Journal of Big Data and Artificial Intelligence

WWW.JBDAI.ORG

ISSN: 2692-7977

JBDAI Vol. 3, No. 1, 2025 DOI: 10.54116/jbdai.v3i1.49

Applications of Analytics in Disease Prediction Types

Cheng-Yi Tsai
Penn State University
czt5442@psu.edu

Satish Mahadevan Srinivasan
Penn State University
sus64@psu.edu

Abhishek Tripathi
The College of New Jersey
tripatha@tcnj.edu

Abstract

Predictive analytics has immense potential in disease-type classifications. The key is to identify the set of genetic and clinical variables that can serve as predictors for disease classification purposes. However, the predictive and the prescriptive models both suffer from high dimensionality of these predictors. Therefore, it becomes important to identify a subset of these genetic and clinical variables that can be used for disease-type predictions. Earlier studies identified a subset of 978 landmark genes that can infer the expression values of the remaining gene in the human genome with $\sim$81% accuracy. This study focused on understanding if there is any significant difference in the characteristics of the landmark and non-landmark genes. Several experiments were performed on diseased tissues that were classified across race, ethnicity, and disease types, with an objective to identify the number of differentially expressed genes within the landmark and non-landmark gene sets. Statistically, there was no conclusive evidence to support the hypothesis that there is a significant difference in the number of differentially expressed genes across the landmark and non-landmark gene sets.

Keywords *L1000 dataset analysis, landmark genes, non-landmark genes, differentially expressed genes, cancer tissues, RNA-Seq data.*

1. Introduction

Cancer is a disease that is characterized by uncontrolled cell growth. It is a heterogeneous disease that consists of many different subtypes. Early diagnosis of cancer type has become a priority for many cancer researchers because it can facilitate the subsequent genetic and clinical management of the patients. Cancer research is mainly focused on primarily identifying the cancer type and, secondarily, on classifying patients into high- or low-risk groups. These two tasks involve analyzing large datasets and building predictive and prescriptive models that can decode the interaction between both the clinical and genetic variables. Therefore, biomedical and bioinformatics research teams have started to rely heavily on machine learning (ML) and artificial intelligence (AI) techniques. These techniques have been proven to model the progression and treatment of cancerous conditions. In addition, these techniques have the ability to detect key features from complex datasets.

Even though ML methods can help in detecting cancer types and help us understand the progression of the disease, an appropriate level of validation is still needed for these methods to be used in clinical practice. Studies in the past (Duncan et al. 2008; Liang et al. 2015; Danaee et al. 2017; Bailey et al. 2018; Huang et al. 2018; Saltz et al. 2018; Way et al. 2019) applied various ML techniques that focused on the impact of the genetic variables (genes) on the clinical responses. These techniques also have applications in cancer research. By using ML techniques, scientists can screen early stages of cancer progression by analyzing the genetic variables to find its nature before the symptoms show up. At the same time, with the advent of new technologies in the field of medicine, large amounts of cancer data have been collected and are available to the biomedical research community. Within the data lie complex patterns that can be mined efficiently by using the current state of the ML techniques. However, an accurate prediction of a disease outcome is one of the most interesting and challenging tasks for the biomedical research community. As a result, ML methods have become a popular tool for medical researchers. These techniques can discover and identify patterns and relationships between them from complex datasets, while they are able to effectively predict future outcomes of a cancer type (Kourou et al. 2015).

Advances in the area of personalized medicine are significantly fueled by advances in ML and AI techniques. Personalized medicine is important because it has increasingly been applied with success in clinical trials. Early detection of cancer increases the survival rate. Determining which genes contribute to decreased survival likelihood in cancer patients can provide clinically relevant biomarkers. This study was an effort in developing data analytics techniques to assess the RNA-sequencing (RNA-Seq) data from the National Cancer Institute (NCI) database (Tomczak et al. 2015). Early diagnosis of cancer, including cancer susceptibility, recurrence, and survival prediction, can be efficiently performed by using various ML and AI techniques. The availability of larger datasets that contain gene expression profiling captured over the period of time can significantly improve our ability for prognosis in cancer patients (Clayman et al. 2020a, 2020b).

The gene expression profiling measures which genes are expressing in a cell at any given moment. Gene expression profiling measures the messenger RNA (mRNA) levels, showing the pattern of genes expressed by a cell at the transcription level (Fielden and Zacharewski 2001). Gene expression profiling is used by a variety of researchers in the area of biomedical engineering, from molecular biologists to environmental toxicologists. This technology can provide accurate information on gene expression for the entire human genome. Different techniques are used to determine gene expressions, including DNA microarrays and sequencing technologies, for example, the RNA-Seq (Hurd and Nelson 2009).

The genome is a collection of biological information, but it is unable to disclose that information on its own. The initial product of the genome expression is the transcriptome. RNA-Seq is a state-of-the-art approach that can determine the quantity and sequences of RNA in a sample by using the next-generation sequencing. It analyzes the transcriptome, which indicates which of the genes in our DNA are turned on and off, and to what extent, and corresponding gene expression levels. RNA-Seq possesses the capability to measure the expression values of the genes across the transcriptome. RNA-Seq also promises to discover de novo transcriptome with high specificity in different species. It is a relatively new method and has already provided unprecedented insights into the transcriptional complexities of a variety of organisms.

RNA-Seq is a relatively modern approach used to generate read counts of complementary DNA in parallel to generate a comprehensive set of corresponding gene expression levels. Some ML models effectively generalize between microarray and RNA-Seq data. RNA-Seq and microarray-based predictive models can predict clinical outcomes with a similar performance. RNA-Seq better represents transcript expression patterns that map onto clinically and genetically generated cancer subgroups compared with microarray data (Zhang et al. 2015).

L1000 is a high-throughput gene expression assay that measures the mRNA transcript abundance of 978 “landmark” genes from human cells. Landmark gene expression levels measured with the L1000 microarray have been assessed in The Library of Integrated Cellular Signatures (LINCS), which uses expression of ~1,000 landmark genes to infer ~21,000 target genes with ~81% accuracy. LINCS measured ~1.4 million gene expression profiles of heterogeneous normal and diseased tissue. Computational analysis of large gene expression indicates that it would be feasible to derive sufficient information about the transcriptional state of a cell by measuring only a subset of expressed genes. In addition to that, the genome-wide expression analysis has shown that gene expression is highly correlated, with a small cluster of genes that exhibit similar expression patterns across cell states. The genes that are part of the landmark genes have an expression profile that has been determined as being informative to characterize the transcriptome and can be directly measured from the L1000 assay. These genes have a good predictive power for inferring the expression of other genes that are not directly measured in the assay (Chen et al. 2016; Clayman et al. 2020a, 2020b).

This study used the The Cancer Genome Atlas Program (TCGA)/NCI Genomic Data Commons (GDC) dataset to select RNA-Seq and clinical outcome data for the present analysis. The TCGA is a comprehensive set of studies compiled through the National Institutes of Health that includes genetic and clinical data within individual patient samples. TCGA data have been thoroughly assessed because TCGA data includes a large depth (large sample size) and breadth (heterogeneity of sample types and clinical data) for various applications, including predictive analytics and ML (Tomczak et al. 2015). The TCGA data, along with other cancer research data, are currently hosted through the GDC, a data repository initiated in June 2016 for its applications in precision medicine.

Previous studies (Clayman et al. 2020a; Liang et al. 2015; Quang et al. 2015; Duan et al. 2016; Chen et al. 2018; Tsagri et al. 2018; Duncan et al. 2008; Danaee et al. 2017; Kursa and Rudnicki 2010; Kogelman and Kadarmideen 2014; Petralia et al. 2016; Chen et al. 2016; Clayman et al. 2020b) used different predictive analytics methods, such as clustering methods, deep learning, and feature selection, to evaluate the impact of genes on clinical responses. Many of these studies (Clayman et al. 2020a; Duan et al. 2016; Chen et al. 2018; Tsagri et al. 2018; Duncan et al. 2008; Danaee et al. 2017; Kursa and Rudnicki 2010; Kogelman and Kadarmideen 2014; Petralia et al. 2016; Chen et al. 2016; Clayman et al. 2020b) have assessed the impact of clinical and genetic variables on clinical results such as metastases possibility and survival time. Some of the studies (Lin et al. 2018; Way et al. 2019; Bailey et al. 2018; Saltz et al. 2018; Malta et al. 2018) assessed heterogeneous datasets, including data from several cancer types. Others evaluated homogeneous datasets with data from a single cancer type.

However, studies in the past failed to understand the significance and role of landmark genes in disease-type predictions. The exact nature and characteristics of landmark genes are still unknown. It is unknown as to how different the landmark genes are when compared with non-landmark genes with respect to predicting disease types. By using statistical techniques, we explored if there is any significant difference in the characteristics of the landmark and non-landmark genes. The present study chose genes based on selection criteria, that is, landmark or non-landmark, and compared the ability of genes and/or gene sets to predict clinical outcomes.

This study sought to understand if there is any significant difference in the characteristics of the landmark and non-landmark genes. Earlier studies (Duncan et al. 2008; Chen et al. 2016; Danaee et al. 2017; Ramaker et al. 2017; Bailey et al. 2018; Chen et al. 2018; Huang et al. 2018; Way et al. 2018; Daoud and Mayo 2019; Clayman et al. 2020a) only focused on using the expression values of the landmark genes to determine the expression values of the non-landmark genes but did not discuss whether the landmark genes are any different from the non-landmark genes, that is, could we find a different set of non-landmark genes and say that they are similar in characteristics to the original set of identified landmark genes.

2. Literature Survey

Personalized medicine can be facilitated by analyzing both the genomic and clinical variables. Genes interact with one another and with the different clinical variables such as survival, cancer stage, gender, and age of diagnosis to determine the disease type. For example, let us consider cancer types, namely, prostate, breast, ovarian, and pancreas cancer. All these cancer types possess several genes in common, including the breast cancer 1, early onset (*BRCA1*) and *BRCA2*. The mutation in *BRCA1* and *BRCA2* is associated with a Gleason score ≥ 8, T3/T4 tumor stage, nodal involvement, and metastases at the time of diagnosis in prostate cancer patients (Castro et al. 2013). With another gene, *TP53,* the presence or absence of a *TP53* mutation has been identified as a predictor of survival in prostate cancer patients (Ecke et al. 2010). However, the clinical variables of prostate cancer patients such as the tumor state can be used to predict treatment resistance. Prostate cancer adenocarcinoma metastases possess greater treatment resistance as opposed to primary tumors and possess more de-differentiation of phenotypes. The prostate adenocarcinomas have a strong inverse relationship between stemness index and reduced leukocyte fractions, indicative of reduced immune response when tissue is more differentiated as indicated by mRNA expression-based stemness index response (Malta et al. 2018).

