1-1-2017

Integration Of Mutation And Gene Expression Data To Identify Disease Subtypes

Sahar Ansari
Wayne State University,

Integration of mutation and gene expression data to identify disease subtypes

by

SAHAR ANSARI

Master's Thesis

Submitted to the Graduate School

of Wayne State University,

Detroit, Michigan

in partial fulfillment of the

requirements for the degree of

Master of Science

2017

MAJOR: COMPUTER SCIENCE

Approved By:

7/5/2017

Advisor Date

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Abstract

Understanding the biological insights hidden in the vast amount of data collected, while investigating a disease, is the main goal for collecting such data in the first place. Changes in the gene expression or the function of proteins are important components in progression of a disease and is a key to understanding the disease mechanism. However, more often than not, the causes of such changes are not easily identified. In many cases, genetic variants may cause some of the observed gene expression changes. In this thesis, we focus on identifying the variants that significantly alter gene expression for an individual by integrating genetic variant data, gene expression data, as well as a priori knowledge about gene-gene interaction networks from multiple databases. Here we show that one can use variants that change gene expression to identify subgroups of patients with significantly different survival profiles. The method is validated on four different cancer types (renal, lung, colorectal cancer and leukemia) from the TCGA database. The results show that this method is able to identify variants that significantly affect the gene expression (and in turn the phenotype), as well as identify disease sub-types that are biologically meaningful as validated by survival and pathway analysis.

Chapter 1

INTRODUCTION

Understanding the biological insights hidden in the vast amount of data collected, while investigating a disease, is the main goal for collecting such data in the first place. The advent of microarrays and more recently next generation sequencing make it much easier to collect different types of information from a variety of angles about the same sample at the same time. Examples of such information include: DNA changes such as single point mutations or copy number variations, DNA methylation, alternative splicing, microRNA expression, post-translational modifications, etc. Even though we have the ability to collect such rich data, analyzing it in order to completely understand the investigated phenotype is still an open challenge.

Gene expression, as the most common type of data collected, can identify the important changes between a disease state and a normal state. However, more often than not, the causes of such changes are not easy to identify. Other types of data, such as single nucleotide polymorphism (SNP), methylation, and copy number variation (CNV) can complement the gene expression data [16, 18, 38, 74]. It is now accepted that the changes in the system are not likely to be captured completely in any one type of data [39, 45]. This is particularly true for complex diseases, such as cancer, which involve many phenomena that affect many levels [64, 91]. More information can be obtained if different types of data were analyzed together, thus the integration of multiple types of data has become a very important problem to solve [78, 79, 80, 81, 82, 83, 84].

Data integration methods can be divided into two main categories: multi-stage analysis and meta-dimensional analysis [64]. In multi-staged methods, different types of data are

integrated one after another in sequential steps [1, 10, 65, 69, 90]. In meta-dimensional methods all types of data are integrated simultaneously to model the complex phenotype [24, 61]. The analysis results, obtained from both categories, have fewer false positives compared to using only one data type, especially if the different data types cover different levels of regulation in the system (e.g. genetic, genomic, proteomic etc). The method introduced in this thesis, belongs to the meta-dimensional category since it simultaneously integrates variants and gene expression data to identify disease subtypes.

Different approaches have been proposed to identify variants that cause changes in gene expression levels [26, 85, 93]. Many of the proposed approaches, mentioned in [85], also integrate other types of information along with mutation and gene expression (e.g. copy number variation (CNV) data, methylation data) to predict the effect of presence of each variant on the expression of its host gene [5, 6, 13, 19, 85]. Furthermore, some methods use a priori knowledge about the network of interactions between genes because a given variant may change the expression of a set of genes rather than a single gene [3, 23]. In this thesis, we obtained gene-gene interaction data from protein-protein interaction databases as well as signaling pathways to investigate the effect of a variant on a subnetwork of genes.

Analyzing different types of data made it clear that the characteristics and progression of different diseases are the results of the interaction between the disease and the immune system of the host [12, 70, 97]. Because of this, different patients may respond differently to the same drug. Therefore, identifying different subtypes of a given phenotype is extremely important in selecting the most appropriate drug[35, 40].

We take advantage of data available in The Cancer Genome Atlas (TCGA) database [77]. TCGA is a collaboration between the National Cancer Institute (NCI) and National Human Genome Research Institute (NHGRI), that includes data for 33 types of cancer covering many dimensions of the genomic changes. This data includes 11,000 patients and is available for research purposes.

The summary of data available in TCGA database is shown in Figure 1.1.

Disease Name	Cohort	Cases
Adrenocortical carcinoma	ACC	92
Bladder urothelial carcinoma	BLCA	412
Breast invasive carcinoma	BRCA	1098
Cervical and endocervical cancers	CESC	307
Cholangiocarcinoma	CHOL	51
Colon adenocarcinoma	COAD	460
Colorectal adenocarcinoma	COADREAD	631
Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	DLBC	58
Esophageal carcinoma	ESCA	185
FFPE Pilot Phase II	FPPP	38
Glioblastoma multiforme	GBM	613
Glioma	GBMLGG	1129
Head and Neck squamous cell carcinoma	HNSC	528
Kidney Chromophobe	KICH	113
Pan-kidney cohort (KICH+KIRC+KIRP)	KIPAN	973
Kidney renal clear cell carcinoma	KIRC	537
Kidney renal papillary cell carcinoma	KIRP	323
Acute Myeloid Leukemia	LAML	200
Brain Lower Grade Glioma	LGG	516
Liver hepatocellular carcinoma	LIHC	377
Lung adenocarcinoma	LUAD	585
Lung squamous cell carcinoma	LUSC	504
Mesothelioma	MESO	87
Ovarian serous cystadenocarcinoma	OV	602
Pancreatic adenocarcinoma	PAAD	185
Pheochromocytoma and Paraganglioma	PCPG	179
Prostate adenocarcinoma	PRAD	499
Rectum adenocarcinoma	READ	171
Sarcoma	SARC	261
Skin Cutaneous Melanoma	SKCM	470
Stomach adenocarcinoma	STAD	443
Stomach and Esophageal carcinoma	STES	628
Testicular Germ Cell Tumors	TGCT	150
Thyroid carcinoma	THCA	503
Thymoma	THYM	124
Uterine Corpus Endometrial Carcinoma	UCEC	560
Uterine Carcinosarcoma	UCS	57
Uveal Melanoma	UVM	80

Figure 1.1: Summary of different cancer types available in TCGA database together with the number of patients included [77].

In this thesis, we focus on identifying the variants (with high or low frequencies) that alter gene expression significantly for each individual. This is done by integrating mutation and gene expression data, as well as a priori knowledge about gene-gene interaction networks from multiple databases. We introduce an algorithm that divides the samples of one cancer type into two subtypes by focusing on the variants that cause significant changes in gene expression between the samples with and without those variants.

The groups of samples identified are validated by comparing the survival data for the two groups. The survival analysis is widely used to validate the identified subtypes in a set of patients [25, 37]. Survival analysis uses the measurements of the last follow-up time from the beginning of treatment or diagnosis of a disease to time of death. Statistical tests are performed to distinguish if there is any significant difference between the survival curves of patients in those subtypes. Once the subtypes are identified, we investigate the difference between them from a biological pathways context. The goal of pathway analysis methods is to identify the most perturbed pathways in a given condition. Pathways are divided in two main categories: i) signaling pathways, that are defined as graphs in which nodes represent genes/proteins and edges are interactions between them, and ii) metabolic pathways in which the nodes represent biochemical compounds and the edges represent reactions, carried out by enzymes which are coded by genes [49]. A pathway describes all the known phenomena involved in a given biological process, to which it is associated. It is part of a larger system that has a set of components interacting with each other. These components work together to achieve a common goal. The name of a pathway usually represents the biological process, phenomenon, or disease process described by the pathway. Different types of interactions are described by different types of edges, or weights in the structure of a pathway. As an example, Figure 1.2 shows the apoptosis signaling pathway, which includes the genes and interactions involved in the known mechanism for cell death. Different types of interactions are shown by different arrows.

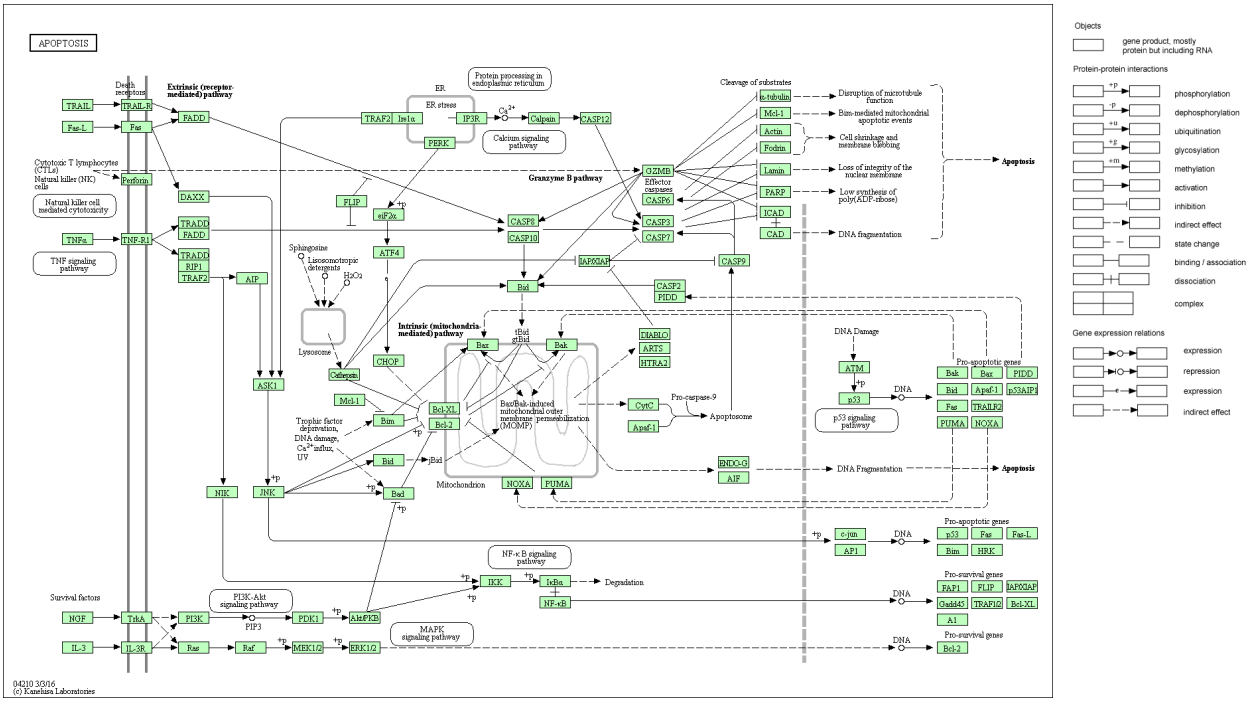


Figure 1.2: KEGG apoptosis signaling pathway (from: `http://www.genome.jp/kegg/pathway/hsa/hsa04210.html`).

Such pathways describe all known phenomena involved in a biological process (e.g. cell cycle), or a disease (e.g. Alzheimer's disease), etc. In this thesis, we focus on signaling pathways to be able to map the measured expression level of the genes to the corresponding nodes in those pathways. Intuitively, the impact of a given phenotype on a given pathway should be determined by the number of differentially expressed (DE) genes on that pathway, the magnitude of the changes in the expression level of the genes, and the type, direction and strength of the interactions between the genes in that pathway.

