Identifying Regulators from Multiple Types of Biological Data in Cancer

Brittany Baur
Marquette University

IDENTIFYING REGULATORS FROM MULTIPLE TYPES OF BIOLOGICAL DATA IN CANCER

by

Brittany Baur

A Dissertation submitted to the Faculty of the Graduate School,
Marquette University,
in Partial Fulfillment of the Requirements for
the Degree of Doctor of Philosophy

Milwaukee, Wisconsin

May 2017

ABSTRACT
IDENTIFYING REGULATORS FROM MULTIPLE TYPES OF BIOLOGICAL DATA IN CANCER

Brittany Baur

Marquette University, 2017

Cancer genomes accumulate alterations that promote cancer cell proliferation and survival. Structural, genetic and epigenetic alterations that have a selective advantage for tumorigenesis affect key regulatory genes and microRNAs that in turn regulate the expression of many target genes. The goal of this dissertation is to leverage the alteration-rich landscape of cancer genomes to detect key regulatory genes and microRNAs. To this end, we designed a feature selection algorithm to identify DNA methylation signals around a gene that would highly predict its expression. We found that genes whose expression could be predicted by DNA methylation accurately were enriched in Gene Ontology terms related to the regulation of various biological processes. This suggests that genes controlled by DNA methylation are regulatory genes. We also developed two tools that infer relationships between regulatory genes and target genes leveraging structural and epigenetic data. The first tool, ProcessDriver integrates copy number alteration and gene expression datasets to identify copy number cancer driver genes, target genes of these drivers and the disrupted biological processes. Our results showed that driver genes selected by ProcessDriver are enriched in known cancer genes. Using survival analysis, we showed that drivers are linked to new tumor events after initial treatment. The second tool was developed to leverage structural and epigenetic data to infer interactions between regulatory genes and targets on a network-level. Our canonical correlation analysis-based approach utilized the DNA methylation or copy number states of potential regulators and the expression states of potential targets to score regulatory interactions. We then incorporated these regulatory interaction scores as prior knowledge in a dynamic Bayesian framework utilizing time series gene expression data. Our results indicated that the canonical correlation analysis-based scores reflect the true interactions between genes with high accuracy, and the accuracy can be further increased by using the scores as a prior in the dynamic Bayesian framework. Finally, we are developing an algorithm to detect cancer-related microRNAs, associated targets and disrupted biological processes. Our preliminary results suggest that the modules of miRNAs and target genes identified in this approach are enriched in known microRNA-gene interactions.

ACKNOWLEDGEMENTS

Brittany Baur

I would first and foremost want to thank my adviser, Dr. Serdar Bozdag. His help, patience and guidance have been an integral part of finishing my dissertation. I would also like to thank my dissertation committee in no particular order, Dr. Daniel Rowe, Dr. Leah Solberg-Woods, Dr. Allison Abbot, Dr. Stephen Merrill and Dr. Mehdi Maadooliat. Thank you all for all the helpful comments and feedback over the years. I am also grateful to the Department of Mathematics, Statistics and Computer Science and Marquette graduate school for all their support over the last six years.

I would like to thank my fiancé, Adrian Haynes, my parents, Amy and Andrew Baur and my brother, Nicholas Baur for always being there whenever I needed it.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Introduction

1.1. Biological background

This section defines the biological background and terms that will be used throughout this dissertation.

1.1.1. Copy number aberration

Copy number refers to the number of copies of a gene. Typically, there are two copies of each gene in a diploid genome, one from each parent. Copy number variation (CNV) is a structural variation in the copy number between human individuals (Henrichsen et al., 2009). CNVs arise from germline cells and are therefore present in all the cells of the organism (Li et al., 2009). CNVs are present in healthy individuals and are responsible for phenotypic variation in humans, but can also cause diseases (Feuk et al., 2006; Henrichsen et al., 2009).

Copy number aberrations or alterations (CNAs) in cancer are somatic changes to copy number that are only present in the tumor (Li et al., 2009). Amplifications will usually lead to an increase in expression of genes within the region that is amplified (Lu et al., 2011). Deletions will usually decrease the expression of genes within the deleted region (Lu et al., 2011). More recently it has been shown that aberrations of regulatory elements can also alter gene expression (Beroukhim et al., 2017). For example, enhancer amplification or a deletion of an insulator element can increase the expression of adjacent genes (Beroukhim et al., 2017). In addition, long-range chromosomal rearrangements and aberrations that place genes closer to enhancers can also alter expression (Beroukhim et al., 2017). CNAs that recurrent in cancer patients generally harbor "driver" genes that confer a fitness advantage for tumorigenesis (Akavia et al.,2010). There is a positive selection for driver genes that promote cancer cell proliferation and survival in tumors (Akavia et al., 2010).

1.1.2. Epigenetic variation

Epigenetics refers to non-genetic influences on gene expression. In other words, gene expression can be altered without a change in the DNA sequence. This dissertation focuses on DNA methylation. DNA methylation is a chemical change to DNA, in which a methyl group is added to the nucleotide cytosine. Heritable DNA methylation of cytosine occurs at a CpG site (Schübeler, 2014). A CpG site is where a cytosine nucleotide is linked to a guanine nucleotide by a single phosphate in the 5' to 3' direction. Approximately, 60% to 90% of CpGs are methylated in human (Tucker, 2001). When CpG sites are clustered together, it is known as a CpG island.

The effect of DNA methylation on gene expression is dependent on the genomic position and CpG island status of the DNA methylation. DNA methylation in promoter regions near the transcription start site (TSS) will lead to a decrease in gene expression, regardless of whether the DNA methylation is in a CpG island (Varley et al., 2013). However, DNA methylation in the gene body, farther away from the TSS, could increase or decrease gene expression depending on whether it is in a CpG island (Varley et al., 2013). DNA methylation in a gene body and not in a CpG island typically increases gene expression. However, if the DNA methylation occurs inside a CpG island, it could increase or decrease gene expression (Varley et al., 2013). DNA methylation of the first exon and near the TSS is tightly linked to decreases in gene expression (Brenet et al., 2011).

CpG islands are present in approximately 70% of mammalian promoters (Dawson and Kouzarides, 2012). Approximately 5 to 10% of promoter CpG islands are hypermethylated in cancer (Dawson and Kouzarides, 2012). Although hypermethylation of promoters is widely studied in cancer, DNA methylation of the gene bodies may activate oncogenes and could be a therapeutic target in cancer (Yang et al., 2014).

1.1.3. microRNAs

microRNAs (miRNAs) are small (~22 nucleotides long), non-coding molecules of RNA (Ambros, 2004). Pre-miRNAs are produced from non-coding DNA, and then is exported to the cytoplasm where it is processed into a mature miRNA (Lima et al., 2011). miRNAs decrease gene expression by base-pairing with complementary mRNA transcripts (Lima et al., 2011). If the miRNA base-pairs with the mRNA transcript with complete complementarity, the mRNA transcript will be cleaved (Lima et al., 2011). If the miRNA base-pairs with partial complementarity, translation of the mRNA transcript into a protein will be repressed or the mRNA transcript will be destabilized (Lima et al., 2011). In all cases, the translation of a mRNA transcript into a functional protein is prevented. miRNAs can target many genes and a gene could be targeted by multiple miRNAs.

miRNAs are largely down-expressed in tumors relative to normal tissue (Di Leva et al., 2014). Several studies have shown the loss of Dicer1, which is involved in the maturation of miRNAs, promotes tumorigenesis (Kumar et al., 2009; Lambertz et al., 2009). These results suggest that miRNAs have mostly tumor suppressor properties, however there are several up-expressed oncomiRNAs in cancer (Di Leva et al., 2014).

1.1.4. Cancer genomes

Genome instability is defined as a high frequency of mutations, such as chromosomal rearrangements, copy number variations and nucleotide changes (Vincent et al., 2014). During normal cell generation, the rates of spontaneous mutations are very low due to “caretaker” genes that resolve defects in DNA replication (Hanahan and Weinberg, 2011). These genes behave as tumor suppressors as their functions can be lost by epigenetic repression or copy number deletions, which can lead to an increased mutation rate and in turn increase the risk of tumor progression (Hanahan and Weinberg, 2011). Structural and epigenetic changes occur randomly, but can by chance affect cancer genes, such as these “caretaker” genes, other tumor suppressors and oncogenes (Stratton, 2011).

Structural and epigenetic changes are inherited over the course of mitotic cell division, allowing deleterious alterations that undermine genome integrity to accumulate and increase the proliferation and invasiveness of cancer cells (Jones and Baylin, 2007; Hanahan and Weinberg, 2011). Epigenetic changes may collaborate with structural changes to evolve cancer cells (Jones and Baylin, 2007). Alterations that promote cell proliferation of the cancer cell have a positive selective advantage in cancer and therefore deleterious alterations are often recurrent in cancer patients (Hanahan and Weinberg, 2011). miRNA genes have often been found to be located at fragile sites of the genome that are prone to alteration, indicating a causative role of miRNAs in cancer progression, as well (Vincent et al., 2014).

1.2. Statement of problem

In 2016, 1.7 million new cancer cases and 595,690 cancer-related deaths were projected to occur in the US (Siegel et al., 2016). Cancer genomes accumulate alterations that confer a fitness advantage for cancer proliferation and survival. These alterations can include copy number amplifications and deletions, aberrant DNA methylation and changes in the expression of microRNAs (miRNAs) compared to non-cancer tissue. The genes and miRNAs that are directly affected by these alterations promote tumorigenesis, and drastically alter the cellular phenotype. These are key regulatory genes and miRNAs that when disrupted, alter the expression of many downstream target genes. Recent technology can generate vast amounts of biological data on the entire genome. Therefore, it is important to gain meaningful information about cancer through high-throughput biological datasets.

The overall goal of this dissertation is to leverage structural, epigenetic and miRNA alterations in the cancer genome to identify key regulators that are disrupted by these alterations and their associated targets by integrating multiple types of biological data. To this end, we have developed several algorithms to address the aims of this dissertation. The following is an overview of the aims that are addressed this dissertation.

1.2.1. Select regions of a gene in which DNA methylation is predictive of its expression

Cancer tissue can exhibit DNA methylation that is too high or too low in critical genes compared to normal tissue (Akhavan-Niaki and Samadani, 2013). Hypermethylation of CpG islands in a gene's promoter in cancer is a typical feature in many cancer genomes (Jones and Baylin, 2007). This type of hypermethylation generally leads to a decrease in the expression of tumor supressors. For example, hypermethlation of the promoter of tumor suppressor genes ITIH5, DKK3 and RASSF1A are biomarkers of breast cancer (Kloten et al., 2013). Since hypermethylation of the promoter usually decreases gene expression, these tumor suppressors are less expressed which in turn allow the tumor to proliferate. Hypomethylation is also a phenomenon in cancer and plays an important role in tumor progression. For example, hypomethylation of Wnt5a, a signaling protein that influences the expression of many other genes, could make the gene more accessible for up-expression and promote aggressiveness in prostate cancer (Wang et al., 2007). Hypomethylation of oncogenes, such as cMYC and H-RAS, may also make them more accessible for upregulation (Akhavan-Niaki and Samadani, 2013).

There are several next-generation sequencing-based assays to measure DNA methylation such as bisulfite sequencing (Chatterjee et al.,2011), MeDIP-seq (Down et al., 2008), and reduced representation bisulfite sequencing (Gu et al., 2011). There are also bisulfite microarray-based assays to measure DNA methylation (Adorján et al., 2002). For humans, the Illumina Infinium HumanMethylation27 BeadChip Kit array contains 27,578 probes for 14,495 genes (Weisenberger et al., 2008). Later, Illumina developed higher-resolution Illumina Infinium HumanMethylation450 BeadChip Kit array, which have an average of 18 probes associated with a gene in various genomic positions and CpG island statuses (Bibikova et al., 2011). Due to its high resolution and low cost, the Illumina Infinium HumanMethylation 450K array has become one of the most frequently used assays to quantify DNA methylation in human. At the time of writing, the Gene Expression Omnibus database (Barrett et al., 2013) had about 30,000 samples that were profiled using the Illumina 450K array.

Choosing representative DNA methylation probes is important for downstream functional analysis, such as determining if a gene has aberrant DNA methylation in cancer (Maeda et al., 2014). DNA methylation probes that are predictive of gene expression may be closer to a functional region of interest. For example, Rhee et al. found that genes that were down-expressed and had hypermethylation in the TSS contain sequences for transcription factor binding (Rhee et al., 2013). Selecting one or two representative probes is also important for predictive models that may integrate other sources of biological data. For example, Li et al. tested various feature selection methods to predict whether a gene is up or down expressed in lung cancer based on DNA methylation and histone features (Li et al., 2015). Using the 450K DNA methylation data, the authors averaged the DNA methylation probes in genomic regions, such as the gene body. The drawback of averaging the value at each probe is that signals can be lost. An alternative approach would be to utilize the most representative probes as features.

However, it is not straightforward to determine which probes to choose from a 450K array that best represent the overall methylation level of the gene and are informative to the gene’s expression level. A simple, but valuable approach may be to choose a single probe based on a metric such as the variation. One approach is to use the standard deviation (SD) across samples and choose the probes with the greatest variation (Selamat et al., 2012; Noushmehr et al., 2010). Other studies restrict the analysis to probes from CpG islands in upstream regions, since DNA methylation blocking transcription factors from binding is a well-studied phenomenon (Li et al., 2014). Several studies restrict the number of probes to those within a certain proximity surrounding the TSS (Farré et al, 2015; Rica et al., 2013). However, both approaches ignore possibly informative DNA methylation in the gene body.

Due to the context-dependent nature of DNA methylation, the need to identify the regions of DNA methylation of interest, and its critical importance to cancer, we proposed an approach that, for a given gene, selected the most "informative" areas of DNA methylation. In this method, "informative" was defined by the probe(s) of the gene where DNA methylation was most predictive of gene expression. Gene expression was binary, indicating whether the sample was up-expressed or down-expressed when treated with a hypomethylating agent versus untreated for breast cancer cell line data (Li et al., 2014). We also used up-expressed and down-expressed samples with respect to the median for the Cancer Genome Atlas (TCGA) luminal A breast cancer data (The Cancer Genome Atlas, 2012). This approach was designed for the 450K DNA methylation array, where there is an average of 18 probes per gene.

Multiple classification and feature selection methods to select the most informative DNA methylation probes for a given gene were evaluated in this aim. Due to the context-dependent nature of DNA methylation, the feature selection was unsupervised and did not consider genomic position of the probes or CpG island status of the genome position.

1.2.2. Infer copy number drivers and associated biological processes

A copy number aberration that is recurrent in cancer patients harbors genes that promote cancer cell proliferation and survival (Hanahan and Weinberg, 2000). There is a positive selection advantage for an aberration that affects genes that allow the tumor to grow and proliferate. These genes, which are oncogenes and tumor suppressors, are known as "drivers." Large aberrations can also harbor genes that do not have a fitness advantage to tumor proliferation which are known as "passenger" genes. Passenger genes that do not have a selective advantage are amplified or deleted along with the drivers due to their proximity to the driver and as a result, have similar changes in expression with respect to copy number. Due to their similar copy number and expression profiles, separating drivers from passengers is an important and difficult challenge.

The goal of this aim is to identify copy number drivers in a large aberration by associating the driver with downstream disrupted biological processes. We proposed a computational pipeline, called ProcessDriver, based on the idea that there are driver genes located within an aberrated region that regulate the expression of genes outside the aberration. Therefore, an aberration can have effects across the genome extending beyond the region undergoing gains and losses via the driver genes inside the region. This is because a driver is influential in changing the pathology of the cell from normal to tumor, and therefore has many target interactions. The driver gene is the link between the aberration and genes affected by the aberration located elsewhere in the genome. This idea was leveraged to separate the passengers from drivers.

Additionally, our method is unique in uncovering the biological processes that are driven by the driver genes. Certain biological processes are known to be disrupted in cancer, such as cell cycle and cell death (Evan and Vousden, 2001). Aberrations that allow the cell to evade cell death and undergo cell cycle more frequently are favored in tumors. ProcessDriver associates a driver with the targets of the biological process(es) that it most likely disrupts.

1.2.3. Infer gene regulatory networks by integrating structural and epigenetic information

One of the challenging and important computational problems in systems biology is to infer networks of genetic interactions. A gene regulatory network is a graph where nodes represent genes and edges between the genes represent an interaction. The interaction, for instance, could be a transcription factor-target relationship. In a directed network, an edge goes from a regulator to a target. Traditionally, gene expression data are used to detect changes in a regulator's expression and examine the corresponding downstream effects on a target's expression (Hecker et al., 2009). However, heterogeneous data sources have improved the inference of gene regulatory networks (Hecker et al., 2009).

The goal of this aim is to infer regulatory networks by integrating copy number and DNA methylation along with gene expression data. We proposed canonical correlation analysis-based approach which utilized the DNA methylation or copy number states of potential regulators and the expression states of potential targets to score interactions. Our algorithm assumes that changes to a regulator's copy number or DNA methylation would lead to downstream changes in a target's expression level. Therefore, changes in the DNA methylation or copy number states of regulators may be seen as a natural perturbation to the regulator that can aid in establishing directionality in the network. Therefore, this approach may be better than using expression states for both regulators and targets. Furthermore, we integrated time series gene expression data with a dynamic Bayesian approach using the scores from our canonical correlation analysis-based algorithm as prior knowledge.

1.2.4. Infer cancer-related miRNA-gene module drivers

The expression of certain key miRNAs is known to be altered in cancer cells (Lu et al., 2005). Since miRNAs regulate the expression of genes, changes in the expression of key miRNAs in cancer could have widespread, downstream effects. A miRNA and its target genes are known as a "driver module" if the effects of a disruption in miRNA expression, and corresponding changes in the expression of its target genes, promote cancer cell survival and proliferation.

The goal of this aim is to associate miRNAs with potential targets via biological processes. Certain biological processes are known to be dysregulated in cancer tumors via miRNAs, such as apoptosis (Lima et al., 2011) and cell cycle (Kim et al., 2009). If a miRNA is disrupted in cancer, and the targets genes are involved in one or more of these processes, that miRNA is more likely to be a driver. Therefore, biological process information could be used to aid in miRNA-gene module driver detection. To our knowledge, no other approach has associated a miRNA with target genes via biological processes. Our approach is able to identify likely miRNA drivers, associated targets and processes that are disrupted as a result of the changes in expression of the miRNA driver and corresponding target genes.

1.3 Status of Problem

1.3.1. Select regions of a gene in which DNA methylation is predictive of gene expression

To our knowledge, there has been no previous algorithm designed to select DNA methylation probes associated with a gene from the Illumina Infinium 450K DNA methylation array that are most informative to a gene's expression level. However, a variety of studies integrate epigenetic factors to explain gene expression and are outlined here.

Rhee et al. provided an extensive analysis of the effects of DNA methylation on gene expression in different molecular subtypes of breast cancer (Rhee et al., 2013). They found that there is more positive correlation of gene expression moving upstream of the TSS in less aggressive subtypes of breast cancer compared to more aggressive subtypes. This study also used decision trees to investigate the combinatorial effects of DNA methylation status in different genomic positions on gene expression and found CpG islands to be the most informative feature.

Li et al. tested various models to predict differential gene expression in normal versus tumor samples using epigenomics data in lung cancer (Li et al., 2015). The model predicts whether an individual gene is up- or down-expressed in lung cancer compared to normal tissue using histone H3 methylation modification, DNA methylation, nucleotide composition and nucleotide composition based features. They found that a model comprised of 67 features chosen with a ReliefF feature selection and random forest classification performed the best. Many of the selected features were related to the CpG methylation status of the promoter suggesting that promoter methylation is an important predictor of differential expression in normal versus tumor samples.

Gevaert et al. developed an algorithm called MethylMix which identifies differentially methylated genes that are predictive of gene expression (Gevaert et al., 2015). MethylMix uses beta-mixture modeling to identify subpopulations of patients with similar DNA methylation levels for each CpG site. For a CpG site, each beta mixture represents a subset of patients where a particular beta distribution of DNA methylation states is observed. Next, the algorithm determines which sites are hypo- or hyper-methylated by comparing the mean of each mixture component of each CpG site with the mean methylation of the normal samples. For hypo- or hyper-methylated genes, linear regression was used to determine if DNA methylation had a significant impact on gene expression. Their analysis found that hyper- and hypo-methylated genes have oncogenic and tumor suppressor properties. For example, they found that tumor suppressor TMEM25 was hypermethylated in many cancers, and the hypermethylation prevents gene expression.

1.3.2. Infer copy number drivers and associated biological processes

Many algorithms have been proposed that identify candidate copy number genes. The problem of computationally separating driver genes that promote tumorigenesis from passenger genes in a large, aberrated region remains a challenging one. The result of these algorithms is a list of candidate driver genes in the recurrent aberration, possibly a ranked list with a score. These lists could be further validated by experimental or computational techniques.

GISTIC2.0 uses segmented copy number data to find regions of the genome that harbor drivers because they are recurrent in cancer samples (Mermel et al., 2011). Segmented copy number data describe the copy number of a particular segment of the genome for a given patient. It could be thought of as a “snapshot” of the copy number at the point in time the data were obtained. However, the segmented data does not describe the underlying alterations that have taken place resulting in a particular segmented datum. Different alterations could have taken place leading up to the “snapshot” value, and alterations can often overlap. Therefore, finding alterations that are recurrent in cancer samples is a challenge.

GISTIC2.0 deconstructs the segmented copy number profile into a set of likely alterations, and then finds the alterations that are recurrent over cancer samples. An algorithm called Ziggurat deconstruction alternates between estimating the background rates of copy number alteration and computing the most likely deconstruction for each copy number profile. The output is the copy number alterations for each cancer sample.

The next step is to define the regions that are undergoing significant alterations. That is to score regions of the genome according to the probability that the observed set of copy number alterations within the region would have occurred by chance. Alterations that harbor a driver gene are likely to be frequently occurring and of higher amplitude, therefore GISTIC2.0's scoring of regions take both into account. Higher scores mean that the region is likely not altered by chance, and is undergoing positive selection because it is harboring a driver. GISTIC2.0's handling of segmented copy number made it a popular choice for future work for pipelines integrating expression data to narrow down the candidate drivers within the regions proposed by GISTIC2.0.

Several studies have focused on integrating cis gene expression (Tamborero et al., 2013; Fan et al., 2012; Pickering et al., 2013). Cis genes are genes that are located within the aberration that would be directly impacted by that aberration. The idea behind these studies is that the cis gene in which copy number has the greatest influence on its expression is a likely candidate driver. For example, Oncodrive-CIS predicts likely drives based on copy number impact on gene expression (Tamborero et al., 2013). For a given gene, a score is calculated for each sample with an aberration that represents the aberration's impact on gene expression in the sample when compared to non-aberrated samples as a reference. The median of these scores is the overall score for the gene. When compared to a background model, when the overall score is higher, the gene is likely a copy number driver. Ambatipudi et al. also selected drivers based on copy number's impact on cis gene expression in gingivobuccal cancers (Ambatipudi et al., 2012). The strength of the correlation between cis copy number and gene expression is also used to detect drivers (Fan et al., 2012; Pickering et al., 2013).

Several other methods associate cis genes with downstream target trans genes. Trans genes are genes located outside of the aberration. The idea behind these studies is that a copy number aberration disrupts a key regulatory cis gene. A regulatory gene will control the expression of target genes elsewhere in the genome. Therefore, when a key regulatory gene is disrupted because of an amplification or deletion, it is expected that there are widespread, downstream affects in trans.

