Identification of significantly mutated subnetworks in the breast cancer genome

by

Rasif Ajwad

A Thesis submitted to the Faculty of Graduate Studies of
The University of Manitoba
in partial fulfillment of the requirements of the degree of

MASTER OF SCIENCE

Department of Computer Science
University of Manitoba
Winnipeg

Abstract

Cancer genome projects aim at identifying the genetic variations that are related to clinical phenotypes. Recent studies showed that somatic cancer mutations target genes that are in specific cellular pathways. However, only a limited number of genes in the pathways in each patient are mutated. The existing pathway approaches consider only existing pathways and ignores the topology of the pathways. For this reason, new efforts have been focused on identifying significantly mutated subnetworks and associating them with cancer survival.

We developed a novel bioinformatics analysis pipeline to identify significantly mutated subnetworks in the breast cancer genome. We took network topology into account for measuring gene-pair mutation similarity to infer significantly mutated subnetworks. Our goals are to evaluate whether the identified subnetworks can be used as biomarkers for predicting breast cancer patient survival and provide the potential mechanisms of the pathways enriched in the subnetworks, to enable breast cancer treatment. Using the copy number variation datasets from the METABRIC (Molecular Taxonomy of Breast Cancer International Consortium) study, we identified a significantly mutated yet clinically and functionally relevant subnetwork using two graph-based clustering algorithms. The mutational pattern of the subnetwork is significantly associated with breast cancer survival. The genes in the subnetwork are significantly enriched in the retinol metabolism KEGG pathway. Our study showed that the new bioinformatics pipeline has the potential to identify new network-based biomarkers, which may be useful for stratifying cancer patients for choosing optimal treatments. The approach can be used for other types of cancer mutation data analysis.

Acknowledgement

Firstly, I would like to start by expressing my gratitude to the Almighty. I was not always sincere remembering Him, made many mistakes throughout the journey but was never let down. My prayers were granted and I was blessed with many sweet memories during my stay here in Canada.

My sincerest gratitude to my supervisors Dr. Pingzhao Hu and Dr. Michael Domaratzki for their continuous support, encouragement, and guidance for my research throughout these two years. Coming from an entirely different environment, both culturally and geographically, settling down would not be easy without their insightful guidance and advice. They were always there when I needed help and motivation, and their passion for research helped me adjust my sailing course during difficult times. It was an honor working in close proximity with both of them.

Special thanks to my colleagues and friends at the Hu-Lab for their support, in the form of friendship and encouragement: Kaiqiong Zhao, Chen Chi, Jiaying You, Svetlana Frenkel, Ye Tian and Md. Mohaiminul Islam. I would also like to thank all the faculty and staff members in the Department of Computer Science who have helped me in various ways.

Shayla Islam Tamanna, love of my life. Thank you for all the support throughout the whole journey. I would never find the strength in tough times without your constant moral support. Numerous times I gained confidence from your supportive words and could convert them into success.

Ibnat Zareen Nysa, my adorable sister, for her constant bugging. She pushed me to the edge to finish my studies as soon as possible and to be with her.

Last but not the least, I would like to dedicate this thesis to my parents, whose constant prayers, encouragement and love have helped me throughout my life. I will forever be indebted to them for everything I have achieved so far in my life.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

CNV	Copy Number Variation
CORUM	The Comprehensive Resource of Mammalian Protein Complexes
DNA	Deoxyribonucleic Acid
ER	Estrogen Receptor
FDR	False Discovery Rate
GBM	Glioblastoma
KEGG	Kyoto Encyclopedia of Genes and Genomes
METABRIC	Molecular Taxonomy of Breast Cancer International Consortium
mRNA	Messenger Ribonucleic Acid
NGS	Next Generation Sequencing
PR	Progesterone Receptor
REACTOME	A database of Reactions, Pathways and Biological Processes
RNA	Ribonucleic Acid
SNP	Single Nucleotide Polymorphism
SNV	Single Nucleotide Variants
TCGA	The Cancer Genome Atlas

Chapter 1

Background and Introduction

1.1. Genetic Variations

Human genetic variation refers to the change or variation in DNA sequence of human genome. Genetic variation maintains the difference between the individuals. Sometimes these variations are passed from one generation to other, which is a key factor in evolution. The accumulation of somatic genetic mutations, which includes single nucleotide polymorphisms (SNPs) and copy number variations (CNVs) (the variations in the copy numbers in a specific DNA region (Futreal et al., 2004; Thapar & Cooper, 2013)) drives cancer progression. Somatic CNVs are changes in the copy numbers (details in Section 1.1.2) of a DNA sequence that arise during the process of cancer development. They have been found to be prevalent in breast cancer (Beroukhim et al., 2007). Genes in the CNV regions that have changes to the chromosome structure in the form of gain or loss in copies of DNA segments, if mutated, can create abnormal proteins with different functions than a normal protein, which can lead to uncontrollable growth of cancer cells.

For example, four different nucleotides construct all the DNAs inside our genome. They are: Adenine (A), Guanine (G), Cytosine (C) and Thymine (T). Variation in the number of DNA copies can correspond to the deletion or duplication of one or more segments of the DNA. If we have a DNA segment that normally has ATGGCTA, this might have a variation such as ATGGCGGCTA (a duplication of "GGC") or ATTA (a deletion of "GGC").

1.1.1. Single Nucleotide Polymorphism

Amongst the genetic variations found in human genome, SNP is the most common. Generally speaking, every SNP is a change or difference in a single nucleotide. For example, a SNP may replace the nucleotide A with the nucleotide T in that specific position in a certain stretch of DNA, which implicates a single variation in the resultant sequence.

SNPs can be found in every 200-300 nucleotides on average, which essentially means there are roughly 10 millions of SNPs in the human genome (Liu, 2007). These are commonly found in DNA between genes. Using the SNPs as biological markers, scientists are now able to locate genes that are associated with a particular type of disease. When SNPs are found within a gene or in the vicinity of a regulatory region near a gene, they can have a somewhat direct role in disease as they can affect the gene’s function.

Although most SNPs do not seem to have much effect on human health development, some of them, however, can cause diseases (Shastry, 2009). SNPs can influence the diversity among people, and regulate the most common traits, such as eye color, hair etc. The difference in drug response of each individual for some common diseases, such as obesity, diabetes etc., may also be influenced by SNPs (Kao, Chong, & Lee, 2000). SNPs may also trigger synonymous and nonsynonymous mutations. Synonymous mutations are just a change in one base pair (mostly the last one in a set of three RNA nucleotides, known as codon) in the RNA copy of the DNA, so it doesn’t change the encoding of amino acids. However, in nonsynonymous mutations there is usually a single nucleotide variation (either insertion, deletion or substitution) during the transcription phase, when the messenger RNA (mRNA) copies the DNA. This variation can affect a whole frame of amino acid sequence, which may lead to lethal mutations. The earlier this occurs in an amino acid sequence, the more the severity of the mutation (Bell, 2002).

1.1.2. Copy Number Variations

Human genes usually have two copies, one copy inherited from each parent. Some cases have been found where the number of copies of a particular gene varies in a person. Some genes have only one copy or more than two copies (3 or more). There are also some cases where one or more genes are completely missing. The incidents are listed as ‘genetic differences’ and they are generally known as copy number variation (CNV) (Redon et al., 2006). CNVs correspond to relatively large segments of a genome that have been deleted (less than the normal number 2) or duplicated (more than the normal number 2). CNV can be classified as copy number gain (the number of copies in a DNA segment is larger than 2) and copy number loss (the number of copies in a DNA segment is smaller than 2). In this thesis, we define a CNV as a DNA segment that is 1 kilobase (kb) or larger and exists at variable copy numbers when compared to a reference genome.

There are cases where having these variations can work in our favor. For example, extra copies of genes create a redundancy in the sequence, which means the extra copies of genes can evolve ‘*de novo*’ (altered) functions for adaption and the original function is also upheld by the remaining copies (Ohno, 1970). However, there are a few too many disadvantages in comparison to the advantages stated above. The primary one is that CNVs have been widely reported to play a role in development of cancer (Vogelstein & Kinzler, 2002). Many a times the change in the copy number leads genomic disorders such as Red-green color blindness, Autism, Crohn’s disease, Alzheimer’s, Hemophilia, Thalassemia, Parkinsons disease etc. (Zhang, Gu, Hurles, & Lupski, 2009). There are a few studies where CNV has been susceptibly associated to diseases (Crespi & Crofts, 2012; Stankiewicz & Lupski, 2010).

It has been known that not all mutations cause damage to the human cells. Depending on the location of the mutations in the gene, the alteration can make no difference or can even be beneficial. In fact, of all the mutations that occur in human genome, only a few, known as driver mutations, can cause disease development (Vandin et al., 2012).

One of the main challenges in cancer genomics research is to identify the driver mutations. In recent years, next generation sequencing (NGS) and microarray techniques have become increasingly popular in characterizing driver mutations. These technologies can analyze cancer genomes for large cohorts of patients in a timely way, which allows more differentiation between driver mutations and passenger mutations than traditional Sanger sequencing technology. Although these are significant advancements, the development and application of new computational methods to analyze and characterize the vast amount of CNV data is still far behind the development of the technologies to generate the data (Chin et al., 2011).

1.2. Breast Cancer

1.2.1. Genetic Changes in Breast Cancer

There are some genes in human genome that control fundamental cell functions such as cell growth, cell division and cell repair. The cell functions need to be regulated strictly to ensure that DNA is copied properly. When these regulatory genes are mutated, they affect the cell functionalities. As a result, the damage of DNA is not repaired and cell growth and division becomes uncontrollable, which leads to formation of a tumor. Cancer occurs when the normal regulation of cells is interrupted. In case of breast cancer, the genetic mutations are obtained during an individual’s lifetime and affect certain breast cells. These changes are known as somatic mutations. In contrast, genetic mutations present in all cells in the human genome can also induce the risk of breast cancer. These genetic changes are known as germline mutations. The primary difference between somatic mutations and germline mutations is that germline mutations are inherited from parents. Those who are affected by germline mutations, in association with environmental (exposure to radiation, chemicals) and lifestyle factors (food habit, exercise), also influence whether a person will develop breast cancer or not (Rizzolo et al. , 2011).

1.2.2. Inherited Susceptibility to Breast Cancer

Studies have shown that around 10% of all breast cancers are related to inherited germline mutations (Willems, 2007). The related genes are divided into three different groups based on their mutation frequency and their significance in breast cancer susceptibility: a) High-penetrance group, b) Moderate-penetrance group and c) Low-penetrance group. *BRCA1* and *BRCA2* are two genes that are in high-penetrance group because of the high-risk association to breast cancer and ovarian cancer in women. *PTEN* and *TP53* genes are also considered as high risk genes (Willems, 2007), although researchers have found they are also associated with Cowden syndrome and Li-Fraumeni syndrome respectively. Even in men, *BRCA1* and *BRCA2* mutations can lead to breast cancer. The proteins that are produced from these genes are directly related to cell formation and DNA damage repair, which means mutations in these genes can lead to damaged DNA and uncontrollable cell growth. More recent studies have associated *CDH1* (hereditary diffuse gastric cancer) and *STK11* (Peutz-Jeghers syndrome) genes to be in high-penetrance group as well (Apostolou & Fostira, 2013). Overall, these high-risk genes account for around 25% of inherited breast cancers (Guttmacher, Collins, Wooster, & Weber, 2003).

There are several other mutations of genes which have been studied as possible risk factors for developing breast cancer. Based on their frequency accounted for the disease, they are in either moderate-penetrance or low-penetrance group. Genes in moderate-penetrance group, such as *CHEK2* (Caligo et al., 2004; Falchetti et al., 2008), *ATM* (Thompson et al., 2005) and more recently *PALB2, BRIP1* due to the involvement of *BRCA2* in Fanconi anemia pathway (Levy-Lahad, 2010), account for around 3-4% of the familial risk (Stratton & Rahman, 2008).

1.2.3. Types of Breast Cancer

Depending on how the cancer cells respond to different receptors, from a diagnostic point of view, breast cancer has been divided into four different subtypes. This division ensures that each type of breast cancer has been treated with different chemotherapy and other hormone therapies.

- **Hormone receptor - positive (ER+ & PR+):** This type of breast cancer is ‘positive’ to estrogen hormone, which means that cancer cells grow depending on its response to the said hormone. Around 80% of all breast cancers are reported to be ER+. Out of these, around 65% grow in response to progesterone hormone, thus known as PR+.

- **Human epidermal growth factor receptor 2 (HER2):** In some cases, the cells that contain tumor produce excess of HER2 protein. This type of cancer grows relatively fast in comparison to other types of breast cancer.

- **Triple positive and triple negative:** Triple positive breast cancer has ER+, PR+ and HER2+ subtypes. For triple negative one, it has subtypes ER-, PR- and HER2-.

Breast cancer can also be classified based on its expression profiles. Generally speaking, it is divided into five molecular subtypes:

- **Luminal A:** Luminal A breast cancer is hormone-receptor positive (ER+ and/or PR+). But it is HER2- and has low levels of the protein Ki-67, which controls the speed of cell growth. Luminal A cancers grow slowly and have the best prognosis (Braun et al., 2013).
- **Luminal B:** Luminal B breast cancer is also ER+ and PR+, and either HER2+ or HER2-. Luminal B breast cancers have high levels of Ki-67. Luminal B breast cancers generally grow slightly faster than Luminal A cancers and have slightly worse prognosis than Luminal A (Braun et al., 2013).
- **Basal-like:** Basal-like breast cancers are triple negative (ER-, PR-, HER2-). This type of cancer is more common among younger and African-American women (Alluri & Newman, 2014; Dietze, Sistrunk, Miranda-Carboni, O'Regan, & Seewaldt, 2015).
- **HER2-enriched:** This subtype of breast cancer is ER- and PR- but HER2+. HER2-enriched cancers grow faster than luminal cancers and can have a worse prognosis (Dai et al., 2015).
- **Normal-like:** This subtype of breast cancer is similar to Luminal A, ER+ and PR+, HER2-, and has low levels of the protein Ki-67. However, its prognosis is slightly worse than Luminal A.

1.3. Mutation Analysis Approaches

NGS and microarray methods allow sequencing and genotyping for a large cohort of patients, so there is much more information to be analyzed for identifying driver genetic mutations. One of the common approaches to analyze the genomic alterations is to look for recurrently mutated genes, which are mutated in a large group of the patients. The main idea behind this approach is that if the genes that are mutated in a large fraction of patients can be identified, they will correspond to non-random mutations. However, this approach is challenging. Although some cancer genes have higher mutation frequency rates (e.g., *TP53*) than non-cancerous genes, most of them have much lower mutation frequencies (Vandin et al., 2010). Instead of a single gene, driver mutations can target groups of genes in networks or pathways (Hanahan & Weinberg, 2002; Vogelstein & Kinzler, 2004).

In the following subsections, we will discuss about both single gene, pathway and network level of mutation analysis approaches.

1.3.1. Single Gene-level Analysis

The basic idea of this approach is to find mutated genes with a significantly higher mutation frequency or recurrently mutated genes in a collection of cancer patients. For example, The Cancer Genome Atlas (TCGA) study (McLendon et al., 2008) applied this method to 91 glioblastoma (GBM) patients and identified eight significantly mutated genes with false discovery rate (FDR) smaller than 0.001. They observed that the TP53 gene is mutated in approximately 38% of the patients. Another TCGA study examined 316 high-grade serous ovarian cancer patients and identified around 302 *TP53* mutations (Bell et al., 2011). However, this approach is challenging. Although some cancer genes have higher mutation frequency rates (e.g., *TP53*) than non-cancerous genes, most of them have much lower mutation frequencies (Vandin et al., 2010).

However, due to the spatial heterogeneity of cancer genomes, studying individual mutated genes is not sufficient to understand the mechanism of cancer. The single gene test sometimes fails to find other significant genes that are responsible for driver mutations. While the main reason behind the failure may be an insufficient number of patients, individual gene interactions that affect biological function may also affect results. The specific functions of the genes are altered due to mutations and in this case, the interaction between mutated genes can reveal information related to the mutations necessary for understanding the progression of cancer. Therefore, instead of using a single gene approach, driver mutations can target groups of genes, which can be broadly defined as gene clusters or subnetworks in networks or pathways (Hahn & Weinberg, 2002; Vogelstein & Kinzler, 2004).

1.3.2. Pathway and Network-level Analysis

Compared to analyzing genetic mutation data at single gene level, pathway and network analyses can extract more information as these methods deal with multiple genes in the same pathway or network, so the probability that a molecular event will pass the statistical threshold is increased and the number of hypotheses tested are reduced (Chi et al., 2014). Another benefit of pathway analysis is that results obtained for different related datasets can be compared easily, as pathway information can ensure the interpretation of the data is done in a common feature space (Creixell et al., 2015). This approach to test associations between mutation and phenotype at the pathway level has been implemented in different studies (Ding et al., 2008; Lin et al., 2007; Parsons et al., 2008). A recent study (Bellmunt et al., 2015) conducted a pathway-level analysis to predict the overall survival of in platinum treated locally advanced urothelial tumors. They found that 35 out of 103 patients had mutations in at least one of the two genes *TP53* and

PIK3CA, which consists of 16% and 9% of the total number of mutations identified in the study. The authors also found that around 65% of the patients had CNV mutations.

Pathway level analysis can increase the statistical power to identify significantly mutated pathways in specific cancers and has better biological interpretation. However, the approach of identifying pathways being mutated in large numbers of patients has its limitations too, because only existing pathways are considered, ignoring the topology of the pathways. Moreover, the pathways are analyzed in isolation but they interact in larger networks, which may neglect many groups of interacting genes that are not in known pathways but have significant association with clinical phenotypes (Vandin et al., 2012).

In recent years, methods to identify mutated subnetworks among cancer genomes have been introduced. Cerami et al. (2010) proposed a network-based approach based on the hypothesis that cellular networks are modular and have inter-connected proteins that perform specific biological functions. The authors used a unified molecular interaction network consisting of both protein-protein interactions and pathways to perform an integrated network analysis for identifying candidate driver mutations. Their approach was a combination analysis of sequence mutations and copy number alterations.

Vandin et al. (2012) introduced another approach to finding subnetworks by considering that mutations in the subnetwork are correlated with the clinical parameter of survival time of patients. They presented an algorithm called HotNet to identify significantly mutated subnetworks determined by the mutation frequency of individual genes along with the interactions between them. They considered the mutation as a source of heat on the network and extracted the 'hot' nodes. The significance of the subnetworks was calculated using both the topology of the networks and the frequency of mutation of the genes. Leiserson et al. (2014)

introduced a modified version of the previously mentioned Hotnet algorithm and called it Hotnet2. They used this updated algorithm to analyze a Pan-Cancer dataset of 3281 samples from 12 cancer types. The authors identified significantly mutated subnetworks with known pathways such as *TP53, RTK, PI3K* etc. However, the authors noted that some genes with very high individual mutation scores were absent from the network analysis results and stated that this is due to the lack of data as well as false negatives in the analyzed data.

