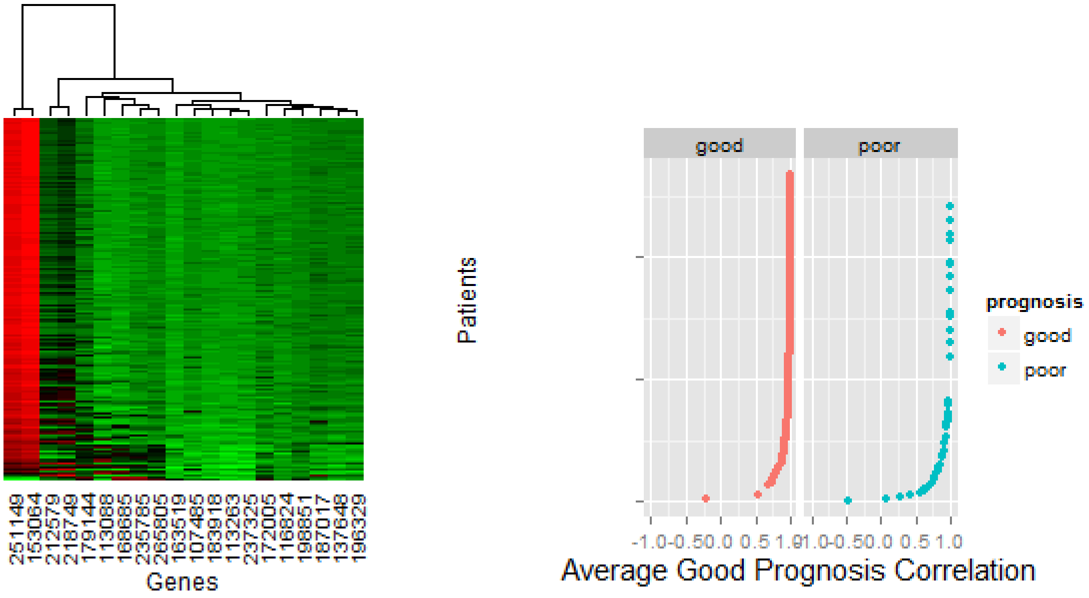
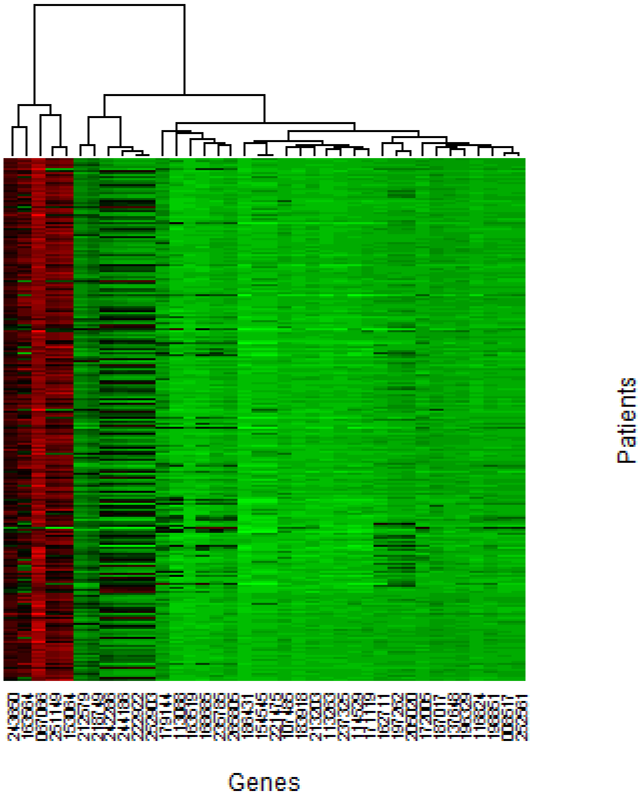


Figure 10: Heat map of the final gene set chosen for the model ordered by the leave-one-out correlation with the good prognosis gene profile. To the right, the correlation values are plotted separately for either of the actual patient outcomes. Outcome binary 0 is associated with a poor prognosis while binary 1 is associated with a good prognosis.