CONEXIC is a computational pipeline that associates a driver cis gene that is disrupted as a result of the aberration with a module of downstream target genes (Akavia et al., 2010). A modified version of GISTIC is used to find significant aberrations that are recurrent in cancer patients. Candidate drivers reside within the regions reported by GISTIC. The next steps in the process are the single modulator step and the network learning step. In the single modulator step, each candidate driver gene is associated with a preliminary module of target genes. For a candidate driver, k-means clustering with k=2 and a normal distribution is used to separate the high and low expressed samples. The initial clusters are the non-aberrated and aberrated samples. The resulting boundary between the clusters is the threshold for the target gene expression. The target gene expression is split into two groups, samples where the driver gene expression is below a threshold and samples where the driver gene expression is above a threshold. If the split is significant then the target gene is associated with the driver's module. Modules have twenty or more target genes to ensure the candidate driver having a large, widespread effect.

In the network learning step of CONEXIC, a regulation program for each module is learned. The modules have more than twenty target genes as determined in the previous step. However, any candidate driver associated with more than twenty target genes is a possible regulator for the module. Therefore, the regulators associated with a module in the network learning step do not necessarily have to contain the candidate driver the module was originally associated with in the previous step. The regulatory program is a regression tree where the decision nodes are the driver and a query on its expression value. The answer to the query is the corresponding expression of the module. A driver that best splits the module of target gene expression into two behaviors is chosen at each step. All possible driver-split value combinations are tested. Finally, after the regulation program is learned, genes can be moved of out or into the module associated with the regulation program if there is an improvement in the score.

Aure et al. proposed a computation pipeline that first identifies in-cis correlated genes and then identified biological processes that the cis-genes were associated with in trans (Aure et al., 2013). First, in-cis correlation analysis was performed where a cis gene was selected if its expression was correlated with its own copy number above a certain threshold with a low false discovery rate. Next, the in-trans correlation analysis was performed. The correlations between the expression of a selected cis gene and all other genes were calculated. The trans genes were ranked by their correlation to a given cis gene and an enrichment score was calculated for each Gene Ontology (GO) term. The enrichment score was the p-value from a minimum hypergeometric test. Background simulations were performed to test this enrichment. If the actual p-value was significant and better than all the p-values from simulations, the in-cis correlated gene was associated with a GO term in trans. Cis genes that are highly correlated to copy number and significantly associated with biological processes are potential driver genes.

1.3.3. Infer gene regulatory networks by integrating structural and epigenetic information

Many different methods have been applied to the problem of inferring gene regulatory networks (Margolin et al., 2006; Husmeier, 2003). Many of the methods are dependent on either time series or steady state gene expression data (Hecker et al., 2009). One of the most popular tools, ARACNE, is based on information theory based on steady-state gene expression data (Margolin et al., 2006). This study defines an edge between genes as an irreducible statistical dependency between the genes. This statistical dependency is defined as the mutual information between two genes, which, unlike Pearson correlation, is invariant and non-zero if a dependency exists. ARACNE has a high true positive rate as well as a high false negative rate.

Dynamic Bayesian network (DBN) is a popular method for inferring gene regulatory networks from time series data (Husmeier, 2003). First, a Bayesian network is described by a graphical structure $G = (V, E)$, where *V* is the set of vertices and *E* is the set of directed edges, a family of conditional probability distributions F and their associated parameters q that together defines a joint distribution over the random variables (genes) of interest. Let $X_1, X_2, \dots, X_n$ be a set of random variables to be nodes in the graph. The joint probability is built on conditional probabilities based on the parents of X_i, $pa(X_i)$:

$$P(X_1, X_2, \dots, X_n) = \prod_{i=1}^{n} P(X_i | pa[X_i]) \tag{1.1}$$

Since the family of conditional probability distributions if fixed, the problem becomes identifying the associated parameters q and the network model *G** by finding:

$$G^* = argmax_G\big(P(G|D)\big) \tag{1.2}$$

And by finding the parameters q* that maximize $P(q|D, G^*)$. That is the maximization of the structure given the expression data. By applying Bayes rule, the posterior probability is:

$$P(G|D) = \frac{1}{Z} P(D|G)P(G) \tag{1.3}$$

Where $Z = \sum_G P(D|G)P(G)$ is a normalization factor and *P(G)* is the prior. The marginal likelihood *P(D|G)* is calculated by integrating out the parameters:

$$P(D|G) = \int P(D|q, G)P(q|G)dq \tag{1.4}$$

When the conditional probabilities are defined by a linear Gaussian distribution or a multinomial distribution and the data is complete, this integral is analytically tractable (Husmeier, 2003). However, multinomial distributions are often preferred because of their ability to capture non-linear relationships between genes although data discretization often leads to information loss (Husmeier, 2003).

However, although the integral in Eq. 1.4 can be solved the posterior distribution in Eq. 1.3 is usually intractable. As the number of nodes in the graph increases, the number of potential graphs also increases which makes an exhaustive search impossible since the denominator in Eq.

1.3 becomes intractable. Additionally, since the data D is sparse, the data also may not be represented well by a single $G*$ with the highest posterior probability and may be better represented by a collection of graphs (Husmeier, 2003). Uncovering the network structure that maximizes the posterior distribution is only feasible if the posterior distribution is sharply peaked (Werhli and Husmeier, 2007). Therefore, algorithms such as greedy hill climbing and Markov Chain Monte Carlo (MCMC) are needed to sample from the posterior probability (Tsamardinos et al., 2006; Hastings, 1970).

Furthermore, one constraint of the Bayesian networks is that they must be acyclic. This is not an acceptable constraint given the prevalence of feedback loops in biology including gene regulation (Husmeier, 2003). However, biological cause and effects such as a transcription factor influencing the expression of a target gene does not occur simultaneously, as there is some time delay (Husmeier, 2003). In Dynamic Bayesian networks, the way around the acyclic constraint is to ‘unfold’ the network across time points (Husmeier, 2003). The amount of time between slices is considered homogenous in most cases because of the increase in model complexity otherwise. However, Zou and Conzen limited the number of potential regulators of a target to regulators that had an earlier or simultaneous expression change to the target (Zou and Conzen, 2004). Therefore, the transcriptional time lag can be zero to several units in this setting.

Due to the intractability of the denominator in Eq. 1.3, an appropriate heuristic approach would be the MCMC sampling with a Metropolis-Hastings acceptance criterion to sample from the posterior distribution (Husmeier, 2003). In the MCMC approach for dynamic Bayesian networks edges can be added or deleted. It is worth noting that a reversal of edge direction is not an option because in the network, which is unfolded in time, would mean that an effect preceded a cause (Husmeier, 2003). Additionally, edges within a time slice are not allowed, as that would mean the events happened simultaneously (Husmeier, 2003). Therefore, a new graph G_{new} is proposed based on the old graph G_{old} by adding or removing an edge between time points. The Metropolis-Hastings acceptance criterion is:

$$P_{MH} = \min\left\{1, \frac{P(G_{new}|D)}{P(G_{old}|D)} \times \frac{Q(G_{old}|G_{new})}{Q(G_{new}|G_{old})}\right\} \quad (1.5)$$

This acceptance criterion cancels out the intractable Z of the posterior probability (Eq. 1.3). Q represents the proposal probability. The Hastings ratio $\frac{Q(G_{old}|G_{new})}{Q(G_{new}|G_{old})}$ is one because without the possibility of an edge reversal move, and with the network unfolded in time, the proposal probabilities are equal. Specifically,

$$\frac{Q(G_{old}|G_{new})}{Q(G_{new}|G_{old})} = \frac{N(G_{old}\ )}{N(G_{new})} \quad (1.6)$$

Where N is the number of neighborhoods, or potential acyclic graphs created by adding or removing an edge from the old or new graph. A potential graph would only be rejected if it's cyclic. Since a dynamic Bayesian network is guaranteed to be acyclic, the number of structures that can be created by adding or removing an edge is the same for both G_{old} and G_{new} and the Hastings ratio is one.

The fact that many interactions must be learned from a small number of time points means that the prior probability would have a large impact on the overall posterior probability (Husmeier, 2003). Several studies have devised priors from various types of biological data (Imoto et al., 2003; Werhli and Husmeier, 2007; Zheng et al., 2011; Chen et al., 2013; Baur and Bozdag, 2015).

One common type of prior for incorporating multiple types of biological data was adapted in several studies (Imoto et al., 2003; Werhli and Husmeier, 2007; Zheng et al., 2011; Chen et al., 2013; Baur and Bozdag, 2015). The prior takes the form of a Gibbs distribution (Eq. 1.7) where prior information was encoded by an energy function (Eq. 1.8), and $Z(\beta)$ was a normalizing constant. The hyperparameter, β, measured the influence of the prior information relative to the time series expression data (Werhli and Husmeier, 2007). Using one source of prior knowledge could easily be extended to incorporating multiple sources of prior knowledge simultaneously as described in Werhli and Husmeier, 2007.

$$P(S|\beta) = \frac{e^{-\beta E(S)}}{Z(\beta)} \tag{1.7}$$

The energy function measured how closely the prior information matched with the network structure at the current step of MCMC (Eq. 1.8). In energy function, *B* is the prior matrix, and *G* is the current network structure (Werhli and Husmeier, 2007). As the energy goes to zero, there is more agreement between the prior and the network structure.

$$E(S) = \sum_{i,j=1}^{N} \left|B_{ij} - G_{ij}\right| \tag{1.8}$$

Imoto et al. used this prior to integrate binding site information, protein-protein interactions and protein-DNA interactions (Imoto et al., 2003). Werhli and Husmeier used it to include binding site information (Werhli and Husmeier, 2007). In a couple of studies, the prior was used to include histone modification data (Chen et al., 2011; Zheng et al., 2013). The idea behind these two studies is that genes with correlated histone modification profiles are more likely to interact. Therefore, the histone modification data can be used as prior information to integrate along with the time series gene expression data.

1.3.4. Infer cancer-related miRNA-gene module drivers

A few studies have developed computational methods to establish miRNA-gene modules (Karim et al., 2016; Jin and Lee, 2015; Zhao et al., 2015). Karim et al. outlined a methodology to infer miRNA-gene modules through collective group relationships (2016). From the correlation matrix of miRNA expression and gene expression, a matrix of collaboration scores was computed for miRNAs, which reflected the similarity or collaboration in regulating the same target genes. Another matrix of collaboration scores was also computed for genes, which reflected their similarity in being regulated by the same miRNAs. Both matrices underwent clustering separately. Groups of miRNAs that regulate the same genes were formed, and groups of genes regulated by the same miRNAs were formed by clustering the collaboration scores. Canonical correlation analysis was used to establish relationships between the groups of miRNAs and the groups of genes, retaining the relationships that had the highest canonical correlation.

Jin and Lee used a Bayesian approach to identify miRNA-gene modules in cancer (Jin and Lee, 2015). First, a biclustering approach was used on the gene expression data to form gene-sample modules. Gene-sample modules were used since cancer is a heterogeneous disease even between patients with the same type of cancer. These gene-sample subsets are likely to be functionally related. A Bayesian network approach was used to connect candidate miRNAs to the genes in the gene-sample module. A network was constructed based on the likelihood of a set of genes and a set of miRNAs as a joint distribution.

1.4 Organization of the Dissertation

Each chapter following the introduction is based on a manuscript that is either published, submitted or in preparation for publication. Some introductory content from each manuscript was moved to this chapter to motivate the work and allow for clarity and elaboration on the current literature. Additionally, some supplemental materials published or submitted along the papers were added to their respective chapters for continuity. The final chapter summarizes the main conclusions of the dissertation and presents future work for integrating multiple types of biological data to infer interactions. Each chapter addresses each aim in the order that they appear in this introduction.

CHAPTER 2

A feature selection algorithm to compute gene centric methylation from probe level methylation data

This chapter appears in Baur and Bozdag, *PLoS ONE*, 2016

Abstract: DNA methylation is an important epigenetic event that affects gene expression during development and various diseases such as cancer. Understanding the mechanism of action of DNA methylation is important for downstream analysis. In the Illumina Infinium HumanMethylation 450K array, there are tens of probes associated with each gene. Given methylation intensities of all these probes, it is necessary to compute which of these probes are most representative of the gene centric methylation level. In this study, we developed a feature selection algorithm based on sequential forward selection that utilized different classification methods to compute gene centric DNA methylation using probe level DNA methylation data. We compared our algorithm to other feature selection algorithms such as support vector machines with recursive feature elimination, genetic algorithms and ReliefF. We evaluated all methods based on the predictive power of selected probes on their mRNA expression levels and found that a K-Nearest Neighbors classification using the sequential forward selection algorithm performed better than other algorithms based on all metrics. We also observed that transcriptional activities of certain genes were more sensitive to DNA methylation changes than transcriptional activities of other genes. Our algorithm was able to predict the expression of those genes with high accuracy using only DNA methylation data. Our results also showed that those DNA methylation-sensitive genes were enriched in Gene Ontology terms related to the regulation of various biological processes.

2.1 Introduction

Methylation of cytosine nucleotides in DNA (hereafter DNA methylation) is involved in cellular differentiation (Meissner et al., 2008), development (Bird, 2002) and has impact in diseases such as cancer (Jones and Baylin, 2007). DNA methylation is typically associated with a decrease in gene expression due to its role in blocking transcription factors from binding (Jones, 2012). It is also speculated that silencing of a gene could precede DNA methylation (Jones, 2012). DNA methylation is also known to have positive correlation with gene expression, as well, particularly in gene bodies (Jones, 2012). Several studies integrate DNA methylation with gene expression to unravel the role of DNA methylation in gene regulation (Brenet et al., 2011; Varley et al, 2013; Rhee et al., 2013; Baur and Bozdag, 2015).

In the Illimina Infinium 450K DNA methylation array, each gene is associated with around 18 DNA methylation probes. In this study, we developed a feature selection algorithm based on sequential forward selection that can utilize various classification methods to select probes that are relevant to gene expression from the 450K array. We also tested this algorithm against more sophisticated approaches such as support vector machines with recursive feature elimination (SVM-RFE), a genetic algorithm and ReliefF. Additionally, we compared our algorithm against several selection methods that do not use gene expression to inform the selection. These methods include choosing the probe with the greatest variation, choosing probes close to the TSS, and choosing probes in upstream CpG islands. Following the selection of probes, we computed several metrics to evaluate the prediction quality of gene expression by the selected probes. These metrics included precision, recall, specificity and Matthew's correlation coefficient. Our results showed that our sequential forward selection algorithm performed best on all metrics when using K-Nearest Neighbors (KNN) where K = 1 (1NN). Our algorithm generally selects one or two probes for each gene, which allows to us identify key regions where DNA methylation changes have impact on gene expression.

We also observed that our algorithm could determine genes whose expression levels are putatively sensitive to the changes in their DNA methylation. We showed that these DNA methylation-sensitive genes were enriched for Gene Ontology (GO) terms related to the regulation of various biological processes. Additional functional analysis clustering showed that DNA methylation-sensitive genes also regulated other genes and proteins by a variety of mechanisms, including DNA-binding, kinase activity, protein degradation and protein synthesis.

2.2. Materials and Methods

2.2.1 Data

Agilent whole genome microarray data and Illumina 450K DNA methylation data of 25 breast cancer lines after treated with the hypomethylating agent, 5-azacitidine (aza) for 72 hours were downloaded from (Li et al., 2014) (GSE57343). Log10 Mock/Aza expression data were normalized to account for the different cell lines using LoEss normalization in the LIMMA package (Schuebel et al., 2007; Smith, 2005). To perform binary prediction of gene expression, the expression data were discretized into up, down and baseline categories using 1.1-fold change threshold for aza-treated cells with respect to mock trials (mock/aza). Baseline mock/aza values were removed. The up and down-expressed mock/aza samples were the binary classifiers in the classification algorithms.

To verify the results of our algorithm on breast cancer cell line, we also downloaded Illumina 450K DNA methylation and Agilent mRNA expression data for 99 Luminal A breast cancer samples from the TCGA repository (The Cancer Genome Atlas Network, 2012). Batch effects were corrected in the mRNA expression data using the LIMMA package (Smith, 2008). Expression data were discretized with a log2 1.2-fold change of the expression level of the sample over the median expression level for that gene across samples. We used the 1.2-fold change threshold instead of 1.1 in tissue samples to reduce potential noise in the discretized data.

Baseline sample expression/median expression values were removed. The up and down-expressed sample expression/median expression were the binary classifiers in the classification algorithm.

2.2.2. A sequential feature selection algorithm for classification methods

We developed a sequential feature selection (SFS) algorithm that can use different classification methods to select the probes that are most relevant to gene expression (Algorithm 1). SFS sequentially adds features until there is no improvement in the prediction. The objective function of the SFS algorithm is the minimization of the mean classification error in a 10-fold cross-validation (CV).

Algorithm 1 describes the process for a single gene and a set of n probes associated with the gene, X. Given the DNA methylation levels of the probes, $M_{k,X}$, and the associated gene expression levels, y_k , each probe is individually tested in a 10-fold cross validation predicting the gene expression based on the DNA methylation levels of the probe (steps 1-5). In each partition of the 10-fold cross validation, the specified classification algorithm (described below) is trained on the training samples. The expression levels of test samples are predicted based on the trained classification algorithm and the methylation levels of the test samples. The number of test samples in which the predicted expression level does not match the true expression level is O. O is computed for every partition and the mean(O) is the classification error, CCE. The probe with the best performance, or minimal CCE, in the 10-fold cross validation is selected (steps 6-8).

Additional probes are sequentially added from the pool of remaining probes if the performance in a 10-fold cross validation improves and more samples are predicted correctly (steps 9-18). If no additional probes lead to increased performance, the algorithm is terminated (steps 19-21).

Algorithm 1. Sequential feature selection with 10-fold CV

Input: y_k: discretized up/down gene expression of sample size *k*
$X=(x_1, x_{2,} \dots . x_n)$: *n* potential probes associated with gene to be added to *S*
$M_{k,X}$*:* DNA methylation values for n probes associated with gene in *k* samples
S: current set of selected probes, initially empty
C: Classification model based on training folds in 10-fold CV
C = Classification $(M_{train,S}\ ,\ y_{train})$*,*
$O(\ M_{test,S}, y_{test}\ \ ) = sum(y_{test} \cong$ *predict(C,* $M_{test,S}$ *))*
Current classification error (CCE): A vector of classification errors for probes being tested, the classification error is mean(O) from a 10-fold CV

```
For i=1:n
    Select probe $x_i$
    Compute 10-fold CV. In each partition, compute C on training and O on test data
    Take mean O as current classification error, CCE(i)
End
Find j s.t. $CCE(j) < CCE(i)$, $1 \le i \le n$, $i \neq j$
Move probe $x_j$ from X to S
Old classification error, $OCE = CCE(j)$
While (true)
    For each $x_i \in X$
        Select probes $\{x_i\} \cup S$
        Compute 10-fold CV. In each partition, compute C on training and O on test data.
        Take mean O as current classification error, CCE(i)
    End For
    Find j s.t. $CCE(j) < CCE(i)$, $1 \le i \le |X|$ , $i \neq j$
    If $CCE(j) < OCE$
        Move probe $x_j$ from X to S
        $OCE = CCE(j)$
    Else:
        Stop search
End While
```

We used the following classification algorithms in combination with sequential feature selection (Algorithm 1).

Support vector machine (SVM): A linear kernel function was used to map the training data to the kernel space (Cortes and Vapnik, 1995). Sequential minimal optimization was used to find the separating hyperplane.

K-Nearest neighbors (KNN): KNN classification algorithm was applied with K = 1,3 and 5 (1NN, 3NN and 5NN, respectively). A Euclidean distance metric was used for all instances of KNN (Friedman et al., 1977).

Decision trees (DT): The minimum parent size (number of observations) was 10 and the minimum leaf size was 1 (Quinlan, 1999).

Naïve Bayes (Bayes): A kernel distribution was specified for predictors in the Naïve Bayes classification algorithm (John et al., 1995).

We also tested other feature selection algorithms, SVM with recursive feature elimination (SVM-RFE), a genetic algorithm feature selection with KNN classification (GA-KNN) and ReliefF.

SVM-RFE: The SVM-RFE algorithm was adapted from (Yan and Zhang, 2015). This study used a correlation bias reduction strategy to deal with highly correlated features. In our adaptation, we also included a modification to deal with class imbalances, such that the weight of misclassifying the minority class was higher. The weights of the penalties were obtained by solving the equation $n0*w0=n1*w1$, where $n0$ and $n1$ were the number of down and up expressed samples, and $w0$ and $w1$ were the respective weights. We used a Gaussian kernel and ranked the features. For each gene, we selected the top k probes where k equals to the number of probes selected in the SFS algorithm.

GA-KNN: A genetic algorithm for selecting features was adapted from (Babatunde et al., 2014). The goal of the GA algorithm was to minimize the fitness function: $\frac{resubLoss}{N-S}$, where $resubLoss$ is the resubstitution loss in a KNN classification (fraction of misclassified data), N is the total number of features and S is the number of selected features. The denominator of the equation penalizes a large number of selected probes. We tested the algorithm using K=1, 3 and 5.

ReliefF: A KNN-based ReliefF implementation from the MATLAB statistics toolbox was also tested. The nearest “hit” of a feature vector for a sample was defined as the closest sample of the same class by Euclidean distance. The nearest “miss” of a feature vector for a sample was defined as the closest sample of the other class. For each iteration, a vector of features from a random instance is selected. The weight of the feature *i* is updated according to the function:

$$W_i = W_i - (x_i - h_i)^2 + (x_i - m_i)^2$$

where x_i is the value of the feature of the randomly selected instance, h_i is the nearest hit and m_i is the nearest miss. Therefore, the weight of a feature decreases if it is more distant from nearby instances of the hits relative to the misses.

We tested this algorithm with *K* = 1, 3 and 5. This implementation ranks the predictors in order of importance. For each gene, we selected the top *k* probes where *k* equals to the number of probes selected in the SFS algorithm.

We also developed two control algorithms namely random and top two correlated.

Random: For a given gene, we randomly selected probes associated with the gene. We set the number of probes randomly selected for a gene equal to the number of probes that were selected in the SFS algorithm that we compared to.

Top two correlated: The two probes most positively or negatively correlated with gene expression were selected.

We tested our algorithm against following probe selection methods, which do not consider gene expression.

All: For a given gene, all the probes associated with the gene are selected.

Upstream CpG Island: For a given gene, we selected probes that are in CpG islands in the upstream regions (TSS200, TSS1500, 5’ UTR and 1st Exon).

TSS: For a given gene, we selected probes within a 2500bp window of the transcription start site.

Top SD: For a given gene, the probe with the highest standard deviation is selected.

2.2.3. Assessment of algorithms

We calculated various metrics to test each algorithm’s ability to predict gene expression based on the selected DNA methylation probes. We applied a leave-one-out cross validation

(LOO-CV) with an appropriate model using the selected probes as predictors and the discretized gene expression as a response. For the SFS algorithm, the classification model used in the feature selection was used in the LOO-CV. For GA-KNN and ReliefF, KNN was used in the LOO-CV. For SVM-RFE, SVM was used in the LOO-CV. For the methods that do not integrate gene expression, namely All, Upstream CpG Island, TSS and Top SD, we evaluated the probe selection with a LOO-CV using KNN, DT, SVM and NB.

Following the LOO-CV, we computed the number of true positive (TP), true negatives (TN), false positives (FP) and false negatives (FN) and calculated various metrics. We considered down-expressed cases positive and up-expressed cases negative outcomes. We calculated the prediction accuracy ((TP + TN)/(TP+TN+FP+FN)), recall (TP/(TP+FN)), precision (TP/(TP+FP)) and specificity (TN/(TN+FP)) for each method. We also computed Matthew’s correlation coefficient (MCC) [Eq 2.1]. MCC can be considered a balanced measure of accuracy even when the class sizes may be different.

$$\frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{2.1}$$

2.2.4. Gene Ontology and functional enrichment

To perform functional analysis on genes whose expression were predicted with high accuracy by DNA methylation, we selected genes that have an MCC > 0.6 in the SFS algorithm. We performed a GO-term enrichment analysis using the web tool GOrilla (Eden et al., 2009), by comparing the list of genes with high MCC to a background of the full list of 17,043 genes in the dataset. To show that the enrichment of GO terms obtained is specific to genes with high MCC, we compared the list of GO terms and p-values for genes with high MCC to the list of GO terms and p-values for genes with MCC < 0.2.