Studies in the past associated with the GDC/TCGA database compared distinct cancer subsets (Bailey et al. 2018) and specifically used deep learning to study immunohistochemical data (Saltz et al. 2018) and mRNA expression-based stemness index (Malta et al. 2018). One study assessed RNA-Seq data available on TCGA (Lim et al. 2020), and one studied a cancer pathway, Ras, across various cancer types by incorporating RNA-Seq data (Way et al. 2018). Also, RNA-RNA interactions have been explored for different cancer subtypes by using deep learning techniques (Dutil et al. 2018). However, not much has been reported with regard to the interactions between the RNA and clinical data for different cancer subtypes. A study on the GDC/TCGA database also compared distinct cancer subsets (Bailey et al. 2018). Twenty-six distinct computational tools and/or algorithms established driver genes that influence distinct cancer and/or cell types and anatomic sites within the TCGA dataset. These include algorithms such as a random forest algorithm used for predicting oncogenes and tumor suppressor genes from somatic mutations. The consensus list/union of the gene sets generated through each of these 26 approaches were pooled for downstream analysis, which included methods that factored in weighting of genes based on performance for distinct cancer types (Bailey et al. 2018).

Genes play a very significant role in the progression of diseases. Some genes are predictive of cancer severity, whereas other genes, including *TP53,* are protective against the development of cancer. Mutations in *TP53* results in alteration in stress and cell-cycle transcriptional regulator genes in few cancer types, and the intensity of the alteration vary across other cancer types. Target genes that are either up- or downregulated in response to a *TP53* mutation involve functions such as cell-cycle inhibition, apoptosis, p53 regulation, and DNA damage response. Genes, for example, *TP53,* that influence pathways that regulate many other genes are especially important to consider when assessing clinical outcomes given that their expression can both directly and indirectly modulate cancer and/or tumor stage progression (Parikh et al. 2014).

The expression level of the genes obtained through RNA-Seq can be used to build predictive models that can predict the outcomes of diseases. A combination of genetic and clinical characteristics can increase the ability to predict overall survival of prostate cancer patients. Personalized medicine is important because it has increasingly been applied with success in clinical trials. In addition, early detection of cancer produces an increase in survival rate and consideration of clinical variables, along with RNA-Seq data, can be used to increase efforts at early detection of cancer (Clayman et al. 2020a). One of the studies focused on developing data analytics techniques to analyze the RNA-Seq and clinical data gathered from the NCI database. By using the data modalities on the genomic and clinical data obtained from the TCGA and by applying integrative clustering, Liang et al. (2015) reported effective differentiation of clinical subgroups for ovarian cancer. A study also investigated the relationship between genetic and clinical variables by accounting for both coding and non-coding genetic variants (Quang et al. 2015).

Computational costs of biological data analysis call for increasingly efficient methods of determining which genetic and clinical factors are most relevant for understanding the overall genetic and clinical profiles of human patients (Duan et al. 2016). This challenge is especially difficult given that distinct individuals can possess distinct profiles of genetic expression, and certain genetic conditions can be more readily captured than others when using a varying number of genetic features. Dimensionality reduction methods, such as random forest analysis, k-means clustering, and principal component analysis (PCA), are often used in tandem to capture essential elements of the data that explain larger datasets when using a subset of relevant features. Dimensionality reduction methods such as PCA are effective methods of data representation when linear relationships are present. PCA can detect multiple types of cancer while also selecting relevant features (Chen et al. 2018). A study to predict cancer outcomes (Tsagri et al. 2018) applied feature selection methods, including PCA and Boruta random forest, for dimensionality reduction. The utility of the *k*-means clustering for protein expression and cancer outcomes is demonstrated in the study by Duncan et al. 2008. Selection methods were further refined by applying the random forest decision tree classifier to determine a smaller subset of important genes to use for downstream analysis in several clustering methods, including *k*-means, partition around medoids, and res-hierarchical clustering. These clustering methods were used to generate subsets within the dataset in an unsupervised manner.

One study used deep learning to assess the entire search space of gene expression levels from RNA-Seq data (Danaee et al. 2017). Another study implemented dimensionality reduction and feature selection methods to reduce computation and model complexity. When dealing with results of gene expression measurements in the context of cancer, identification of a minimal-optimal set of genes related to cancer is often useful for establishing genetic markers (Kursa and Rudnicki 2010). In this way, Boruta analysis can be applied specifically to the approach of identifying the minimal-optimal set of genes as a selection method to restrict analysis to relevant genes. Previous studies implemented thresholding, weighting (Kogelman and Kadarmideen 2014), and networks analysis (Petralia et al. 2016) to assess whether biologically relevant interactions can improve model performance.

A set of 978 landmark genes has been established as predictors of the remaining genes in a microarray dataset analyzed by Chen et al. (2016). When applying 978 landmark genes as inputs, a deep learning method (D-GEX) results in lower error compared with linear regression in predicting expression of 81.31% of target genes in an independent RNA-Seq–based GTEx dataset (Chen et al. 2016). As an extension of the analysis performed by Chen et al. (2016), these 978 landmark genes from the L1000 dataset were selected from the GDC's RNA-Seq dataset to assess whether 978 landmark genes improve clustering (Clayman et al. 2020a).

Dimensionality reduction methods such as PCA are used to select relevant features. In addition to that, the unsupervised learning technique, *k*-means clustering performs well when applied to data with low effective dimensionality. Our previous study (Clayman et al. 2020b) showed that 978 landmark genes better differentiated *k*-means clusters compared with 978 randomly selected non-landmark genes. *K*-means clusters generated from the landmark genes show more separation of cluster groups when plotted against the first two principal components, which capture a greater proportion of variation for the 978 landmark genes (Clayman et al. 2020b). Analysis of these results suggests that the 978 landmark genes better represent the overall genetic profile of these heterogeneous samples. However, clustering results varied when using the 978 landmark genes versus the 978 non-landmark genes as features, depending on whether clustering was performed on the heterogeneous versus the homogeneous datasets. For the heterogeneous dataset, the percentage of variation captured by each of the first two principal components was greater for the 978 landmark genes (PCA1, 13.1%; PCA2, 9.2%) versus the 978 non-landmark genes (PCA1, 9.4%; PCA2, 6.2%), with similar results for the homogeneous dataset. Variability, depending on the set of genes selected, is also depicted based on the distinct appearance of cluster plots, which possess more visual overlap and greater between-cluster sum of squares for the non-landmark genes compared with the landmark genes for both the homogeneous and heterogeneous datasets. *K*-means clustering results coincide with the clinical variable of the Ann Arbor cancer stage to a greater extent when using non-landmark genes as features compared with landmark genes (Clayman et al. 2020b).

The study by Clayman et al. (2020a) depicted the use of 978 landmark genes as a more effective method of identifying distinct clusters of individuals according to visualization of data clusters against the first two principal components of the data when assessing large heterogeneous datasets. Clusters in these plots are more distinct compared with cluster plots generated by using 978 randomly selected non-landmark genes in the dataset, which supports the use of these landmark genes as a representation of the genetic profile of these samples when assessing heterogeneous datasets (Clayman et al. 2020a; Chen et al. 2016). In contrast, non-landmark genes capture more of the variation in the data for homogeneous and heterogeneous datasets studied here. Despite this, the non-landmark genes allow for clustering into groups more consistent with clinical variables for the homogeneous dataset compared with the 978 landmark genes. Certain genes or clinical variables can be more predictive of clustering results than others. When assessing the separation of groups, the role of sets of individual genes and clinical variables can be examined further. Cluster analysis can be used to inform future studies on the ability of genes to predict clinical variables as well as the ability of clinical variables to characterize clusters derived from gene expression results, as examined in this study. This can be especially relevant toward applications for personalized medicine such as treatment responsiveness, depending on the combination of genetic and clinical variables (Clayman et al. 2020a). Predictive models of cancer outcomes can be built by specific protein expression levels with RNA-Seq. The overall survival of prostate cancer patients can be precisely predicted by genetic and clinical characteristics (Clayman et al. 2020a, 2020b). Personalized medicine has become a new trend because there are many successful cases in clinical trials with personalized medicine. Moreover, the survival rate can be increased by early detection of cancer with clinical variables and RNA-Seq data (Clayman et al. 2020a, 2020b).

Other studies evaluated histopathologic imaging data (Ash et al. 2018; Saltz et al. 2018), multi-omics data (Liang et al. 2015; Chaudhary et al. 2018; Lin et al. 2018; Way et al. 2019), mRNA data (Azarkhalili et al. 2018), microarray data (Daoud and Mayo 2019), or RNA-Seq data (Danaee et al. 2017) from the TCGA commons to predict clinical outcomes by using deep learning methods, including convolutional and variational autoencoders. Studies implemented other ML techniques (Huang et al. 2018), support vector machines (Bailey et al. 2018), ensemble methods (Way et al. 2019; Bailey et al. 2018), construction of latent dimensionalities and PCA (Way et al. 2019), feature selection, and clustering (Liang et al. 2015) to assess the impact of genes on clinical responses. A study by Duncan et al. (2008) applied *k*-means clustering to assess protein expression and cancer outcomes. PCA can be applied in predictive analysis of multiple types of cancer by selecting relevant features and capturing linear relationships in the data to reduce the dimensionality of data (Chen et al. 2018). Some of these studies evaluated homogeneous datasets that contain data from a single cancer type, such as prostate cancer (Saltz et al. 2018), breast cancer (Danaee et al. 2017), liver cancer (Chaudhary et al. 2018), lung adenocarcinoma (Chaudhary et al. 2018), and acute myeloid leukemia (Lin et al. 2018). Other studies assessed heterogeneous datasets with data from multiple tumor types (Lin et al. 2018) or multiple cancer types (Ash et al. 2018; Azarkhalili et al. 2018; Bailey et al. 2018; Huang et al. 2018; Way et al. 2018). A study by Petralia et al. (2016) evaluated gene and protein networks within TCGA breast cancer data using the random forest classifier. Previous studies of the GDC used feature selection to reduce the set of genes used for predicting clinical outcomes. Landmark genes have not been extensively used for assessing the GDC dataset and have not been assessed for further feature selection approaches to further reduce this set of genes for predictive analysis (Clayman et al. 2020a, 2020b).