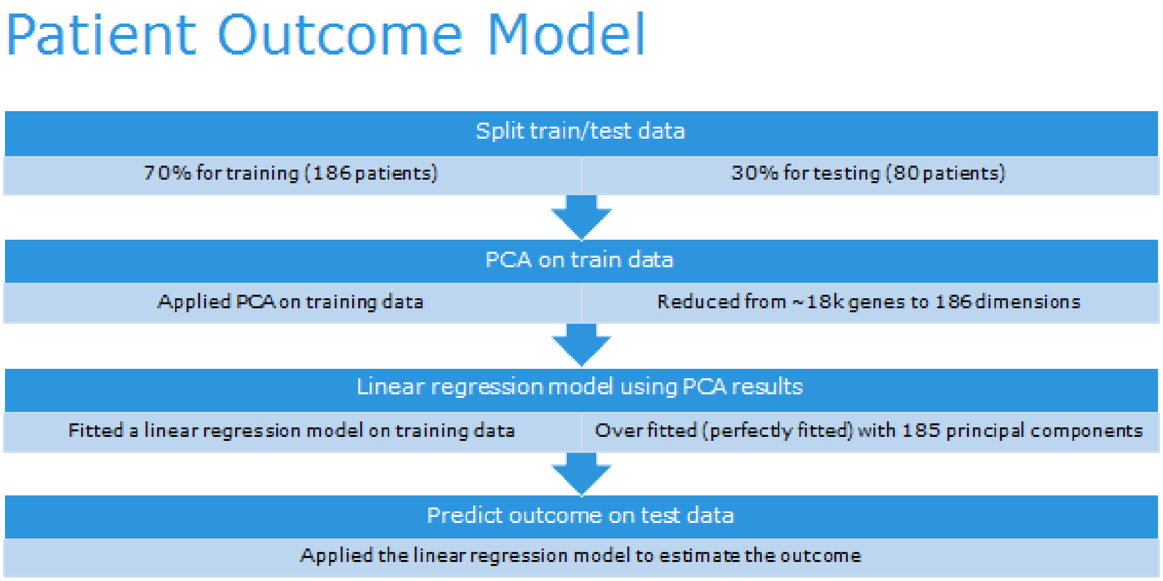
There are two key results of this model. First, one can predict the outcome of the patient with the knowledge of the genetic data. Second, the model identifies important genes for further research. In a later section, the gene set identified here is shown to have relevance in current research for Chronic lymphocytic leukemia (CLL). Furthermore, as discussed below, even if different patients are predicted to have a positive outcome, they may have different genes which are dominantly expressed, suggesting different treatment options.

Some improvements which require further research are to include the clinical data into the classification model to improve accuracy, overlay the correlation with the outcome on the cluster results in Figure 8 to understand if the model should be done within clusters, and use different significance tests to identify different gene sets. We note that if the treatment data was available for this data set, the model should be applied to each treatment type, providing more customization.



Linear Regression Model

In this section PCA analysis was performed on train data to reduce the dimensionality of the problem, and secondly a linear regression model was used to predict patient outcome. Patient outcome was labeled in a continuous scale: 1 for complete remission, 2 for partial remission, 3 for stable, 4 for progression, and 5 for relapse. The 18,000 genes were reduced to 186 principal components and patient outcome was fitted in a linear regression model to predict outcome.



Patient Outcome Model

Split train/test data

70% for training (186 patients)

30% for testing (80 patients)