The pathway databases, like KEGG [55], BioCarta [9] and Reactome [29], provide complex graphs for each signaling pathway, in which each node is a gene/protein and each edge is an interaction between two such genes or proteins. There are many methods that have been proposed that are able to fully take into consideration all the interactions between genes in signaling pathways to find which pathway is most impacted by a given phenotype [14]. These are sometimes referred to as "topology-aware" or "third generation" pathway analysis methods [36, 49]. The method, used in this manuscript, is Impact Analysis, which belongs

to this latest generation of pathway analysis methods, inasmuch as it considers the topology of the pathways, as well as the changes in expression level of the genes [14]. The results of pathways analysis can lead to better understanding of the mechanisms that cause the differences between subtypes.

In summary, this thesis presents a novel method for identifying subtypes of a disease by integrating variant and gene expression data. The method was validated on four different cancer types (renal, lung, colorectal cancer and leukemia) from the TCGA database. The survival analysis shows significant differences between the identified subtypes in all the investigated diseases. To further understand the identified subtypes, we performed a pathway analysis on the identified subtypes. We observed that the pathways that are significantly perturbed between the subtypes are strongly known to be associated to each disease.

Chapter 2

Methods

In this section, we propose a novel method that integrates gene expression and genetic variant data to sub-type patients diagnosed with the same disease. The goal is to identify those sub-groups that might share a common mechanism in order to be able to develop novel drugs specific to each sub-group. First, we will introduce the method that integrates variant and gene expression data to identify the subtypes. Then, we will describe the assessment methods, survival and pathway analysis, used to quantify the quality of the subtypes identified.

2.1 Subtype identification by integrating mutation and gene expression data

The method proposed here aims to detect variants that significantly affect gene expression and use these variants to define novel sub-groups of patients. We start by focusing on the gene in which each variant occurs. Henceforth, we will refer to this as the host gene. We use gene expression data collected using either microarrays or RNA-Seq to identify if there is an expression change between samples that have and do not have a given variant. However, the expression of each gene evolves with time and because we collect data at a given point in time, we might not get the full picture. To overcome this limitation, we decided to enhance our search space and not only consider the effect of a variant on its host gene, but also the effect of the variant on the genes that interact with its host gene. Hence, we use a priori knowledge about gene-gene interactions obtained from different protein-protein interaction networks and available signaling pathways. We acquired the information from

Human Protein database Reference (HPRD) (Release 9) [57, 58], BioGRID [73] and Kyoto Encyclopedia of Genes and Genomes (KEGG) (Release 72.0) [30, 31, 32]. Using information coming from these databases, we construct a "neighborhood" of each host gene that includes all the genes that are directly connected to it in any of the considered interaction networks, as well as the host gene itself.

For each dataset, we start by building a *variant matrix*, in which the rows represent the existing variants in all the given samples, and columns represent the samples:

$$VAR = \begin{pmatrix} var_{11} & var_{12} & ... & var_{1p} \\ var_{21} & var_{22} & ... & var_{2p} \\ ... & ... & ... & ... \\ var_{m1} & var_{m2} & ... & var_{mp} \end{pmatrix}$$

where: $var_{ij} = 1$ if var_i occurs in $sample_j$, 0 otherwise. In this manuscript a variant refers to a combination of a specific allele at a given position.

We then construct a *gene expression matrix*, in which the rows represent the host genes for any variant in the *variant matrix* and their neighbors, and columns represent the samples:

$$EXP = \begin{pmatrix} exp_{11*} & exp_{12*} & ... & exp_{1p*} \\ exp_{21*} & exp_{22*} & ... & exp_{2p*} \\ ... & ... & ... & ... \\ exp_{m1*} & exp_{q2*} & ... & exp_{qp*} \end{pmatrix}$$

Here, exp_{ij*} is a vector including the expressions of the host genes and all the genes in the neighborhood of the host gene for variant var_i in $sample_j$ (see, Figure 2.1).

For all the variants that are present in a sample (all variants marked with red in Figure 2.1), we compute a z-score for every gene expression associated with the variant

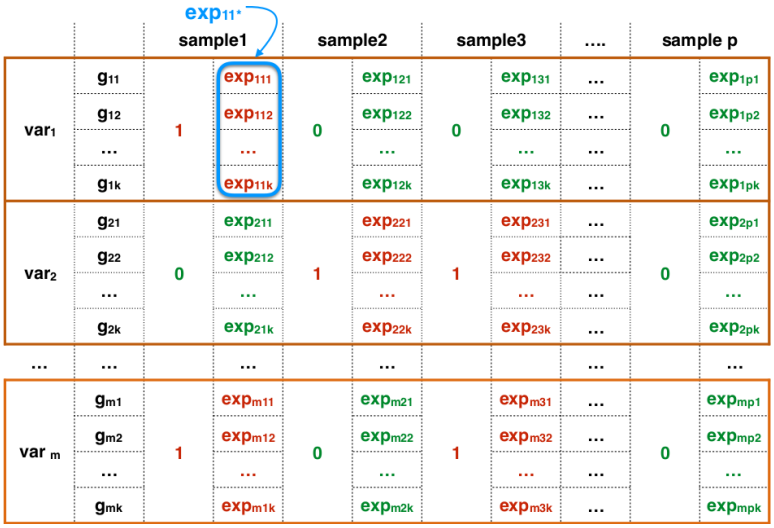


Figure 2.1: Expression values grouped by presence of variants: For each variant, we consider the set of genes that belong to the neighborhood of the host genes. If the variant is present in the given sample (marked with red), then we compute a z-score for each gene and its neighborhood against all expressions in that row that are not associated to the presence of the variant (marked with green). The values in the blue box represent exp_{11*} from the EXP matrix.

association against all the expressions corresponding to samples not exhibiting the variant:

$$Z = \begin{pmatrix} z_{11*} & z_{12*} & \dots & z_{1p*} \\ z_{21*} & z_{22*} & \dots & z_{2p*} \\ \dots & \dots & \dots & \dots \\ z_{m1*} & z_{q2*} & \dots & z_{qp*} \end{pmatrix}$$

where: z_{ij*} is a vector including the z-scores of the host gene and all the genes in the neighborhood of the host gene for variant var_i calculated for $sample_j$ as follows:

$$z_{ijk} = \frac{exp_{ijk} - mean(exp_{iMk})}{sd(exp_{iMk})} \tag{2.1}$$

The z_{ijk} and exp_{ijk} are the z-scores and expressions for the association between variant i and the neighborhood of its host gene (g_{i1}, g_{i2}, ..., g_{ik}) in $sample_j$, $var_{ij} = 1$, where $var_{iM} = 0$; $M \in [1, p]$ (see Figure 2.1 and 2.2).

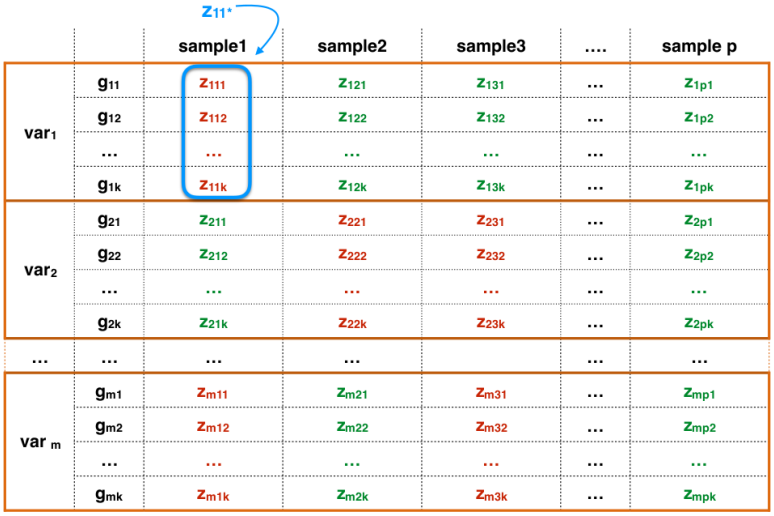


Figure 2.2: Selecting significant variants based on z-score: If a variant is present in a sample, we compute a z-score for every gene in its host gene's neighborhood (z-scores marked with red). We define a variant to be significant in a sample if any of the z-scores computed for a gene in the host gene's neighborhood is significant. For example, we declare variant var_1 to be significant in $sample_1$ if any of the z-scores in the blue box are significant (z_{11*} from Z matrix).

We define a variant to be significant in one sample if any of the genes in the host gene's neighborhood is significant at 1% in a two-tail testing framework (see Figure 2.3). Hence, the resulting matrix will include the *significant variants* in each sample:

$$SIG\ VAR = \begin{pmatrix} sig.var_{11} & sig.var_{12} & ... & sig.var_{1p} \\ sig.var_{21} & sig.var_{22} & ... & sig.var_{2p} \\ ... & ... & ... & ... \\ sig.var_{m1} & sig.var_{m2} & ... & sig.var_{mp} \end{pmatrix}$$

where: $sig.var_{ij} = 1$ if var_i is significant in $sample_j$ (i.e., if p-values associated to any of the z-scores z_{ij*} are less than 0.5% for either tail), 0 otherwise. Based on this *significant variant* matrix, we define two subgroups based on the count of significant variants in each sample. We define the high risk group as the one including the samples with larger number of significant variants (more than the median) and the low risk group as the one including the remaining samples (fewer variants than the median). We choose this separation with the

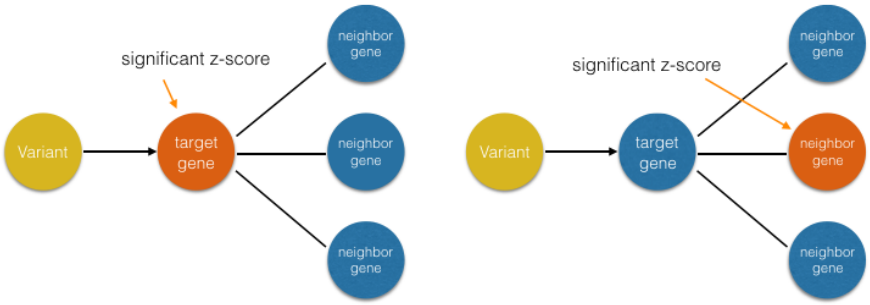


Figure 2.3: The two possible situations for a variant in a sample to be declared as significant: The yellow nodes represent the variants identified in the sample. The blue nodes are the host genes of the given variant and its neighbors. The orange nodes are the genes with significant z-scores.

expectation that a large number of variants will produce a higher disruption and therefore a lower survival rate.