3. Methods and Materials

A total of three datasets were used in this study. The clinical and RNA-Seq dataset (dataset 1) was obtained from the NCI's GDC repository. This dataset consists of clinical and genetic information for tissues of 55 cancer types. In total, there are 13,122 observations (instances) of 83 clinical variables and >20,000 genetic variables. Two L1000 datasets, namely, the microarray version of the L1000 dataset (dataset 2) and the RNA-Seq version of the L1000 dataset (dataset 3) were also analyzed in this study. A brief introduction to all three datasets is provided here.

Table 1: The number of tissue samples per disease types in dataset 1.

Disease (Cancer) Type	No. Samples
Breast	1,485
Kidney	1,448
Brain	759
Colon	675
Prostate gland	660
Bladder	488
Skin	474
Stomach	460
Pancreas	188
Testis	165

3.1. Datasets

3.1.1. Dataset 1

Of 13,122 diseased tissues of 55 disease types, a random subset of 6,802 diseased tissues across 10 different disease types were analyzed in this study. A total of ∼22k genetic variables were considered as predictors for each diseased tissue. The total numbers of tissue samples across each disease type considered in this study are listed in Table 1.

Due to the high dimensionality of the datasets, the descriptions of the individual predictors are not provided here.

3.1.2. Dataset 2

The L1000 microarray-based dataset in the GCTx format contained expression data of 22,268 (∼22k) genes (rows) across 129,158 tissues (columns). Of the ∼22k rows, the first 978 rows were the landmark genes and the remaining 21,290 rows were the non-landmark genes whose expression values were predicted by using the landmark genes. A sample dataset that consisted of ∼22k genes across 6,802 diseased tissues was obtained by matching the tissue identifier in both dataset 1 and dataset 2. To eliminate the variability, the sample dataset was quantile normalized into the numerical range between 4 and 15, that is, the expression values of the genes were in the range between 4 and 15.

3.1.3. Dataset 3

The L1000 RNA-Seq-based dataset contained expression data of 22,268 (∼22k) genes (rows) across 129,158 tissues (columns). Of the ∼22k rows, the first 978 rows were the landmark genes and the remaining 21,290 rows were the non-landmark genes. A sample dataset that consisted of ∼22k genes across 6,802 diseased tissues was obtained by matching the tissue ids in both dataset 1 and dataset 3. To eliminate the variability, the sample dataset was quantile normalized across all samples such that they know all have the same distribution (e.g., same mean ± standard deviation [SD]).

3.2. Tools and Techniques

3.2.1. Analysis of variance

It is a statistical tool used to detect differences between experimental group means. Analysis of variance (ANOVA) is performed in experimental designs with one dependent variable that is a continuous parametric numerical outcome measure, and multiple experimental groups within one or more independent (categorical) variables. The independent variables are called factors, and groups within each factor are referred to as levels. ANOVA, similar to linear regression and general linear models, quantifies the relationship between the dependent variable and the independent variable(s). There are three different general linear models for ANOVA: the fixed effects model, which makes inferences that are specific and valid only to the populations and treatments of the study; the random effects model, which makes inferences about levels of the factor that are not used in the study, that is, this model pertains to random effects within levels, and makes inferences about a population's random variation; and the mixed effects model, which contains both the fixed and the random effects (Sawyer 2009).

Assumptions for ANOVA: a data set should meet the following criteria before performing ANOVA (Sawyer 2009):

Parametric data: A parametric ANOVA requires parametric data (ratio or interval measures). There are nonparametric, one-factor versions of ANOVA for nonparametric ordinal (ranked) data, specifically the Kruskal-Wallis test for independent groups and the Friedman test for repeated measures analysis.

Normally distributed data within each group: The fundamental assumption of parametric ANOVA is that each group of data (each level) be normally distributed. The Shapiro-Wilk test is commonly used to test for normality for group sample sizes (N) < 50 and the D'Agnostino modification is useful for larger samplings (N > 50).

Homogeneity of variance within each group: Because ANOVA compares normal distribution curves of datasets, these curves need to be similar to each other in shape and width for the comparison to be valid. In other words, the amount of data dispersion (variance) needs to be similar between groups. Two commonly invoked tests of homogeneity of variance are by Levene and by Brown and Forsthye.

Independent observations: A general assumption of parametric analysis is that the value of each observation for each subject is independent of the value of any other observation. For independent groups designs, this issue is addressed with random sampling, random assignment to groups, and experimental control of extraneous variables.

Most commercially available statistics programs perform normality and homogeneity of variance tests. Determination of the parametric nature of the data and soundness of the experimental design is the responsibility of the investigator, reviewers, and critical readers of the literature (Sawyer 2009).

Robustness of ANOVA to violations of normality and variance assumptions: ANOVA tests can handle moderate violations of normality and equal variance if there is a large enough sample size and a balanced design. The validity of ANOVA is said to be "robust" in the face of violations of normality assumptions if there is an adequate sample size. ANOVA is more sensitive to violations of the homogeneity of variance assumption, but this is mitigated if sample sizes of factors and levels are equal or nearly so. If normality and homogeneity of variance violations are problematic, then there are three options: transform the data to best mitigate the violation; use one of the nonparametric ANOVAs, but at the cost of reduced power and being limited to one-factor analysis; or identify outliers in the dataset by using formal statistical criteria. In that case, use caution in deleting outliers from the dataset; such decisions need to be justified and explained. Removal of outliers will reduce deviations from normality and homogeneity of variance (Sawyer 2009).

3.2.2. Kruskal Wallis H test

This test is a nonparametric alternative to the one-way ANOVA. The Kruskal Wallis H test is used when the assumptions for ANOVA are not met. This test is also referred to as one-way ANOVA on ranks because the ranks of the data values are used in the test rather than the actual data points. This test determines whether the medians of two or more groups are different. The hypotheses for the test are the following:

H0: Population medians are equal.

H1: Population medians are not equal.

The Kruskal-Wallis H test is more suitable for analysis of the dataset in which the sample size is small (<30). For the dataset that is not normally distributed and contains some strong outliers, it is more appropriate to use ranks rather than actual values to avoid the testing being affected by the presence of outliers or by the non-normal distribution of data. This test also assumes that the observations are independent of each other. The Kruskal Wallis H test will determine if there is a significant difference between groups. However, this test cannot determine which groups are different. To determine which groups are significantly different, a post hoc test needs to be performed.

The assumptions for the Kruskal Wallis H test are the following:

- The test is more commonly used when an independent variable has three or more levels.
- The scales for the dependent variable are either ordinal, ratio, or interval.
- All observations should be independent; in other words, there should be no relationship between the members in each group or between groups.
- All groups should have the same shape distributions.

3.2.3. Paired Wilcoxon signed-rank test

The nonparametric analog of the *t*-test is the Wilcoxon signed-rank test and is used when the one-sample *t*-test assumptions are violated. The pairwise Wilcoxon signed-rank test is performed as a post hoc test to determine which groups are significantly different from other groups. The assumptions of the Wilcoxon signed-rank test are as follows:

- The differences between the data values are continuous (not discrete).
- The distribution of each difference (of the data values) is symmetric.
- The differences of the data values are mutually independent.
- The differences of the data values all have the same median.
- The measurement scale of the data value is interval.

In summary, parametric tests are more commonly used than are nonparametric tests. However, parametric tests require an important assumption, which is the assumption of normality. This means that the distribution of sample means is normally distributed. But, when this assumption is not satisfied, the parametric tests can be misleading. In such situations, nonparametric tests are the available alternative. The nonparametric tests are statistical methods based on signs and ranks. When used, nonparametric tests convert the original data into the order size instead of using the original data value and only uses the rank or signs. Although this can result in the loss of information, but, when the data are not normal, the nonparametric analysis has more statistical power than the parametric analysis. In particular, when the means of the sample group are not normally distributed and when the variances are equal across groups, then nonparametric statistical techniques are excellent alternatives. Another advantage of using the nonparametric test is that it is not sensitive to outliers (Nahm 2016).

3.2.4. Experimental design strategies

This study focused on developing data analytics techniques to assess the genetic and clinical data gathered from the GDC. This is a relevant area of research given that these research techniques have applications in analysis of bioinformatics datasets in general.

3.2.5. Data collection and integration of genetic and clinical GDC data

Data were downloaded from the GDC by using the GDC Data Portal. RNA-Seq data for each cancer type were appended into a dataframe, which included a corresponding sample id for later integration of RNA-Seq and clinical data, which allowed for individual subject-level data analysis. The microarray (LINCS L1000) dataset, which consisted of the expression values of >20,000 genes across 129k samples was collected from the Gene Expression Omnibus (GEO) repository. This dataset is curated by the Broad Institute, which is publicly available in the GEO repository.

3.2.6. Data normalization

The DeSeq Bioconductor package in R (Love et al. 2014) was used to normalize the RNA-Seq dataset for all the cancer types. This normalization method accounts for each gene length as well as the number of observations in the dataset.

3.3. Experiments

3.3.1. Experiment A: Perform ANOVA and the Kruskal Wallis H test to compare the number of differentially expressed genes within the landmark and non-landmark gene sets across different tissue samples (6,802 tissue samples)

Initially, ANOVA is performed on all the genes within the landmark gene set and the non-landmark gene sets to determine if a gene within the set is or is not differentially expressed (Koch et al. 2018). The null and the alternate hypothesis of the ANOVA is given as follows:

H_0 = the gene is not differentially expressed

H_a = the gene is differentially expressed

A significant *p*-value ($p < 0.05$, given $\alpha = 0.05$) that results on an ANOVA test would indicate that the gene is differentially expressed (Koch et al. 2018). For each of the gene sets (landmark and non-landmark), the total numbers of genes that are differentially expressed is identified by performing ANOVA. Then, when using the Kruskal Wallis H test, it is determined if the number of genes that is differentially expressed within the different sets are or are not similar. If the null hypothesis is rejected, then it can be concluded that at least one of the sets has a different number of differentially expressed genes compared with the others. The design strategy for this experiment is summarized in Figure 1.

Note here that, due to the small sample size (n = 15), the nonparametric test was performed because we could not assume that the distribution of sample means is normal. The nonparametric Kruskal Wallis H test was performed here because the variables were measured on a continuous scale, the independent variable consists of two or more categorical, independent, or unrelated group, and there is no relationship between the observations in each group (Nahm 2016).