To investigate if there are any functional differences between genes that have gene body and upstream methylation, we performed gene functional classification clustering using DAVID (Huang et al., 2008). Given an input gene list, the DAVID's functional clustering tool generates a gene-to-gene similarity matrix based on shared functional annotations from different sources (Huang et al., 2007). A clustering algorithm classifies the genes into functionally related clusters. Each functional cluster contains certain related terms shared between the genes in the group. We separated all genes with MCC > 0.6 based on whether the selected probes by the SFS algorithm were exclusively from upstream regions (gene had probes only in 5' UTR, 1st Exon, TSS200 or TSS1500 as defined by Illumina) or exclusively from the gene body applied functional clustering using DAVID for each group of genes.

2.2.5. Implementation

Our algorithm is unbiased as it does not restrict analysis by CpG status or genomic position. We implemented the tool in MATLAB. The source code is freely available under the MIT Open Source license (https://github.com/brittanybaur/genecentricmethylation)

2.3. Results and Discussion

2.3.1. KNN-SFS algorithm resulted in higher recall and specificity

We calculated the prediction accuracy, specificity, recall, precision and Matthews Correlation Coefficient (MCC) for the SFS algorithm using the four different classification algorithms on 31,171 transcripts on the breast cancer cell line data obtained from Li et al., 2014. We calculated various metrics such as precision, recall, specificity and MCC due to the class imbalance of up/down expressed samples. We found that the 1NN-SFS algorithm resulted in the highest MCC, recall and specificity, and the third highest precision (Table 2.1, Fig. 2.1).

Table 2.1. Mean performance of SFS algorithms and controls on the breast cancer cell line data.

	1NN	**3NN**	**5NN**	**Bayes**	**DT**	**SVM**	**1NN Random**	**1NN Top Two**
Accuracy	0.79	0.80	0.74	0.78	0.77	0.74	0.66	0.67
Precision	0.70	0.74	0.65	0.75	0.68	0.64	0.53	0.54
Recall	0.68	0.65	0.53	0.59	0.65	0.63	0.53	0.53
Specificity	0.70	0.67	0.56	0.62	0.66	0.63	0.55	0.55
MCC	0.40	0.40	0.16	0.35	0.33	0.26	0.08	0.08

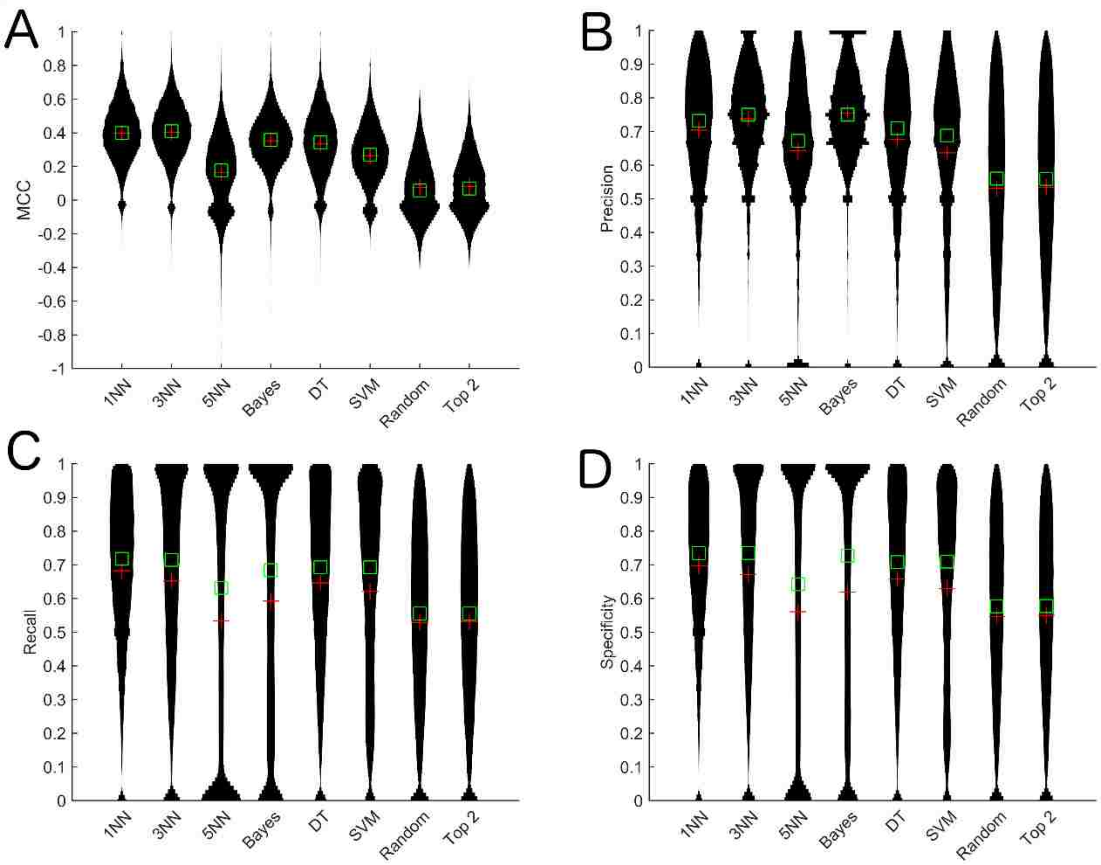


Figure 2.1. Violin plots of performance metrics for the algorithm when utilizing different classification methods in the SFS algorithm and controls on the breast cancer cell line data. A) MCC, B) Precision, C) Recall, D) Specificity. Green squares specify the median and the red pluses specify the mean. Bayes: Naive Bayes, DT: Decision tree, SVM: Support Vector Machine.

The 1NN algorithm also resulted in the second highest accuracy (Fig. 2.2A). We compared the 1NN-SFS algorithm to the random and top two correlated selection methods and evaluated the predictive performance of the probe selection with a 1NN-based LOO-CV. To ensure a fair comparison, we set the number of probes selected for a gene in the 1NN-Random method equal to the number of probes selected for that gene in the 1NN-SFS algorithm. We found that all these controls resulted in worse performance than our algorithm (Fig. 2.1., Fig. 2.2A).

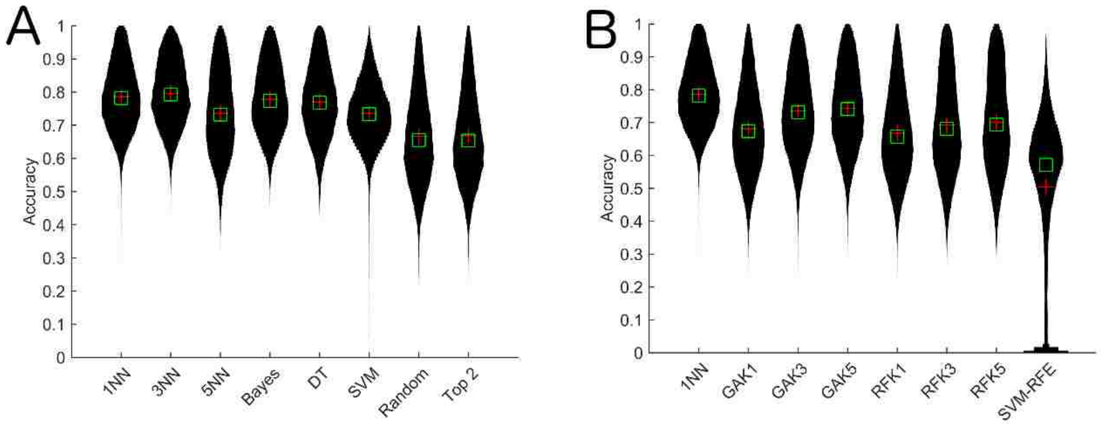


Figure 2.2. Violin plots of accuracy. A) SFS algorithms using various classification algorithms, B) GA and ReliefF algorithms.

We also compared 1NN-SFS algorithm to GA-KNN and ReliefF algorithms for K=1, 3 and 5, and to the SVM-RFE algorithm. We set the number of top ranked probes selected in ReliefF and SVM-RFE equal to the number of probes selected by 1NN-SFS. We observed that the 1NN-SFS algorithm performed better than GA-KNN and ReliefF algorithms for K=1, 3 and5, and the SVM-RFE algorithm by all metrics (Fig. 2.3, Fig 2.2B). Taken together, these results indicate that the 1NN-SFS feature selection method chooses more relevant probes than other algorithms.

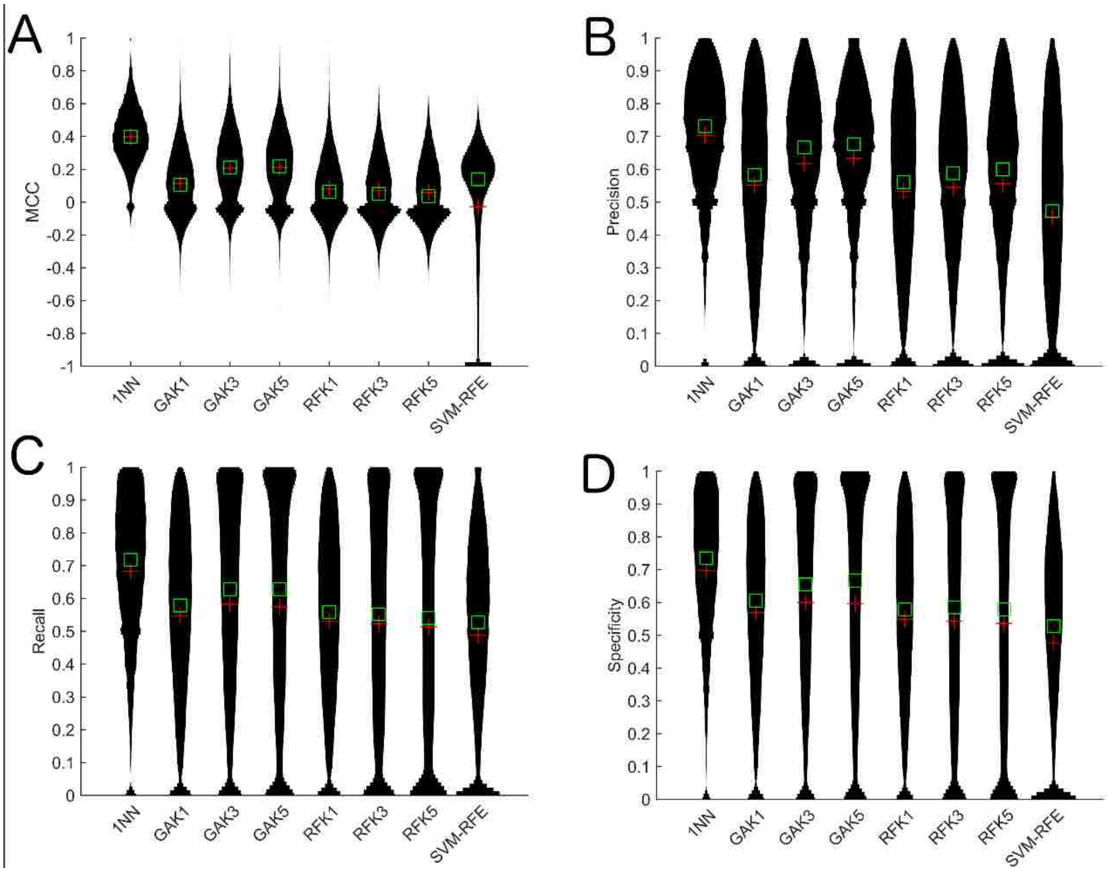


Figure. 2.3. Violin plots of performance metrics for 1NN-SFS algorithm against other algorithms on the breast cancer cell line data. A) MCC, B) Precision, C) Recall, D) Specificity. Random: KNN random, Top 2: KNN top two (see Methods). GAK: GA-KNN algorithm with varying K-nearest neighbors. RFK: Relief-F algorithm with varying K nearest neighbors.

We compared the 1NN-SFS algorithm to probe selection methods that do not consider gene expression. These approaches to probe selection resulted in significantly lower performance when compared to the 1NN-SFS algorithm that integrate gene expression, suggesting the importance of integrating gene expression data to inform the probe selection (Fig. 2.4).

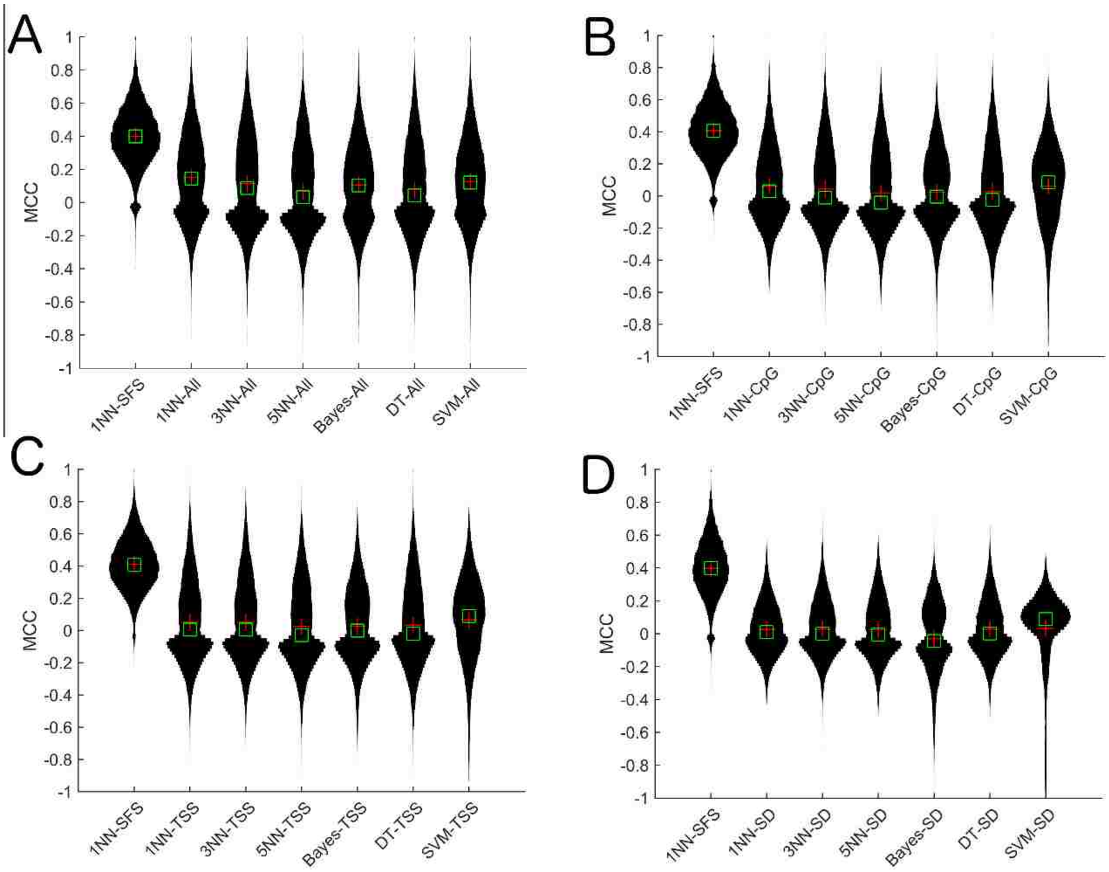


Figure 2.4. Violin plots of MCC for 1NN-SFS algorithm against other probe selection methods on the breast cancer cell line data. A) All, B) Upstream CpG Island, C) TSS, D) Top SD.

We observed the 1NN algorithm usually only selected one or two probes per gene (Fig. 2.5). Out of the 31,171 transcripts tested, 11,833 transcripts had one probe selected and an additional 9,411 transcripts had two probes selected. Since selecting all the probes (no feature selection) leads to significantly poorer performance, the selection of the best one or two probes is important to the algorithm’s good performance. This shows that our algorithm was able to reduce the number of probes for a given gene to a limited number of key informative probes.

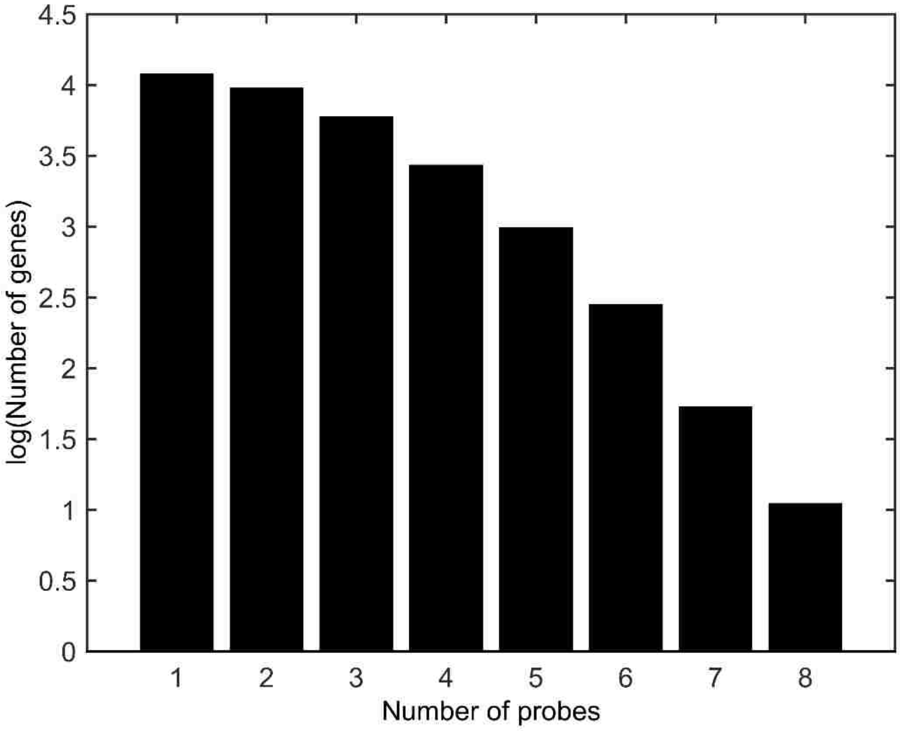


Figure 2.5. Number of probes selected per gene by 1NN-SFS algorithm on the breast cancer cell line data.

2.3.2. KNN algorithm resulted in consistent prediction accuracy

To check the consistency of the algorithm on smaller subsets of the data, we ran the algorithm five additional times on half of the dataset, in which the samples were randomly chosen each execution. For each of the five executions, we compared 1NN-SFS algorithm to random selection method and top two correlated method. Fig. 2.6 shows a heatmap comparison of the MCC for the five runs of the 1NN algorithm compared to the random selection and top two correlated selection. The 1NN consistently gave higher MCC values over the random selection and top two correlated selection. Additionally, the MCC values were consistent across runs.

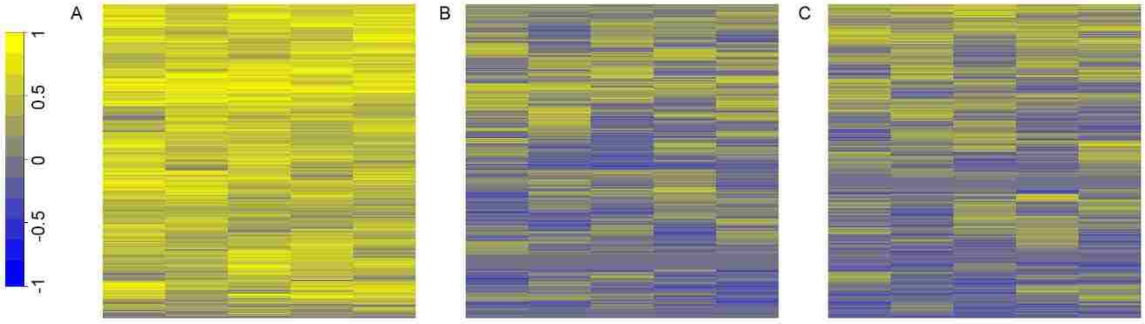


Figure 2.6. Heatmap clustering of MCC values. Heatmap clustering of MCC values for five executions of the algorithm on random halves of the breast cancer cell line data for A) 1NN algorithm and B) random selection of probes C) Top two correlated approach.

2.3.3. DNA methylation-sensitive genes were enriched for regulation-based GO terms

We investigated if there are any common functional property on genes whose transcription levels are sensitive to DNA methylation changes by analyzing genes where the selected probes predict gene expression well. 3,084 genes had MCC > 0.6 in the 1NN-SFS algorithm. The GOrilla results are summarized in Table 2.2, showing that DNA methylation-sensitive genes were enriched for GO terms related to the regulation of various biological processes. The table encompasses only the top 30 significant GO terms.

Table 2.2. Top 30 GO Terms for genes with MCC >0.6 by 1NN-SFS algorithm on the breast cancer cell line data.

Description	FDR q-value
regulation of multicellular organismal process	4.43E-19
regulation of developmental process	2.51E-17
regulation of multicellular organismal development	9.31E-17
positive regulation of biological process	1.16E-16
movement of cell or subcellular component	1.23E-16
positive regulation of cellular process	1.41E-16
negative regulation of biological process	2.3E-16
anatomical structure development	1.38E-15
negative regulation of cellular process	2.72E-15
regulation of cell differentiation	2.85E-15
cell migration	6.81E-15
negative regulation of metabolic process	2.48E-14
anatomical structure morphogenesis	3.53E-14
organ development	5.3E-14
transmembrane receptor protein tyrosine kinase signaling pathway	6.02E-14
cell motility	7.21E-14
Locomotion	1.7E-13
developmental process	1.71E-13
enzyme linked receptor protein signaling pathway	1.75E-13
single-organism developmental process	1.76E-13
regulation of cell development	2.88E-13
regulation of anatomical structure morphogenesis	4.5E-13
negative regulation of macromolecule metabolic process	6.04E-13
intracellular signal transduction	8.58E-13
single-multicellular organism process	2.36E-12
multicellular organismal process	5.86E-12
regulation of localization	1.06E-11
positive regulation of multicellular organismal process	1.07E-11
signal transduction	1.27E-11
cellular component organization or biogenesis	1.39E-11
positive regulation of developmental process	3.2E-11

To verify that this result is specific to well-predicted genes, we compared the result to poorly-predicted genes. We performed GO analysis on 2,880 genes that have MCC < 0.2. We chose MCC thresholds carefully to ensure a fair comparison to GO analysis by having comparable gene set sizes. Table 2.3 shows that only immune response and stimulus detection terms are reported as significant. This result suggests that enrichment of regulation-related GO terms is specific to genes with high MCC values.

Table 2.3. GO terms with MCC < 0.2 for genes by 1NN-SFS algorithm on the breast cancer cell line data.

Description	FDR q-value
detection of chemical stimulus involved in sensory perception of smell	5.62E-11
detection of chemical stimulus involved in sensory perception	5.74E-11
detection of chemical stimulus	5.62E-8
detection of stimulus involved in sensory perception	1.07E-7
detection of stimulus	1.95E-3
immune response	1.23E-2

We applied DAVID's functional classification tool on genes with MCC > 0.6 to determine functional enrichment differences for genes with selected gene body probes and genes with selected promoter probes. 1035 genes had exclusively upstream probes selected, resulting in 33 functional clusters. 699 genes had exclusively gene body probes selected, resulting in 27 functional clusters. We found that in both the promoter and gene body group, many of the clusters suggested that the genes are involved in the regulation of other genes and proteins via a variety of mechanisms. The most enriched clusters are shown in Tables 2.4 and 2.5.