PCA on train data

Applied PCA on training data

Reduced from ~18k genes to 186 dimensions

Linear regression model using PCA results

Fitted a linear regression model on training data

Over fitted (perfectly fitted) with 185 principal components

Predict outcome on test data

Applied the linear regression model to estimate the outcome

Figure 11: Steps for linear regression model

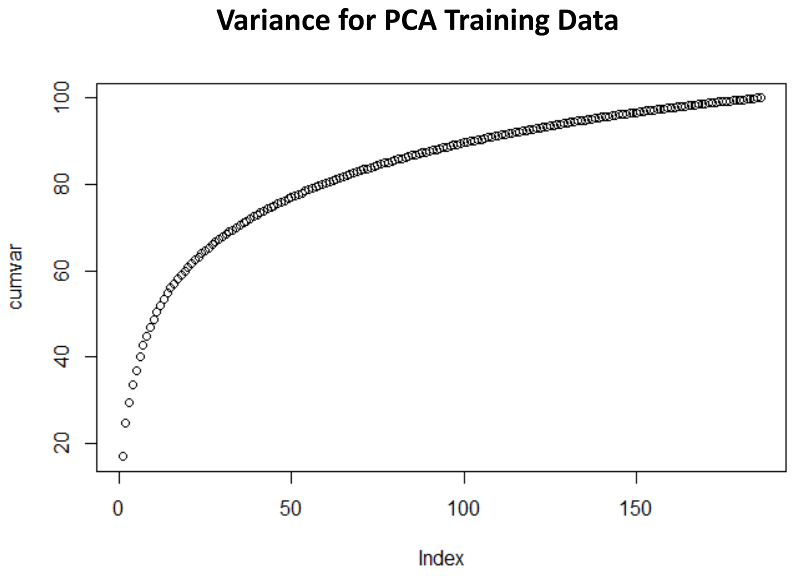


Figure 12: 95% of genetic variation can be represented among 186 principal components

Based on the regression model, patient outcome is correctly predicted 55% percent of the time. If the patient outcome were randomly assigned it would be correctly predicted for only 20% of the patients, which indicates some predictive power in the model for patient outcome. Potentially, in the future, treatment data can be used in order to add predictive power to the model.

Analytical Model

The models shown above have great potential when applied to this data set. However, even more value and accuracy can be gained when they are operationalized over an ever-increasing data set. As new patients are diagnosed and microarray data collected, the model can be fit on an ongoing basis, identifying slightly different sets of genes based on the increasing population in total or as applied to clusters of the population. An organization's research team would desire a system which captures data, applies the necessary transformations, and applies the model on a regular basis using different parameters. The team can refine the model, incorporating cutting-edge research to create a set of analytic data which can be consumed by physicians to augment their own knowledge base.

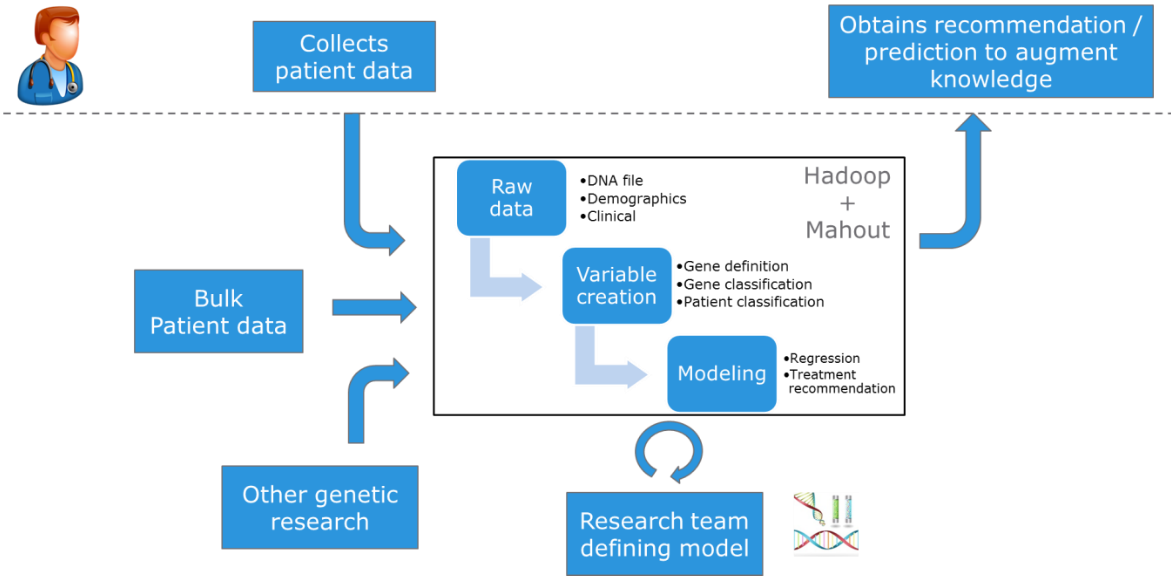


Figure 13: Potential IT and business architecture to operationalize the diagnosis, treatment, and research process flows.

Figure 13 illustrates a potential architecture to implement this system. At the top, the physician and lab technicians diagnose and collect the microarray data from each patient, as well as any clinical information. An ingestion layer supplements that with bulk patient data, including clinical and genetic data, as well as other genetic research findings from the literature. Once integrated, a big data platform such as Hadoop and Mahout could execute the three-step process of loading the raw data, variable creation, and model fitting. The research team at the bottom would be responsible for defining these variables, parameters for the model, executing ad-hoc analysis over the model results, and trying out new models for production. The physician would access the predictions through the patient's chart and make decisions based on their own knowledge augmented with the model results.