One key limitation of the current subnetwork-based approaches is that they cannot assign the same mutated genes into different subnetworks although overlapped subnetworks are possible. This motivates us to apply two network-based clustering algorithms to analyze breast cancer CNV mutation data for identifying significantly mutated subnetworks. The identified subnetworks can be used to test the association of mutation status of the genes in the subnetworks with breast cancer patients' survival. The first approach is HotNet2 (Leiserson et al. (2014) while the other approach is called ClusterOne (Nepusz et al., 2012), which has not been applied to analyze cancer mutation data before, but was developed and applied to identify overlapped protein complexes in protein interaction networks. We adopted this approach since a gene can be assigned to multiple subnetworks and genes are usually involved in multiple complexes and pathways. We developed a new mutation analysis pipeline by taking network topology into account for measuring gene-pair's mutation similarity to infer the significantly mutated subnetworks.

Chapter 2

Motivation and Research Objectives

2.1. Motivation

Since genes (proteins) usually interact with other genes to execute their functions, networks can be modular and divided into subnetworks. It is reasonable to assume that clinically relevant mutations of breast cancer occur in closely interacting genes and cancer is an outcome of coordinated dysfunction of these closely connected subnetworks enriched with clinically informative cancer mutations. The proposed thesis thus aims at developing a new bioinformatics analysis pipeline to identify the mutated subnetworks in the breast cancer genome. We used two available clustering algorithms to establish the pipeline.

2.2. Hypothesis

We hypothesize that the significantly mutated subnetworks identified in the breast cancer genome using a novel network-based mutation analysis approach will lead to the finding of new network-based biomarkers for breast cancer survival prognosis, which may be useful for stratifying cancer patients for choosing optimal treatments. The proposed mutation analysis pipeline can also be applied to other cancer data analysis.

2.3. Research Objectives

Our research has two major objectives: first, we will use graph-based clustering algorithms to identify the significantly mutated subnetworks in cancer genome. We will test the algorithm using breast cancer patient mutation data and human protein interaction networks collected from different sources; second, we will associate the mutation patterns of the genes in the subnetworks with patient’s survival. The subnetworks which have significant association with patient’s survival will be used as prognostic biomarkers for breast cancer.

Chapter 3

Materials and Methods

3.1. Data Collection and Filtering

Identification of significantly mutated subnetworks in breast cancer genome requires both patient-specific mutations and gene interaction networks. For patient-specific mutation data, we used CNV data obtained from the METABRIC (Molecular Taxonomy of Breast Cancer International Consortium) (Curtis et al., 2012). The CNV calls were identified from approximately 2000 clinically annotated primary fresh frozen breast cancer specimens along with a portion of normal specimens from different North American and European tumor banks. Initially, a set of 997 samples including paired DNA and RNA profiles were analyzed. The authors represented these 997 female patient data as ‘Discovery set’. A second group of 995 samples were represented as ‘Validation set’, for which the paired DNA and RNA profiles were not available during the initial study. The purpose of the validation set was to test the reproducibility of the clinical outcome associations. There were 5 discrete somatic states: HOMD, HETD, NEUT, GAIN and AMP. Out of these somatic states, HOMD and HETD were broadly called as loss and GAIN and AMP were broadly called as gain and NEUT was represented by neutral. The CNV calls were represented for gain, loss and neutral as 1, -1 and 0 respectively. We filtered out the data where the CNV length were larger than or equal to 5 KB and the number of probes was larger than or equal to 10. In the Discovery set, after filtering there were a total of 131,956 calls (38,647 were CNV loss and 93309 were CNV gain). In the

Validation set, after filtering there were a total of 137,896 calls (42824 were CNV loss and 95072 were CNV gain).

The gene interaction network data was taken from Menche et al. (2015), which included 141,296 interactions between 13,460 human proteins (genes). The authors have stated that they did not include interactions from gene expression data. The interactions are primarily protein-protein interactions (PPI), but also included other types of physical interactions such as regulatory interactions, protein complexes from the comprehensive resource of mammalian protein complexes (CORUM) database and signaling interactions (Ruepp et al., 2010). The authors treated the interactions as an undirected network.

3.2. Methods

To identify the clinically relevant and statistically significantly mutated subnetworks, we developed an analysis pipeline as shown in **Figure 1**. Briefly speaking, we first retrieved gene information in each sample-specific individual CNV region; then we calculated gene-specific mutation frequency and gene-pair specific mutation similarity score, which, together with the gene interaction network, were passed to Hotnet2 and ClusterONE tools to identify statistically significantly mutated subnetworks and gene clusters, respectively; finally, we evaluated the clinical and pathway significance of the overlapped genes identified in both the mutated subnetworks and gene clusters.

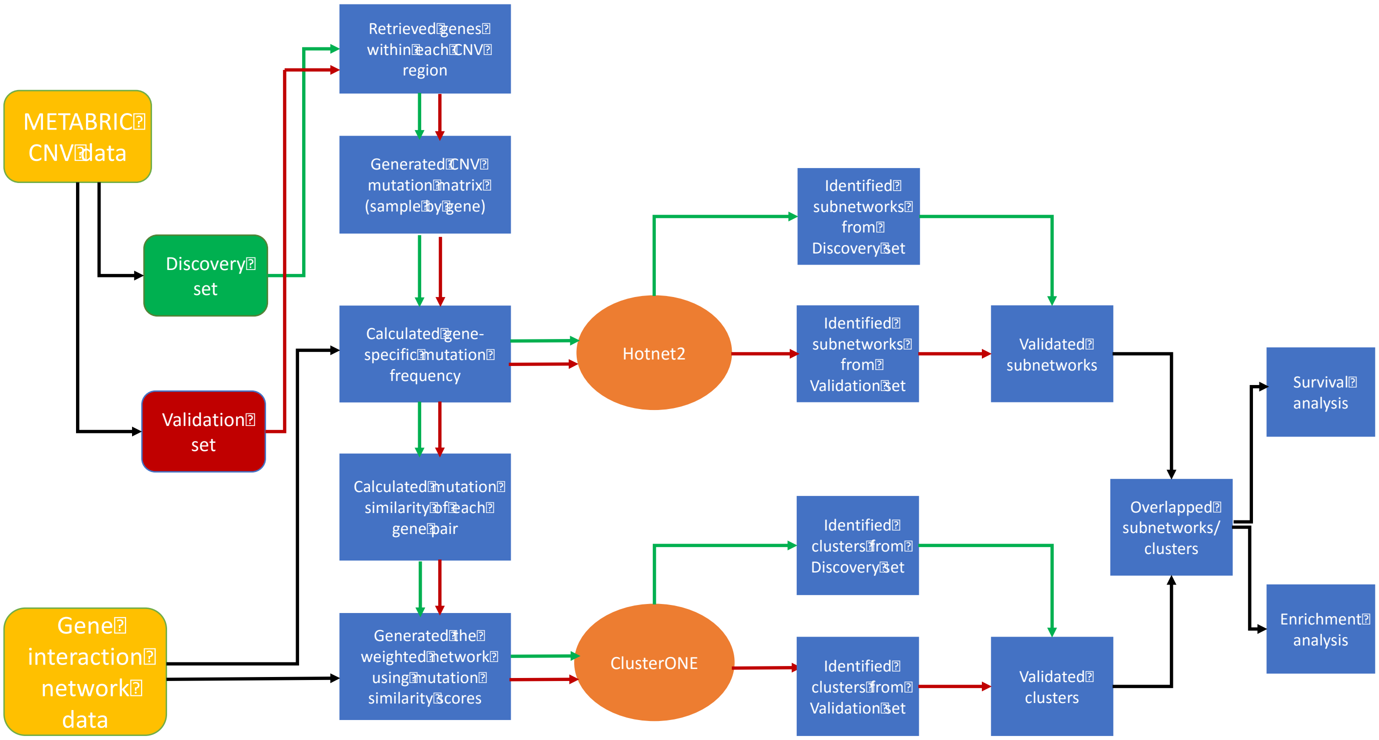

METABRIC CNV data
Discovery set
Validation set
Gene interaction network data
Retrieved genes within each CNV region
Generated CNV mutation matrix (sample by gene)
Calculated gene-specific mutation frequency
Calculated mutation similarity of each gene pair
Generated the weighted network using mutation similarity scores
Hotnet2
ClusterONE
Identified subnetworks from Discovery set
Identified subnetworks from Validation set
Identified clusters from Discovery set
Identified clusters from Validation set
Validated subnetworks
Validated clusters
Overlapped subnetworks/clusters
Survival analysis
Enrichment analysis

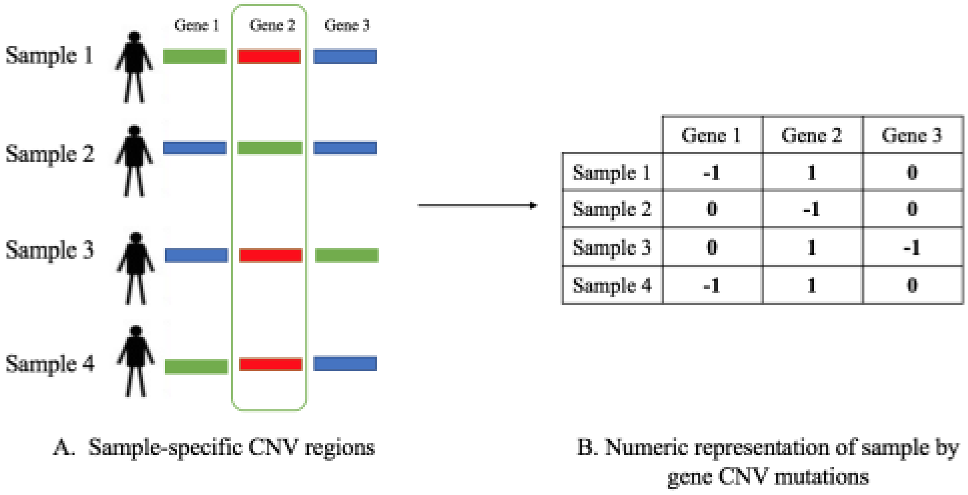
Figure 1. An analysis pipeline to identify significantly mutated gene subnetworks

Hotnet2 and ClusterONE were used to identify significantly mutated gene subnetworks or clusters from a curated gene interaction network and CNV mutation data in METABRIC Discovery and Validation sets, respectively. The clinical and biological pathway significance of the overlapped genes in the subnetworks (clusters) identified by both algorithms in both Discovery and Validation sets were evaluated.

3.2.1. Retrieve CNV-specific genes

We retrieved gene information for each of the patient-specific CNV regions using BiomaRt R package (Smedley et al., 2015) (**Figure 2A**). We used 'hsapiens_gene_ensembl' dataset from the ENSEMBL database, which contains human genes. We used the hg19 version of the database to retrieve the CNV-specific genes. The parameters we used are from the filtered CNV data: the chromosome ID, and the start and end locations of the CNVs in the chromosome.

After we retrieved the genes for each CNV region, we generated a sample-by-gene CNV mutation matrix, where the rows were the filtered samples and the columns were the genes found in the CNV regions. The gene- and sample- specific CNV mutation matrix was generated by expressing the non-mutated genes as ‘0’ and mutated genes as ‘1’ for gain and ‘-1’ for loss, respectively (**Figure 2B**).



A. Sample-specific CNV regions

	Gene 1	Gene 2	Gene 3
Sample 1	-1	1	0
Sample 2	0	-1	0
Sample 3	0	1	-1
Sample 4	-1	1	0

B. Numeric representation of sample by gene CNV mutations

Figure 2. From sample specific CNV regions to gene- and sample-specific CNV mutations

Gene information was retrieved for each sample-specific CNV region (**Figure 2A**), where regions labeled as blue represent no CNV change, regions labeled as red represent as CNV gain and regions labeled as green represent as CNV loss. The non-mutated genes are expressed as ‘0’ and mutated genes are expressed as ‘1’ for gain and ‘-1’ for loss, respectively (**Figure 2B**).

3.2.2. Calculate gene-specific mutation frequency

The next step after generating the mutation matrix was to calculate the gene-specific mutation frequency. The main motivation behind calculating gain and loss frequencies was to have a measure for gene-specific mutation score. We used this mutation frequency as the ‘heat score’, which is required to run the Hotnet2 software, as described in Section 3.2.4.

The mutation frequency for each gene *i* was calculated by:

$$f_{ig} = \frac{n_g}{N_s} \tag{1}$$

$$f_{il} = \frac{n_l}{N_s} \tag{2}$$

Here N_s denotes the total number of samples in the dataset, whereas n_g and n_l are the total number of CNV gains and CNV losses of gene *i* in all the samples, respectively. f_{ig} and f_{il} are the mutation frequencies of CNV gain and loss in gene *i*, respectively.

3.2.3. Calculate gene-pair specific mutation similarity

We calculated the pairwise gene mutation similarity in the gene interaction network by taking the genes' mutation frequencies into account. We used this measure of gene-pair specific mutation score as the weight for each interaction in the network. This similarity score for each pair is used as an input to ClusterONE, as described in Section 3.2.4. The original ClusterONE algorithm had a constant weight (=1) for all the interactions, and the authors stated that using a weighted approach may yield improved results (Nepusz et al., 2012).

For simplicity, we used the cosine similarity (Jiang et al. 2015) and equations (1) and (2) to measure the gene-pair mutation similarity:

$$sim(i,j) = \frac{\sum_{k\epsilon\{g,l\}} f_{ik} f_{jk}}{\sum_{k\epsilon\{g,l\}} (f_{ik})^2 \sum_{k\epsilon\{g,l\}} (f_{jk})^2} \tag{3}$$

The similarity measure $sim(i,j)$ for genes i and j is defined by mutation frequency f. In our case, we have two types of mutation frequencies: gain ($k = g$) and loss ($k = l$). For our work, we treated the gene-pair similarity as a network edge weight. As the gene interaction network we obtained from Menche et al. (2015) is unweighted, we have assigned the gene similarity to the edges of the network.

3.2.4. Identify significantly mutated subnetworks

We have used two different algorithms to identify significantly mutated subnetworks from our collected breast cancer CNV data and the gene interaction network. The first algorithm is **Hotnet2**, which identifies mutated subnetworks of a genome-scale interaction network (Leiserson et al., 2014), and the second one is **ClusterONE**, which identifies clusters or groups of interacted genes in the gene interaction network (Nepusz et al. , 2012). We briefly discuss these two algorithms as follows.

Hotnet2: HotNet2 (HotNet diffusion oriented subnetworks) identifies subnetworks that are mutated more frequently than the general rate of mutation by chance. The authors used an insulated heat diffusion approach, which not only considers the mutation score for each gene, but also leverages the topology of interactions between the genes. The inputs to HotNet2 are: a heat score vector h that contains the mutation score for each gene, and a graph $G = (V, E)$, where each node $v \in V$ corresponds to a gene/protein and each edge $e \in E$ corresponds to an interaction between the corresponding genes/proteins. In the first step, the algorithm performs 'heat diffusion' to extract the local topology of the interaction network. At each iteration, the nodes (genes) in the network send and collect heat from the neighboring nodes. The authors define an insulating parameter β, which denotes the fraction of the heat retained by each node. The iterations terminate when it reaches its equilibrium. HotNet2 identifies strongly connected components in the network and returns a list of subnetworks, each containing at least k genes. The statistical significance of the returned list of subnetworks is then calculated for the number of subnetworks with at least k genes that are returned and the false discovery rate for the subnetworks are also estimated. In our study, the gene-specific mutation score is calculated based on the mutation frequency f_i $\left(f_i = f_{ig} + f_{il}\right)$.

ClusterONE: The ClusterONE (Clustering with overlapping neighborhood expansion) algorithm uses a greedy growth process to find groups of genes with high cohesiveness in a gene interaction network. The authors generalized two structural properties of a protein complex represented by a subgraph to define cohesiveness for each groups: the interaction between its subunits and the separation from the rest of the network. For a group of proteins P, the cohesiveness $C(P)$ is defined by:

$$C(P) = \frac{total\ weight\ for\ internal\ edges}{total\ weight\ for\ internal\ edges + total\ weight\ for\ boundary\ edges + p} \tag{4}$$

The term ‘internal edges’ is used to define the edges consisting of entirely a given group and ‘boundary edges’ are the edges that has connection with the rest of the network. The constant p is a penalty term to model the uncertainty in the data.

The ClusterONE algorithm mainly has three steps. First, the algorithm follows a greedy approach to select the protein which has the highest degree as the first seed, and starts to grow a cohesive group from the initial seed. While selecting the next seed, the algorithm considers all the proteins that are not currently included in any other networks (protein complexes). Then the node with the highest degree is taken again. This step continues until there are no more proteins left to consider. The growth process ensures that any vertex from the group can be removed in later iterations if necessary. This includes the initial seed vertex as well. The seed vertex is selected as an outlier if it is not included in the final group. This means that the vertex will not be included in any of the clusters.

In the next step, highly overlapping pairs are merged based on the optimal cohesive groups. The authors merge pairs of groups which has an overlap score (ω) larger than 0.8. For two protein sets X and Y, the authors defined ω as:

$$\omega(X, Y) = \frac{|X \cap Y|^2}{|\, X \,||\, Y|} \tag{5}$$

The authors state that the merges may be performed one after another or concurrently. For the first approach, the problem is that the overlap scores need to be recalculated in each iteration, after each merging occurs. To avoid this problem, ClusterONE uses the concurrent approach. From a collection of cohesive groups, ClusterONE constructs an overlap graph. Each node in the overlap graph finds a cohesive group, and two nodes are connected if the overlap score is more than the selected threshold ($\omega \geq 0.8$). Nodes that have a direct (one-to-one) or indirect (via paths of edges) connection are then merged to convert them into protein complex. If a node does not have an edge, no merging is done and it is promoted to a protein complex candidate.

In the final step, the algorithm discards complexes which have a density below a particular threshold δ. In our study, we consider the gene network is weighted and the weight of each gene pair is based on the gene-pair mutation similarity $sim(i, j)$.