Table 2.4. Functional clusters of genes with MCC > 0.6 with upstream probes selected by 1NN-SFS algorithm on the breast cancer cell line data.

Cluster Num	Size	Enrichment	Most significant terms (p-val)	Other representative terms (p-val) and notes
1	40	4.39	Atp-binding (4.4E-45), Nucleotide-binding (4.6E-38), adenyl ribonucleotide binding (1.7E-37)	Helicase (4E-12), kinase (5.8E-6), protein kinase activity (3.7E-4)
2	4	3.67	Repeat:ANK 1 (1.7E-6), Repeat:ANK 2 (1.8E-6), Ankyrin (2.9E-6)	Genes coding for ankyrin proteins
3	45	3.46	Kinase (1.8E-56), Protein Kinase – ATP binding site (2.0E-56), domain: protein kinase (2.1E-53)	Phosphorylation (1.7E-51), transferase (1.1E-47), nucleotide binding (2.1E-34)
4	13	3.42	Microtubule cytoskeleton (9.6E-15), cytoskeleton (9.1E-14), cytoskeletal part (4.1E-12)	Centrosome (2.3E-8), genes involved in regulation of cell motility
5	5	3.18	Nucleolus (8.8E-6), nuclear lumen (1.6E-4), intracellular organelle lumen (3.7E-4)	Membrane enclosed lumen (4.4E-4)
6	6	3.05	Regulation of actin filament polymerization (8.4E-13), regulation of actin filament polymerization or depolymerization (1.6E-12), regulation of actin filament length (1.9E-12)	Regulation of protein complex assembly (1.2E-11), negative regulation of actin filament depolymerization (6.3E-11)

7	4	2.91	binding site:S-adenosyl-L-methionine (1.8E-8), s-adenosyl-l-methionine (1.5E-7), methyltransferase (4.3E-7)	Genes coding for methyltransferases
8	5	2.83	Microfilament motor activity (22.0E-12), actin filament-based movement (6.3E-12), domain:Myosin head-like (9.4E-12)	Genes coding for myosin proteins
9	6	2.66	Anti-apoptosis (7.8E-12), negative reglation of apoptosis (1.2E-8), negative regulation of programmed cell death (1.3E-8)	Genes predominately related to BCL2 (BAG3, BAG4, BCL2A1, BL210). Also includes MCL1 and TNFRSF10D
10	16	2.54	Nucleotide phosphate-binding region:GTP (4.7E-28), gtp-binding (2.3E-27), Ras (2.7E-16)	Genes predominately related to the RAS oncogene family
11	13	2.48	Mitosis (2.5-22), nuclear division (2.5E-22), M phase of mitotic cell cycle (3.2E-22)	Organelle fission (4.1E-22), cell division (1.3E-17)
12	8	2.38	Guanine-nucleotide dissociation stimulator, CDC4, conserved site (1E-14), guanyl-nucleotide exchange factor activity (1.6E-14), Dbl homology (DH) domain (2.8)	Regulation of Ras protein signal transduction (2.0E-13), regulation of small GTPase mediated signal transfuction (7.2E-13), regulation of Rho protein signal transduction (9.2E-12)
13	59	2.29	Transcription regulator activity (2.7E-50), transcription regulation (2.2E-47), regulation of transcription, DNA dependent (2.2E-47)	Sequence specific DNA-binding (3.1E-29), repressor (6.0E-22)
14	8	2.26	LIM domain (6.9E-18), Zinc finger, LIM-type (2.3E-17), zinc (2.5E-7)	Metal-binding (2.1E-6)
15	5	2.23	ABC transporter-like (9E-8), ABC transporter, conserved site (1.5E-7), ATPase activity (4.2E-7)	Members of ATP-binding cassette sub-family (ABC)
16	4	1.85	Negative regulation of translation (1.5E-8), translation regulation (4.0E-8), mRNA 5'-UTR binding (2.7E-7)	Insulin-like growth factor 2 (IGF2) mRNA binding proteins
17	5	1.84	Protein tyrosine phosphatase activity (3.9E-9), tyrosine-specific phosphatase (1.5E-8), dephosphorylation (1.6E-8)	Phosphatases
18	7	1.73	Purine ribonucleoside triphosphate biosynthetic process (1.1E-10), purine nucleoside triphosphate biosynthetic process (1.1E-10) ribonucleotide triphosphate biosynthetic process (1.1E-10),	Various ATPase coding genes
19	10	1.72	Ribosomal protein (6.7E-19), structural constituent of ribosome (8.2E-18), cytostolic ribosome (1.6E-17)	Genes coding for ribosomal proteins
20	14	1.62	Wd repeat (8.2E-25), WD40 repeat (3.3E-24), WD40 repeat, conserved site	Genes with WD domain.

Table 2.5. Functional clusters of genes with MCC > 0.6 with gene body probes selected by 1NN-SFS algorithm on the breast cancer cell line data.

Cluster Num	Number of genes	Enrichment	Most significant terms (p-val)	Other representative terms (p-val) and notes
1	48	2.84	Atp-binding (1.1E-51), Nucleotide-binding (6.5E-47), adenyl ribonucleotide binding (4.2E-45)	phosphorylation (4.8E-33), kinase (7.6E-40), transferase(1.9E-29)
2	12	2.36	Nucleolus (1.2E-14), nuclear lumen (3.9E-11), intracellular organelle lumen (3.7E-10)	
3	11	2.06	Transcription regulation (1.6E-10), transcription(2.1E-10), regulation of transcription (6.8E-8)	
4	9	1.83	Ribosomal protein (7.2E-17), ribonucleoprotein (1.8E-15), ribosome (5.6E-15)	RNA binding (2.8E-4)
5	8	1.64	Cytoskeleton (1.7E-7), microtubule cytoskeleton (2.8E-6), intracellular non-membrane-bounded organelle (1.4E-5)	
6	9	1.62	GTP-binding (6.7E-15), guanyl nucleotide binding (5.2E-13), small GTPase mediated signal transduction (2.8E-12)	Ras oncogene related genes (RHOF, RAB3B, RAB3D, NKIRAS2, ERAS)
7	7	1.47	RNA-recognition motif, RNP-1 (3,8E-12), nucleotide-binding, alpha-beta plait (4.1E-12), RNA binding (4.8E-10)	RNA binding proteins and ribonucleoproteins
8	6	1.39	Negative regulation of ubiquitin-protein ligase activity during mitotic cell cycles (2.2E-12), negative regulation of ubiquitin-ligaase activity (2.6E-12)	Genes coding for proteasomes and ubiquitin
9	66	1.38	Regulation of transcription (1.1E-34), transcription (2.4E-24), transcription regulation (5.0E-32)	
10	6	1.28	Homeobox (30E-10), Homebox, conserved site (5.0E-10), homeodomain-related (5.7E-10)	Homeobox proteins
11	7	1.28	Tpr-repeat (3.0E-13), tetratricopeptide-like helical (6.4E-13), tetratricopeptide region (3.9E-8)	
12	4	1.17	SNF2-reated (6.4E-9), domain:Helicase C-terminal (1.6E-7), domain:Helicase ATP-binding (1.9E-7)	Chromodomain helicase DNA binding protein family

13	4	1.09	Protein import into nucleus, docking (1.6E-19), nuclear pore (2.3E-7_ nuclear import (2.7E-7)	Exportin 1, nucleoporin, transportin 2, importin 5

For genes with probes selected from the promoter regions (Table 2.4), the most enriched cluster contains genes involved in ATP-binding, nucleotide-binding, helicase and protein kinase activity. Additionally, cluster 3 also contains many kinase, phosphorylation and nucleotide binding terms. A common theme is that these terms are all mechanisms by which other genes and proteins can be regulated. Importantly, these functions may be related to the regulation-based GO terms represented in the GOrilla analysis. Other possible mechanisms of regulation of other genes and proteins include an enrichment of DNA-methyltransferases (cluster 7) and regulation of protein synthesis via ribosomal protein (cluster 19). DNA methylation may also play a role in the regulation of apoptosis-related genes (cluster 9) and cell motility (cluster 8). A group of 59 genes were enriched in terms related to transcription regulator activity (cluster 13).

Similar results were obtained for genes where the probes were selected from gene body regions (Table 2.5). The first and third cluster involve transcription regulation and protein kinase activity. Cluster 4 contains additional genes coding for ribosomal proteins. Cluster 8 contains genes coding for proteasomes and ubiquitin, suggesting that protein degradation may also be under the control of DNA methylation of certain genes. Additionally, 66 genes were enriched in terms related to transcription regulation (cluster 9).

Together, these results suggest that if DNA methylation is a good predictor of gene expression (MCC > 0.6) than that gene may likely be involved in the regulation of other genes and proteins through a variety of mechanisms including DNA binding, protein kinase activity, protein synthesis and protein degradation. We did not find a significant functional difference between genes where gene body probes are selected and genes where upstream probes are selected. This suggests that a gene under strong epigenetic control via DNA methylation is more likely to be a regulatory gene, regardless of the genomic position of the predictive DNA methylation.

2.3.4. Verification in TCGA luminal A breast cancer data

To verify our work in another dataset, we performed the 1NN-SFS algorithm on 99 luminal A breast cancer samples from the TCGA database. We computed the performance metrics, and found the average to be 0.7 for all metrics (Fig. 2.7).

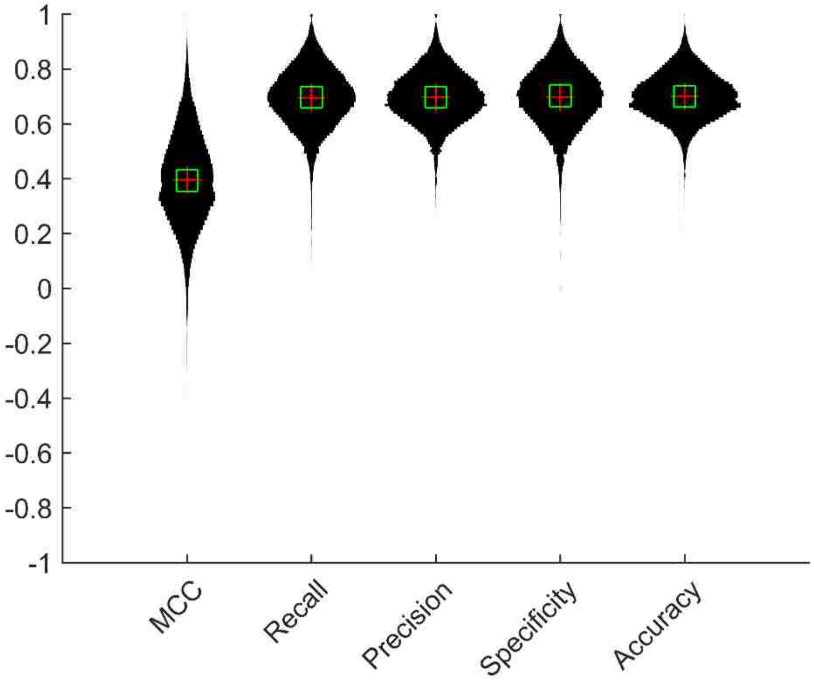


Figure 2.7. Performance metrics of 1NN-SFS algorithm on TCGA data.

We performed the same GO-term analysis for luminal A data that we performed in the cell line data. We chose 1,823 and 1,407 genes that were predicted with an MCC > 0.6 and MCC < 0.2, respectively. 534 of the genes with MCC > 0.6 in the TCGA data overlapped with the genes with MCC > 0.6 in the cell line data (hypergeometric p-value < 2.01 e-41). Table 6 shows only the top 30 GO terms for genes with high MCC and Table 7 shows all of the GO terms for genes with low MCC. Similar to our previous result for the cell line data, we found that genes that predicted well were again enriched in GO-terms related to the regulation of various biological processes while genes that were predicted poorly were not. We note here that the poorly-predicted genes had GO-terms involved in the detection of a chemical stimulus and smell. This was due to a single family (olfactory receptor family) where almost all of the members of the family had their expression predicted poorly. This was not the case for the regulation-based terms in the well-predicting gene set.

Table 2.6. Top 30 GO terms with MCC > 0.6 for genes by 1NN-SFS algorithm on TCGA data.

Description	FDR q-value
positive regulation of cellular process	3.75E-8
positive regulation of biological process	2E-7
RNA metabolic process	3.6E-7
regulation of metabolic process	7.55E-7
regulation of transcription from RNA polymerase II promoter	8.6E-7
cellular macromolecule metabolic process	9.69E-7
regulation of gene expression	1.16E-6
regulation of macromolecule metabolic process	1.19E-6
regulation of macromolecule biosynthetic process	1.35E-6
regulation of cellular macromolecule biosynthetic process	1.36E-6
RNA biosynthetic process	1.45E-6
regulation of primary metabolic process	1.54E-6
regulation of biosynthetic process	1.56E-6
macromolecule metabolic process	2.36E-6
aromatic compound biosynthetic process	2.48E-6
regulation of cellular biosynthetic process	2.52E-6
positive regulation of RNA biosynthetic process	3.02E-6
regulation of RNA biosynthetic process	3.12E-6
nucleobase-containing compound biosynthetic process	3.4E-6
nucleic acid metabolic process	3.44E-6
regulation of cellular metabolic process	3.45E-6
regulation of transcription, DNA-templated	3.63E-6
cellular process	3.73E-6
heterocycle biosynthetic process	3.93E-6
cellular nitrogen compound biosynthetic process	4.29E-6
positive regulation of macromolecule biosynthetic process	4.35E-6
regulation of nucleic acid-templated transcription	5.11E-6
nucleobase-containing compound metabolic process	6.76E-6
regulation of nucleobase-containing compound metabolic process	1.04E-5
positive regulation of RNA metabolic process	1.07E-5

Table 2.7. GO terms with MCC < 0.2 for genes by 1NN-SFS algorithm on TCGA data.

Description	FDR q-value
detection of chemical stimulus involved in sensory perception	1.27E-42
detection of chemical stimulus	6.29E-41
detection of chemical stimulus involved in sensory perception of smell	8.16E-41
detection of stimulus involved in sensory perception	3.18E-38
detection of stimulus	7.93E-31
G-protein coupled receptor signaling pathway	1.44E-21
sensory perception of smell	1.16E-19
sensory perception of chemical stimulus	4.86E-14
cell surface receptor signaling pathway	7.68E-7
sensory perception	7.02E-6
response to stimulus	5.47E-5
drug metabolic process	4.77E-3
signal transduction	1.12E-2

We performed DAVID’s functional classification analysis on genes with probes exclusively selected from the promoter and genes with probes exclusively selected from the gene body as previously described. 659 genes with MCC > 0.6 contained selected probes exclusively from the upstream regions, resulting in 22 total clusters. 396 genes with MCC > 0.6 contained selected probes exclusively from the gene body, resulting 23 clusters. For genes with selected probes from the promoter (Table 2.8), cluster 2 contained genes involved with RNA splicing, which is another mechanism by which other genes can be regulated. Similar to functional clustering results on cell line data, cluster 4 contained genes coding ribosomal proteins and cluster 1 and 5 contained transcriptional regulation genes. For genes with probes selected from the gene body (Table 2.9), clusters 1 and 3 had terms involved with protein regulation and cluster 2 contained genes involved with nucleotide-binding. For both the cell line and TCGA data for genes with selected gene body probes, chromodomain helicase and GTP-binding clusters were observed.

Table 2.8. Functional clusters of genes with MCC > 0.6 with upstream probes selected by 1NN-SFS algorithm in TCGA data.

Cluster number	Number of genes	Enrichment	Top terms (pval)	Other representative terms and notes
1	5	4.73	Nucleolus (8.8E-6), nuclear lumen (1.6E-4), intracellular organelle lumen (3.7E-4)	Transcription, DNA-dependent (4.3E-2)
2	24	4.08	RNA splicing (1.0E-29), RNA processing (8.0E-29), mRNA processing (1.1E-28)	Spliceosome (6.8E-23), rna-binding (2.3E-10)
3	13	2.48	Cytoskeleton (1.5E-18), cytoplasm (7.2E-10), microtubule cytoskeleton (4.7E-9)	
4	11	2.25	Ribosomal protein (6.3E-21), ribonucleoprotein (3.5E-19), ribosome (1.5E-18)	Group of genes coding for mitochondrial ribosomal proteins
5	134	2.2	Transcription regulation (1.9E-45), zinc (4.1E-45), transcription (1.3E-43)	Transcription regulation
6	13	2.03	Ubl conjugation pathway (1E-19), modification-dependent protein catabolic process (3E-17), modification-dependent macromolecule catabolic process (3E-17)	Ubiquitin proteins, proteolysis (4.7E-14)
7	5	1.84	Repeat: ANK1 (2.1E-8), repeat ANK2 (2.1E-8), ank repeat(2.4E-8	Ankyrin proteins
8	9	1.68	Mitosis (5.8E-17), cell division (1.1E-15), nuclear division (4.3E-15)	
9	9	1.4	Repeat:WD3 (1.2E-15), repeat:WD 2 (1.6E-15), repeat: WD1 (1.6E-15)	WD containing proteins
10	6	1.17	Kelch repeat (1.7E-10), repeat:Kelch 4 (8.2E-10), repeat:Kelch 1 (8.7E-10)	
11	4	1.13	Aminoacyl-tRNA synthetase (7.7E-9), tRNA aminoacylation (3.7E-8), amino acid activation (3.7E-8)	tRNA synthetases
12	4	1.09	Protein tyrosine phosphatase (1.1E-7), protein tyrosine phosphatase, active site (1.5E-7), protein tyrosine phosphatase activity (5E-7)	Protein tyropsine phosphatases
13	19	1.02	Transport (7.2E-14), mitochondrial envelope (4.3E-13), mitochondrion (5.8E-13)	

Table 2.9. Functional clusters of genes with MCC > 0.6 with gene body probes selected by 1NN-SFS algorithm in TCGA data.

Cluster Number	Number of genes	Enrichment	Most significant terms (p-val)	Other representative terms (p-val) and notes
1	4	3.3	GTPase activation (5.5E-7), domain:PH (1.9E-6), Pleckstrin homology (4.5E-6)	Rho GTPases
2	5	2.5	Atp-binding (2.2E-5), nucleotide-binding(5.9E-5), adenyl ribonucleotide binding (1.8E-4)	
3	17	2.14	Protein kinase – core (8.7E-23), kinase (2.7E-21), protein kinase – atp binding site (1.2E-20)	Phosphorylation (1.9E-20), nucleotide-binding (1.9E-15), transferase (7.3E-16)
4	5	1.86	Zinc (1.7E-4), metal-binding (5.7E-4), zinc ion binding (1E-3)	
5	5	1.59	GTP-binding (8.4E-8), guanyl nucleotide binding (7.4E -7), guanyl ribonucleotide binding (4.7E-7)	
6	4	1.59	Guanine nucleotide dissociation stimulator, CDC24, conserved site (5.2E-8), Dbl homology (DH) domain (6.8E-8), Rho guanyl nucleotide exchange factor activity (1.8E-7)	Regulation of apoptosis (2.1E-4)
7	4	1.41	EGF-like, type 3 (1.6E-6), egf-like domain (1.7E-6), EGF-like (1.7E-6)	
8	5	1.26	DNA/RNA helicase (4.7E-8), domain:Helicase C-terminal (6.4E-7), Helicase:ATP-binding (7.4E-7)	Chromodomain helicases
9	8	1.18	Repeat:WD 3 (9.1E-14), repeat:WD2 (1.2E-13), repeat:WD1 (1.2E-13)	WD containing proteins
10	4	1.17	Nucleoplasm (3.3E-4), transcription regulation (1.2E-3), transcription (1.2E-3)	

2.4. Conclusions

We developed an algorithm, which utilizes different classification and regression methods to select DNA methylation probes from the Illumina Infinium HumanMethylation450

BeadChip Kit array that are most relevant to expression of their corresponding gene. We tested the algorithms based on their ability to predict up/down expressed samples. We found that the 1NN-SFS algorithm performed the best compared to other methods tested (Fig. 2.1-2.3) and random selection (Fig. 2.1). We demonstrated that this algorithm led to consistent results (Fig. 2.6). The 1NN-SFS has the advantages of selecting a certain number of probes as opposed to ranking the probes.

We also observed that genes whose expression was predicted by DNA methylation with high accuracy were enriched in GO terms related to the regulation of various biological processes in both datasets. The overlap between highly predicted genes in both datasets was also significantly higher. Genes whose expression was accurately predicted by DNA methylation may be more sensitive to changes in DNA methylation. Therefore, genes that are sensitive to changes in DNA methylation may be more likely to be involved in the regulation of various biological processes.

Additionally, functional clustering revealed that many genes that were sensitive to DNA methylation were regulators of other genes and proteins through a variety of mechanisms including DNA-binding, protein kinase activity, protein degradation and protein synthesis. These results suggest that these functions may answer how genes under the control of DNA methylation regulate the various biological processes. There were no significant differences in function between genes with gene body probes selected and genes with upstream probes selected. This suggests that genes under the control of DNA methylation are more likely to be a regulatory gene regardless of the genomic position of the most predictive DNA methylation.

To verify results on cell line dataset, we also applied 1NN-SFS on a breast cancer dataset obtained from TCGA. The overall prediction accuracy in breast cancer data was lower than the accuracy in cell line data (Fig. 2.1 and 2.7). This could be due to the heterogeneity of the tissue samples. The expression of the tissue samples might be affected by other factors such as copy number alteration and mixed cell population in the tissues. On the other hand, cell line data contain more homogenous cells in each sample.

These methods will help researchers evaluate which probes are most involved in gene expression and which genes are sensitive to changes in DNA methylation. Future work should be aimed at studying other DNA methylation platforms to find the best methods for choosing regions of where DNA methylation has a significant impact on gene expression. The ideas in this paper could be extended to bisulfite sequencing and other commonly used platforms. Methylation-seq data could work if the data is converted to segment data. Additionally, the combinatorial effects of DNA methylation in different regions on gene expression can be studied with approaches similar to methods here.

CHAPTER 3

ProcessDriver: A computational pipeline to identify copy number drivers and associated disrupted biological processes in cancer

A manuscript on the project described in this chapter was submitted to a peer-reviewed journal.

Abstract: Copy number amplifications and deletions that are recurrent in cancer samples harbor genes that confer a fitness advantage to cancer tumor proliferation and survival. One important challenge in computational biology is to separate the causal, driver genes from passenger genes in large, aberrated regions. Many previous studies focus on the genes within the aberration (i.e., cis genes), but do not utilize the genes that are outside of the aberrated region and dysregulated as a result of the aberration (i.e., trans genes). We propose a computational pipeline, called ProcessDriver, that prioritizes candidate drivers by relating cis genes to dysregulated trans genes and biological processes. ProcessDriver assumes that a driver cis gene should be closely associated with the disrupted trans genes and biological processes, as opposed to previous studies that assume a driver cis gene should be the most correlated gene to the copy number of an aberrated region. We applied our method on breast, bladder and ovarian cancer data from the TCGA database. Our results included previously known driver genes and cancer genes, as well as potentially novel driver genes. Additionally, many genes in the final set of drivers were linked to new tumor events after initial treatment using survival analysis. Our results highlight the importance of selecting driver genes based on their widespread, downstream effects in trans.

3.1. Introduction

Copy number amplifications and deletions that are recurrent in cancer samples harbor driver genes that confer a fitness advantage to cancer tumor proliferation and survival (Hanahan and Weinberg, 2001). Passenger genes that do not have a selective advantage are amplified or deleted along with the drivers due to their proximity to the driver and as a result, have similar changes in expression with respect to copy number. Due to their similar copy number and expression profiles, separating drivers from passengers is an important and difficult challenge.