Biological Role of Identified Significant Genes

Genes between the patients with different clinical outcomes were analyzed to observe any genes with significant differing expression levels. Twenty-five genes were identified and their biological functions were analyzed to determine its potential role in chronic lymphocytic leukemia's development. Many of the genes identified play a role in cell regulation and immune system activity and are implicated in multiple types of cancers and immune dysregulation disorders. Table 1 lists all of the significant genes identified with a functional description of some genes that have an important role in the immune system.

Table 1: Identified Significant Genes

IL7R	AL109767.1	GZMK	MAL	IL32
CD2	GIMAP7	SH2D1A	GATA3	GIMAP1
LTK	TRAT1	AP000563.2	BANK1	DMD-AS1
Metazoa_SRP	TMPRSS4	GIMAP5	RP3-340B19.5	PYHIN1
CD3E	ESPN	SNORA40	MTND5P5	SP100

IL7R [5] - *IL7R* codes for a receptor protein found on the cell surface that plays a critical role in the development of lymphocytes. Studies suggest that IL7R functions to block apoptosis and low expression of *IL7R* are associated with a better prognosis for acute leukemia. However, high expression levels of the gene have been associated with lung cancer, renal carcinoma, colorectal cancer, breast cancer, chronic lymphocytic leukemia, and acute myeloid leukemia. Patients have a potential positive response to cancers exhibiting this mutation if treated with *Ruxolitinib*.

CD2 [6] - CD2 molecule is a cell adhesion molecule on the surface of T cells and natural killer cells and contributes to their activation. T cells play a critical role in adaptive cell-mediated immunity and create signaling molecules, cytokines, which stimulate immune system activity. Natural killer (NK) cells play an essential role in innate immunity which defends the host from infection. Both T cells and NK cells are essential for defending the body against pathogens and any compromise to the function of these cells exposes the host to an array of immune disorders. Increased CD2 expression is a predictor for even free survival in children with acute lymphoblastic leukemia after chemotherapy. The role of this gene suggests that patients with a poor prognosis have a lower expression of CD2 than patients with a favorable outcome.

LTK [8] - IL 2 is an inducible T-cell tyrosine kinase receptor which helps to control pathways leading to cell growth and differentiation (NCBI). The *LTK* gene is reported to be overexpressed in human leukemia. This suggests patients with higher than normal expression levels have a poor prognosis. *ALK* and *LTK* share a high degree of similarity and a study shows cells with mutant *LTK* were susceptible to *Crizotinib*.

CD3E [26] - The CD3E molecule forms the T cell receptor-CD3 complex which plays a role in coupling antigen recognition to intracellular signal transduction pathways. T cell receptor plays an essential role in immune system activation. Mutations in *CD3E* are linked to immunodeficiency and type 1 diabetes in women.

GIMAP7 [9] - This gene makes up the GTPase IMAP family member 7 and is overexpressed in certain types of cancers. Mutations that lead to a continual expression of GTPase proteins have been implicated in cancer development. GTPase proteins are critical to many biological roles and play a role in signal transduction, cell differentiation, and protein and vesicle translocation through membranes.

TMPRSS4 [10, 23] - *TMPRSS4* is a serine-4 transmembrane protease and involved in the mechanism of peptide bond hydrolysis. Protease dysregulation has been implicated in different diseases such as cancer, arthritis, cardiovascular disease, and neurodegeneration. Overexpression of this gene has been reported in ovarian, lung, pancreatic, cervical, liver, and breast cancer. Knockout of this gene in mice results in development defects in mice. Increase in this gene expression is associated with poor outcome and the gene has been proposed as a future potential therapeutic target. In 2012 a 2-hydroxydiarylamide derivative was published to have an inhibitory activity against *TMPRSS4* and could have potential promising anti-cancer activity. A patent was filed in 2012 for this compound.

ESPN [11] - This gene codes for acting-bundling protein and plays a role in regulating dimension, organization, and signaling of sensory transduction. Mutations of this gene are associated with deafness.

GZMK [12] - *GZMK* is granzyme K protein and related to serine proteases from cytoplasmic granules of cytotoxic lymphocytes (CTLs) which contribute their activation. It is a member of the granzyme family which promotes rapid cell death of viruses and cancer cells, and the blockade of granzymes has been implicated as a method of immune escape for cancer. CTLs are important in the immune system for the defense of viruses and specific tumor cells due to their ability to recognize, bind, and lyse specific target cells [7].