3.2.5. Parameters used to run Hotnet2 and ClusterONE

Hotnet2 runs in four consecutive steps: The first step is to create an influence matrix that defines an "influence score" for each gene pair in the network based on known gene interactions and a heat diffusion process. The second step is heat score generation, where the tool creates a JSON file containing heat scores for each gene. In our case, the heat scores were directly calculated from the mutation score. The third step is delta selection, which uses permutated data sets to select thresholds that should be used for edge weight removal in the final step. The output of this step includes a list of selected thresholds for each permutation test. We took the median of the deltas across all permutation tests we performed. The final step identifies the mutated subnetworks based on the influence matrix and heat score, removing edges with weight less than delta, and extracting the resulting connected components. This step does not need any additional parameters. The parameters we used in each of the steps are shown in **Table 1**.

To run ClusterONE, we only need a weighted network. We did not use any additional parameters.

3.2.6. Validating Results and Retrieving Overlapped Clusters

We obtained the subnetworks by running the Hotnet2 software step by step using the parameters described in **Table 1**. After obtaining results using both Discovery and Validation datasets, we validated the subnetworks that were identified in both datasets. The validation is done by finding the overlapped subnetworks from Validation set with the subnetworks identified from Discovery set. The overlap score threshold was set to larger than or equal to 50%. The clusters identified by running ClusterONE from both the Discovery and Validation sets were also validated by the same approach.

We then performed survival analysis and pathway enrichment analysis on the validated subnetworks and clusters which is discussed in the following subsections.

Table 1. Parameters used to run Hotnet2

Step 1: Influence Matrix Creation			
Dataset	**Parameter Name**	**Value**	**Description**
Discovery	Edge swap constant, Q	115	The software performs $Q \times no\ of\ edges$ swaps
	Number of permutations	1000	Number of permuted networks to create
Validation	Edge swap constant, Q	115	The software performs $Q \times no\ of\ edges$ swaps
	Number of permutations	800	Number of permuted networks to create
Step 2: Heat Score Generation			
Discovery	Minimum Heat Score	1	The minimum score for the gene-specific mutation frequency
Validation	Minimum Heat Score	1	The minimum score for the gene-specific mutation frequency
Step 3: Delta Selection			
Discovery	Minimum Network Size	3	Minimum size of the connected components that should be returned
Validation	Minimum Network Size	3	Minimum size of the connected components that should be returned

3.2.7. Survival Analysis

To identify potential network biomarkers for breast cancer prognosis, we evaluated the association of mutation patterns of the genes in the identified subnetworks using HotNet2 and clusters using ClusterONE with breast cancer survival using Kaplan-Meier plots (Collett, 2003). To do this, the sample-specific mutation score was first quantified using an unweighted and weighted approaches, respectively.

The first approach is an unweighted version. From the mutation influence matrix generated previously, we extracted a submatrix that includes the genes that we found in the identified subnetwork. We then counted the number of mutated occurrences (either gain or loss, 1 or -1 from the mutation matrix) for each sample, which is called as sample-specific unweighted mutation score. For example, we can generate the mutation submatrix as shown in **Table 2** from the mutation matrix we generated (Subsection 3.2.1.). The calculated mutation score is also shown in the **Table 2**.

Table 2. Calculation of sample-specific unweighted mutation score

	KCNRG	*TRIM13*	*DGCR6L*	*RIMBP3*	*NOXA1*	*RAC2*	*ICMT*	Unweighted mutation Score
Sample 1	1	0	0	-1	0	0	1	3
Sample 2	0	1	-1	1	1	0	0	4
Sample 3	1	1	0	0	0	1	1	4
Sample 4	0	-1	1	0	-1	-1	-1	5
Sample 5	0	0	0	1	0	0	0	1
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
Sample n	-1	0	0	0	1	-1	0	3

The second approach is a weighted version. The mutation rate calculated for each gene in Subsection 3.2.2 can be treated as weight for the mutation score calculation in **Table 2**. For each gene i in a given subnetwork, we have a gene-specific mutation frequency f_i $\left(f_i = f_{ig} + f_{il}\right)$ as shown in **Equations 1** and **2**, so we can calculate the sample-specific weighted mutation score as:

$$p_j = \sum_{i=1}^{G} f_i \ \cdot |v_{ji}| \tag{6}$$

Here p_j is the mutation score for sample j, G is the number of genes in the subnetwork, f_i is the mutation frequency score for the gene i, and v_{ji} is the copy number variation status of gene i in sample j (see **Figure 2B**). For example, from the generated mutation matrix shown in **Table 3**, we can calculate the sample-specific weighted mutation score for sample 1 as follows:

$$p_1 = (0.3\times1)+(0.38\times|-1|)+(0.5\times1) = 1.18$$

Table 3. Calculation of sample-specific weighted mutation score

	KCNRG *f = 0.3*	*TRIM13* *f = 0.6*	*DGCR6L* *f = 0.25*	*RIMBP3* *f = 0.38*	*NOXA1* *f = 0.2*	*RAC2* *f = 0.4*	*ICMT* *f = 0.5*	Weighted mutation score
Sample 1	1	0	0	-1	0	0	1	1.18
Sample 2	0	1	-1	1	1	0	0	1.43
Sample 3	1	1	0	0	0	1	1	1.80
Sample 4	0	-1	1	0	-1	-1	-1	1.95
Sample 5	0	0	0	1	0	0	0	0.38
.	.	.	.	.	.	.	.	.
Sample n	-1	0	0	0	1	-1	0	0.9

We used the product-limit method, also known as the Kaplan-Meier (KM) method, to estimate a survival function. KM method is a nonparametric technique that uses the exact survival time for each individual in a sample instead of grouping the times into intervals. The Kaplan-Meier estimator can be written as follows:

For $t \in [t_j , t_{j+1})$, j=1,2,3,.., we have

$$\widehat{S}\,(\mathrm{t}) = \prod_{i=1}^{j} \left(1 - \frac{d_i}{n_i}\right) \tag{7}$$

Here, $\widehat{S}$ (t) = Probability of surviving at least to time t

t = actual time of death of the n individuals in the study;

t_j = actual time of death at time interval *j*

d_i = # of deaths that occur at time interval *i*

n_i = # of patients remaining at time interval *i*

To perform Kaplan-Meier analysis, we categorized the breast cancer patients into high mutation group (patients with the mutation score larger than or equal to the median of the mutation scores of all patients) and low mutation group (patients with mutation score less than median of the mutation scores of all patients). We used *survival,* an R package for this analysis.

3.2.8. Pathway Enrichment Analysis

Using the gene list from each of the identified subnetworks, we performed the enrichment analysis of gene ontologies (biological processes and molecular functions) and biological pathways (REACTOME and KEGG) using the Enrichr (Kuleshov et al., 2016) R package. The

Enrichr software contains a diverse and up-to-date collection of over 100 gene set libraries available for analysis and download. It is used to perform pathway enrichment analysis on the identified overlapped genes from both ClusterONE clusters and Hotnet2 subnetworks to identify which pathways are over represented in the overlapped genes.

Pathway enrichment analyses use a predefined (a priori) knowledge of gene sets, regarding the involvement of genes in biological pathways or biological functions. The main focus is to analyze whether majority of the input gene in a gene list is over-represented in the gene sets. The degree of the over-representation is quantified by an enrichment score (Subramanian et al., 2005). The analysis is mainly divided into three parts: First, the enrichment score is calculated based on the number of over-represented genes in a given gene set for both the input gene list and the background gene list (e.g. the whole human genome) and , then the statistical significance of the calculated score is estimated. Finally, the false discovery rate is calculated after normalizing the scores for each gene set.

Chapter 4

Results and Discussions

4.1. Clinical Characteristics

The Discovery and Validation data sets have very similar distributions in age, and expression levels of progesterone receptor (PR) and HER-2 (*ERB-B2*) (*P*>0.05) (Table 4). On the other hand, the two sets have statistically significant differences in PAM50 subtypes, grade, the tumour stage with breast cancer patients in the Discovery set having a much higher prevalence in stage 0 compared to the Validation set (49% vs 12%), and in expression of estrogen (ER) (*P*<0.05).

Table 4. Clinical characteristics

Characteristic	METABRIC Discovery	METABRIC Validation	*P*†
Age	61(51, 70)*	63(52,71)	0.2107
Subtype			0.0005
Normal	58(6%)‡	144(15%)	
LumA	454(46%)	255(26%)	
LumB	266(27%)	222(23%)	
Her2	84(9%)	153(15%)	
Basal	118(12%)	211(21%)	
Grade			0.0095
1	68(7%)	98(11%)	
2	407(42%)	356(40%)	
3	505(51%)	444(49%)	
Stage			0.0005
0	480(49%)	12(2%)	
1	177(18%)	185(34%)	
2	274(28%)	296(55%)	
3	40(4%)	46(9%)	
4	9(1%)	1(0.1%)	
ER-expr	784(80%)	712(72%)	<0.0001

PR-expr	517(53%)	517(52%)	0.9282
Her2-expr	112(11%)	132(13%)	0.1940

* For continuous variables (Age), quantiles are presented.

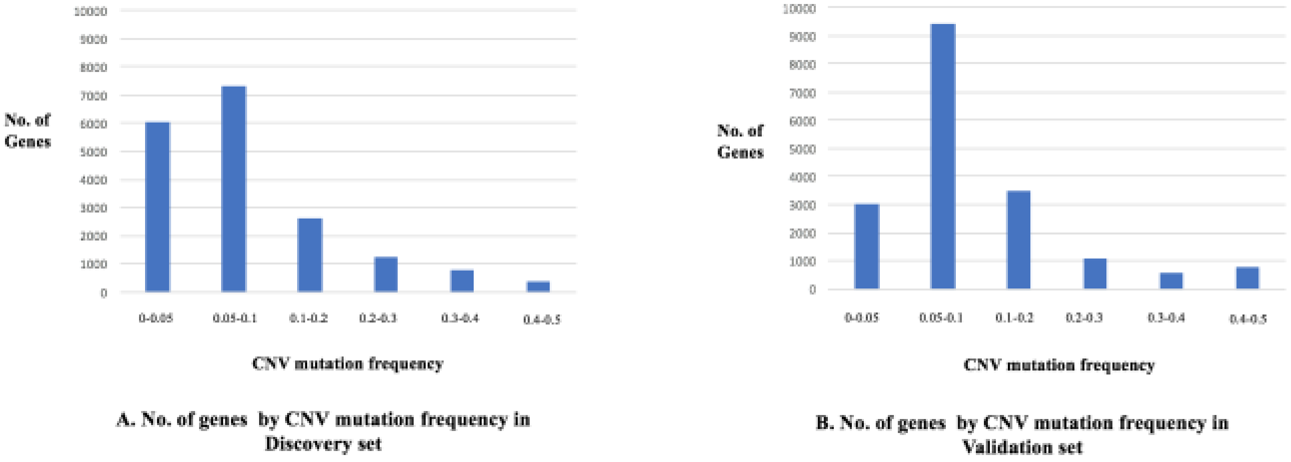
† *P*-values were determined by Wilcoxon rank sum test for continuous variables and Chi-square test for categorical variables

‡ The number of patients in each category and its proportion are presented. In ER-expr, PR-expr and Her2-expr, only the number of positive case are presented.

4.2. Mutation frequency and mutation similarity

Based on the individual-specific CNV positions, we retrieved 18,341 genes in both the Discovery and Validation sets. Based on the gene- and sample-specific mutation matrix (**Figure 2B**), the genes with the highest number of CNV mutations were *SLC41A1* and *LEMD1*, both with a CNV gain mutation frequency of approximately 46%. In both of the Discovery (**Figure 3A**) and Validation (**Figure 3B**) sets, approximately 70% of the genes have a mutation frequency lower than 10%. There are 6.1% and 7.3% of the genes with a mutation frequency higher than 30% in the Discovery (**Figure 3A**) and Validation (**Figure 3B**) sets, respectively. The gene-specific mutation frequencies were treated as mutation scores in Hotnet2 tool to identify significantly mutated subnetworks.

We calculated the mutation similarity for the gene pairs of the 141,296 gene interactions from the gene interaction network. We divided the interactions into 3 different groups: high mutation group (interactions with a mutation similarity score at least 0.9), medium mutation group (interactions with a mutation similarity score less than 0.9 but at least 0.5) and low mutation group (mutation similarly scores below 0.5). Based on these thresholds, we found a total of 51,953, 37,251 and 52,092 interactions in the high, medium and low mutation groups, respectively. The gene-pair specific mutation similarities were treated as weights for the gene interaction network in ClusterONE, to identify significantly mutated gene clusters



A. No. of genes by CNV mutation frequency in Discovery set

B. No. of genes by CNV mutation frequency in Validation set

.

Figure 3. Distribution of number of genes by CNV mutation frequency

Number of genes by CNV mutation frequency in Discovery set (**Figure 3A**) and Validation set (**Figure 3B**).

4.3. Significantly mutated subnetworks identified by Hotnet2

We found a total of 99 and 79 subnetworks, that have at least 3 interacting genes, from the Discovery and Validation data sets, respectively. Ten of these subnetworks were identified as significantly mutated subnetworks based on (1) adjusted *P*-value < 0.1; (2) the overlapping rate (number of overlapped genes divided by the maximum number of genes in either Discovery or Validation set) being greater than or equal to 50% (**Table 5**). All of these 10 subnetworks have 3-7 interacting genes.

Survival analysis showed that the mutation patterns of the genes in the four identified subnetworks (the subnetworks with bold color in **Table 5**) are significantly associated with breast cancer survival (Detailed in Section 4.6).

Table 5. Significantly mutated subnetworks identified by HotNet2

The subnetworks shown in this table were selected based on (1) adjusted *P*-value < 0.1; and (2) the overlapping rate (number of overlapped genes divided by the minimum number of genes in either Discovery or Validation set) being larger than or equal to 50%.

Discovery set		Validation set		No. of overlapped genes	Overlapped subnetwork ID
Subnetwork ID	No. of genes	Subnetwork ID*	No. of genes		
18	**5**	**9**	**6**	**5**	**S1**
32	4	6	7	4	S2
38	4	20	5	4	S3
49	**4**	**8**	**7**	**4**	**S4**
50	**4**	**21**	**5**	**4**	**S5**
79	3	23	4	3	S6
81	**3**	**33**	**4**	**3**	**S7**
89	3	16	6	3	S8
93	3	35	4	3	S9
96	3	13	6	3	S10

* Subnetworks identified in Validation set with the largest number of overlapped genes for a given subnetwork identified in Discovery set. The bold subnetworks are significantly associated with cancer survival (*P*-value<0.01) discussed in Section 4.6.

4.4. Significantly mutated gene clusters identified by ClusterONE

We identified 18 significantly mutated gene clusters in both Discovery and Validation sets, which have (1) adjusted p-value < 0.1; and (2) overlapping rate greater than or equal to 50% (**Table 6**). Five of the 18 clusters have at least 50 genes. Ten of the 18 clusters have fewer than 30 genes. The heatmaps in **Figure 4** and **Figure 5** shows the overlapping rate between the 18 clusters from both Discovery and Validation set, respectively. It appears that some of the clusters have shared genes (that is, the same gene can be assigned in multiple clusters). For example, cluster C1 has shared genes with clusters C2, C3 and C12 in both Discovery and Validation sets, respectively. **Figure 6** shows the heatmap of the overlapping rate between Discovery and Validation clusters. It is clear that all of the 18 clusters identified in the Discovery set were validated by the Validation set (see the score on the diagonal line). Survival analysis showed that the mutation patterns of the genes in the identified clusters #1, #6, #8, #9, #13, and #16 (the clusters with bold color in **Table 6**) are significantly associated with breast cancer survival (Detailed in Section 4.6).

Table 6. Clusters identified by ClusterONE in Discovery and Validation sets

The clusters shown in this table were selected based on (1) adjusted p-value < 0.1; and (2) the overlapping rate (number of overlapped genes divided by the minimum number of genes in either Discovery or Validation set) being greater than or equal to 50%. The bold subnetworks are significantly associated with cancer survival (*P*-value<0.01) discussed in Section 4.6.

Discovery set			Validation set			No. of Overlapped genes
Cluster ID	No. of genes in clusters	Adjusted p-value	Cluster ID	No. of genes in clusters	Adjusted p-value	
C1	**275**	**0**	**C1**	**280**	**0**	**269**
C2	154	0	C2	154	0	154
C3	77	0	C3	77	0	77
C4	58	0	C4	59	0	58
C5	55	0	C5	55	0	55
C6	**40**	**0**	**C6**	**40**	**0**	**40**
C7	24	4.95E-07	C7	24	4.94E-07	24
C8	**35**	**2.57E-05**	**C8**	**37**	**2.33E-05**	**33**
C9	**46**	**8.31E-05**	**C9**	**45**	**3.54E-05**	**41**
C10	19	2.35E-04	C10	19	2.02E-04	19
C11	19	2.96E-04	C11	26	2.05E-04	19
C12	26	2.96E-04	C12	51	2.18E-03	26
C13	**13**	**6.29E-03**	**C13**	**12**	**9.3E-03**	**12**
C14	12	8.73E-03	C14	20	1.33E-02	12
C15	16	3.93E-02	C15	16	2.95E-02	16
C16	**8**	**5.18E-02**	**C16**	**8**	**5.15E-02**	**8**
C17	16	6.04E-02	C17	16	5.25E-02	16
C18	17	7.11E-02	C18	17	7.8E-02	17

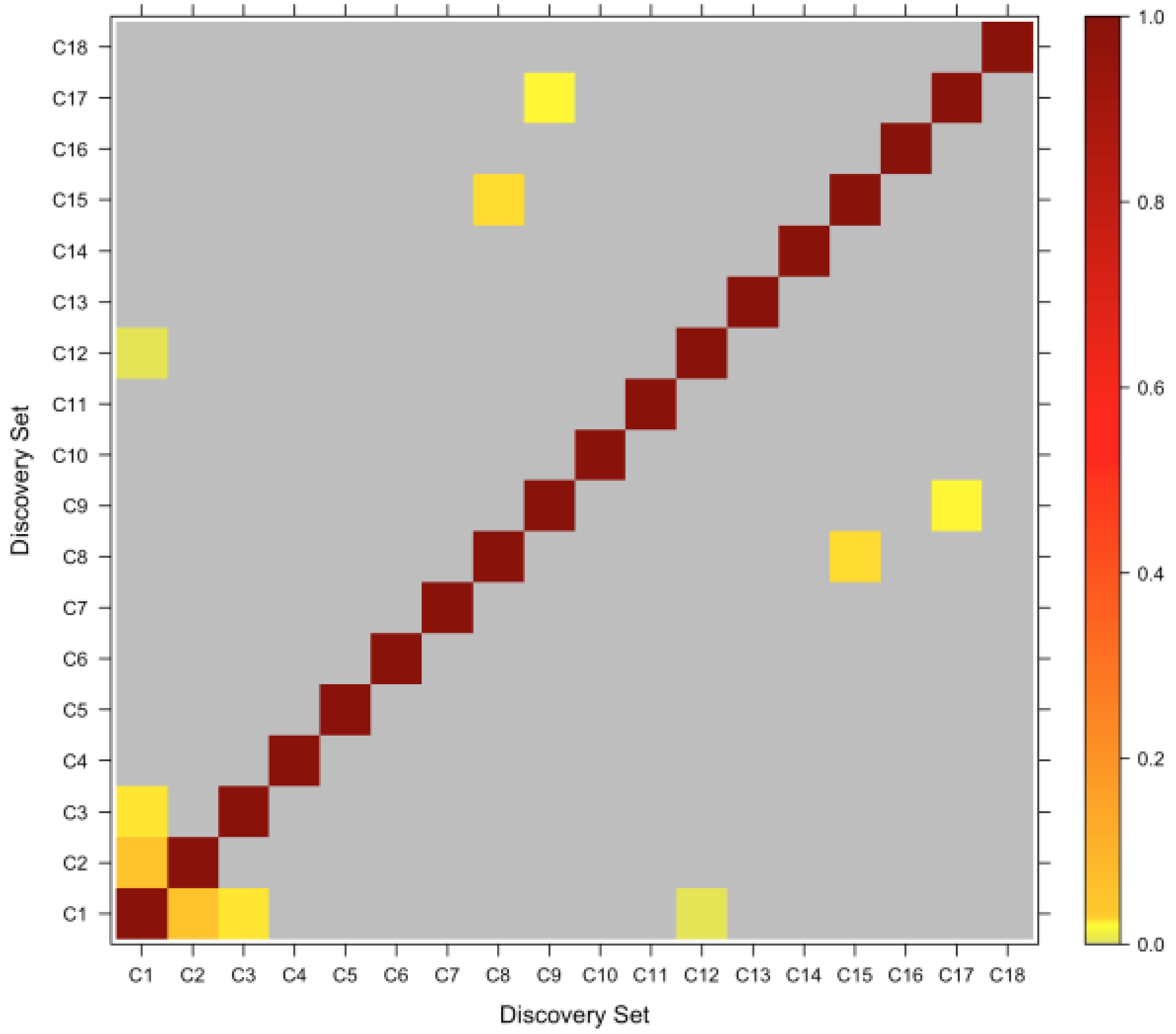


Figure 4. Heatmap of overlapping score for each pair of clusters identified by ClusterONE in Discovery set

The overlapping score is defined as the number of overlapped genes divided by the minimum number of genes in the ClusterONE cluster). The score is represented with the bar in the right, red being the highest overlapping score (1.0) and space grey being the least.