One of the tools to compute significant recurrent copy number alterations in a given set of samples is GISTIC. GISTIC relies on copy number data to detect regions of the genome that harbor likely drivers (Mermel et al., 2011; Beroukhim et al., 2007). GISTIC leveraged the notion that a region containing a driver gene should be altered significantly more than expected by chance. This method has proven useful in identifying regions that likely harbor candidate driver genes. However, it is difficult to distinguish passengers from drivers in large regions based on copy number data alone.

Some studies have integrated copy number and gene expression data to determine the effects of copy number on gene expression for genes within a copy number aberration, known as cis genes (Tamborero et al, 2013; Ambatipudi et al., 2011; Fan et al., 2012; Pickering, et al., 2013). The underlying assumption is that driver genes will have a more altered expression due to a copy number aberration than passenger genes. For example, Oncodrive-CIS is a method to score the cis genes as drivers by comparing the gene expression of samples with the aberration to the gene expression of samples without the aberration (Tamborero et al., 2013). The strength of the correlation between copy number and gene expression is also used to detect drivers (Fan et al, 2012; Pickering et al., 2013).

Some studies have identified drivers by considering the wider impact of a driver on downstream target genes located outside of the aberration, known as trans genes. For instance, Akavia et al. had the underlying assumption that copy number influences the driver gene expression, which in turn alters the expression of a group of downstream, trans genes (2010). Aure et al. determined which cis genes were highly correlated to their own copy number (2013). The authors then determined which of these cis genes played a network perturbing role in cancer through expression correlation to all other genes.

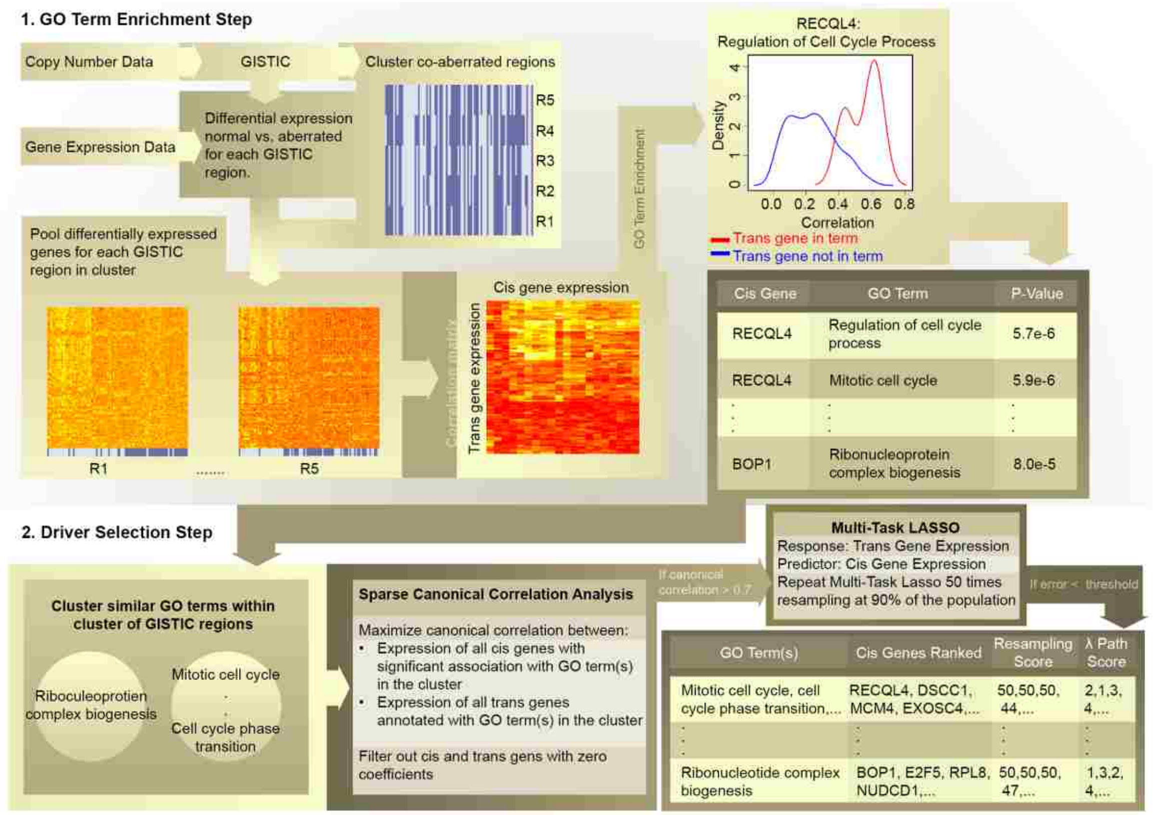
Certain biological processes are known to be disrupted in cancer such as apoptosis and cell cycle (Evan and Vousden, 2001). Therefore, identifying modules of cis and trans genes based on biological processes would allow for additional insight into the specific biological processes that the driver disrupts. Additionally, a driver cis gene changes the pathology of the cell and therefore influences the expression of many other genes in trans. Therefore, the cis genes in the module can also be narrowed down to a set of likely drivers based on the strength of the association of the cis genes with the downstream trans genes, as opposed to the strength of a cis gene's association with its own copy number.

In this study, we propose a pipeline called ProcessDriver that detects driver cis genes, associated trans genes and disturbed biological processes. We first find all of the differentially expressed cis and trans genes with respect to an aberration. For a given aberration, the pipeline creates modules of differentially expressed cis genes and differentially expressed trans genes based on biological processes. The module is subject to further refinements to determine likely drivers from the cis genes based on the relationship between cis gene expression and trans gene expression. The pipeline is therefore able to determine which biological processes and trans genes are dysregulated by the driver gene. We found that our selected drivers were more enriched in cancer genes and were associated with a higher risk of new tumor events after initial treatment. Additionally, consistent with previous studies, we found that the selected drivers were more correlated with their own copy number.

3.2. Materials and Methods

3.2.1 ProcessDriver

We implemented a computational pipeline called ProcessDriver in R to compute candidate copy-number driven driver genes by relating cis genes to dysregulated trans genes and biological processes. ProcessDriver utilizes gene expression, copy number alteration data and GO database. ProcessDriver consists of two main steps, namely GO term enrichment step and driver selection step. The entire pipeline of ProcessDriver is illustrated in Figure 1. In what follows, we describe each main step of ProcessDriver. ProcessDriver is licensed under MIT License and freely available upon request.


1. GO Term Enrichment Step
Copy Number Data
GISTIC
Cluster co-aberrated regions
Gene Expression Data
Differential expression normal vs. aberrated for each GISTIC region.
R5
R4
R3
R2
R1
Pool differentially expressed genes for each GISTIC region in cluster
R1
....
R5
Correlation matrix
Cis gene expression
Trans gene expression
GO Term Enrichment
RECQL4:
Regulation of Cell Cycle Process
Density
0
1
2
3
4
0.0
0.2
0.4
0.6
0.8
Correlation
Trans gene in term
Trans gene not in term
2. Driver Selection Step
Cluster similar GO terms within cluster of GISTIC regions
Riboculeoprotien complex biogenesis
Mitotic cell cycle
Cell cycle phase transition
Sparse Canonical Correlation Analysis
If canonical correlation > 0.7
Multi-Task LASSO
If error < threshold

Cis Gene	GO Term	P-Value
RECQL4	Regulation of cell cycle process	5.7e-6
RECQL4	Mitotic cell cycle	5.9e-6
. . .	. . .	. . .
BOP1	Ribonucleoprotein complex biogenesis	8.0e-5

Maximize canonical correlation between:
- Expression of all cis genes with significant association with GO term(s) in the cluster
- Expression of all trans genes annotated with GO term(s) in the cluster

Filter out cis and trans gens with zero coefficients

Response: Trans Gene Expression
Predictor: Cis Gene Expression
Repeat Multi-Task Lasso 50 times resampling at 90% of the population

GO Term(s)	Cis Genes Ranked	Resampling Score	λ Path Score
Mitotic cell cycle, cell cycle phase transition,...	RECQL4, DSCC1, MCM4, EXOSC4,...	50,50,50, 44,...	2,1,3, 4,...
. . .	. . .	. . .	. . .
Ribonucleotide complex biogenesis	BOP1, E2F5, RPL8, NUDCD1,...	50,50,50, 47,...	1,3,2, 4,...

Figure 3.1. Flowchart of ProcessDriver. In the GO term association step, cis and trans genes that were differentially expressed with respect to a copy number aberration were computed. Each cis gene was associated with up to ten biological processes by performing a Kolmogorov-Smirnov test using the correlation between the expression of the cis gene and every trans gene as a score. In the driver selection step, a GO term module containing similar GO terms and associated cis and trans genes was formed. The sparse CCA and multi-task LASSO were performed to narrow down potential drivers of the biological processes in the module from the cis genes.

3.2.1.1. GO term enrichment step

The GO term enrichment step first identifies differentially expressed cis and trans genes for a given aberration. Next, cis genes are associated with biological processes through the trans genes.

A. Computing GISTIC regions and differential expressed genes on GISTIC regions

GISTIC 2.0 was used to detect significant recurrent somatic copy number alterations (GISTIC regions hereafter) (Mermel et al., 2011). A GISTIC region with a log2 ratio above 0.1 was considered amplified, and a GISTIC region with a log2 ratio below -0.1 was considered deleted. A confidence level of 0.75 was used to calculate the GISTIC region. The differential expression analysis was performed using DESeq2 for each GISTIC region between samples with no significant deletions or amplifications versus amplified or deleted samples (Love et al., 2014) (p-value < 0.001). Genes were considered differentially expressed with respect to an aberration if their adjusted p-value was less than .001 in DESeq2 in one or more of the GISTIC regions within an aberration. Aberrations with greater than 50 differentially expressed genes were considered. These are aberrations of interest suitable for our algorithm because of the widespread effects of the aberration in trans, as well as the need to determine which cis genes are drivers. Batch effects were taken into account using the TCGA batch IDs as a covariate in DESeq2.

B. Clustering GISTIC regions into aberrations

To account for co-occurring aberrations, GISTIC regions were clustered together such that more similar regions were considered as a single aberration containing the individual GISTIC regions. Throughout the rest of the manuscript, a cluster of GISTIC regions will be referred to as an aberration. To cluster GISTIC regions into aberrations, a distance matrix was calculated where each entry was 1 minus the Pearson correlation of the copy number of two different GISTIC regions across all samples. Hierarchical clustering was performed on the distance matrix using average linkage using the stats package in R and the resulting dendrogram was cut at half of the maximum distance between the inter-cluster pairs.

The set of differentially expressed genes as determined by DESeq2 for each GISTIC region within the aberration were pooled together. Aberrations with greater than 50 differentially expressed genes were considered. These are aberrations of interest suitable for our algorithm because of the widespread effects of the aberration in trans, as well as the need to determine which cis genes are drivers. For each aberration, a differentially expressed gene is hereafter called cis gene if its chromosomal position was within a GISTIC region of that aberration, or called trans gene otherwise.

C. Computing aberration-adjusted expression

In the remaining steps of ProcessDriver algorithm, we related expression changes between cis genes and trans genes beyond the effects of copy number aberration. Both cis and trans genes expression are potentially under the influence of the copy number aberration of interest to varying degrees, and possibly other copy number aberrations in cis and trans. Due to the confounding effects of copy number aberration on gene expression, correlation between all gene expression will be high, making it difficult to establish relationships based solely on gene expression. To alleviate these copy number effects on gene expression, we computed aberration-adjusted expression. First we computed the variance stabilizing regularized log (rlog) transformation of the RNA-seq data. Then we applied principal component regression (PCR) between a gene's expression as a response and the copy number of all the GISTIC regions as predictors. The aberration-adjusted expression was the residual expression after PCR. We chose the PCR method as it is a suitable model to address the multicollinearity issue between the copy number of the GISTIC regions. All the remaining steps in ProcessDriver used the aberration-adjusted expression data.

D. GO term association

To link cis genes in aberrations to possible dysregulated biological processes in trans, each cis gene was associated with up to ten GO biological process terms through the trans genes. For a given aberration, the correlation between each cis gene’s expression and each of the trans gene’s expression in that aberration was calculated. A cis gene’s correlation to all trans genes was used as a score in a Kolmogorov-Smirnov (KS) test to determine significant GO terms using the TopGO package in R (Alexa and Rahnenfuhrer, 2010). The KS test examined whether trans genes annotated with a particular GO term were more correlated to the cis gene than trans genes not related to that GO term. KS test repeated for each cis gene in each aberration and up to ten GO terms with *p-value < .05* were chosen to be associated with each cis gene.

3.2.1.2 Driver selection step

The driver selection step clusters cis and trans genes to form modules based on associated biological processes. Next, expression data are utilized in a sparse canonical correlation analysis to filter cis and trans genes with canonical correlation greater than 0.7. Finally, cis genes are ranked as drivers using two multi-task LASSO-based methods.

A. Clustering of significant GO terms into GO term modules

Since some of the GO terms are semantically similar to each other and closely related in the GO term hierarchy, for each aberration, the set of GO terms associated with the cis genes were clustered using the getTermSim function with the relevance measure in the GOSim package in R (Frohlich et al., 2007). For each GO term cluster, we defined *GO term module* as the collection of cis genes that were significantly associated with at least one GO term in that GO term cluster, and the trans genes that were annotated with at least one GO term in that GO term cluster.

B. Applying sparse CCA to refine GO term modules

To further refine a GO term module to determine likely drivers, we performed sparse canonical correlation analysis (SCCA) between the expression of p cis genes and the expression of K trans genes (Witten et al., 2009). Let X_{ij} and Y_{ij} be the expression for patient i for cis and trans gene j, respectively. The goal of CCA is to maximize the canonical correlation between two groups of variables X and Y, by finding a linear combination Yu and Xv called canonical variates, where $u = (u_1, \dots, u_K),\ v = (v_1, \dots, v_p),$ are weight vectors (Hotelling, 1936).

$$\rho = \frac{v'X'Yu}{\sqrt{v'X'Xv}\sqrt{u'Y'Yu}} \tag{3.1}$$

SCCA maximizes this correlation while also applying penalties to *u* and *v* such that some of the weights become zero resulting in $q < p$ cis genes and $M < K$ trans genes (Witten et al., 2009).

If the canonical correlation was greater than 0.7, cis and trans genes that had non-zero coefficients were left in the GO term module while those with zero coefficients were removed. If the canonical correlation was less than 0.7, the module was no longer considered.

C. Applying multi-task LASSO to computer driver cis genes

Multi-task LASSO was performed with the expression of the remaining trans genes as a response and the expression of the remaining cis genes as the predictors in order to rank the cis genes based on their influence on trans gene expression. Let X and Y now represent the remaining *q* cis and *M* trans gene expression, respectively. Multi-task LASSO is the multi-response version of LASSO (Friedman et al., 2010). Friedman et al., defines the multi-task LASSO model [Eq. 3.2] for *q* cis genes, *M* trans genes and *N* patients as:

$$\min_{(\beta_0, \beta) \in \mathbb{R}^{(q+1) \times M}} \frac{1}{2N} \sum_{i=1}^{N} ||Y_{i,1:M} - \beta_0 - \beta^T X_{i,1:q}||_F^2 + \lambda \sum_{j=1}^{q} ||\beta_j||_2 \tag{3.2}$$

In Eq. 3.2, $Y_{i,1:M}$ is a vector corresponding to the expression values of the trans genes in patient *i* and $X_{i,1:q}$ is the covariate vector of cis genes. β_j is the *j*th row of the *q* x *M* coefficient matrix corresponding to *j*th cis gene and λ is the tuning parameter controlling the strength of the penalty.

We ranked cis genes as drivers based on the order of appearance of each of the cis gene predictors in the model as λ goes from largest to smallest. As λ gets smaller, more cis genes will be non-zero and included in the model. The multi-task sharing portion involves which variables are selected. For each variable, a separate coefficient is fit for each response, resulting in the *q + 1* x *M* coefficient matrix (Friedman et al., 2010). Therefore, for all the trans genes, the coefficient for a given cis gene is either zero or non-zero, although the value of the non-zero coefficients will vary between trans genes. Therefore, this ranking will be the same for every trans gene, regardless of the non-zero coefficient value for the included cis genes

As an additional ranking system, the multi-task LASSO was rerun fifty times, each time resampling 90% of the samples without replacement. For a single resample, the value of λ used was the simplest model where the cross-validation error was within one standard error of the minimum cross-validation error. The number of times a cis gene was selected out of fifty resamples was used as a system to rank cis genes within the module. This ranking system would identify potential drivers that are robust to sample variation.

3.2.2. Datasets to assess ProcessDriver

To assess the performance of ProcessDriver, we used Illumina HiSeq 2000 RNA sequencing and level 3 segmented copy number inferred from Affymetrix Genome-Wide Human SNP 6.0 copy number data were downloaded for 92 luminal A breast cancer samples, 120 ovarian cancer samples and 120 bladder cancer samples from the TCGA repository (The Cancer Genome Atlas Research Network, 2011, 2012, 2014).

3.3. Results

We downloaded RNA-seq and segmented copy number data from the TCGA repository for 92 luminal A breast cancer, 120 bladder cancer and 120 ovarian cancer samples. We used GISTIC 2.0 to identify recurrent copy number aberrated GISTIC regions using segmented copy number data from each cancer type and clustered them into aberrations (see Materials and Methods). For breast, ovarian and bladder cancer, 175, 116 and 156 GISTIC regions were clustered into 66, 82 and 79 aberrations, respectively. DESeq2 was used to compute differentially expressed cis and trans genes for each aberration. Tables 3.1-3.3 contain information about the cytoband locations and number of differentially expressed cis and trans genes for the aberrations considered in each cancer type.

Table 3.1. Ovarian cancer co-aberrated regions. Amplified (Amp) and deleted (Del) regions in each co-aberrated region with the number of cis genes and the number of trans genes.

Co-aberrated regions	**Num cis**	**Num trans**
Amp 1q24.2, Amp 14q11.2, Del 3p25.1	40	49
Amp 3p12.3, Amp 19q13.42	31	20
Amp 8q24.21, Del 8p23.3	102	287
Amp 10p12.1, Amp 20q13.33	23	50
Amp 10q21.3, Del 6q27	9	49
Amp 11q14.1	14	64
Amp 12p13.2, Del 10p11.23	42	30
Amp 14q32.33, Del 9q34.3	56	13
Amp 19p13.12, Amp 19q12	75	48
Del 4p16.3, Del 4q13.2	64	69
Del 4q34.1	7	68
Del 5p15.2	19	74
Del 5q11.2, Del 5q13.2	33	44
Del 7p22.1, Del 11p15.4	113	147
Del 8p23.1, Del 12q23.1	23	140
Del 9p24.3, Del 9p11.2	43	24
Del 12p13.33, Del 12p13.2	37	17
Del 13q13.1, Del 13q14.3	50	71
Del 14q24.3, Del 16p13.3	147	77
Del 15q14	8	75
Del 16q23.1, Del 18q22.2	23	141
Del 17p11.2	6	114
Del 19q13.31, Del 19q13.41, Del 19q13.42	192	106

Table 3.2. Bladder cancer co-aberrated regions. Amplified (Amp) and deleted (Del) regions in each co-aberrated region with the number of cis genes and the number of trans genes.

Co-aberrated regions	Num cis	Num trans
Amp 1q23.3, Amp 6p22.3, Amp 7p21.1, Amp 7p11.2, Amp 10p14	26	137
Amp 3p25.2, Amp 8q22.3	43	540
Amp 7q31.1, Del 3p22.2, Del 3p14.2, Del 3p12.3	59	133
Amp 12q15, Amp 19q13.42	47	52
Amp 18p11.32, Del 16p13.3, Del 16p12.2	53	68
Del 2q34, Del 2q37.3	40	31
Del 5q11.2, Del 5q31.3	78	86
Del 6q12, Del 6q27, Del 14q24.3	95	141
Del 9p21.3, Del 9p11.2, Del 9q22.33	63	989
Del 11p15.4, Del 11p11.12, Del 11q25	31	78
Del 13q13.1, Del 13q14.2, Del 13q14.3	32	273
Del 15q13.2, Del 15q24.3	47	42
Del 17p11.2, Del 17p11.2, Del 17p11.2	16	69

Table 3.3. Breast cancer co-aberrated regions. Amplified (Amp) and deleted (Del) regions in each co-aberrated region with the number of cis genes and the number of trans genes.

Co-aberrated regions	Num cis	Num trans
Amp 1p13.3, Amp 16q12.2, Del 1p13.3, Del 16q12.2	4	54
Amp 2p16.3, Amp 17q23.2, Del 4q34.1	26	126
Amp 5p15.2, Del 5p15.2, Del 5p15.1, Del 12q23.1	3	285
Amp 6p12.1, Amp 8p11.23	16	114
Amp 7p14.1, Del 7p22.1, Del 7q11.21, Del 7q34	15	112
Amp 8q12.1, Amp 8q12.3, Amp 8q22.1, Amp 8q23.3, Amp 8q24.21, Del 8q11.21, Del 8q13.3, Del 8q24.3	187	429
Amp 9p11.2, Amp 9q34.3, Del 9q34.3	3	118
Amp 11p15.1, Amp 19q13.41, Del 1p36.11	8	43
Amp 12p13.2, Del 12p13.33, Del 12p13.2, Del 12p13.2	18	52
Amp 13q21.33, Del 13q13.1	25	52
Amp 14q24.3, Del 3p21.1, Del 14q24.3, Del 17p12	30	234
Amp 16p13.3, Amp 16p11.1	175	62
Amp 17q21.31, Del 11p11.12, Del 17q21.31	17	187
Amp 20q13.33, Del 15q24.3, Del 20q13.2, Del 20q13.33	76	243
Amp 22q11.23, Del 22q11.21, Del 22q11.23, Del 22q11.23	35	124
Del 2q11.2, Del 2q11.2, Del 17p11.2, Del 17p11.2	18	91
Del 11q14.3, Del 11q22.3	16	117
Del 16q22.1, Del 16q23.1	96	287

For each cis gene in each aberration, associated dysregulated GO biological process terms were computed (Section 3.2.1.1). For each aberration, GO term modules were formed (Section 3.2.1.2A and then the cis and trans genes were filtered with SCCA (Section 3.2.1.2B). Finally, the cis genes were ranked as likely drivers with two multi-task LASSO-based ranking methods (Section 3.2.1.2C). The number of GO terms, and the average number of cis and trans genes per module before and after SCCA are summarized in table 3.4.

Table 3.4. Number of modules and average number of cis and trans genes per module before and after SCCA.

	Breast cancer	**Ovarian cancer**	**Bladder cancer**
Number of modules	188	119	140
Average of number of cis genes pre-SCCA	11.6	12.1	10.9
Average number of cis genes after-SCCA	6.1	5.6	6.4
Average of number of trans genes pre-SCCA	35.1	21.7	39.5
Average number of trans genes after-SCCA	17.3	11.3	19.2

In the following sections, to evaluate the performance of ProcessDriver, we categorize cis genes into various groups namely, multiple driver, driver, semi-driver, last in λ path, and filtered. A *driver gene* is a cis gene that was selected 50 out of 50 times during resampling of multi-task LASSO and appears as the first gene in the λ path in at least one GO term module. A *multiple driver gene* is a gene that was selected as a driver in more than one GO term module. A cis gene that is *last in λ path* is a gene that was selected last in λ path in every GO term module it appeared in. A *semi-driver* was never selected as a driver gene, but was not last in λ path in at least one module. A cis gene that in the *filtered* group was filtered because the canonical correlation of the GO term module was < 0.7 (Figure 3.2) or its coefficient was 0 in a GO term module with canonical correlation > 0.7, and otherwise never appeared in the multi-task LASSO phase (Section 3.2.1.2C).

