IDENTIFYING PATTERNS OF CANCER DISEASE MECHANISMS BY MINING ALTERNATIVE REPRESENTATIONS OF GENOMIC ALTERATIONS

by

Vicky Chen

B.S. in Biomedical Engineering, University of Virginia, 2008

M.S. in Biomedical Informatics, University of Pittsburgh, 2012

Submitted to the Graduate Faculty of

School of Medicine in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2016

UNIVERSITY OF PITTSBURGH

SCHOOL OF MEDICINE

This dissertation was presented

by

Vicky Chen

It was defended on

September 13, 2016

and approved by

Rebecca Jacobson, MD, MS, Department of Bioinformatics, University of Pittsburgh

Harry Hochheiser, PhD, Department of Bioinformatics, University of Pittsburgh

John Paisley, PhD, Department of Electrical Engineering, Columbia University

Dissertation Advisor: Xinghua Lu, MD, PhD, Department of Biomedical Informatics, University of Pittsburgh

IDENTIFYING PATTERNS OF CANCER DISEASE MECHANISMS BY MINING ALTERNATIVE REPRESENTATIONS OF GENOMIC ALTERATIONS

Vicky Chen, PhD

University of Pittsburgh, 2016

Cancer is a complex disease driven by somatic genomic alterations (SGAs) that perturb signaling pathways and consequently cellular function. Identifying combinatorial patterns of pathway perturbations would provide insights into common disease mechanisms shared among tumors, which is important for guiding treatment and predicting outcome. However, identifying perturbed pathways is challenging, because different tumors can have the same perturbed pathways that are perturbed by different SGAs.

We started off by designing a novel semantic representation that captures the functional similarity of distinct SGAs perturbing a common pathway in different tumors. This representation was used alongside the nested hierarchical Dirichlet process topic model in order to identify combinatorial patterns in altered signaling pathways. We found that the topic model was able to capture the functional relationships between topics. It was also able to identify cancer subtypes composed of tumors from different tissues of origin that exhibit different survival rates.

These results led us to investigate the performance of the methodology on pan-cancer data, as well as in conjunction with cancer driver data. The results revealed that the framework was still able to identify clinically relevant features in pan-cancer. However, the addition of driver data decreased the noise in the data and improved the separation of tumors in the clustering results. This provided support for including the use of driver data in our methodology.

In order to have gene representations independent of literature, we developed a biological representation that could identify functionally related genes. Its performance when used alongside topic modeling was tested. We found that the topic association patterns separated tumors by their tissue of origin. But, analyzing some of the cancer types on an individual basis still led to significant differences in survival.

Our studies show the potential for using alternative representations in conjunction with topic modeling to investigate complex genomic diseases. With further research and refinement of this methodology, it has the potential to capture the relationship between pathways involved in cancer. This would contribute to a better understanding of cancer disease mechanisms and treatment.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SUPPLEMENTAL TABLES

LIST OF SUPPLEMENTAL FIGURES

PREFACE

When reflecting on these past years of graduate school, I feel grateful to many people. I would first like to thank my advisor, Dr. Xinghua Lu, for his support and advice throughout my PhD study. He always motivated and encouraged me throughout my dissertation research, and was always there whenever I encountered problems during my experiments. I would also like to thank the members of my committee members - Dr. Rebecca Jacobson and Dr. Harry Hochheiser of the University of Pittsburgh, and Dr. John Paisley of Columbia University - for monitoring my progress and providing prudent advice.

I would also like to thank the members of Dr. Lu's laboratory: Dr. Lujia Chen, Dr. Chunhui Cai, Dr. Jonathan Young, Michael Ding, and Xueer Chen. Their feedback and openness has made my time at the office a much more rewarding experience. In particular, I would like to thank Lujia for her support and companionship both professionally and personally. I would also like to thank Toni Porterfield for all of her help and support during my studies here at DBMI. Without her help scheduling and coordinating everything and tracking my milestones along the way, my time here would have been much more difficult. I'm so grateful to be a part of the DBMI family.

Finally, I would like to dedicate my dissertation to my family for their unwavering support. I thank you all for supporting me throughout this wonderful journey in Pittsburgh.

1.0 INTRODUCTION

1.1 CANCER

Cancer is one of the leading causes of death worldwide, responsible for 8.2 million deaths in 2012 [1]. It is a disease where abnormal cells have uncontrolled cell growth and division, and is able to invade other tissues [2]. There are many different exogenous and endogenous factors related to cancer, which may lead to genetic alterations or otherwise impact cellular functions. Factors that induce altered cellular functions include viruses [3-5] and chronic inflammation [6-8]. However, the disease is more often the result of genetic alterations that lead to altered cellular function [9, 10]. These may be constitutional or somatic genetic alterations (SGAs). Constitutional alterations, which are also referred to as germline alterations, are genetically inherited; thus each constitutional alteration is present in every cell. On the other hand, SGAs occur after conception and thus each SGA is not present in every cell. While constitutional genetic alterations may result in a predisposition towards developing cancer [11-15], inherited cancer makes up only a small proportion of the total cancer cases [16]. Most cancers are the result of SGAs that are accumulated over time [9].

Understanding how SGAs contribute to the development of cancer would shed light on the mechanisms that underlie the disease and guide treatment. Currently, cancers are first classified by the organ of origin and then further classified into subtypes or grades according to their clinical and/or molecular characteristics [17]. For example, cancers are first classified as lung cancer, breast cancer, etc, where breast cancers can be further classified as LumA, LumB, Basal and Her2 subtypes according to their molecular characteristics through gene expression [18, 19]. Classification according to anatomical location is a natural option from a taxonomic perspective, and molecular classification of tumors based on gene expression provides valuable clinical information in terms of prognosis. However, current classifications do not reflect the underlying disease mechanisms. In other words, current classification of cancers is mainly based on the observed phenotype rather the driving causes of the disease. As such, the current classifications have limited utility in terms of guiding precision medicine targeting the cause of the disease. Here we hypothesize that, by finding the combinatorial patterns of pathway perturbations among cancers from different tissue of origin, we may classify cancers according to the disease mechanisms. This, in turn, will lead to a better understanding of cancer and better therapy.

1.1.1 Biological Processes of Cancer

Decades of cancer research revealed that a number of biological processes are involved in the development of cancer. Hanahan and Weinberg have summarized these as six different hallmarks of cancer (sustained proliferative signaling, evading growth suppressors, tissue invasion and metastasis, enabling replicative immortality, sustained angiogenesis, and resisting cell death), and more recently added two emerging hallmarks (deregulated metabolism and avoiding immune destruction) and two enabling characteristics (tumor-promoting inflammation and genome instability and mutation) [7, 10]. The theory is that normal cellular functions need to be altered in order to attain these hallmarks and for normal cells to develop into malignant and metastatic tissue. These changes in cellular function are, in turn, the result of altered functions of cellular signaling pathways regulating these processes. Alterations of signaling pathways that lead to cancer are often the result of SGAs affecting signaling proteins. For example, mutations in the RB1 pathway lead to aberrant activation of cell proliferation pathways [20-22], and SGAs resulting in aberrant activation of mTOR pathways enable cancer cells to adapt to changed nutrition or hypoxic environments [23, 24]. It is believed that different combinations and degrees of perturbation of signaling pathways underlying the hallmark processes lead to the heterogeneous phenotypes associated with individual tumors.

Projects such as The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium have made a large volume of cancer genomic data available. One goal of analyzing SGA data is to reveal the pathways perturbed by SGA events and therefore shed light on the disease mechanisms of cancers. Analyses of this data have lead researchers to notice that the prevalence of distinct SGA varies across different types of cancer, both within and across different tissue types [25-34]. For example, it is noted that TP53 mutations are highly prevalent in ovarian cancer but relatively infrequent in kidney cancers [25, 31, 35]. Most current TCGA publications mainly concentrate on investigating the prevalence of mutations and often report the results as a ranked list of most frequently mutated genes. However, such a list does not reflect the combinatorial patterns of SGAs, such as where a set of SGAs commonly co-occurs in a subset of tumors, as an indication that these tumors share a common disease mechanism. It is noted that specific combination of SGAs, thereby the combination of perturbed signaling pathways, determines the aggressiveness of cancer and its responsiveness to therapy [18, 36]. Different combinations of perturbed pathways can occur in cancers from the same organ of origin, which is what results in the heterogeneity of cancer [9, 18]. However, to the best of our knowledge, few studies try to develop and apply statistical methods specifically aimed to identify **combinatorial patterns of SGAs or pathways** as a means for discovering distinct disease mechanisms underlying heterogeneity of cancers [37, 38]. This is the main task to be addressed in this dissertation.

There are two challenges in detecting patterns in signaling pathway perturbations based on genomic alteration data from TCGA or the International Cancer Genome Consortium (ICGC). The first of these lies in the fact not all mutations that occur are “driver” mutations that lead to the development of cancer. Rather, a cancer cell may have multiple “passenger” mutations that do not contribute to cancer development [9, 39, 40]. Thus, it is necessary to be able to distinguish between driver and passenger mutations. The second challenge lies in the fact that signaling pathways are composed of multiple proteins. It is often the case that perturbing any one of the proteins along a pathway would have the same impact on the entire pathway. For example, Ras, Raf, MEK and MAPK are all part of the MAPK signaling pathway, and mutating any one of these genes would lead to altered cell proliferation [41-44]. As discussed before, this phenomenon leads to the apparent high degree of heterogeneity of SGAs, i.e., very few tumors share a common set of SGAs even if a common set of pathways are perturbed among them [9, 18, 45]. This is what makes it difficult to determine that the same signaling pathway is affected in different tumors, since different SGAs may be observed. When it is already difficult to determine that an individual signaling pathway is perturbed, it becomes even more challenging to learn combinations of pathway perturbations directly based on genomic alteration data. In this dissertation, we will investigate methods to overcome these challenges and detect patterns in pathway perturbations in cancer based on SGAs.

1.1.2 Identifying Driver Genes

As cancer genomic sequencing data became increasingly available, it is now widely appreciated that most of the mutations in a cancer sample are passenger mutations and only a few of them are driver mutations [7, 35, 46]. It is important to identify driver genes in order to reduce the amount of noise when trying to classify cancers. Since identifying driver genes is vital to understanding cancer development, many researchers have focused on this problem. Multiple approaches have been implemented in order to distinguish between driver mutations and passenger mutations and evaluate the resulting identifications. The current driver identification approaches falls under three large categories, with each concentrating on one of the following characteristics: frequency of gene alteration (recurrence), mutual exclusivity of mutations within a pathway, and functional impact of mutations.

1.1.2.1 Frequency-based approaches for identifying driver genes

The frequency-based approaches aim to identify the mutations that are observed in a population of tumors with a prevalence above random chance determined according to a background mutation frequency [35, 39, 47-50]. The assumption is that the mutations that increase the chances of cancer developing would occur at a higher prevalence than random mutations due to “positive selection”. Thus the mutations that occur at a higher frequency are more likely to be driver mutations that result in oncogenic advantages in cancer cells [7, 9, 10]. Most of these methods identify driver genes based on somatic mutations [35, 39, 47-50]. Although researchers also investigated the landscape of copy number alterations to identify cancer driver genes [51], there is no established criterion to determine if a copy number alteration of a gene or chromosome region is a driving event.

In order to accurately predict which genes are mutated at a higher than expected frequency, a background mutation frequency needs to be established. Based on the idea that silent mutations do not undergo selective pressure, Greenman et al. implemented one of the most basic methods of calculating background mutation frequency by using the rate of silent mutations [39]. This method has been used to evaluate mutated genes in both breast cancer and across the entire human cancer genome [52, 53]. However, having a one-fit-all background mutation rate can result in both false positive and false negative calls for driver mutations. Researchers noticed that mutation rate can be highly heterogeneous across the human genome, dependent on genome location, sequence patterns (e.g., nucleotide repeats), and transcription activity of the region [48, 54]. In order to compensate for the fact that the mutation rate of genes isn't consistent in individual patients or in individual genes, Lawrence et al. introduced an adjustment to this calculation by taking into account the patient-specific mutation frequency and mutation spectrum, and gene-specific background mutation rates [47, 48]. They implemented this method and made it available in the MutSigCV tool [48]. Since it is also possible for mutation rates to vary based on the type of mutation, another method of calculating background mutation frequency is to separate the mutations into different mutational mechanism categories [35, 49, 50]. This is relevant in situations where a certain type of mutation mechanism is more common. For example, smoking increases the prevalence of the mutation mechanism G:C-to-T:A transversions. This means there would be more passenger genes with this type of mutation. Thus, a gene with this type of mutation would have to recur at a higher rate in order to be of interest when compared to a gene with a different mutation mechanism. The background rate was calculated for each mutation mechanism category by dividing the number of mutations in a category by the total number of bases where such a mutation can occur [35, 49, 50].

While frequency-based approaches can detect genes that have high mutation frequencies as drivers with relatively good accuracy, they are not suited to differentiate between driver and passenger mutations that have relatively low frequency. For example, Lawrence et al. examined mutation rates and performed saturation analysis using pan-cancer data from TCGA, and they concluded that current methods cannot reliably detect drivers if the mutation rate is below 20% [47], however over 90% of genes mutated in TCGA samples have a mutation rate below this threshold. Even though methods have been developed to better detect the genes that are mutated at intermediate frequencies [47], being able to use recurrence to detect driver genes that occur at low frequencies remains difficult.

Cancers can also be driven by somatic copy number alterations (amplification or deletion) of certain genes. The best known examples are the copy number alteration of MYC [55], ERBB2 (Her2) [56] and PTEN [57]. Her2 amplification is prevalent in breast cancers and is associated with specific molecular characteristics, such that a molecular subtype of breast cancer is labeled as Her2-positive [19, 58]. While many genes affected by somatic copy number variations are classified as drivers through biological experiments, setting a guideline by which to detect drivers in copy number variations has been difficult. In order to take steps toward detecting drivers, Zack et al. established a way to estimate the background frequency of copy number alterations [51]. They did so by calculating the frequency that copy number alteration events of similar length and amplitude occur across the entire TCGA dataset, while correcting for the length and chromosomal location of the copy number alteration [51]. However, they were still unable to set a threshold by which to determine which genes are drivers.

1.1.2.2 Finding driver genes on a pathway based on their mutual exclusivity pattern

Mutation patterns in tumors have been used to identify driver genes or pathways based on the characteristic that SGAs affecting members of a common pathway are less likely to co-occur in the same tumor, a phenomenon referred to as mutual exclusivity. The explanation for this phenomenon is that when a tumor only needs one driver gene to be altered in a driver pathway, altering a second driver is unnecessary [45]. Thus, driver genes with the same functional impact would tend to, but does not have to, be mutually exclusive. Both somatic gene mutations [45, 49, 59, 60] and copy number alterations [25, 45, 51, 59] have been used to detect mutual exclusivity. One method of determining mutual exclusivity is to calculate the correlation or anticorrelation of genes when compared with a background co-occurrence rate [49, 51]. Like the background mutation rate, the background co-occurrence rate can be calculated in different ways. Zack et al. calculated the background co-occurrence rate for copy number alterations by permuting the copy number profiles of chromosomes across samples [51]. When calculating the rate for gene mutations, Dees et al. permuted the observed gene mutations, while keeping both the distribution of the number of gene mutations per sample and the number of mutations in each gene constant [49].

Mutual exclusivity can also be identified independent of a background co-occurrence rate. Ciriello et al. used gene interaction networks to identify gene subsets that are fully connected [60]. These subsets are then assessed for highly recurrent and mutually exclusive genes, by comparing to randomly generated networks [60]. The Dendrix algorithm was developed to determine mutual exclusivity by identifying genes that have a maximal coverage of the tumors while minimizing the overlap of the genes [59]. This has also been extended to simultaneously identify multiple driver pathways [45].

Mutual exclusivity is not based on the frequency of alterations, but rather based on the assumption that the mutual exclusivity of genes is the result of them being part of the same driver pathways. As a result, methods based on mutual exclusivity are capable of detecting driver genes that have a lower rate of occurrence. However, there are many other situations that could result in genes being mutually exclusive besides being part of the same driver pathway. Since the algorithms are unable to differentiate between mutually exclusive drivers and other situations that result in mutual exclusivity, this would impact the accuracy of the driver genes predicted by these methods [61].

1.1.2.3 Identifying driver genes using the impact of mutations on individual gene function

Driver mutations have also been identified by detecting genes hosting mutation events that likely affect their function. For example, a gene that is affected by mutations that tend to cause loss or gain of function, e.g., truncation mutations or relatively large insertions or deletions, or mutations in the evolutionally conserved regions. The theory behind searching for driver genes based on the functional impact of their mutation is that driver genes would have **more mutations with greater functional impact**, because drivers need to be altered in order for cancer to develop. Many different methods of assessing the functional impact of a mutation have been developed. For example, Reva et al. used evolutionary information to predict functional impact, with the idea that residues that are evolutionarily conserved are more likely to be functional than those that are not conserved [62]. This allows predictions to be made as long as protein family sequences are available to generate sequence homologs [62].

Mutation locations can also be used to assess functional impact. Gain of function genetic alterations often cluster together in specific protein regions [63], which can be utilized to detect driver genes that undergo gain of function alterations. These genes can be identified by searching for genes with a bias towards having clusters of protein-affecting mutations, when compared to a baseline clustering rate. Tamborero et al. used silent mutations as a measure of baseline clustering [63], in lieu of using a homogeneous mutation probability [49]. Another location-based method is to determine the rate of mutation at functional residues. Reimand et al. evaluated mutation rates at phosphorylation-sites in order to predict signaling-specific cancer driver genes [64]. In this case, the background mutation rates used were non-synonymous mutations that were not phosphorylation-related [64].

These types of functional impact measures can be used to identify the genes that have a tendency to accumulate mutations with high functional impact [50]. As such, they are aimed at detecting a different set of driver genes than the previous two approaches. The benefit of functional impact measures is that they limit the noise incorporated from mutations that do not impact function. However, the accuracy of these driver identification methods is limited by the accuracy of the functional impact predictions.

1.1.2.4 Lack of gold standard driver genes

Despite all the driver identification methods available, the lack of a gold standard of knowledge has made the evaluation of these methods limited. The data available that most closely approaches a gold standard are cancer driver databases [46, 65]. As such, some driver identification methods have been run on these cancer driver databases to determine if they are capable of detecting known drivers [62, 63]. These cancer driver databases can also be used as an evaluation metric when running the algorithms on cancer datasets. Generally the algorithms are run on cancer datasets to identify both known and new cancer genes, with the new genes being considered potential novel drivers [35, 39, 45, 47-51, 62-64, 66]. Using the databases to calculate a positive predictive proxy measure makes it possible to approximate the positive predictive value of the returned driver gene lists, and compare the results of different methods [50]. Another evaluation method that doesn’t require cancer datasets or databases is the use of simulated data [45]. Generating artificial data allows researchers to know what the correct prediction results should be. However, simulated data are generated to fit the designed algorithm, and can be difficult to apply and compare to other algorithms.

1.1.3 Cancer Subtyping

Cancers can differ based on many different features, such as the tissue of origin or its disease mechanisms. The process of cancer subtyping is to separate cancers based on these features. Because these features can have an effect on a cancer’s response to treatment, the resulting cancer subtypes respond differently to different treatments. This is why cancer subtyping is an important step in the diagnostic process, and plays a role in cancer treatment. Therefore, the goal of improving cancer subtyping methods is to improve cancer treatment through precision medicine, which will be expanded upon in section 1.1.4.

Current medical guidelines focus on classifying cancer based on observable features, which include the primary site, histology, cell differentiation, and the extent of the disease [17]. The primary site, or the location where the cancer is first developed, is used to separate the cancer into broad categories, such as breast cancer or lung cancer. This is also used to track where a metastasized secondary cancer originated from. Histological features are used to further subtype primary cancer. For example, lung cancer can be further separated into small cell lung cancer, adenocarcinoma, squamous-cell carcinoma, and large-cell carcinoma. Cell differentiation is used to grade the tumor and determine its aggressiveness. Tumors that bear a closer resemblance to normal tissue would be less aggressive and have a lower grade than those that are poorly differentiated. The extent of the disease can then be used to stage the disease, with tumors that have spread further receiving a higher stage.

In last two decades, molecular features of some cancers have been discovered, allowing them to be subtyped further. For example, receptor expression and gene expression can be used to further separate breast cancer into basal, HER-2, and luminal A and B tumors [19, 67, 68]. The limitation of these current methods is that there are many different genetic alterations that can result in these different disease presentations. As a result, the same treatment has different degrees of effectiveness in patients with the same cancer subtype. It is important to identify the genetic changes that result in these different features, in order to learn what causes the varying degrees of treatment response. Contemporary molecular subclassification does not directly provide insights into the cellular or disease mechanisms underlying the subtypes or guide treatment.

There has been recent research focusing on separating cancers into different classes based on their disease mechanisms, which can be shared across different tissues of origin. The most frequently used data types are somatic mutations [35, 37, 38, 69] and copy number alterations [38, 69], however other types of expression data are also used [38, 69]. Different clustering methods have been applied to these input data in order to separate the tumors into classes. Unsupervised hierarchical clustering was directly applied to significantly mutated genes to separate the tumors into separate classes based on their mutation patterns [35]. Similarly, Hoadley et al. also clustered the tumors based on significantly mutated genes [38]. However, they took this process a step further by also clustering by copy number alteration as well as 4 other data types, which resulted in a total of 6 different clustering results [38]. They then separated the tumors into classes by clustering based on 5 clustering results, after excluding somatic mutations from their analysis [38]. This cluster-of-cluster analysis is an effective way to combine the input from multiple data types. However, if some of the data types contain overlapping information, their combined signal may overpower the signal provided by other data types. Thus, care needs to be exercised when selecting and processing the data types used for input. Ciriello et al. used somatic mutation, copy number alteration, methylation, and gene expression data in order to generate recurrent functional events, which are functional alterations that occur at a high frequency [69]. These functional events should contain information about the functions altered in a cancer sample. Generating a network where samples and functional events are nodes, and samples are connected to events using edges allowed them to represent the alterations in each sample. This also makes it possible to use graph clustering to group samples with similar alterations into classes [69]. All of the previously listed methods used the altered genes directly, and so were unable to represent the fact that different genes can impact the same pathway. This restriction makes it difficult to learn the disease mechanisms that lead to the development of cancer.

In order to help identify the disease mechanisms of cancer, signaling pathway information should be incorporated. Hofree et al. noticed that mutation data are highly heterogeneous among tumors, making it difficult to find mutation patterns [37]. To address this problem, they mapped mutations to gene interaction networks as a means to map mutations to function space. Then tumors with similar regions in the gene interaction network perturbed can be identified, thus stratifying the tumors [37]. While this is a method to represent the functions a gene is involved in, it is limited by the accuracy and scope of the gene interaction networks available. However, an increased incorporation of pathway perturbation information into cancer subtyping should result in more specific diagnoses and more precise treatment than is currently available.

1.1.4 The need for disease-mechanism-based classification of cancers

Research that leads to a better understanding of cancer and its disease mechanisms would have significant impact in different areas of cancer treatment. Two prominent areas would be in precision medicine and drug development. Current methods of cancer classification are based on phenotype [17]. While some of the contemporary cancer classifications contain information for prognosis, they usually do not provide information about the underlying disease mechanisms. This is because the gene expression profile of a tumor reflects a convoluted outcome of all cellular signal pathways that actively regulate expression in tumor cells. Therefore, it is difficult to map a profile to the activation states of individual pathways. However, if we can identify the SGAs that underlie a specific gene expression profile, the results will shed light on the disease mechanisms of a tumor.

The idea behind precision medicine is to tailor treatment to an individual patient using patient-specific information, including their genomic data. The critical task is to understand how perturbations of signaling pathways, particularly **combinations of pathway aberrations**, contribute to the development of different subtypes of cancers. Such an understanding will further guide the development of distinct treatment strategies for tumors with different disease mechanisms. A better understanding of what treatment methods work for different cancer patients would allow an individual to be given treatments that have the best chances for success [70, 71]. This would decrease the amount of time spent ruling out ineffective treatments and limit the drugs and potential side effects that a patient is exposed to. Another approach to using this understanding to improve cancer treatment is through drug development. An aim in cancer treatment is to identify drugs that can target cancer cells while limiting the impact on normal tissues. Cancer drivers and disease mechanisms that are identified are potential targets for drug development. Alternatively, this information can be used for drug repurposing, if an existing drug is already known to affect these drug targets.

The overarching goal of this study is to investigate the distinct disease mechanisms of cancers by mining the combinatorial patterns of pathway aberrations, which in turn can be inferred by mining the combinatorial patterns of SGAs perturbing the pathways. However, this task is challenging because a pathway can be perturbed by distinct SGAs that affect different member proteins of the pathway, which leads to the well-known phenomenon that few tumors share common SGAs even though a common set of pathways are perturbed. The key task that needs addressing is to identify the common pathways that are affected by distinct SGAs in different tumors. By starting with the SGAs in tumors, it is necessary to determine the pathways that they lie on, and thus the functions they are involved in. Since gene names themselves do not convey the function of a gene, thereby being insufficient for inferring the functional similarity of distinct genes, an alternative representation of genes is needed to assess the functional similarity of genes with distinct names. For example, in a representation we call semantic embedding, a gene can be represented by a set of words from literature that describe the function of the gene. If a significant number of words are shared in the two word sets describing two distinct genes, the functions of the two genes are likely to be similar. Different sets of words are used to describe different gene functions, and each of these word sets is associated with a topic. Once tumors are composed of these representations, we would need to detect the patterns in their altered pathways through the patterns in the words. Such a task can be accomplished through the application of topic modeling.

1.2 TOPIC MODELING

In text mining, a common method of capturing the statistical structure of a corpus of texts is to capture the tendency of certain words to co-occur when a semantic topic is discussed. By capturing such structures, one can represent a text as a mixture of words from such topics, which serve as a more concise and abstract representation of documents in the corpus than individual words. The goal of identifying representative features, like topics, for samples has been approached in many different ways. Two of the methods that involve matrix factorization are singular value decomposition (SVD) [72] and nonnegative matrix factorization (NMF) [73], which are similar to principal component analysis. When given a corpus represented as a document-by-word matrix, these methods factorize the matrix into a word-by-feature matrix and a feature-by-document matrix. The word-by-feature matrix identifies what words are associated with a given feature, while the feature-by-document matrix identifies what features are associated with a given document. Both SVD and NMF are methods to perform latent semantic indexing that also results in dimensionality reduction. One of the limitations of SVD is that it requires the discovered features to be orthogonal. This means that the features must be independent of each other. NMF does not have this restriction, allowing dependencies to exist between features. NMF also limits the matrixes from having negative values, which SVD permits. By allowing feature matrices to only have non-negative values means that the resulting semantic features have greater interpretability. However, both of these methods are not probabilistic, and so they can’t capture the probability that a feature is associated with a document in the corpus.