2.2 Survival Analysis

Once we identified the subtypes, our goal is to assess if there is any difference in the survival rates between the two groups. We use the well known Kaplan-Meier [33] estimator to plot the survival curves based on lifetime data obtained from TCGA. The survival function, S(t), is the probability of patients in a particular group to survive at the given time (t). The probability of surviving in each period of time is a function of number patients alive at the beginning of the period and the number of deaths in that period of time. The probability of a patient to survive k or more periods depends on the survival rates in all the previous periods [7].

$$S(t) = p_i \cdot p_2 \cdot p_3 \cdot ... \cdot p_k \tag{2.2}$$

where:

$$p_i = \frac{\text{number of alive patients at the beginning of period i - number of death in period i}}{\text{number of alive patients at the beginning of period i}} \tag{2.3}$$

Patient number	Survival time (days)	Number known to be alive (r_i)	Deaths (d_i)	Proportion surviving (p_i)	Cumulative proportion surviving (S [t])
	0				1
1	1	8			
2	1	8	2	(8 - 2)/8 = 0.750	1 × 0.750 = 0.750
3	4	6	1	(6 - 1)/6 = 0.833	0.750 × 0.833 = 0.625
4	5	5	1	(5 - 1)/5 = 0.800	0.625 × 0.800 = 0.500
5	6+				
7	9	3	1	(3 - 1)/3 = 0.667	0.500 × 0.667 = 0.333
8	9+				
11	22	1	1	(1 - 1)/1 = 0.00	0.333 × 0.00 = 0.000

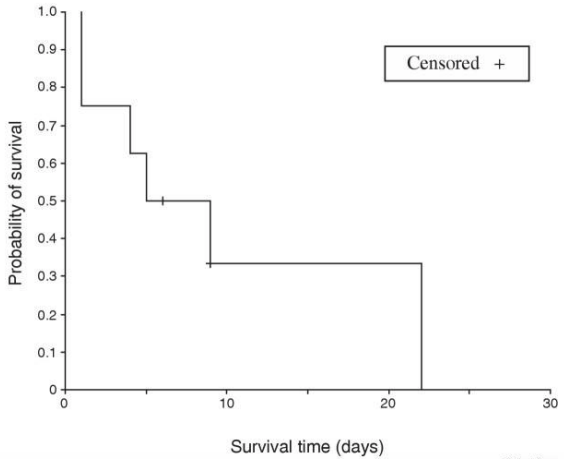


Figure 2.4: An example of survival table available for a group of patients [7]. The table shows the number of patients alive and number of death for each period of time. The probability of survival is calculated for each period based on the survival rates in all the previous times. The curve shows the probability calculated for each period. The censored patients are the ones for whom the death event is not available due to the fact that they dropped out of the study and the available data only shows their last follow-ups.

The method is explained with an example provided by [7] shown in Figure 2.4. We used the R package *survminer* for plotting the survival curves of the two identified subtypes [34].

In addition, we compute the Cox p-value to assess if there is a difference between the survival rates of the two groups.

2.3 Pathway Analysis

The subtypes identified are also validated using pathway analysis. We obtained all the signaling pathways defined in the KEGG database [30, 31, 32]. We choose the Impact Analysis (IA) [14] since it is the most widely used topology-based pathway analysis method. This method takes into consideration the type and position of each gene, the magnitude of expression change for each gene and the types of interactions between them. This was the first

proposed method that includes all the mentioned factors. Previous methods, only considered the pathways as a set of genes and ignored all the interaction information that was provided by the predefined pathways. We are using an extended version of IA that takes as input the entire list of genes and it does not need a subset of genes as differentially expressed [88]. We refer to this method as IA-all genes. The IA-all genes gives the opportunity to use the entire set of measurements provided rather than just selecting a subset of them as differentially expressed. Typically, the selection of DE genes will cause loss of data for more than 20,000-30,000 genes, and the DE selection will only have information for about 300 genes. It was shown that IA-all gene yields significantly better results compared to the classical IA [88]. In IA-all genes, the "perturbation factor" was calculated for all the genes in each pathway as follows:

$$PF(g_i) = \Delta E(g_i) + \sum_{u \in US_{g_i}} \frac{\beta_{ug_i} \cdot PF(u)}{N_{ds}(u)} \tag{2.4}$$

where ΔE is the measured gene expression of the gene, US is a set of all the genes that are upstream of the gene of interest (g_i) in the predefined pathways, N_{ds} is the number of downstream genes for each of the genes upstream of the gene g_i, and β_{ug_i} represent the type of interaction between gene u and g_i. In IA-all genes, $\beta_{ug_i} = 1$ if the type of interaction is activation, or activation like and $\beta_{ug_i} = -1$ if the type of interaction is inhibition, or inhibition like.

The score for pathway k is calculated as the sum of the absolute values of perturbation factors of all the genes in the pathway, $totalPF$:

$$totalPF_k = \sum_{i \in pathway_k} |PF(g_i)| \tag{2.5}$$

The quantity $totalPF$ of a pathway represents the amount of disruption of the whole pathway in the condition under study. The significance of each pathway is assessed by computing the probability of obtaining just by chance a $totalPF$ value more extreme than

the one observed. This probability is estimated using a bootstrap approach, where the null distribution for $totalPF$ for each pathway is generated by sampling random gene expression changes from the original set of expression changes. The number of bootstraps used was 2,000. This process is repeated for all pathways and yields a p-value for each pathway. Subsequently, the set of p-values for all pathways are corrected for multiple comparisons using the false discovery rate (FDR). Here, we used the ROntoTools version 2.0.0 , which is an implementation of IA-all genes as an R package available in Bioconductor [89].

Chapter 3

Results

3.1 Lung Cancer

Lung cancer is one of the leading cause of death among all types of cancer [92]. According to American Lung association (`http://www.lung.org/`), a very small number of patients (17.7 percent), diagnosed in very advanced stages, survive more than five years. Lung squamous cell carcinoma is a very common subtype of lung cancer that causes more than 400,000 deaths every year [79]. We use gene expression and mutation data for lung squamous cell carcinoma (LUSC) from TCGA database [77] to subtype the cohort of patients into two groups based on their molecular profile: short survival (more aggressive cancer) and long term survival (less aggressive cancer). There are 178 LUSC patients that have mutation profiles available, 154 LUSC patients that have gene expression profiles available, and 89 patients that have both. Since the proposed method focuses on the integration of gene expression with mutation data, we focus on the 89 patients that have both mutation and gene expression profiles available for further analysis.

We pre-process the mutation data to include the variants that might have affected the gene expression and in turn the protein expression, which presumably determined the phenotype. Out of the 29,565 variants that belong to at least one of the selected 89 patients, we eliminate the *silent* mutations as these are unlikely to influence the phenotype. The remaining 22,913 variants belong to 21 different predicted classifications (frame shift, missense, non-sense, insertion of a stop codon, etc.). The gene expression data is pre-processed to eliminate outliers, by considering only the genes that are expressed in more than 50% of the samples. This reduces the number of genes considered from 12,042 to 11,883.

To compute the neighborhood of each variant's host gene, we obtained gene-gene interaction information from KEGG [30, 31, 32], HPRD [57, 58] and BioGRID [73] databases. The integrated network from these three sources includes 9,693 genes and 37,193 interactions. Adding existing information about known interactions allows us to estimate the potential effect of a variant on groups of interacting genes.

For each variant in each sample, the z-scores for the host gene and its neighbors are calculated as explained in the Methods section. The z-scores higher than 2.6, corresponding to a p-value lower than 1%, are considered as significant. In each sample, we consider the variants that cause a significant change in expression (significant z-scores) of its host gene and/or its neighbors as significant variants (sig.vars). We divide the samples into two groups based on the number of significant variants. The first group includes the samples that have a number of significant variants higher than the median, while the second groups includes samples with a number of significant variants lower than or equal to the median. Our expectation is that the patients with higher number of significant variants will suffer of an increase disruption and therefore have a lower survival rate.

We validate the identified subtypes in two ways. First, we use the information from clinical data to perform a survival analysis. In this analysis, we compare the survival curves of patients in the identified subtypes. Survival analysis aims to identify the proportion of the population that survive during a given time [7]. To analyze the survival of the investigated patients, we used the clinical data provided by TCGA. The survival time associated to each patient is the time from beginning of their treatments until their deaths or their last follow-ups. The state of dead/alive for the patients are also extracted from the clinical data. As expected, the patients with higher number of significant variants exhibit a lower survival rate than those with fewer (see red curve in Figure 3.1). The Cox p-value representing the significance of difference between these two subtypes is 0.045. This significant p-value validates that the molecular signature of the patients in two subtypes changes their survival

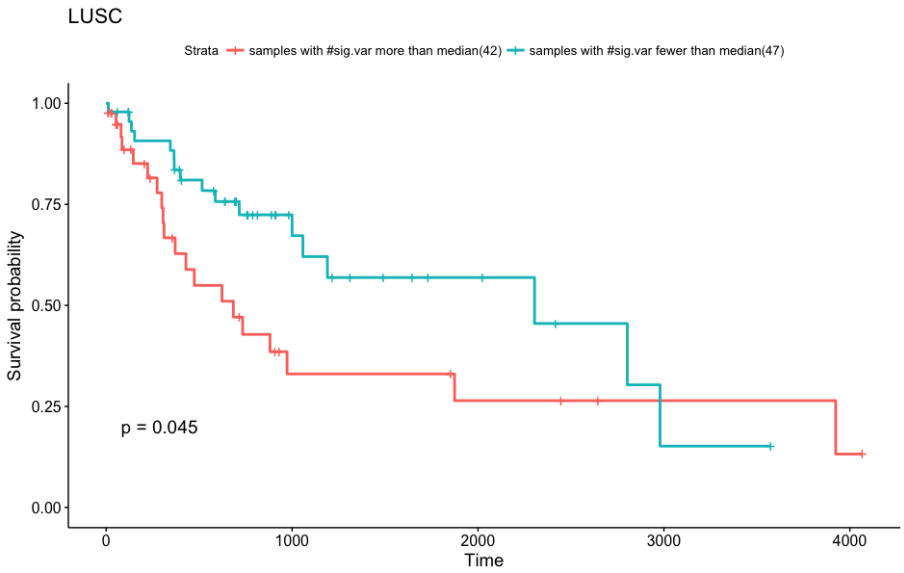


Figure 3.1: The survival curve of patients in the two identified subtypes in LUSC. The red curve represents the patients with higher number of significant variants and the blue curve represents the patients with a lower number of significant variants. As expected the patients with higher number of significant variants have a lower survival rate. The set of low survival patients contains 42 individuals while the other contains 47. The Cox p-value representing the significance of differences between two curves is 0.045.

rates. The subtypes could be investigated further to identify different reactions to different drugs and could warrant more aggressive treatments to patients in the low survival group.

Second, we validate the subtypes by performing pathway analysis comparing the patients in low versus high survival. Understanding the differences between the active mechanisms involved in each group helps to design better drugs and improve the knowledge of the investigated disease. Pathway analysis results identify the networks that describe the potential mechanisms that differentiate the two subtypes. The ranked list of pathways according to the Impact Analysis [14, 88] is shown in Table 3.1. The top significant pathway, *chemokine signaling pathway* has an important role in evolving many cancer diseases such as non-small lung cancer [11, 76]. Also, interestingly we identified *staphylococcus aureus infection* pathway as one of the top significant pathways. Staphylococcus aureus includes mechanisms involved in different types of infections such as superficial skin infections, food poisoning and life-threatening infections [32]. There are many studies that show the impact

	names	totalPertNorm	pPert	pPert.fdr	references
path:hsa04062	Chemokine signaling pathway	6.52049	0.00050	0.01099	[11, 76]
path:hsa05150	Staphylococcus aureus infection	5.79590	0.00050	0.01099	[60, 75]
path:hsa04512	ECM-receptor interaction	5.30600	0.00050	0.01099	[42, 47]
path:hsa05144	Malaria	5.25846	0.00050	0.01099	
path:hsa05323	Rheumatoid arthritis	5.15320	0.00050	0.01099	[27]
path:hsa04060	Cytokine-cytokine receptor interaction	5.06846	0.00050	0.01099	[87]
path:hsa05146	Amoebiasis	4.69088	0.00050	0.01099	
path:hsa04510	Focal adhesion	4.63666	0.00050	0.01099	[15, 42, 54]
path:hsa04080	Neuroactive ligand-receptor interaction	-3.92057	0.00050	0.01099	[43]
path:hsa04145	Phagosome	4.76257	0.00099	0.01522	

Table 3.1: The top 10 ranked pathways when comparing two subtypes from LUSC. The highlighted pathways have strong associations with lung cancer based on literature.

of infection on survival of patients with lung cancer [60, 75]. Multiple studies show strong association between the variants in *ECM-receptor interaction* pathway, which is one of the top three significant pathways and risk for non-small cell lung cancer (LUSC) [42, 47]. In summary, many of the top 10 significant pathways have known associations to lung cancer (see Table 3.1).