SH2D1A [13] - *SH2D1A* plays a role in bidirectional B and T cell stimulation. Mutations in *SH2D1A* is associated with lymphoproliferative syndrome X-linked type 1 and Duncan disease, a rare immunodeficiency categorized by extreme susceptibility to Epstein-Barr virus infection, with symptoms including malignant lymphoma and severe mononucleosis.

GIMAP5 [14] - Gene is part of the GTPase IMAP family members which plays a role in multiple cell functions. Loss of this gene function causes spontaneous death of T cells. The gene is expressed in lymphocytes and hematopoietic system (blood system). If implicated in leukemia it suggests a down regulation of this gene.

MAL [27] - Gene codes for T cell differentiation protein and is important for T cell signal transduction. Past studies demonstrate the expression of this gene in esophageal cancer suppresses invasion and tumorigenicity. *MAL* expression is reduced esophageal cancer, and loss of function of this gene suggests a poor prognosis in patients.

GATA3 [15] - *GATA* binding protein 3-transcription factor maintains mammary luminal epithelial cell function. *GATA3* expression is lost in breast cancer and correlates to a poor prognosis in patients and expression of this gene is shown to suppress breast cancer metastasis and alter tumor microenvironment.

IL32 [16, 17] - Interleukin 32 promotes breast and gastric cancer invasion. IL32 is a cytokine excreted from activated T cells and induces expression of TNF alpha and IL8 which are associated with inflammatory processes and breast cancer and gastric cancer proliferation via angiogenesis. An IL32 inhibitor could be a potential drug target for preventing the spread of cancer.

GIMAP1 [18] - GTPase IMAP family member. Gene is critical for mature B and T cell lymphocytes survival.

DMD-AS1 [19] - DMD antisense RNA. Low expression of this gene is implicated in Duchene muscular dystrophy.

SP100 [20] - SP100 nuclear antigen. Implicated as a tumor suppressor gene.

Standard Chronic Lymphocytic Leukemia Diagnosis and Treatment Protocol

Chronic lymphocytic leukemia (CLL) is diagnosed through a variety of tests. Common tests include blood tests where a complete blood count is taken and if the white blood cell count is higher than normal, the person may have CLL. Blood tests are a common diagnosis method but a bone marrow biopsy is sometimes performed before treatment in order to determine the stage of the cancer. After a blood test, in order to confirm a CLL diagnosis flow cytometry is perform where fluorescent dyes are added to the cells to analyze the surface proteins. CLL contain distinctive cell surface protein markers on the outside of their cells which is called an immune phenotype and differentiates CLL from other types of leukemia.

When a diagnosis of CLL is confirmed treatment is given according to stage and age. Patients who are considered stable and low /intermediate risk are generally observed and not treated unless they exhibit negative side effects of the disease. For patients at a more advanced stage of CLL the standard treatment therapies are similar. The most common drugs used for CLL consist of combinations of *fludarabine*, *cyclophosphamide*, *rituximab*, *pentostatin*, *chlorambucil, alemtuzumab, ofatumumab, ibrutinib,* and *idelalisib*. Treatment is administered as a combination of 2-3 of these drugs in monthly cycles depending on if the patient has been untreated or has relapsed. The most utilized combinational drug protocol is a FCR therapy which consists of *fludarabine, cyclophosphamide, and rituximab*. *Fludarabine* and cyclophosphamide are both drugs used for the treatment of blood cancer like leukemia by interfering with DNA synthesis and affect rapidly growing cells and normal resting cells. Rituximab is a monoclonal antibody that inhibits CD20, a protein that is expressed on the surface of B cells, and is used to treat conditions characterized by dysfunctional or excessive B cells such as in leukemia. Similar to Rituximab, *Alemtuzumab* is also a monoclonal antibody which targets cell expression CD52, a protein expressed on mature lymphocytes surface, for destruction.

Treatment Personalization Based on Identified Significant Genes

Regression analysis was done to identify genes that were significantly different between the CLL patients with a favorable and poor prognosis. Twenty-five genes were identified as having a significant difference between the prognostic groups and of those 25 genes, 16 were identified as having important biological roles in the development of different types of cancers and immunological disorders. Mutations in these genes contribute to a range of disease such as lung cancer, chronic lymphocytic leukemia, deafness, Duchene muscular dystrophy, breast cancer, and esophageal cancer. Three of the genes, *IL7R, LTK, and TMPRSS4*, were discovered to have drugs that treat patients with those mutations. The drugs that target those genes are not part of the standard therapies for chronic lymphocytic leukemia.