1.2.1 Probabilistic Topic Models

Topic models were originally developed for use in text mining, and are capable of learning the major topics associated with a document based on its words. In this setting, a topic is represented as a probabilistic distribution over a space of words, which capture the tendency of words to co-occur when discussing a specific semantic topic. The idea is that a document, such as an article or a paper, is about more than one topic. When a topic is discussed, words related to that topic would occur more frequently. For example, there is a corpus of documents related to cancer. One of the topics in the corpus is related to cell death and has words such as “cell”, “death”, and “apoptosis” associated with it. When a document in this corpus has the topic “cell death” then these associated words would occur more frequently. The distributions of these topics also vary by document, and so two documents with different distributions of the same topics would also have different distributions of words. Thus, by examining a corpus of documents, it would be possible to use the words to learn what the topics in the corpus are, and what distribution of topics each document contains. Continuing from the example above, another topic in the corpus is related to cancer treatment. One document in the corpus can be focused on research of a pathway that results in cell death, and also covers it potential application in cancer treatment. A second document can be focused on different treatments of cancer, and also mentions the resulting death of cancer cells. As such, even though both documents have the same topic associations, they have different distributions of the two topics, which is reflected by the different distributions of associated words.

One of the earlier topic models is probabilistic latent semantic indexing (PLSI) [74]. It is a generative model that represents the probability of topic and word co-occurrences as mixtures of conditionally independent multinomial distributions. In this representation, the words in a document are assumed to be generated independently, which means that the ordering of the words does not matter. As such, each document is treated as a bag-of-words. Another assumption made is that the words in a document are conditionally independent given the associated topics. PLSI allows a mixture of topics to be assigned to a single document, where the topics are characterized as distributions of words, and is able to assign topic labels to previously unseen documents after the model is already learned. However, this model does not make assumptions about how to assign topic weights to documents, and so only learns topic distributions for the documents in the training set. This makes it difficult to generalize the model to new documents added to a corpus.

The most commonly used topic model is latent Dirichlet allocation (LDA), which is also the basis for many other topic models [75]. LDA is a Bayesian model of a collection of documents that is an extension of PLSI. Therefore, it is based on the same concepts as PLSI, and represents topics and documents in the same manner. In addition to the assumptions made in PLSI, LDA assumes that the topic distribution has a Dirichlet prior. For each document in a corpus, LDA sets the prior for the words to be a Poisson distribution, and the topic mixture weights to be a Dirichlet distribution. In other words, LDA is able to regenerate a document in the corpus through the following process (taken from [75]):

1. Choose N ~ Poisson(λ)
2. Choose Θ ~ Dirichlet(α)
3. For each of the N words w_n:
    a. Choose a topic z_n ~ Multinomial(Θ)
    b. Choose a word w_n from $p(w_n|z_n, \beta)$, a multinomial probability conditioned on the topic z_n.

In this notation, N is the number of words (w) in the document; z is a topic; and Θ, α, and β are model parameters. The addition of this Dirichlet prior allows LDA to generate topic probabilities for new documents. This means that LDA is a generative model that overcomes the difficulty of generalization to new documents. Yet, like PLSI, the topics learned by LDA have a flat structure to them, which means that the presence of a topic in a document is independent of the other topics that may exist in the document (Figure 1-1A). However, this independence isn't true in the altered signaling pathways of cancer. Thus, a model that is capable of capturing the relationship between topics is necessary.

While unable to identify topic relationships, one type of model that has been used to capture the relationship between documents is the relational topic model. A relational topic model takes as input documents and their associated document network, where a document is linked through citations or other means. Topics and topic assignments are then learned and used to generate a network structure, with the aim of recreating both the documents and the document network. By requiring the topics to be used to recreate the document network, the learned topics would be able to explain the original network structure. In comparison, models that do not require this coupling between the topic and the network may result in two subsets of topics, one that is used to explain the networks and the other used to explain the words. As a result, the topics would not be useful for predicting the relationship between edges and words. However, relational topic models are still unable to capture relationships between topics.

In reality, the topics that occur in a document are not independent as assumed in LDA. Instead, they have hierarchical relationships where topics that are more general and appear more frequently in a corpus are higher in the hierarchy, and topics that are more specific and appear less frequently are deeper in the hierarchy (Figure 1-1B). For example, with the document focused on a pathway that results in cell death, the topics related to this document can include the most general topic of biology, followed by increasingly more specific topics of cellular processes, cell death, and apoptosis. In this corpus there would be more documents that contain the general topic of biology than the more specific topics. Extensions of LDA have been developed that are capable of learning these hierarchical relationships between topics [76-79]. These models generate a structured prior on topics contained in a corpus of documents, which results in topics that have a hierarchical structure. This structure matches our interests because, as mentioned earlier, different alterations have different degrees of specificity. Therefore, the alterations that occur in many different cancer types would be like the topic "biology" from the example earlier, while the alterations limited to specific cancer types would be like the topic "apoptosis". The nested hierarchical Dirichlet process (nHDP) allows a document to access the entire tree and be labeled with topics that lie along different branches of the tree [77]. This gives the model greater flexibility, and does not force documents to be associated with unrelated topics. We hypothesize that establishing a framework that uses the nHDP would allow us to identify altered pathways in tumors and generate new knowledge.

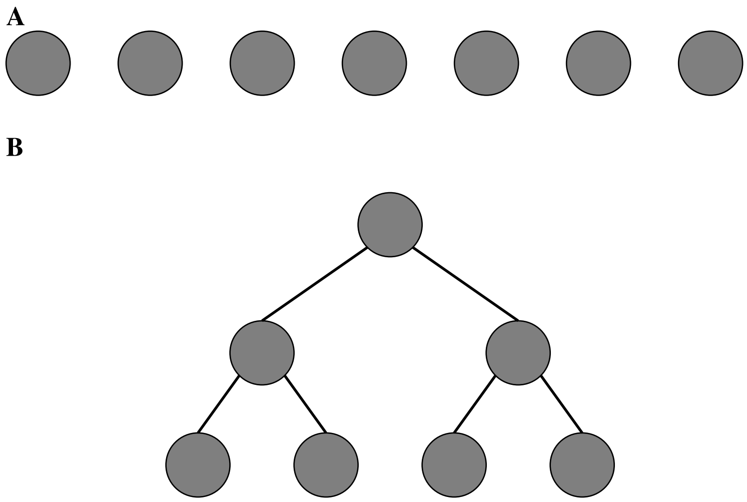


Figure 1-1. Example topic structures for the latent Dirichlet allocation and nested hierarchical Dirichlet process. Examples of two possible topic structures that can be learned through topic modeling using A) latent Dirichlet allocation or B) nested hierarchical Dirichlet process. Circles represent topics and lines represent relationships between topics. Topics higher in the hierarchical tree are more general, and become more specific as they go deeper into the tree.

To summarize, while PLSI makes it possible to generate probabilistic mixtures of topics for documents, its limited assumptions makes the model unable to generalize to documents that it is not trained on. LDA overcomes this by adding an assumption about the topic distribution for a document. Both of these models generate topics that are independent of each other, which may not be true in all datasets. Hierarchical topic models overcome this limitation, and nHDP has an increased flexibility in the generated topic tree structure over other hierarchical models.

1.2.2 Application to Biomedicine

Due to its capabilities, topic modeling has been applied to some areas of biomedical research. Since these algorithms were first developed with textual data in mind, it has been used to process biomedical text. This can include processing biomedical literature for different purposes, such as identifying the important concepts in the document themselves [80] or their associated genes [81]. Analysis of gene expression data is another area where topic models are more frequently used. The goal can be to simply classify or cluster samples based on changes in gene expression under different conditions [82-84]. Lee et al. used toxicogenomic gene expression data to try to find common functional features affected by different drugs or tested using different conditions [83]. Similarly, this has also been used to analyze high content screening data generated from different drugs [85]. In these situations, where actual words and documents do not exist, adaptations are used. Oftentimes each sample is treated as a document, each gene or endpoint as a word, and the measurement associated with the gene or endpoint as the word count [82-85]. Other less common applications also exist, such as applying topic models to study the functional core in taxonomy [86]. In most cases, LDA was the topic model used to analyze the data [80-83, 85, 86]. However, the hierarchical Dirichlet process model has been applied to gene expression clustering, because the structure of the model better captures the hierarchical structure of biological functions [84].

In a prior instance, a relational topic model was applied with the goal of identifying genomic features that can explain the phenotypic similarities of different cancer tumors. Working with Glioblastoma Multiforme tumors, Cho and Przytycka used gene expression data as a representation of disease phenotype to create a phenotype similarity network [87]. Mutations, gene copy number variations, and microRNA dysregulation were used as genomic features, where a “word” variable was created for each gene and each variation observed for the gene. These word variables were then used to generate the tumor documents. The topics and network generated were used to identify the genomic features that could explain the phenotypic similarities [87]. However, the model used is unable to capture hierarchical structure of functions. Given the nature of the representation of genes, which does not capture their function, it is difficult to identify genes with similar functions that can be shared across different subtypes.

1.3 FUNCTIONAL SIMILARITY OF GENES

The goal of using an alternative representation of genes is to be able to assess their functional coherence or similarity. This is due to the fact that gene names themselves do not provide much information. Thus, additional information about gene function is necessary in order to be able to identify similar genes. As such, a computational algorithm would not be able to find functional similarities between two different genes if we directly used gene names as words. For example, if we just had the words PI3K and AKT, they seem independent. However, once we represent the fact that they are both part of a signaling pathway involved in apoptosis their functional similarity is apparent. Through this idea of using additional information, different measures of functional coherence have been developed. Two of the larger types of methods are by using gene annotations or by using associated literature.

1.3.1 Using gene annotations to identify functionally similar genes

Databases containing controlled vocabularies of terms, such as the Gene Ontology (GO) [88], were created in order to have a unified language when annotating genes. This means that gene annotations related to gene function can be a resource for determining functional coherence. The most frequently used method is to determine if a specific annotation term is enriched in a list of genes. Thus algorithms often are based on statistical tests of over-representation such as the hypergeometric distribution [89] or Fisher's exact test [90]. Gene annotations can also be used to generate annotation profiles, which can then be used to measure the similarity between genes. For example, a *kappa* statistic co-occurrence score can be measured between genes based on their annotations [91]. These score distributions could be used to group similar genes [91].

These annotation-based methods can also be developed further by including methods such as multiple hypothesis correction [89], or incorporating GO structure into the analysis [90]. The GO has a hierarchical structure where parent nodes encompass all the information held within children nodes. Since genes are only annotated with the most specific term applicable, ignoring this structure would make it more difficult to accurately measure the functional similarity of genes. One method of incorporating the GO structure is to adjust the weight of the gene annotations used when calculating enrichment scores [90]. However, even with these improvements, annotation based methods are often restricted to analysis of gene lists. They are also limited by the amount, and accuracy, of gene annotations and annotation terms available.

1.3.2 Using literature to identify functionally similar genes

Methods that directly use the literature associated with genes are able to bypass the need for gene annotations. This requires the genes to be represented in a way that captures the information contained in the literature. The most frequently used method to do this is through a bag-of-words, vector space representation [81, 92-96]. Such a representation ignores the order that terms appear in a document, and just captures its presence in a document. These vector space representations can then be used to identify the important semantic features of genes. Methods using associated literature would also be limited by the amount of literature available. However, they do avoid the complications involved with the extra annotation step.

1.3.2.1 Representing genes in the vector space

The main variations in vector space representations lies in what method is used to track, or index, the importance of a term in a document. The simplest indexing method is binary, which assigns a 1 if the term occurs in a document and a 0 if it does not [93]. Term frequency can be directly used as an indexing measure to capture importance based on the number of times a term occurs in a document [93]. However, if a term appears in many documents in a corpus, then its presence in an individual document would not convey as much information. As such, most methods use term frequency-inverse document frequency (tf-idf) [92, 93, 96], which normalizes term frequency based on the number of documents the term appears in. Example vectors using these three indexing methods can be seen in Table 1-1. Other indexing methods that correct values based on document distributions are also available [95], as well as those that are based on if a document refers to other documents [93].

Table 1-1. Example vectors using three different indexing methods

Binary	
activ	1
apoptosi	1
associ	1
cancer	1
cell	1
codon	1
dna	1
express	1
mdm2	1
mutat	1
neurogenet	0
p53	1
protein	1
regul	1
tp53	1
tumor	1
vasodil	0

Term Frequency	
activ	79351.05
apoptosi	41463.75
associ	69460.42
cancer	104378.64
cell	150998.09
codon	10760.11
dna	37610.01
express	92302.17
mdm2	8741.60
mutat	90268.99
neurogenet	0
p53	310367.08
protein	54966.58
regul	56595.32
tp53	33462.43
tumor	63526.75
vasodil	0

tf-idf	
activ	255.15
apoptosi	307.14
associ	200.17
cancer	407.73
cell	371.00
codon	283.16
dna	287.10
express	225.13
mdm2	364.23
mutat	423.80
neurogenet	0
p53	3979.07
protein	116.70
regul	158.98
tp53	984.19
tumor	319.23
vasodil	0

1.3.2.2 Identifying semantic features associated with genes

Different methods have been used to identify the important semantic features associated with a specific gene out of the entire vocabulary space, which is generally at least a few thousand terms. Methods that have been used to identify these features include latent semantic indexing [95], non-negative matrix factorization [92, 96], and latent Dirichlet allocation [81]. All of these methods result in a gene-to-feature adjacency matrix. These adjacency matrices can then be used in a number of ways, such as clustering genes based on their feature similarities to identify related gene sets [92, 95, 96], or to rank genes based on given queries [95]. The matrices have also been used to create bipartite graphs, and the functional coherence of genes was then measured based on graph connectivity [81]. The generated features have to be further analyzed in order to determine what they represent, and identify the meanings of the featured shared between gene sets.

1.3.3 Evaluation

The evaluation methods of functional coherence algorithms have mainly focused on three different aspects: the accuracy of the algorithms, the quality of the generated gene clusters, and the stability of the algorithms. Measuring the accuracy of an algorithm involves determining if it is capable of correctly determining the functional coherence of genes. Since clustering is often used to identify clusters of functionally coherent genes, some methods have evaluated the coherence of the generated clusters. The stability of the algorithms is measured to determine their performance under different situations.

1.3.3.1 Algorithm Accuracy

Two different evaluation methods have been used to determine the accuracy of algorithms: evaluation through the use of datasets or expert evaluation. In order to correctly measure accuracy, the true functions of the genes in a dataset are necessary. Therefore, yeast datasets are frequently used, since they are well-studied and the gene functions are better understood. The results generated using yeast datasets are qualitatively analyzed to determine if they are reasonable and the gene clusters correspond with established functions [89, 91-93]. In order to ensure that the results obtained are based on the data, they can be compared with random datasets generated by shuffling the data matrix [92]. Datasets can also be generated in order to have a gold standard for evaluation. This can be done by selecting genes with known functions [95] or just artificially generating data [90]. Knowing the genes' function makes it possible to calculate the precision and recall for an algorithm that can rank functionally related genes based on a query [95]. When evaluating if an algorithm can correctly identify functionally enriched terms, artificially enriching the terms makes it possible to evaluate and compare the performance of different algorithms [90]. Some authors have run datasets without well-established gene functions, which can be used for qualitative evaluations of the algorithms [81, 91, 92]. Expert evaluation is a potential method for qualitative evaluation. Chagoyen et al. used it to determine if the semantic features generated were coherent to a specific function [92]. However, individuals may differ in opinion, and multiple experts would be needed to be able to get a more accurate measure.

1.3.3.2 Clustering Performance

In order to evaluate the performance of clustering to create cohesive gene clusters, a couple of different metrics are available. Two of the more traditional measures are the Rand index [97] and the silhouette score [98]. The Rand index is used to measure the similarity of two data clusters, and so it can be used to compare a generated cluster with a set of genes treated as the ground truth [93]. In order to correct for random partitions, an adjusted Rand index was developed which uses a hypergeometric distribution as a model for randomness [99]. If there is no gold standard for comparison, the silhouette score can be used. This score measures how well each data point lies within its cluster, and so it can also be used to determine appropriate cluster sizes for a dataset. One other method that has been used is to treat the clusters as class labels, and to calculate a misclassification score for the resulting clusters [93]. Similar to the Rand index, a gold standard is needed in order to accurately calculate the misclassification scores.

1.3.3.3 Algorithm Stability

Measuring the stability of an algorithm means determining how sensitive it is to the data and other factors. Highly sensitive algorithms are less reliable because they would give very different results under slightly different conditions. When an algorithm contains stochastic or adjustable components, such as initializations, repeated executions while randomizing the components can be performed. This evaluation method can be used to determine if the results of the algorithm are produced by chance [96]. However, this does not reveal how sensitive an algorithm is to the input data. One method of determining this is to add noise to the data, and evaluating if the algorithm is still able to perform under these conditions [92]. Another method of reducing the signal-to-noise ratio is to reduce the amount of signal available, such as by decreasing the threshold for differentially expressed genes and evaluating the stability of the results with the smaller gene sets of interest [90]. Only a combination of tests would be able to determine the stability of algorithms given different input and variables. A gold standard for data used for these tests would be needed in order to properly evaluate the results, which is difficult given the limitations of current biological knowledge.

1.4 RESEARCH OVERVIEW

It is important to identify the disease mechanisms related to cancer in order to have a better understanding of its development. Being able to understand the combinations of perturbed pathways in cancer may also lead to better treatment of the disease. Thus the identification and incorporation of driver genes to identifying the perturbed pathways is important. Since topic modeling can be used to identify common themes across different documents, it can potentially be used to identify perturbed pathways shared across different tumors. To do this, it is necessary to use a representation of genes in order to be able to identify genes that are functionally similar and part of the signaling pathway.

The aim of the research is to identify patterns in the signaling pathways perturbed in cancer. We used two types of alternative representations of genes, semantic and biological, to capture the perturbed pathways in individual tumors. Driver genes were predicted for each cancer tumor using somatic genomic alteration and gene expression data. Topic modeling was then used to identify the pathway perturbations that occur and are shared in different tumors. The methodology to do this was developed and applied to cancer data, and shows the potential to identify features using topics that can be used to separate patients into classes with different survival outcomes.

2.0 REVEALING COMMON DISEASE MECHANISMS SHARED BY CANCERS OF DIFFERENT TISSUES OF ORIGIN THROUGH SEMANTIC REPRESENTATION OF GENOMIC ALTERATIONS AND TOPIC MODELING

2.1 INTRODUCTION

Cancer is a complex disease involving multiple hallmark processes [7, 10], and aberrations in these processes are caused by somatic genomic alterations (SGAs) that perturb pathways regulating these processes. Different combinations of pathways lead to heterogeneous oncogenic behaviors of cancer cells, which impact patient outcomes and response to treatment. Identification of combinatorial patterns of pathway perturbations can reveal common disease mechanisms shared by a tumor subtype and such information can guide targeted therapy.

Transcriptomic data have been widely used to reveal different cancer subtypes among tumors of the same tissue of origin, and such studies have identified many clinically relevant subtypes, which have significant prognostic value [25-28, 67, 100-103]. However, transcriptomics-based subtyping does not provide insight into the disease mechanisms underlying each subtype, that is, transcriptomics-based subtyping does not reveal the causative pathways underlying the development of subtypes. As such, such subtyping does not provide guidance for targeted therapy. Another limitation of transcriptomics-based subtyping is that tissue-specific gene expression prevents discovery of transcriptomic patterns across cancer types.

Recent pan-cancer studies found that tumors are invariably clustered according to tissue of origins when using features that are related to transcriptomics [35, 38]. Therefore, studying common disease mechanism of cancers should be addressed from new perspectives.

In order to gain a better insight into cancer disease mechanisms, an alternative approach is to study SGAs that perturb signaling pathways with the goal of identifying which perturbed pathways underlie each of the subtypes. It can be hypothesized that each cancer subtype is likely driven by a specific combination of perturbed pathways, and identification of such common disease mechanisms would provide guidance for targeted therapy.

However, the direct use of SGA data to identify these signaling pathways is challenging. This is because pathways are composed of multiple genes, and in different tumors the same pathway can be perturbed by distinct SGAs affecting different members of the pathway. As such, two tumors sharing common pathway perturbations may exhibit totally different sets of SGAs, making it difficult to detect similarities between tumors. Thus individual tumors may present itself with different genomic alterations, while undergoing the same pathway perturbations [45]. This effect is amplified by the fact that multiple pathways need to be perturbed for cancer to develop. All of this results in highly heterogeneous mutation patterns in tumors with common pathway perturbations.

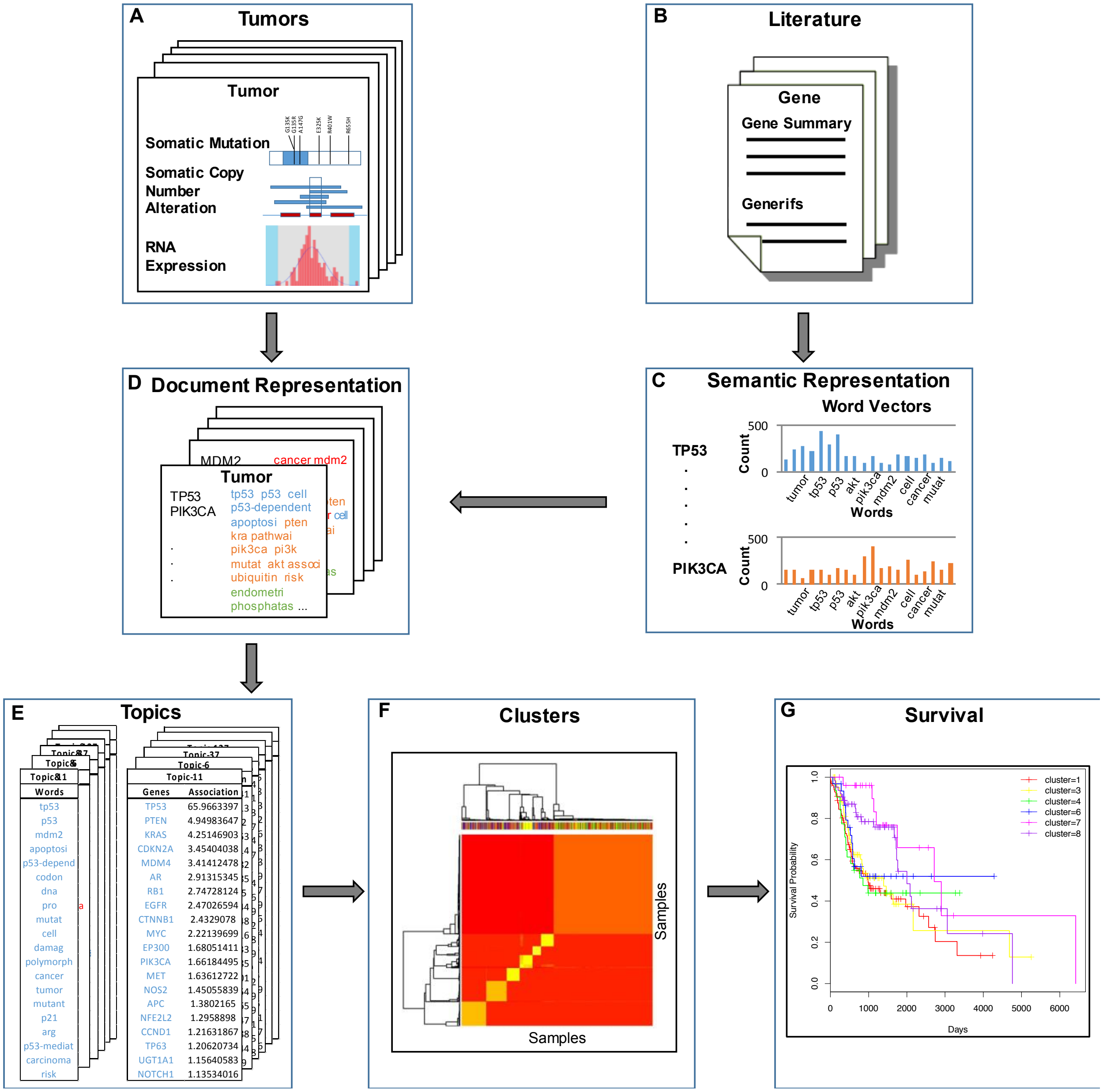


Figure 2-1. Conceptual overview of research. A) Somatic mutation, copy number alteration, and gene expression data for each tumor was collected. **B)** GeneRIF and gene summaries associated with genes were collected. **C)** The semantic data associated with each gene was processed to create a word vector representation (note the differences in the word frequency profile for different genes). **D)** A document representation for each tumor was created by combining the word vectors of each SGA associated with the tumor. **E)** The document representations were used as input for a hierarchical topic model, which identified topics associated with each tumor. **F)** The tumors were represented in topic space, and clustering analysis was applied to group tumors with

similar topic allocations. **G)** These clusters were then used to perform survival analysis on tumors of the same cancer type.

In order to tackle this problem, we have developed a novel semantic representation of tumors that captures the similarity of functions of distinct genes. This representation would help us identify functionally related genes whose alterations result in similar changes in signaling pathways. We also chose to use topic modeling to identify patterns in these altered signaling pathways based on the semantic representations. The tumors were clustered based on these patterns, and a survival analysis was performed on the results. The conceptual overview of our research is shown in Figure 2-1.

2.2 METHODS

2.2.1 Data Processing

2.2.1.1 Cancer Genomic Data

Cancer somatic mutation data was downloaded (July, 2013) from The Cancer Genome Atlas (TCGA) and copy number variation and gene expression data was downloaded from The UCSC Cancer Genomics Browser [104, 105]. Data from five different cancer types was used: breast invasive carcinoma (BRCA), head and neck squamous cell carcinoma (HNSC), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), and ovarian serous cystadenocarcinoma (OV), where the LUAD and LUSC data was combined into one large lung cancer (LUNG) dataset for processing.

2.2.1.1.1 Somatic Mutations

PolyPhen-2 was used to determine which somatic mutations for each cancer sample had a potential effect on protein function, where each cancer sample was a different cancer tumor [106]. We considered a mutation event that was labeled either "possibly damaging" or "probably damaging" to be a functional mutation. Since the tool can only analyze single nucleotide polymorphisms, it was used to analyze all the missense mutations. The frame shift, nonsense, splice site, and multiple nucleotide mutations were considered functional mutations, because of their tendency to have a larger impact on protein function. This analysis was used to determine the functionally mutated genes for each cancer sample for each cancer type.

2.2.1.1.2 Copy Number Variation

We only considered the genes whose copy number variations resulted in an altered gene expression. In order to determine if the expression of a sample was altered, we first calculated the mean and variance of the samples with no copy number variation. These values were then used to calculate the probability of a gene to be differentially expressed using a one-tailed test on a normal distribution. If the probability fell below the threshold, then we considered the expression to be altered and kept the sample for further analysis. In this analysis, we only considered the instances where the gene was marked as +/-2 in copy number, and a probability threshold of 0.01 was used. For each cancer type, we utilized the gene expression data that contained the most samples.