Furthermore, as an alternative method of grouping the patients, we compare significant variants with non-significant variants that share the same location in genome. In this data, there are 6 such variants. We grouped the samples that contain the significant variants in one group and samples containing the non-significant variants in the other. The list of the variants, their position in the genome, and reference alleles are shown in Table 3.2. The survival analysis shows that the survival rate of patients with significant variants is drastically affected when comparing with those without (Figure 3.2).

Based on the assessments performed, our method is able to identify variants that significantly affect the gene expression, and therefore the phenotype. Patients with significant variants, as defined by our method, generally have a lower survival rate.

host gene	chrom	pos	ref	sig.var	classification	not sig.var	classification
TP53	17	7577550	C	-	Frame Shift Del	T	Missense
TP53	17	7578177	C	G	Splice Site	T	Splice Site
TP53	17	7577538	C	T	Missense	A	Missense
TP53	17	7577538	C	G	Missense	A	Missense
TP53	17	7578458	G	-	Frame Shift Del	C	Missense
ARHGEF2	1	155921325	C	T	Splice Site	A	Splice Site

Table 3.2: Significant variants that share a position with non-significant variants in LUSC: For each significant variant (red), we show the host gene and the associated non-significant variant (blue). For each variant, we include the associated classification. Notice that even variants with same classification have different effect on gene expression.

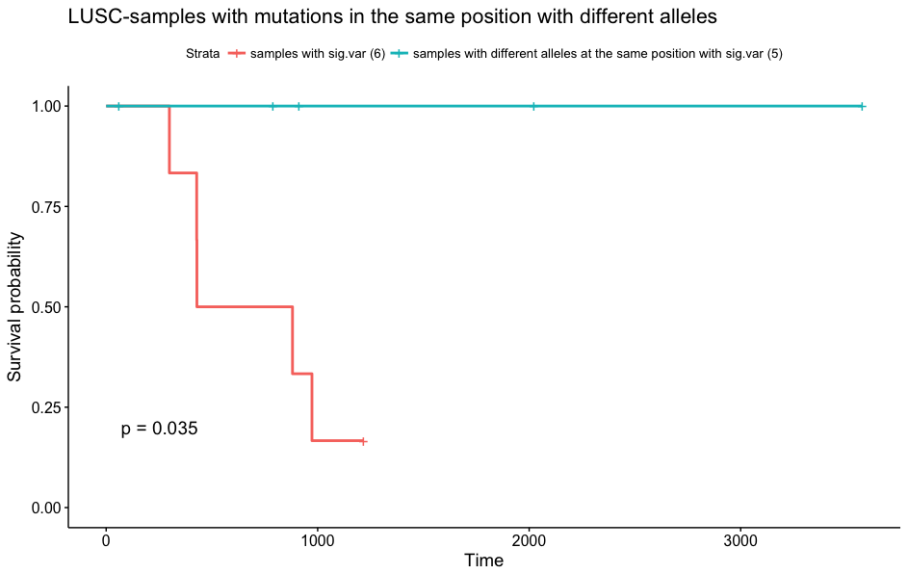


Figure 3.2: The survival curve of patients with significant variants (red) compared to patients that have a non-significant variants (blue) at the same location in LUSC. The Cox p-value (0.035) shows there is a significant difference between the survival rates.

3.2 Kidney Cancer

We analyzed data that come from experiments studying renal cell carcinoma (KIRC), which is the most common type of kidney cancer [44]. This type of cancer can easily spread to other organs such as lungs. In most of the patients, the cancer has already spread when the cancer is diagnosed and the survival of the patients highly depends on the spread of the cancer. We downloaded the gene expression and variant data from TCGA database. The gene expression data includes measurements for 533 patients and the variant data includes profiles of 417 patients. We focus on 324 patients for whom both types of data are available. The pre-processing of variant data by eliminating silent variants and variants that do not appear in the 324 patients resulted in 40,211 variants.

The gene expression data is pre-processed as well by removing any gene that did not have measurements in at least half the samples. This resulted in 13,002 genes to be used further in the analysis. The same gene-gene interactions downloaded from KEGG, HPRD and BioGRID are used here, as well.

After the subtyping, the group with expected low survival (i.e., number of significant variants more than the median) exhibited an even more significant difference than the lung cancer comparison. The Cox p-value representing the significance of difference between two curves is 0.0067 (see Figure 3.3).

The top significant pathways resulted from the pathway analysis are known to be associated to kidney cancer, which further validates the identified subtypes (see Table 3.3). The *rap1 signaling pathway*, which is the top significant pathway found by our method, is also identified as the most perturbed pathway by independent researchers [51]. This is a validation of the involvement of this pathway in progression of kidney cancer. Most of the top pathways include mechanisms that have the potential to explain different survival of samples in kidney cancer.

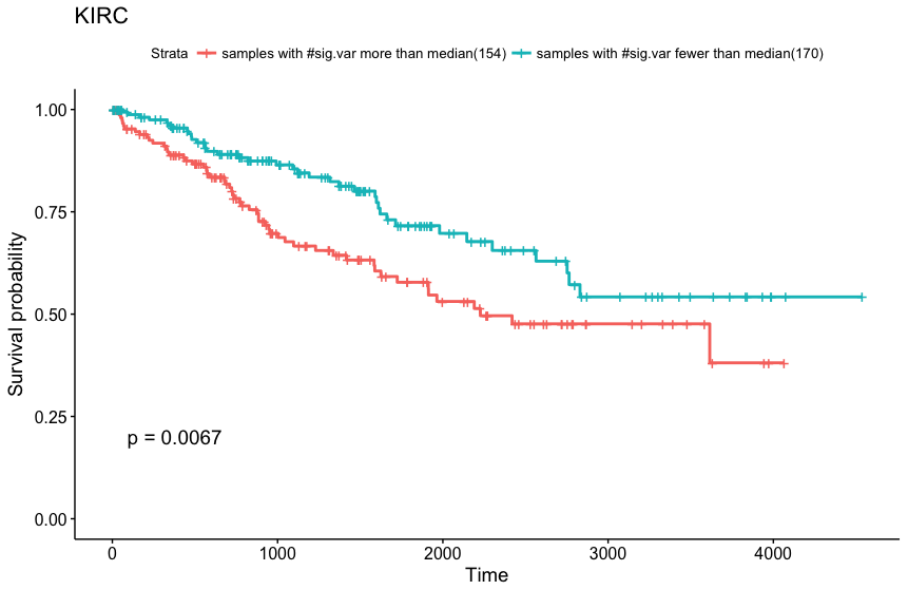


Figure 3.3: The survival curve of patients in the two identified subtypes in KIRC. The red curve represents the patients with higher number of significant variants and the blue curve represents the patients with a lower number of significant variants. As expected the patients with higher number of significant variants have a lower survival rate. The set of low survival patients contains 154 individuals while the other contains 170. The Cox p-value representing the significance of differences between two curves is 0.0067.

	names	totalPertNorm	pPert	pPert.fdr	
path:hsa04015	Rap1 signaling pathway	8.77739	0.00050	0.00275	[51]
path:hsa04151	PI3K-Akt signaling pathway	8.54322	0.00050	0.00275	[4, 21]
path:hsa04080	Neuroactive ligand-receptor interaction	7.99071	0.00050	0.00275	[44, 96]
path:hsa04060	Cytokine-cytokine receptor interaction	7.23171	0.00050	0.00275	[28, 56]
path:hsa04510	Focal adhesion	7.15629	0.00050	0.00275	[66]
path:hsa04614	Renin-angiotensin system	6.65325	0.00050	0.00275	[48, 50]
path:hsa04014	Ras signaling pathway	6.03074	0.00050	0.00275	[4]
path:hsa04022	cGMP-PKG signaling pathway	5.78521	0.00050	0.00275	
path:hsa05166	HTLV-I infection	5.77079	0.00050	0.00275	
path:hsa04066	HIF-1 signaling pathway	5.74182	0.00050	0.00275	[68]

Table 3.3: The top 10 ranked pathways when comparing two subtypes from KIRC. The highlighted pathways have strong associations with kidney cancer based on literature.

host gene	chrom	pos	ref	sig.var	classification	not sig.var	classification
PIK3CA	3	178952085	A	G	Missense	T	Missense
TCEB1	8	74858968	T	C	Missense	A	Missense
VHL	3	10183794	G	T	Missense	A	Nonsense
VHL	3	10191470	G	A	Splice Site	T	Splice Site
VHL	3	10191470	G	C	Splice Site	T	Splice Site
VHL	3	10191480	T	C	Missense	-	Frame Shift Del
VHL	3	10191480	T	G	Missense	A	Missense
VHL	3	10191570	T	G	Missense	C	Missense

Table 3.4: Significant variants that share a position with non-significant variants in KIRC: For each significant variant (red), we show the host gene and the associated non-significant variant (blue). For each variant, we include the associated classification. Notice that even variants with same classification have different effect on gene expression.

When comparing significant variants with non-significant ones that share the same location (see Table 3.4), the survival comparison exhibits significant difference between the groups with a p-value of 0.011 (see Figure 3.4).

Overall, the renal cancer study confirms as well that having a large number of significant variants, as defined by our method, implies a lower survival rate.

3.3 Colorectal Cancer

The colorectal cancer generally occurs in the colon or rectum, which are parts of the digestive system. Colorectal adenocarcinoma is the most common colorectal cancer that often begins with a polyp growth formed on the inner part of the colon [20]. Colorectal cancer is the fourth common cancer, however the survival rate has improved due to early detection through colonoscopies and blood tests [72]. The gene expression data including 457 samples, and variant data including 154 samples studying colon adenocarcinoma (COAD) are downloaded from TCGA database. The analysis is performed on 137 number of samples for whom both types of data is available. After removing the variants with *silent* classification, the variant data includes 41,115 variants.