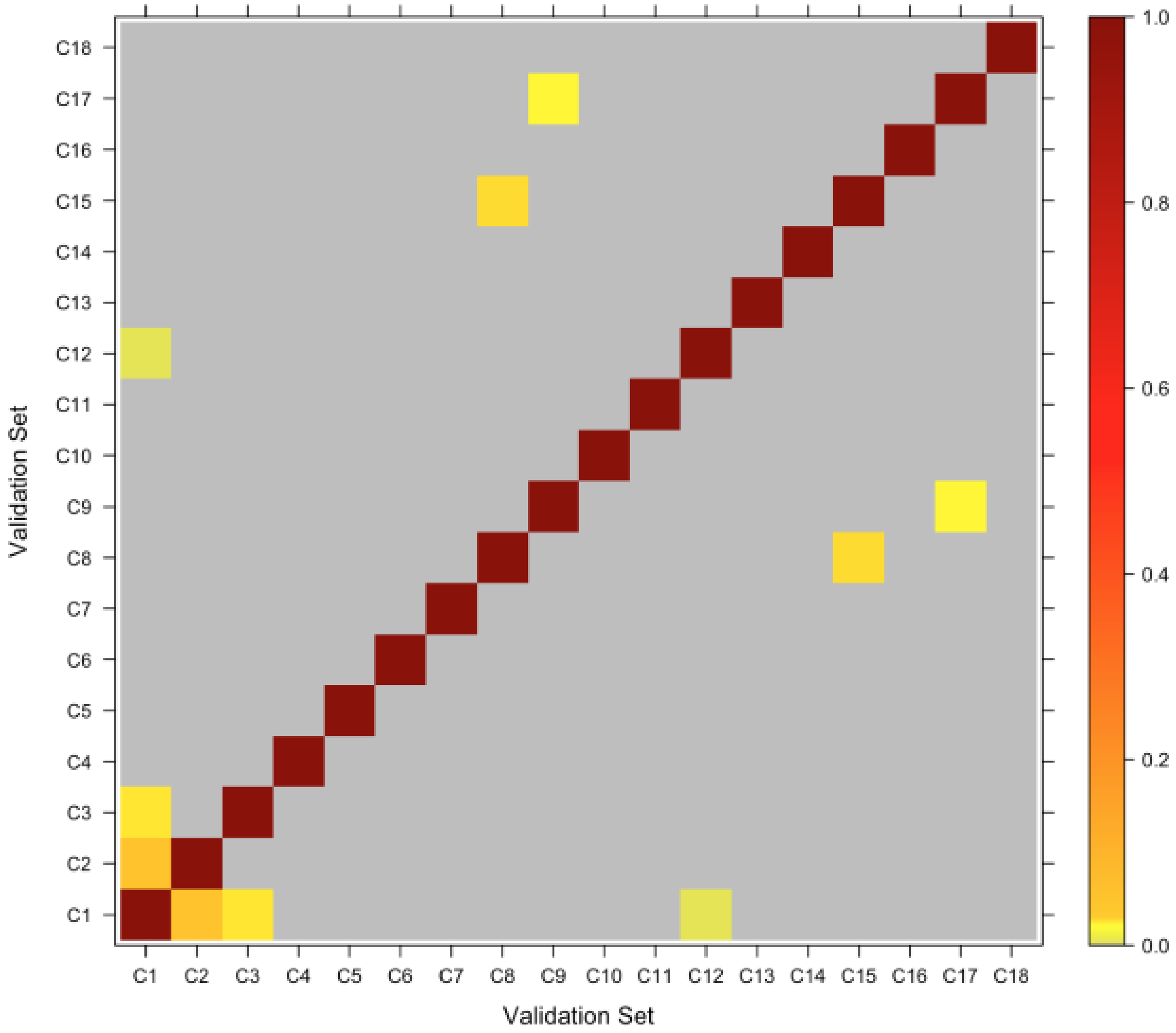


Figure 5. Heatmap of overlapping score for each pair of clusters identified by ClusterONE in Validation set

The overlapping score is defined as the number of overlapped genes divided by the minimum number of genes in the ClusterONE cluster). The score is represented with the bar in the right, red being the highest overlapping score (1.0) and space grey being the least.

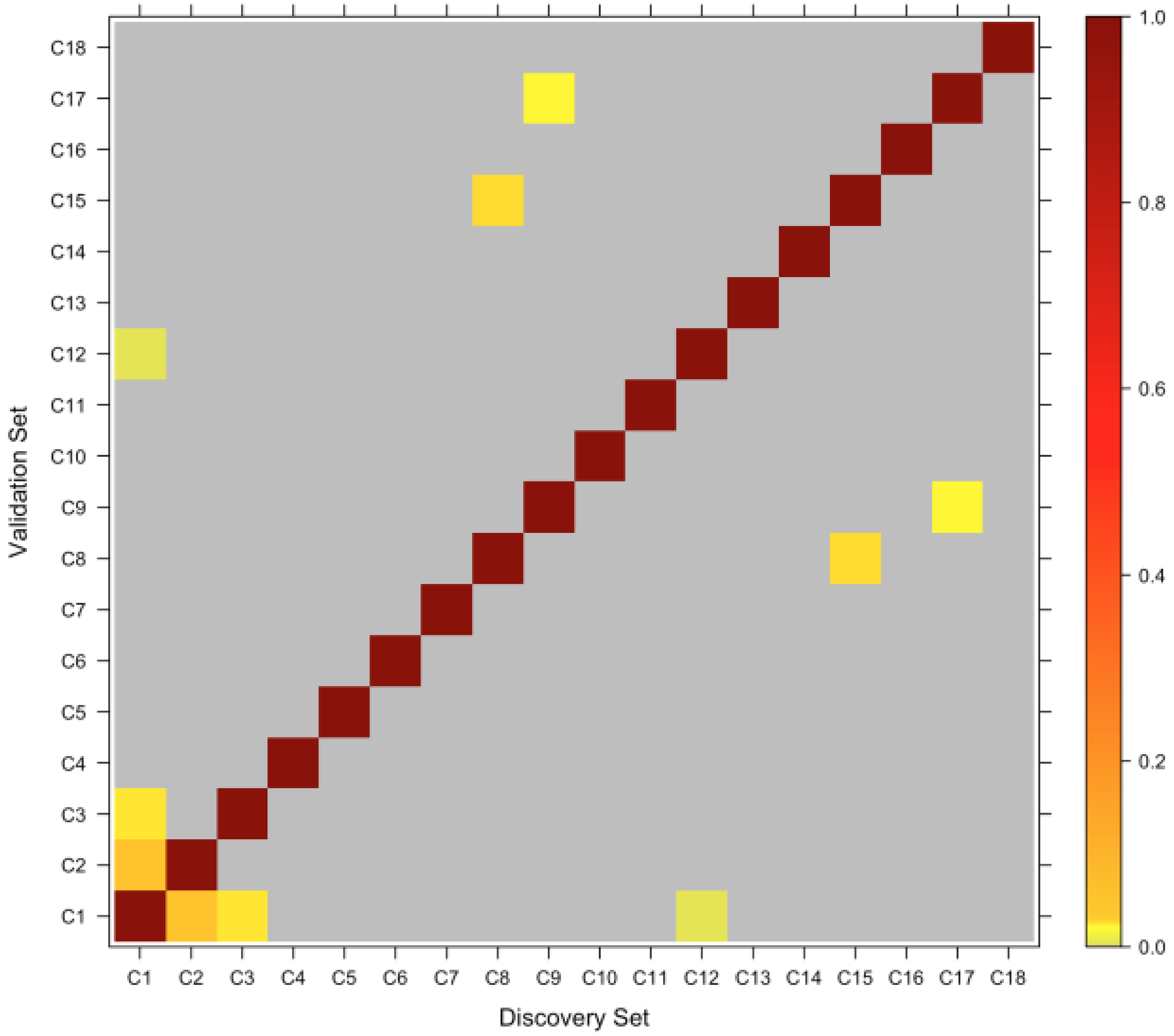


Figure 6. Heatmap of overlapping score for the clusters in Discovery and Validation sets identified by ClusterONE

The overlapping score is defined as the number of overlapped genes divided by the minimum number of genes in either the clusters found in Discovery or the Validation set by ClusterONE. The score is represented with the bar in the right, dark red being the highest overlapping score (1.0) and light yellow being the least. The X-axis in the figure represents the clusters (C1-C18) in Discovery set and the Y-axis represents the clusters (C1-C18) in the Validation set (See **Table 6**).

4.5. Comparison of subnetworks identified by Hotnet2 and clusters identified by ClusterONE

We compared the significantly mutated subnetworks (shown in bold in **Table 5**) identified by Hotnet2 with the significantly mutated clusters (shown in bold in **Table 6**) obtained by ClusterONE (**Figure 7**). Three significantly mutated subnetworks S2, S3 and S4 have overlapping genes with significantly mutated clusters C9,C5 and C16, respectively, but there is only one significantly mutated subnetwork (subnetwork S4 in **Table 5** and cluster C16 in **Table 6**), which has an overlapping rate larger or equal to 50%. All the 4 genes in the subnetwork S4 are in the cluster C16 with 8 genes. The gene cluster C16 identified by ClusterONE includes 8 genes *RDH5, RDH8, RDH10, RDH11, RDH12, RDH13, RDH14, SDR16C5* in both Discovery and Validation sets. The subnetwork S4 identified by Hotnet2 includes 4 genes *RDH8, RDH11, RDH12* and *RDH14* in Discovery set and 7 genes *RDH5, RDH8, RDH11, RDH12, RDH13, RDH14* and RLBP*1* in Validation set. As shown in **Figure 8**, the two genes *RDH11* and *RDH12* are highly mutated in Discovery set (**Figure 8A**) and the six genes *RDH8, RDH11, RDH12, RDH13, RDH14* and RLBP*1* are highly mutated in Validation set (**Figure 8B**). The mutation distribution heatmap for the genes that were identified are shown in **Figure 9** (for Discovery set) and **Figure 10** (for Validation set).

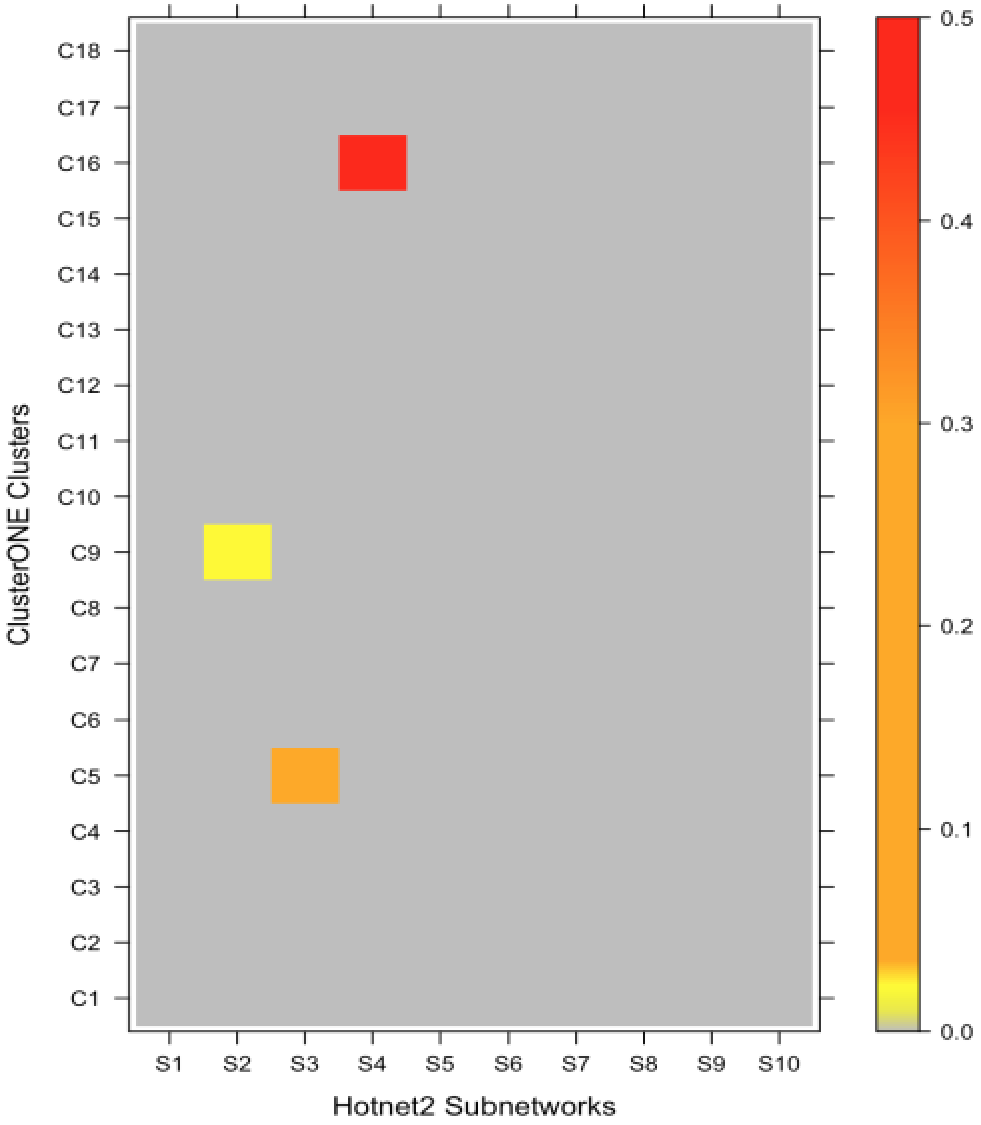
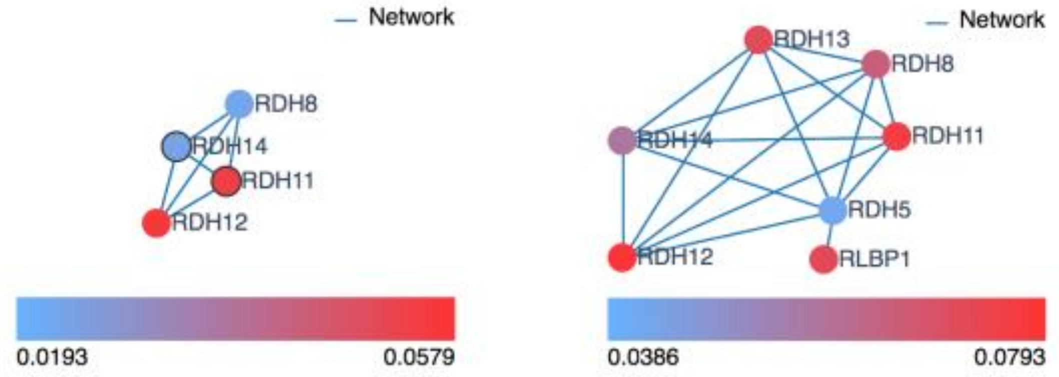


Figure 7. Heatmap of overlapping score for each pair of clusters identified by ClusterONE and subnetworks identified by Hotnet2

The overlapping score is defined as the number of overlapped genes divided by the maximum number of genes in either the Hotnet2 subnetwork or the ClusterONE cluster). The score is represented with the bar in the right, red being the highest overlapping score (1.0) and space grey being the least.



A. Subnetwork in Discovery set

B. Subnetwork in Validation set

Figure 8. Subnetworks identified by Hotnet2

The subnetworks (S4 in **Table 5**) were identified by Hotnet2 in Discovery and Validation sets, respectively. For the identified subnetworks, the heat (mutation level) is shown in the bottom, blue being the least mutated genes and red being the highest mutated gene in the network.