2.2.1.1.3 Combined Data

The somatic mutation and copy number variation data were combined in order to get a more comprehensive view of the genes that are altered in each tumor. Thus a combined sample to SGA matrix was created. This matrix contains, for each cancer sample, each gene that was either functionally mutated or had a copy number variation that resulted in an altered gene expression. In order to reduce the sizes of the datasets and decrease the chances of including passenger mutations, a frequency threshold was set and any SGA that occurred less frequently than that threshold was eliminated. The cancer types were combined and a threshold of 20 was used. This threshold was selected because we had a total of 2,396 samples and this value is close to 0.01% of the total samples.

2.2.2 Representation of Genes and Tumors

2.2.2.1 Semantic Representation of Genes

Semantic data was obtained from three different sources, which could be used as independent data sources: PubMed articles, GeneRIFs, and gene summaries. PubMed articles were downloaded from PubMed on April 10, 2013. The rest of the data was downloaded from NCBI Gene: gene to article associations was downloaded on April 9, 2013, and both GeneRIFs and gene summaries were downloaded on September 16, 2013. This text was preprocessed by removing stop words, tokenization, and Porter stemming [107].

2.2.2.1.1 Tf-Idf Calculation

We calculated the term frequency-inverse document frequency (tf-idf) of each word to determine which words contained information pertinent to a gene. To do so, we tried two different methods of generating the corpus used when calculating tf-idf score: 1) treating the gene list for each cancer type as a separate corpus whose tf-idf scores were generated separately, and 2) treating the entire list of genes as one large corpus. For each corpus generated, each gene represented a document. The term frequency (tf) and document frequency (df) were calculated for each word in each gene, with the term frequency being the number of times the word is associated with the sample, and the document frequency being the number of samples the word is associated with. Using these values, we then calculated the tf-idf for a specific word with:

$$\mathrm{tfidf}(w, d, D) = \mathrm{tf}(w, d) * \log_{10} \frac{|D|}{\mathrm{df}(w, D)}$$

where *w* represents the word, *d* is the cancer sample (or document), and *D* is the entire corpus. Thus $|D|$ represents the total number of tumors. The cumulative tf-idf for each word was calculated by summing the tf-idf score across all documents.

2.2.2.2 Semantic Representation of Tumors

2.2.2.2.1 Word Vector Creation

Word vectors containing relevant words and their term frequencies are needed for the topic modelling process. The calculated tf-idf and cumulative tf-idf scores were used to limit the vocabulary size across the entire dataset as well as for each gene. Only the 20,000 words with the highest cumulative tf-idf scores were included in the vocabulary. A word vector was then created for each gene by going through its list of 200 words with the highest tf-idf scores and including only the ones that occur in the vocabulary. We also tested the idea of altering some of the word vectors created by altering the tf-idf score for each gene's gene name and aliases. These scores were altered by setting the tf-idf score for each gene's gene name and aliases equal to the highest tf-idf score associated with that gene.

In order to obtain the word vector associated with a cancer sample, we utilized only the genes altered in that sample. For each sample, we combined the word vectors for all of the genes altered in the cancer sample. The values for each word in a sample word vector were set by summing the tf-idf scores for the word across all genes with word vectors containing the word.

2.2.3 Topic Modeling

2.2.3.1 Nested Hierarchical Dirichlet Process

The nested hierarchical Dirichlet process (nHDP) is a hierarchical topic model [108], which uses Bayesian nonparametric prior to model the covariance of topics in a training corpus. nHDP represents the relations among topics using a tree, in which a node represents a topic and a path in a tree indicates that the topics on the path have a high tendency to co-occur in documents. When modeling the topics present in a text document, nHDP allows each document to access the entire tree [108] (considering all possible topics) and places a high probability on multiple paths. The nHDP algorithm was run on the word vectors created using the cancer data, with each tumor treated as a separate document. The returned results contain a topic matrix containing the words associated with each topic, and a document-topic distribution matrix containing the number of words from each tumor (document) that are associated with each topic. We used the parameter value β_0 = 0.01, and we define the maximal level of the tree to be 3 and initialized the tree with the parameters 10, 5, and 3. The nHDP algorithm was run 10 times to generate 10 different topic models for each dataset.

2.2.3.1.1 Topic Model Selection

The model that had the highest cumulative document likelihood was selected as the best-fitting topic model. To calculate the cumulative document likelihood for a specific model, we used its outputs to obtain two matrices: 1) the matrix containing the number of words from each document associated with each topic and 2) the matrix containing the probability of each word occurring in each topic. We first calculated the probability of each topic (*t*) being associated with each document (*d*) by dividing the number of words in *d* that is associated with *t* by the total number of words in *d*. The likelihood for *d* is calculated by first calculating the word probability for each word (*w*) by summing the probability of *w* given *t* times the probability of *t* given *d*. The document log likelihood is then calculated by summing up the log of the probability of each word.

The pseudocode used to calculate cumulative document log likelihood can be found below:

```
Input:
Matrix containing number of words from each document associated with each topic
Matrix containing probability of each word occurring in each topic

Initialize:
cumulativeDocProb - a variable to store the cumulative document log likelihood

cumulativeDocProb = 0
foreach document (d)
    docProb = 0
    Calculate probability of each topic being associated with d
    foreach word (w) in d
```

```
        wordProb = 0
        foreach topic (t)
            wordProb = wordProb + p(w|t) * p(t|d)
        docProb = docProb + log(wordProb)
    cumulativeDocProb = cumulativeDocProb + docProb
```

2.2.4 Analysis

2.2.4.1 Calculating Topic to Gene Associations

Since the topics in our setting reflect the functions that are repeatedly perturbed by SGAs among all tumors, it would be interesting to know which SGAs are associated with each functional theme. However, the nHDP model only captures the association of words with topics. Further calculations were needed to determine the SGAs associated with each topic. Utilizing the topic to document association and topic to word association matrices generated by the topic model, we calculated the topic to gene associations for each topic (*t*). The topic to document association matrix contains the number of words from each document that is associated with each topic; this was used to calculate the probability of each topic being associated a document by dividing the number of words associated with a topic by the total number of words in the document. The strength of association to each gene for topic *t* was then calculated by cycling through each document (*d*). We then further cycled through each word (*w*) associated with each gene (*g*) that is in *d*. The strength of association of *g* was then calculated by summing the count for *w* times the probability of *t* given *d* and the probability of *w* given *t*.

The pseudocode used to calculate topic to gene associations can be found below:

```
Input:
Matrix containing probability of each word occurring in each topic
Word vector representation of each gene
Matrix containing number of words from each document associated with each topic

Initialize:
topicGeneMatrix - an empty topic by gene matrix

foreach topic (t)
    foreach document (d)
        Calculate probability of t being associated with d
        foreach gene (g) in d
            foreach word (w) in g's word vector
                topicGeneMatrix[t][g] = topicGeneMatrix[t][g] + w count in g * p(t|d) *
                     p(w|t)
```

2.2.4.2 Clustering Tumors

In order to determine if the topics obtained through the nHDP learned additional relationship information from the data, we performed consensus clustering to cluster the tumors using either the altered genes or the word count per topic association as features. We used partitioning around medoids (PAM) as the clustering method. The algorithm was run for cluster sizes 4-6 using 10 repetitions when clustering based on the altered genes; it was run for cluster sizes 4-10 using 20 repetitions when clustering based on the word count per topic association. Consensus clustering was performed using the clusterCons package version 1.0 in R [109].

2.2.4.3 Visualization of Tumor Clusters

In addition to consensus clustering, we also chose to visualize the tumors (documents) in order to see how clearly our topic model was able to separate the different samples. The t-Distributed Stochastic Neighbor Embedding (t-SNE) technique of dimensionality reduction was used to plot the points in a two-dimensional space [110]. This allowed us to directly use the word counts per topic for each tumors as input for plotting, after first removing the topics that do not have any associated samples. We used the Matlab implementation downloaded from http://homepage.tudelft.nl/19j49/t-SNE.html.

2.2.4.4 Calculating Cluster to Topic Associations

The proportion of samples (documents) in a cluster associated with each topic was calculated to see how topic associations vary between different clusters. In order to determine which documents are associated with each topic, the proportion of words from each document associated with each topic was calculated. Any topic that was associated with at least 0.01 of the words in a document was considered to be associated with the document. This threshold was used to remove associations that are the result of noise. We then obtained the proportion of documents in each cluster that are associated with each topic.

The pseudocode used to calculate the cluster-to-topic association can be found below:

```
Initialize:
topicClusterMatrix - an empty topic by cluster matrix

foreach cluster (c)
    Get all documents in cluster
    foreach topic (t)
```

```
        Counter number of times t is associated with documents in c
foreach cluster (c)
    Divide counts by the number of documents associated with c
```

2.2.4.5 Survival Analysis

In order to determine if there was a biological impact in subtyping the tumors based on clustering, we chose to perform a survival analysis. Kaplan-Meier survival analysis was done on the tumors with the same cancer type. These samples were separated into subsets based on the clustering results obtained previously. Survival data for the tumors were obtained on May 2, 2016 from the clinical data available on TCGA [25]. The analysis was performed twice for each cancer type: once using all tumors, and once after excluding all clusters that contained less than 25 samples. We used the survival package version 2.38.3 in R to conduct the analysis [111, 112].

2.3 RESULTS

2.3.1 Cancer Data

The combined somatic mutation and copy number variation data resulted in datasets of the following sizes (Table 2-1): BRCA with 779 samples and 15,517 genes; HNSC with 324 samples and 14,548 genes; LUAD with 398 samples and 11,851 genes; LUSC with 331 samples and 10,874 genes; and OV with 562 samples and 10,235 genes. This resulted in a dataset with 2,396 samples and 20,760 genes after combining all four cancer datasets, and 2,396 samples with 2,733 unique genes after applying a threshold.

Table 2-1. Number of tumors and number of genes for each cancer type

Cancer Type	Number of Samples	Number of Genes
BRCA	779	15,517
HNSC	324	14,548
LUAD	398	11,851
LUSC	331	10,874
OV	562	10,235
Combined	2,396	20,760
Thresholded	2,396	2,733

2.3.2 Semantic Data

Word vectors were created using the different semantic data sources. Four different combinations of data were used to generate word vectors: 1) PubMed articles, 2) GeneRIFs, 3) gene summaries, and 4) GeneRifs and gene summaries combined. The vocabulary sizes of these resulting word vectors were: 357,577 word for PubMed articles; 54,755 words for GeneRIFs; 7,933 words for gene summaries; and 57,035 words for GeneRifs and gene summaries combined.

2.3.2.1 Construction of Gene Word Vectors

We created a word vector for each gene using the different semantic data sources. Since the words used to represent a gene are related to the gene's function, the word vectors highlight the similarities and differences between two genes. A subset of words with their tf-idf scores from the word vectors of three genes are given as examples in Table 2-2. Both TP53 and MDM2 are known cancer genes. TP53 is a tumor suppressor that is involved in apoptosis and DNA repair, and MDM2 is a proto-oncogene that inhibits TP53. On the other hand, the TTN gene encodes for a protein that is important in muscles. The shared words in the word vectors for TP53 and MDM2 reflects their similarity, especially when compared against the word vector for TTN.

The number of words in a gene word vector provides information on if there is enough data to fully represent the altered gene. As such, the distribution of the length of the gene word vectors informs us about how well the genes are represented using the semantic data. The main factor that impacts the gene word vector length is the data source, because the variations applied to the word vectors are all based off of the tf-idf calculated from these datasets. The resulting distributions tell us that no single data source is able to fully represent all of the 2,726 altered genes (S Figure 1). This is especially true for the gene vectors generated using only gene summary data, as around one third of the word vectors do not even have any words. It is only after combining the GeneRIF and gene summary data that the number of genes without any words falls below 400 and a larger portion of genes are represented by a full 200 words. When using PubMed articles, the number of genes without any words associated is much smaller, but we also have fewer genes that are represented by a full 200 words.

Another factor that needs to be considered about the representations is the quality of the information of the data sources used to create the word vectors. The GeneRIFs and gene summaries provide direct information regarding the gene and its function. While this means that such information is not provided for all genes, this does limit the amount of noise. On the other hand, Pubmed articles cover a greater number of genes. However, there is a lot more noise because most of the sentences do not provide information about the genes themselves.

Table 2-2. Subset of words from word vectors for three different genes created using GeneRIFs and gene summaries with altered tf-idf scores for gene names.

TP53	
Word	**Tf-Idf**
p53	4084
tp53	4084
cell	1443
cancer	890
express	887
mutat	788
activ	683
gene	615
associ	614
protein	602
tumor	563
regul	505
carcinoma	465
role	456
apoptosi	418
result	405
function	397
pathwai	387
dna	384
suggest	371

MDM2	
Word	**Tf-Idf**
hdm2	629
mdm2	629
hdmx	629
p53	363
cell	150
cancer	136
associ	117
regul	113
activ	97
express	95
snp309	95
protein	90
risk	83
suggest	76
result	74
tumor	73
polymorph	70
ubiquitin	69
interact	66
degrad	66

TTN	
Word	**Tf-Idf**
ttn	88
titin	88
domain	31
pevk	18
region	17
protein	16
muscl	15
mutat	15
structur	14
elast	12
mechan	12
heart	11
interact	11
molecular	11
express	10
stiff	10
cardiomyopathi	10
studi	10
famili	10
sarcomer	10

2.3.2.2 Construction of Tumor Word Vectors

Word vectors were created for each of the 2,396 samples in the dataset by combining the word vectors for the genes associated with each sample. The average length of the word vector for each tumor varies depending on the semantic data source used, and can be found in Table 2-3. If we consider the average length of the word vectors as a measure of the amount of information they contain, then we can gauge the relative quality of the data sources. By comparing the numbers we can see that gene summaries on their own provide the least information, and GeneRIF and gene summaries combined provide the most. Based on this, and the gene word vectors, we decided to focus on using either PubMed articles or the combination of GeneRIF and gene summaries as a data source.

Table 2-3. Average length of cancer sample word vectors using different data sources

Data Source	Average Length
PubMed	980.1703
GeneRIF	1483.7859
Gene Summary	508.7003
GeneRIF and Gene Summary	1627.5776

2.3.3 Topic Modeling Results

The goal of using topic modeling is to capture recurrent semantic themes (defined by a set of commonly co-occurring words) that exist in text documents representing SGAs in a collection of tumors. Presence of such a theme in the corpus usually is due to the repeated occurrence of SGAs in tumors that share a common functional description (although different genes). The settings used to generate the topic models resulted in a hierarchical tree that contains a maximum of 210 topics. However, since the algorithm may not utilize all of the topics when learning the hierarchical structure, the actual number of topics used can vary across the models. Table 2-4 lists the average number and standard deviation of topics used over 10 runs of the nHDP algorithm for the two data sources, and the two word vector generation methods described in section 2.2.2.2.1.

Table 2-4. Average number and standard deviation of topics used across different sources and word vector generation methods

Data Source and Word Vectors	Average Length	Standard Deviation
PubMed Tf-Idf	201.1	4.508
PubMed Adjusted Tf-Idf	186.4	6.381
GeneRIF and Gene Summary Tf-Idf	195.6	5.680
GeneRIF and Gene Summary Adjusted Tf-Idf	202.4	4.222

Each of the topics has different word associations, which can be used to gain a better understanding of the types of functions and genes that are associated with the topic. We inspected the words that constitute the topics and the SGAs associated with them, and an example topic is shown in Figure 2-2. It is clear this topic is related to *BRCA1/2* genes and their relationship to cancer, particularly breast and ovarian cancers. The main function of *BRCA1/2* is related to DNA repair, and we found words related to DNA repair in the topic but they did not rank high enough to be shown in the figure, which only shows the top 20 words. Interestingly, *RAD51* gene, another DNA-repair gene that binds with *BRCA2* [113] and is regulated by *BRCA1* [114], is ranked high, indicating that the nHDP model was able to capture the DNA-repair theme. Similarly, three genes that are strongly associated with this topic are *BRCA1*, *BRCA2* and *TP53*; all are related to DNA repair, and they commonly occur in breast and ovarian cancers.

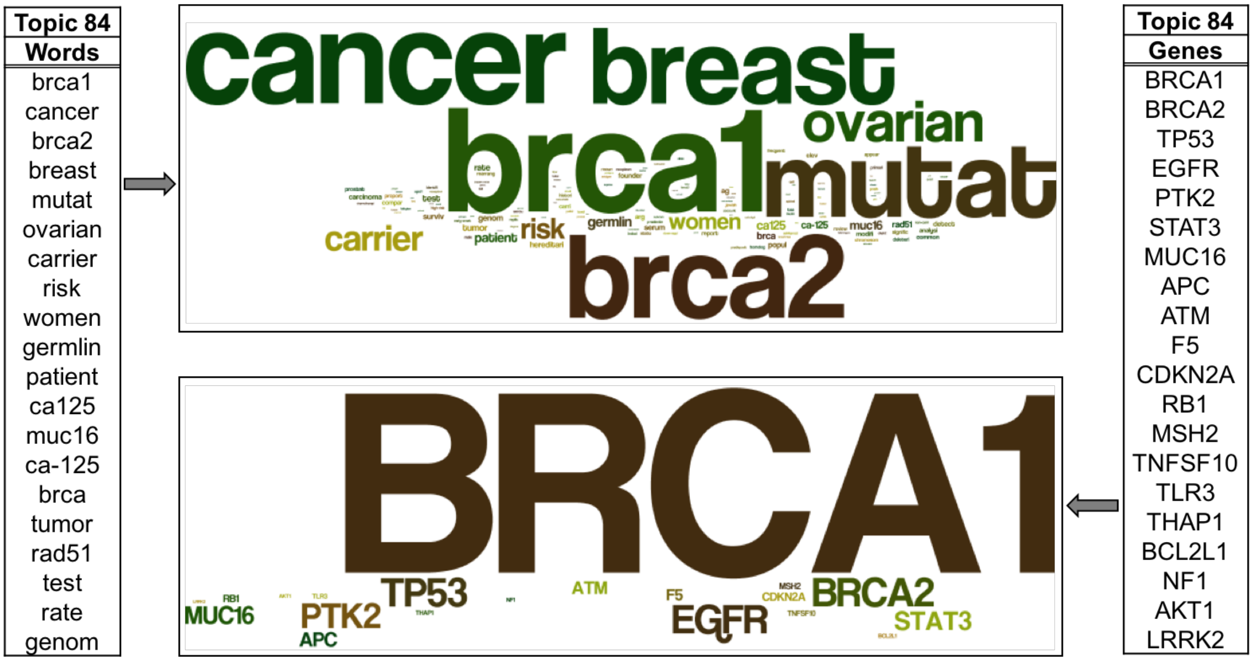


Figure 2-2. Example topic associations. The top 20 words and SGAs for topic #84 are shown. On the left are the words associated with the topic, and on the right are the SGAs that are associated with the topic. In the center are the word cloud representations of the words and genes, on the top and bottom respectively.

2.3.4 Clustering Tumors

Clustering tumors allowed us to compare between using genetic alterations directly and using the topic model. We found that clustering based on altered genes did not result in clean clusters for any of the cluster sizes (Figure 2-3A). In comparison, there was much clearer separation in the tumors for all of the different datasets used when clustering by topic associations. This indicates that there was too much noise and variability to be able to cluster samples using the genetic data directly. On the other hand, the clearer separation using the topic association indicates that the topic generated contains additional information learned about the relationship between genes. The clustering result using topics generated with GeneRif and gene summary data with altered tf-idf scores for gene names shown in Figure 2-3B. The results shown has 8 clusters, however one cluster only contains 2 samples.

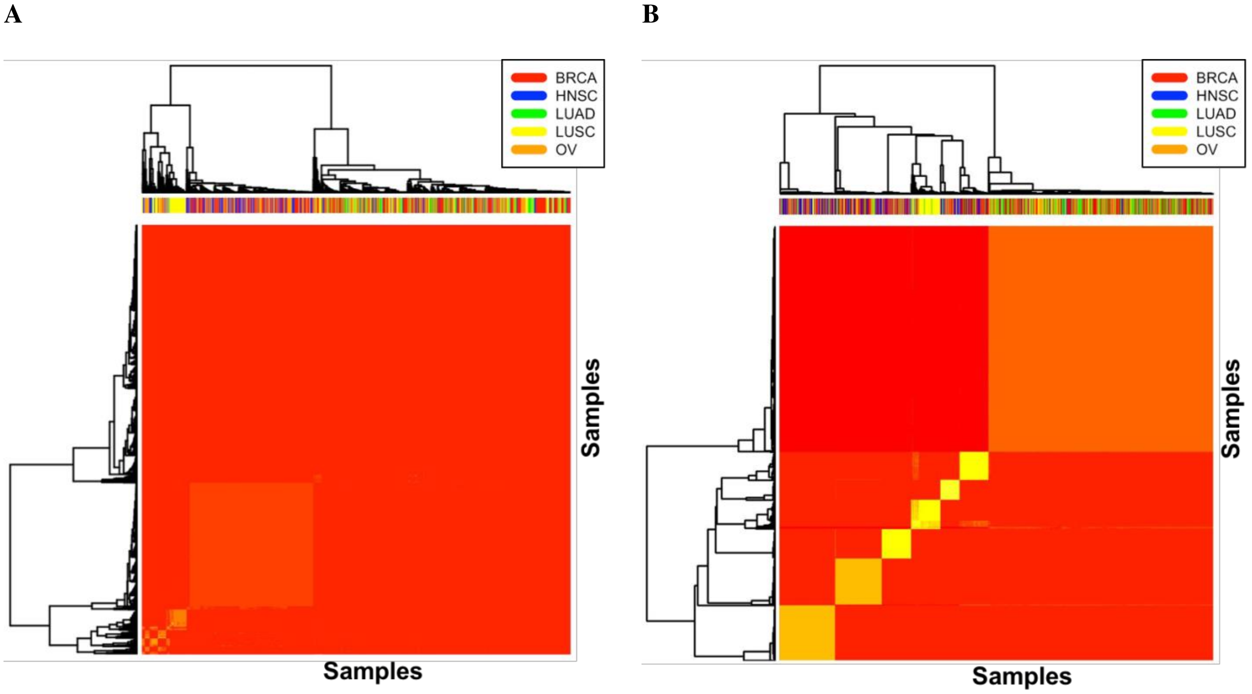


Figure 2-3. Clustering of tumors. A) Samples were clustered using genomic alteration data. **B)** Samples were clustered using topic associations generated using GeneRif and gene summary data with altered tf-idf scores for gene names.

2.3.5 Visualizing Tumors

Using t-SNE to visualize the tumors allowed us to see how well the topic representation was able to separate them. We visualized the tumors using the topics generated with GeneRif and gene summary data with altered tf-idf scores for gene names. When we labeled the tumors based on the 8 clusters generated, there was a fairly clear separation of the different clusters (Figure 2-4A). In comparison, labeling the tumors by cancer type shows that there is not a clear separation of cancer types (Figure 2-4B). This supports the theory that different tumors of the same cancer type may have different disease mechanisms that lead to the development of cancer, and these disease mechanisms may instead be shared by tumors of other cancer type.

2.3.6 Topic to Cluster Associations

A key motivation of employing nHDP, instead of other probabilistic topic models such as the LDA model, is that nHDP not only detects recurrent themes but also, importantly, the covariance structure of topics. In other words, if a topic represents a pathway perturbed by SGAs, nHDP can capture the combinatorial patterns of pathway perturbations. We examined and illustrated the example topic allocation trees (Figure 2-5). Apparently, the pattern of topics associations differed between clusters, and certain subtrees are strongly associated with one cluster but not the other. This implies that the combination of semantic (functional) themes, rather than the possession of unique functional themes, is what separates the different clusters. While we found that many topics would show up in multiple clusters, there are other more specific topics that are exclusive to one cluster. This was expected, because the topics that are higher in the hierarchy are more general and could be shared across clusters. However, the topics deep in the hierarchy are more specific and so should appear in fewer clusters.

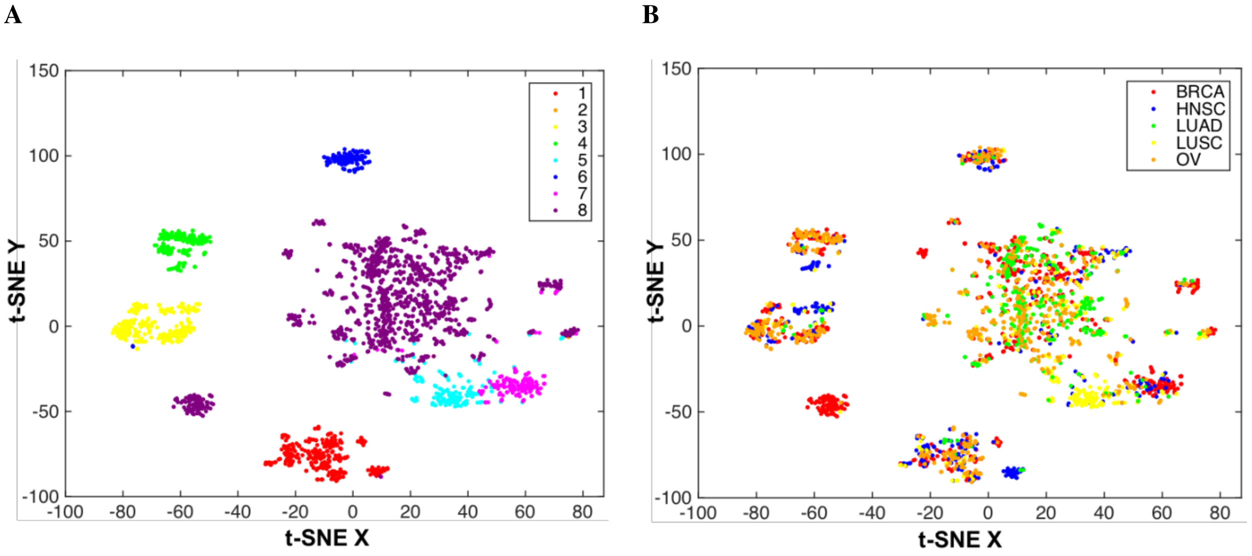


Figure 2-4. Visual representation of distance between tumors. Topic representations were used to calculate the t-sne distance between individual tumors. **A)** Tumors are labeled based on the clusters identified in section 0 and seen in Figure 2-3B. **B)** Tumors are labeled based on their cancer type.

2.3.7 Survival Analysis

Assuming that different clusters consist of tumors sharing common disease mechanisms, we performed survival analysis to determine if such subtyping reveals clinical differences. Using the 8 clusters generated in section 2.3.4 to group the tumors, we performed survival analysis on each of the different cancer types. Of the five cancer types, BRCA, HNSC, and LUSC were all found to be significant (S Table 1). This was true both when all samples and clusters were used, and when only the clusters containing at least 25 samples were used. The resulting survival curves can be seen in Figure 2-6. These results indicate that semantic representation and clustering revealed cancer subtypes that have significantly different tumors with biologically different features, which were identified using their topic associations.