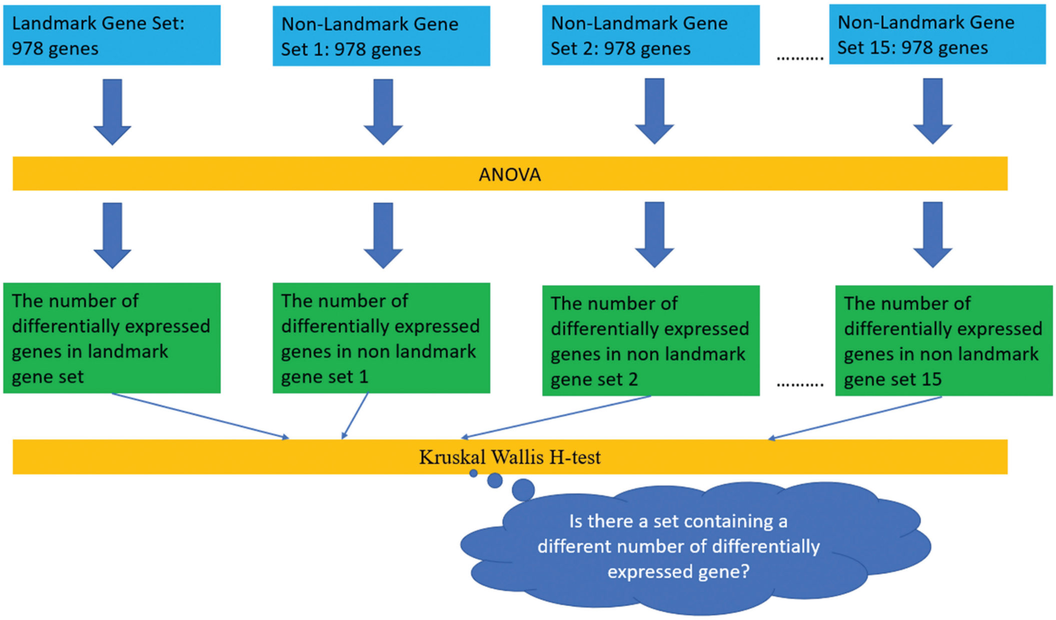
Landmark Gene Set: 978 genes
Non-Landmark Gene Set 1: 978 genes
Non-Landmark Gene Set 2: 978 genes
....
Non-Landmark Gene Set 15: 978 genes
ANOVA
The number of differentially expressed genes in landmark gene set
The number of differentially expressed genes in non landmark gene set 1
The number of differentially expressed genes in non landmark gene set 2
....
The number of differentially expressed genes in non landmark gene set 15
Kruskal Wallis H-test
Is there a set containing a different number of differentially expressed gene?

Figure 1: Design strategy for experiment A.

3.3.2. Experiment B: Perform an ANOVA and a Kruskal Wallis H test to compare the number of differentially expressed genes within the landmark and non-landmark gene sets across different tissue samples classified by race, ethnicity, and disease types

To begin with, the diseased tissues were classified (split) by categories either by race or ethnicity, or by disease types. Within subcategories (e.g., Asian, White, Black) of each category (e.g., race), ANOVA was performed on all the genes within the landmark gene set and the non-landmark gene sets to determine if a gene within the set was or was not differentially expressed. For each subcategory within a category, ANOVA was performed to identify the number of differentially expressed genes within each of the gene sets (landmark and non-landmark) (Koch et al. 2018). For each subcategory, within each category, the Kruskal Wallis H test was used to determine if there was a significant difference in the number of differentially expressed genes. If the null hypothesis is rejected, then it can be concluded that at least one or more subcategories have a different number of differentially expressed genes. Also, the Kruskal Wallis H test was performed within each category to determine if the number of genes that were differentially expressed within the different gene sets across different subcategory were or were not similar. If the null hypothesis is rejected, then it can be concluded that at least one of the sets within the gene sets (landmark or non-landmark gene sets) has a different number of differentially expressed genes across different subcategories. Finally, the pairwise Wilcox signed-rank test was used to identify the subcategory or the gene set that was different from the others in terms of the number of differentially expressed genes. The pairwise Wilcox signed-rank test is a post hoc test because the Kruskal Wallis H test is an omnibus test statistic. It cannot tell which specific groups of the independent variable are statistically significantly from each other. The design strategy for the experiments conducted in this section is summarized in Figure 2.

3.3.3. Experiment C: Perform correlation studies to determine the pairwise correlation range of the genes in the landmark and non-landmark gene sets

In this experiment, the expression values of the 978 landmark genes are compared with the expression values of the 15 different randomly selected 978 non-landmark genes set across ∼129k tissue samples. An ANOVA was performed on a dataset that consisted of randomly selected 100 correlation values between gene pairs from both the landmark gene set and the 15 different non-landmark gene set. Here, rejecting the null hypothesis would indicate that there is no significant difference in the correlation values of the gene pairs across the landmark and non-landmark gene sets. Here, the parametric test ANOVA is performed on the dataset (16 gene sets), which consisted of the randomly selected 100 correlation values between gene pairs because the sample size (n = 100) was good enough to assume that the sample was taken from the normally distributed population, that is, each sample was drawn independently of the other samples, the variance in different groups (gene sets) was the same, and the correlation values in each group was continuous. Visualizations, such as boxplots, would highlight the variations in the correlations of the gene pairs across the different set of genes. The design strategy for the experiment conducted in this section is summarized in Figure 3.

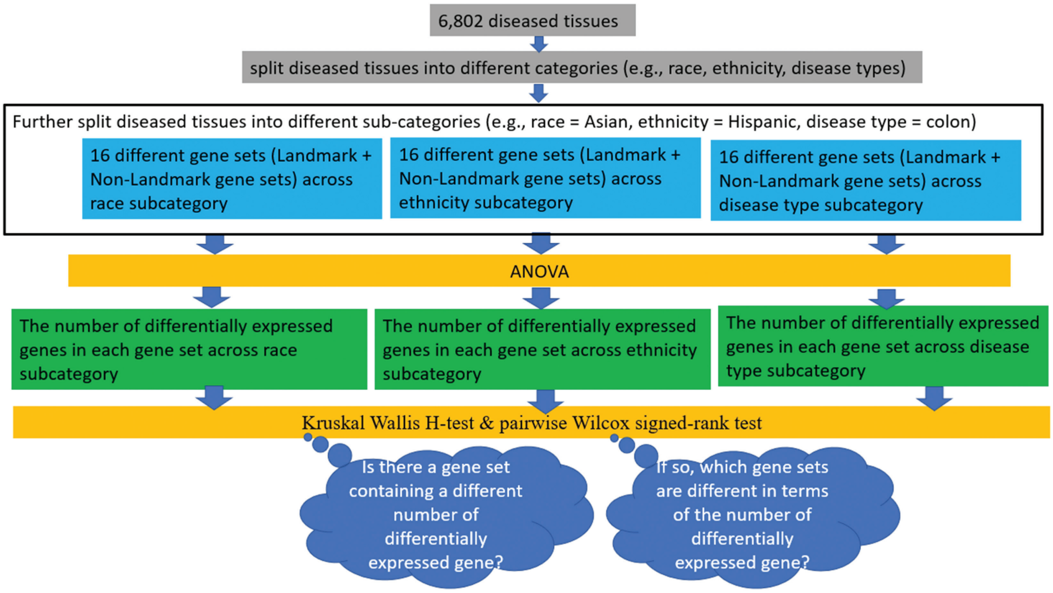
6,802 diseased tissues
split diseased tissues into different categories (e.g., race, ethnicity, disease types)
Further split diseased tissues into different sub-categories (e.g., race = Asian, ethnicity = Hispanic, disease type = colon)
16 different gene sets (Landmark + Non-Landmark gene sets) across race subcategory
16 different gene sets (Landmark + Non-Landmark gene sets) across ethnicity subcategory
16 different gene sets (Landmark + Non-Landmark gene sets) across disease type subcategory
ANOVA
The number of differentially expressed genes in each gene set across race subcategory
The number of differentially expressed genes in each gene set across ethnicity subcategory
The number of differentially expressed genes in each gene set across disease type subcategory
Kruskal Wallis H-test & pairwise Wilcox signed-rank test
Is there a gene set containing a different number of differentially expressed gene?
If so, which gene sets are different in terms of the number of differentially expressed gene?

Figure 2: Design strategy for experiment B.

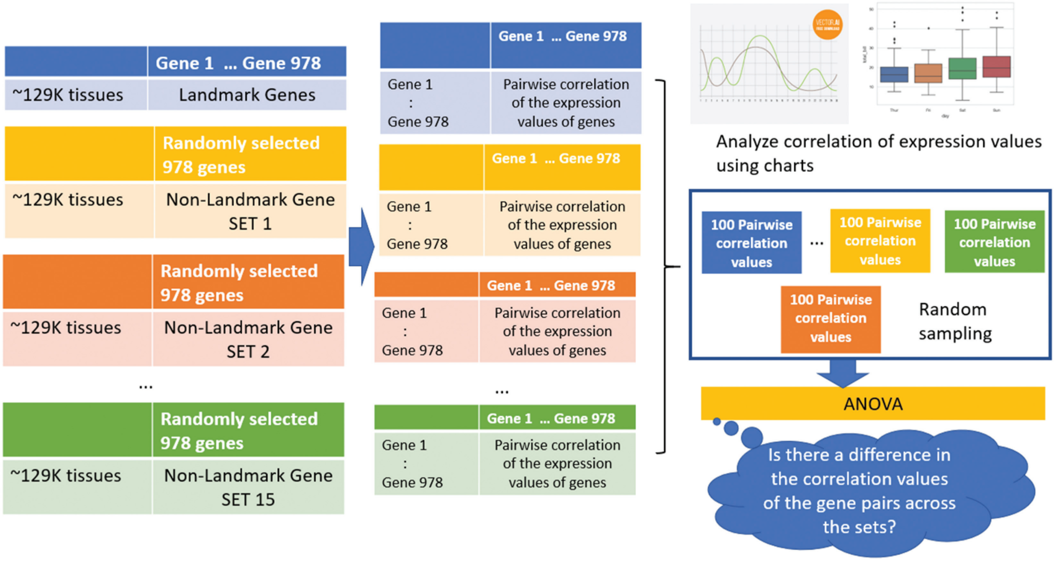
Gene 1 ... Gene 978
~129K tissues
Landmark Genes
Randomly selected 978 genes
~129K tissues
Non-Landmark Gene SET 1
Randomly selected 978 genes
~129K tissues
Non-Landmark Gene SET 2
...
Randomly selected 978 genes
~129K tissues
Non-Landmark Gene SET 15
Gene 1 ... Gene 978
Gene 1 : Gene 978
Pairwise correlation of the expression values of genes
...
Analyze correlation of expression values using charts
100 Pairwise correlation values
...
100 Pairwise correlation values
100 Pairwise correlation values
100 Pairwise correlation values
Random sampling
ANOVA
Is there a difference in the correlation values of the gene pairs across the sets?