The gene expression data includes measurements for 17,062 genes. We removed the genes that do not have any expression for more than 50% of the samples. The number of genes that satisfies this condition is 12,387. The samples, as explained before, are divided

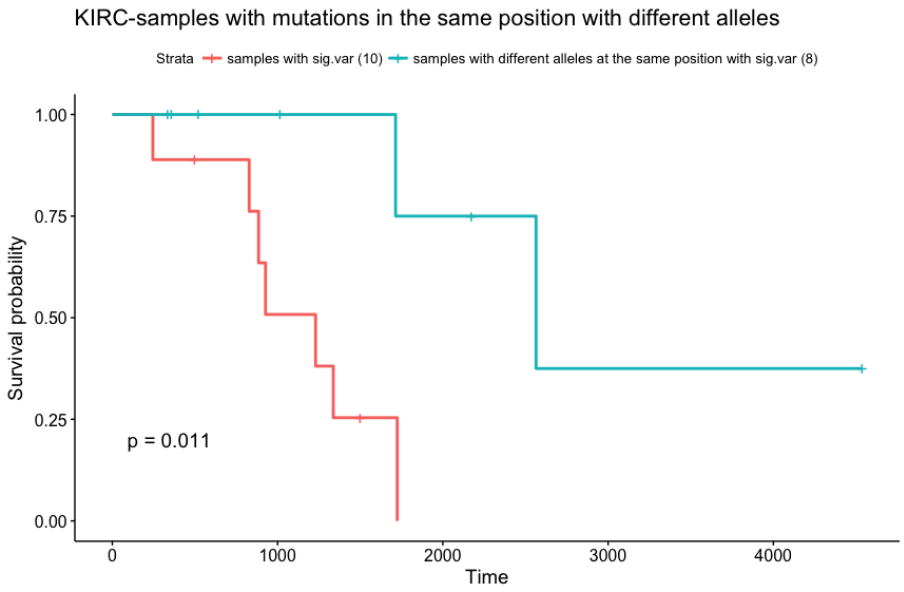


Figure 3.4: The survival curve of patients with significant variants (red) compared to patients that have a non-significant variants (blue) at the same location in KIRC. The Cox p-value (0.011) shows there is a significant difference between the survival rates.

in two groups based on the number of significant variants. The survival rate of each group is shown in Figure 3.5. The calculated cox p-value (0.024) indicates that the survival rates of two identified groups are significantly different.

Same as before, we validate the identified groups by performing pathway analysis comparing the gene expression in each group. The results in Table 3.5 show that the ranked list of significant pathway are strongly associated to colorectal cancer. The results in [52] indicate that there are significant changes in *renin-angiotensin system*, which is the top significant pathway by our method, in colorectal cancer metastases. Neo et. al suggest that a blockade of the renin-angiotensin system decreased tumor growth in colorectal cancer. This pathway is significantly perturbed when comparing one identified subtype versus the other. Many of the top ten pathways, resulted from our method, have known associations to colorectal cancer based on literature and they include the mechanisms that have potential to explain the differences between the survival of different samples in each group.

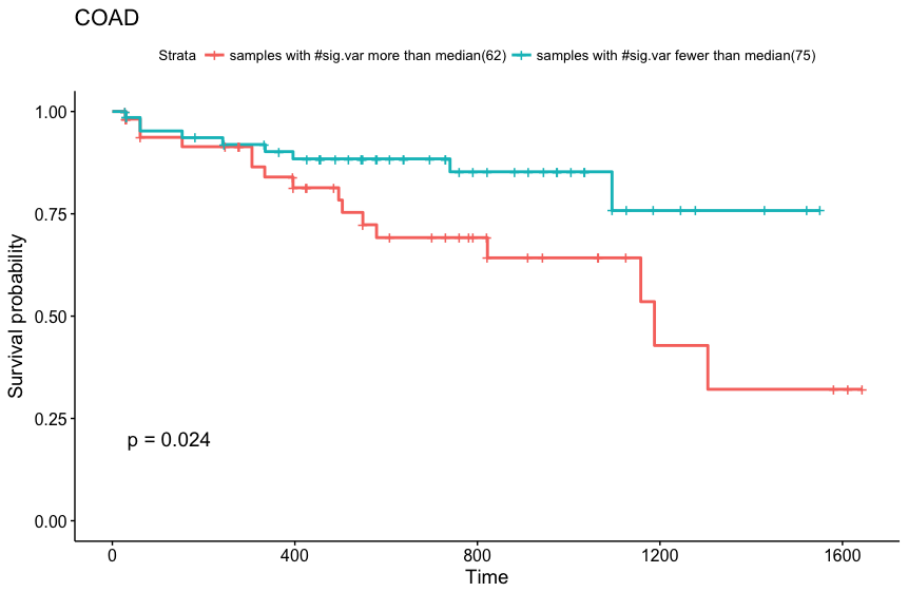


Figure 3.5: The survival curve of patients in the two identified subtypes in COAD. The red curve represents the patients with higher number of significant variants and the blue curve represents the patients with a lower number of significant variants. As expected the patients with higher number of significant variants have a lower survival rate. The set of low survival patients contains 62 individuals while the other contains 75. The Cox p-value representing the significance of differences between two curves is 0.024.

	names	totalPertNorm	pPert	pPert.fdr	references
path:hsa04614	Renin-angiotensin system	4.23970	0.00100	0.09895	[2, 52]
path:hsa04978	Mineral absorption	3.61844	0.00100	0.09895	[59, 98]
path:hsa04972	Pancreatic secretion	4.53224	0.00150	0.09895	
path:hsa04310	Wnt signaling pathway	3.69741	0.00300	0.11309	[8, 53, 63]
path:hsa04260	Cardiac muscle contraction	-3.16972	0.00300	0.11309	
path:hsa04918	Thyroid hormone synthesis	3.14207	0.00350	0.11309	
path:hsa04976	Bile secretion	3.27800	0.00400	0.11309	[62, 22]
path:hsa03320	PPAR signaling pathway	2.96947	0.00500	0.12369	[95]
path:hsa04145	Phagosome	-2.63990	0.00600	0.13193	
path:hsa05012	Parkinson's disease	-2.59650	0.00700	0.13853	

Table 3.5: The significant pathways when comparing two subtypes from COAD. The highlighted pathways have strong associations with colorectal cancer based on literature.

host gene	chrom	pos	ref	sig.var	classification	not sig.var	classification
FBXW7	4	153247289	G	A	Missense	C	Missense
KRAS	12	25398284	C	A	Missense	G	Missense
KRAS	12	25398284	C	T	Missense	G	Missense
PIK3CA	3	178936091	C	A	Missense	C	Missense
NRAS	1	115258747	C	T	Missense	G	Missense

Table 3.6: Significant variants that share a position with non-significant variants in COAD: For each significant variant (red), we show the host gene and the associated non-significant variant (blue). For each variant, we include the associated classification. Note that even variants with same classification have different effect on gene expression.

We analyzed the significant variants that share the chromosomal position with non-significant variants (see Table 3.6). The subtypes are identified by samples with such significant variants and samples with non-significant variants at the same position. The Figure 3.6 shows the survival rates of the two subtypes. The p-value is not significant, which maybe due to the low number (4) of samples that are associated to the non-significant variant.

3.4 Acute Myeloid Leukemia

Acute Myeloid Leukemia (LAML) is the most common type of acute leukemia. It is caused by an increase in the number of myeloid cells, which have not matured. These cells will not develop and will not be able to prevent infections. Recently, with diagnosis of subtypes of LAML, and improvements of drug treatments the survival rates have increased [46]. The gene expression data for LAML from TCGA includes 179 samples, while the variant data includes 197. The number of common samples with both data types is 162. After removing the variants with *silent* predicted classification, the variant data includes 1,618 number of variants.

The gene expression includes measurements for 16,818 genes and after filtering the genes without any expression in more than 50% of the samples the data includes 12,204 genes. The samples are divided in two subtypes as explained before. The survival rates of the two subtypes are shown in Figure 3.7. The significant Cox p-value indicates significant differences between the survival rates meaning that the identified subtypes are meaningful.

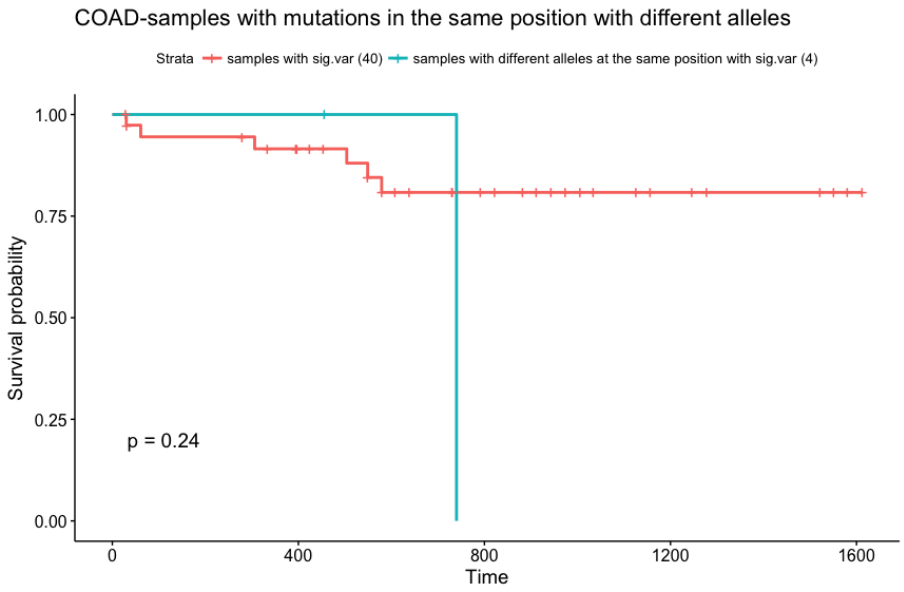


Figure 3.6: The survival curve of patients with significant variants (red) compared to patients that have a non-significant variants (blue) at the same location in COAD. The survival curves shows there is a difference between the survival rates even though the p-value is not significant.

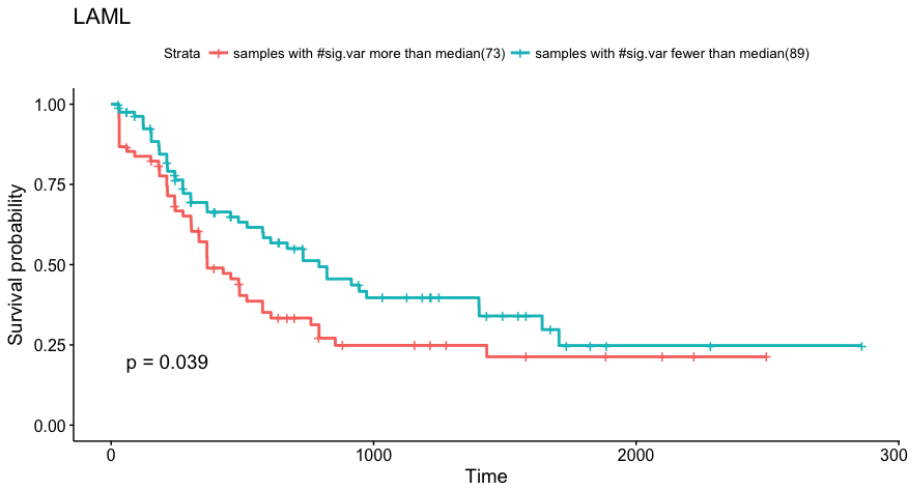


Figure 3.7: The survival curve of patients in the two identified subtypes in LAML. The red curve represents the patients with higher number of significant variants and the blue curve represents the patients with a lower number of significant variants. As expected the patients with higher number of significant variants have a lower survival rate. The set of low survival patients contains 73 individuals while the other contains 89. The Cox p-value representing the significance of the differences between two curves is 0.039.