IL7R is a receptor protein expressed on the cell surface and important for lymphocyte development. *IL7R* is overexpressed in multiple types of cancer including chronic lymphocytic leukemia. Patients with mutations in this gene have exhibited a positive clinical response if treated with *Ruxolitinib*, a drug primarily used to treat myelofibrosis, a disease of the bone marrow [5]. The *LTK* gene helps control pathways leading to cell growth and differentiation and is reported to be overexpressed in human leukemias. *LTK* shares a high degree of structural homology with *ALK* which has been implicated in the growth of non-small cell lung cancer (NSCLC). Patients with an *ALK* mutation have shown a positive response to the drug *Crizotinib* which is an *ALK* inhibitor[8]. Since *ALK* and *LTK* share a high degree of similarity, mutant *LTK* cells were shown to respond to *Crizotinib* and this could suggest an off-label use for the drug. Mutation in TMPRSS4 where it is overexpressed has been associated with multiple cancers such as lung, cervical, breast, ovarian, and pancreatic. Increase in *TMPRSS4* expression is connected with a poor clinical outcome and has been studied as a potential therapeutic target. In 2012 a 2-hydroxydiarylamide derivative was published to exhibit inhibitory activities against *TMPRSS4* with promising anti-cancer properties. In 2012 a patent was filed for this compound to develop it into an accessible drug for patients [23]. Based off the genetic profiling done on the patients in the use case, those that express mutations in *IL7R, LTK, and TMPRSS4* could benefit from additional therapy with the drugs that target their specific mutations.

Return on Investment Potential

Using genetic profiling in cancer care creates an opportunity to benefit the economic aspects of hospitals and insurance companies through:

- Reducing the cost of treatment
- Identifying potential genetic targets for pharmaceutical companies
- Reducing patients' financial/personal burden
- Improving patient quality of life

Advances in medicine allow for better treatment and outcome of cancer patients. However, it adds to the increasing cost of healthcare in the United States. The yearly drug treatment cost for chronic lymphocytic leukemia in the United States ranges from 83,000-120,000 dollars with many patients requiring multiple years of treatment [31]. This does not include other health resources such as hospital equipment, medical staff, palliative care, and laboratory tests. Many hospitals are in a fragile financial state due to the increasing cost of providing 24/7 healthcare and increasing government funding shortfalls. It is estimated that 60% of hospitals lose money providing patient care and 30% of hospitals lose money overall [25]. The majority of patients diagnosed with CLL is above the age of 60 and is Medicaid-eligible. The actual cost of treatment is more than the reimbursement for Medicaid. In reality, more effective treatment can translate to a reduced revenue loss.

Table 2: Current Chronic Lymphocytic Leukemia Treatment Cost for one patient.

Type	Cost Per Year(USD) for 1 Patient
Drug Treatment only	83,000-120,000
Hospital Expenses	20,000
Hospital Staff	30,000

Genetic profiling can also identify genetic targets which can then be used by the pharmaceutical industry to develop a targeted therapy. For patients, precision medicine can decrease treatment duration, avoid unnecessary treatment/side effects, enable disease prevention/prediction of disease rather than reaction to it, and reduce their personal cost.

Conclusions

This article demonstrated the use of genomic microarray outcome data from Chronic Lymphocytic Leukemia patients for predicting patient outcome. Additionally, specific genes were identified that had significant differing expression levels between prognostic groups. These genes were identified along with their biological role in the mechanism of cancer development. Some of the genes identified were prominent in multiple types of cancers suggesting common mutational pathways in disease development. In addition, a few treatment options are commercially available or in stages of development to target some of the genes identified. These treatments are not listed as a form of standard therapy for Chronic Lymphocytic Leukemia patients and off-label use of the drugs can benefit patients with the genetic mutations. This relates back to the concept that not only the physical location of the cancer matters, but also the genetic mechanism for why it develops should be taken into consideration. Since cancer is not a one size fits all disease and each patients' cancer is influenced by different mutations, personalized treatment based on specific mutations that drive cancer growth for each patient has the ability to add a layer of precision in the oncology field. Genetically profiling a person's cancer with the utilization of big data methods to analyze the information at a large scale is quickly being recognized as a more effective way to treat the disease and paving a way to the future of cancer treatment.