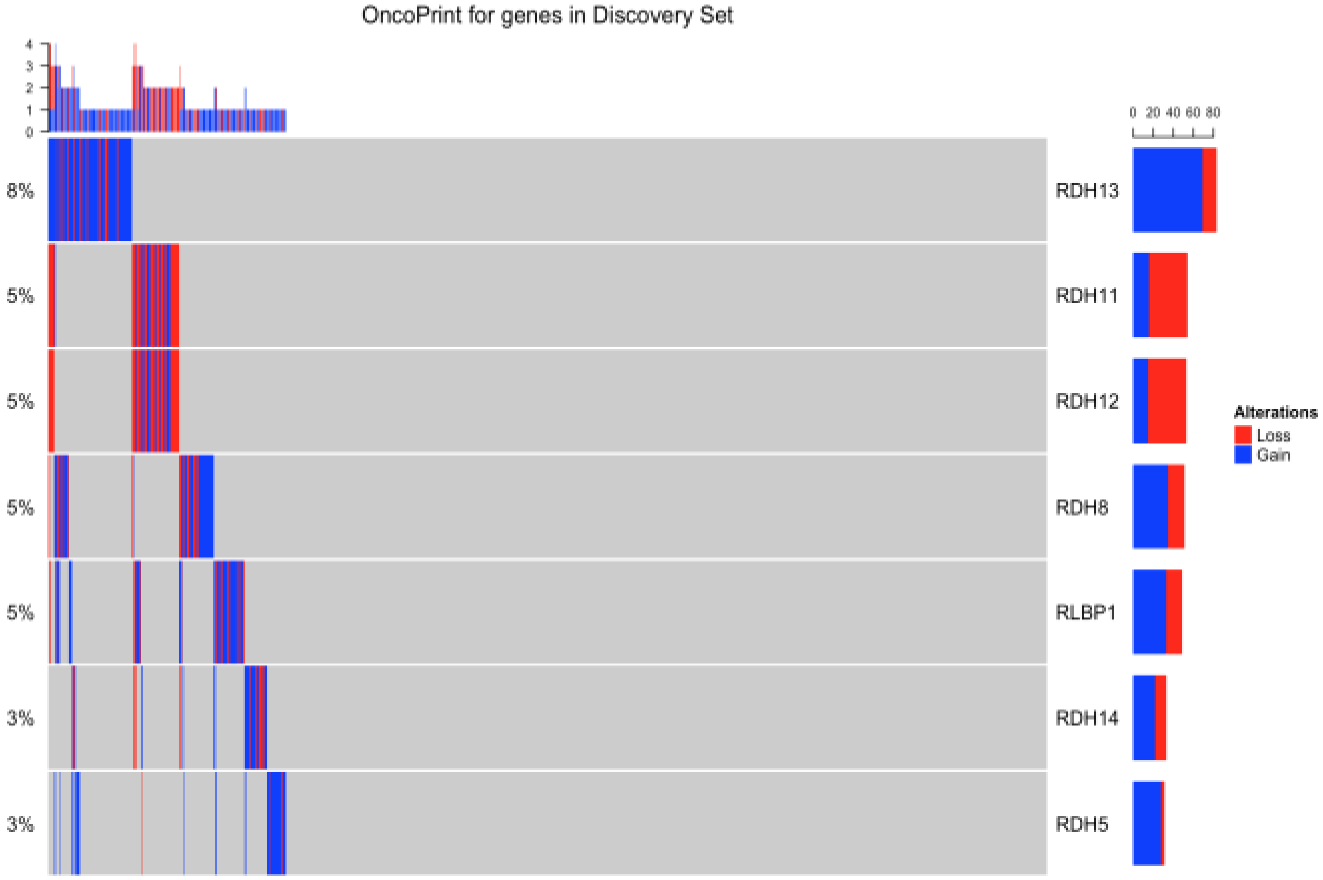


Figure 9. Mutation heatmap for significantly mutated subnetwork S4 – Discovery set

Mutation distribution heatmap for the overlapped genes identified from ClusterONE clusters and Hotnet2 subnetworks for Discovery set. X-axis is sorted based on the mutation frequency of the alterations in all samples and Y-axis is sorted to visualize the mutation frequencies across genes. Bar plots at both sides of the heatmap show numbers of different alterations for each sample and for each gene. Red represents CNA loss mutations and blue represents CNA gain mutations.

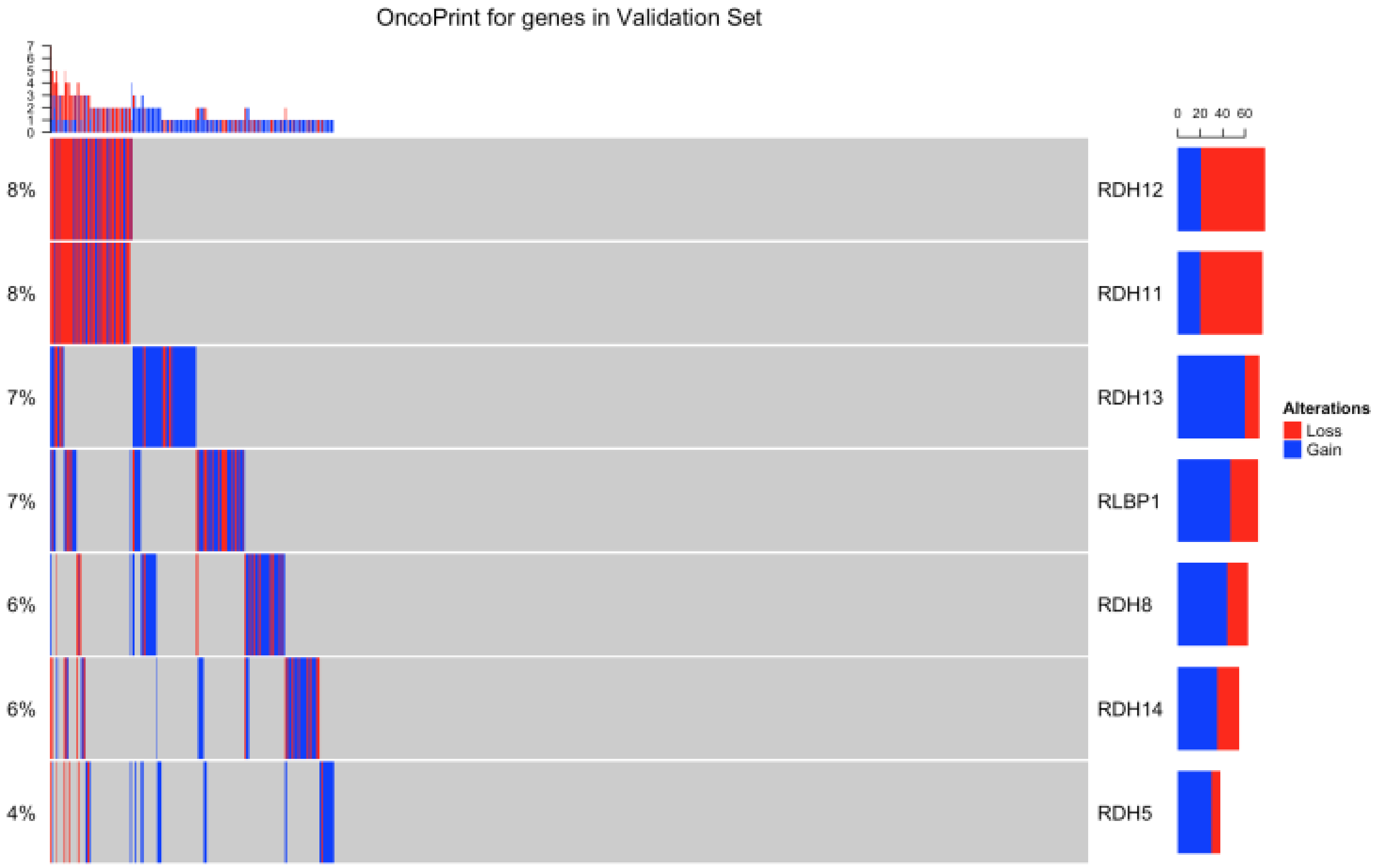


Figure 10. Mutation heatmap for significantly mutated subnetwork S4 –Validation set

Mutation distribution heatmap for the overlapped genes identified from ClusterONE clusters and Hotnet2 subnetworks for Validation set. Rows are sorted based on the frequency of the alterations in all young-specific samples and columns are sorted to visualize the mutual exclusivity across genes. Bar plots at both sides of the heatmap show numbers of different alterations for each sample and for each gene. Red represents CNA loss mutations and blue represents CNA gain mutations.

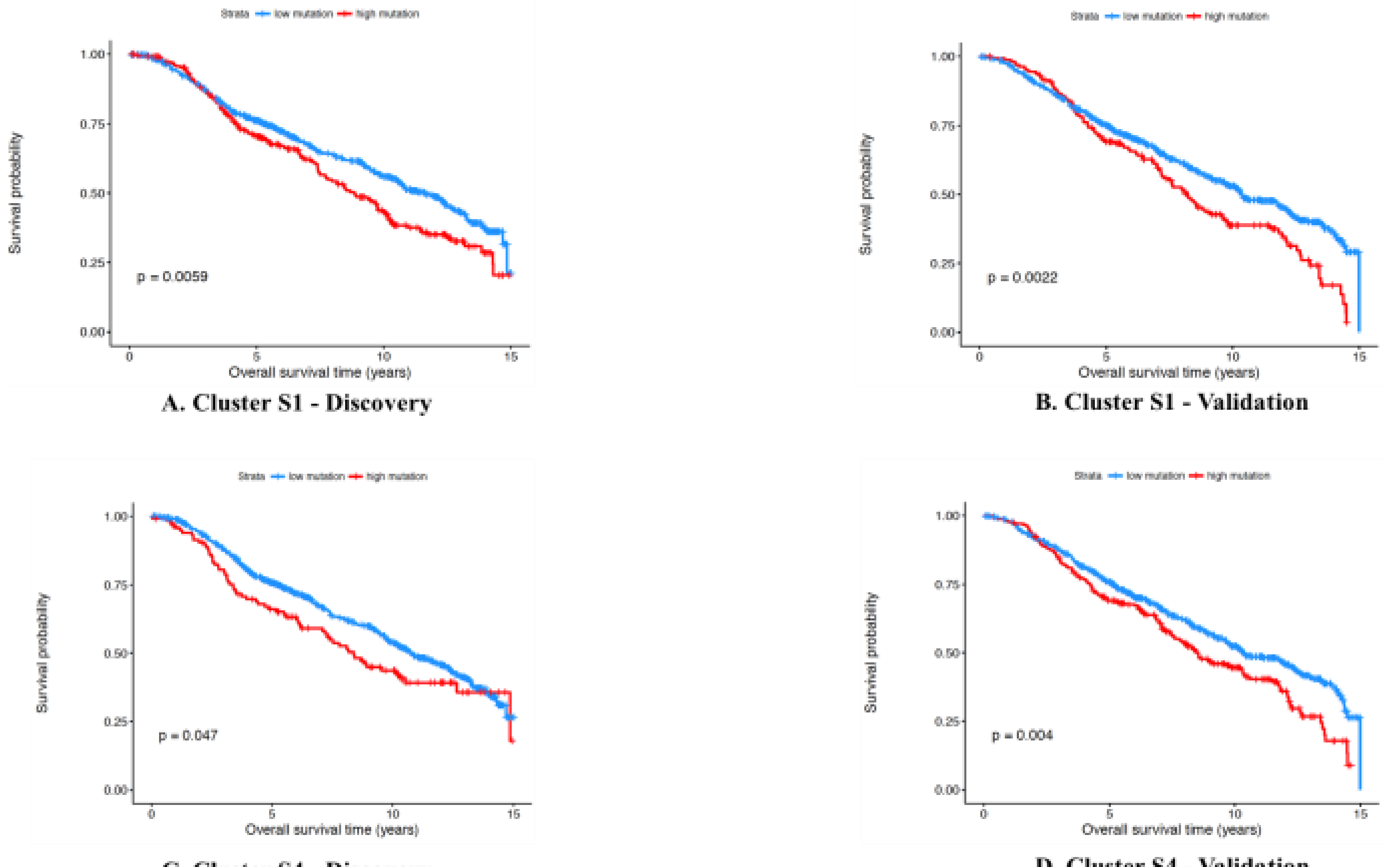
4.6. Survival Analysis

We performed Kaplan-Meier survival analysis using both the unweighted and weighted approach discussed in section 3.2.7. The first step was to calculate the unweighted and weighted score for the genes that were validated from the results of both ClusterONE and Hotnet2. We generated the unweighted and weighted matrices which contained the genes that were identified in the clusters/subnetworks. These unweighted and weighted matrices are essentially submatrices of the sample-by-gene matrices we generated previously. The unweighted and weighted score was then calculated to be used as a predictor in the survival analysis.

4.6.1. Unweighted approach

4.6.1.1. Overall survival:

The unweighted survival analysis performed on all the samples showed that the genes that are identified in high mutation group in ClusterONE yields to significantly lower survival probability in the span of 15 years. Figure 11-15A and 11-15B demonstrates that most of the survival plots of the overall samples has a significant p-value (p<0.01). Same trend is noticed on the genes that were in high mutation group based on the results of Hotnet2 (Figure 11-15C and 11-15D).



A. Cluster S1 - Discovery

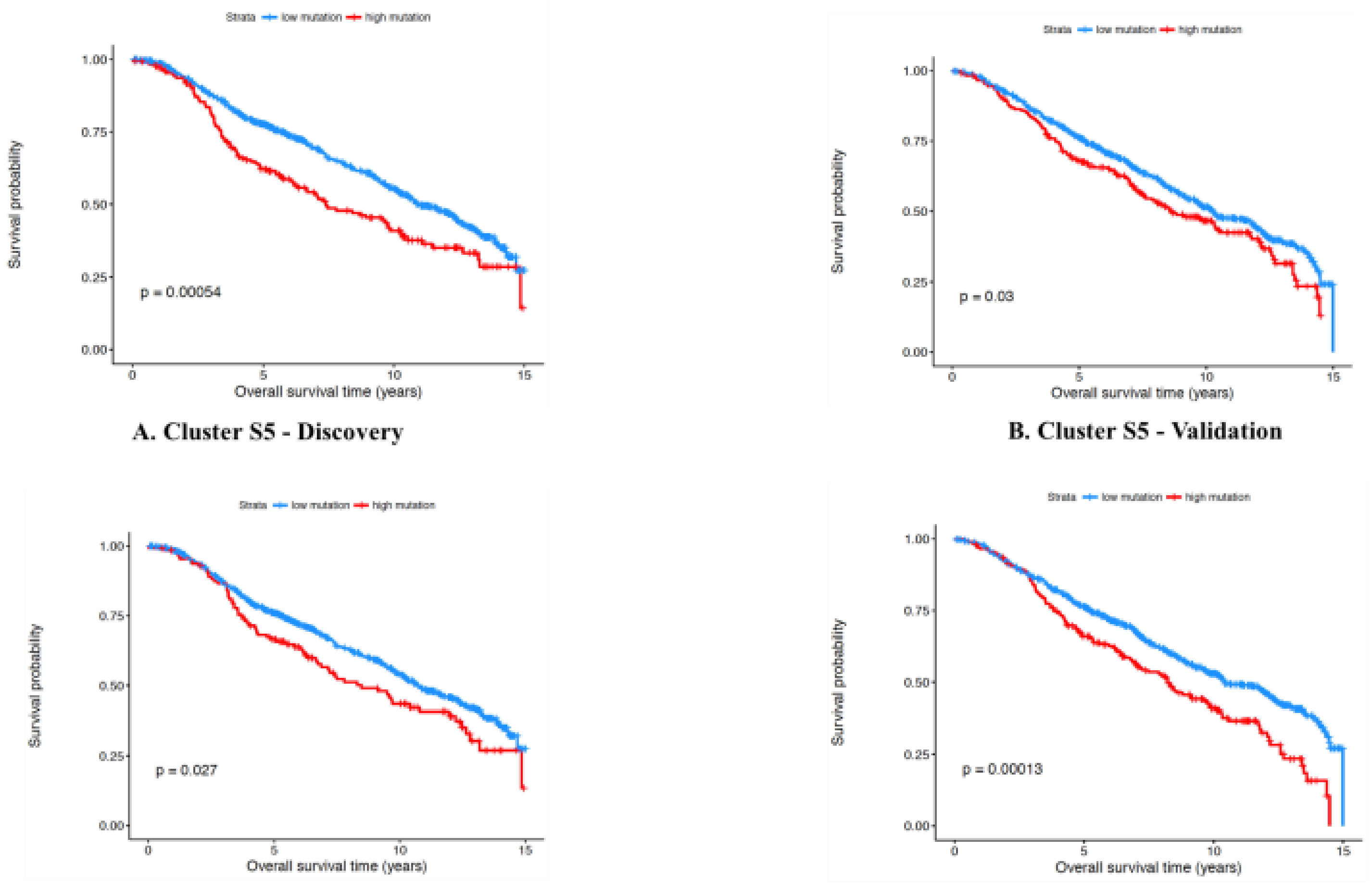
B. Cluster S1 - Validation

C. Cluster S4 - Discovery

D. Cluster S4 - Validation

Figure 11. Kaplan-Meier survival plots for subnetworks S1 and S4 (Unweighted approach)

Survival plots based on the unweighted approach based on the results of Hotnet2. Patients were stratified into those with high mutation scores and those with low mutation scores in the gene cluster. The blue curve shows the survival probability for the patients with low mutation scores and the red curve shows the survival probability for the patients with high mutation scores. The corresponding p-value is also shown in each plot. Figure 11A, 11B, 11C and 11D shows the survival plots for the subnetworks S1 Discovery, S1 Validation, S4 Discovery and S4 Validation (**bold clusters in Table 5**) respectively.