Figure 3: Design strategy for experiment C.

4. Results and Discussions

This study sought to address if there is any significant difference in the characteristics of the landmark and non-landmark genes. In an attempt to address the above-mentioned objectives, a total of three experiments were designed. Here, we present the results obtained from all three experiments and also discuss the inferences gathered from the results obtained.

To begin with, 16 different sets of 978 genes were obtained from dataset 1. One of the sets included the 978 landmark genes and the remaining 15 sets included the randomly chosen 978 genes out of the pool of non-landmark genes. When randomly choosing the genes for each set, it was ensured that none of the genes were duplicated within and across the sets (Figure 4). The tissue samples within dataset 1 were further portioned across race, ethnicity, and disease (cancer) types.

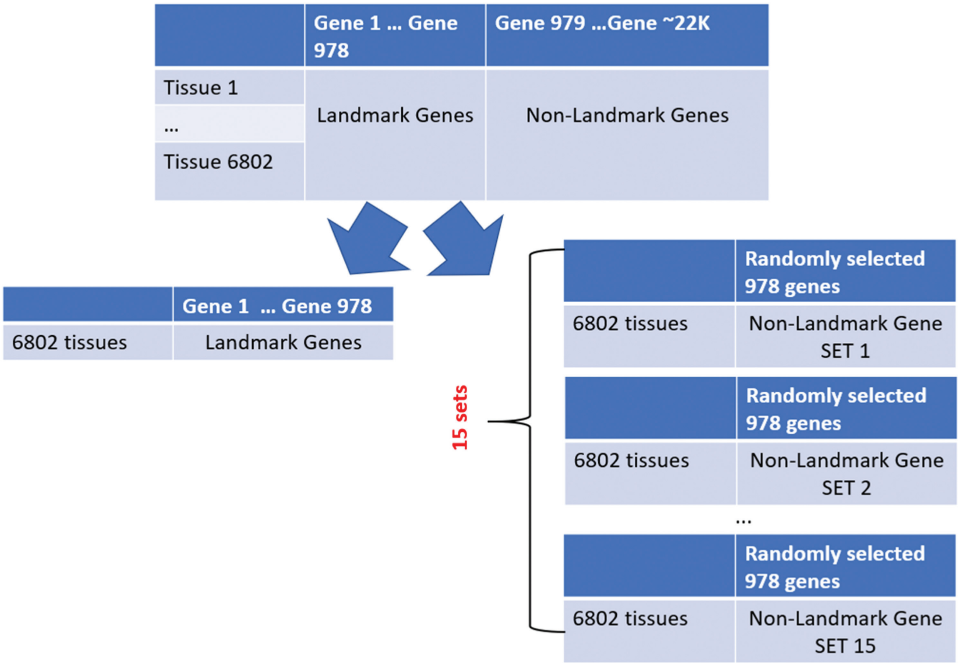

Figure 4: Sixteen sets of 978 genes used in experiments A–C.

Across the 16 sets of genes, ANOVA was performed to analyze the expression values of the genes across 6,802 diseased tissue samples of 10 cancer types. The objective was to determine how many genes in each of the sets were differentially expressed (refer to experiment A). Identifying the differentially expressed genes is critical because they are assumed to be the driving force and/or the molecular biomarkers of different phenotypes (Zhao et al. 2018). Within the landmark and the non-landmark sets of 978 genes, it is important to determine if there is a statistical difference in the number of differentially expressed genes. If there is a statistical difference in the number of differentially expressed genes between one or more sets, then it would indicate that the characteristics of one set is different from the other. The total number of genes, of the 978 genes, that were differentially expressed across the landmark set and the 15 different non-landmark sets in three different samples, namely, sample 1, sample 2, and sample 3, are highlighted in Table 2. The number of differentially expressed genes across each sample for both the landmark gene set and the non-landmark gene sets were obtained by considering different p-values for the ANOVA, that is, for sample 1, sample 2, and sample 3, the p-values were <0.05, <0.01, and <0.1, respectively.

In the landmark gene set, 99.9% of the genes were differentially expressed across the 6,802 diseased tissue samples. ANOVA resulted in a p-value of all the 977 genes to be < 0.1. However, the number of genes that were differentially expressed across the 15 sets of non-landmark genes ranged between 973 and 978 (Table 2). The Kruskal Wallis H test was performed to determine if there is a significant difference in the number of differentially expressed genes across the 16 gene sets. Here, three different samples, namely, sample 1 ($p < 0.05$), sample 2 ($p < 0.01$), and sample 3 ($p < 0.1$), were considered for the analysis (Table 2). Across the 16 different gene sets, no significant difference in the number of differentially expressed genes was observed at $\alpha = 0.05$. The Kruskal-Wallis H test resulted in a p-value of 0.1759. This implies that both the landmark gene set and the 15 different set of randomly chosen non-landmark gene sets, both have similar numbers of differentially expressed genes.

The total number of differentially expressed genes across the 16 gene sets (the landmark set and 15 non-landmark sets) at $p < 0.05$ for the different races, namely, Asian, White, and Black, are recorded in Table 3. Descriptive statistics across the races indicated that relatively more genes were differentially expressed in the White race (mean ± SD, 977 ± 1.61) than in the Black (mean ± SD, 715 ± 117) and Asian (mean ± SD, 656 ± 71.9) races.

The Kruskal Wallis H test was performed to determine if there was a significant difference in the number of differentially expressed genes across the three races. A significant difference was observed in the number of differentially expressed genes across the three races at $\alpha = 0.05$. The Kruskal-Wallis H test resulted in a p-value of 7.767e-08. In addition to that, a pairwise Wilcox signed-rank test was performed to determine which group of races differed from each other in terms of the number of differentially expressed genes. At $\alpha = 0.05$, a significant difference in the number of differentially expressed genes was observed for the White race, with the p-value of 1.9e-06 against the Asian and Black race (refer to experiment B).

Table 2: The number of genes that were differentially expressed across the landmark and non-landmark gene sets.

Gene Sets	Sample 1: $p < 0.05$	Sample 2: $p < 0.01$	Sample 3: $p < 0.1$
L	977	977	977
NL set 1	976	976	976
NL set 2	973	973	973
NL set 3	977	977	977
NL set 4	978	978	978
NL set 5	978	978	978
NL set 6	978	978	978
NL set 7	977	976	978
NL set 8	978	976	978
NL set 9	978	978	978
NL set 10	977	975	977
NL set 11	977	975	977
NL set 12	976	976	976
NL set 13	976	974	976
NL set 14	976	975	976
NL set 15	976	974	977

L, landmark genes; NL, non-landmark genes.

Table 3: The number of genes that were differentially expressed across the landmark and non-landmark gene sets for different races at $p < 0.05$.

	$p < 0.05$		
Gene Sets	Asian	White	Black
L	738	978	870
NL set 1	778	978	811
NL set 2	784	977	848
NL set 3	711	978	824
NL set 4	670	978	799
NL set 5	643	978	754
NL set 6	592	977	609
NL set 7	553	974	522
NL set 8	590	978	570
NL set 9	702	978	799
NL set 10	591	973	636
NL set 11	630	976	644
NL set 12	699	978	823
NL set 13	624	976	746
NL set 14	606	976	631
NL set 15	588	975	550

The total number of differentially expressed genes across the 16 gene sets (the landmark set and 15 non-landmark sets) at $p < 0.01$ for the different races, namely, Asian, White, and Black, is recorded in Table 4. Descriptive statistics across the races indicates that relatively more genes were differentially expressed in the White race (mean ± SD, 975 ± 2.73) than in the Black (mean ± SD, 603 ± 132) and Asian (mean ± SD, 515 ± 77.1) races.

The Kruskal Wallis H test was performed to determine if there was a significant difference in the number of differentially expressed genes across the three races. A significant difference was observed in the number of differentially expressed genes across the three races at $\alpha = 0.05$. The Kruskal-Wallis H test resulted in a p-value of 6.584e-08. In addition to that, a pairwise Wilcox signed-rank test was performed to determine which group of races differed from

Table 4: The number of genes that were differentially expressed across the landmark and non-landmark gene sets for different races at $p < 0.01$.

	$p < 0.01$		
Gene Sets	Asian	White	Black
L	589	978	785
NL set 1	648	978	717
NL set 2	665	976	741
NL set 3	574	978	746
NL set 4	510	978	696
NL set 5	472	978	641
NL set 6	450	976	479
NL set 7	413	970	385
NL set 8	456	974	444
NL set 9	575	977	694
NL set 10	460	971	523
NL set 11	502	972	530
NL set 12	556	978	718
NL set 13	487	975	629
NL set 14	447	973	505
NL set 15	436	975	422

each other in terms of the number of differentially expressed genes. At $\alpha = 0.05$, a significant difference in the number of differentially expressed genes was observed for the White race, with the *p*-value of 2.1e-06 against the Asian and Black races (refer to experiment B).

The total number of differentially expressed genes across the 16 gene sets (the landmark set and 15 non-landmark sets) at $p < 0.1$ for the different races, namely, Asian, White, and Black, are recorded in Table 5. Descriptive

Table 5: The number of genes that were differentially expressed across the landmark and non-landmark gene sets for different races at $p < 0.1$.

	$p < 0.1$		
Gene Sets	Asian	White	Black
L	799	978	912
NL set 1	835	978	850
NL set 2	840	978	881
NL set 3	779	978	863
NL set 4	743	978	838
NL set 5	713	978	801
NL set 6	659	978	663
NL set 7	642	975	581
NL set 8	665	978	631
NL set 9	761	978	840
NL set 10	674	976	693
NL set 11	700	976	709
NL set 12	764	978	863
NL set 13	712	977	797
NL set 14	676	976	697
NL set 15	670	977	620

statistics across the races indicated that relatively more genes were differentially expressed in the White race (mean ± SD, 977 ± 1.01) than in the Black (mean ± SD, 765 ± 107) and Asian (mean ± SD, 727 ± 63.6) races.

The Kruskal Wallis H test was performed to determine if there was a significant difference in the number of differentially expressed genes across the three races. A significant difference was observed in the number of differentially expressed genes across the three races at $\alpha = 0.05$. The Kruskal-Wallis H test resulted in a *p*-value of 9.975e-08. In addition to that, a pairwise Wilcox signed-rank test was performed to determine which group of races differed from each other in terms of the number of differentially expressed genes. At $\alpha = 0.05$, a significant difference in the number of differentially expressed genes was observed for the White race, with the *p*-value of 1.6e-06, against the Asian and Black races (refer to experiment B).