References

1) "Treatment Types." Types of Cancer Treatment. American Cancer Society, n.d. Web. 29 Jan. 2016.

2) Cho, Sang-Hoon, Jongsu Jeon, and Seung Il Kim. "Personalized Medicine in Breast Cancer: A Systematic Review." J Breast Cancer Journal of Breast Cancer 15.3 (2012): 265. Web.

3) "Targeted Cancer Therapies." National Cancer Institute. National Cancer Institute, n.d. Web. 29 Jan. 2016.

4) Tsimberidou, A.-M., N. G. Iskander, D. S. Hong, J. J. Wheler, G. S. Falchook, S. Fu, S. Piha-Paul, A. Naing, F. Janku, R. Luthra, Y. Ye, S. Wen, D. Berry, and R. Kurzrock. "Personalized Medicine in a Phase I Clinical Trials Program: The MD Anderson Cancer Center Initiative." Clinical Cancer Research 18.22 (2012): 6373-383. Web.

5) Kim, Min Sung, Nak Gyun Chung, Myung Shin Kim, Nam Jin Yoo, and Sug Hyung Lee. "Somatic Mutation of IL7R Exon 6 in Acute Leukemias and Solid Cancers." Human Pathology 44.4 (2013): 551-55. Web.

6) Uckun, FM, PG Steinherz, H. Sather, M. Trigg, D. Arthur, D. Tubergen, P. Gaynon, and G. Reaman. "CD2 Antigen Expression on Leukemic Cells as a Predictor of Event-free Survival after Chemotherapy for T-lineage Acute Lymphoblastic Leukemia: A Children's Cancer Group Study." Bloodjournal (1996): n. pag. Print.

7) "Genes and Mapped Phenotypes." National Center for Biotechnology Information. U.S. National Library of Medicine, n.d. Web. 29 Jan. 2016.

8) Roll, J. Devon, and Gary W. Reuther. "ALK-Activating Homologous Mutations in LTK Induce Cellular Transformation." PLoS ONE 7.2 (2012): n. pag. Web.

9) "Expression of GIMAP7 in Cancer - Summary - The Human Protein Atlas." Expression of GIMAP7 in Cancer - Summary - The Human Protein Atlas. Human Protein Atlas, n.d. Web. 29 Jan. 2016.

10) Aberasturi, A. L De, and A. Calvo. "TMPRSS4: An Emerging Potential Therapeutic Target in Cancer." Br J Cancer British Journal of Cancer 112.1 (2014): 4-8. Web.

11) "ESPN Gene." Gene Cards. N.p., n.d. Web. 29 Jan. 2016.

12) "Genes and Mapped Phenotypes." National Center for Biotechnology Information. U.S. National Library of Medicine, n.d. Web. 29 Jan. 2016.

13) Booth, C., K. C. Gilmour, P. Veys, A. R. Gennery, M. A. Slatter, H. Chapel, P. T. Heath, C. G. Steward, O. Smith, A. O'meara, H. Kerrigan, N. Mahlaoui, M. Cavazzana-Calvo, A. Fischer, D. Moshous, S. Blanche, J. Pachlopnik Schmid, S. Latour, G. De Saint-Basile, M. Albert, G. Notheis, N. Rieber, B. Strahm, H. Ritterbusch, A. Lankester, N. G. Hartwig, I. Meyts, A. Plebani, A. Soresina, A. Finocchi, C. Pignata, E. Cirillo, S. Bonanomi, C. Peters, K. Kalwak, S. Pasic, P. Sedlacek, J. Jazbec, H. Kanegane, K. E. Nichols, I. C. Hanson, N. Kapoor, E. Haddad, M. Cowan, S. Choo, J. Smart, P. D. Arkwright, and H. B. Gaspar. "X-linked Lymphoproliferative Disease Due to SAP/SH2D1A Deficiency: A Multicenter Study on the Manifestations, Management and Outcome of the Disease." Blood 117.1 (2010): 53-62. Web.

14) Chen, Xi-Lin, Daniel Serrano, Marian Mayhue, Kasper Hoebe, Subburaj Ilangumaran, and Sheela Ramanathan. "GIMAP5 Deficiency Is Associated with Increased AKT Activity in T Lymphocytes." PLOS ONE PLoS ONE 10.10 (2015): n. pag. Web.