A. Cluster S5 - Discovery

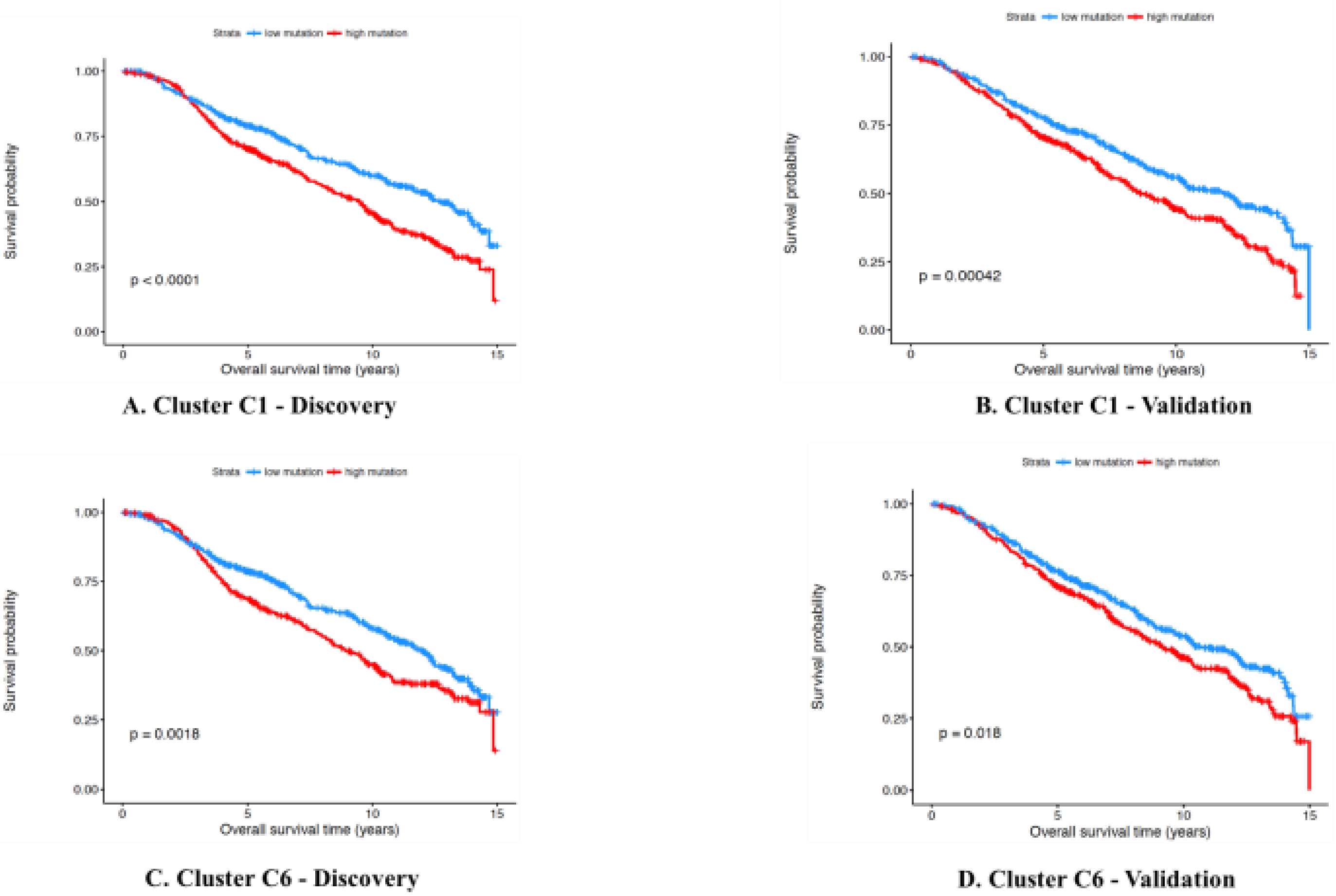
B. Cluster S5 - Validation

C. Cluster S7 - Discovery

D. Cluster S7 - Validation

Figure 12. Kaplan-Meier survival plots for subnetworks S5 and S7 (Unweighted approach)

Survival plots based on the unweighted approach based on the results of Hotnet2. Patients were stratified into those with high mutation scores and those with low mutation scores in the gene cluster. The blue curve shows the survival probability for the patients with low mutation scores and the red curve shows the survival probability for the patients with high mutation scores. The corresponding p-value is also shown in each plot. Figure 12A, 12B, 12C and 12D shows the survival plots for the subnetworks S5 Discovery, S5 Validation, S7 Discovery and S7 Validation (**bold clusters in Table 5**) respectively.



A. Cluster C1 - Discovery

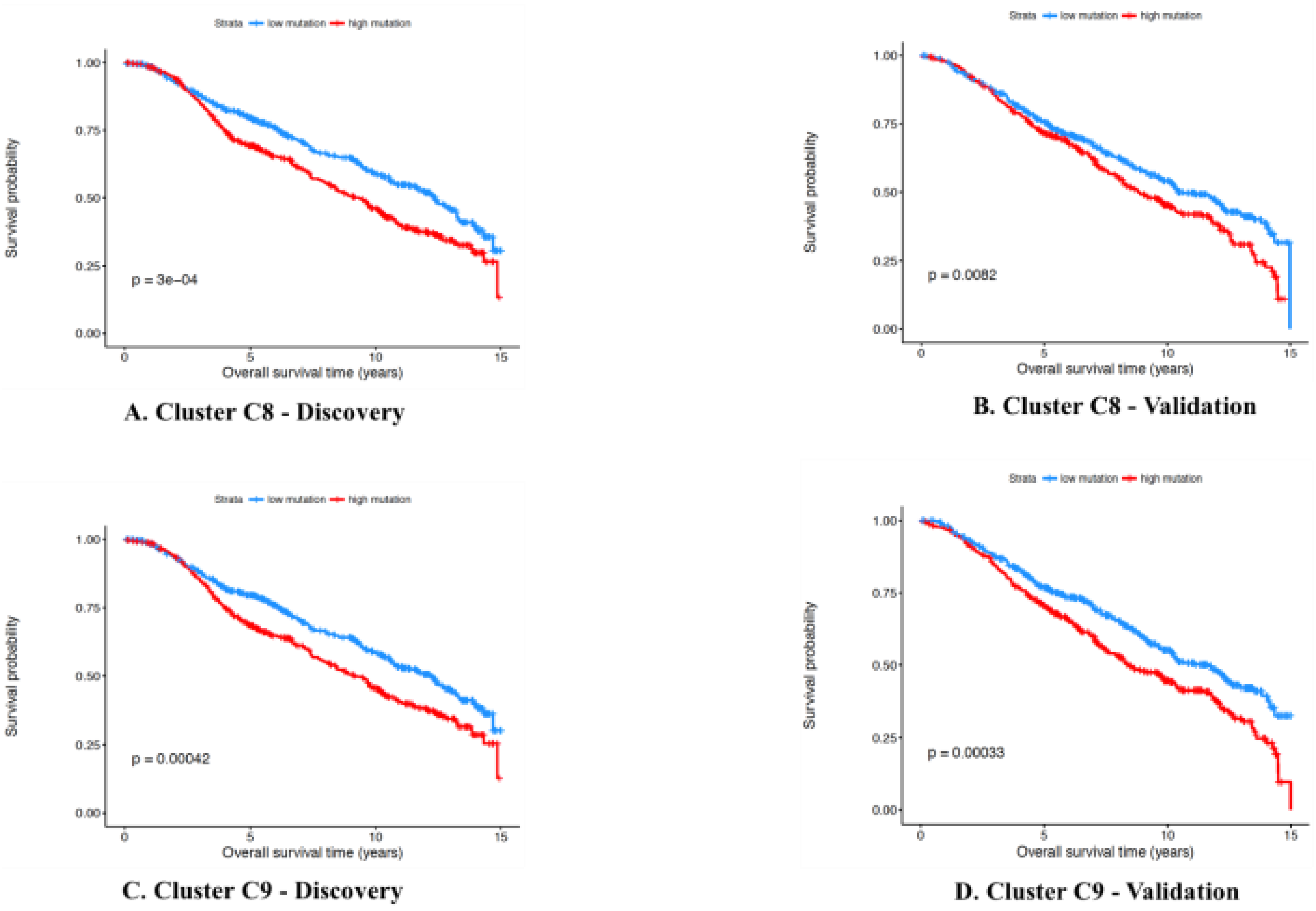
B. Cluster C1 - Validation

C. Cluster C6 - Discovery

D. Cluster C6 - Validation

Figure 13. Kaplan-Meier survival plots for clusters C1 and C6 (Unweighted approach)

Survival plots based on the unweighted approach based on the results of ClusterONE. Patients were stratified into those with high mutation scores and those with low mutation scores in the gene cluster. The blue curve shows the survival probability for the patients with low mutation scores and the red curve shows the survival probability for the patients with high mutation scores. The corresponding p-value is also shown in each plot. Figure 13A, 13B, 13C and 13D shows the survival plots for the clusters C1 Discovery, C1 Validation, C6 Discovery and C6 Validation (**bold clusters in Table 6**) respectively.



A. Cluster C8 - Discovery

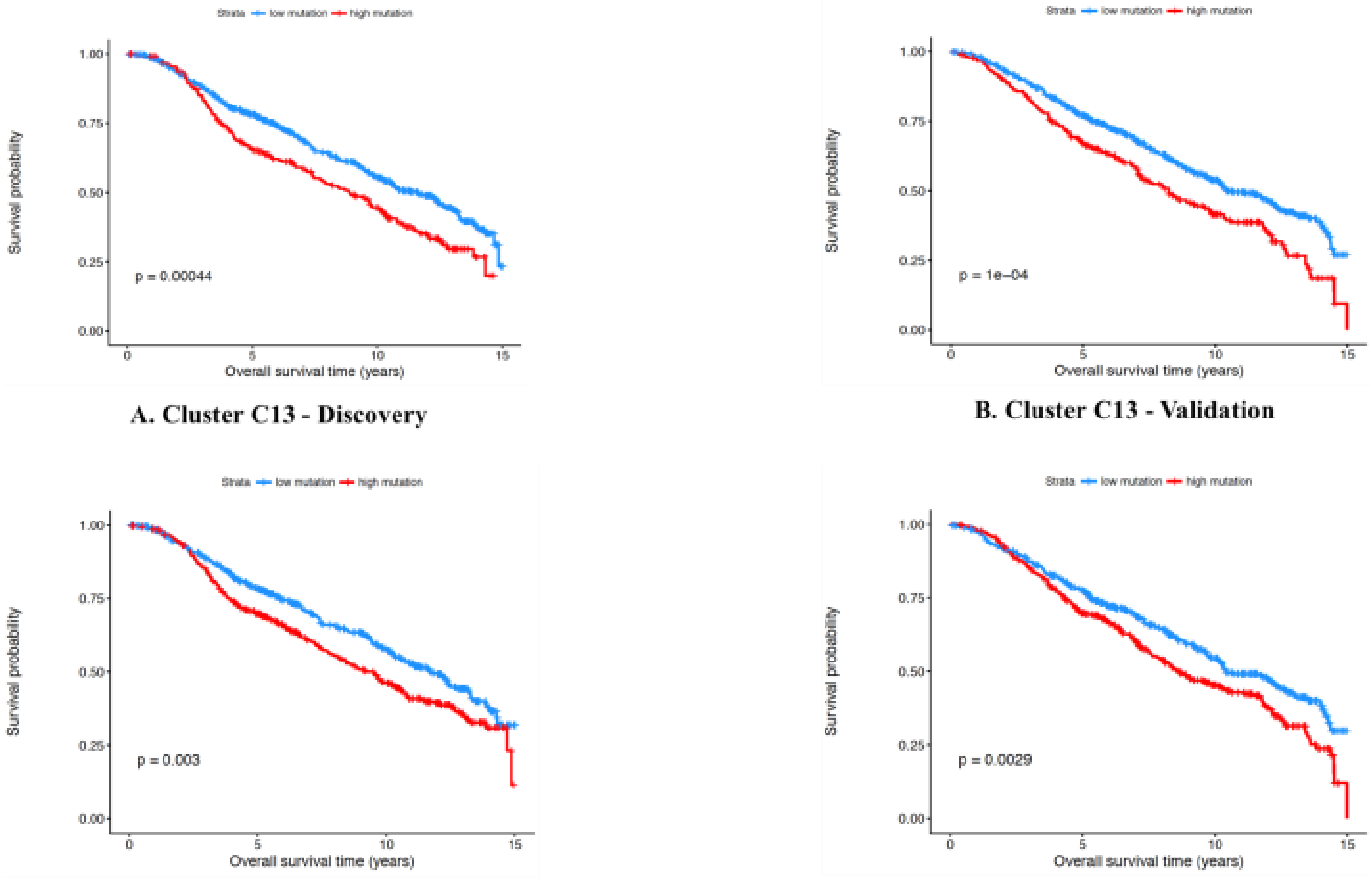
B. Cluster C8 - Validation

C. Cluster C9 - Discovery

D. Cluster C9 - Validation

Figure 14. Kaplan-Meier survival plots for clusters C8 and C9 (Unweighted approach)

Survival plots based on the unweighted approach based on the results of ClusterONE. Patients were stratified into those with high mutation scores and those with low mutation scores in the gene cluster. The blue curve shows the survival probability for the patients with low mutation scores and the red curve shows the survival probability for the patients with high mutation scores. The corresponding p-value is also shown in each plot. Figure 14A, 14B, 14C and 14D shows the survival plots for the clusters C8 Discovery, C8 Validation, C9 Discovery and C9 Validation (**bold clusters in Table 6**) respectively.



A. Cluster C13 - Discovery

B. Cluster C13 - Validation

C. Cluster C16 - Discovery

D. Cluster C16 - Validation

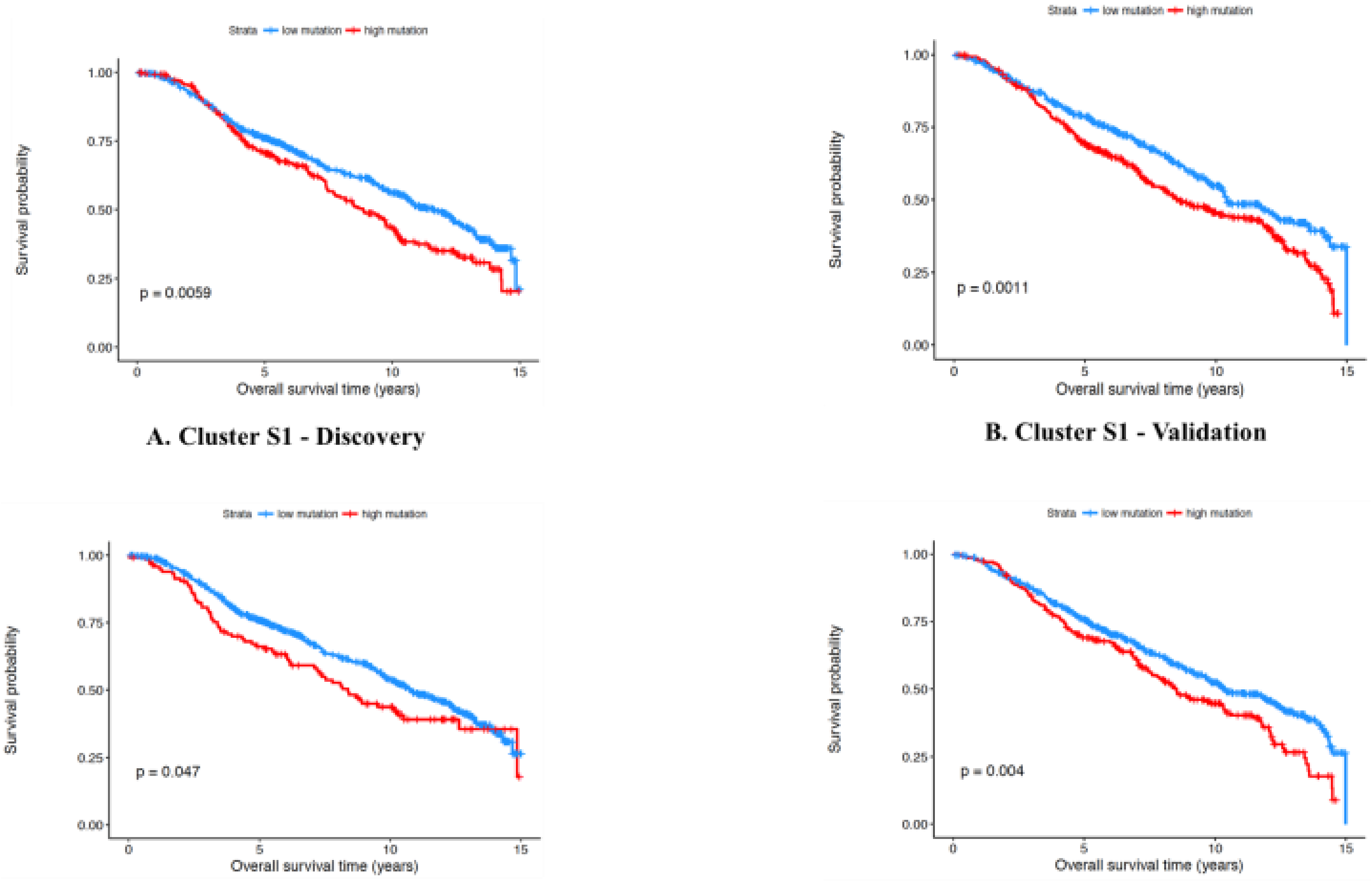
Figure 15. Kaplan-Meier survival plots for clusters C13 and C16 (Unweighted approach)

Survival plots based on the unweighted approach based on the results of ClusterONE. Patients were stratified into those with high mutation scores and those with low mutation scores in the gene cluster. The blue curve shows the survival probability for the patients with low mutation scores and the red curve shows the survival probability for the patients with high mutation scores. The corresponding p-value is also shown in each plot. Figure 15A, 15B, 15C and 15D shows the survival plots for the clusters C13 Discovery, C13 Validation, C16 Discovery and C16 Validation (**bold clusters in Table 6**) respectively.

4.6.2. Weighted approach

4.6.2.1. Overall survival:

For the weighted approach, we considered the gene-specific mutation frequency for the genes identified previously. This was done to get a weighted score for each gene that were identified by both ClusterONE and Hotnet2. This mutation frequency is then multiplied with the mutation status (either mutated: 1/-1 or not: 0). This weighted approach ensures that the score for each gene are distributed based on their mutation frequency over all the samples. Figure 16-20A and 16-20B demonstrates that most of the survival plots of the overall samples has a significant p-value (p<0.01). Same trend is noticed on the genes that were in high mutation group based on the results of Hotnet2 (Figure 16-20C and 16-20D).