The results of pathway analysis are shown in Table 3.7. The references mentioned in the table show that the significant pathways are associated to acute myeloid leukemia. An independent study of acute myeloid leukemia [17] has also identified the *neuroactive ligand-receptor interaction* pathway to be significantly perturbed (with p-value 10^{-5}). This is the second top significant pathway resulted from our method.

	names	totalPertNorm	pPert	pPert.fdr	references
path:hsa05144	Malaria	10.13814	0.00050	0.00660	
path:hsa04080	Neuroactive ligand-receptor interaction	9.40327	0.00050	0.00660	[17]
path:hsa04610	Complement and coagulation cascades	9.13439	0.00050	0.00660	
path:hsa05134	Legionellosis	8.62989	0.00050	0.00660	[67]
path:hsa05322	Systemic lupus erythematosus	8.54986	0.00050	0.00660	[86]
path:hsa05150	Staphylococcus aureus infection	7.85037	0.00050	0.00660	[41]
path:hsa04060	Cytokine-cytokine receptor interaction	7.15228	0.00050	0.00660	[94]
path:hsa05202	Transcriptional misregulation in cancer	6.92006	0.00050	0.00660	
path:hsa05140	Leishmaniasis	6.51377	0.00050	0.00660	
path:hsa04620	Toll-like receptor signaling pathway	6.24701	0.00050	0.00660	[71]

Table 3.7: The top 10 ranked pathways when comparing two subtypes from LAML. The highlighted pathways have strong associations with kidney cancer based on literature.

Same as all previous studies, we analyzed the significant variants that share a position with non-significant variants with different alleles. The list of such variants, their position in the genome, and reference alleles are shown in Table 3.8. The survival curves of patients with these variants are shown in Figure 3.8.

host gene	chrom	pos	ref	sig.var	classification	not sig.var	classification
DNMT3A	2	25457242	C	T	Missense	G	Missense
IDH1	2	209113113	G	A	Missense	T	Missense
U2AF1	2	,44524456	G	T	Missense	A	Missense

Table 3.8: Significant variants that share a position with non-significant variants in LAML: For each significant variant (red), we show the host gene and the associated non-significant variant (blue). For each variant, we include the associated classification. Notice that even variants with same classification have different effect on gene expression.

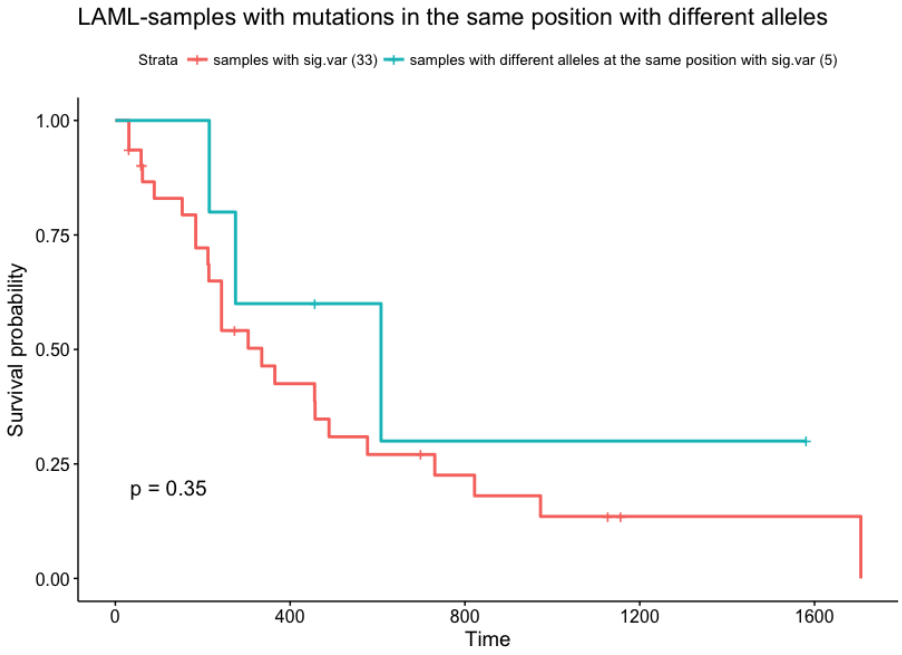


Figure 3.8: The survival curve of patients with significant variants (red) compared to patients that have a non-significant variants (blue) at the same location in LAML. The survival curves shows there is a difference between the survival rates even though the p-value is not significant.

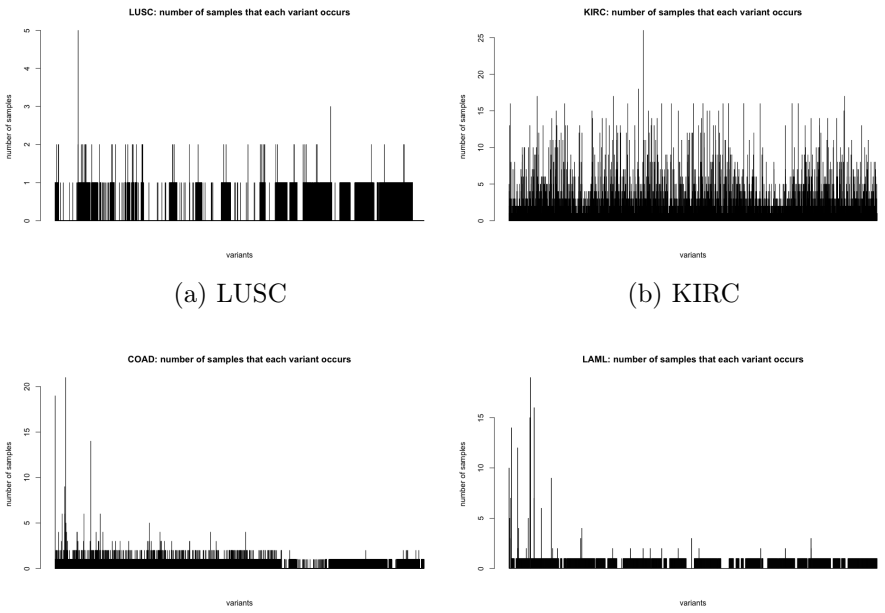
Chapter 4

Discussion and conclusions

We proposed a novel method that integrates multiple types of data to subtype patients diagnosed with the same disease. The goal was to identify those subgroups that share the same mechanisms and the same molecular profile in order to be able to develop novel drugs specific to each subgroup. We showed the effectiveness of the method on four of the most common cancer types (renal, lung, colorectal cancer and leukemia) using data from the The Cancer Genome Atlas (TCGA). The survival analysis between the groups identified showed a significant differences in the survival rates in all four patient cohorts studied. In addition, by comparing the long term versus the short term survival groups using pathway analysis, we identified that the top pathways in each study are strongly associated in literature with the conditions under study.

One of the main challenges when studying cancer samples is the heterogeneity of the molecular profiles even within the same phenotype. This aspect was clear from the beginning of our study when we compared the number of samples that share each variant (see Figure 4.1). In best case scenario the percentage of samples that share a variant ranges from 5.6% to 15% in the samples considered. On average this is much worse, with percentages ranging from 0.03% to 1%. Even with this limitation, our method is able to extract information from each variant and increase the power of the analysis by integrating gene expression.

In the proposed method, the main differentiator between the groups identified is the number of significant variants in each sample. It is only natural to ask the question if this method is biased towards patients with more variants. In other words, perhaps our results



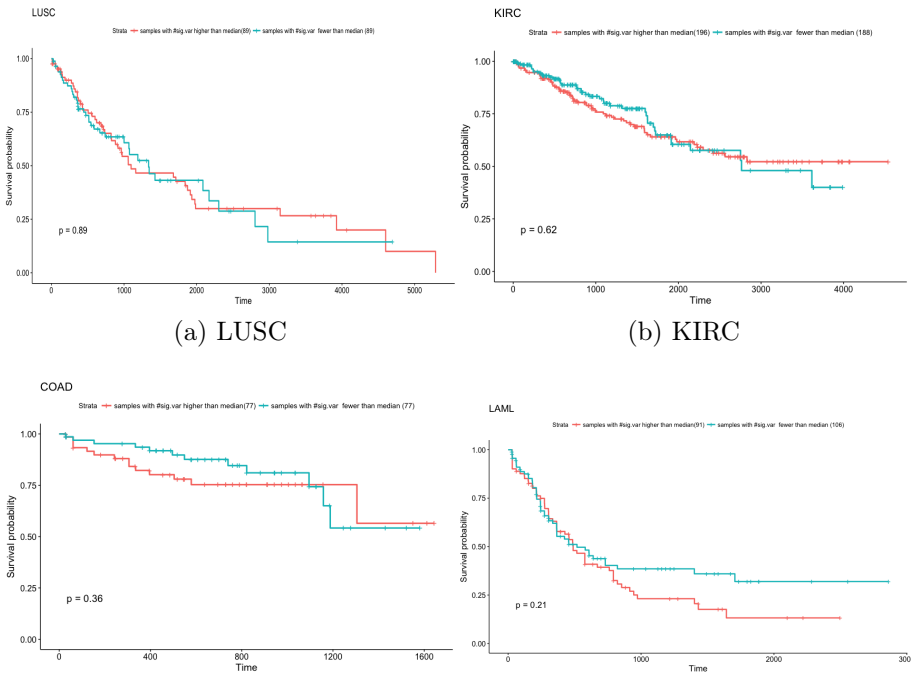
(a) LUSC (b) KIRC

(c) COAD (d) LAML

Figure 4.1: The number of samples having a given variant. In most of the cases, each variant only occurs in one sample. The maximum numbers of samples in LUSC, KIRC, COAD and LAML are 5, 26, 21 and 19, respectively.

showing that one of the subgroups has a significantly shorter survival may be due to the fact that those patients simply have much more mutations or genetic instability. To investigate this, we assumed that the number of variants (as obtained from TCGA) is the differentiator. We grouped the patients that have more than the median number of variants and fewer than the median in each one of the investigated diseases. We performed the same survival analysis and none of the pairs of groups chosen this way were significant (see Figure 4.2). This confirms that the number of variants alone cannot divide the patients in significantly different groups. In other words, the shorter survival of the groups we identified is not simply due to the presence of more mutations.

In summary, here we proposed a novel method for integrating gene expression and variant data with the purpose of subtyping patients in long and short term survival. Based on our study of four of the most common cancer types, the proposed method is able to split in subgroups each study cohort with significant differences in rate of survival.



(a) LUSC (b) KIRC

(c) COAD (d) LAML

Figure 4.2: Survival rates of identified groups based on the number of variants in the samples. The survival curves show that the rates in divided groups are not significant.

ACKNOWLEDGMENT

This research was supported in part by the following grants: NIH R01 DK089167, R42 GM087013 and NSF DBI-0965741, and by the Robert J. Sokol Endowment in Systems Biology in Reproduction. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of any of the funding agencies.

REFERENCES

[1] GR Abecasis, LR Cardon, and WOC Cookson. A general test of association for quantitative traits in nuclear families. *The American Journal of Human Genetics*, 66(1):279–292, 2000.

[2] Eleanor I Ager, Jaclyn Neo, and Christopher Christophi. The renin–angiotensin system and malignancy. *Carcinogenesis*, 29(9):1675–1684, 2008.

[3] Frank W Albert and Leonid Kruglyak. The role of regulatory variation in complex traits and disease. *Nature Reviews Genetics*, 16(4):197–212, 2015.

[4] Gowrishankar Banumathy and Paul Cairns. Signaling pathways in renal cell carcinoma. *Cancer Biology & Therapy*, 10(7):658–664, 2010.