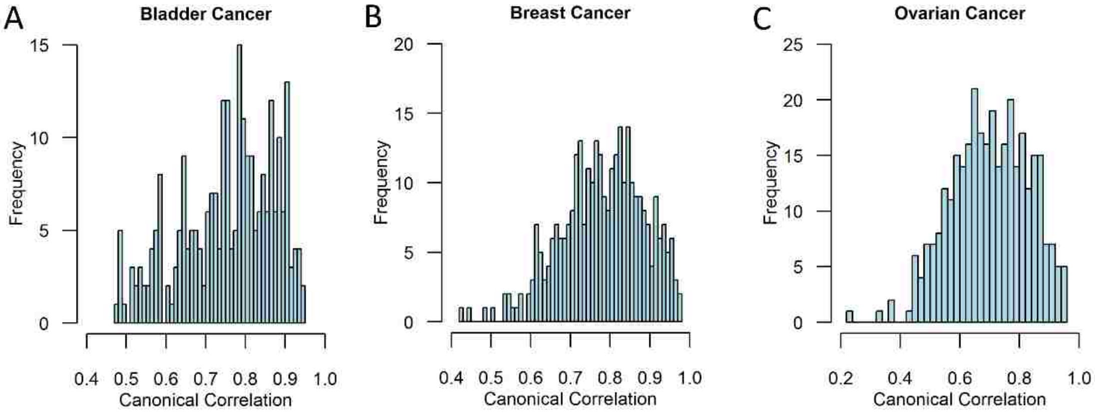


Figure 3.2. Histograms of canonical correlations for GO term modules in (A) bladder cancer, (B) breast cancer and (C) ovarian cancer.

For comparison purposes, we imitated some of the existing methods and selected drivers based solely on the magnitude of correlation between their gene expression and their copy number. For each GO module, cis genes with highest correlation between their expression and copy number were selected as *top correlated* group. This group served to highlight the differences between methods that consider the relationship between trans gene expression and cis gene expression to select drivers and existing methods that selected drivers based on gene expression correlation to cis copy number.

3.3.1. Multiple drivers are enriched in known cancer genes

Table 3.5 lists the entire multiple driver genes computed by ProcessDriver using breast cancer data and Tables 3.6 and 3.7 lists the multiple driver genes in ovarian and bladder cancer, respectively. For breast cancer, 19 out of 44 of the multiple driver genes were associated with cancer in the literature using the tool OncoSearch (Lee et al., 2014), as one or more publications describe their involvement in a cancer. Additionally, we found articles associating five more genes with cancer (Braig and Bosserhoff, 2013; Furic et al., 2012; Yamaguchi et al., 2014; Chen et al., 2013; Jensen et al., 2014). Seven multiple drivers were known cancer genes in the AGCOH or intOgen database (Huret et al., 2013; Gundem et al., 2010). Additionally, we used the BioGRID database to find genes the multiple driver interacts with and then determined which of the interacting genes are cancer genes in the AGCOH or intOgen database (Chatr-aryamontri et al., 2015). Overall, 26 out of the 44 breast cancer multiple drivers are a likely cancer gene or connected to a known cancer gene.

Table 3.5. Multiple driver genes associated with cancer in breast cancer. Number of articles is the number of articles found with OncoSearch tool plus additional literature references found manually. For cancer type (CT) column, BC – breast cancer, C – cancer, * indicates the multiple driver is cancer gene in AGCOH or intOgen databases.

Gene	**GO Terms**	**# Articles**	**CT**	**Cancer gene interactions**
AURKA	mitotic cell cycle, cell cycle	71	BC*	BRCA1, CDKN2A, TP53, CDC20, PLK1, TACC1, NIN, CHFR
SMARCB1	macromolecule metabolic process, RNA biosynthetic process	59	C*	AKT1, ARID1A, ARID1B, ATM, BRCA1, CCNE1, CREBBP, ING1, MLL, MLL3, MYC, NCOR1, SIN3A, SMARCA4, SMARCB1, TP53, XPO1, CDX2, GATA1, RELB, SS18, XPC, MLLT10, MCPH1
ADAM17	positive regulation of cellular process	32	BC	
TRADD	purine nucleoside metabolic process	10	C	CASP8, CAV1, TNF, BCL10
CUL5	carbohydrate metabolic process	6	BC	VHL, RNF7, RBX1
ELAC2	cellular component organization	5	C*	CUX1
PSMA7	mitotic cell cycle process	5	C	CUL1, EGFR, PLK1, TIMP2
RBM5	cellular response to endogenous stimulus	5	BC*	
COPS3	Cellular component organization,	4	C	COPS2, CUL1, CUL2, CUL3, HSP90AA1, DDB1, PTGS2
TBX21	T cell receptor signaling pathway	2	C	CREBBP, EP300, GATA3
APPBP2	cell cycle process, cellular protein localization	1	C	PCSK5, MLLT3
ARFGAP1	mitotic cell cycle process	1	C*	
BOP1	ribonucleoprotein complex biogenesis	1	C*	

DDT	macromolecule metabolic process	1	C	
HAGH	regulation of RNA metabolic process	1	C	
MED17	cellular response to DNA damage stimulus	1	C*	BRCA1, SMARCA4, TP53, BARD1, ESR2, GATA1, BRD4
PTDSS1	G2/M transition of mitotic cell cycle	1	C	
RBM38	regulation of protein metabolic process	1	C	
RRS1	mitotic cell cycle, regulation of protein complex assembly	1	C	
DIDO1	phosphorus metabolic process, phosphorylation	1	C	HNRNPK, WWP1
EIF4ENIF1	macromolecule metabolic process, RNA biosynthetic process	1	C	
DSCC1	mitotic cell cycle, cell cycle phase transition	1	C	
AZIN1	cellular cation homeostasis	1	C	FANCA, FANCC
BCL2L13	gene expression,	1	C	-
COG4	nucleobase-containing compound metabolic process	-	-	APC
PSMD7	cellular response to DNA damage stimulus	-	-	PSMD11

Table 3.6. Multiple driver genes associated with cancer in ovarian cancer. Number of articles is the number of articles found with OncoSearch tool plus additional literature references found manually. For cancer type (CT) column, OC – ovarian cancer, C – cancer, * indicates the multiple driver is cancer gene in the OCGene databases.

Gene	GO Terms	# Articles	CT	Cancer gene interactions
CASP3	macromolecule metabolic process, single-organism metabolic process	492	OC*	MCL1, BIRC2, BIRC3, CTTN, BIRC5, MAP3K14, BCL2, DCC, CASP10, CASP8, CFLAR, HSPD1, HSPE1, BIRC7, CASP3, XIAP
RAF1	phosphorus metabolic process, nitrogen compound metabolic process	192	C*	PRKCZ, SFN, PRKG1, BIRC2, BIRC3, HRAS, PAK1, RRAS2, KRAS, PEBP1, RB1, SPRY2, AKT1, MAP2K1, MAPK3, RBL2, MAPK7,…
XRCC1	heterocycle biosynthetic process, aromatic compound biosynthetic process	19	C*	CHD1L, PARP1, APEX1, PARP2, LIG3, TP53, PCNA, OGG1, POLB
ALKBH1	regulation of transcription, cellular response to stress	14	C	-
NUMB	cellular component biogenesis, protein complex assembly	2	C	MDM2, TP53, NOTCH1, L1CAM
ARHGAP 35	cellular component movement	2	C	RHOA
SPTLC2	organelle organization, cellular component assembly	2	C	-
MRPL36	protein modification process	1	C	-
UTP20	cell cycle, protein modification process	2	C	-
GSDMD	lymphocyte activation, response to cytokine	1	C	-
PHRF1	cellular metabolic process	1	C	-
PWP1	modification-dependent macromolecule cat...	1	C	-
DACT3	multicellular organismal process	1	C	-
RFC3	cell division, nuclear division	1	OC*	PCNA
STK33	biosynthetic process	1	C	-
SNW1	cellular response to stress	1	C	MEN1, VDR, RB1, SMAD3, RBL2, SMAD4, NOTCH3, NCOA1, RBL1, ASCC2, SKIL,…
MRPS31	macromolecule catabolic process,	-	*	EIF6
MED6	response to organic substance	-	-	VDR, ESR2, MED1, MED25, ESR1

Table 3.7. Multiple driver genes associated with cancer in bladder cancer. Number of articles is the number of articles found with OncoSearch tool plus additional literature references found manually. For cancer type (CT) column, BC – bladder cancer, C – cancer, * indicates the multiple driver is cancer gene in AGCOH or intOgen databases.

Gene	GO terms	Num Articles	CT	Cancer gene interactions
DEK	mitotic cell cycle, cell cycle phase transition	15	BC*	-
PPARG	ribonucleoside monophosphate metabolic process, DNA repair	214	C*	CREBBP, EP300, HDAC3, MED24, NFE2L2, RB1, PML, NCOA4, NCOA3
VHL	epithelial cell proliferation, nucleoside monophosphate catabolic process	244	C*	ATM, CUL2, EP300, FN1, HDAC3, IREB2, RASGRP1, TP53, RHOC, CSTB, FSCN1, RBX1
HSPA9	cellular response to stress	14	C	TP53
SDHC	cellular protein modification process	6	C*	-
NUMB	RNA metabolic process	2	C	MDM2, NOTCH1, TP53, L1CAM
FANCC	DNA metabolic process	1	C*	HSP90AA1, HSPA8, SPTAN1, FANCA, FANCE, FANCF, FANCG
ERH	viral process, symbiosis	1	C	HSPA8, SETDB1, TP53, SH3GL2
SNW1	macromolecule localization	1	C	MEN1, MLL, MYC, NCOR2, NOTCH1, RB1, SIN3A, SMAD2, SMAD4, NOTCH3, RBL2
DRG2	cellular protein localization, regulation of transcription from RNA polymerase	1	C	-
ANP32B	DNA-dependent DNA replication, RNA splicing	1	C	-
FEM1B	single-organism cellular process	1	C	-
ARPC2	cellular nitrogen compound biosynthetic process	1	C	CDH1, CTTN
CAB39L	mitotic cell cycle process	1	C	STK11
SYNJ2BP	protein modification by small protein conjugation	1	C	ACVR2A
HARS	translation	-	-	EEF1B2
MED6	cellular response to stress	-	-	ESR2

For ovarian cancer, 18 out of 33 multiple driver genes were associated with cancer through the literature or an interactor with a known cancer gene. Articles for nine genes were found with Oncosearch and supporting literature was found for seven more (Saeki et al., 2009; Ettahar et al., 2013; Honoré et al., 2002; Jiang et al., 2008; Shen et al., 2014; Scholl et al., 2009; Sato et al, 2015). The remaining two were found to have interactions with known cancer genes in the OCGene ovarian cancer database (Liu et al., 2015). For bladder cancer, 17 out of 26 multiple driver genes were a likely cancer gene or an interactor with one. Eight drivers had articles found by OncoSearch and supporting literature was found for seven more (Sato et al., 2015; Xu et al., 2016; Yang et al., 2016; Subauste et al., 2010; Rauhala et al., 2013; Choi et al., 2016; Lui et al., 2016). The remaining two had interactions with known cancer genes in the AGCOH or intOgen databases (Huret et al., 2013; Gundem et al., 2013).

Our methods associated cis genes with disrupted biological process in trans. Many of the multiple driver genes in all three datasets were appropriately associated with biological processes that they are known to be involved in. For example, in breast cancer, BOP1 is required for the maturation of ribosomal RNAs (Lapik et al., 2004) and was associated in our algorithm with "ribosome biogenesis" (Table 3.5). In ovarian cancer, candidate GSDMD is involved in the release of Interleukin 1-Beta, and was associated with out methods with "lymphocyte activation" and "response to cytokines" (Table 3.6). HSPA9 in bladder cancer is a heat shock protein and was associated "cellular response to stress" (Table 3.7). These genes and others are all involved in cancer, and are candidate copy number drivers and respective candidate disrupted processes.

In order to compute the enrichment of cis gene categories in known cancer gene lists, we created a list of cancer genes by combining 727 known cancer genes from the AGCOH database (Huret et al., 2013) and 475 known cancer genes from the intOgen database (Gundem et al., 2013). The overlap between all cis genes and the cancer gene list in ovarian cancer was poor (hypergeometric p -value = 0.28). Thus, for ovarian cancer, we used a more specific cancer list from the OCGene ovarian cancer database (Liu et al., 2015). The OCGene ovarian cancer database had a stronger, but marginal overlap with the cis genes (hypergeometric p-value = 0.09). Cis genes in breast and bladder cancer had sufficient overlap with the intOgen and AGCOH database (p-value = 0.0025 for bladder and 0.11 for breast cancer). We found that drivers and multiple drivers had lower p-values than genes that were filtered out by ProcessDriver and cis genes that were the most correlated with their own copy number (Table 3.8). Although some of the p-values were marginal, the enrichment for drivers and/or multiple drivers was higher than for cis genes that were filtered out. The marginal p-values could be due the incompleteness of the databases. As shown in Table 3.5-3.7, additional literature was found via a manual search for some multiple drivers supporting their involvement in cancer, despite not being present in the databases, yet.

Table 3.8. Enrichment of cis genes with known cancer genes. Hypergeometric p-values for the enrichment of known cancer genes in selected drivers, cis genes that were filtered out by ProcessDriver, and cis genes that were the most correlated with their own copy number.

Bladder Cancer		Multiple Driver (26)	Driver (89)	Semi-Driver (197)	Last in λ path (43)	Filtered (197)	Top Cor (120)
	AGCOH, IntOgen	**0.06**	0.12	0.86	0.52	0.77	0.6
Breast Cancer		Multiple Driver (44)	Driver (116)	Semi-Driver (266)	Last in λ path (51)	Filtered (259)	Top Cor (128)
	AGCOH, IntOgen	**0.01**	0.15	0.96	0.19	0.27	0.52
Ovarian Cancer		Multiple Driver (33)	Driver (82)	Semi-Driver (184)	Last in λ path (45)	Filtered (398)	Top Cor (138)
	OCGenes	0.25	**0.07**	0.35	0.71	0.85	0.8

3.3.2. SCCA filters cis genes with a lower correlation of expression to their own copy number

The underlying assumption in many previous studies on cancer drivers is that driver gene expression has a higher correlation to their own copy number than passenger genes (Tamborero et al., 2013). Although we did not use correlation of cis gene expression to its copy number to narrow down likely drivers, we expect that our drivers would have a higher correlation between their gene expression and copy number than the correlation of other genes' expression to their own copy number. Figure 3.3. illustrates the distribution of the correlation of cis copy number to gene expression in the different groups of cis genes for bladder and breast cancer data and Figure 3.4. shows the same distribution for the ovarian cancer data. Cis genes that were filtered by SCCA had a significantly lower average correlation of expression with copy number than driver genes in all three cancers (Wilcoxon rank-sum p-value < 0.001 for ovarian and breast cancer and < 0.05 for bladder cancer). We also observed that for cis genes filtered by SCCA, there were still genes with extremely high correlation between expression and copy number. These results suggest that utilizing correlation between gene expression and copy number to select potential driver genes could make false positive selections.

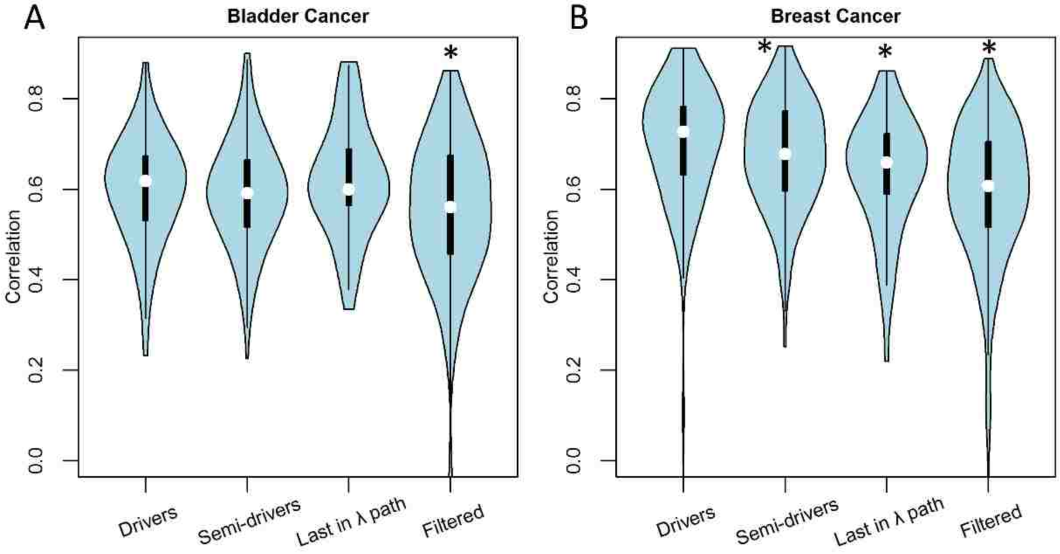


Figure 3.3. Correlation of copy number to cis gene expression. Violin plots representing the correlation of cis genes to their own copy number for selected drivers and cis genes filtered-out by ProcessDriver for (A) bladder cancer and (B) breast cancer. Definition of each group is in the results section. Asterisk indicates p < 0.05 in a Wilcoxon rank-sum test compared to the drivers group.

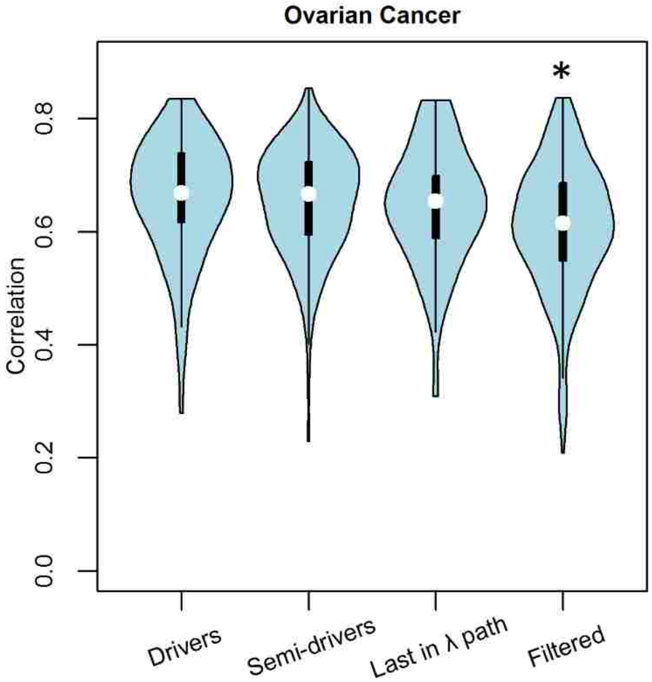


Figure 3.4. Correlation of copy number to cis gene expression in ovarian cancer. Violin plots representing the correlation of cis genes to their own copy number for selected drivers and cis genes filtered-out by ProcessDriver. Asterisk indicates $p < 0.05$ in a Wilcoxon rank-sum test compared to the driver group.

3.3.3. Driver genes are associated with a higher risk of new tumor events after initial treatment

In order to evaluate if the driver genes could predict new tumor events after initial treatment, we performed survival analysis on cis genes. We fit a univariate Cox proportional hazard model for each cis gene for the number of days to a new tumor event after the initial treatment and used the cis gene expression as a covariate. If a patient did not experience a new tumor event after the initial treatment, the days until the last follow-up were used and the patient was censored. In the bladder cancer cohort, 97 out of 120 patients have had new tumor events after the initial treatment and in the ovarian cancer cohort 86 out of 120 patients have had new tumor events. Only two out of 92 of the luminal A patients had new tumor events after initial treatment, therefore luminal A was not included in this analysis.

A hazard ratio > 1 implies that an increase of expression of the cis gene increases the risk of a new tumor event, while a hazard ratio < 1 implies that an increase of the cis gene expression decreases the risk of a new tumor event. Overall in bladder cancer, drivers had hazard ratios greater than one (Figure 3.5A). We compared the mean of the hazard ratios of each group using the Wilcoxon rank-sum test. We observed that the mean of the hazard ratios was significantly higher in the driver group compared to the top correlated, filtered and last in λ path groups with $p < 0.05$.

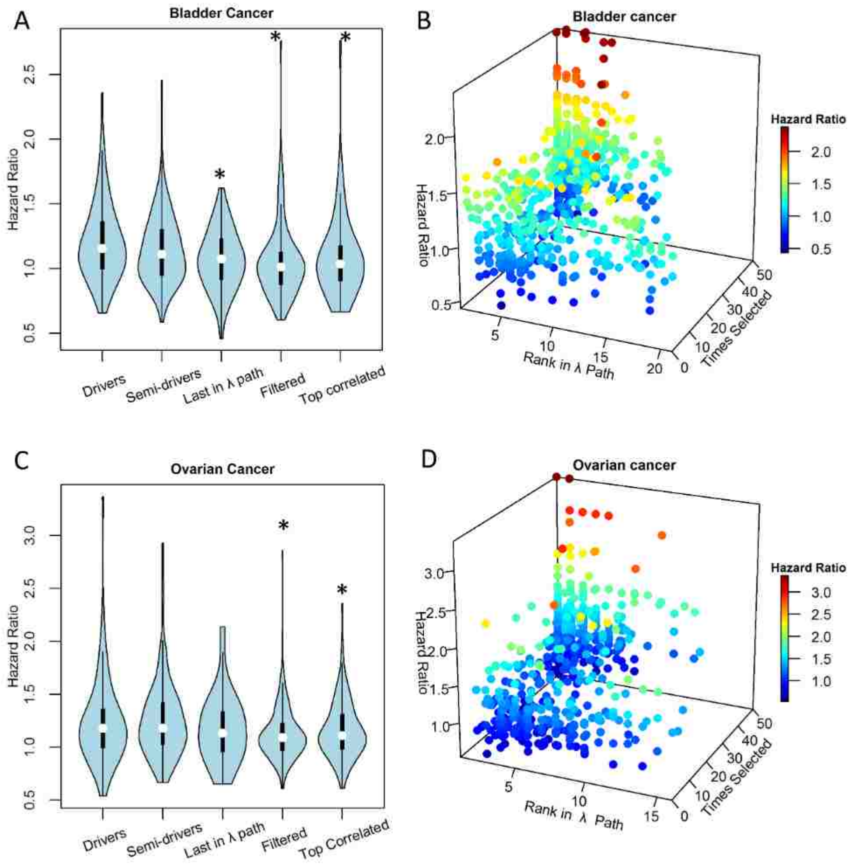


Figure 3.5. Hazard ratios for new tumor events in a univariate Cox proportional hazards model. Violin plots of hazard ratios for genes filtered out or selected at various stages of the driver selection step for (A) bladder cancer and (C) ovarian cancer. Asterisk indicates $p < 0.05$ in a Wilcoxon rank-sum test in bladder cancer and F-test of variances in ovarian cancer compared to the driver group. Hazard ratios were plotted for genes in the multi-task LASSO stage against the number of times they were selected by resampling and the rank in the λ path for (B) bladder cancer and (D) ovarian cancer.

In ovarian cancer, multiple driver RAF1, a putative oncogene, had the highest hazard ratio of 3.2. However, multiple driver CASP3, which promotes apoptosis and is in a deleted region, had the lowest hazard ratio of 0.55. This highlights that the hazard ratio could be dependent on the drivers oncogenic or tumor suppressor activities since a lower hazard ratio implies lower risk with increased expression. We found that the driver group ($\sigma^2 = 0.16$) had a significantly higher variance than the top correlated ($\sigma^2 = 0.07$), and filtered ($\sigma^2 = 0.055$) groups (Levenne's test p-value < 0.05). Although not significant, drivers also had a larger variance than the last in λ path group. This suggests that drivers of ovarian cancer have a higher or lower hazard ratio due to tumor suppressor and oncogenic activities (Figure 3.5C).