For the above observations, it is conclusive that the number of differentially expressed genes across the races was significantly different, at $p < 0.01$ (see Table 4), $p < 0.05$ (see Table 3), and at $p < 0.1$ (see Table 5). Within the three races, the number of differentially expressed genes was significantly different for White race when compared with the Asian and Black races, at $p < 0.01$ (see Table 4), $p < 0.05$ (see Table 3), and at $p < 0.1$ (see Table 5).

Finally, the Kruskal Wallis H test was performed to determine if there was a significant difference in the number of differentially expressed genes across the 16 gene sets across the three different races. The Kruskal-Wallis H test resulted in a *p*-value of 0.7449 for $p < 0.01$, which indicated that there was no significant difference in the number of differentially expressed genes across the 16 different gene sets, that is, irrespective of landmark or non-landmark genes, both types of the genes are similar across the races in terms of the constitution of the number of differentially expressed genes. Similarly, for the $p < 0.05$ and $p < 0.1$, the Kruskal-Wallis H test resulted in a *p*-value of >0.749, which indicated that, across the races, there was no significant difference in the number of differentially expressed genes across the 16 different gene sets (refer to experiment B).

The total number of differentially expressed genes across the 16 gene sets (the landmark set and 15 non-landmark sets) at $p < 0.05$ for the two ethnic groups, namely, Hispanic and non-Hispanic, is recorded in Table 6. Descriptive statistics across the ethnic groups indicate that relatively more genes are differentially expressed in the non-Hispanic group (mean ± SD, 977 ± 1.31) than in the Hispanic group (mean ± SD, 506 ± 106).

The Kruskal Wallis H test was performed to determine if there was a significant difference in the number of differentially expressed genes across the two ethnic groups. A significant difference was observed in the number of differentially expressed genes across the two ethnic groups at $\alpha = 0.05$. The Kruskal-Wallis H test resulted in a *p*-value of 1.207e-06. Thus, at $\alpha = 0.05$, a significant difference in the number of differentially expressed genes was observed between the Hispanic and non-Hispanic ethnic groups (refer to experiment B).

Table 6: The number of genes that were differentially expressed across the landmark and non-landmark gene sets for different ethnic groups at $p < 0.05$.

	$p < 0.05$	
Gene Sets	Hispanic	Non-Hispanic
L	658	978
NL set 1	620	978
NL set 2	650	976
NL set 3	601	977
NL set 4	555	978
NL set 5	484	978
NL set 6	397	977
NL set 7	331	976
NL set 8	402	977
NL set 9	581	978
NL set 10	414	974
NL set 11	441	975
NL set 12	612	978
NL set 13	516	975
NL set 14	447	978
NL set 15	391	977

Finally, the Kruskal Wallis H test was performed to determine if there was a significant difference in the number of differentially expressed genes across the 16 gene sets across the two ethnic groups. The Kruskal-Wallis H test resulted in a *p*-value of 0.989 for $p < 0.05$, which indicated that there was no significant difference in the number of differentially expressed genes across the 16 different gene sets, that is, irrespective of landmark or non-landmark genes, both types of the genes are similar across ethnic groups in terms of the constitution of the number of differentially expressed genes (refer to experiment B).

The total numbers of differentially expressed genes across the 16 gene sets (the landmark set and 15 non-landmark sets) at $p < 0.01$ for the two ethnic groups, namely, Hispanic and non-Hispanic, are recorded in Table 7. Descriptive statistics across the ethnic groups indicated that relatively more genes were differentially expressed in the non-Hispanic group (mean ± SD, 976 ± 2.28) than in the Hispanic group (mean ± SD, 361 ± 97.8).

The Kruskal Wallis H test was performed to determine if there was a significant difference in the number of differentially expressed genes across the two ethnic groups. A significant difference was observed in the number of differentially expressed genes across the two ethnic groups at $\alpha = 0.05$. The Kruskal-Wallis H test resulted in a *p*-value of 1.289e-06. Thus, at $\alpha = 0.05$, a significant difference in the number of differentially expressed genes was observed between the Hispanic and non-Hispanic ethnic groups (refer to experiment B).

Finally, the Kruskal Wallis H test was performed to determine if there was a significant difference in the number of differentially expressed genes across the 16 gene sets across the two ethnic groups. The Kruskal-Wallis H test resulted in a *p*-value of 0.979 for $p < 0.01$, which indicated that there was no significant difference in the number of differentially expressed genes across the 16 different gene sets, that is, irrespective of landmark or non-landmark genes, both types of the genes are similar across ethnic groups in terms of the constitution of the number of differentially expressed genes (refer to experiment B).

The total number of differentially expressed genes across the 16 gene sets (the landmark set and 15 non-landmark sets) at $p < 0.1$ for the two ethnic groups, namely, Hispanic and non-Hispanic, are recorded in Table 8. Descriptive statistics across the ethnic groups indicated that relatively more genes were differentially expressed in the non-Hispanic group (mean ± SD, 977 ± 0.931) than in the Hispanic group (mean ± SD, 590 ± 107).

The Kruskal Wallis H test was performed to determine if there was a significant difference in the number of differentially expressed genes across the two ethnic groups. A significant difference was observed in the number of differentially expressed genes across the two ethnic groups at $\alpha = 0.05$. The Kruskal-Wallis nonparametric test resulted in a *p*-value of 1.109e-06. Thus, at $\alpha = 0.05$, a significant difference in the number of differentially expressed genes was observed between the Hispanic and non-Hispanic ethnic groups (refer to experiment B).

Table 7: The number of genes that were differentially expressed across the landmark and non-landmark gene sets for different ethnic groups at $p < 0.01$.

	$p < 0.01$	
Gene Sets	Hispanic	Non-Hispanic
L	503	978
NL set 1	479	978
NL set 2	501	976
NL set 3	434	977
NL set 4	419	978
NL set 5	297	978
NL set 6	268	976
NL set 7	214	974
NL set 8	273	977
NL set 9	432	978
NL set 10	269	971
NL set 11	278	974
NL set 12	454	978
NL set 13	368	973
NL set 14	310	975
NL set 15	275	973

Table 8: The number of genes that were differentially expressed across the landmark and non-landmark gene sets for different ethnic groups at $p < 0.1$

	$p < 0.1$	
Gene Sets	Hispanic	Non-Hispanic
L	752	978
NL set 1	703	978
NL set 2	720	977
NL set 3	682	977
NL set 4	638	978
NL set 5	582	978
NL set 6	471	978
NL set 7	410	976
NL set 8	477	977
NL set 9	661	978
NL set 10	503	976
NL set 11	530	977
NL set 12	700	978
NL set 13	605	975
NL set 14	533	978
NL set 15	476	977

Finally, the Kruskal Wallis H test was performed to determine if there was a significant difference in the number of differentially expressed genes across the 16 gene sets across the two ethnic groups. The Kruskal-Wallis H test resulted in a p-value of 0.992 for $p < 0.1$, which indicated that there was no significant difference in the number of differentially expressed genes across the 16 different gene sets, that is, irrespective of landmark or non-landmark genes, both types of the genes are similar across ethnic groups in terms of the constitution of the number of differentially expressed genes (refer to experiment B).

The total number of differentially expressed genes across the 16 gene sets (the landmark set and 15 non-landmark sets) at $p < 0.05$ for the 10 disease (cancer) types, namely, colon, brain, bladder, skin, breast, kidney, prostate, stomach, testis, and pancreas, are recorded in Table 9. Descriptive statistics across the disease type are recorded in Table 10.

The Kruskal Wallis H test was performed to determine if there was a significant difference in the number of differentially expressed genes across the 10 disease types. A significant difference was observed in the number of differentially expressed genes across the 10 disease types at $\alpha = 0.05$. The Kruskal-Wallis H test resulted in a p-value of <2.2e-16. Thus, at $\alpha = 0.05$, a significant difference in the number of differentially expressed genes was observed between the 10 different disease types (refer to experiment B).

Finally, the Kruskal Wallis H test was performed to determine if there was a significant difference in the number of differentially expressed genes across the 16 gene sets across the 10 disease types. The Kruskal-Wallis H test resulted in a p-value of 0.999 for $p < 0.1$, which indicated that there was no significant difference in the number of differentially expressed genes across the 16 different gene sets, that is, irrespective of landmark or non-landmark genes, both types of the genes were similar across the different disease types in terms of the constitution of the number of differentially expressed genes (refer to experiment B).

A correlation study was performed to differentiate the characteristics of the landmark and the non-landmark genes in the L1000 dataset (dataset 3). The pairwise correlation of the expression values of the 978 landmark genes across $\sim$129,000 tissue samples were compared against the expression values of the 978 non-landmark genes across 15 different randomly selected set of non-landmark genes across the $\sim$129,000 tissue samples. The results of the correlation study are demonstrated in Figure 5. The blue-colored line represents the range of the correlation values between a pair of genes in the landmark set. The remaining colored lines represent the range of the correlation values between the pair of genes in the different non-landmark gene sets. The range of the correlation values of the gene pairs within the landmark set were between [–0.8, –0.4] and [0.4, 0.8]. However, for the other sets of non-landmark genes, the range of correlation values between the pair of genes were almost similar, without any distinctive patterns (Figure 5).

Table 9: The number of genes that were differentially expressed across the landmark and non-landmark gene sets for different disease types at $p < 0.05$.

Gene Set	Colon	Brain	Bladder	Skin	Breast	Kidney	Prostate	Stomach	Testis	Pancreas
L	79	8	130	6	4	53	13	86	1	1
NL set 1	99	12	102	4	3	52	10	88	3	2
NL set 2	73	7	112	7	7	72	13	110	3	1
NL set 3	86	17	109	5	7	66	14	106	10	4
NL set 4	92	19	122	8	7	41	40	95	9	3
NL set 5	80	23	109	16	17	63	27	95	15	3
NL set 6	91	14	85	16	11	48	20	67	29	5
NL set 7	96	29	73	24	18	37	18	48	24	4
NL set 8	79	19	70	22	10	40	7	63	28	1
NL set 9	103	7	100	10	8	58	25	91	15	0
NL set 10	94	16	60	17	6	54	18	69	25	3
NL set 11	110	28	77	18	9	54	16	78	18	3
NL set 12	102	11	106	9	1	70	15	99	10	2
NL set 13	102	16	78	11	8	42	22	62	7	4
NL set 14	91	20	71	16	9	45	21	74	24	4
NL set 15	96	19	73	16	6	54	16	47	19	4

Table 10: Descriptive statistics for different disease types.