[5] Ali Bashashati, Gholamreza Haffari, Jiarui Ding, Gavin Ha, Kenneth Lui, Jamie Rosner, David G Huntsman, Carlos Caldas, Samuel A Aparicio, and Sohrab P Shah. DriverNet: uncovering the impact of somatic driver mutations on transcriptional networks in cancer. *Genome Biology*, 13(12):R124, 2012.

[6] Denis Bertrand, Kern Rei Chng, Faranak Ghazi Sherbaf, Anja Kiesel, Burton KH Chia, Yee Yen Sia, Sharon K Huang, Dave SB Hoon, Edison T Liu, Axel Hillmer, et al. Patient-specific driver gene prediction and risk assessment through integrated network analysis of cancer omics profiles. *Nucleic Acids Research*, 43(7):e44–e44, 2015.

[7] Viv Bewick, Liz Cheek, and Jonathan Ball. Statistics review 12: Survival analysis. *Critical Care*, 8(5):389–394, 2004.

[8] Mariann Bienz and Hans Clevers. Linking colorectal cancer to Wnt signaling. *Cell*, 103(2):311–320, 2000.

[9] BioCarta. BioCarta - Charting Pathways of Life. http://www.biocarta.com.

[10] Alan P Boyle, Eurie L Hong, Manoj Hariharan, Yong Cheng, Marc A Schaub, Maya Kasowski, Konrad J Karczewski, Julie Park, Benjamin C Hitz, Shuai Weng, et al. Annotation of functional variation in personal genomes using regulomedb. *Genome Research*, 22(9):1790–1797, 2012.

[11] Zeng-hui Cheng, Yu-xin Shi, Min Yuan, Dan Xiong, Jiang-hua Zheng, and Zhi-yong Zhang. Chemokines and their receptors in lung cancer progression and metastasis. *Journal of Zhejiang University. Science. B*, 17(5):342, 2016.

[12] Lisa M Coussens and Zena Werb. Inflammation and cancer. *Nature*, 420(6917):860–867, 2002.

[13] Li Ding, Gad Getz, David A Wheeler, Elaine R Mardis, Michael D McLellan, Kristian Cibulskis, Carrie Sougnez, Heidi Greulich, Donna M Muzny, Margaret B Morgan, et al. Somatic mutations affect key pathways in lung adenocarcinoma. *Nature*, 455(7216):1069–1075, 2008.

[14] Sorin Drăghici, Purvesh Khatri, Adi Laurentiu Tarca, Kashayp Amin, Arina Done, Călin Voichiţa, Constantin Georgescu, and Roberto Romero. A systems biology approach for pathway level analysis. *Genome Research*, 17(10):1537–1545, 2007.

[15] Grace K Dy, Lourdes Ylagan, Saraswati Pokharel, Austin Miller, Elizabeth Brese, Wiam Bshara, Carl Morrison, William G Cance, and Vita M Golubovskaya. The Prognostic Significance of Focal Adhesion Kinase Expression in Stage I Non–Small-Cell Lung Cancer. *Journal of Thoracic Oncology*, 9(9):1278–1284, 2014.

[16] Gerda Egger, Gangning Liang, Ana Aparicio, and Peter A Jones. Epigenetics in human disease and prospects for epigenetic therapy. *Nature*, 429(6990):457–463, 2004.

[17] Mirko Francesconi, Daniel Remondini, Nicola Neretti, John M Sedivy, Leon N Cooper, Ettore Verondini, Luciano Milanesi, and Gastone Castellani. Reconstructing networks of

pathways via significance analysis of their intersections. *BMC Bioinformatics*, 9(4):S9, 2008.

[18] Moritz Gerstung, Andrea Pellagatti, Luca Malcovati, Aristoteles Giagounidis, Matteo G Della Porta, Martin Jädersten, Hamid Dolatshad, Amit Verma, Nicholas CP Cross, Paresh Vyas, et al. Combining gene mutation with gene expression data improves outcome prediction in myelodysplastic syndromes. *Nature Communications*, 6, 2015.

[19] Olivier Gevaert, Victor Villalobos, Branimir I Sikic, and Sylvia K Plevritis. Identification of ovarian cancer driver genes by using module network integration of multi-omics data. *Interface Focus*, 3(4):20130013, 2013.

[20] Maria Gonzalez-Pons and Marcia Cruz-Correa. Colorectal cancer biomarkers: where are we now? *BioMed Research International*, 2015, 2015.

[21] Huifang Guo, Peter German, Shanshan Bai, Sean Barnes, Wei Guo, Xiangjie Qi, Hongxiang Lou, Jiyong Liang, Eric Jonasch, Gordon B Mills, et al. The PI3K/AKT pathway and renal cell carcinoma. *Journal of Genetics and Genomics*, 42(7):343–353, 2015.

[22] MJ Hill. Bile flow and colon cancer. *Mutation Research/Reviews in Genetic Toxicology*, 238(3):313–320, 1990.

[23] Matan Hofree, John P Shen, Hannah Carter, Andrew Gross, and Trey Ideker. Network-based stratification of tumor mutations. *Nature Methods*, 10(11):1108–1115, 2013.

[24] Emily R Holzinger, Scott M Dudek, Alex T Frase, Sarah A Pendergrass, and Marylyn D Ritchie. ATHENA: the analysis tool for heritable and environmental network associations. *Bioinformatics*, page btt572, 2013.

[25] David W Hosmer, Stanley Lemeshow, and Susanne May. Applied survival analysis. *Institute of Translational Health Sciences*, 2011.

[26] Joanna MM Howson, Wei Zhao, Daniel R Barnes, Weang-Kee Ho, Robin Young, Dirk S Paul, Lindsay L Waite, Daniel F Freitag, Eric B Fauman, Elias L Salfati, et al. Fifteen new risk loci for coronary artery disease highlight arterial-wall-specific mechanisms. *Nature Genetics*, 2017.

[27] Jianguang Ji, Xiangdong Liu, Kristina Sundquist, and Jan Sundquist. Survival of cancer in patients with rheumatoid arthritis: a follow-up study in sweden of patients hospitalized with rheumatoid arthritis 1 year before diagnosis of cancer. *Rheumatology*, 50(8):1513–1518, 2011.

[28] Valerie Sloane Jones, Ren-Yu Huang, Li-Pai Chen, Zhe-Sheng Chen, Liwu Fu, and Ruo-Pan Huang. Cytokines in cancer drug resistance: cues to new therapeutic strategies. *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer*, 1865(2):255–265, 2016.

[29] G Joshi-Tope, Marc Gillespie, Imre Vastrik, Peter D'Eustachio, Esther Schmidt, Bernard de Bono, Bijay Jassal, GR Gopinath, GR Wu, Lisa Matthews, Suzanna Lewis, Ewan Birney, and Lincoln Stein. REACTOME: a knowledgebase of biological pathways. *Nucleic Acids Research*, 33(Database issue):D428–432, 2005.

[30] Minoru Kanehisa and Susumu Goto. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30, 2000.

[31] Minoru Kanehisa, Susumu Goto, Shuichi Kawashima, and Akihiro Nakaya. The KEGG databases at GenomeNet. *Nucleic Acids Research*, 30(1):42–46, 2002.

[32] Minoru Kanehisa, Susumu Goto, Shuichi Kawashima, Yasushi Okunom, and Masahiro Hattori. The KEGG resource for deciphering the genome. *Nucleic Acids Research*, 32(Database Issue):277–280, Jan 2004.

[33] Edward L Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481, 1958.

[34] Alboukadel Kassambara and Marcin Kosinski. *survminer: Drawing Survival Curves using 'ggplot2'*, 2017. R package version 0.3.1.

[35] Gary J Kelloff and Caroline C Sigman. Cancer biomarkers: selecting the right drug for the right patient. *Nature Reviews Drug Discovery*, 11(3):201–214, 2012.

[36] Purvesh Khatri, Marina Sirota, and Atul J Butte. Ten years of pathway analysis: current approaches and outstanding challenges. *PLOS Computational Biology*, 8(2):e1002375, 2012.

[37] John P Klein and Melvin L Moeschberger. *Survival analysis: techniques for censored and truncated data*. Springer Science & Business Media, 2005.

[38] Roland P Kuiper, Marjolijn JL Ligtenberg, Nicoline Hoogerbrugge, and Ad Geurts van Kessel. Germline copy number variation and cancer risk. *Current Opinion in Genetics & Development*, 20(3):282–289, 2010.

[39] Gert RG Lanckriet, Tijl De Bie, Nello Cristianini, Michael I Jordan, and William Stafford Noble. A statistical framework for genomic data fusion. *Bioinformatics*, 20(16):2626–2635, 2004.

[40] Brian D Lehmann, Joshua A Bauer, Xi Chen, Melinda E Sanders, A Bapsi Chakravarthy, Yu Shyr, and Jennifer A Pietenpol. Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *The Journal of Clinical Investigation*, 121(7):2750–2767, 2011.

[41] T Lehrnbecher, D Varwig, J Kaiser, D Reinhardt, T Klingebiel, and U Creutzig. Infectious complications in pediatric acute myeloid leukemia: analysis of the prospective multi-institutional clinical trial AML-BFM 93. *Leukemia*, 18(1):72–77, 2004.

[42] Bi-Qing Li, Jin You, Tao Huang, and Yu-Dong Cai. Classification of non-small cell lung cancer based on copy number alterations. *PLOS One*, 9(2):e88300, 2014.

[43] Qianping Li, Junyi Hou, Zhaoyan Hu, Biao Gu, and Yan Shi. Multiple mutations of lung squamous cell carcinoma shared common mechanisms. *Oncotarget*, 7(48):79629, 2016.

[44] Xiaoxia Liu, Jinling Wang, and Guiling Sun. Identification of key genes and pathways in renal cell carcinoma through expression profiling data. *Kidney and Blood Pressure Research*, 40(3):288–297, 2015.

[45] Brenton Louie, Peter Mork, Fernando Martin-Sanchez, Alon Halevy, and Peter Tarczy-Hornoch. Data integration and genomic medicine. *Journal of Biomedical Informatics*, 40(1):5–16, 2007.

[46] Bob Lowenberg, James R Downing, and Alan Burnett. Acute myeloid leukemia. *New England Journal of Medicine*, 341(14):1051–1062, 1999.

[47] Pengfei Lu, Valerie M Weaver, and Zena Werb. The extracellular matrix: a dynamic niche in cancer progression. *The Journal of Cell Biology*, 196(4):395–406, 2012.

[48] Rana R McKay, Gustavo E Rodriguez, Xun Lin, Marina D Kaymakcalan, Ole-Petter R Hamnvik, Venkata S Sabbisetti, Rupal S Bhatt, Ronit Simantov, and Toni K Choueiri. Angiotensin system inhibitors and survival outcomes in patients with metastatic renal cell carcinoma. *Clinical Cancer Research*, 21(11):2471–2479, 2015.

[49] Cristina Mitrea, Zeinab Taghavi, Behzad Bokanizad, Samer Hanoudi, Rebecca Tagett, Michele Donato, Călin Voichiţa, and Sorin Drăghici. Methods and approaches in the topology-based analysis of biological pathways. *Frontiers in Physiology*, 4:278, 2013.

[50] Akira Miyajima, Satoshi Yazawa, Takeo Kosaka, Nobuyuki Tanaka, Suguru Shirotake, Ryuichi Mizuno, Eiji Kikuchi, and Mototsugu Oya. Prognostic impact of renin–angiotensin system blockade on renal cell carcinoma after surgery. *Annals of Surgical Oncology*, 22(11):3751–3759, 2015.