A. Cluster S1 - Discovery

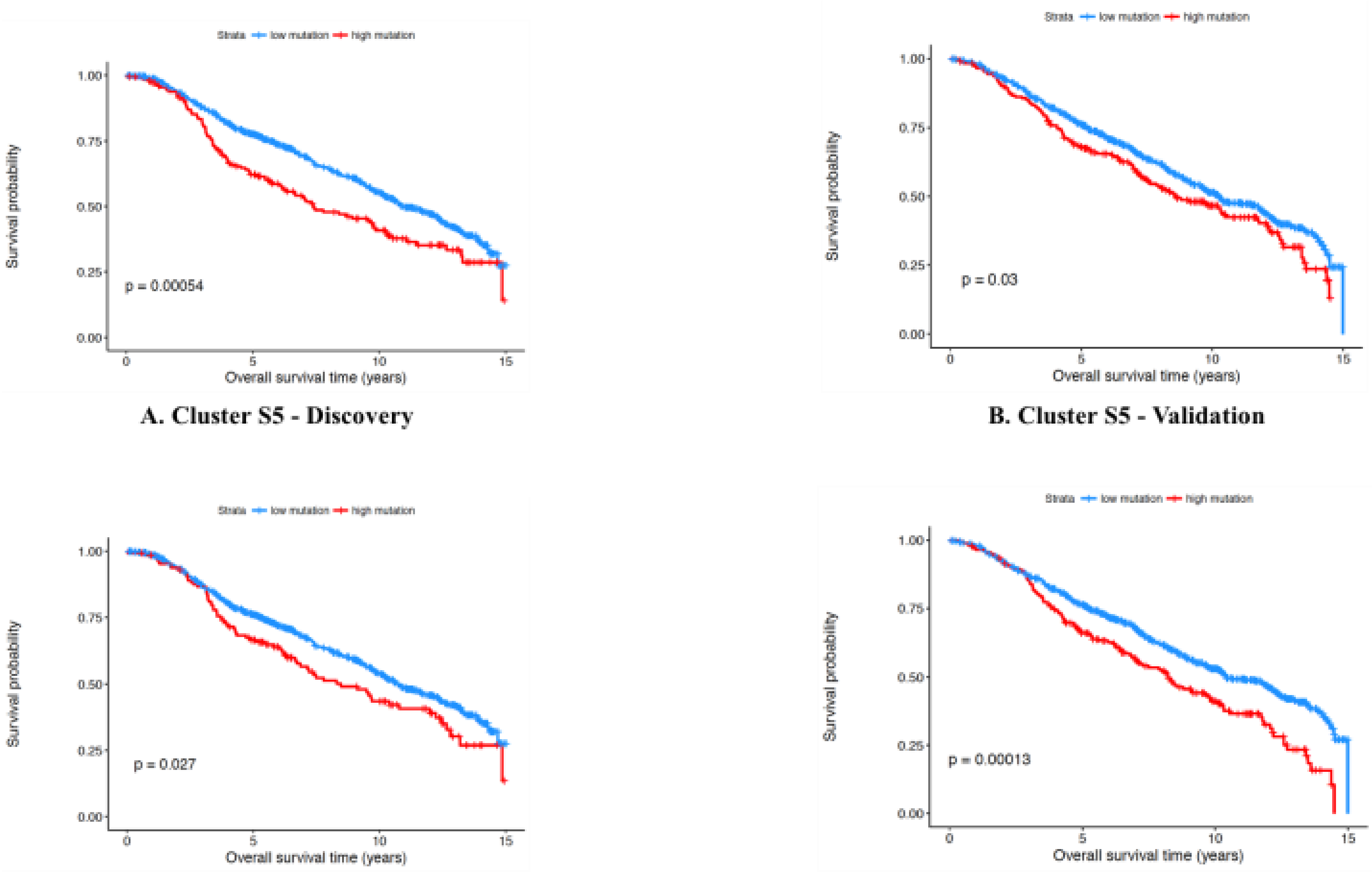
B. Cluster S1 - Validation

C. Cluster S4 - Discovery

D. Cluster S4 - Validation

Figure 16. Kaplan-Meier survival plots for subnetworks S1 and S6 (Weighted approach)

Survival plots based on the weighted approach based on the results of Hotnet2. Patients were stratified into those with high mutation scores and those with low mutation scores in the gene cluster. The blue curve shows the survival probability for the patients with low mutation scores and the red curve shows the survival probability for the patients with high mutation scores. The corresponding p-value is also shown in each plot. Figure 16A, 16B, 16C and 16D shows the survival plots for the subnetworks S1 Discovery, S1 Validation, S4 Discovery and S4 Validation (**bold clusters in Table 5**) respectively.



A. Cluster S5 - Discovery

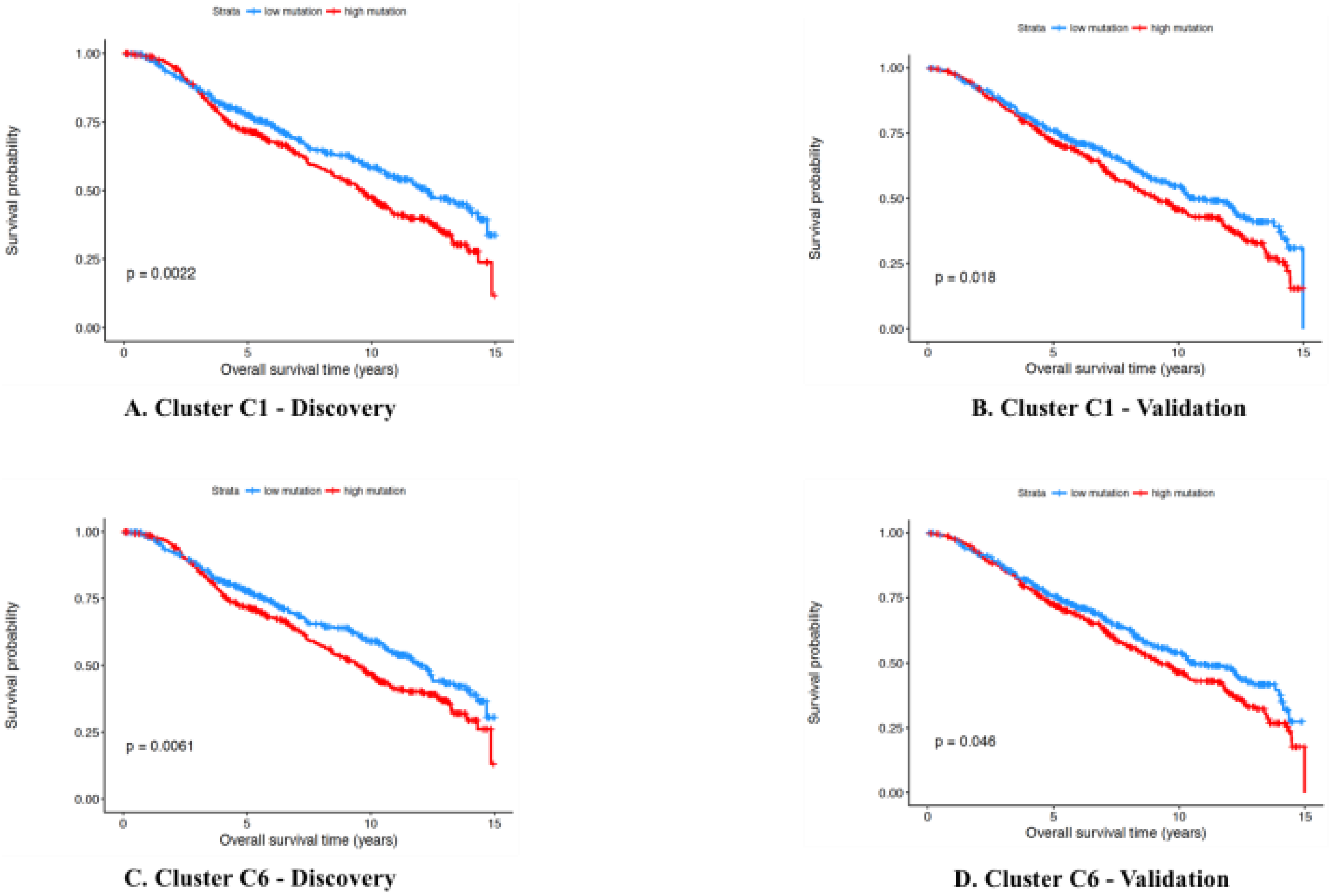
B. Cluster S5 - Validation

C. Cluster S7 - Discovery

D. Cluster S7 - Validation

Figure 17. Kaplan-Meier survival plots for subnetworks S5 and S7 (Weighted approach)

Survival plots based on the weighted approach based on the results of Hotnet2. Patients were stratified into those with high mutation scores and those with low mutation scores in the gene cluster. The blue curve shows the survival probability for the patients with low mutation scores and the red curve shows the survival probability for the patients with high mutation scores. The corresponding p-value is also shown in each plot. Figure 17A, 17B, 17C and 17D shows the survival plots for the subnetworks S5 Discovery, S5 Validation, S7 Discovery and S7 Validation (**bold clusters in Table 5**) respectively.



A. Cluster C1 - Discovery

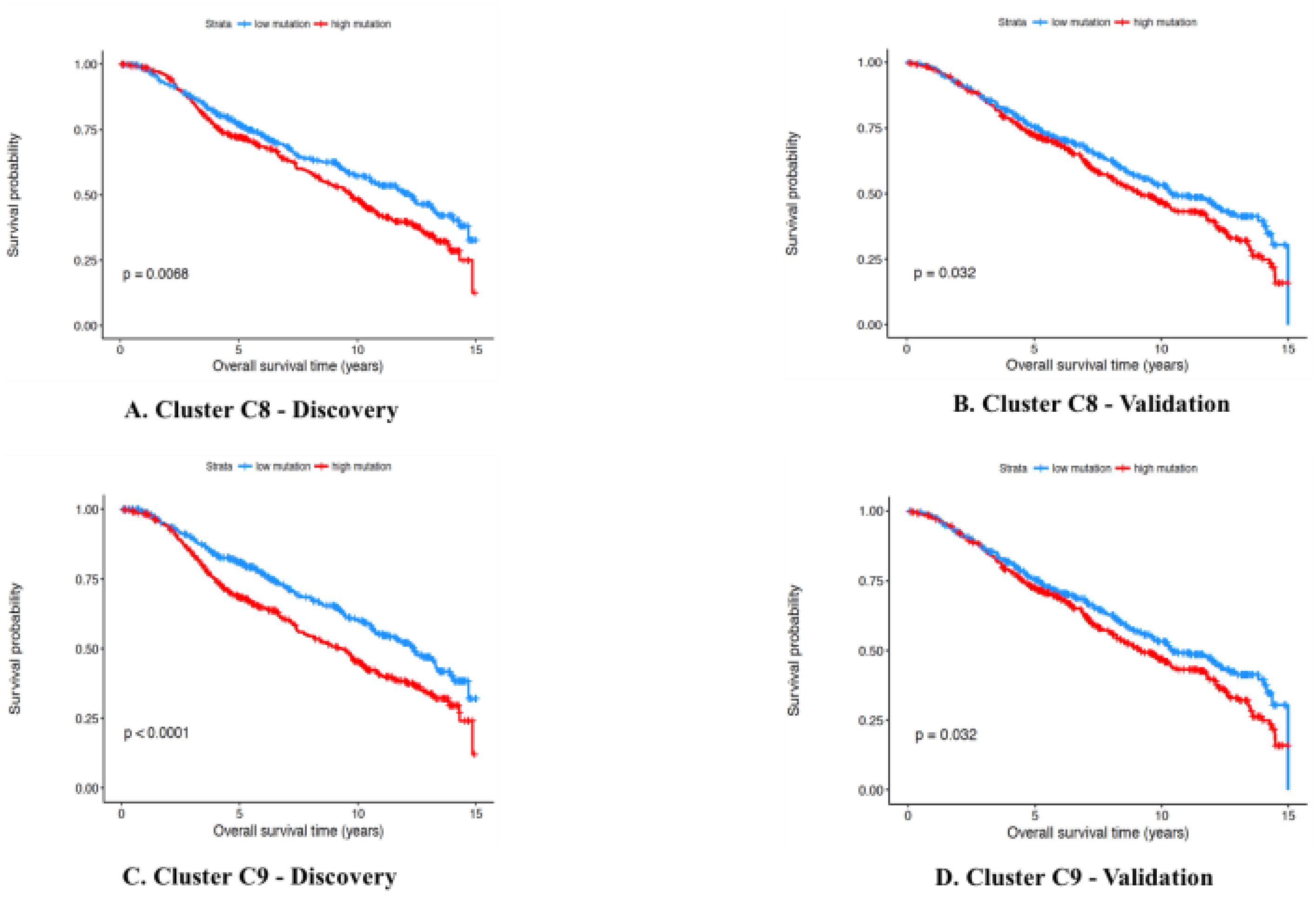
B. Cluster C1 - Validation

C. Cluster C6 - Discovery

D. Cluster C6 - Validation

Figure 18. Kaplan-Meier survival plots for clusters C1 and C6 (Weighted approach)

Survival plots based on the weighted approach based on the results of ClusterONE. Patients were stratified into those with high mutation scores and those with low mutation scores in the gene cluster. The blue curve shows the survival probability for the patients with low mutation scores and the red curve shows the survival probability for the patients with high mutation scores. The corresponding p-value is also shown in each plot. Figure 18A, 18B, 18C and 18D shows the survival plots for the clusters C1 Discovery, C1 Validation, C6 Discovery and C6 Validation (**bold clusters in Table 6**) respectively.



A. Cluster C8 - Discovery

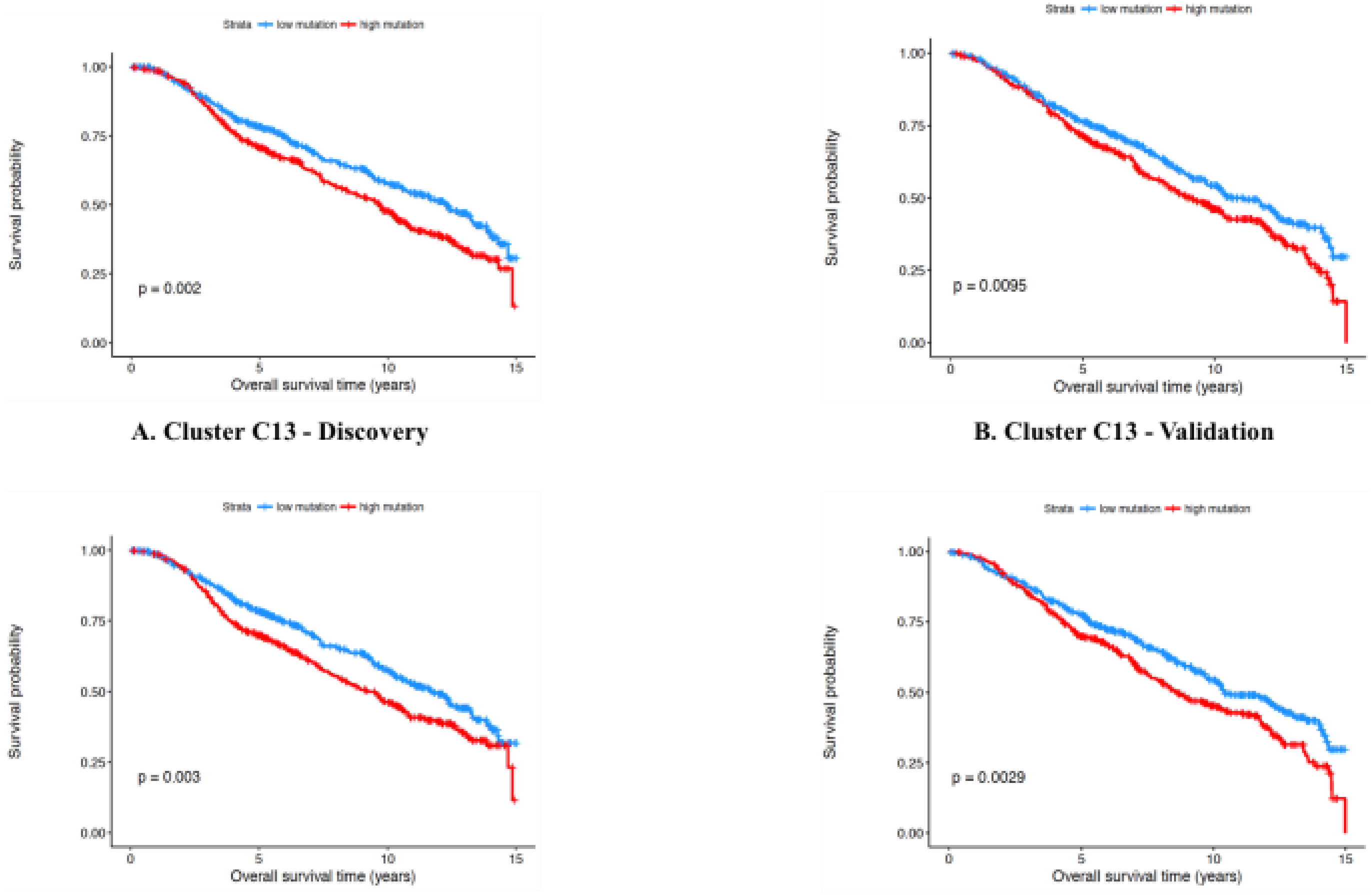
B. Cluster C8 - Validation

C. Cluster C9 - Discovery

D. Cluster C9 - Validation

Figure 19. Kaplan-Meier survival plots for clusters C8 and C9 (Weighted approach)

Survival plots based on the unweighted approach based on the results of ClusterONE. Patients were stratified into those with high mutation scores and those with low mutation scores in the gene cluster. The blue curve shows the survival probability for the patients with low mutation scores and the red curve shows the survival probability for the patients with high mutation scores. The corresponding p-value is also shown in each plot. Figure 19A, 19B, 19C and 19D shows the survival plots for the clusters C8 Discovery, C8 Validation, C9 Discovery and C9 Validation (**bold clusters in Table 6**) respectively.



A. Cluster C13 - Discovery

B. Cluster C13 - Validation

C. Cluster C16 - Discovery

D. Cluster C16 - Validation

Figure 20. Kaplan-Meier survival plots for clusters C13 and C16 (Weighted approach)

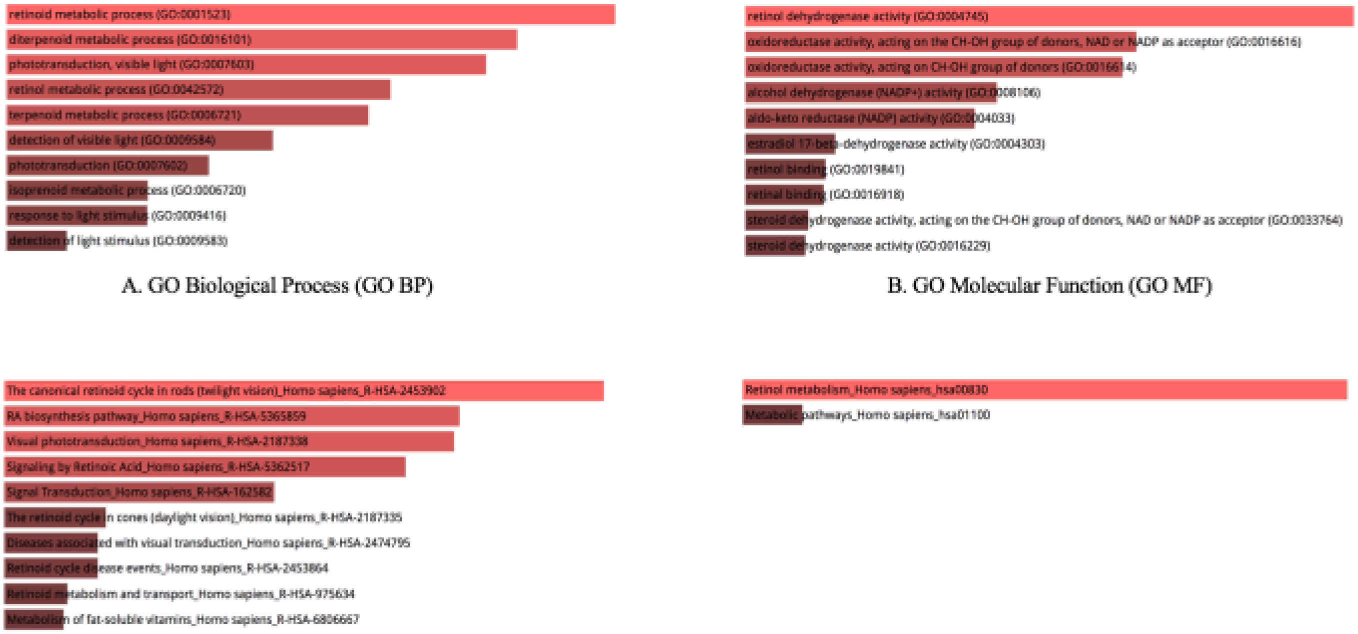
Survival plots based on the weighted approach based on the results of ClusterONE. Patients were stratified into those with high mutation scores and those with low mutation scores in the gene cluster. The blue curve shows the survival probability for the patients with low mutation scores and the red curve shows the survival probability for the patients with high mutation scores. The corresponding p-value is also shown in each plot. Figure 20A, 20B, 20C and 20D shows the survival plots for the clusters C13 Discovery, C13 Validation, C16 Discovery and C16 Validation (**bold clusters in Table 6**) respectively.