Bladder cancer also contains drivers with low hazard ratios. For example, multiple driver FEM1B has a hazard ratio of 0.8 and is a pro-apoptotic protein (Subauste et al., 2010). Figure 3.5B and 3.5D illustrates the hazard ratio for new tumor events after initial treatment for cis genes that appeared in the multi-task LASSO phase in bladder and ovarian data sets, respectively. The results show that cis genes with the highest hazard ratios were selected close to 50 out of 50 times during resampling and had a relatively low rank in the λ path.

3.4. Conclusions

We designed and implemented ProcessDriver in three different cancer sets and found consistently that the most likely candidate drivers are more enriched in known cancer genes. For each dataset, more than half of the multiple drivers are known to be involved in cancer. Biological processes are associated with each driver through the trans genes, and all the trans genes are differentially expressed as a result of the aberration. Therefore, the processes associated with a driver are the ones that are likely disrupted.

We also found that the selected drivers have more extreme hazard ratios for new tumor events after initial treatment with respect to new tumor events compared to cis genes filtered out by ProcessDriver and cis genes selected based on their correlation of expression to their own copy number. Since drivers promote tumorigenesis, it is expected that drivers would be linked to new tumor events.

Aside from ensuring that all cis genes and trans genes are differentially expressed with respect to an aberrated region, we do not use the correlation of copy number to cis gene expression in our filtering of drivers. However, as expected, the cis genes that were selected as drivers had expression that was more correlated to their own copy number compared to cis genes filtered by SCCA. This result suggests that drivers tend to have higher correlation to copy number. However, when we selected the cis genes that are most correlated to their own copy number for each GO term module, it results in a lower enrichment of known cancer genes and lower hazard ratios with respect to new tumor events compared to drivers selected by ProcessDriver. These results highlight the importance of selecting drivers based on the relationship between cis gene expression and trans gene expression, as opposed to selecting the cis genes based on correlation to their own copy number as in previous studies (Tamborero et al., 2013).

While a couple of studies relate cis genes to other genes in trans, our approach differs from previous approaches in several ways. The statistical approaches outlined in this pipeline strongly emphasize a close relationship between a potential driver and downstream target trans genes and provide insight into disrupted biological processes. Akavia et al. relates the expression of cis genes to downstream targets, but does not integrate information about biological processes (2013). Aure et al. associates cis genes with biological processes in trans. However, all other genes are used as trans genes (2013). In this study, all trans genes must be differentially expressed with respect to the aberration. In Aure et al., 2013 the correlation with cis genes to their own copy number is to first narrow down cis genes. Here, we demonstrate that the relationship between cis gene expression and trans gene expression is more valuable in selecting drivers than the correlation of cis genes to their own copy number.

ProcessDriver will narrow down a list of driver genes from many genes that are cis-affected by copy number. This could help find drivers which could be therapeutic targets of drugs. Additionally, the algorithm associates drivers with biological processes through the trans genes, which could aid in gaining insight into the widespread, downstream effects of the driver.

CHAPTER 4

A Canonical Correlation Analysis Based Dynamic Bayesian Network Prior to Infer Gene Regulatory Networks from Multiple Types of Biological Data

Reprinted with permission from JOURNAL OF COMPUTATIONAL BIOLOGY 22/4, published by Mary Ann Liebert, Inc., New Rochelle, NY.

ABSTRACT

One of the challenging and important computational problems in systems biology is to infer gene regulatory networks of biological systems. Several methods that exploit gene expression data have been developed to tackle this problem. In this study, we propose the use of copy number and DNA methylation data to infer gene regulatory networks. We developed an algorithm that scores regulatory interactions between genes based on canonical correlation analysis. In this algorithm, copy number or DNA methylation variables are treated as potential regulator variables and expression variables are treated as potential target variables. We first validated that the canonical correlation analysis method can infer true interactions in high accuracy. We showed that the use of DNA methylation or copy number datasets leads to improved inference over steady-state expression. Our results also showed that epigenetic and structural information could be used to infer directionality of regulatory interactions. Additional improvements in gene regulatory network inference can be gleaned from incorporating the result in an informative prior in a dynamic Bayesian algorithm. This is the first study that incorporates copy number and DNA methylation into an informative prior in dynamic Bayesian framework. By closely examining top-scoring interactions with different sources of epigenetic or structural information, we also identified potential novel regulatory interactions.

4.1. Introduction

Gene regulatory networks (GRNs) are graphs where nodes represent genes and edges represent regulatory interactions between genes. Several methods have been developed in fields such as Bayesian statistics (Werhli and Husmeier, 2007), information theory (Margolin et al., 2004, Bozdag et al., 2010) and regression (Setty et al., 2011) to infer GRNs.

Among the several methods to infer GRNs, the dynamic Bayesian network (DBN) framework is a popular one because of its ability to handle noisy input data (Husmeier, 2003). However, inference of networks based on time-series microarrays could be difficult because the interactions are complex and there may be very few time points due to experimental limitations. This has led to poor reconstruction of GRNs and a lack of scalability to more complex organisms.

More recent studies have attempted to integrate other sources of biological knowledge, such as literature, protein-protein interactions and DNA binding data. One type of informative prior was implemented in a variety of studies (e.g., Imoto et al., 2003; Werhli and Husmeier, 2007 and Zheng et al., 2011). This prior takes the form of a Gibbs distribution in which prior information is encoded by an energy function. These studies have shown that informative priors improve the inference of GRNs.

A few studies have integrated epigenetic and structural data types such as DNA methylation, copy number and histone modification into various inference frameworks (Setty et al., 2012; Zheng et al., 2011). Setty et al. (2012) used a regression-based framework to infer regulatory programs, which after taking DNA methylation and copy number into account, explains differential gene expression in terms of transcription factors and miRNAs. In Zheng et al. (2011), the correlation of histone features was used as informative prior for a DBN-based method.

Canonical correlation analysis (CCA) (Hotelling, 1936) has been used on a genome-wide scale in combination with a penalization method to identify co-expressed or co-regulated genes and associated DNA markers (Waaijenborg et al., 2008). These studies have been successful in identifying drivers of gene expression by their ability to reduce the number of variables. These studies highlight the potential use of CCA for inferring GRNs.

In the present study, we developed a CCA-based algorithm to score potential regulatory relationships in a set of genes. We used DNA methylation or copy number variables to represent potential regulators, and expression variables to represent potential targets. We used the scores from the CCA algorithm as a prior for a DBN-based algorithm (Werhli and Husmeier, 2007) to infer GRNs in breast cancer tissues. Our algorithm is based on the assumption that changes to a regulator's DNA methylation or copy number level will lead to corresponding changes in its downstream targets. We investigated the use of the DBN method to improve the results of the CCA algorithm. This is the first study that makes use of these data types in an informative prior for the DBN framework.

We tested our algorithm to infer known GRNs in human based on a breast cancer dataset. Our results showed that the CCA algorithm is able to infer GRNs with high accuracy. We also showed that the DBN method could use the CCA results to obtain a higher accuracy.

4.2. Methods

4.2.1 Canonical correlation analysis based algorithm

We applied a CCA-based algorithm between DNA methylation or copy number and gene expression datasets to compute potential regulatory interactions between genes. The brief explanation of CCA is as follows. Consider a matrix X with $q \times p$ expression observations and a matrix Y with $q\times p$ DNA methylation or copy number observations, where p is the number of variables and q is the number of samples. Without loss of generality, we assume that the number of variables and samples in each set is the same. CCA computes canonical variates, $\gamma = Yu$ and $\delta = Xv$, $u = (u_1, \dots, u_p)$, $v = (v_1, \dots, v_p)$, where *u* and *v* are weight vectors that maximize the canonical correlation ρ between X and Y (Equation 4.1). In this work, X and Y matrices were generated using gene expression, DNA methylation and copy number datasets of 175 breast cancer samples. We assumed that the variables that had higher weights would be more likely to have regulatory interactions between each other.

$$\rho = \frac{v'X'Yu}{\sqrt{v'X'Xv}\sqrt{u'Y'Yu}} \tag{4.1}$$

We designed an algorithm to resample a small portion of the total set of genes iteratively and applied CCA on this subset (Algorithm 1). At each iteration of resampling, we took a subset of expression variables and a non-overlapping set of copy number or DNA methylation variables. The copy number or DNA methylation variables represented potential regulators and the expression variables represented potential targets. The two sets must be non-overlapping so that large weights were not due to genes where DNA methylation or copy number was highly correlated to their own expression (resulting in a false self-loop). We performed CCA between the two subsets to compute weight vectors for both DNA methylation (copy number) and expression variables. The top *r* out of *p* potential regulators were selected out of the weights in *u*. Genes with large weights in the *u* vector were the most likely candidates to be regulators of some or all of the *p* genes in the *v* vector. Genes with low weights in *u* were not potential regulators of the set of selected genes in *v*.

We computed the regulatory interaction score by taking into account the absolute weight of the regulator and target, as well as the canonical correlation of the canonical variate (Algorithm 4.1). The scores were continuously summed whenever the potential target and regulator pair were selected together. The largest theoretical addition to the score per iteration could be one when the potential target and potential regulator both had the maximum combination of weights and the canonical correlation is one. After subsampling iterations were over, the score between each gene pair was divided by the total number of times the pair was selected together to scale the score between zero and one.

Algorithm 4.1: Scoring interactions based on CCA

```
*n*: number of genes, *m*: number of resamples, *p*: resampling subset size
*S(j, k)*: interaction score between regulator *j* and target *k*
*C(j, k)*: number of times, gene *j* was selected as regulator for gene *k*
for *i* =1:*m*
Select subset of methylation/copy number variables and subset of expression variables
Perform CCA between two subsets
Select weights $u$ (methylation/copy number) and $v$ (expression) that maximize canonical correlation $\rho$
Select $r$ top 80th percentile of weights in $u$
for $j$ = 1:$r$
    for $k$=1:$p$
        $S(j,\ k) = S(j,k) + \frac{(abs(u_j\,) + abs(v_k))\ * \rho}{(max(abs(u)) + max(abs(v))\,)}$
        $C(j,k) = C(j,k) + 1$
$S = \frac{S}{C}$
```

The CCA computes weights for both regulator and target, and a canonical correlation. The weight for the regulator could quantify the strength or relevance of the regulator to the selected targets, and the weight of the target could quantify how much the target's expression is affected. The canonical correlation could quantify strength of the association. We computed our score by taking into account the weight of the regulator and target, as well as the canonical correlation of the canonical variate.

4.2.2 Dynamic Bayesian network analysis

We implemented the DBN-based algorithm as described in Werhli and Husmeier, 2007. We applied a Markov Chain Monte Carlo (MCMC) search heuristic with a Metropolis Hastings acceptance criterion as described in Werhli and Husmeier, 2007. The prior (Equation 2.2) took the form of a Gibbs distribution where prior information (results from the CCA algorithm) was encoded by an energy function (Equation 2.3), where $Z(\beta)$ was a normalizing constant. The parameter, β, measured the influence of the prior information relative to the time series expression data.

$$P(G|\beta) = \frac{e^{-\beta E(G)}}{Z(\beta)} \tag{4.2}$$

The energy function measured how closely the prior information matched with the network structure at the current step of MCMC (Equation 4.3). In energy function, *B* is the prior matrix, which was calculated by the CCA algorithm and *G* is the current network structure. As the energy goes to zero, there is more agreement between the prior and the network structure.

$$E(G) = \sum_{i,j=1}^{N} |B_{ij} - G_{ij}| \tag{4.3}$$

At each MCMC step, a move was made in which an edge was added or deleted. In this method, the Metropolis Hastings acceptance criterion (Eq. 1.5) is expanded so that both the network structure *G* and the hyperparameter β can be sampled from the posterior distribution, $P(G, \beta|D)$ (Wehrli and Husmeier, 2007):

$$P_{MH} = \min\left\{\frac{\left(P(D|G_{new})P(G_{new}|\beta_{new})P(\beta_{new})Q(G_{old}|G_{new})R(\beta_{old}|\beta_{new})\right)}{P(D|G_{old})P(G_{old}|\beta_{old})P(\beta_{old})Q(G_{new}|G_{old})R(\beta_{new}|\beta_{old})}\right\} \tag{4.4}$$

Hasting's ratio, $Q(G_{old} \mid G_{new})/Q(G_{new} \mid G_{old})$, is equal to one as described in Section 1.2.3 of this dissertation and in Husmeier, 2003. $R(\beta_{new}|\beta_{old})$ is the proposal distribution for the hyperparameter. Based on the study by Werhli and Husmeier, Eq. 4.4 is broken up into two sub-moves of applying the acceptance criterion to a new graph structure, G_{new}, while holding β fixed:

$$P_{MH}(G_{new}\,|G_{old}) = \min\left\{1, \frac{P(D|\, G_{new})P(G_{new}|\beta)Q(G_{old}|G_{new})}{P(D|G_{old})P(G_{old}|\beta)Q(G_{new}|G_{old})}\right\} \tag{4.5}$$

And then sampling a new hyperparameter while holding *G* fixed:

$$P_{MH}(\beta_{new}|\beta_{old}) = \min\left\{1, \frac{P(G|\beta_{new})P(\beta_{new})R(\beta_{old}|\beta_{new})}{P(G|\beta_{old})P(\beta_{old})R(\beta_{new}|\beta_{old})}\right\} \tag{4.6}$$

Assuming a uniform prior distribution and symmetric proposal distribution on the hyperparameter, Eq 4.6 simplifies to:

$$P_{MH}(\beta_{new}|\beta_{old}) = \min\left\{1, \frac{P(G|\beta_{new})}{P(G|\beta_{old})}\right\} \tag{4.7}$$

This method can extend to incorporate multiple types of biological priors with their own hyperparameters as described in Werhli and Husmeier, 2007. In our study, the MCMC procedure was initialized, starting with an empty matrix, using a burn-in phase of 100,000 steps. Following the burn-in phase, the network was sampled every 50 steps for another 100,000 steps.

4.2.3 Datasets

We obtained DNA methylation, copy number and gene expression datasets for 175 breast cancer samples from the TCGA repository (The Cancer Genome Atlas, 2012). The datasets were generated in Illumina Infinium 450k methylation array, Affymetrix Genome-Wide Human SNP 6.0 array, and Agilent mRNA expression array platforms, respectively. Additionally, for the DBN algorithm, Affymetrix mRNA time-series data for MCF7 breast cancer cells were downloaded from Nagashima et al. 2007, seven experiments with seven time points were used.

Batch effects were corrected for the mRNA expression data using the LIMMA package in R (Smyth, 2005)**.** For copy number, the normalized segmented means from TCGA were used as the variables with no additional processing. We determined the segment each gene was located within using the start and end sites of the gene provided by UCSC genome browser annotations (Karolchik et al., 2014). Interactions between genes in the same copy number segment and from a gene to itself were treated as uninformative prior for the DBN algorithm.

For the 450k DNA methylation data, each gene was associated with an average of 18 DNA methylation probes. We calculated the correlation between each probe and the expression level of its associated gene. For every gene studied, we took the intensity of the top negatively or positively correlated probe as a measure of the DNA methylation.

4.2.4 Networks and evaluation

We collected interactions of a subset of genes from the Human Transcription Regulation Interaction (HTRI) database (Bovolenta et al., 2012) and the Transcription Regulatory Element

Database (Zhao et al., 2005). These databases are composed largely of transcription factors that bind near genes. We note that some transcription factor binding events do not result in expression changes in a target gene. However, transcription factor binding is widely studied in humans. In the absence of a gold standard, these TF-binding events are the most useful tool for identifying interactions. Interactions found in either or both databases were included in our collection of interactions (Table 4.1). We used this collection to assess the CCA- and DBN-based algorithm's ability to recall these interactions.

Table 4.1. Three networks used to assess the recall of regulatory interactions by CCA and DBN.

Network name	Genes in the network	# of genes	# of interactions
GATA3	GATA3, ESR1, ETS1, FOXA1, FOXP3, MYC, SP1, STAT1, TFAP2A, CDK2NA, TMEM2, PRDM4, MID2, TEK, RBMS1, SERPINF1, EDN1, ATP2B1, PPARG, VGLL4, APP, ATOX1, BTG2, STAT4, STX3	25	92
BRCA	CHEK2, MAP3K3, NEK2, BARD1, BRCA1, BRCA2, FHIT, ESR1, ELK4, BCL2, STAT3, PCNA, POU4F2, TP53, ELK1, ERG, DDB2, GADD45A, IGF1R, TAF10, JAK1, JAK2	22	32
FOXA1	FOXA1, BCL2, BCL6, CDKN2A, FOXA3, STAT1, STAT4, TRIM25, IRF1, GATA3, TP73, PRDM14, STAT6, PCNA, CDKN1A, FASN, PTGS2, CCL17, FCGRT, ICAM1	20	32

We ran the CCA algorithm on DNA methylation and copy number datasets independently. The CCA algorithm computed an interaction score scaled between zero and one for every gene pair in the network. We used CCA results as priors for the dynamic Bayesian algorithm. We computed receiver operating characteristic (ROC) curves and the area under each ROC curve (AUC) to assess the performance of the CCA and DBN algorithms. We counted misdirected edges as false positives.

4.2.5. Implementation

We implemented our tool in MATLAB. The foundation for the code written for this study was made possible by the Bayes Net Toolbox (Murphy, 2002). Additionally, code obtained from Husmeier, 2003 aided in the development by providing the MCMC portion of the algorithm. The source code of this tool is freely available upon request.

4.3. Results and discussion

4.3.1 Computing regulatory interactions in breast cancer by utilizing the CCA algorithm

For each of the three networks studied, we computed the receiver operating characteristic (ROC) curve and the area under the curve (AUC) for the CCA algorithm. The sensitivity is the true positive rate (y-axis) and the inverse specificity is the false positive rate (x-axis). Therefore, a larger AUC for an ROC curve indicates a greater number of correctly inferred edges between genes relative to incorrectly inferred edges. We ran the algorithm using DNA methylation, copy number and expression variables as regulator variables. The resampling step was iterated 10,000 times in which a subset of five variables was selected. Target variables were always expression variables. Figure 4.1 A, B and C and Table 4.2 show that using DNA methylation variables as potential regulators achieved higher accuracy than the accuracy obtained when copy number or expression variables were used as potential regulators in GATA3 and FOXA1 networks. Copy number based results performed best in BRCA network. Using expression variables as regulators performed worst in all three networks. If expression is used on both sides of the canonical correlation analysis, it may be more difficult to separate regulator from target. The DBN results (Figure 4.1 D, E and F) are discussed in section 4.3.3.

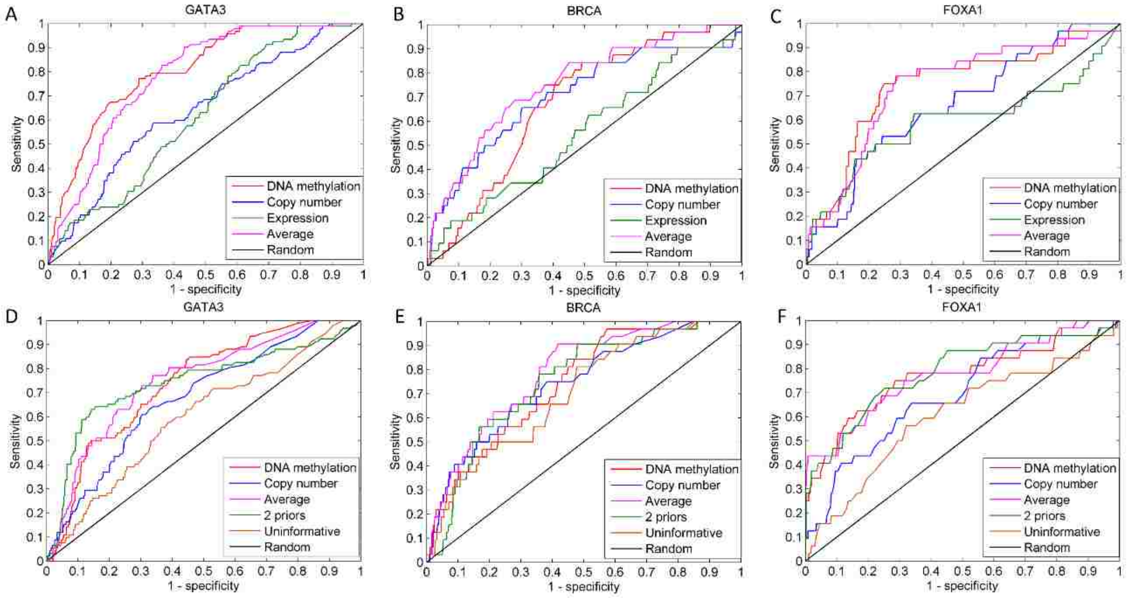


Figure 4.1. **ROC curves for the CCA algorithm in A) GATA3, B) BRCA and C) FOXA1 networks and the DBN algorithm in D) GATA3, E) BRCA and F) FOXA1 networks.**

Table 4.2. AUC for the networks presented in Figure 4.1.

	GATA3		**BRCA**		**FOXA1**	
	CCA	DBN	CCA	DBN	CCA	DBN
DNA methylation	0.81	0.73 ± .01	0.68	0.74 ± .01	0.75	0.77 ± .01
Copy number	0.64	0.67 ± .01	0.70	0.72 ± .01	0.67	0.69 ± .01
Expression	0.62		0.55		0.60	
Averaged	0.78	0.74 ± .01	0.75	0.78 ± .02	0.74	0.78 ± .01
Uninformative		0.60 ± .01		0.70 ± .01		0.60 ± .01
Both		0.75 ± .01		0.75 ± .01		0.78 ± .02

We tested the convergence of the CCA algorithm by running the algorithm over 20,000 steps with five variables. Every 1,000 steps, we calculated the absolute difference between the entries in the scoring matrix at the current step and the scoring matrix from 1,000 steps ago. We then summed all of these differences. We find that the sum of the differences becomes relatively small at around 2,000 steps. suggesting our algorithm converges quickly. Figure 4.2. shows the results for BRCA DNA methylation. Similar results were obtained for other datasets and networks.

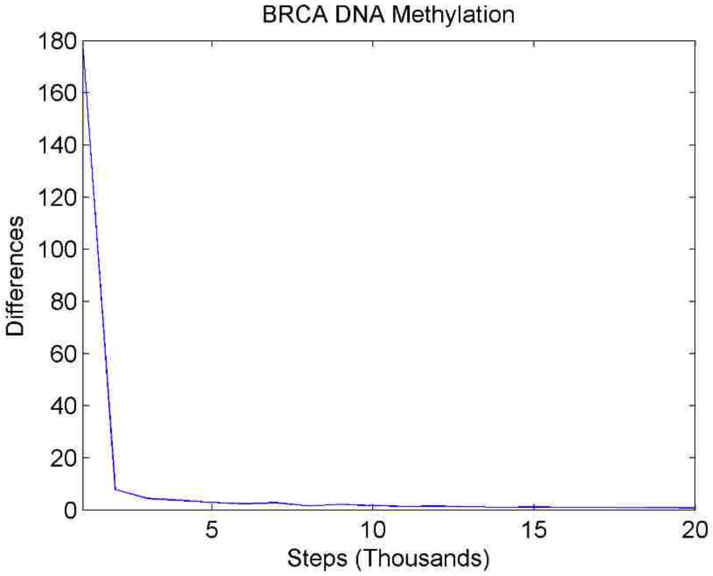


Figure 4.2. Convergence of CCA algorithm.

We also tested the algorithm using different subset sampling sizes. We found that in general, sampling subset sizes between 3 and 5 lead to a higher increase in AUC (Table 4.3).