[51] Richard Neapolitan, Curt M Horvath, and Xia Jiang. Pan-cancer analysis of TCGA data reveals notable signaling pathways. *BMC Cancer*, 15(1):516, 2015.

[52] Jaclyn H Neo, Eleanor I Ager, Peter W Angus, Jin Zhu, Chandana B Herath, and Christopher Christophi. Changes in the renin angiotensin system during the development of colorectal cancer liver metastases. *BMC Cancer*, 10(1):134, 2010.

[53] Laura Novellasdemunt, Pedro Antas, and Vivian SW Li. Targeting Wnt signaling in colorectal cancer. A review in the theme: cell signaling: proteins, pathways and mechanisms. *American Journal of Physiology-Cell Physiology*, 309(8):C511–C521, 2015.

[54] Sebahat Ocak, Heidi Chen, Clay Callison, Adriana L Gonzalez, and Pierre P Massion. Expression of focal adhesion kinase in small-cell lung carcinoma. *Cancer*, 118(5):1293–1301, 2012.

[55] Hiroyuki Ogata, Susumu Goto, Kazushige Sato, Wataru Fujibuchi, Hidemasa Bono, and Minoru Kanehisa. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 27(1):29–34, 1999.

[56] Marina Parton, Martin Gore, and Tim Eisen. Role of cytokine therapy in 2006 and beyond for metastatic renal cell cancer. *Journal of Clinical Oncology*, 24(35):5584–5592, 2006.

[57] Suraj Peri, J Daniel Navarro, Ramars Amanchy, Troels Z Kristiansen, Chandra Kiran Jonnalagadda, Vineeth Surendranath, Vidya Niranjan, Babylakshmi Muthusamy, TKB Gandhi, Mads Gronborg, et al. Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Research*, 13(10):2363–2371, 2003.

[58] Suraj Peri, J Daniel Navarro, Troels Z Kristiansen, Ramars Amanchy, Vineeth Surendranath, Babylakshmi Muthusamy, TKB Gandhi, KN Chandrika, Nandan Deshpande,

Shubha Suresh, BP Rashmi, K Shanker, N Padma, Vidya Niranjan, HC Harsha, Naveen Talreja, BM Vrushabendra, MA Ramya, AJ Yatish, Mary Joy, HN Shivashankar, MP Kavitha, Minal Menezes, Dipanwita R Choudhury, Neelanjana Ghosh, R Saravana, Sreenath Chandran, Sujatha Mohan, Chandra K Jonnalagadda, CK Prasad, Chandan Kumar-Sinha, Krishna S Deshpande, and Akhilesh Pandey. Human protein reference database as a discovery resource for proteomics. *Nucleic Acids Research*, 32(Suppl 1):D497–D501, 2004.

[59] Marinos Pericleous, Dalvinder Mandair, and Martyn E Caplin. Diet and supplements and their impact on colorectal cancer. *Journal of Gastrointestinal Oncology*, 4(4):409–423, 2013.

[60] Elliott Perlin, Ki Moon Bang, A Shah, Phyllis D Hursey, Wayne L Whittingham, Kabeeruddin Hashmi, Lyle Campbell, and OO Kassim. The impact of pulmonary infections on the survival of lung cancer patients. *Cancer*, 66(3):593–596, 1990.

[61] Andrea Peters, Torsten Hothorn, and Berthold Lausen. ipred: Improved predictors. *R news*, 2(2):33–36, 2002.

[62] Bandaru S Reddy and Ernst L Wynder. Metabolic epidemiology of colon cancer: fecal bile acids and neutral sterols in colon cancer patients and patients with adenomatous polyps. *Cancer*, 39(6):2533–2539, 1977.

[63] Tannishtha Reya and Hans Clevers. Wnt signalling in stem cells and cancer. *Nature*, 434(7035):843–850, 2005.

[64] Marylyn D Ritchie, Emily R Holzinger, Ruowang Li, Sarah A Pendergrass, and Dokyoon Kim. Methods of integrating data to uncover genotype-phenotype interactions. *Nature Reviews Genetics*, 16(2):85–97, 2015.

[65] Joel Rozowsky, Alexej Abyzov, Jing Wang, Pedro Alves, Debasish Raha, Arif Harmanci, Jing Leng, Robert Bjornson, Yong Kong, Naoki Kitabayashi, et al. Alleleseq: analysis

of allele-specific expression and binding in a network framework. *Molecular Systems Biology*, 7(1):522, 2011.

[66] Shinichi Sakamoto, Steven Schwarze, and Natasha Kyprianou. Anoikis disruption of focal adhesion-Akt signaling impairs renal cell carcinoma. *European Urology*, 59(5):734–744, 2011.

[67] David Schlossberg and Jose Bonoan. Legionella and immunosuppression. In *Seminars in Respiratory Infections*, volume 13, pages 128–131, 1998.

[68] Gregg L Semenza. HIF-1: upstream and downstream of cancer metabolism. *Current Opinion in Genetics & Development*, 20(1):51–56, 2010.

[69] Andrey A Shabalin. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics*, 28(10):1353–1358, 2012.

[70] Sreenath V Sharma and Jeffrey Settleman. Oncogene addiction: setting the stage for molecularly targeted cancer therapy. *Genes & Development*, 21(24):3214–3231, 2007.

[71] Evelien LJM Smits, Nathalie Cools, Eva Lion, Kirsten Van Camp, Peter Ponsaerts, Zwi N Berneman, and Viggo FI Van Tendeloo. The toll-like receptor 7/8 agonist resiquimod greatly increases the immunostimulatory capacity of human acute myeloid leukemia cells. *Cancer Immunology, Immunotherapy*, 59(1):35–46, 2010.

[72] The American Cancer Society. What is colorectal cancer?

[73] Chris Stark, Bobby-Joe Breitkreutz, Teresa Reguly, Lorrie Boucher, Ashton Breitkreutz, and Mike Tyers. BioGRID: a general repository for interaction datasets. *Nucleic Acids Research*, 34(Suppl 1):D535–D539, 2006.

[74] Barbara E Stranger, Matthew S Forrest, Mark Dunning, Catherine E Ingle, Claude Beazley, Natalie Thorne, Richard Redon, Christine P Bird, Anna De Grassi, Charles

Lee, et al. Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science*, 315(5813):848–853, 2007.

[75] Stephen E Straus, Philip A Pizzo, and Larry I Lutwick. Infectious complications of lung cancer. *Lung Cancer: Clinical Diagnosis and Treatment. Second ediction, edited by MJ Straus, Grune and Stratton Inc., New York*, pages 293–314, 1983.

[76] Toshiyuki Tanaka, Zhongbin Bai, Yuttana Srinoulprasert, BoGi Yang, Haruko Hayasaka, and Masayuki Miyasaka. Chemokines in tumor progression and metastasis. *Cancer Science*, 96(6):317–322, 2005.

[77] TCGA Research Network. The Cancer Genome Atlas Research Network. `http://cancergenome.nih.gov/`.

[78] The Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. *Nature*, 474(7353):609–615, 2011.

[79] The Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. *Nature*, 489(7417):519–525, 2012.

[80] The Cancer Genome Atlas Research Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, 487(7407):330–337, 2012.

[81] The Cancer Genome Atlas Research Network. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, 2012.

[82] The Cancer Genome Atlas Research Network. Integrated genomic characterization of endometrial carcinoma. *Nature*, 497(7447):67–73, 2013.

[83] The Cancer Genome Atlas Research Network. Integrated genomic characterization of papillary thyroid carcinoma. *Cell*, 159(3):676–690, 2014.

[84] The Cancer Genome Atlas Research Network. Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature*, 517(7536):576–582, 2015.

[85] Louise B Thingholm, Lars Andersen, Enes Makalic, Melissa C Southey, Mads Thomassen, and Lise Lotte Hansen. Strategies for integrated analysis of genetic, epigenetic, and gene expression variation in cancer: Addressing the challenges. *Frontiers in Genetics*, 7, 2016.

[86] Yukiko Tsunematsu, Ryo Koide, Michiko Sasaki, and Hirotaka Takahashi. Acute myeloid leukemia with preceding systemic lupus erythematosus and autoimmune hemolytic anemia. *Japanese Journal of Clinical Oncology*, 14(1):107–113, 1984.

[87] Alison L Van Dyke, Michele L Cote, Angie S Wenzlaff, Wei Chen, Judith Abrams, Susan Land, Craig N Giroux, and Ann G Schwartz. Cytokine and cytokine receptor single-nucleotide polymorphisms predict risk for non–small cell lung cancer among women. *Cancer Epidemiology and Prevention Biomarkers*, 18(6):1829–1840, 2009.

[88] Călin Voichiţa, Michele Donato, and Sorin Drăghici. Incorporating gene significance in the impact analysis of signaling pathways. *Proceedings of the International Conference on Machine Learning Applications (ICMLA)*, December 2012.

[89] Calin Voichita, Sahar Ansari, and Sorin Draghici. *ROntoTools: R Onto-Tools suite*, 2016. R package version 2.0.0.

[90] Kai Wang, Mingyao Li, and Hakon Hakonarson. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, 38(16):e164–e164, 2010.

[91] Lin Wang, Fuhai Li, Jianting Sheng, and Stephen TC Wong. A computational method for clinically relevant cancer stratification and driver mutation module discovery using personal genomics profiles. *BMC Genomics*, 16(7):S6, 2015.

[92] Richard Wender, Elizabeth TH Fontham, Ermilo Barrera, Graham A Colditz, Timothy R Church, David S Ettinger, Ruth Etzioni, Christopher R Flowers, G Scott Gazelle,

Douglas K Kelsey, et al. American Cancer Society lung cancer screening guidelines. *CA: A Cancer Journal for Clinicians*, 63(2):106–117, 2013.

[93] Rohan BH Williams, Eva KF Chan, Mark J Cowley, and Peter FR Little. The influence of genetic variation on gene expression. *Genome Research*, 17(12):1707–1716, 2007.

[94] Shuling Wu, Reinhard Geßner, Arend von Stackelberg, Renate Kirchner, Guenter Henze, and Karl Seeger. Cytokine/cytokine receptor gene expression in childhood acute lymphoblastic leukemia. *Cancer*, 103(5):1054–1063, 2005.

[95] Wan-Lin Yang and Harold Frucht. Activation of the PPAR pathway induces apoptosis and COX-2 inhibition in HT-29 human colon cancer cells. *Carcinogenesis*, 22(9):1379–1383, 2001.

[96] William Yang, Kenji Yoshigoe, Xiang Qin, Jun S Liu, Jack Y Yang, Andrzej Niemierko, Youping Deng, Yunlong Liu, A Keith Dunker, Zhongxue Chen, et al. Identification of genes and pathways involved in kidney renal clear cell carcinoma. *BMC Bioinformatics*, 15(17):S2, 2014.

[97] Hua Yu, Marcin Kortylewski, and Drew Pardoll. Crosstalk between cancer and immune cells: role of STAT3 in the tumour microenvironment. *Nature Reviews Immunology*, 7(1):41–51, 2007.

[98] Huawei Zeng, Darina L Lazarova, and Michael Bordonaro. Mechanisms linking dietary fiber, gut microbiota and colon cancer prevention. *World Journal of Gastrointestinal Oncology*, 6(2):41–51, 2014.