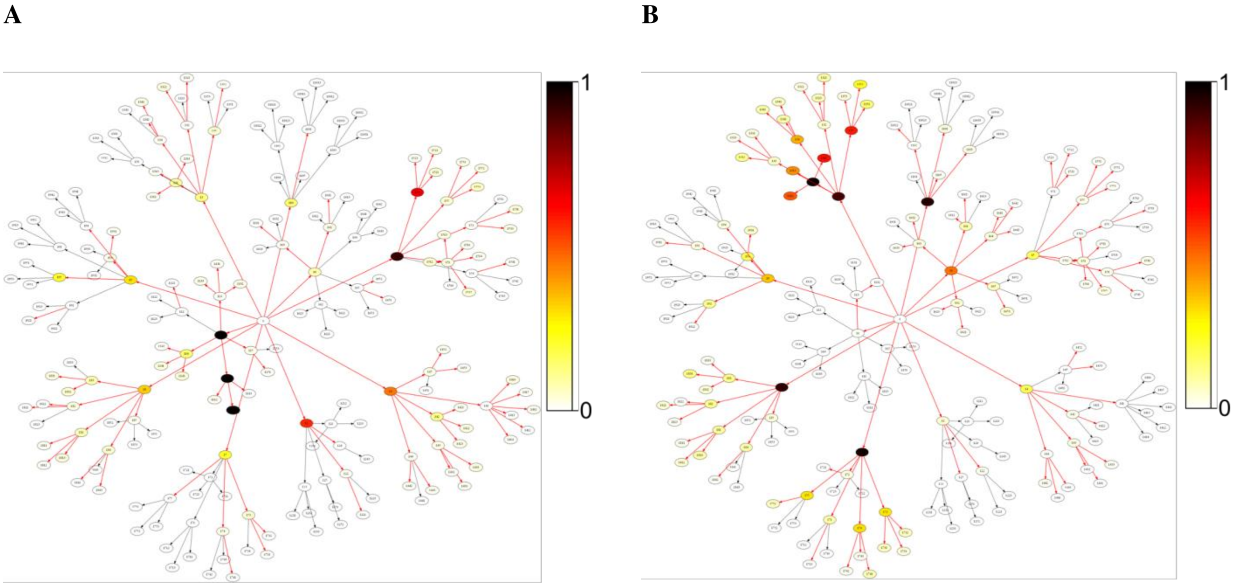


Figure 2-5. Graphical visualization of cluster-to-topic associations. The calculated degree of cluster-to-topic associations for two of the clusters using the clustering results seen in Figure 2-3B. These visualizations show the structure of the topic tree, where each node represents a topic. The color scale denotes the proportion of tumors in a cluster associated with each topic, where white means that none of the tumors in the clusters are associated and black means that all of the tumors are associated with the topic. The visualization for the topics associated with clusters 4 and 5 are shown in A and B respectively.

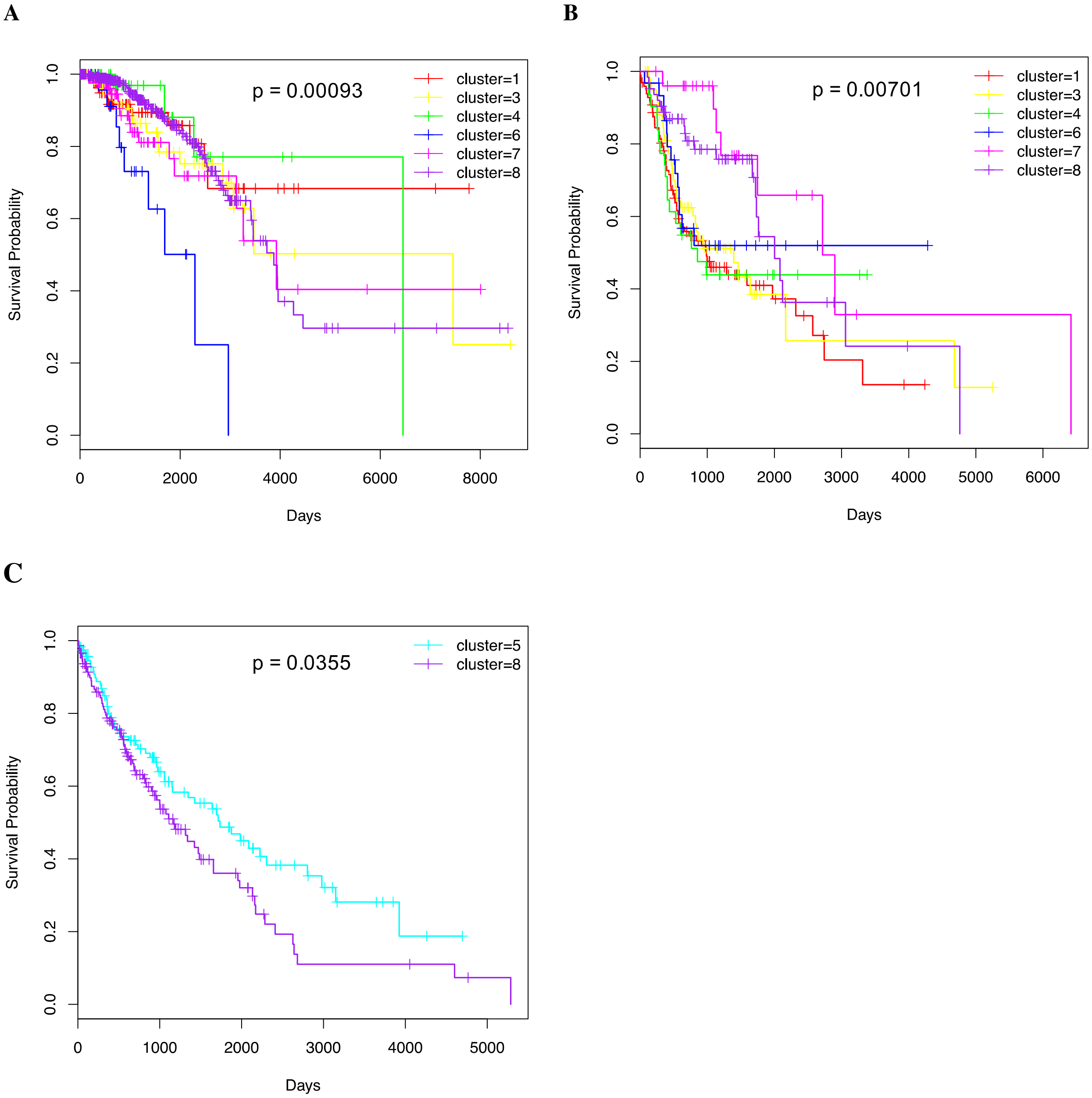


Figure 2-6. Survival analysis of tumors. The survival analysis curves calculated using only the clusters that contain at least 25 samples. A, B, and C correspond to cancer types BRCA, HNSC, and LUSC respectively.

2.4 DISCUSSION

In this study, we investigated the utility of semantic representation and topic modeling for identifying combinatorial patterns in signaling pathway perturbations in different tumors. Our results show that semantic representation of SGAs makes it possible to detect the functional similarity of different genes, which in turn enabled nHDP to detect recurrent combinatorial patterns of pathway perturbation. Interestingly, this approach enabled us to identify cancer subtypes (clusters) consisting of tumors with quite diverse tissues of origin, which exhibit significantly different clinical outcomes (survival).

To our knowledge, this is a novel approach to studying common disease mechanisms using genomic alteration data. Our approach is the first to generate semantic representations to capture the functional information of tumors. We conjecture that the existence of topics in this new representation is due to recurrent SGAs that perturb genes involved in a common biological process or pathway. As such, one can further hypothesize that the presence of a topic in a tumor represents that a specific pathway is perturbed in the tumor. Following the same vein of thinking, one can hypothesize that tumors within a cluster identified in this study share a common disease mechanism, i.e., they share a particular combinatorial pattern of pathway perturbation. Further in depth analysis of topics and associated SGAs is needed to examine if such a hypothesis is supported by the results. If proved to be the case, our finding can potentially guide therapy targeting specific combination of pathways.

This study also has its limitations. Semantic data is limited by the amount and breadth of research available, so genes that are not well research or functions that have not been discovered would not be properly represented. This was seen with some of the semantic datasets tested, where there may be too much noise or being represented by a limited number of words. Using semantic data also means that the topics generated are composed of words, which makes it difficult to identify the underlying genes that led to these associations. While the fact that we were able to identify patterns shared across multiple cancer types is promising, the input data is limited to only five cancer types.

2.5 CONCLUSION

Our research is the first time semantic representations are applied in this way to represent cancer samples, as well as the first use of a hierarchical topic model in this aspect of biomedical research. Applying topic modeling to the semantic representations of tumors made it possible to identify combinatorial patterns of perturbed pathways in cancer tumors. This enabled the identification of cancer subtypes containing different tissues of origin that exhibit significantly different survival outcomes. If these subtypes are shown to share combinatorial patterns of pathway perturbations, then these methods can potentially be used to guide targeted therapy of cancer.

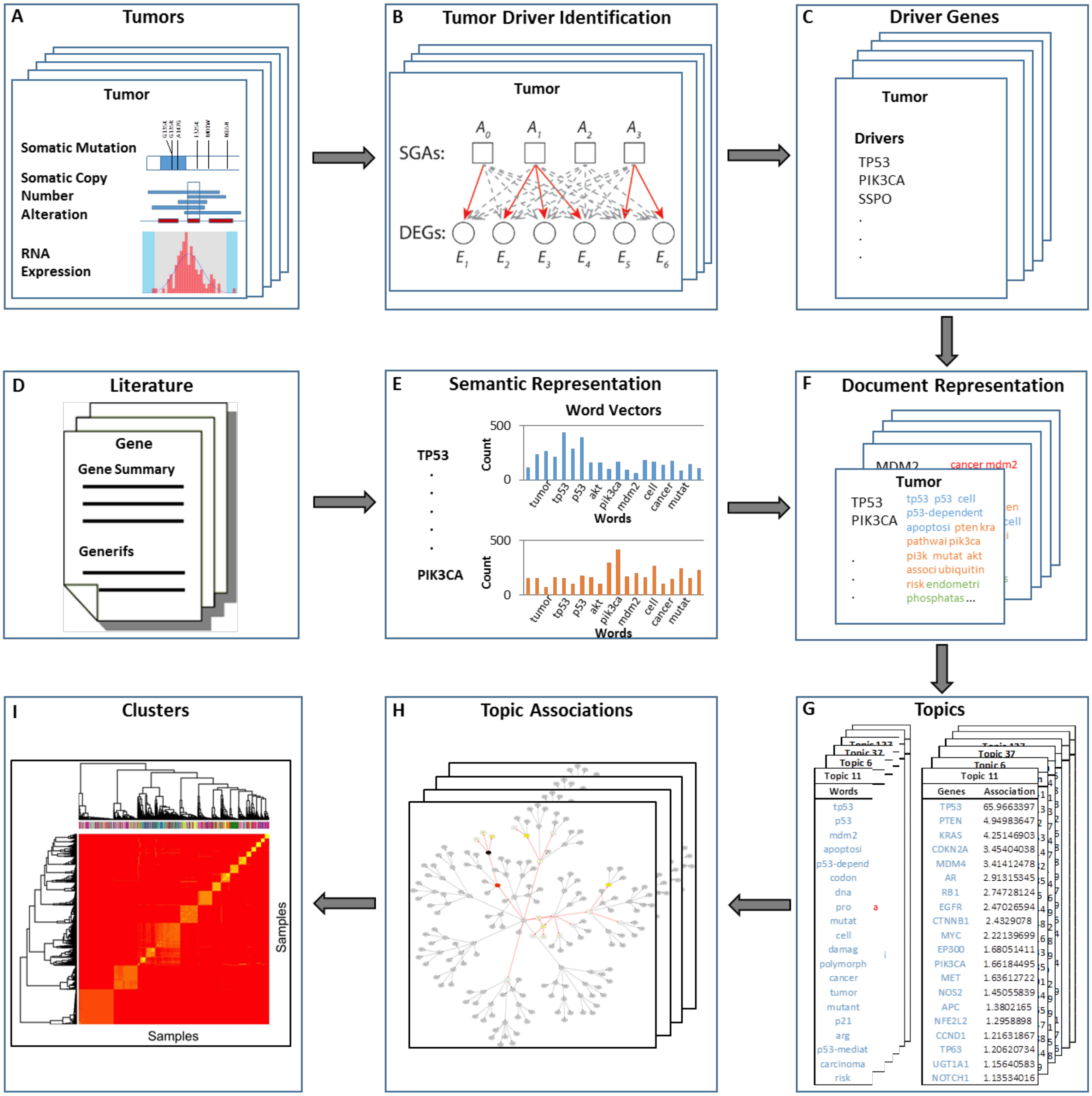
3.0 SEMANTIC MODELING OF DRIVER GENOMIC ALTERATIONS TO IDENTIFY PATHWAY PERTURBATION PATTERNS IN CANCER

3.1 INTRODUCTION

In the previous study, we found that it was possible to identify cancer subtypes that exhibit distinct clinical outcomes through the application of semantic representation and topic modeling to tumors of a limited number of cancer types. We wanted to determine if the methods were capable of identifying pathway perturbations shared across different cancer types for a larger pan-cancer dataset. This would simplify the process of identifying shared features and treatment methods across different cancer types. However, increasing the number of cancer types means that the number of somatic genomic alterations (SGAs) would also increase.

We had tried to limit the mutated genes to potential drivers using Polyphen-2, by only keeping the genes that were predicted to have a functional impact. However, just because a mutation may impact the function of a gene, that does not mean the resulting functional change would drive the development of cancer. This tool also cannot be used to evaluate copy number alterations. It appeared that the topic modeling algorithm was able to further screen the SGAs by identifying those that had functions with a biological impact. However, the topic modeling algorithm isn’t designed to identify driver genes, and so it is only able to filter out the less common passenger alterations. Even though topic modeling should continue to filter out nondriver SGAs, adding additional cancer types increases the difficulty of identifying combinations of altered pathways in tumors. Therefore, it may be useful to limit the input to cancer drivers, which would allow the topic modeling algorithm to work with an input that contains less noise.

To focus on these questions, we continued using both the semantic representations we developed previously and topic modeling. We expanded the dataset from 5 cancer types to a total of 17 different cancer types. In order to study the impact of using drivers as input in lieu of all SGAs, drivers were identified using a Bayesian network-based framework. The results were compared using cluster and survival analysis. The conceptual overview of our research for the driver dataset is shown in Figure 3-1, while the overview of our research for the SGA dataset is the same as that seen in Figure 2-1.



Words
tp53
p53
mdm2
apoptosi
p53-depend
codon
dna
pro
mutat
cell
damag
polymorph
cancer
tumor
mutant
p21
arg
p53-mediat
carcinoma
risk

Genes	Association
TP53	65.9663397
PTEN	4.94983647
KRAS	4.25146903
CDKN2A	3.45404038
MDM4	3.41412478
AR	2.91315345
RB1	2.74728124
EGFR	2.47026594
CTNNB1	2.4329078
MYC	2.22139699
EP300	1.68051411
PIK3CA	1.66184495
MET	1.63612722
NOS2	1.45055839
APC	1.3802165
NFE2L2	1.2958898
CCND1	1.21631867
TP63	1.20620734
UGT1A1	1.15640583
NOTCH1	1.13534016

Figure 3-1. Conceptual overview of research for the driver dataset. A) Somatic mutation, copy number alteration, and gene expression data for each tumor was collected. **B-C)** The genomic alteration data was used as input for the tumor driver identification algorithm in order to identify the drivers associated with each tumor. **D)** Generif and gene summaries associated with genes were collected. **E)** The semantic data associated with each gene was processed to create a word vector representation. **F)** A document representation for each tumor was created by combining the word vectors of each driver associated with the tumor. **G)** The document representations were used as

input for a hierarchical topic model, which identified topics associated with each tumor. **H-I)** The generated topic associations were used to cluster the tumors.

3.2 METHODS

3.2.1 Data Processing

3.2.1.1 Cancer Genomic Data

Cancer somatic mutation data was downloaded from The Cancer Genome Atlas (TCGA) [25]. Copy number variation GISTIC2 results and three different platforms (RNASeqV2, RNASeq, and Microarray) for gene expression data were downloaded from Broad GDAC Firehose [115]. Pan-cancer data covering 17 different cancer types was used: bladder urothelial carcinoma (BLCA), breast invasive carcinoma (BRCA), colon adenocarcinoma (COAD), esophageal carcinoma (ESCA), glioblastoma multiforme (GBM), head and neck squamous cell carcinoma (HNSC), kidney renal clear cell carcinoma (KIRC), kidney renal papillary cell carcinoma (KIRP), liver hepatocellular carcinoma (LIHC), lung adenocarcinoma (LUAD), lung squamous cell (LUSC), ovarian serous cystadenocarcinoma (OV), prostate adenocarcinoma (PRAD), rectum adenocarcinoma (READ), stomach adenocarcinoma (STAD), thyroid carcinoma (THCA), and uterine corpus endometrial carcinoma (UCEC).

3.2.1.1.1 Somatic Mutations

The somatic mutation data downloaded was classified into 11 different types, with three general categories: insertions, deletions, and mutations. We considered all non-synonymous and non-silent mutations to be functional mutations (S Table 2). As such, the only type of mutation that was treated as non-functional was silent mutations. Based on the functional call of each gene, a binary vector representation for each cancer sample was generated. Each cancer sample was represented by multiple genes, where a 1 represents that a functional mutation has occurred, and a 0 represents that a functional mutation has not occurred.

3.2.1.1.2 Copy Number Variation

GISTIC2 thresholds the copy number variation data for each gene in a cancer sample to one of five levels: -2, -1, 0, 1, or 2. These levels represent homozygous deletion, single copy deletion, diploid normal copy, low copy number amplification, and high copy number amplification respectively. Since we wanted to consider genes that are more likely to have a functional impact, we only treated the genes with homozygous deletions (-2) or high copy number amplifications (2) as copy number alterations.

These genes were further filtered to eliminate the genes that were not consistently deleted or amplified across different tumors in a specific cancer type. The reasoning behind this is that if a gene perturbation has a functional impact on cancer development, then this gene should be consistently altered. For example, deletion of a tumor suppressor gene would promote cancer development, therefore we would expect to see that the gene has consistently undergone deletions in tumors instead of amplifications. As such, we discarded genes that were not consistently deleted or amplified. We calculated the ratio of number of tumors amplified to number of tumors deleted, and discarded any gene that had a value greater than 1:3 or smaller than 3:1. A binary vector representation for each cancer sample was generated. Each cancer sample was represented by multiple genes, where a 1 represents that a copy number alteration has occurred, and a 0 represents that a copy number alteration has not occurred.

3.2.1.1.3 Gene Expression Data

There is gene expression data for three different platforms available on TCGA: RNASeqV2, RNASeq, and Microarray. RNASeqV2 is the most frequently used platform for measuring gene expression of TCGA tumors, with a smaller portion of tumors being measured by the other two platforms. All of the gene expression datasets for each cancer type were downloaded. For each cancer type, we picked the platform that covered the largest number of tumors and also contained measurements for normal samples. Measurements for normal samples were needed for their use in determining which genes were differentially expressed. Prior to analysis, any gene expression value in RNASeqV2 or RNASeq data that was less than 20 were considered noise. In normal data, we set these values to 20 before calculation; in tumor data, we considered these genes to have normal expression and excluded them from further analysis.

For each cancer sample, we were only interested in the genes that had an altered expression. In order to identify the differentially expressed genes (DEGs) for each sample, the expression level of genes in tumors were compared against normal cells of the corresponding tissue type. For the genes whose expression in normal cells followed a Gaussian distribution, we used their mean and variance to calculate the p-values for each gene in each cancer sample. If the p-value fell within a 0.05 threshold on either tail, then the gene was considered to be differentially expressed in the corresponding cancer sample. For the genes that did not follow a Gaussian distribution due to low variance (less than 0.1), we used fold change to determine differential expression. Fold change was calculated by dividing the gene expression of the tumor cell by the average expression of the normal cells. Any gene that was determined to have undergone a 3-fold change was considered differentially expressed.

We wanted to identify the DEGs that were driven by somatic mutations. Therefore, all genes that underwent amplification or deletion were not considered differentially expressed. This eliminated the chances that gene differential expression was the result of copy number alterations. A Pearson correlation analysis was performed to identify tissue specific DEGs that were correlated with cancer type or tissue of origin. Any DEG with a correlation value larger than 0.9 was removed. A binary matrix representation for all tumors was generated. Each cancer sample was represented by multiple genes, where a 1 means that the gene is differentially expressed, and a 0 represents that the gene is not differentially expressed.

3.2.1.1.4 Combined Data

The somatic mutation and copy number variation representation data were combined to form a binary somatic genomic alteration (SGA) matrix, where 1 represents a somatic mutation, copy number alteration, or both, and 0 represent that neither have occurred. In order to minimize redundant SGA information, SGAs with similar patterns were grouped together. We first merged neighboring genes to form a SGA unit when the co-occurrence of their SGAs over the union of tumors was larger than 0.9. We then grouped genes or SGA units into SGA groups if they shared the exact same SGA pattern.

3.2.2 Cancer Driver Data

We used driver data that was calculated and generated by other members in our lab. This algorithm to predict the causal relationship between SGAs and DEGs for individual tumors is a Bayesian network-based framework developed in our lab [116]. We assume that individual DEGs may be caused by individual SGAs, though it is possible that some are caused by non-

SGA factors. Using initial stricter assumptions that SGAs on a common pathway are mutually exclusive and a DEG is the result of one altered pathway, then it can be assumed that a DEG is most likely caused by one SGA. The problem of identifying driver SGAs and their associated DEGs was represented as a tumor-specific model using a bipartite causal Bayesian network with two sets of variables A_{set} and G_{set}, where causal edges can be added from variables in A_{set} to G_{set} (Figure 3-2). A_{set} is composed of all SGAs in the given tumor, as well as a "leak node" to represent non-SGA factors. G_{set} is composed of all DEGs in the given tumor. The algorithm learns the structure of the Bayesian network for each tumor, based on the SGAs and DEGs for the tumor. The posterior probability for a structure *M* given the data *D* is:

$$P(M'|D) = \frac{P(D,M')}{P(D)} = \frac{P(D,M')}{\sum_M P(D,M)} = \frac{P(D|M') * P(M')}{\sum_M P(D|M) * P(M)}$$

where the sum is taken over all possible models. The term $P(M)$ denotes the prior probability that the Bayesian network has *M* as its structure. The term $P(D|M)$ can be derived as follows:

$$P(D|M) = \int_\theta P(D|M,\theta) * P(\theta|M) d\theta$$

where θ represents the parameters associated with *M*.

The driver identification algorithm was run using the generated SGA and DEG binary matrices as input. A binary matrix representation for all tumors was generated using the drivers identified by the algorithm. Each cancer sample was represented by driver genes, where a 1 means that the gene is considered a driver for that sample, and a 0 means that the gene is not considered a driver.

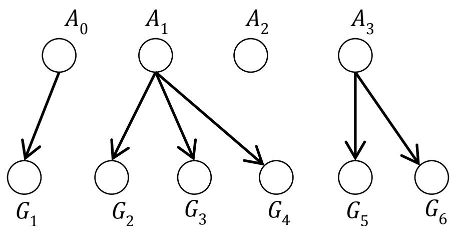


Figure 3-2. Tumor-specific driver identification structure. A bipartite Bayesian network that represents a hypothesis about which SGAs are causing which DEGs. In this network, A_1 represents a driver gene for G_2, G_3, and G_4. A_2 represents a passenger gene.

3.2.3 Representation of Genes and Tumors

3.2.3.1 Semantic Representation of Genes

The data that was downloaded and used to generate the semantic representation of genes were the same as those described in section 2.2.2.1. The processing and calculations were also performed in the same way.

3.2.3.2 Semantic Representation of Tumors

Two different sets of semantic representations were generated for the tumors. The first set used all of the SGAs associated with each cancer tumor. The second set used only the identified driver genes associated with each cancer tumor. Here we expanded the driver SGA units or groups back to their component genes. This is because we only had semantic representations of individual genes, and would not be able to directly represent a set of genes. The tumor representations generated for these datasets were different, because of the differences in the total number of genes in the dataset.

Word vectors containing relevant words and their term frequencies were generated for the topic modelling process. Word document (gene) frequency and normalized tf-idf scores were used to limit the vocabulary size both across the entire dataset and for each gene. The normalized tf-idf score for a word was calculated by dividing the cumulative tf-idf score by the number of genes the word appears in. We first trimmed the words that occurred in less than a set threshold number of genes. The thresholds we used were 26 words for the SGA dataset, and 3 words for the driver dataset. We then further trimmed the vocabulary by removing the 1,500 words with the smallest normalized tf-idf scores. Trimming the vocabulary allowed us to remove words that were too common and too rare to provide useful information in the topic modeling process; it also limits the number of features in the modeling process.

In order to create the word vector associated with a tumor, we utilized only the genes or drivers positively associated with that sample (has a value of 1). For each sample, we combined the word vectors for all of the positively associated genes in the tumor. During the process, if a gene word vector contained its own gene name or alias, then the tf-idf score was altered. We set this altered score to be equal to the smaller of the following two values: the highest tf-idf score associated with that gene, or 1.5 times the second highest tf-idf score. The values for each word in a tumor word vector were set by summing the tf-idf scores for the word across all the genes with word vectors containing the word.

3.2.4 Topic Modeling

Both the topic modeling algorithm and the topic model selection process were the same as those described in section 2.2.3. The algorithm was run on the two corpuses of tumor word vectors, which resulted in a set of word-to-topic and document-to-topic distribution matrices for each corpus.

3.2.5 Analysis

3.2.5.1 Evaluating Semantic Representation of Genes

We wanted to determine if genes with similar functions had word vector representations that were closer in similarity, in order to ensure that the semantic representation captures the functional similarity of genes. Cosine similarity was used to measure the similarity between gene word vectors. We used the genes on the KEGG pathway hsa05200 (pathways in cancer) to obtain our list of functionally related genes [117, 118]. An equal number of randomly selected genes was used as our list of functionally unrelated genes. The cosine similarity of each pair of genes for each list was calculated. The cosine similarity distribution for both lists were compared using the Wilcoxon rank sum test to determine if there was a significant difference between functionally related and unrelated genes.

3.2.5.2 Calculating Topic to Gene Associations

The method used to calculate the topic to gene associations was the same as in section 2.2.4.1. Only the top 20 genes associated with a topic that have an association score of at least 0.001 were used for further analysis.