Disease Type	Mean ± Standard Deviation
Colon	92.1 ± 10.4
Brain	16.6 ± 6.69
Bladder	92.3 ± 21.3
Skin	12.8 ± 6.12
Breast	8.19 ± 8.19
Kidney	53.1 ± 10.7
Prostate gland	18.4 ± 7.78
Stomach	79.9 ± 19.4
Testis	15 ± 9.26
Pancreas	2.75 ± 1.44

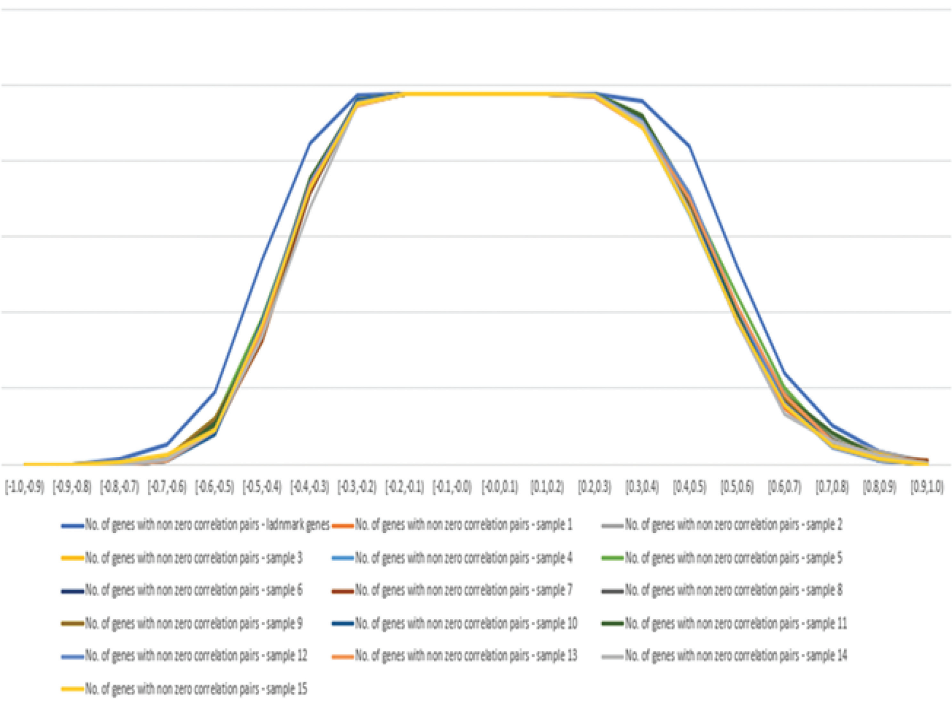

Figure 5: Pairwise correlation of genes in the landmark and non-landmark gene sets.

One-way ANOVA was performed on a dataset that contained randomly selected 100 correlation values between the pair of genes from both the landmark gene set and the 15 non-landmark gene sets. At $\alpha = 0.05$, one-way ANOVA resulted in a p-value of 0.999, which suggests not to reject the null hypothesis and conclude that there was no significant difference in the correlation values of the gene pairs across the 16 gene sets, that is, there was no evidence that the correlation values of the gene pairs in both the landmark set and the non-landmark sets are any different (refer to experiment C).

The boxplot of the correlation values of the gene pairs across the 16 gene sets are shown in Figure 6. The red-colored boxplot represents the landmark gene set, and the remaining colored boxplots represent the different non-landmark gene sets. The boxplots of the different gene set clearly highlights a slight variation in the correlation values of the gene pairs. However, there are no definitive patterns to clearly differentiate the correlation values of the gene pairs in both the landmark and non-landmark gene sets.

Based on all the experiments conducted so far, there were no observations that support the fact that landmark and non-landmark genes are different from each other.

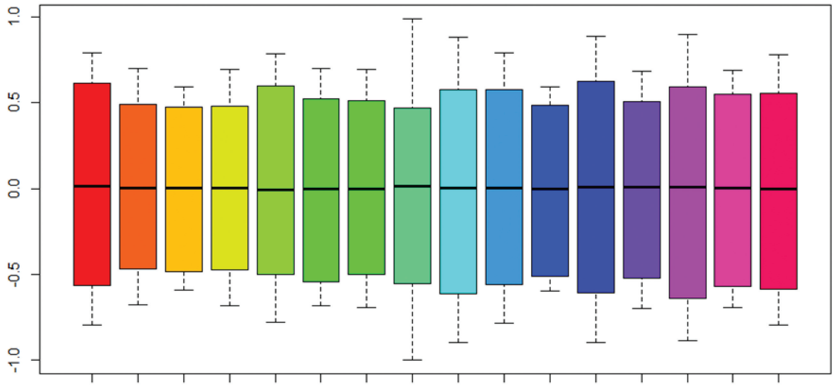

Figure 6: A boxplot of the correlation values of the gene pairs across the 16 gene sets.

5. Conclusion and Future Direction

This study aimed at understanding if there is any significant difference in the characteristics of the landmark and the non-landmark genes. Studies in the past (Duncan et al. 2008; Chen et al. 2016; Danaee et al. 2017; Ramaker et al. 2017; Bailey et al. 2018; Chen et al. 2018; Huang et al. 2018; Way et al. 2018; Daoud and Mayo, 2019; Clayman et al. 2020a) only focused on using the expression values of the landmark genes to determine the expression values of the non-landmark genes but did not discuss whether the landmark genes are any different from the non-landmark genes, that is, could we find a different set of non-landmark genes and say that they are similar in characteristics to the original set of identified landmark genes. The two experiments, namely the experiment A (see Figure 1) and experiment B (see Figure 2), indicated that there is no significant difference in the characteristics of the landmark and non-landmark genes. Across the landmark gene set and the 15 different randomly chosen non-landmark gene sets of similar size, no significant difference was observed in the number of differentially expressed genes across race, ethnicity, and disease types. On analyzing the correlation of the gene pairs within the landmark gene set and the 15 different randomly chosen non-landmark gene sets of similar size, it was observed that landmark gene pairs had slightly more range of correlation values compared with the other 15 sets of non-landmark gene pairs. However, the statistical test concluded that there was no evidence that the correlation values of the gene pairs in both the landmark gene set, and the non-landmark gene sets were any different (refer to experimental design C).

In this study, we only considered 16 sets of 978 genes, that is, one set of landmark genes identified in the work by Chen et al. (2016) and Clayman et al. (2020a, 2020b), and the 15 sets of randomly chosen genes labeled as non-landmark genes. The 15 sets, each contained 978 genes, were randomly chosen of the remaining ∼21,000 genes. Choosing a set of random 978 genes without repetition (non-landmark) of the remaining ∼21,000 genes is a complex combinatorial problem, and comparing all of those combinations against the established set of landmark genes is computationally intensive. Therefore, we chose a sample set of 15 equally sized non-landmark gene sets and used the non-parametric statistical tests to determine if there was any significant difference in the characteristics of the landmark and the non-landmark genes.

This study made a significant contribution to the field of personalized medicine. This field strives on the objective of determining the contributing genetic variables or the genes that are clinically relevant biomarkers for diseases. Large-scale availability of the gene expression profiling and clinical data related to carcinomas diseases provided us with the opportunity to explore and identify the significant variables (clinical and genetic) that could clearly characterize one or more diseases. At the same time, the advances in ML and AI techniques have made it possible to model the clinical and genetic variables to understand the relationships between these variables and the clinical outcomes. Even though both clinical and genetic variables are important to understand the clinical outcomes, the goal was to focus on identifying the genetic variables that can serve as clinically relevant biomarkers. This is because the clinical variables are mainly associated with the manifestations of the disease; that is, they are highly corelated with the disease types, and, also, it is a measurement that is captured after the fact. However, the genetic characteristics of an individual organism in a species or population, that is, genetic predisposition has a direct influence on disease development under the influence of environmental conditions.

Future studies can assess the potentiality of the linear combination or the principal components of the landmark and non-landmark gene clusters for disease-type predictions. This will be performed by implementing statistical, data mining, and ML techniques to extract patterns from the data as well as building predictive models. Effects of gene-

gene interactions for various types of cancer diseases should also be further assessed by using survival analysis, given that gene interactions are predictive of clinical outcomes. Various cancer types should be assessed to determine genes relevant to specific disease or cancer types.

Future studies may also build on this analysis by using predictive analytics techniques to further develop the understanding of how to investigate the relationship between genetic and clinical variables by accounting for both coding and noncoding genetic variants (Quang et al. 2015). This may be especially relevant toward applications for personalized medicine such as treatment responsiveness, depending on the combination of genetic and clinical variables. Future studies can assess whether clustering results based on gene expression levels can predict various disease types.

References

Ash, J. T., G. Darnell, D. Munro, and B. E. Engelhardt. 2018. "Joint Analysis of Gene Expression Levels and Histological Images Identifies Genes Associated with Tissue Morphology." Accessed March 12, 2025. https://doi.org/10.1101/458711

Azarkhalili, B., A. Saberi, H. Chitsaz, and A. Sharifi-Zarchi. 2018. "DeePathology: Deep Multi-Task Learning for Inferring Molecular Pathology from Cancer Transcriptome." Preprint, submitted August 2019. http://arxiv.org/abs/1808.02237

Bailey, M. H., C. Tokheim, E. Porta-Pardo, S. Sengupta, D. Bertrand, A. Weerasinghe, et al. 2018. "Comprehensive Characterization of Cancer Driver Genes and Mutations." *Cell* **173**, no. 2: 371–385. doi: 10.1016/j.cell.2018.02.060

Castro, E., C. Goh, D. Olmos, E. Saunders, D. Leongamornlert, M. Tymrakiewicz, et al. 2013. "Germline *BRCA* Mutations Are Associated with Higher Risk of Nodal Involvement, Distant Metastasis, and Poor Survival Outcomes in Prostate Cancer." *J Clin Oncol* **31**, no. 14: 1748–1757. doi: 10.1200/JCO.2012.43.188

Chaudhary, K., O. B. Poirion, L. Lu, and L. X. Garmire. 2018. "Deep Learning–Based Multi-Omics Integration Robustly Predicts Survival in Liver Cancer." *Clinical Cancer Research* **24**, no. 6: 1248–1259. doi: 10.1158/1078-0432.CCR-17-0853