4.6.3. Survival analysis for overlapped genes from Hotnet2 and ClusterONE

We also performed survival analysis on the 7 overlapped genes (**Figure 9**) found from the identified subnetwork S4 (**Table 5**) using Hotnet2 and cluster C16 (**Table 6**) using ClusterONE. The analysis showed that the mutation pattern of the genes in the significantly mutated subnetwork is significantly associated with breast cancer survival in both Discovery set and Validation set (*P*-value<0.01). The same trend was also observed in the cluster identified by ClusterONE in Discovery and Validation sets (KM plots are not shown).

4.7. Enrichment analysis

We further performed enrichment analysis of the seven genes *RDH5, RDH8, RDH10, RDH11, RDH12, RDH13, RDH14* (**Figure 9** and **Figure 10**) via the Enrichr software. Our analysis revealed that the genes are significantly over-represented in retinoid metabolic biological process, retinol dehydrogenase activity molecular function and retinol metabolism KEGG pathway (*P*-value <1E-06). The results are shown in **Figure 21**.



A. GO Biological Process (GO BP)

B. GO Molecular Function (GO MF)

C. REACTOME

D. Kyoto Encyclopedia of Genes and Genomes (KEGG)

Figure 21. Enrichment analysis of the genes in significantly mutated subnetwork S4

Graph bars are sorted by p-value ranking. The length of the bar represents the significance of that specific gene set. Light red colored bars have a p-value < 1E-06.

Many studies have found that retinoid receptors modulate various effects of retinoids, including estrogen metabolism in human breast carcinomas (Ginestier et al., 2009; Suzuki et al., 2001). Interestingly, retinoids (such as vitamin A and its natural and synthetic analogs) have been used as potential chemotherapeutic or chemopreventive agents because of their differentiative, anti-proliferative, pro-apoptotic properties. Our results show that breast cancer treatment with reninoids may be a potential personalized therapy for breast cancer patients since the CNV patterns of the breast cancer patients can imply whether the retinoids pathway is altered.

Chapter 5

Significance and Conclusions

Since genes usually interact with other genes to execute their functions, gene networks can be modular and divided into subnetworks. It is reasonable to assume that clinically relevant mutations in breast cancer occur in closely interacting genes and breast cancer is an outcome of coordinated dysfunction of these closely connected subnetworks enriched with clinically informative cancer mutations.

We designed a novel analysis pipeline to identify significantly mutated gene subnetworks using breast cancer copy number variations. To decrease the potential false positives in identifying the mutated subnetworks, we applied two graph-based clustering algorithms to analyze the data. We identified a significantly mutated yet clinically and functionally relevant subnetwork. The mutational pattern of the subnetwork is significantly associated with breast cancer survival. The genes in the subnetwork are significantly enriched in retinol metabolism KEGG pathway.

The proposed pipeline has some unique features that address the limitations in the current methods. One obvious limitation of the current methods is the failure to identify overlapped mutated subnetworks. Our method addresses this problem. This makes more biological sense because a gene can be assigned to multiple subnetworks and genes are usually involved in multiple complexes and pathways. Furthermore, the proposed method takes a gene's mutation similarity to infer mutated subnetworks and the network-specific mutation score is used to predict cancer survival, which provides an innovative way to identify breast cancer prognostic biomarkers.

Our analysis has some limitations. For example, we only considered the CNV mutations due to a lack of single nucleotide variants (SNVs) in METABRIC study. In the future, we will apply the same pipeline to analyze the breast cancer CNVs and SNVs together in TCGA study (Koboldt et al., 2012). We expect to replicate the findings from our study using the TCGA data.

References

Alluri, P., & Newman, L. A. (2014). Basal-like and triple-negative breast cancers: searching for positives among many negatives. *Surgical Oncology Clinics of North America*, *23*(3), 567–77. http://doi.org/10.1016/j.soc.2014.03.003

Apostolou, P., & Fostira, F. (2013). Hereditary breast cancer: the era of new susceptibility genes. *BioMed Research International*, *2013*, 747318. http://doi.org/10.1155/2013/747318

Bell, D., Berchuck, A., Birrer, M., Chien, J., Cramer, D. W., Dao, F., … Thomson, E. (2011). Integrated genomic analyses of ovarian carcinoma. *Nature, 474*(7353), 609–615. http://doi.org/10.1038/nature10166

Bell, J. I. (2002). Single nucleotide polymorphisms and disease gene mapping. *Arthritis Research, 4 Suppl 3*(Suppl 3), S273-8. http://doi.org/10.1186/ar555

Bellmunt, J., Werner, L., Leow, J. J., Mullane, S. A., Fay, A. P., Riester, M., … Rosenberg, J. (2015). Somatic Copy Number Abnormalities and Mutations in PI3K/AKT/mTOR Pathway Have Prognostic Significance for Overall Survival in Platinum Treated Locally Advanced or Metastatic Urothelial Tumors. *PLOS ONE, 10*(6), e0124711. http://doi.org/10.1371/journal.pone.0124711

Beroukhim, R., Getz, G., Nghiemphu, L., Barretina, J., Hsueh, T., Linhart, D., … Sellers, W. R. (2007). Assessing the significance of chromosomal aberrations in cancer: Methodology and application to glioma. *Proceedings of the National Academy of Sciences*, *104*(50), 20007–20012. http://doi.org/10.1073/pnas.0710052104

Braun, L., Mietzsch, F., Seibold, P., Schneeweiss, A., Schirmacher, P., Chang-Claude, J., … Aulmann, S. (2013). Intrinsic breast cancer subtypes defined by estrogen receptor signalling—prognostic relevance of progesterone receptor loss. *Modern Pathology, 26*(9), 1161–1171. http://doi.org/10.1038/modpathol.2013.60

Caligo, M. A., Agata, S., Aceto, G., Crucianelli, R., Manoukian, S., Peissel, B., … Montagna, M. (2004). The CHEK2 c.1100delC mutation plays an irrelevant role in breast cancer predisposition in Italy. *Human Mutation*, *24*(1), 100–101. http://doi.org/10.1002/humu.20051

Cerami, E., Demir, E., Schultz, N., Taylor, B. S., & Sander, C. (2010). Automated Network Analysis Identifies Core Pathways in Glioblastoma. *PLoS ONE*, *5*(2), e8918. http://doi.org/10.1371/journal.pone.0008918

Chi, Y.-Y., Gribbin, M. J., Johnson, J. L., & Muller, K. E. (2014). Power calculation for overall hypothesis testing with high-dimensional commensurate outcomes. *Statistics in Medicine*, *33*(5), 812–827. http://doi.org/10.1002/sim.5986

Chin, L., Hahn, W. C., Getz, G., & Meyerson, M. (2011). Making sense of cancer genomic data. *Genes & Development*, *25*(6), 534–555. http://doi.org/10.1101/gad.2017311

Collett, D. (2003). *Modelling survival data in medical research*. *Texts in statistical science* (Vol. 46). http://doi.org/10.1198/tech.2004.s817

Creixell, P., Reimand, J., Haider, S., Wu, G., Shibata, T., Vazquez, M., … Stein, L. D. (2015). Pathway and network analysis of cancer genomes. *Nature Methods*, *12*(7), 615–621. http://doi.org/10.1038/nmeth.3440

Crespi, B. J., & Crofts, H. J. (2012). Association testing of copy number variants in schizophrenia and autism spectrum disorders. *Journal of Neurodevelopmental Disorders*, *4*(1), 1. http://doi.org/10.1186/1866-1955-4-15

Curtis, C., Shah, S. P., Chin, S., Turashvili, G., Rueda, O. M., Dunning, M. J., … Aparicio, S. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, *486*(7403), 346–352. http://doi.org/10.1038/nature10983.The

Dai, X., Li, T., Bai, Z., Yang, Y., Liu, X., Zhan, J., & Shi, B. (2015). Breast cancer intrinsic subtype classification, clinical use and future trends. *American Journal of Cancer Research*, *5*(10), 2929–43. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/26693050

Dietze, E. C., Sistrunk, C., Miranda-Carboni, G., O'Regan, R., & Seewaldt, V. L. (2015). Triple-negative breast cancer in African-American women: disparities versus biology. *Nature Reviews - Cancer*, *15*(4), 248–254. http://doi.org/10.1038/nrc3896

Ding, L., Getz, G., Wheeler, D. A., Mardis, E. R., McLellan, M. D., Cibulskis, K., … Wilson, R. K. (2008). Somatic mutations affect key pathways in lung adenocarcinoma. *Nature*, *455*(7216), 1069–1075. http://doi.org/10.1038/nature07423

Falchetti, M., Lupi, R., Rizzolo, P., Ceccarelli, K., Zanna, I., Calò, V., … Ottini, L. (2008). BRCA1/BRCA2 rearrangements and CHEK2 common mutations are infrequent in Italian male breast cancer cases. *Breast Cancer Research and Treatment*, *110*(1), 161–167. http://doi.org/10.1007/s10549-007-9689-2

Futreal, P. A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., … Stratton, M. R. (2004). A census of human cancer genes. *Nature Reviews Cancer*, *4*(3), 177–183. http://doi.org/10.1038/nrc1299

Ginestier, C., Wicinski, J., Cervera, N., Monville, F., Finetti, P., Bertucci, F., … Charafe-Jauffret, E. (2009). Retinoid signaling regulates breast cancer stem cell differentiation. *Cell Cycle (Georgetown, Tex.)*, *8*(20), 3297–302. http://doi.org/10.4161/cc.8.20.9761

Guttmacher, A. E., Collins, F. S., Wooster, R., & Weber, B. L. (2003). Breast and Ovarian Cancer. *New England Journal of Medicine*, *348*(23), 2339–2347. http://doi.org/10.1056/NEJMra012284

Hahn, W. C., & Weinberg, R. A. (2002). Modelling the molecular circuitry of cancer. *Nature Reviews. Cancer*, *2*(5), 331–41. http://doi.org/10.1038/nrc795

Jiang, M., Chen, Y., & Chen, L. (2015). Link Prediction in Networks with Nodes Attributes by Similarity Propagation. Retrieved from http://arxiv.org/abs/1502.04380

Kao, S. L., Chong, S. S., & Lee, C. G. (2000). The role of single nucleotide polymorphisms (SNPs) in understanding complex disorders and pharmacogenomics. *Annals of the Academy of Medicine, Singapore*, *29*(3), 376–82. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10976393

Koboldt, D. C., Fulton, R. S., McLellan, M. D., Schmidt, H., Kalicki-Veizer, J., McMichael, J. F., … Palchik, J. D. (2012). Comprehensive molecular portraits of human breast tumours. *Nature*, *490*(7418), 61–70. http://doi.org/10.1038/nature11412

Kuleshov, M. V, Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., … Ma'ayan, A. (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res*, *44*(W1), W90–W97. http://doi.org/10.1093/nar/gkw377

Leiserson, M. D. M., Vandin, F., Wu, H.-T., Dobson, J. R., Eldridge, J. V, Thomas, J. L., … Raphael, B. J. (2014). Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nature Genetics*, *47*(2), 106–114. http://doi.org/10.1038/ng.3168

Levy-Lahad, E. (2010). Fanconi anemia and breast cancer susceptibility meet again. *Nature Genetics*, *42*(5), 368–369. http://doi.org/10.1038/ng0510-368

Lin, J., Gan, C. M., Zhang, X., Jones, S., Sjöblom, T., Wood, L. D., … Velculescu, V. E. (2007). A multidimensional analysis of genes mutated in breast and colorectal cancers. *Genome Research*, *17*(9), 1304–18. http://doi.org/10.1101/gr.6431107

Liu, Z. (2007). *Aquaculture Genome Technologies*. John Wiley & Sons. Retrieved from https://books.google.ca/books?hl=en&lr=&id=4lJIVRkfrVMC&oi=fnd&pg=PA59&dq=single+nucleotide+polymorphisms+snps&ots=aZGF6A7HOq&sig=90tw2fPI3yeOJQxBaKLBU5Mrvkg#v=onepage&q=single nucleotide polymorphisms snps&f=false

McLendon, R., Friedman, A., Bigner, D., Van Meir, E. G., Brat, D. J., M. Mastrogianakis, G., … Thomson, E. (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, *455*(7216), 1061–1068. http://doi.org/10.1038/nature07385

Menche, J., Sharma, A., Kitsak, M., Ghiassian, S. D., Vidal, M., Loscalzo, J., & Barabasi, A.-L. (2015). Uncovering disease-disease relationships through the incomplete interactome. *Science*, *347*(6224), 1257601–1257601. http://doi.org/10.1126/science.1257601

Nepusz, T., Yu, H., & Paccanaro, A. (2012). Detecting overlapping protein complexes in protein-protein interaction networks. *Nature Methods*, *9*(5), 471–472. http://doi.org/10.1038/nmeth.1938

Ohno, S. (1970). Evolution by Gene Duplication. *(1970)*. http://doi.org/10.1007/978-3-642-86659-3

Parsons, D. W., Jones, S., Zhang, X., Lin, J. C., Leary, R. J., Angenendt, P., … Kinzler, K. W. (2008). An Integrated Genomic Analysis of Human Glioblastoma Multiforme. *Science*, *321*(5897), 1807–1812. http://doi.org/10.1126/science.1164382

Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., … Hurles, M. E. (2006). Global variation in copy number in the human genome. *Nature*, *444*(7118), 444–454. http://doi.org/10.1038/nature05329

Rizzolo, P., Silvestri, V., Falchetti, M., & Ottini, L. (2011). Inherited and acquired alterations in development of breast cancer. *The Application of Clinical Genetics*, *4*, 145–58. http://doi.org/10.2147/TACG.S13226

Ruepp, A., Waegele, B., Lechner, M., Brauner, B., Dunger-Kaltenbach, I., Fobo, G., … Mewes, H.-W. (2010). CORUM: the comprehensive resource of mammalian protein complexes--2009. *Nucleic Acids Research*, *38*(Database issue), D497-501. http://doi.org/10.1093/nar/gkp914

Shastry, B. S. (2009). SNPs: impact on gene function and phenotype. *Methods in Molecular Biology (Clifton, N.J.)*, *578*, 3–22. http://doi.org/10.2217/14622416.8.8.1075

Smedley, D., Haider, S., Durinck, S., Pandini, L., Provero, P., Allen, J., … Kasprzyk, A. (2015). The BioMart community portal: an innovative alternative to large, centralized data repositories. *Nucleic Acids Research*, *43*(W1), W589-98. http://doi.org/10.1093/nar/gkv350

Stankiewicz, P., & Lupski, J. R. (2010). Structural variation in the human genome and its role in disease. *Annu Rev Med*, *61*, 437–455. http://doi.org/10.1146/annurev-med-100708-204735

Stratton, M. R., & Rahman, N. (2008). The emerging landscape of breast cancer susceptibility. *Nature Genetics*, *40*(1), 17–22. http://doi.org/10.1038/ng.2007.53

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., … Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(43), 15545–50. http://doi.org/10.1073/pnas.0506580102

Suzuki, T., Moriya, T., Sugawara, A., Ariga, N., Takabayashi, H., & Sasano, H. (2001). Retinoid Receptors in Human Breast Carcinoma: Possible Modulators of in Situ Estrogen Metabolism. *Breast Cancer Research and Treatment*, *65*(1), 31–40.

http://doi.org/10.1023/A:1006433929792

Thompson, D., Duedal, S., Kirner, J., McGuffog, L., Last, J., Reiman, A., … Easton, D. F. (2005). Cancer Risks and Mortality in Heterozygous ATM Mutation Carriers. *JNCI Journal of the National Cancer Institute*, *97*(11), 813–822. http://doi.org/10.1093/jnci/dji141

Vandin, F., Clay, P., Upfal, E., & Raphael, B. J. (2012). Discovery of mutated subnetworks associated with clinical data in cancer. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 55–66. http://doi.org/9789814366496_0006

Vandin, F., Upfal, E., & Raphael, B. J. (2010). Algorithms for detecting significantly mutated pathways in cancer. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *6044 LNBI*(3), 506–521. http://doi.org/10.1007/978-3-642-12683-3_33

Vogelstein, B., & Kinzler, K. W. (2002). *The genetic basis of human cancer*. McGraw-Hill, Medical Pub. Division. Retrieved from https://books.google.ca/books/about/The_genetic_basis_of_human_cancer.html?id=pYG09OPbXp0C&redir_esc=y

Vogelstein, B., & Kinzler, K. W. (2004). Cancer genes and the pathways they control. *Nature Medicine*, *10*(8), 789–99. http://doi.org/10.1038/nm1087

Willems, P. J. (2007). Susceptibility genes in breast cancer: more is less? *Clinical Genetics*, *72*(6), 493–496. http://doi.org/10.1111/j.1399-0004.2007.00909.x

Zhang, F., Gu, W., Hurles, M. E., & Lupski, J. R. (2009). Copy number variation in human health, disease, and evolution. *Annual Review of Genomics and Human Genetics*, *10*, 451–81. http://doi.org/10.1146/annurev.genom.9.081307.164217