3.2.5.3 Topic Analysis

In order to have a quantitatively comparable method of measuring the functional similarity of the genes associated with the generated topics, we chose to use the protein-protein interaction (PPI) ratio. For a list of genes, this ratio measures the number of existing PPIs over the total number of possible interactions. The idea is that functionally similar genes would have a greater number of PPI when compared to randomly selected genes, and so would have a larger PPI ratio. This ratio was calculated using the following equation:

$$R_{PPI} = \frac{I}{g(1-g)},$$

where I is the number of interactions in the gene set, and g is the total number of genes in the gene set. Human PPI data version 3.4.127 was downloaded from BioGrid [119].

The PPI ratio was calculated for the genes associated with each topic. For each gene list length, we generated random gene lists of equal length by randomly selecting from the list of all SGAs for the SGA dataset and the list of all drivers for the driver dataset. A total of 10,000 random draws were obtained for each length, which was used to create a PPI ratio distribution for random genes. For each topic, we determined where the PPI ratio fell in the random distribution by counting the number of values that were smaller than the calculated topic PPI ratio. Dividing this count by the number of samples gave us the proportion of randomly generated PPI values that the topic PPI ratio was larger than.

3.2.5.4 Clustering Tumors

We performed consensus clustering on the tumors to determine if relationship information was learned from the data by using topic modeling. As such, clustering was performed while using as input the SGA dataset directly, the driver dataset directly, the word count per topic associations generated using the SGA dataset, and the word count per topic associations generated using the driver dataset. We used partitioning around medoids (PAM), k-means, and merging the two for consensus as the clustering methods. The algorithm was run for cluster sizes 10-25 when clustering directly based on SGAs or drivers, and cluster sizes 10-30 when clustering based on the word count per topic associations. Consensus clustering was performed using the clusterCons package version 1.0 in R [109].

3.2.5.5 Cluster Evaluation

We used the gene alterations associated with each cluster in order to evaluate if the samples share common combinations of pathways. These gene lists were used to calculate protein-protein interaction ratios. The reasoning is that genes sharing pathways would have a larger number of protein-protein interactions than unrelated genes. As such, if tumors share common combinations of pathways, then there should be more PPI between their genes than samples with combinations of pathways that are unrelated. For the clusters generated using either dataset, the SGA list for each cluster was obtained by compiling all of the SGAs associated with each cancer sample in a cluster. For the clusters generated using the driver dataset, the driver list for each cluster was obtained by compiling all of the drivers associated with each cancer sample in a cluster. The method of calculating the PPI ratio and comparing against the random distribution was the same as described in section 3.2.5.3. A total of 10,000 random draws was used to generate the random distribution for gene list length.

3.2.5.6 Visualization of Tumor Clusters

The topic representation of the tumors was visualized in a two-dimensional space using the same methods as those described in section 2.2.4.3.

3.2.5.7 Survival Analysis

The biological impact of subtyping the tumors based on clustering was measured by performing survival analysis on each cancer subtype separately. Survival data for the tumors was obtained from the TCGA project clinical data using the National Cancer Institute's Genomic Data Commons: http://gdc.nci.nih.gov/. The survival analysis was performed using the same method as those described in section 2.2.4.5. The only change made was to exclude any cluster that contained less than 20 samples when analyzing each individual cancer type.

3.3 RESULTS

3.3.1 Cancer Data

For the 17 different cancer types downloaded, when we kept the tumors that had both somatic mutations and copy number alteration data available, we had a total of 5608 tumors. The breakdown of the number of samples in each cancer type is listed in S Table 3. This data resulted in a total of 38,004 SGAs. This final SGA to sample mapping was used as the somatic genomic alteration (SGA) dataset.

When we restricted the samples by the copy number alteration data, we had a total 4468 tumors. This also resulted in the removal of THCA data, because of the quality copy number alteration data. The breakdown of number of samples in each cancer type is listed in S Table 3. This data resulted in a total number of 26,203 SGAs. From this genomic data we obtained a final list of 721 identified drivers, with any SGA unit or group being treated as an individual driver. If we expand these SGA units or groups back to their component genes, we obtain a list of 733 genes. This final gene to sample mapping was used as the driver dataset. Given that two different cancer sample gene association datasets were used, the SGA dataset and the driver dataset, a different semantic representation was generated for each.

3.3.2 Semantic Data

The initial vocabulary size for the SGA dataset was 167,314 words. Of the 26,203 SGAs there were 8,724 that had less than 5 words associated. After trimming the vocabulary, we had a final size of 6,396 words. This also resulted in 859 new genes that had less than 5 words associated. However, we felt that this increase was acceptable because it accounted for less than 5% of the genes that previously had more than 5 word associations (4.91%).

Similarly, the initial vocabulary size for the driver dataset was 31,869 words. Of the 733 genes there were 146 that had less than 5 words associated. After trimming the vocabulary, we had a final size of 6,029 words. This resulted in 21 new genes that had less than 5 words associated. Once again, this accounted for less than 5% of the genes that previously had more than 5 word associations (3.78%).

3.3.2.1 Semantic Representation Evaluation

The vector representations of genes were used to calculate and compare the distribution of cosine similarities for random gene pairs when compared to the genes on the KEGG pathway named pathways in cancer [117, 118]. We found that there was a significant difference between the random and KEGG distributions both when we used the SGA dataset and when we used the driver dataset for calculation (Figure 3-3). The Wilcoxon rank sum p-value were 0 and 5.8e-124 for the SGA and driver dataset respectively. Having the cosine similarity distribution for functionally related gene pairs be greater than the random gene pairs supports the idea that the semantic representation captures functional information. We believe that the driver genes would result in a less significant difference because the driver genes have all been predicted to be relevant to the development of cancer. As such, randomly selecting driver gene pairs would be more likely to have functional relevance than when randomly selecting from all SGAs. A subset of the words, with their tf-idf scores, from an example driver gene pair with a high cosine similarity score is given in Table 3-1. Words that occur in both vectors are highlighted in red. Both PIK3CA and PTEN are involved in the same pathway, with PIK3CA being a known oncogene and PTEN being a known tumor suppressor. This highlights the ability of semantic vector representations to identify functionally related genes.

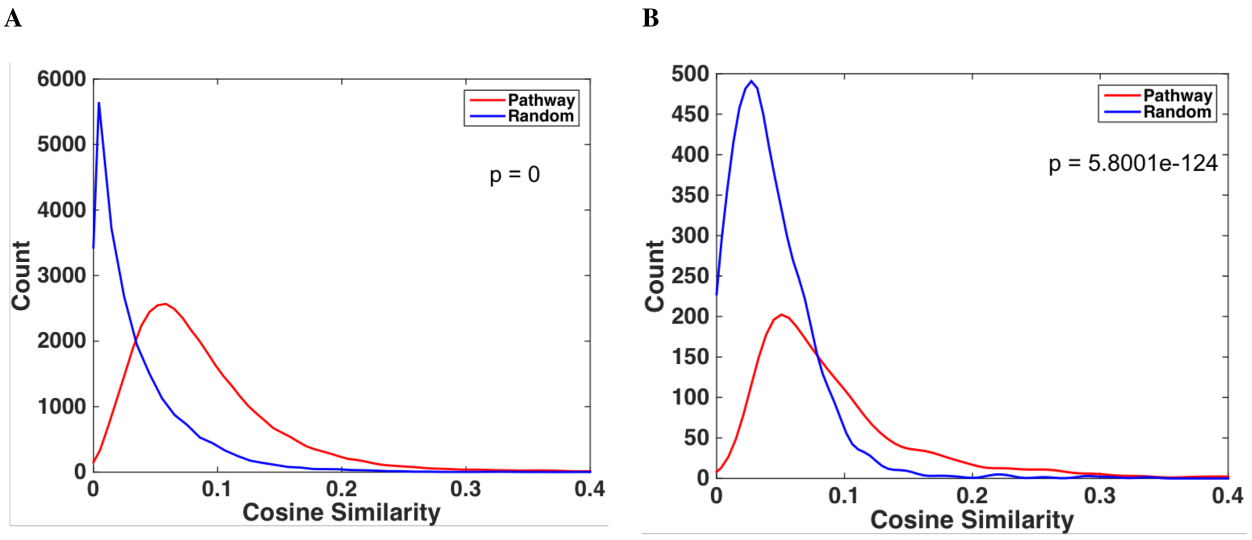


Figure 3-3. Cosine similarity distributions of semantic representations of genes. The cosine similarity was calculated for pairs of genes on a KEGG pathway, and the same number of pairs of random genes. The resulting distribution for the KEGG genes can be seen in red, and random genes can be seen in blue. **A)** All SGAs in the pathway were compared with genes randomly selected from all SGAs. **B)** All drivers in the pathway were compared with genes randomly selected from all drivers.

Table 3-1. Subset of words from two word vectors with one of the highest cosine similarity scores

PIK3CA		PTEN	
Word	**Tf-Idf**	**Word**	**Tf-Idf**
pik3ca	416	pten	262
pi3k	416	akt	175
mutat	220	cancer	156
akt	155	cell	153
cancer	118	tumor	130
pathwai	100	express	129
pten	82	carcinoma	109
activ	76	mutat	107
kra	73	pi3k	98
cell	65	pathwai	93
signal	63	activ	81
breast	59	endometri	78
carcinoma	53	signal	75
3-kinas	53	prostat	75
tumor	52	breast	67
oncogen	50	phosphatas	65
braf	49	pik3ca	57
akt1	45	associ	57
patient	41	promot	54
endometri	38	patient	51

3.3.3 Topic Modeling Results

The structure of the tree and the topics were inspected, and some example topics from the topic associations calculated using the driver dataset are shown in Figure 3-4. Figure 3-4A shows the full structure of the topic model tree, while Figure 3-4B is a look at one specific branch. The topics shown (Figure 3-4C-E) span one branch from the root to the leaf. This reveals that the words in topics closer to the root are more general, containing common words such as cancer, tumor, and cell. As we move further down the branch, the words get more specific. One other aspect to note is that GATA proteins are known to regulate mucin genes [120]. The fact that these words appear in topics along the same branch indicates that the hierarchical structure is also capturing the functional relationship between topics.

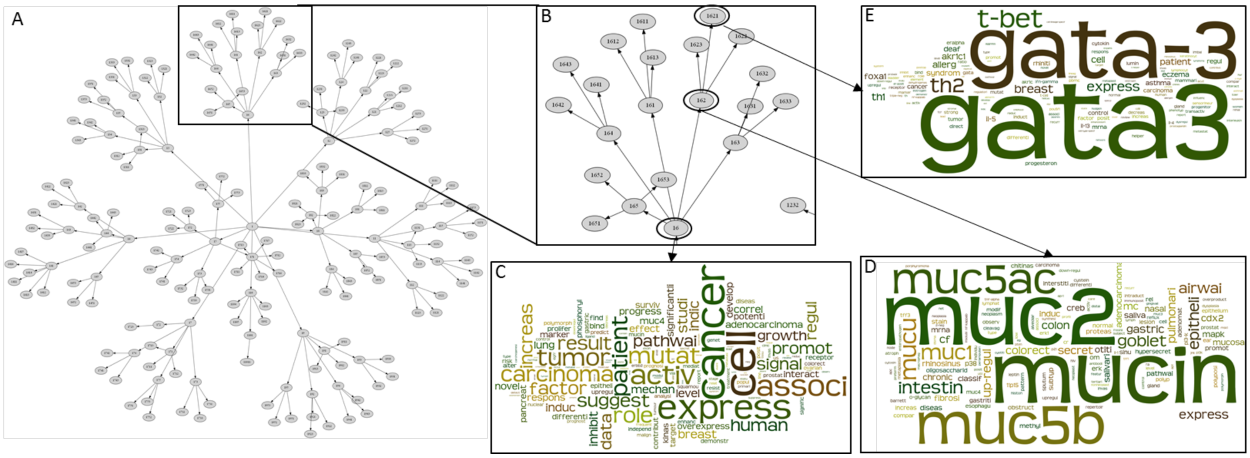


Figure 3-4. Topic model structure and topic associations. The hierarchical structure and example topics of a topic model generated using the semantic driver dataset as input is shown. **A)** The visualization shows the full hierarchical structure of the topic tree, where each node represents a topic. **B)** One branch along the topic tree. **C-E)** Word clouds showing three example topics that are progressively deeper in the topic tree, and further away from the root. These words associated with the topics also are progressively more specific.

3.3.3.1 Topic Analysis

In order to confirm that the topics have grouped together functionally similar genes, their PPI ratios were calculated and compared to the PPI ratios of random gene lists. Functionally related genes would have a greater number of PPI than unrelated genes, so the PPI ratio of topic genes should be higher if they are functionally related. We found that 98% of the topics for the SGA dataset had a PPI ratio larger than at least 95% of those generated at random (S Table 4), and the driver dataset had 91% of the topics meet that threshold (S Table 5). This indicates that the topic model is able to identify and capture functional information, and their topic association patterns can then be used to separate tumors.

3.3.4 Clustering Tumors

Clustering was used to compare the performance of the different datasets in capturing relevant information. The tumors were clustered using genomic alterations directly (S Figure 2), drivers directly (S Figure 3), and using the topic associations generated based on the SGA dataset and the driver dataset (Figure 3-5). We found that clustering based on the genomic alterations was unable to generate clean clusters at any of the cluster sizes that were searched. On the other hand, clustering directly based on drivers generated clean clusters at all of the cluster sizes that were searched. This indicates that a lot of noise was removed through driver identification. However, clustering directly using drivers is not capable of grouping together drivers with the same functional impact. As for the topic associations, the fact that both topic associations were able to generate cleaner clusters indicates that it decreases the amount of noise in the data and captures information about the relationship between genes. However, when comparing the driver topic association clusters with the driver clusters, it suggests that the semantic representation may be adding in some additional noise. The clustering results for the SGA dataset at k = 14 is shown in Figure 3-5A. The clustering results for the driver dataset at k = 16 is shown in Figure 3-5B. These two clustering results were used to generate 16 and 23 clusters, respectively.

3.3.4.1 Cluster Evaluation

With topics grouping together functionally similar genes, when we clustered the tumors based on their topic associations the resulting clusters should contain samples that have common perturbed pathways. Tumors with common perturbed pathways would contain genes that have a greater number of PPI. As such, a cluster of genes with a PPI ratio greater than the PPI ratios for random genes would indicate that the tumors were more likely to share common perturbed pathways. We found that all of the clusters for the SGA dataset had PPI ratios that were larger than at least 95% of those generated at random (S Table 6). For the driver dataset, 95% of the clusters had a driver PPI ratio that was larger than at least 95% of those generated at random (S Table 7). However, when we used all associated SGAs instead of just drivers, we found only 13% to be above the 0.95 threshold (S Table 8). This is likely because the tumor representations were based on drivers, and so the topics identified drivers that have similar functions. However, the SGAs associated with these tumors contained more noise and extraneous functions. This means that the driver representation would not account for these functions. On the other hand, the SGA dataset was calculated based on capturing the functional similarity of all genes, even if the genes are not associated with the development of cancer.

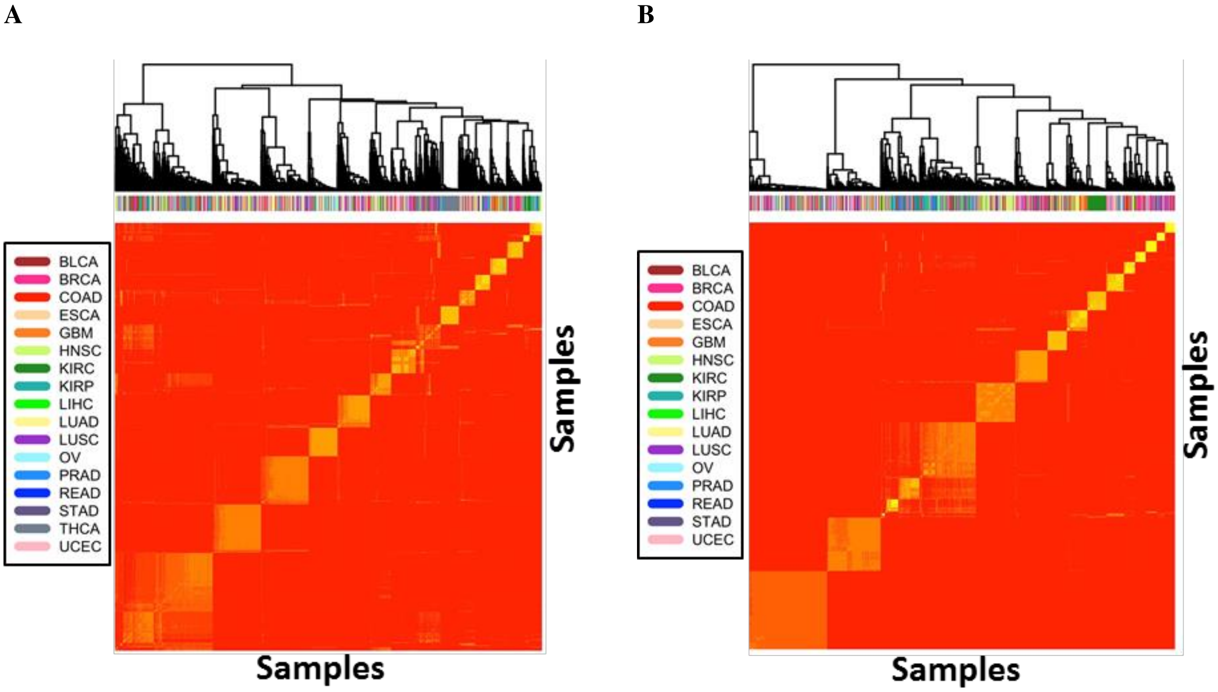


Figure 3-5. Clustering of tumors. Samples were clustered using topic associations generated on the somatic genomic alteration (SGA) dataset and driver dataset. **A)** The clustering results for the SGA dataset at k = 14, which was cut to 16 clusters. **B)** The clustering results for the driver dataset at k = 16, which was cut to 23 clusters.

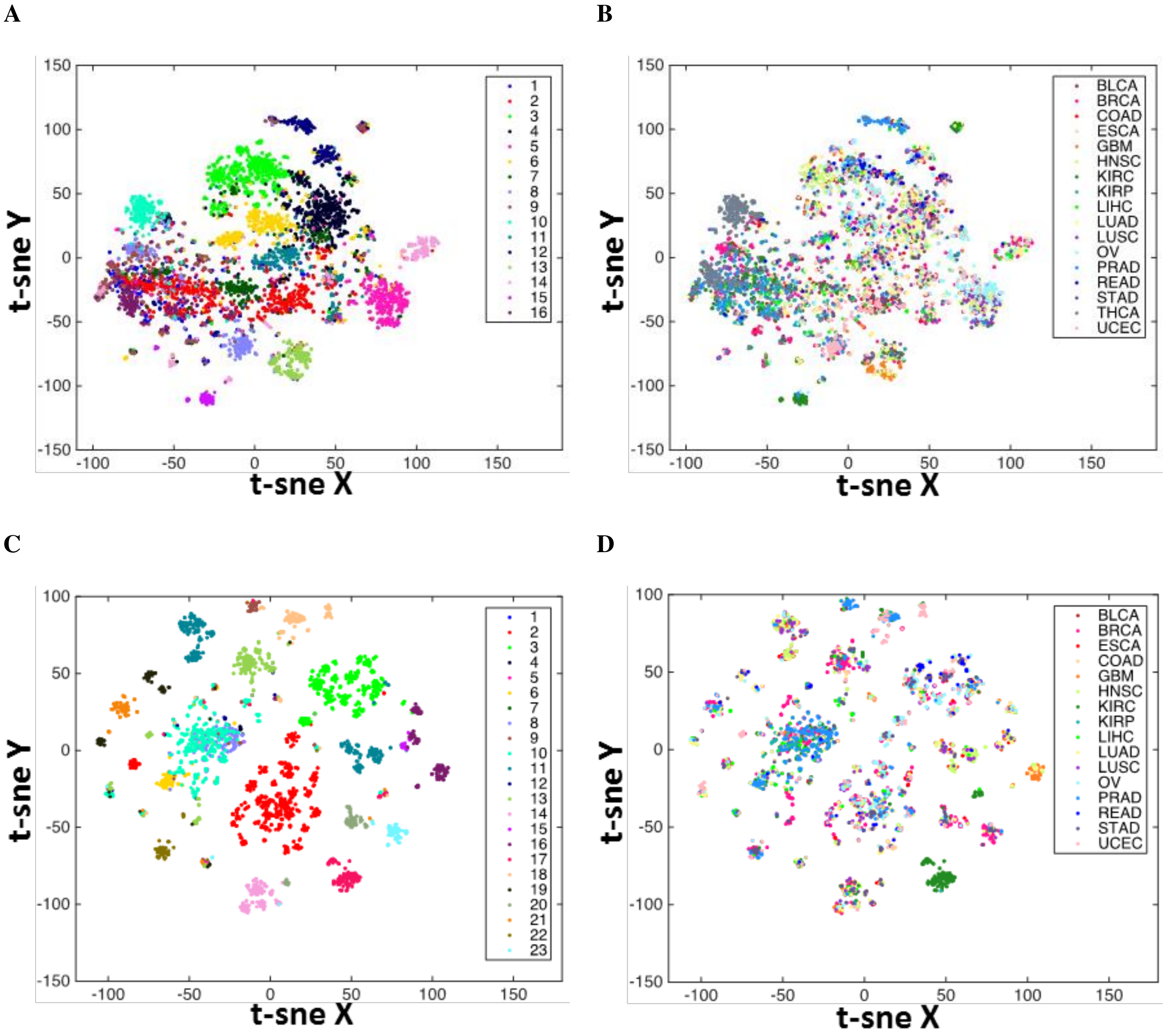


Figure 3-6. Visual representation of distance between tumors. Topic representations were used to calculate the t-sne distance tumors. A and B were generated using the topic associations for the SGA dataset. C and D were generated using the topic associations for the driver dataset. **A, C)** Tumors are labeled based on the clusters identified in section 3.2.5.4 and seen in Figure 3-5. **B, D)** Tumors are labeled based on their cancer type.

3.3.5 Visualization of Tumor Clusters

In order to see how well the topic representations could separate the tumors, they were visualized in a 2D space using t-SNE. The tumors were labeled either based on the clusters generated (section 3.3.4) or by their cancer type and the results can be seen in Figure 3-6. The clusters are not well separated when using the topic associations generated by the SGA dataset (Figure 3-6A). In comparison, there is a much cleaner separation between clusters when visualizing the tumors using the topic associations generated by the driver dataset (Figure 3-6C). This indicates that there may be too much noise in the SGA dataset for the topic model to handle. Limiting the dataset to the identified driver genes decreased the amount of noise, which would make it easier to capture the relevant information about genes. As such, we feel that further research should focus on using driver genes, rather than all SGAs, as input. Another observation we had was that both representations contained a mixture of cancer types across tumors, which can be seen in Figure 3-6B and D. This indicates that disease mechanisms can be shared across cancer types.

3.3.6 Survival Analysis

Survival analysis was performed on each of the different cancer types for both the SGA and the driver datasets, using 16 and 23 clusters respectively. We found that the majority of the tumors did not have a significant difference in survival rates. For the SGA dataset, only KIRC and LUSC were found to be significant (S Figure 4); for the driver dataset, only BRCA, LUSC, and UCEC were found to be significant (S Figure 5). Different combinations of biological features may have been identified using the clustering method, as indicated by a difference in survival rates for some cancer types. The fact that different cancer types had significantly different survival rates for the two datasets suggests that different features were found. However, the way many of the clusters seem to mix in the t-SNE representation of the SGA dataset makes the results less compelling and difficult to interpret. A lack of difference in survival data for some cancer types was expected, due to either a lack of adequate survival data or no known difference in survival performance. However, the fact that so few cancer types had a significant difference still indicates that it may be difficult to detect these different features in a pan-cancer dataset.

3.4 DISCUSSION

In this study, we applied and analyzed the use of semantic representation and topic modeling on a pan-cancer dataset, and also compared the results obtained when using the SGA dataset and the driver dataset as input. Analysis of the results indicates that these methods are capable of grouping functionally related genes and finding functional relationships between topics. This then allows the clustered samples to be grouped with other samples that have similar functions altered. The results indicate that using driver genes leads to a better separation of samples, and a different set of survival results.

Building upon our previous results in Chapter 2.0, we provided further results supporting the hypothesis that the semantic representations used detect the functional similarity of genes. We also took steps to validate and evaluate our topic and clustering results in a quantitative manner. Using t-SNE visualization allowed us to observe the clean separation of clusters using the driver dataset, in contrast with the full SGA dataset. This highlights the importance of using driver genes in research, even if functional similarity can be captured for all genes. Using driver genes decreases the dimensionality of the problem, since prior to driver identification tumors could have as many as 500 or more alterations. A focus on driver genes also means that the genes and functions grouped by topics would be related to cancer development, which is not true when all SGAs are used. Applying the methods on a pan-cancer scale demonstrated its ability to perform on a large scale, which would make it useful for drawing interpretations about cancers that span different tissue types.

Given that this study continues to use semantic data, it is once again limited by the amount and breadth of research available. This limitation means that it is difficult to accurately represent genes with newly discovered or poorly understood functions. As a result, the grouping of some genes to certain topics may not accurately reflect the biological truth. Interpretation of the results is also limited by the amount of insight that is gained from the generated topics. Without an understanding of what functions are captured by the topics associated with individual clusters or samples, it is difficult to draw conclusions about what functions are being altered. The quality of the results generated using the driver dataset are also dependent on the accuracy of the identified drivers.