15) Chou, Jonathan, Jeffrey H. Lin, Audrey Brenot, Jung-Whan Kim, Sylvain Provot, and Zena Werb. "GATA3 Suppresses Metastasis and Modulates the Tumour Microenvironment by Regulating microRNA-29b Expression." Nature Cell Biology Nat Cell Biol 15.2 (2013): 201-13. Web.

16) Wang, Shouman, Feiyu Chen, and Lili Tang. "IL-32 Promotes Breast Cancer Cell Growth and Invasiveness." Oncology Letters Oncol Lett (2014): n. pag. Web.
17) Tsai, C.-Y., C.-S. Wang, M.-M. Tsai, H.-C. Chi, W.-L. Cheng, Y.-H. Tseng, C.-Y. Chen, C. D. Lin, J.-I. Wu, L.-H. Wang, and K.-H. Lin. "Interleukin-32 Increases Human Gastric Cancer Cell Invasion Associated with Tumor Progression and Metastasis." Clinical Cancer Research 20.9 (2014): 2276-288. Web.
18) Saunders, A., L. M. C. Webb, M. L. Janas, A. Hutchings, J. Pascall, C. Carter, N. Pugh, G. Morgan, M. Turner, and G. W. Butcher. "Putative GTPase GIMAP1 Is Critical for the Development of Mature B and T Lymphocytes." Blood 115.16 (2010): 3249-257. Web.
19) Blake, Derek J., Andrew Weir, Sarah E. Newey, and Kay E. Davies. "Function and Genetics of Dystrophin and Dystrophin-Related Proteins in Muscle." Physiol Rev Physiological Reviews 82.2 (2002): 291-329. Web.
20) Negorev, D. G., O. V. Vladimirova, A. V. Kossenkov, E. V. Nikonova, R. M. Demarest, A. J. Capobianco, M. K. Showe, F. J. Rauscher, L. C. Showe, and G. G. Maul. "Sp100 as a Potent Tumor Suppressor: Accelerated Senescence and Rapid Malignant Transformation of Human Fibroblasts through Modulation of an Embryonic Stem Cell Program." Cancer Research 70.23 (2010): 9991-10001. Web.
21) Grogg, Matthew W., and Yi Zheng. "Rho GTPase-Activating Proteins in Cancer." The Rho GTPases in Cancer (2009): 93-107. Web.
22) Roll, J. Devon, and Gary W. Reuther. "ALK-Activating Homologous Mutations in LTK Induce Cellular Transformation." PLoS ONE 7.2 (2012): n. pag. Web.
23) Kang, Sunghyun, Hye-Jin Min, Min-Seo Kang, Myung-Geun Jung, and Semi Kim. "Discovery of Novel 2-hydroxydiarylamide Derivatives as TMPRSS4 Inhibitors." Bioorganic & Medicinal Chemistry Letters 23.6 (2013): 1748-751. Web.
24) van 't Veer L, et. al., "Gene Expression Profiling Predicts Clinical Outcome Of Breast Cancer," Nature (2002).
25) "The Fragile State of Hospital Finances." American Hospital Association(n.d.): 1-9. Web. 1 Jan. 2016.
26) "CD3E Molecule." National Center for Biotechnology Information. U.S. National Library of Medicine, n.d. Web. 15 Feb. 2016.
27) Mimori, Koshi, Takeshi Shiraishi, Kohjiro Mashino, Hideto Sonoda, Keishi Yamashita, Keiji Yoshinaga, Takaaki Masuda, Tohru Utsunomiya, Miguel A. Alonso, Hiroshi Inoue, and Masaki Mori. "MAL Gene Expression in Esophageal Cancer Suppresses Motility, Invasion and Tumorigenicity and Enhances Apoptosis through the Fas Pathway." Oncogene 22.22 (2003): 3463-471. Web.
28) "The Case for Personalized Medicine." PMC.com : The Case for Personalized Medicine. N.p., n.d. Web. 16 Feb. 2016.
29) "What Are the Key Statistics for Chronic Lymphocytic Leukemia?" American Cancer Society. N.p., n.d. Web. 16 Feb. 2016.
30) "DNA Microarray Technology." National Human Genome Research Institute. N.p., n.d. Web. 16 Feb. 2016.
31) Shanafelt, Tait D., Heidi Gunderson, and Timothy G. Call. "Commentary: Chronic Lymphocytic Leukemia—The Price of Progress." The Oncologist 15.6 (2010): 601–602. PMC. Web. 16 Feb. 2016.