Chen, X., J. Xie, and Q. Yuan. 2018. "A Method to Facilitate Cancer Detection and Type Classification from Gene Expression Data using a Deep Autoencoder and Neural Network." Accessed March 12, 2025. *ArXiv* 1812.08674 [Cs, Stat]. https://api.semanticscholar.org/CorpusID:56517070

Chen, Y., Y. Li, R. Narayan, A. Subramanian, and X. Xie. 2016. "Gene Expression Inference with Deep Learning." *Bioinformatics* **32**, no. 12: 1832–1839. doi: 10.1093/bioinformatics/btw074

Clayman, C. L., S. M. Srinivasan, R. S. Sangwan. 2020a. "Cancer Survival Analysis Using RNA Sequencing and Clinical Data." *Procedia Computer Science* **168**: 80–87. doi: 10.1016/j.procs.2020.02.261

Clayman, C. L., S. M. Srinivasan, R. S. Sangwan. 2020b. "K-Means Clustering and Principal Components Analysis of Microarray Data of L1000 Landmark Genes." *Procedia Computer Science* **168**: 97–104. doi: 10.1016/j.procs.2020.02.265

Danaee, P., R. Ghaeini, and D. A. Hendrix. 2017. "A Deep Learning Approach for Cancer Detection and Relevant Gene Identification." *Pacific Symposium on Biocomputing 2017*. Accessed March 12, 2025. https://doi.org/10.1142/9789813207813_0022

Daoud, M., and M. Mayo. 2019. "A Survey of Neural Network-based Cancer Prediction Models from Microarray Data." *Artificial Intelligence in Medicine* **92**: 204–214. doi: 10.1016/j.artmed.2019.01.006

Duan, Q., S. P. Reid, N. R. Clark, Z. Wang, N. F. Fernandez, A. D. Rouillard, et al. 2016. "L1000CDS^2: LINCS L1000 Characteristic Direction Signatures Search Engine." *NPJ Systems Biology and Applications* **2**: 16015. doi: 10.1038/npjsba.2016.15

Duncan, R., B. Carpenter, L. C. Main, C. Telfer, and G. I. Murray. 2008. "Characterisation and Protein Expression Profiling of Annexins in Colorectal Cancer." *British Iournal of Cancer* **98**, no. 2: 426–433. doi: 10.1038/sj.bjc.6604128

Dutil, F., J. P. Cohen, M. Weiss, G. Derevyanko, and Y. Bengio. 2018. "Towards Gene Expression Convolutions using Gene Interaction Graphs." International Conference on Machine Learning Workshop on Computational Biology, 2018. Preprint, submitted Jun 18. Accessed March 12, 2025. http://arxiv.org/abs/1806.06975

Ecke T. H., H. H. Schlechte, K. Schiemenz, M. D. Sachs, S. V. Lenk, B. D. Rudolph, et al. 2010. "TP53 Gene Mutations in Prostate Cancer Progression." *Anticancer Research* **30**, no. 5: 1579–1586.

Fielden, M. R., and T. R. Zacharewski. 2001. "Challenges and Limitations of Gene Expression Profiling in Mechanistic and Predictive Toxicology." *Toxicological Sciences* **60**, no. 1: 6–10. doi: 10.1093/toxsci/60.1.6

Huang, S., N. Cai, P. P. Pacheco, S. Narrandes, Y. Wang, and W. Xu. 2018. "Applications of Support Vector Machine (SVM) Learning in Cancer Genomics." *Cancer Genomics & Proteomics* **15**, no. 1: 41–51. doi: 10.21873/cgp.20063

Hurd, P. J., and C. J. Nelson. 2009. "Advantages of Next-Generation Sequencing Versus the Microarray in Epigenetic Research." *Briefings in Functional Genomic & Proteomics* **8**, no. 3: 174–183. doi: 10.1093/bfgp/elp013

Koch, C. M., S. F. Chiu, M. Akbarpour, A. Bharat, K. M. Ridge, E. T. Bartom, et al. 2018. "A Beginner's Guide to Analysis of RNA Sequencing Data." *American Journal of Respiratory Cell and Molecular Biology* **59**, no. 2: 145–157. doi: 10.1165/rcmb.2017-0430TR

Kogelman, L. J. A., and H. N. Kadarmideen. 2014. “Weighted Interaction SNP Hub (WISH) Network Method for Building Genetic Networks for Complex Diseases and Traits Using Whole Genome Genotype Data.” *BMC Systems Biology* **8**, no. Suppl 2: S5. doi: 10.1186/1752-0509-8-S2-S5

Kourou, K., T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis. 2015. “Machine Learning Applications in Cancer Prognosis and Prediction.” *Computational and Structural Biotechnology Journal* **13**: 8–17. doi: 10.1016/j.csbj.2014.11.005

Kursa, M. B., and W. R. Rudnicki. 2010. “Feature Selection with the Boruta Package.” *Journal of Statistical Software* **36**, no. 11: 1–13. doi: 10.18637/jss.v036.i11

Liang, M., Z. Li, T. Chen, and J. Zeng. 2015. “Integrative Data Analysis of Multi-Platform Cancer Data with a Multimodal Deep Learning Approach.” *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **12**, no. 4: 928–937. doi: 10.1109/TCBB.2014.2377729

Lim, S., S. Lee, I. Jung, S. Rhee, and S. Kim. 2020. “Comprehensive and Critical Evaluation of Individualized Pathway Activity Measurement Tools on Pan-Cancer Data.” *Briefings in Bioinformatics* **21**, no. 1, 36–46. doi: 10.1093/bib/bby097

Lin, M., V. Jaitly, I. Wang, Z. Hu, L. Chen, M. A. Wahed, et al. 2018. “Application of Deep Learning on Predicting Prognosis of Acute Myeloid Leukemia with Cytogenetics, Age, and Mutations.” Preprint, submitted Oct 30. Accessed March 12, 2025. https://arxiv.org/abs/1810.13247

Love, M. I., W. Huber, and S. Anders. 2014. “Moderated Estimation of Fold Change and Dispersion for RNA-seq Data with DESeq2.” *Genome Biology* **15**, no. 550: 1–21. doi: 10.1186/s13059-014-0550-8

Malta, T. M., A. Sokolov, A. J. Gentles, T. Burzykowski, L. Poisson, J. N. Weinstein, et al. 2018. “Machine Learning Identifies Stemness Features Associated with Oncogenic Dedifferentiation.” *Cell* **173**, no. 2: 338–354. doi: 10.1016/j.cell.2018.03.034

Nahm, F. S. 2016. “Nonparametric Statistical Tests for the Continuous Data: The Basic Concept and the Practical Use.” *Korean Journal of Anesthesiology* **69**, no. 1: 8–14. doi: 10.4097/kjae.2016.69.1.8

Parikh, N., S. Hilsenbeck, C. J. Creighton, T. Dayaram, R. Shuck, E. Shinbrot, et al. 2014. “Effects of *TP53* Mutational Status on Gene Expression Patterns Across 10 Human Cancer Types.” *The Journal of Pathology* **232**, no. 5: 522–533. doi: 10.1002/path.4321

Petralia, F., W.-M. Song, Z. Tu, and P. Wang. 2016. “New Method for Joint Network Analysis Reveals Common and Different Coexpression Patterns among Genes and Proteins in Breast Cancer.” *Journal of Proteome Research* **15**, no. 3: 743–754. doi: 10.1021/acs.jproteome.5b00925

Quang, D., Y. Chen, and X. Xie. 2015. “DANN: A Deep Learning Approach for Annotating the Pathogenicity of Genetic Variants.” *Bioinformatics* **31**, no. 5: 761–763. doi: 10.1093/bioinformatics/btu703

Ramaker, R. C., B. N. Lasseigne, A. A. Hardigan, L. Palacio, D. S. Gunther, R. M. Myers, et al. 2017. “RNA Sequencing-Based Cell Proliferation Analysis Across 19 Cancers Identifies a Subset of Proliferation-Informative Cancers with a Common Survival Signature.” *Oncotarget* **8**, no. 24: 38668–38681. doi: 10.18632/oncotarget.16961

Saltz, J., R. Gupta, L. Hou, T. Kurc, P. Singh, V. Nguyen, et al. 2018. “Spatial Organization and Molecular Correlation of Tumor-Infiltrating Lymphocytes Using Deep Learning on Pathology Images.” *Cell Reports*, **23**, no. 1: 181–193. doi: 10.1016/j.celrep.2018.03.086

Sawyer, S. F. 2009. “Analysis of Variance: The Fundamental Concepts.” *Journal of Manual & Manpulative Therapy* **17**, no. 2: 27–38. doi: 10.1179/jmt.2009.17.2.27E

Tomczak, K., P. Czerwińska, and M. Wiznerowicz. 2015. “The Cancer Genome Atlas (TCGA): An Immeasurable Source of Knowledge.” *Contemporary Oncology (Pozn)* **19**, no. 1A: A68–A77. doi: 10.5114/wo.2014.47136

Tsagri, M., Z. Papadovasilakis, K. Lakiotaki, and I. Tsamardinos. 2018. “Efficient Feature Selection on Gene Expression Data: Which Algorithm To Use? Accessed March 12, 2025. https://www.biorxiv.org/content/10.1101/431734v1

Way, G. P., F. Sanchez-Vega, K. La, J. Armenia, W. K. Chatila, A. Luna, et al. 2018. “Machine Learning Detects Pan-Cancer Ras Pathway Activation in the Cancer Genome Atlas.” *Cell Reports* **23**, no. 1: 172–180. doi: 10.1016/j.celrep.2018.03.046

Way, G. P., M. Zietz, V. Rubinetti, D. S. Himmelstein, and C. S. Greene. 2019. “Sequential Compression of Gene Expression Across Dimensionalities and Methods Reveals no Single Best Method or Dimensionality.” Accessed March 12, 2025. https://www.biorxiv.org/content/10.1101/573782v2.full.pdf+html

Zhang, W., Y. Yu, F. Hertwig, J. Thierry-Mieg, W. Zhang, D. Thierry-Mieg, et al. 2015. “Comparison of RNA-Seq and Microarray-Based Models for Clinical Endpoint Prediction.” *Genome Biology* **16**: 133. doi: 10.1186/s13059-015-0694-1

Zhao, B., A. Erwin, and B. Xue. 2018. “How Many Differentially Expressed Genes: A Perspective from the Comparison of Genotypic and Phenotypic Distances.” *Genomics* **110**, no. 1: 67–73. doi: 10.1016/j.ygeno.2017.08.007