3.5 CONCLUSION

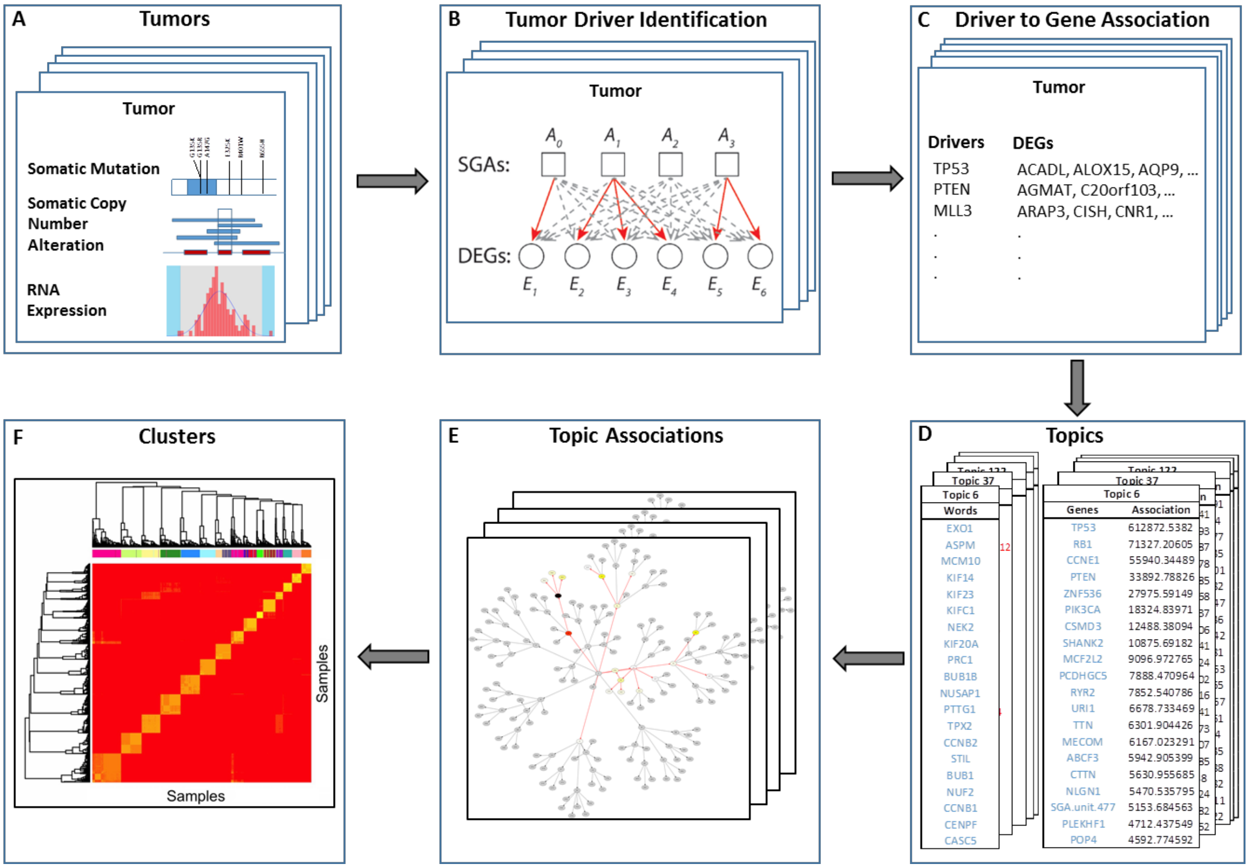
Our research applied both semantic representation and hierarchical topic modeling to pan-cancer data. While semantic representations made it possible to identify functionally similar genes, their performance is dependent on available research and literature. Therefore, it may be useful to pursue other forms of representation that could capture gene function without this dependency. Our results show that driver data lead to an improved performance in cluster separation. Being able to better identify the similarities and differences between tumors is potentially useful for understanding and treating cancer. As such, the application of driver data in identifying cancer disease mechanisms may be beneficial.

4.0 BIOLOGICAL REPRESENTATION OF DRIVER GENOMIC ALTERATIONS TO IDENTIFY PATHWAY PERTURBATION PATTERNS IN CANCER

4.1 INTRODUCTION

Our prior studies used semantic representations to capture the functional information of genes and their associated tumors. However, as mentioned previously, these representations are knowledge-driven and dependent on available literature. As such, genes would not be accurately represented without literature available covering the relevant functions. This makes it difficult to make new discoveries about the functional similarities of genes. Therefore, we wanted to develop a representation that is independent of literature and determine how well it performs on pan-cancer data.

In order to tackle this problem and continue to capture the functional similarity of distinct genes, we developed a novel biological representation of genes that takes advantage of the driver identification algorithm developed in our lab. Gene expression is the compilation of the signaling state of cells. Therefore, an alteration that has an impact on gene function would result in a signature of differentially expressed genes (DEGs). This means that the functions of driver somatic genomic alterations (SGAs) could be represented by the DEGs they have a causal relationship with. Using a driver-based representation that is calculated directly from the tumors is a data-driven approach, and allows the representation to be more biologically relevant and better fit the dataset than a knowledge-driven approach. It also avoids the bias that is inherent in a semantic representation, where the quality depends on the available literature and focus of research.



Topic 6
Words
EXO1
ASPM
MCM10
KIF14
KIF23
KIFC1
NEK2
KIF20A
PRC1
BUB1B
NUSAP1
PTTG1
TPX2
CCNB2
STIL
BUB1
NUF2
CCNB1
CENPF
CASC5

Topic 6	
Genes	Association
TP53	612872.5382
RB1	71327.20605
CCNE1	55940.34489
PTEN	33892.78826
ZNF536	27975.59149
PIK3CA	18324.83971
CSMD3	12488.38094
SHANK2	10875.69182
MCF2L2	9096.972765
PCDHGC5	7888.470964
RYR2	7852.540786
URI1	6678.733469
TTN	6301.904426
MECOM	6167.023291
ABCF3	5942.905399
CTTN	5630.955685
NLGN1	5470.535795
SGA.unit.477	5153.684563
PLEKHF1	4712.437549
POP4	4592.774592

Figure 4-1. Conceptual overview of research. A) Somatic mutation, copy number alteration, and gene expression data for each tumor was collected. **B-C)** The genomic alteration data was used as input for the tumor driver identification algorithm in order to identify the driver to gene association for each tumor. **D)** The driver to gene association information was used to generate tumor representations that were used as input for a hierarchical topic model, which identified topics associated with each tumor. **E-F)** The generated topic association were used to cluster the tumors.

In order to determine if biological representations can be used in conjunction with topic modeling to identify combinatorial patterns of pathway perturbations, we applied our novel biological representation to pan-cancer data in lieu of the previously used semantic representation. We continued using pan-cancer data, and obtained the relationships between SGAs and DEGs using the tumor-specific driver identification algorithm. The results were evaluated using cluster and survival analysis. The conceptual overview of our research is shown in Figure 4-1.

4.2 METHODS

4.2.1 Data Processing

The somatic mutation data was downloaded from The Cancer Genome Atlas (TCGA) [25], while the copy number variation and gene expression data were downloaded from Broad GDAC Firehose [115]. Since all three data types were needed for analysis, only 16 cancer types were used: bladder urothelial carcinoma (BLCA), breast invasive carcinoma (BRCA), colon adenocarcinoma (COAD), esophageal carcinoma (ESCA), glioblastoma multiforme (GBM), head and neck squamous cell carcinoma (HNSC), kidney renal clear cell carcinoma (KIRC), kidney renal papillary cell carcinoma (KIRP), liver hepatocellular carcinoma (LIHC), lung adenocarcinoma (LUAD), lung squamous cell (LUSC), ovarian serous cystadenocarcinoma (OV), prostate adenocarcinoma (PRAD), rectum adenocarcinoma (READ), stomach adenocarcinoma (STAD), and uterine corpus endometrial carcinoma (UCEC). The 16 downloaded cancer types resulted in a total of 4468 tumors, and 26,203 SGAs. The breakdown of the samples per cancer type is listed in S Table 3. The methods used for data processing were the same as those listed in section 3.2.1.

4.2.2 Cancer Driver Data

The driver data containing the driver to gene association for these tumors were calculated in the same way as described in section 3.2.2. This resulted in a final list of 721 unique drivers and 15,902 DEGs.

4.2.3 Representation of Genes and Tumors

4.2.3.1 Biological Representation of Genes

The calculated driver data contained information regarding the relationship between somatic genomic alterations (SGAs) and differentially expressed genes (DEGs). This relationship was used to generate the biological representation of genes, where each driver SGA, SGA unit, or SGA group was represented by its associated DEGs. The SGA units and groups were not expanded back to their component genes in order to avoid over-representing these drivers since they would all be represented by the same DEGs. A binary vector of DEGs was created where a value of 0 meant that the DEG was not associated with the driver, and a value of 1 meant that the DEG was associated with the driver. Since the driver and DEG associations differed depending on the tumor, the representations for the same driver could be different for different tumors.

4.2.3.2 Biological Representation of Tumors

Vectors containing the relevant DEGs and their frequencies were needed for the topic modelling process. Since the cancer driver identification process already excluded the DEGs that were predicted to be irrelevant to cancer development, further limitation of the "vocabulary" size was not performed. This resulted in a final vocabulary size of 15,902 genes. For each tumor, the DEG vector representation was created by using count vectors and summing up the values of each of its drivers.

4.2.4 Topic Modeling

The topic modeling algorithm and the topic model selection process were the same as those described in section 2.2.3. The nHDP algorithm was run on the corpus of DEG vectors representing the tumors in order to generate the resulting word-to-topic and document-to-topic distribution matrices.

4.2.5 Analysis

4.2.5.1 Evaluating Biological Representation of Genes

In order to ensure that the biological representation captures the functional similarity of genes, we wanted to determine that gene vector representations for drivers with similar functions were closer in similarity. Since the driver representation was different for each tumor, a comprehensive representation for each driver was generated. This representation was a count vector, where the value for each DEG was calculated by counting the number of times the DEG was associated with the driver across all tumors. However, the drivers that we have representations of are all predicted to be cancer drivers, which means they are already likely to be functionally related. Therefore, for each driver we generated a corresponding random “driver” by keeping the counts and replacing the associated DEGs with the DEGs randomly selected from the entire “vocabulary”. This random selection represents the situation where the identified DEGs do not have any relationship with the driver, and therefore would not be functionally related. The subset of an example driver gene vector representation as well as its corresponding randomly generated vector can be seen in Table 4-1.

Table 4-1. Example biological representation vector and corresponding randomly generated vector

SSPO	
DEG	Count
PCOLCE2	131
GYG2	112
SNHG12	100
DCHS1	92
TPSB2	85
TTLL7	81
HIP1	78
PLD4	78
COLEC12	78
ALDH1L2	77
ELMO1	75
IPCEF1	71
MAP7D2	69
KRT17	68
XKR9	67
PYGM	66
EPM2A	65
AKAP12	64
PLIN4	64
SLCO2B1	62

Random	
DEG	Count
SNORD116.20	131
ANKRD37	112
NEK10	100
AGTPBP1	92
LOC84856	85
CD34	81
COL29A1	78
REEP2	78
NTS	78
HSPA2	77
GPR4	75
C7orf10	71
FAM95B1	69
GTF2IRD2B	68
GPCPD1	67
COL2A1	66
AMZ1	65
RASSF3	64
NCF1C	64
SYT9	62

Cosine similarity was used to measure the similarity for both the comprehensive and the random DEG count vectors. The cosine similarity of each pair of genes within the two collections of DEG count vectors were calculated. The cosine similarity distribution for both collections were compared using the Wilcoxon rank sum test to determine if there was a significant difference between functionally related and randomly generated drivers.

4.2.5.2 Calculating Topic to Driver Associations

Since we used the input DEG count vectors as documents, the "words" that the topic model selected were actually genes. However, further calculations were necessary in order to determine the drivers associated with each topic. The topic to driver associations for each topic (*t*) were calculated using the generated word-to-topic and document-to-topic matrices. The word-to-topic matrix was normalized by topic to get the probability of each word being associated with a topic. For each document (*d*) and each word (*w*) in the document, we identified the driver that was most strongly associated by extracting row *w* from the normalized word-to-topic matrix, and performing element-wise multiplication of it with row *d* of the document-to-topic matrix. The topic *t* with the largest value in this resulting vector was identified, and the driver associated with word *w* was assigned to this topic. If the driver was already associated with topic *t*, then the value assigned to the driver would be increased by 1. We chose to use an incremental value of 1 because it helps to decrease the chance of topics becoming dominated by common drivers, and drowning out the signal of rarer drivers. Only the top 20 drivers associated with a topic were used for further analysis.

The pseudocode used to calculate topic to driver associations can be found below:

```
Input:
wordTopicMatrix - matrix containing probability of each word occurring in each topic
documentTopicMatrix - matrix containing number of words from each document associated with each topic
DEG vector representation of each SGA for each document

Initialize:
topicDriverMatrix - an empty topic by driver matrix

foreach document (d)
    foreach word (w)
        Perform element-wise multiplication of row w in wordTopicMatrix with row d in documentTopicMatrix
        Find the topic (t) with the largest value in resulting vector
        Identify driver (s) associated with w for d
        topicDriverMatrix[t][s] = topicDriverMatrix[t][s] + 1
```

4.2.5.3 Topic Analysis

The method for performing topic analysis was the same as that described in section 3.2.5.3. Prior to running the analysis, any SGA unit or SGA group was expanded, and the component genes were used. This is because SGA units and SGA groups cannot be directly used to identify protein-protein interactions. As a result, some of the gene lists had a length greater than 20.

4.2.5.4 Clustering Tumors

Consensus clustering was performed on the tumors to determine if the use of topic modeling allowed us to learn additional relationship information about the data. As such, clustering was performed using the generated topic associations. We used partitioning around medoids, k-means, and merging the two for consensus as the clustering methods. The algorithm was run for cluster sizes 10-30. Consensus clustering was performed using the clusterCons package version 1.0 in R [109].

4.2.5.5 Cluster Evaluation

We used the drivers associated with each cluster to evaluate if the samples share common combinations of pathways. The driver gene list for each cluster was obtained by taking the union of all of the drivers associated with each tumor in a cluster. A total of 10,000 random draws was used to generate the random distribution for each driver list length. The reasoning for this analysis was provided in section 3.2.5.5, and the method for calculating the PPI ratio and comparing against the random distribution was described in section 3.2.5.3.

4.2.5.6 Visualization of Tumor Clusters

The topic representations of the tumors were visualized in a two-dimensional space using the same methods as those described in section 2.2.4.3.

4.2.5.7 Calculating Cluster to Topic Associations

We visualized the proportions of samples in a cluster associated with each topic using the same method as section 2.2.4.4.

4.2.5.8 Survival Analysis

Kaplan-Meier survival analysis was performed on two different cancer subtypes as a measure of biological impact in subtyping the tumors based on clustering: GBM and OV.

Survival data for the tumors was obtained from the TCGA project clinical data using the National Cancer Institute's Genomic Data Commons: http://gdc.nci.nih.gov/. Due to the fact that clustering the pan-cancer data resulted in clusters mainly being separated by cancer types, consensus clustering was performed on the tumors for these individual cancer types separately for cluster sizes 4-20. The tumors were separated into subsets based on these newly generated clustering results. We used the survival package version 2.38.3 in R to conduct the analysis [111, 112].

4.3 RESULTS

4.3.1 Biological Representation Evaluation

The DEG vector representations of drivers and the randomly generated representations were used to calculate and compare the resulting cosine similarity distributions. We found that there was a significant difference between the random and driver distributions (Figure 4-2). The Wilcoxon rank sum p-value was 3.05e-269. This result supports the idea that the DEGs associated with a driver SGA have a relationship, and so a biological representation based on these DEGs would be able to capture functional information about the drivers. We also identified driver pairs that have high cosine similarity, and found that one of the top pairs was KEAP1 and NFE2L2 (Table 4-2). These two genes are on the same pathway and play a role in response to oxidative stress [121]. This further highlights the ability of biological gene representations to identify functionally related genes.

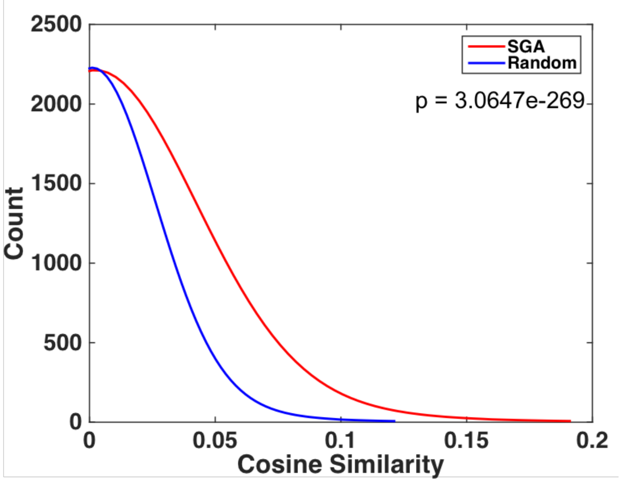


Figure 4-2. Cosine similarity distribution of biological representations of genes. Cosine similarity was calculated for each pair of driver SGAs and each pair of randomly generated DEG vectors. The resulting distribution for the SGAs can be seen in red, and random genes can be seen in blue.

4.3.1.1 Topic Analysis

To confirm that the topics generated based on biological representations could group together functionally similar drivers, the PPI ratios of topic driver SGAs were calculated and compared to the PPI ratios of randomly selected driver SGAs. We found that 80% of the topics had a PPI ratio larger than at least 95% of those generated at random (S Table 9). This indicates that the topic model is able to capture functional information when working with biological representations. However, the fact that this is a relatively lower percentage may mean that it is more difficult to identify the functional signal using this current representation.

Table 4-2. Subset of words from two word vectors with one of the highest cosine similarity scores

KEAP1		NFE2L2	
Word	Count	Word	Count
SLC7A11	121	SLC7A11	129
AKR1C1	118	PANX2	115
PGR	114	AKR1C1	104
NQO1	109	LRP8	100
CYP4F11	108	SLC12A8	96
VGLL1	107	TRIM16L	93
FREM1	107	WNT5A	92
COL13A1	106	CABYR	91
ADRB2	106	AKR1B15	87
CHRNA5	103	AKR1C3	86
CKMT1B	102	TXNRD1	85
C6orf97	102	GCLM	78
WISP2	99	VSIG10L	77
CABYR	99	GCLC	74
RBMS3	98	C1orf31	74
KIAA1529	95	CBR3	73
TRIM16L	94	ABCC1	73
CBFA2T3	94	SRXN1	72
SIGLEC1	89	OSGIN1	72
AKR1C3	88	CBR1	71

4.3.2 Clustering Tumors

Consensus clustering was performed to see if the cancer samples could be separated along common disease mechanisms identified using biological representations. We found that clustering based on the topic associations (Figure 4-3) resulted in cleaner clusters than using SGAs directly (S Figure 2), but less so than those using drivers directly (S Figure 3). One interesting feature about the clusters generated using biological topic associations is that the resulting clusters are generally dominated by one cancer type. The fact that this is seen in clusters generated using topic associations, but not those generated directly from driver data indicates that the DEGs associated with each tumor still contains some tissue type specific information. This information would get captured through biological representation, but would not be picked up when looking at the driver SGAs directly.

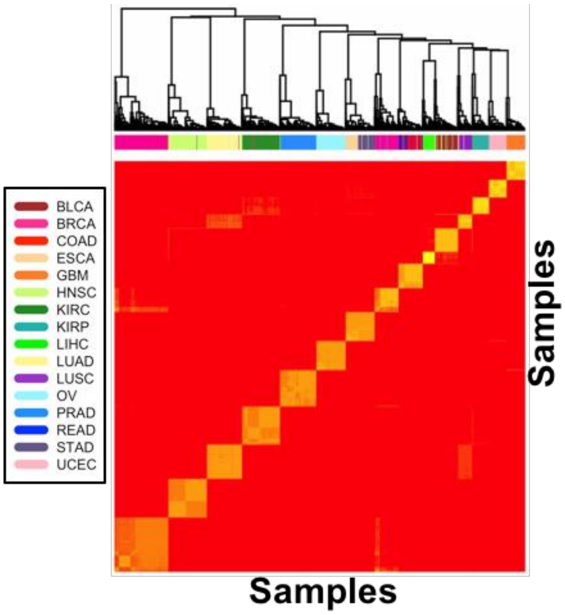


Figure 4-3. Clustering of tumors. Samples were clustered using topic associations generated using biological representations of drivers. The clustering result at k = 15 is shown.

4.3.2.1 Cluster Evaluation

Since we found that topics group together functionally similar genes, clustering samples based on these topic associations should result in clusters with common perturbed pathways. We found that all of the clusters generated using topic associations had PPI ratios that were greater than at least 99% of those generated at random (S Table 10). This indicates that even if the topics identified using biological representations do not perform as well at capturing functional information, it is still possible to find the overall combinatorial patterns of pathway perturbations when clustering tumors based on these topics.

4.3.3 Visualization of Tumor Clusters

Similar to what we saw with the clustering results, the t-SNE projection also revealed a clean separation between clusters (Figure 4-4). It even more clearly highlights how the different cancer types dominate each individual cluster. This indicates that though the topics used genes to capture distinct recurrent themes, these resulting themes were heavily influenced by the tissue of origin. This can be due to the fact that the biological representation uses DEGs to represent tumors, and some of the DEGs are still influenced by their tissue of origin despite attempts to filter out tissue-specific signals. The fact that these tumors were separated based on tissue and not just cancer type can be observed where cancers with similar origins, such as colon adenocarcinoma (COAD) and rectum adenocarcinoma (READ) or esophageal carcinoma (ESCA) and stomach adenocarcinoma (STAD) occur in the same cluster. This just means that the tissue-specific expression is the predominate signal identified using the biological representation, and it would take a closer look at individual tissue types to see if another functional signal can be found.

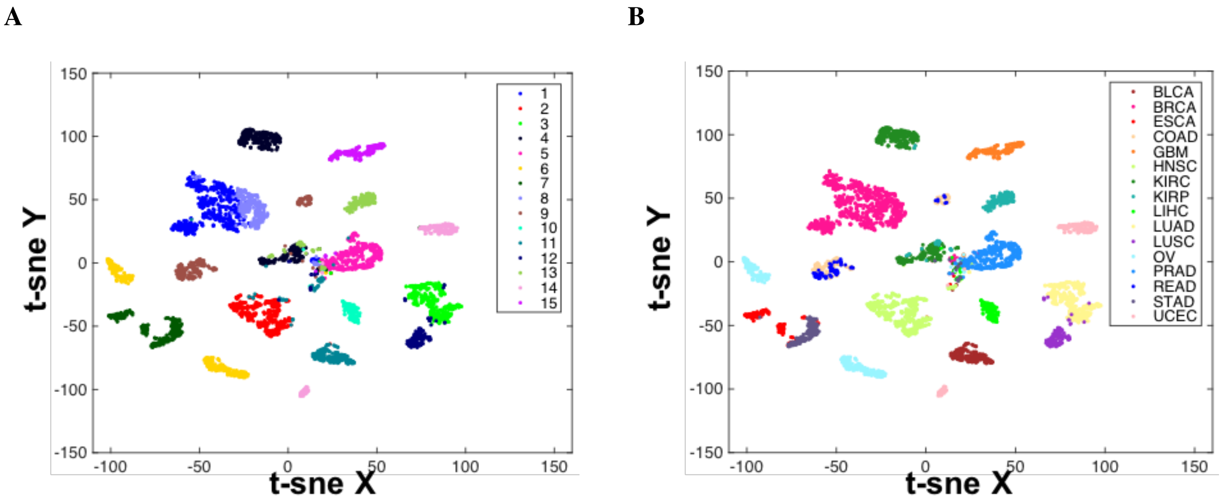


Figure 4-4. Visual representation of distance between tumors. Topic representations were used to calculate the t-sne distance between individual tumors. **A)** Tumors are labeled based on the clusters identified in section 0 and seen in Figure 4-3. **B)** Tumors are labeled based on their cancer type.

4.3.4 Topic to Cluster Associations

Since nHDP can detect the covariance structure of topics, we can use it to observe the combinatorial patterns of perturbations in different tumors. In Figure 4-5, we illustrated some of the example topic allocation trees. When examining these allocation trees, we found that the topic association patterns differed between clusters, and that some clusters are mainly composed of topics from just one subtree. Since the clusters were dominated by a specific tissue type, this implies that the topic model has identified the relationship between tissue-related genes. In these topic allocation trees, we see that the general topics that are strongly shared by the samples in a specific cluster, whereas the more specific topics deep in the hierarchy are only associated with a smaller portion of samples in the cluster.

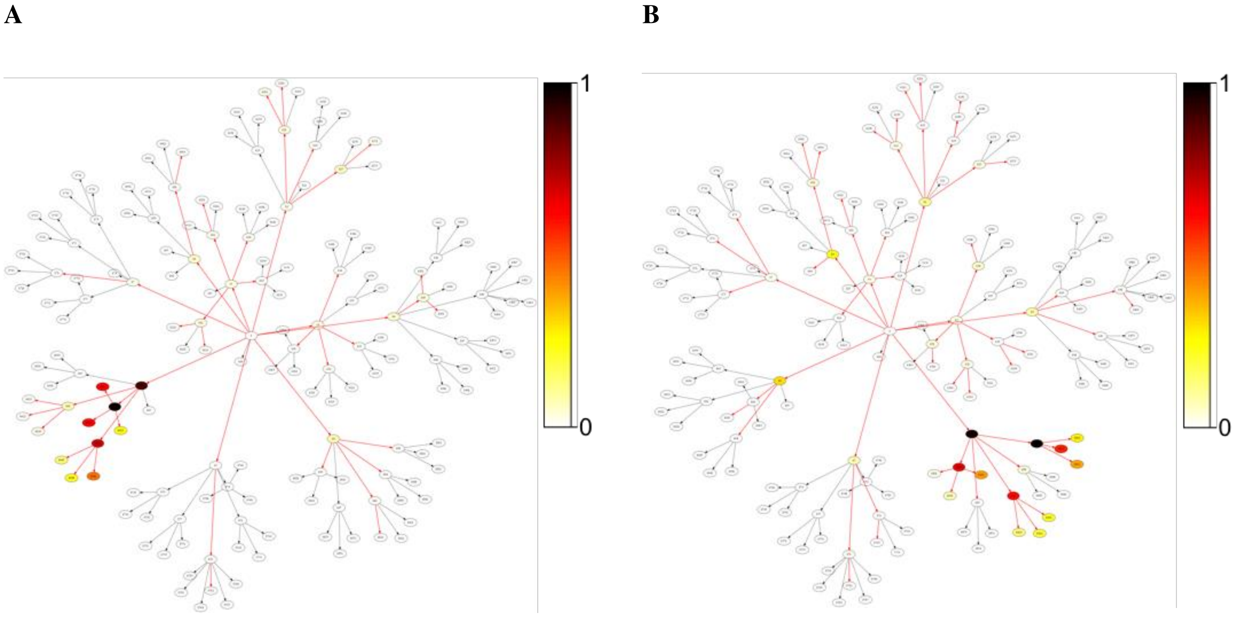


Figure 4-5. Graphical visualization of cluster-to-topic associations. The calculated degree of cluster-to-topic associations for two clusters using the clustering results seen in Figure 4-3 are shown here. These visualizations show the structure of the topic tree, where each node represents a topic. The color scale denotes the proportion of tumors in a cluster associated with each topic, where white means that none of the tumors in the clusters are associated and black means that all of the tumors are associated with the topic. **A)** The visualization for the topics associated with cluster 6. **B)** The visualization for the topics associated with cluster 9.

4.3.5 Survival Analysis

Even though clustering of pan-cancer data resulted in clusters dominated by individual cancer types, we still wanted to see if biological representations could be used to separate tumors by disease mechanisms that result in clinical differences. Since different cancer types have different survival rates and cannot be compared directly, we chose to look further into GBM and OV and clustered these samples separately before performing survival analysis on the results. We found that the two cancer types had a p-value of 0.0356 and 0.0826 respectively. These results indicate that the topic associations may have captured some clinically relevant features. However, since clustering was performed on the topic associations generated using pan-cancer data, the topics may not be able to capture the information relevant to a specific cancer type as cleanly. This means that some clinically relevant features may end up not being captured by the generated topic model.

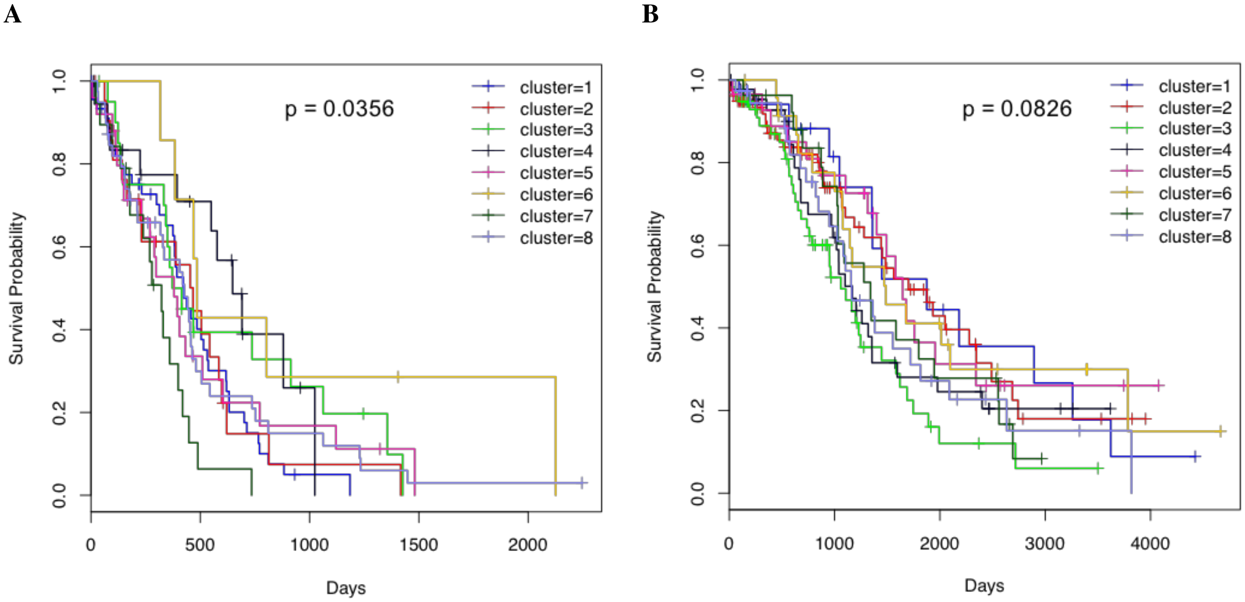


Figure 4-6. Survival analysis of tumors. The survival analysis curves calculated using the clusters generated when clustering GBM and OV tumors separately. The curves correspond to **A)** GBM and **B)** OV.

4.4 DISCUSSION

We investigated the ability to use biological representations and topic modeling to identify combinatorial patterns in perturbed signaling pathways of different tumors. Our results show that it is possible to detect the functional similarity of different drivers using a biological representation. This allowed nHDP to identify patterns in pathway perturbations across different cancer types. In contrast with the semantic representations, the tumors were generally clustered based on tissue of origin. As such, two of the cancer types were separately clustered and GBM was found to have significantly different survival rates.

To our knowledge, this is the first time a biological representation has been used to represent genes or tumors. We also show its ability to differentiate between functionally similar and random genes. Despite the fact that a smaller portion of the generated topics were considered functionally coherent, the topic associations could still be used to find combinatorial patterns in the pathway perturbations of tumors. This is where we found a very contrasting difference from when we used semantic representations, clusters were generally dominated by a specific cancer type or tissue of origin. These results highlight a problem that would need to be tackled if biological representations will be used in pan-cancer analysis: the differential expression of genes is affected by both tissue of origin and cell type. Until this issue is handled, biological representations can only be effectively used on a smaller-scale analysis of cancer types. However, the fact that there was a significant different in survival for one of the two cancer types we evaluated indicates that the biological representation is still capable of detecting differences between tumors of the same cancer type.

The biological representation that we developed was based on identified drivers and their causal relationships with DEGs. As such, the accuracy of the representation is dependent on the accuracy of the identification results. An inaccurate identification could result in either the wrong driver being represented by the DEGs, or a driver being represented by the wrong set of DEGs. The current method of representation also has done little to trim or limit the DEGs that are used.

As such, the current iteration may result in a situation where the tumors are represented by almost all of their DEGs. This may help explain why there is such a clear separation between the different cancer types. Also when we worked on analyzing individual cancer types, we used the topic associations generated using pan-cancer data. Since the topic model was capturing the features from a larger dataset, the topics generated may be too general and have a hard time capturing the finer differences between tumors of a specific cancer type.

4.5 CONCLUSION

In this study we developed a biological representation of genes and tumors that we paired with hierarchical topic modeling and applied to pan-cancer data. The biological representations were data-driven and able to identify functionally similar genes, with their accuracy dependent on the predicted causal relationships between SGAs and DEGs. This representation seems to be dominated by tissue-specific signals, which resulted in clusters that were dominated by a single tissue of origin. However, analyzing two of the cancer types individually lead us to find that one of them still had significantly different survival rates. Therefore, a biological representation could still be used to identify clinically relevant features.

5.0 OVERALL DISCUSSION

This dissertation focused on studying the utility of using alternative representations of genes and tumors in conjunction with hierarchical topic models for identifying combinatorial patterns of pathway perturbations in cancer. The effect of using driver data in the process was also studied. The results show that both semantic and biological representations are valid methods of capturing the functional similarity of genes, and the methodology we developed is worthy of further exploration with multiple examples of tumors in a cancer type being separated into groups with different survival outcomes.

We initially started off with an evaluation of the feasibility of using semantic representations of tumors and topic modeling for cancer analysis. While vector representation of genes had previously been used, to our knowledge, this is the first time that they have been combined and used to represent tumors. This was also the first time hierarchical topic modeling was used for cancer research, and further the first use of the nested hierarchical Dirichlet process (nHDP) in the biomedical field. We found that these methods made it possible to separate tumors into clear clusters based on their topic association patterns. If these topics represent pathways perturbed by somatic genomic alterations (SGAs), then nHDP makes it possible to detect patterns of pathway perturbations. The fact that the majority of the cancer types have significantly different clinical outcomes indicates that clinically relevant features are being captured by this methodology.

In order to further explore the abilities of this methodology, we applied it to pan-cancer data. At this time our lab had developed an algorithm for identifying cancer drivers, which is a major aspect of cancer research. We were interested in seeing what type of impact the inclusion of driver data would have, and so we compared the results obtained when including and excluding this data. We found that while the genes associated with the topics and clusters were generally functionally similar, this change in input data resulted in a different set of features being captured by the topics. This was highlighted by the fact that the two representations had different cancer types with significantly different survival results, and that the clusters had a much cleaner separation when using the driver data. The cleaner separation also indicates that driver data is useful for filtering out noise. Therefore, our results support the use of driver data in our methods.

With the known limitations of semantic representations, we wanted to explore a more data-driven approach. This led to the development of a biological representation using the causal relationship between driver SGAs and differentially expressed genes (DEGs). To our knowledge, this is the first time a biological representation has been used to capture the functional information of genes or tumors. Our evaluation showed that this representation is capable of differentiating between functionally related and random genes. Application of topic modeling allowed us to identify topic association patterns that aligned with tissue of origin. When the topic associations was able to separate patients into groups with significantly different survival rates, it showed that, like semantic representations, biological representations could also be used to capture clinically relevant features.

There are a number of limitations inherent to the studies, one of which is that the results are all dependent upon the quality and accuracy of the representations. As such, we tried to ensure that the representations used were capable of capturing the functional similarity of genes. It still stands to reason that any improvements to the representation would only be beneficial to the overall results. The accuracy of the results is also dependent on the sample size available as input, as driver predictions and the generated topics only improve when sample sizes increase. Therefore, this methodology would not be appropriate for small sample sizes. Another limitation is the difficulty inherent in interpreting and comparing the biological relevance of topics generated through topic modeling. While gene enrichment tools can be used to evaluate gene lists associated with a topic, these lists are only readily available when using our biological representation. All other situations would require an extra extrapolation step to identify the genes associated with a topic. Even if a gene list is readily available, it is difficult to quantitatively compare gene enrichment results. This problem also directly ties in with another limitation. Without an accurate understanding of what the topics are representing, it is difficult to interpret the biological implications of the topic associations generated.

Despite these limitations, the methodologies developed here show promise and are worth further investigation. While mainly limited to text data, topic modeling has been applied to biomedical data with an increasing frequency. Its ability to identify topics and associations directly from the data is useful when dealing with large datasets containing many features (words). If it turns out that the hierarchical tree generated using nHDP can capture the pathways involved in cancer and their relationships, then there is the potential for its use to help guide cancer treatment. This would allow treatments to be designed based on a patient’s perturbed signaling pathways obtained using their genomic data.

6.0 FUTURE WORK

The work in this dissertation is an initial exploration into the applicability of alternative representations and topic modeling to the analysis of cancer genomic data. Following this work, there are two major directions that can be focused on in future work. One focus would be on improving or exploring different aspects related to the methodology that we have used. The second focus would be to apply this framework to other datasets.

6.1 METHOD REFINEMENT

One of the areas that requires additional work is developing a metric or method that can be used to analyze and evaluate the generated topics. While we have a method of measuring the coherency of the topics, this does not allow us to compare the functions associated with the topics. As such, without such a measure it is difficult to compare between the information captured using two different models.

If such a measure was established, then this would allow for the exploration and evaluation of other aspects of the framework. It would be possible to evaluate the stability of the model, and determine how much the functions captured in a topic changes depending on the input data. For example, this would allow us to determine how the use of driver data changes the functions captured. Another aspect that could be explored would be how the use of different topic models impact the functions captured. This would give us a better idea of how the hierarchical nature of the nested hierarchical Dirichlet process captures the relationship between functions in a way that flat topic models are unable to.

Another aspect that could be refined would be the gene and tumor representations. After the first study, we made changes to the semantic representations used. However, there are still other potential knowledge sources that can be explored, and variations of the vector representations that can be used. On the other hand, we have not had a chance to explore the biological representation more fully. As such, it is possible that the representation could be improved through further adjustments. For example, setting a threshold that limits the number of differentially expressed genes used in the representation. This could potentially limit the tissue-specific genes included, and also has the additional benefit of decreasing the vocabulary size.

6.2 ADDITIONAL APPLICATIONS

The other aspect that could be explored is applying these methods to other datasets. One area that can be explored is analyzing a subset of the cancer types. Since we were interested in seeing if there were functions shared across different cancer types, we mainly focused on pan-cancer data. However, it may also be of value to explore study cancer on a smaller scale, such as a specific cancer type, tissue of origin, or cell type. This is especially true for the current biological representation, which separates the tumors by cancer type. Studying a subset of cancer types would result in a more refined and detailed look at the subset being studied, which may make it possible to pick up the subtler differences between tumors in different subtypes.

APPENDIX A

SUPPLEMENTAL MATERIALS

A.1 SUPPLEMENTAL TABLES

S Table 1. Survival analysis results of all five cancer types using semantic representation

Cancer Type	All Samples (P-Value)	Minimum 25 Samples (P-Value)
BRCA	0.00398	0.00093
HNSC	0.0126	0.00701
LUAD	0.456	N/A
LUSC	0.038	0.0355
OV	0.211	0.256

S Table 2. Somatic mutation classifications and associated functional implications

Mutation Classification	Functional Implication
Frame_Shift_Del	Yes
Frame_Shift_Ins	Yes
In_Frame_Del	Yes
In_Frame_Ins	Yes
Missense_Mutation	Depends
Nonsense_Mutation	Yes
Nonstop_Mutation	Yes
RNA	Depends
Silent	No
Splice_Site	Yes
Translation_Start_Site	Yes

S Table 3. Number of tumors used for each cancer type in pan-cancer data

Cancer Type	Abbreviation	Number of Samples in SGA Dataset	Number of Samples in Driver Dataset
Bladder urothelial carcinoma	BLCA	234	200
Breast invasive carcinoma	BRCA	972	851
Colon adenocarcinoma	COAD	182	182
Esophageal carcinoma	ESCA	209	149
Glioblastoma multiforme	GBM	234	201
Head and neck squamous cell carcinoma	HNSC	491	459
Kidney renal clear cell carcinoma	KIRC	446	426
Kidney renal papillary cell carcinoma	KIRP	169	168
Liver hepatocellular carcinoma	LIHC	194	147
Lung adenocarcinoma	LUAD	465	383
Lung squamous cell carcinoma	LUSC	178	136
Ovarian serous cystadenocarcinoma	OV	449	322
Prostate adenocarcinoma	PRAD	419	398
Rectum adenocarcinoma	READ	81	77
Stomach adenocarcinoma	STAD	227	176
Thyroid carcinoma	THCA	399	N/A
Uterine corpus endometrial carcinoma	UCEC	239	193
Pan-cancer	**PANCAN**	**5608**	**4468**

S Table 4. Proportion of topics meeting the PPI ratio threshold for semantic somatic genomic alteration dataset

Threshold	Count	Percentage
0.9	201	98.05%
0.95	201	98.05%
0.99	193	94.15%

S Table 5. Proportion of topics meeting the PPI ratio threshold for semantic driver dataset

Threshold	Count	Percentage
0.9	177	90.77%
0.95	177	90.77%
0.99	173	88.72%

S Table 6. Proportion of PPI ratios smaller than cluster gene list for semantic somatic genomic alteration dataset

Cluster	Proportion Smaller	# of Genes
1	1	25304
2	1	31693
3	1	23513
4	1	26550
5	1	23829
6	1	25107
7	1	23664
8	1	14552
9	1	16547
10	1	3783
11	1	23382
12	1	19067
13	1	14068
14	1	20806
15	1	10853
16	1	9755

S Table 7. Proportion of PPI ratios smaller than cluster driver list for semantic driver dataset

Cluster	Proportion Smaller	# of Drivers
1	0.9995	12
2	0.9995	1937
3	0.9977	1651
4	1	126
5	0.9867	371
6	0.9994	1196
7	0	13
8	0.9948	1395
9	1	974
10	0.9961	1954
11	0.9993	1330
12	0.9705	329
13	0.9999	1555
14	0.9937	1307
15	0.9979	484
16	1	1071
17	1	958
18	1	1163
19	0.9987	1027
20	1	1122
21	1	1132
22	1	838
23	0.9871	1102

S Table 8. Proportion of PPI ratios smaller than cluster gene list for semantic driver dataset

Cluster	Proportion Smaller	# of Genes
1	0.9864	540
2	0.9951	15954
3	0.492	15255
4	0.0001	5849
5	0	8428
6	0	12623
7	0.136	499
8	0	12603
9	0	11878
10	0.0072	14644
11	0.9977	15160
12	0	8296
13	0.0629	13769
14	0.1309	14507
15	0.2354	10988
16	0.1929	13575
17	0.0002	8451
18	0	12793
19	0.01	12315
20	0.7802	13634
21	0	12554
22	0	11338
23	0.0003	11843

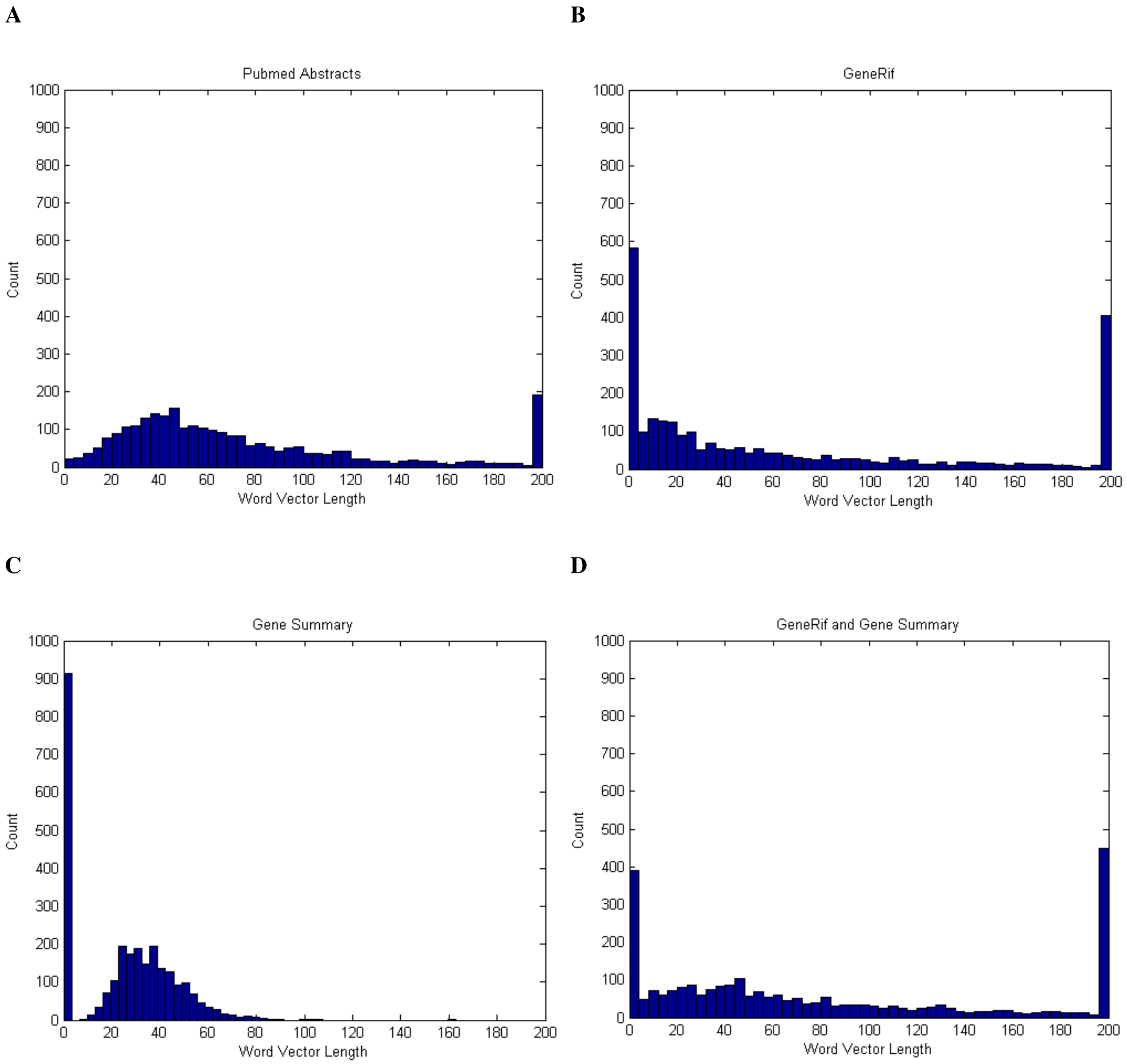
S Table 9. Proportion of topics meeting the PPI ratio threshold for biological driver dataset

Threshold	Count	Percentage
0.9	144	83.24%
0.95	139	80.35%
0.99	122	70.52%

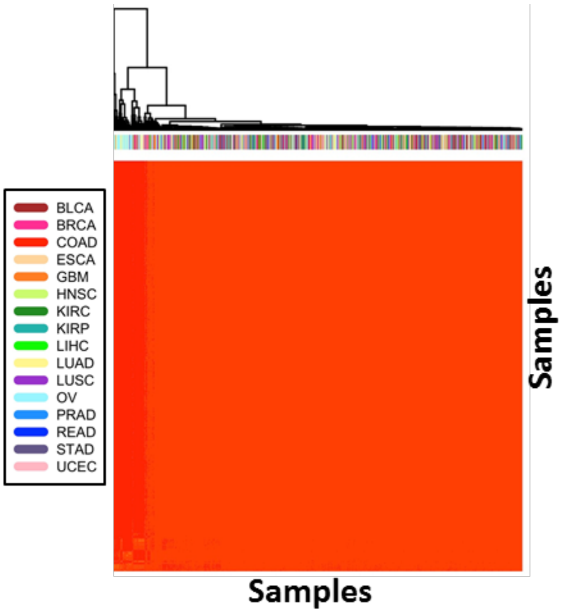
S Table 10. Proportion of PPI ratios smaller than cluster driver list for biological driver dataset

Cluster	Proportion Smaller	# of Drivers
1	1	2005
2	1	1387
3	0.9998	1416
4	1	1589
5	0.9983	1481
6	0.9954	1243
7	0.9999	1279
8	1	1425
9	1	1207
10	1	1176
11	0.9997	1450
12	0.9998	1144
13	1	1297
14	1	879
15	0.9951	927

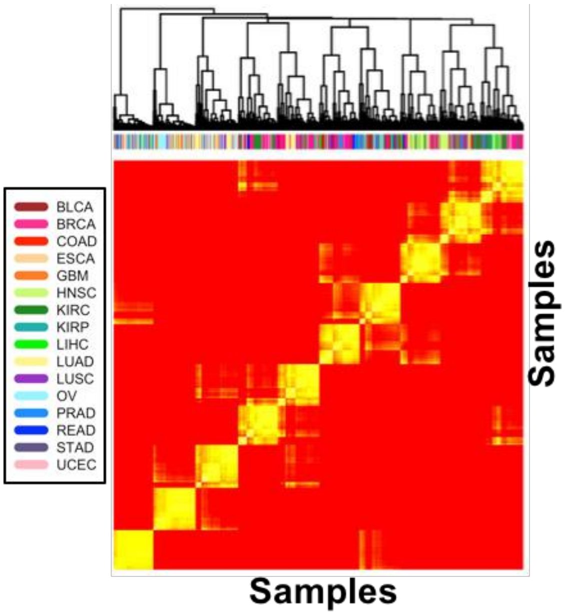
A.2 SUPPLEMENTAL FIGURES



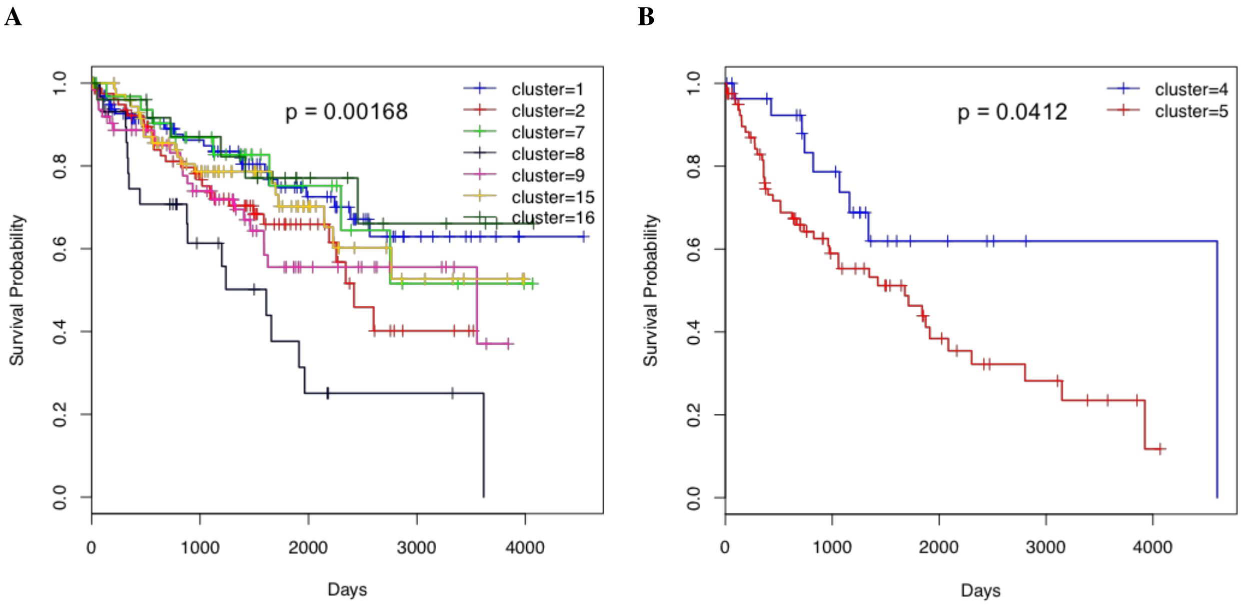
S Figure 1. Word vector length distributions. Distribution of the length of the gene word vectors generated using different semantic datasets. A) Pubmed articles, B) GeneRIFs, C) gene summaries, D) GeneRIFs and gene summaries



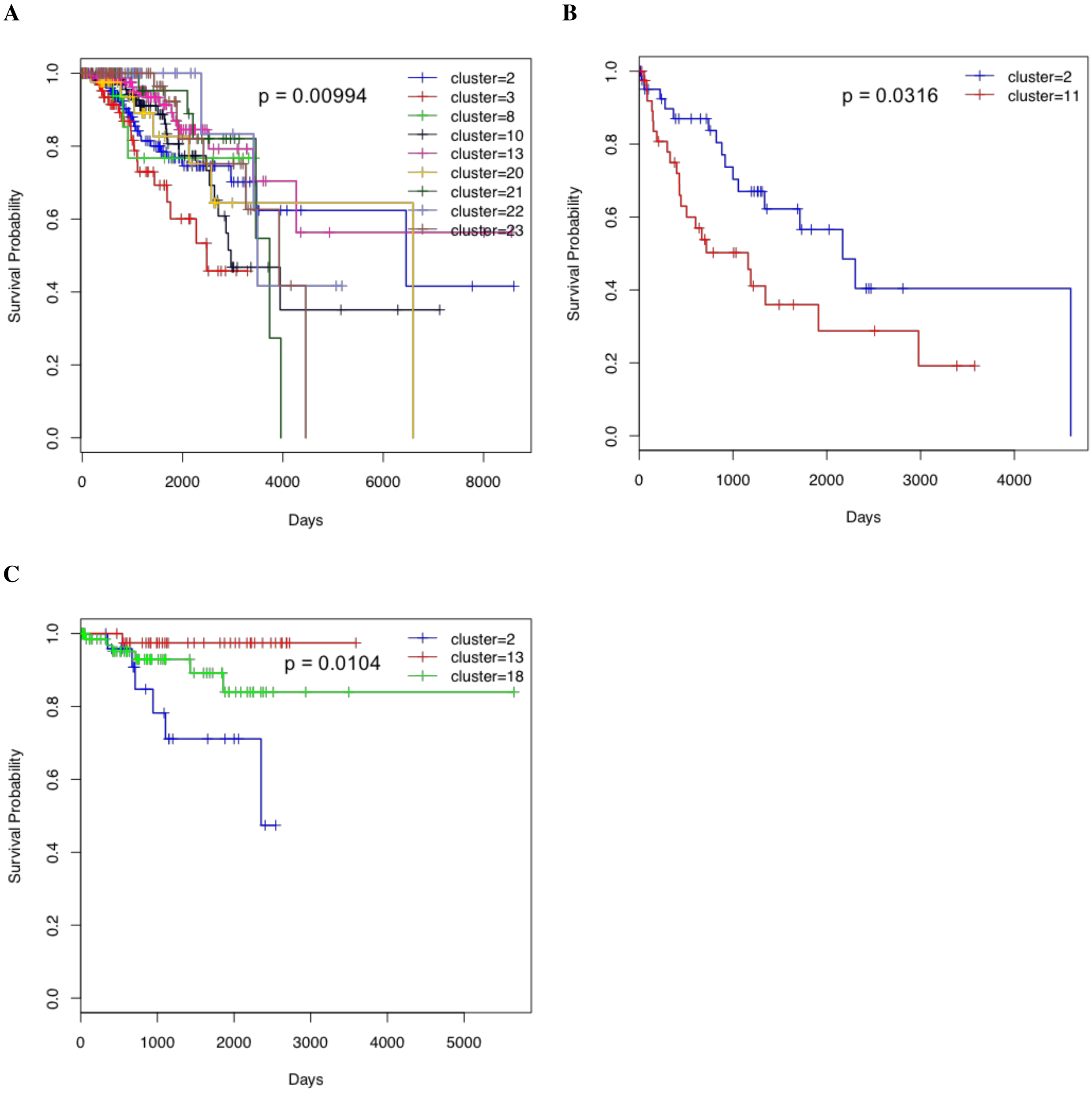
S Figure 2. Clustering of tumors using genomic alterations. Samples were clustered using genomic alteration data at k = 10.



S Figure 3. Clustering of tumors using drivers. Samples were clustered using driver data at k = 10.



S Figure 4. Survival analysis of tumors for semantic SGA dataset. The survival analysis curves calculated using only the clusters that contain at least 20 samples. Figures A and B correspond to cancer types KIRC and LUSC respectively.



S Figure 5. Survival analysis of tumors for semantic driver dataset. The survival analysis curves calculated using only the clusters that contain at least 20 samples. Figures A, B, and C correspond to cancer types BRCA, LUSC, and UCEC respectively.

BIBLIOGRAPHY

1. Torre LA, Bray F, Siegel RL, Ferlay J, Lortet-Tieulent J, Jemal A: **Global cancer statistics, 2012**. *CA: A Cancer Journal for Clinicians* 2015:n/a-n/a.
2. Cooper GM: **Elements of Human Cancer**: Jones and Bartlett Publishers; 1992.
3. Evans AS, Mueller NE: **Viruses and cancer causal associations**. *Annals of Epidemiology* 1990, **1**(1):71-92.
4. Butel JS: **Viral carcinogenesis: revelation of molecular mechanisms and etiology of human disease**. *Carcinogenesis* 2000, **21**(3):405-426.
5. zur Hausen H: **Viruses in human cancers**. *Science* 1991, **254**(5035):1167-1173.
6. Coussens LM, Werb Z: **Inflammation and cancer**. *Nature* 2002, **420**(6917):860-867.
7. Hanahan D, Weinberg RA: **Hallmarks of Cancer: The Next Generation**. *Cell* 2011, **144**(5):646-674.
8. Mantovani A, Allavena P, Sica A, Balkwill F: **Cancer-related inflammation**. *Nature* 2008, **454**(7203):436-444.
9. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Kinzler KW: **Cancer Genome Landscapes**. *Science* 2013, **339**(6127):1546-1558.
10. Hanahan D, Weinberg RA: **The Hallmarks of Cancer**. *Cell* 2000, **100**(1):57-70.
11. Narod SA, Madlensky L, Tonin P, Bradley L, Rosen B, Cole D, Risch HA: **Hereditary and familial ovarian cancer in southern ontario**. *Cancer* 1994, **74**(8):2341-2346.
12. Ford D, Easton DF, Stratton M, Narod S, Goldgar D, Devilee P, Bishop DT, Weber B, Lenoir G, Chang-Claude J *et al*: **Genetic Heterogeneity and Penetrance Analysis of the BRCA1 and BRCA2 Genes in Breast Cancer Families**. *The American Journal of Human Genetics* 1998, **62**(3):676-689.
13. Easton DF, Ford D, Bishop DT: **Breast and ovarian cancer incidence in BRCA1-mutation carriers. Breast Cancer Linkage Consortium**. *American Journal of Human Genetics* 1995, **56**(1):265-271.
14. Ford D, Easton DF, Bishop DT, Narod SA, Goldgar DE: **Risks of cancer in BRCA1-mutation carriers**. *The Lancet* 1994, **343**(8899):692-695.
15. Liaw D, Marsh DJ, Li J, Dahia PLM, Wang SI, Zheng Z, Bose S, Call KM, Tsou HC, Peacoke M *et al*: **Germline mutations of the PTEN gene in Cowden disease, an inherited breast and thyroid cancer syndrome**. *Nature Genetics* 1997, **16**(1):64-67.
16. Heald B, Church JM: **Cancer and Genetic Counseling**. In: *Inherited Cancer Syndromes: Current Clinical Management*. Edited by Ellis CN, 2 edn: Springer Science & Business Media; 2010: 23-34.
17. **The Merck manual** [http://www.merckmanuals.com/professional/index.html]
18. Taherian-Fard A, Srihari S, Ragan MA: **Breast cancer classification: linking molecular mechanisms to disease prognosis**. *Briefings in Bioinformatics* 2014.

19. Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA *et al*: **Molecular portraits of human breast tumours**. *Nature* 2000, **406**(6797):747-752.
20. Bosco EE, Wang Y, Xu H, Zilfou JT, Knudsen KE, Aronow BJ, Lowe SW, Knudsen ES: **The retinoblastoma tumor suppressor modifies the therapeutic response of breast cancer**. *Journal of Clinical Investigation* 2007, **117**(1):218-228.
21. Hayashi R, Goto Y, Ikeda R, Yokoyama KK, Yoshida K: **CDCA4 is an E2F transcription factor family-induced nuclear factor that regulates E2F-dependent transcriptional activation and cell proliferation**. *Journal of Biological Chemistry* 2006, **281**(47):35633-35648.
22. Li W, Sanki A, Karim RZ, Thompson JF, Soon Lee C, Zhuang L, McCarthy SW, Scolyer RA: **The role of cell cycle regulatory proteins in the pathogenesis of melanoma**. *Pathology* 2006, **38**(4):287-301.
23. Brugarolas J, Lei K, Hurley RL, Manning BD, Reiling JH, Hafen E, Witters LA, Ellisen LW, Kaelin WG, Jr.: **Regulation of mTOR function in response to hypoxia by REDD1 and the TSC1/TSC2 tumor suppressor complex**. *Genes & development* 2004, **18**(23):2893-2904.
24. Huang J, Zhu H, Haggarty SJ, Spring DR, Hwang H, Jin F, Snyder M, Schreiber SL: **Finding new components of the target of rapamycin (TOR) signaling network through chemical genetics and proteome chips**. *Proceedings of the National Academy of Sciences of the United States of America* 2004, **101**(47):16594-16599.
25. Cancer Genome Atlas Research Network: **Integrated genomic analyses of ovarian carcinoma**. *Nature* 2011, **474**(7353):609-615.
26. Cancer Genome Atlas Network: **Comprehensive molecular portraits of human breast tumours**. *Nature* 2012, **490**(7418):61-70.
27. Cancer Genome Atlas Research Network: **Integrated genomic characterization of papillary thyroid carcinoma**. *Cell* 2014, **159**(3):676-690.
28. Cancer Genome Atlas Research Network: **Comprehensive molecular characterization of gastric adenocarcinoma**. *Nature* 2014, **513**(7517):202-209.
29. Cancer Genome Atlas Network: **Comprehensive molecular characterization of human colon and rectal cancer**. *Nature* 2012, **487**(7407):330-337.
30. Cancer Genome Atlas Research Network: **Comprehensive genomic characterization of squamous cell lung cancers**. *Nature* 2012, **489**(7417):519-525.
31. Cancer Genome Atlas Research Network: **Comprehensive molecular characterization of clear cell renal cell carcinoma**. *Nature* 2013, **499**(7456):43-49.
32. Cancer Genome Atlas Research Network: **Comprehensive molecular profiling of lung adenocarcinoma**. *Nature* 2014, **511**(7511):543-550.
33. Cancer Genome Atlas Research Network: **Comprehensive molecular characterization of urothelial bladder carcinoma**. *Nature* 2014, **507**(7492):315-322.
34. Cancer Genome Atlas Network: **Comprehensive genomic characterization of head and neck squamous cell carcinomas**. *Nature* 2015, **517**(7536):576-582.
35. Kandoth C, McLellan MD, Vandin F, Ye K, Niu B, Lu C, Xie M, Zhang Q, McMichael JF, Wyczalkowski MA *et al*: **Mutational landscape and significance across 12 major cancer types**. *Nature* 2013, **502**(7471):333-339.

36. Gross AM, Orosco RK, Shen JP, Egloff AM, Carter H, Hofree M, Choueiri M, Coffey CS, Lippman SM, Hayes DN *et al*: **Multi-tiered genomic analysis of head and neck cancer ties TP53 mutation to 3p loss**. *Nature genetics* 2014, **46**(9):939-943.
37. Hofree M, Shen JP, Carter H, Gross A, Ideker T: **Network-based stratification of tumor mutations**. *Nature methods* 2013, **10**(11):1108-1115.
38. Hoadley KA, Yau C, Wolf DM, Cherniack AD, Tamborero D, Ng S, Leiserson MDM, Niu B, McLellan MD, Uzunangelov V *et al*: **Multiplatform Analysis of 12 Cancer Types Reveals Molecular Classification within and across Tissues of Origin**. *Cell* 2014, **158**(4):929-944.
39. Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, Bignell G, Davies H, Teague J, Butler A, Stevens C *et al*: **Patterns of somatic mutation in human cancer genomes**. *Nature* 2007, **446**(7132):153-158.
40. Ji X, Tang J, Halberg R, Busam D, Ferriera S, Pena M, Venkataramu C, Yeatman T, Zhao S: **Distinguishing between cancer driver and passenger gene alteration candidates via cross-species comparison: a pilot study**. *BMC Cancer* 2010, **10**(1):426.
41. Zhang W, Liu HT: **MAPK signal pathways in the regulation of cell proliferation in mammalian cells**. *Cell Res* 2002, **12**(1):9-18.
42. Calipel A, Lefevre G, Pouponnot C, Mouriaux F, Eychène A, Mascarelli F: **Mutation of B-Raf in Human Choroidal Melanoma Cells Mediates Cell Proliferation and Transformation through the MEK/ERK Pathway**. *Journal of Biological Chemistry* 2003, **278**(43):42409-42418.
43. Bromberg-White JL, Andersen NJ, Duesbery NS: **MEK genomics in development and disease**. *Briefings in Functional Genomics* 2012, **11**(4):300-310.
44. Stacey DW, Degudicibus SR, Smith MR: **Cellular ras activity and tumor cell proliferation**. *Experimental Cell Research* 1987, **171**(1):232-242.
45. Leiserson MDM, Blokh D, Sharan R, Raphael BJ: **Simultaneous Identification of Multiple Driver Pathways in Cancer**. *PLoS Computational Biology* 2013, **9**(5):e1003054.
46. Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N, Stratton MR: **A census of human cancer genes**. *Nature Reviews Cancer* 2004, **4**(3):177-183.
47. Lawrence MS, Stojanov P, Mermel CH, Robinson JT, Garraway LA, Golub TR, Meyerson M, Gabriel SB, Lander ES, Getz G: **Discovery and saturation analysis of cancer genes across 21 tumour types**. *Nature* 2014, **505**(7484):495-501.
48. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, Carter SL, Stewart C, Mermel CH, Roberts SA *et al*: **Mutational heterogeneity in cancer and the search for new cancer-associated genes**. *Nature* 2013, **499**(7457):214-218.
49. Dees ND, Zhang Q, Kandoth C, Wendl MC, Schierding W, Koboldt DC, Mooney TB, Callaway MB, Dooling D, Mardis ER *et al*: **MuSiC: Identifying mutational significance in cancer genomes**. *Genome Research* 2012, **22**(8):1589-1598.
50. Tamborero D, Gonzalez-Perez A, Perez-Llamas C, Deu-Pons J, Kandoth C, Reimand J, Lawrence MS, Getz G, Bader GD, Ding L *et al*: **Comprehensive identification of mutational cancer driver genes across 12 tumor types**. *Scientific Reports* 2013, **3**.
51. Zack TI, Schumacher SE, Carter SL, Cherniack AD, Saksena G, Tabak B, Lawrence MS, Zhang C-Z, Wala J, Mermel CH *et al*: **Pan-cancer patterns of somatic copy number alteration**. *Nature Genetics* 2013, **45**(10):1134-1140.

52. Pleasance ED, Cheetham RK, Stephens PJ, McBride DJ, Humphray SJ, Greenman CD, Varela I, Lin M-L, Ordonez GR, Bignell GR *et al*: **A comprehensive catalogue of somatic mutations from a human cancer genome**. *Nature* 2010, **463**(7278):191-196.

53. Stephens PJ, Tarpey PS, Davies H, Van Loo P, Greenman C, Wedge DC, Nik-Zainal S, Martin S, Varela I, Bignell GR *et al*: **The landscape of cancer genes and mutational processes in breast cancer**. *Nature* 2012, **486**(7403):400-404.

54. Nachman MW, Crowell SL: **Estimate of the Mutation Rate per Nucleotide in Humans**. *Genetics* 2000, **156**(1):297-304.

55. Nesbit CE, Tersak JM, Prochownik EV: **MYC oncogenes and human neoplastic disease**. *Oncogene* 1999, **18**(19):3004-3016.

56. Yu D, Hung MC: **Overexpression of ErbB2 in cancer and ErbB2-targeting strategies**. *Oncogene* 2000, **19**(53):6115-6121.

57. Li J, Yen C, Liaw D, Podsypanina K, Bose S, Wang SI, Puc J, Miliaresis C, Rodgers L, McCombie R *et al*: **PTEN, a Putative Protein Tyrosine Phosphatase Gene Mutated in Human Brain, Breast, and Prostate Cancer**. *Science* 1997, **275**(5308):1943-1947.

58. Sotiriou C, Pusztai L: **Gene-Expression Signatures in Breast Cancer**. *New England Journal of Medicine* 2009, **360**(8):790-800.

59. Vandin F, Upfal E, Raphael BJ: **De novo discovery of mutated driver pathways in cancer**. *Genome Research* 2012, **22**(2):375-385.

60. Ciriello G, Cerami E, Sander C, Schultz N: **Mutual exclusivity analysis identifies oncogenic network modules**. *Genome Research* 2012, **22**(2):398-406.

61. Lu S, Lu KN, Cheng S-Y, Hu B, Ma X, Nystrom N, Lu X: **Identifying Driver Genomic Alterations in Cancers by Searching Minimum-Weight, Mutually Exclusive Sets**. In: *PLoS Computational Biology*. 2015.

62. Reva B, Antipin Y, Sander C: **Predicting the functional impact of protein mutations: application to cancer genomics**. *Nucleic Acids Research* 2011, **39**(17).

63. Tamborero D, Gonzalez-Perez A, Lopez-Bigas N: **OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes**. *Bioinformatics* 2013, **29**(18):2238-2244.

64. Reimand J, Wagih O, Bader GD: **The mutational landscape of phosphorylation signaling in cancer**. *Scientific Reports* 2013, **3**.

65. Forbes SA, Tang G, Bindal N, Bamford S, Dawson E, Cole C, Kok CY, Jia M, Ewing R, Menzies A *et al*: **COSMIC (the Catalogue of Somatic Mutations in Cancer): a resource to investigate acquired mutations in human cancer**. *Nucleic Acids Research* 2010, **38**(suppl 1):D652-D657.

66. Chen R, Khatri P, Mazur PK, Polin M, Zheng Y, Vaka D, Hoang CD, Shrager J, Xu Y, Vicent S *et al*: **A Meta-analysis of Lung Cancer Gene Expression Identifies PTK7 as a Survival Gene in Lung Adenocarcinoma**. *Cancer Research* 2014, **74**(10):2892-2902.

67. Sørlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, Hastie T, Eisen MB, van de Rijn M, Jeffrey SS *et al*: **Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications**. *Proceedings of the National Academy of Sciences* 2001, **98**(19):10869-10874.

68. Lam SW, Jimenez CR, Boven E: **Breast cancer classification by proteomic technologies: Current state of knowledge**. *Cancer Treatment Reviews* 2014, **40**(1):129-138.

69. Ciriello G, Miller ML, Aksoy BA, Senbabaoglu Y, Schultz N, Sander C: **Emerging landscape of oncogenic signatures across human cancers**. *Nature Genetics* 2013, **45**(10):1127-1133.
70. Dudley JT, Chen R, Butte AJ: **Matching cancer genomes to established cell lines for personalized oncology**. *Pacific Symposium on Biocomputing* 2011:243-252.
71. Shaikh AR, Butte AJ, Schully SD, Dalton WS, Khoury MJ, Hesse BW: **Collaborative Biomedicine in the Age of Big Data: The Case of Cancer**. *Journal of Medical Internet Research* 2014, **16**(4):e101.
72. Deerwester S, Dumais ST, Furnas GW, Landauer TK, Harshman R: **Indexing by Latent Semantic Analysis**. *Journal of the American Society for Information Science* 1990, **41**(6):391-407.
73. Lee DD, Seung S: **Learning the parts of objects by non-negative matrix factorization**. *Nature* 1999, **401**:788-791.
74. Hofmann T: **Probabilistic latent semantic indexing**. *Proceedings of the 22nd Annual International SIGIR Conference* 1999.
75. Blei DM, Ng AY, Jordan MI: **Latent Dirichlet Allocation**. *Journal of Machine Learning Research* 2003, **3**:993-1022.
76. Blei DM, Griffiths TL, Jordan MI, Tenenbaum JB: **Hierarchical Topic Models and the Nested Chinese Restaurant Process**. *Advances in Neural Information Processing* 2003, **16**.
77. Paisley J, Wang C, Blei DM, Jordan MI: **Nested Hierarchical Dirichlet Processes**. *ArXiv e-prints* 2014.
78. Perotte A, Bartlett N, Elhadad N, Wood F: **Hierarchically Supervised Latent Dirichlet Allocation**. *Neural Information Processing Systems* 2012.
79. Teh YW, Jordan MI, Beal MJ, Blei DM: **Hierarchical Dirichlet Processes**. *Journal of the American Statistical Association* 2006, **101**(476):1566-1581.
80. Zheng B, McLean DC, Lu X: **Identifying biological concepts from a protein-related corpus with a probabilistic topic model**. *BMC Bioinformatics* 2006, **7**:58-58.
81. Zheng B, Lu X: **Novel metrics for evaluating the functional coherence of protein groups via protein semantic network**. *Genome Biology* 2007, **8**(7):R153-R153.
82. Bicego M, Lovato P, Olibani B, Perina A: **Expression Microarray Classification using Topic Models**. *Proceedings of the 2010 ACM Symposium on Applied Computing* 2010:1516-1520.
83. Lee M, Liu Z, Kelly R, Tong W: **Of text and gene – using text mining methods to uncover hidden knowledge in toxicogenomics**. *BMC Systems Biology* 2014, **8**(93).
84. Wang L, Wang X: **Hierarchical Dirichlet process model for gene expression clustering**. *EURASIP Journal on Bioinformatics and Systems Biology* 2013, **2013**(1):5.
85. Bisgin H, Chen M, Wang Y, Kelly R, Fang H, Xu X, Tong W: **A systems approach for analysis of high content screening assay data with topic modeling**. *BMC Bioinformatics* 2013, **14**(Suppl 14):S11.
86. Chen X, Hu X, Lim TY, Shen X, Park EK, Rosen GL: **Exploiting the Functional and Taxonomic Structure of Genomic Data by Probabilistic Topic Modeling**. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 2012, **9**(4):980-991.
87. Cho D-Y, Przytycka TM: **Dissecting cancer heterogeneity with a probabilistic genotype-phenotype model**. *Nucleic Acids Research* 2013, **41**(17):8011-8020.

88. Ashburner M, Ball C, Blake J, Botstein D, Butler H, Cherry M, Davis A, Dolinski K, Dwight S, Eppig J *et al*: **Gene Ontology: tool for the unification of biology**. *Nature Genetics* 2000, **25**(1):25-29.

89. Boyle EI, Weng S, Gollub J, Jin H, Botstein D, Cherry JM, Sherlock G: **GO::TermFinder--open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes**. *Bioinformatics* 2004, **20**(18):3710-3715.

90. Alexa A, Rahnenfuhrer J, Lengauer T: **Improved scoring of functional groups from gene expression data by decorrelating GO graph structure**. *Bioinformatics* 2006, **22**(13):1600-1607.

91. Huang DW, Sherman BT, Tan Q, Collins JR, Alvord WG, Roayaei J, Stephens R, Baseler MW, Lane HC, Lempicki RA: **The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists**. *Genome Biology* 2007, **8**(9):R183.

92. Chagoyen M, Carmona-Saez P, Shatkay H, Carazo J, Pascual-Montano A: **Discovering semantic features in the literature: a foundation for building functional associations**. *BMC Bioinformatics* 2006, **7**(1):41.

93. Glenisson P, Antal P, Mathys J, Moreau Y, Moor BD: **Evaluation of the Vector Space Representation in Text-Based Gene Clustering**. In: *Pacific Symposium on Biocomputing*. 2003: 391-402.

94. Glenisson P, Coessens B, Van Vooren S, Mathys J, Moreau Y, De Moor B: **TXTGate: profiling gene groups with text-based information**. *Genome Biology* 2004, **5**(6):R43-R43.

95. Homayouni R, Heinrich K, Wei L, Berry MW: **Gene clustering by Latent Semantic Indexing of MEDLINE abstracts**. *Bioinformatics* 2005, **21**(1):104-115.

96. Vazquez M, Carmona-Saez P, Nogales-Cadenas R, Chagoyen M, Tirado F, Carazo JM, Pascual-Montano A: **SENT: semantic features in text**. *Nucleic Acids Research* 2009, **37**(Web Server issue):W153-W159.

97. Rand WM: **Objective Criteria for the Evaluation of Clustering Methods**. *Journal of the American Statistical Association* 1971, **66**(336):846-850.

98. Rousseeuw PJ: **Silhouettes: A graphical aid to the interpretation and validation of cluster analysis**. *Journal of Computational and Applied Mathematics* 1987, **20**(0):53-65.

99. Hubert L, Arabie P: **Comparing partitions**. *Journal of Classification* 1985, **2**(1):193-218.

100. Verhaak RGW, Hoadley KA, Purdom E, Wang V, Qi Y, Wilkerson MD, Miller CR, Ding L, Golub T, Mesirov JP *et al*: **An integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR and NF1**. *Cancer cell* 2010, **17**(1):98.

101. The Cancer Genome Atlas Research Network: **Comprehensive, Integrative Genomic Analysis of Diffuse Lower-Grade Gliomas**. *New England Journal of Medicine* 2015, **372**(26):2481-2498.

102. Cancer Genome Atlas Research Network, Kandoth C, Schultz N, Cherniack AD, Akbani R, Liu Y, Shen H, Robertson AG, Pashtan I, Shen R *et al*: **Integrated genomic characterization of endometrial carcinoma**. *Nature* 2013, **497**(7447):67-73.

103. Parker JS, Mullins M, Cheang MCU, Leung S, Voduc D, Vickery T, Davies S, Fauron C, He X, Hu Z *et al*: **Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes**. *Journal of Clinical Oncology* 2009, **27**(8):1160-1167.
104. TCGA Research Network: **Integrated genomic analyses of ovarian carcinoma**. *Nature* 2011, **474**(7353):609-615.
105. Goldman M, Craft B, Swatloski T, Ellrott K, Cline M, Diekhans M, Ma S, Wilks C, Stuart J, Haussler D *et al*: **The UCSC Cancer Genomics Browser: update 2013**. *Nucleic Acids Research* 2013, **41**(D1):D949-D954.
106. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR: **A method and server for predicting damaging missense mutations**. *Nature methods* 2010, **7**(4):248-249.
107. Porter MF: **An algorithm for suffix stripping**. *Program* 1980, **14**(3):130-137.
108. Paisley J, Wang C, Blei DM, Jordan MI: **Nested Hierarchical Dirichlet Processes**. *ArXiv e-prints* 2012.
109. Simpson TI, Armstrong JD, Jarman AP: **Merged consensus clustering to assess and improve class discovery with microarray data**. *BMC Bioinformatics* 2010, **11**:590.
110. van der Maaten L, Hinton GE: **Visualizing Data using t-SNE**. *Journal of Machine Learning Research* 2008, **9**(November):2579-2605.
111. Terry M. Therneau, Patricia M. Grambsch: **Modeling Survival Data: Extending the Cox Model**. New York: Springer; 2000.
112. Therneau TM: **A Package for Survival Analysis in S**. In.; 2015.
113. Lord CJ, Ashworth A: **RAD51, BRCA2 and DNA repair: a partial resolution**. *Nature Structural & Molecular Biology* 2007, **14**(6):461-462.
114. Cousineau I, Abaji C, Belmaaza A: **BRCA1 Regulates RAD51 Function in Response to DNA Damage and Suppresses Spontaneous Sister Chromatid Replication Slippage: Implications for Sister Chromatid Cohesion, Genome Stability, and Carcinogenesis**. *Cancer Research* 2005, **65**(24):11384-11391.
115. **Broad Institute TCGA Genome Data Analysis Center (2014): Firehose stddata__2014_10_17 run. Broad Institute of MIT and Harvard. doi:10.7908/C1K64H78**. In.
116. Cai C, Lu KN, Chen V, Chen L, Lu S, Dutta-Moscato J, Clark N, Wang QJ, Lee A, Cooper GM *et al*: **A function map of cancer genome alterations derived by tumor-specific causal inference**. In.
117. Kanehisa M, Goto S: **KEGG: Kyoto Encyclopedia of Genes and Genomes**. *Nucleic Acids Research* 2000, **28**(1):27-30.
118. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M: **KEGG as a reference resource for gene and protein annotation**. *Nucleic Acids Research* 2016, **44**(Database issue):D457-D462.
119. Stark C, Breikruetz B-J, Reguly T, Boucher L, Brietkreutz A, Tyers M: **BioGRID: a general repository for interaction datasets**. *Nucleic Acids Research* 2006, **34**(Database Issue):D535-539.
120. Abba MC, Nunez MI, Colussi AG, Croce MV, Segal-Eiras A, Aldaz CM: **GATA3 protein as a MUC1 transcriptional regulator in breast cancer cells**. *Breast Cancer Research* 2006, **8**(6):1.

121. Wakabayashi N, Itoh K, Wakabayashi J, Motohashi H, Noda S, Takahashi S, Imakado S, Kotsuji T, Otsuka F, Roop DR: **Keap1-null mutation leads to postnatal lethality due to constitutive Nrf2 activation**. *Nature Genetics* 2003, **35**(3):238-245.