Table 4.3. Resampling subset size vs. AUC. Meth – DNA methylation, CN – Copy number, Exp – Expression for regulators.

p	GATA3			BRCA			FOXA1		
	Meth	CN	Exp	Meth	CN	Exp	Meth	CN	Exp
3	0.77	0.65	0.64	0.7	0.69	0.6	0.74	0.63	0.62
4	0.77	0.65	0.65	0.7	0.69	0.6	0.74	0.66	0.63
5	0.81	0.64	0.63	0.68	0.7	0.55	0.75	0.67	0.6
6	0.76	0.62	0.6	0.6	0.62	0.55	0.68	0.68	0.6
7	0.7	0.6	0.59	0.55	0.56	0.55	0.65	0.68	0.55

Additionally, we modified the algorithm to resample one target and five regulators, and then chose the best regulator(s) in the 80th percentile of weights. The formula for scoring the algorithm was the same. We computed ROC curves for this method (Figure 4.3). The performance with this modification was worse than the original algorithm. This result, in addition to the simple correlation approach, suggests that the ability to detect regulators may be amplified by relating them to multiple targets. To test this, we modified our algorithm again to select one regulator and five targets. Since there is only one regulator, we do not select the top 80th percentile of weights. The rest of the algorithm remained the same. We again computed ROC curves for this method (Figure 4.3). This approach performs very well in some cases, although our original approach performs slightly better. This suggests that selecting top regulator(s) by weight, in light of other potential candidates, contributes to the robustness of the algorithm.

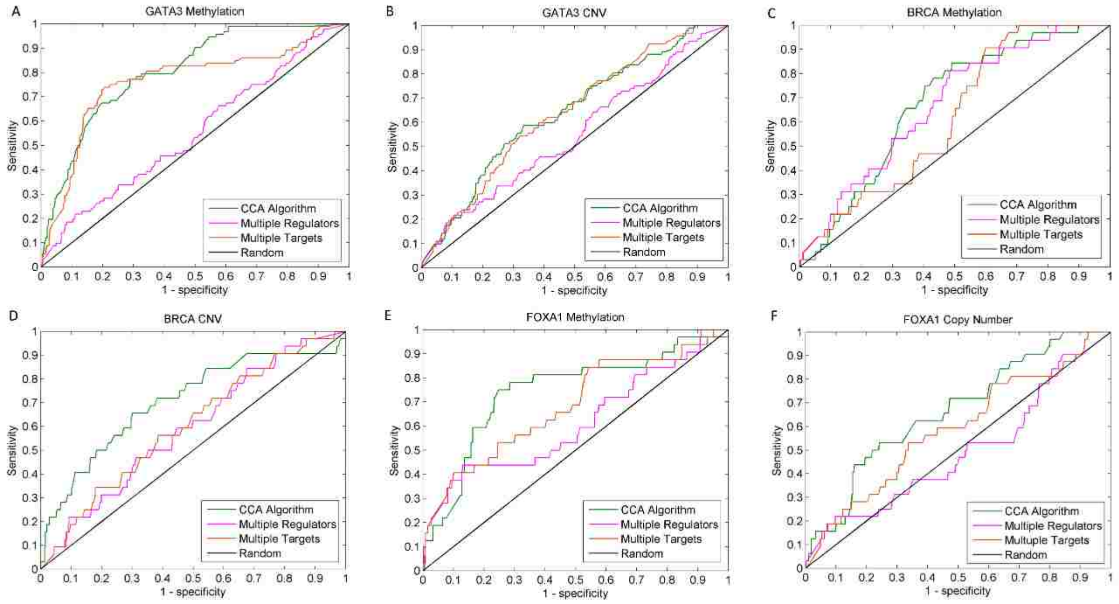


Figure 4.3. CCA algorithm vs. algorithm performed with a single regulator and multiple targets and the algorithm performed with a single target and multiple regulators.

Table 4.4. AUC for ROC curves in Fig.4.3.

	GATA3		**BRCA**		**FOXA1**	
	Meth	CN	Meth	CN	Meth	CN
CCA Algorithm	0.81	0.64	0.68	0.70	0.75	0.67
Multiple Targets	0.76	0.64	0.62	0.60	0.69	.59
Multiple Regulators	0.55	0.55	0.67	0.59	0.61	0.5

Additionally, we tested the hypothesis that methylation of a potential target establishes directionality in the network, contributing to the overall better performance of methylation data compared to expression data. We looked at interactions with only one direction. We subtracted the score in the correct direction (CD) from the score in the incorrect direction (ID). Therefore, positive directionality scores represent correct directionality, zero scores represent no directionality and a negative score represent directionality in the wrong direction. Supplementary Figure 4.4. shows the results for the GATA3, BRCA and FOXA1 network. We find that for the GATA3 and FOXA1 networks, but not the BRCA network, the directionality score was significantly more positive for DNA methylation data compared to expression data. This suggests that the better performance of DNA methylation data may be due in part by its ability to establish directionality.

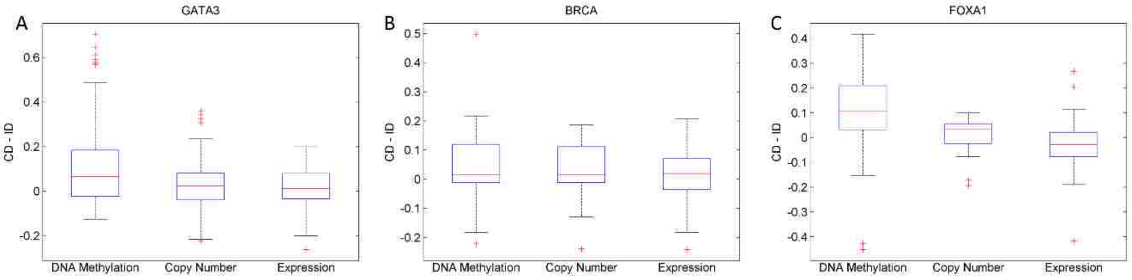


Figure 4.4. Directionality in the CCA algorithm.

We calculated the correlation between the methylation state, copy number or expression state of a potential regulator and the expression of a potential target individually. We took the absolute value of the correlation as a score and computed an ROC curve and the corresponding AUC for each network (Fig. 4.5 and Table 4.5). We found that in all cases, a simple correlation performs worse than our algorithm. We note that using the DNA methylation level for potential regulators performed better than the copy number or expression in the simple correlation approach. This also suggests that DNA methylation data may be more useful in establishing genetic relationships.

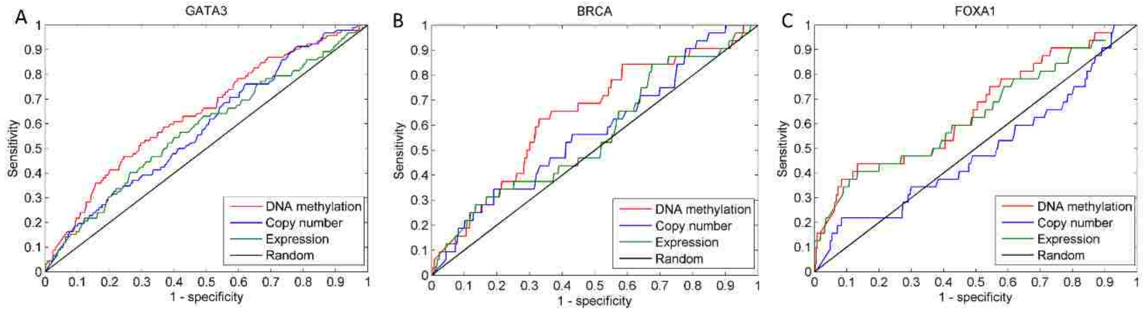


Figure 4.5. ROC curves for correlation-based approach.

Table 4.5. AUC for ROC curves in Fig 4.5.

	GATA3	**BRCA**	**FOXA1**
DNA methylation	0.64	0.64	0.64
Copy number	0.58	0.57	0.46
Expression	0.58	0.56	0.62

We tested for a significant separation of experimentally validated versus non-experimentally validated genes. Figure 4.6 shows the CCA algorithm scores for interacting and non-interacting pairs of genes in the GATA3 network and Figure 4.7 shows the scores for the BRCA and FOXA1 networks. We observed that for DNA methylation and copy number datasets, the CCA algorithm scores were significantly higher for interacting pairs compared to non-interacting pairs (Wilcoxon rank-sum test $p < .001$). This was not the case for the gene expression datasets.

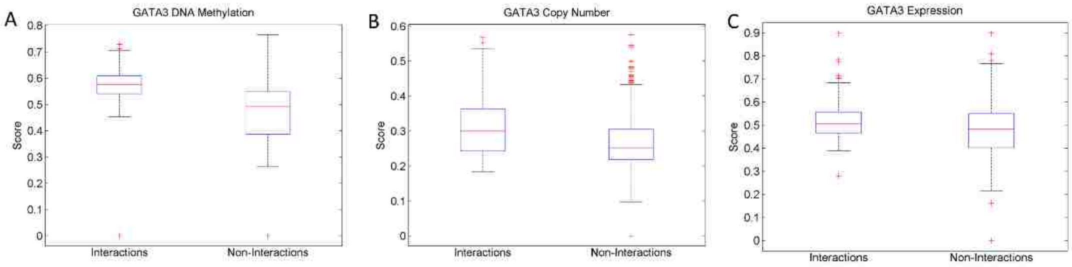


Figure 4.6. **Regulatory interaction scores by the CCA algorithm for interacting vs. non-interacting pairs in the GATA3 network using A) DNA methylation, B) Copy number and C) Expression datasets as regulator variables.**

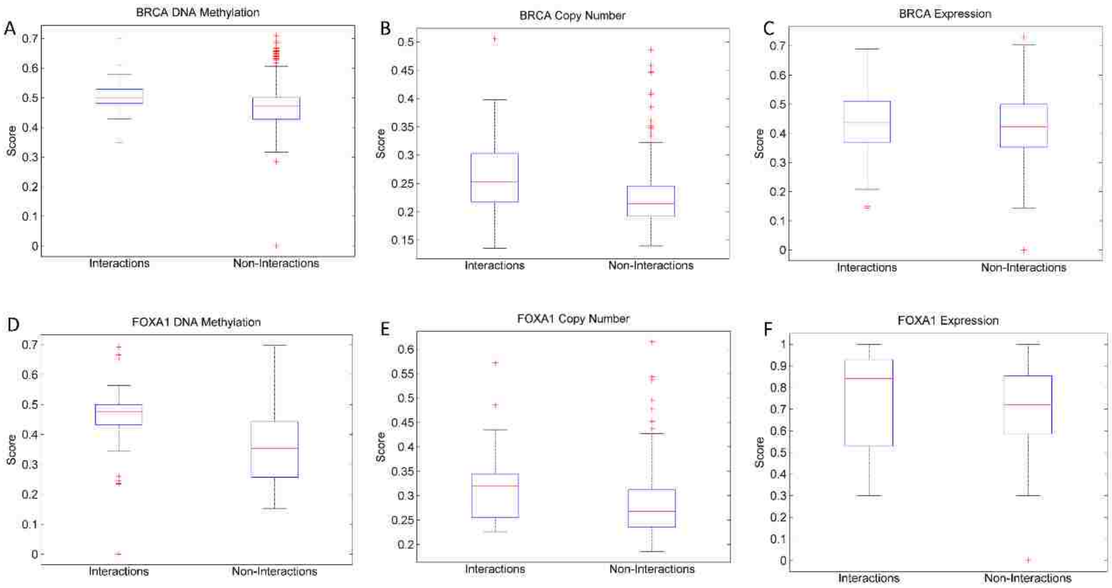


Figure 4.7. Regulatory interaction scores by the CCA algorithm for interacting vs. non-interacting pairs in the BRCA and FOXA1 network using A) DNA methylation, B) Copy number and C) Expression datasets as regulator variables.

4.3.2 Validation of false positive interactions of the CCA algorithm in literature

False positives are interactions that were scored high by the CCA algorithm, but were not present in the databases. False positive interactions may represent unknown interactions since interaction databases are often incomplete. Tables 4.6 and 4.7 show the top 15 scoring false positive interactions computed by the CCA algorithm using DNA methylation and copy number, respectively. In Table 4.6, several of the regulators are known to be affected by DNA methylation in cancer cells. Specifically, BCL2 has a hyper-methylation biomarker (Stone et al., 2013). Inactivation of IRF1 via DNA methylation has been implicated in the tumorigenesis of gastric cancers (Yamashita et al., 2010). TP73 is known to be controlled by promoter hypermethylation (Dong et al, 2002). In Table 4.7, aberrant expression of ATOX1 has been recently linked to breast cancer (Choong et al., 2010). IRF1 undergoes structural changes and has been linked to frequent loss of heterozygosity in breast cancer (Cavalli et al., 2010). These results suggest that the performance of a regulator may indicate the extent the regulator is affected by DNA methylation and/or copy number abnormalities.

Table 4.6. Top 15 false positive interactions computed by the CCA algorithm utilizing DNA methylation data.

GATA3			BRCA			FOXA1		
Regulator	Target	Score	Regulator	Target	Score	Regulator	Target	Score
GATA3	ESR1	0.76	BCL2	ESR1	0.71	TP73	STAT4	0.7
FOXA1	ESR1	0.75	BCL2	IGF1R	0.70	BCL2	FOXA1	0.7
FOXA1	ETS1	0.71	BCL2	BARD1	0.68	IRF1	STAT4	0.67
GATA3	ETS1	0.69	BCL2	GADD45A	0.67	TP73	IRF1	0.65
ESR1	ETS1	0.68	BCL2	MAP3K3	0.67	IRF1	ICAM1	0.64
TFAP2A	STAT4	0.68	BRCA2	IGF1R	0.67	TP73	CCL17	0.62
APP	ESR1	0.67	BCL2	FHIT	0.66	BCL2	GATA3	0.61
ESR1	RBMS1	0.66	BCL2	PCNA	0.66	IRF1	FOXA1	0.61
FOXA1	FOXP3	0.65	BCL2	ERG	0.66	IRF1	CCL17	0.6
ETS1	ESR1	0.65	BCL2	CHEK2	0.66	TP73	ICAM1	0.58
GATA3	FOXP3	0.65	BCL2	DDB2	0.65	TP73	STAT1	0.57
FOXA1	MYC	0.64	BCL2	BRCA1	0.65	FOXA1	FCGRT	0.56
TFAP2A	ETS1	0.63	IGF1R	CHEK2	0.63	GATA3	PTGS2	0.56
TMEM2	ESR1	0.63	IGF1R	ESR1	0.62	BCL2	ICAM1	0.55
FOXA1	TEK	0.63	JAK1	ESR1	0.62	GATA3	FASN	0.55

Table 4.7. Top 15 false interactions computed by the CCA-base algorithm utilizing copy number data.

GATA3			BRCA			FOXA1		
Regulator	Target	Score	Regulator	Target	Score	Regulator	Target	Score
SP1	FOXA1	0.58	STAT3	FHIT	0.49	IRF1	GATA3	0.62
ATOX1	GATA3	0.58	BRCA1	NEK2	0.46	STAT6	FOXA1	0.54
SP1	GATA3	0.55	GADD45A	JAK1	0.45	IRF1	FOXA1	0.54
ATOX1	ESR1	0.54	MAP3K3	BRCA2	0.45	IRF1	BCL2	0.5
SP1	ATP2B1	0.50	JAK2	BRCA2	0.41	STAT6	GATA3	0.48
SP1	RBMS1	0.48	GADD45A	ESR1	0.41	IRF1	FCGRT	0.45
ATOX1	FOXA1	0.48	STAT3	BRCA1	0.38	CDKN1A	GATA3	0.44
ATOX1	BTG2	0.48	BCL2	TAF10	0.36	IRF1	CDKN2A	0.43
SP1	PRDM4	0.48	ERG	ESR1	0.36	IRF1	STAT4	0.42
SP1	ATOX1	0.47	BRCA2	CHEK2	0.35	GATA3	FCGRT	0.41
GATA3	ESR1	0.47	ELK4	NEK2	0.35	GATA3	BCL2	0.41
ATP2B1	FOXA1	0.46	POU4F2	CHEK2	0.35	STAT6	TP73	0.41
SP1	BTG2	0.46	BRCA1	BARD1	0.34	IRF1	ICAM1	0.41
PRDM4	FOXA1	0.46	POU4F2	JAK1	0.33	IRF1	PCNA	0.41
FOXA1	ESR1	0.45	TAF10	DDB2	0.32	IRF1	TP73	0.41

The CCA algorithm computed interactions between IRF1 and BCL2, CDKN2A and PCNA that are supported by literature (Saneau et al., 2000; Coccia et al., 1999; Frontini et al., 2009). Interactions of SP1 with FOXA1, GATA3, and RBMS1 are also supported by our results and by literature (Chavez et al., 2009; Gilli et al., 2004; Haigermoser et al., 1996). Tables 4.8 and 4.9. provide a complete summary of supporting literature.

Table 4.8. Experimental validation for false positives inferred with DNA Methylation CCA.

GATA3			BRCA			FOXA1		
Reg	Target	Ref	Reg	Target	Ref	Reg	Target	Ref
GATA3	ESR1		BCL2	ESR1		TP73	STAT4	
FOXA1	ESR1	Bernardo et al., 2010	BCL2	IGF1R		BCL2	FOXA1	
FOXA1	ETS1		BCL2	BARD1		IRF1	STAT4	
GATA3	ETS1		BCL2	GADD45A		TP73	IRF1	
ESR1	ETS1		BCL2	MAP3K3		IRF1	ICAM1	
TFAP2A	STAT4		BRCA2	IGF1R		TP73	CCL17	
APP	ESR1	Von Arnim et al., 2006	BCL2	FHIT		BCL2	GATA3	
ESR1	RBMS1		BCL2	PCNA		IRF1	FOXA1	
FOXA1	FOXP3		BCL2	ERG		IRF1	CCL17	
ETS1	ESR1		BCL2	CHEK2		TP73	ICAM1	
GATA3	FOXP3	Wang et al., 2011	BCL2	DDB2		TP73	STAT1	
FOXA1	MYC	Ni et al., 2013	BCL2	BRCA1		FOXA1	FCGRT	
TFAP2A	ETS1		IGF1R	CHEK2		GATA3	PTGS2	
TMEM2	ESR1		IGF1R	ESR1	Foulstone et al., 2013	BCL2	ICAM1	
FOXA1	TEK		JAK1	ESR1		GATA3	FASN	

Table 4.9. Experimental validation for false positives inferred with Copy Number CCA.

GATA3			BRCA			FOXA1		
Reg	Target	Ref	Reg	Target	Ref	Reg	Target	Ref
SP1	FOXA1	Chavez et al., 2009	STAT3	FHIT		IRF1	GATA3	
ATOX1	GATA3		BRCA1	NEK2	Wang et al., 2004	STAT6	FOXA1	
SP1	GATA3	Gilli et al., 2004	GADD45A	JAK1		IRF1	FOXA1	
ATOX1	ESR1		MAP3K3	BRCA2		IRF1	BCL2	Saneau et al., 2000
SP1	ATP2B1		JAK2	BRCA2		STAT6	GATA3	Stockinger et al., 2007
SP1	RBMS1	Haigermoser et al., 1996	GADD45A	ESR1		IRF1	FCGRT	
ATOX1	FOXA1		STAT3	BRCA1		CDKN1A	GATA3	
ATOX1	BTG2		BCL2	TAF10		IRF1	CDKN2A	Coccia et al, 1999
SP1	PRDM4		ERG	ESR1		IRF1	STAT4	
SP1	ATOX1		BRCA2	CHEK2		GATA3	FCGRT	
GATA3	ESR1		ELK4	NEK2		GATA3	BCL2	Tsarovina et al., 2010
ATP2B1	FOXA1		POU4F2	CHEK2		STAT6	TP73	
SP1	BTG2		BRCA1	BARD1	Rodriquez et al., 2004	IRF1	ICAM1	
PRDM4	FOXA1		POU4F2	JAK1		IRF1	PCNA	Frontini et al., 2009
FOXA1	ESR1	Bernardo et al., 2010	TAF10	DDB2		IRF1	TP73	

4.3.3 Inferring GRNs by a DBN-based approach utilizing priors by the CCA-algorithm

We used the results from the CCA algorithm as a prior in the DBN algorithm. We first analyzed the effect of the parameter, β, which is a measure of the agreement between the time series data and the prior (Werhli and Husmeier, 2007) by holding β constant throughout the MCMC learning process. Table 4.10 shows the values of β used and the AUC achieved after averaging 5 executions of the DBN algorithm using DNA methylation- and copy number-based priors, respectively. In every case, except for GATA3 DNA methylation, the DBN improved the overall accuracy for some values of β. DNA methylation tended to perform better with higher values of β, while copy number tended to perform better at lower values for β.

Table 4.10. AUC for various values of $\boldsymbol{\beta}$.

β	**GATA3**		**BRCA**		**FOXA1**	
	Meth	CN	Meth	CN	Meth	CN
CCA	0.81	0.64	0.68	0.70	0.75	0.67
1	0.64	0.60	0.60	0.60	0.69	0.56
3	0.68	0.61	0.60	0.81	0.70	0.63
5	0.69	0.63	0.62	0.75	0.73	0.64
7	0.72	0.63	0.62	0.75	0.73	0.69
9	0.73	0.73	0.72	0.73	0.72	0.64
11	0.72	0.70	0.70	0.65	0.70	0.65
13	0.73	0.60	0.70	0.62	0.76	0.64
15	0.76	0.60	0.74	0.68	0.79	0.63
17	0.73	0.55	0.72	0.68	0.73	0.60
19	0.73	0.56	0.72	0.65	0.73	0.60

In the following experiments, we used β as a hyperparameter as in Werhli and Husmeier (2007). We used the CCA results on DNA methylation and copy number independently in two DBN algorithm runs. We also used DNA methylation and copy number results as two separate priors in a single DBN algorithm run. Finally, we averaged together the CCA results of DNA methylation and copy number and used it as a single prior. For each network, results with an uninformative prior were also obtained.

Figure 4.1 D, E and F shows the ROC curve for the DBN method, with the AUC reported in Table 4.2. When the CCA results for DNA methylation or copy number were used as a prior,

there were improvements in the AUC over just CCA alone. One exception was in GATA3 network in which DNA methylation-based prior had higher accuracy than in the DBN algorithm with this prior. This was potentially due to limitations of time series expression data. The results suggest that using the average prior or two priors leads to more improvement over using a single prior alone. In order to compute significance of the DBN algorithm's improvement over the CCA algorithm, we ran the CCA algorithm five times for each network for both DNA methylation and copy number data. We then used each result as a prior in the DBN algorithm. We performed a paired t-test between the AUC in the CCA algorithm results and the DBN algorithm results. We found that the improvement of DBN was significant (< .05) in all cases except for GATA3 network results when utilizing DNA methylation data. We report the significance in Table 4.11.

Table 4.11. T-Test for DBN runs.

GATA3 Copy Number		**BRCA Copy Number**		**FOXA1 Copy Number**	
CCA	DBN	CCA	DBN	CCA	DBN
0.64	0.66	0.68	0.72	0.67	0.69
0.64	0.66	0.68	0.73	0.68	0.70
0.64	0.66	0.68	0.72	0.67	0.70
0.64	0.67	0.70	0.70	0.66	0.68
0.64	0.65	0.69	0.72	0.67	0.69
p = .003		**p = .02**		**p < .001**	
GATA3 Methylation		**BRCA Methylation**		**FOXA1 Methylation**	
CCA	DBN	CCA	DBN	CCA	DBN
0.80	0.74	0.68	0.72	0.75	0.75
0.80	0.74	0.69	0.74	0.73	0.75
0.79	0.73	0.68	0.74	0.73	0.77
0.80	0.73	0.68	0.75	0.73	0.75
0.8	0.73	0.69	0.74	0.74	0.77
		p < .001		**p = .03**	

4.3.4. Validation of false positives in literature for DBN

Tables 4.12 and 4.13 show the top 15 false positive interactions computed by the DBN algorithm for each prior in GATA3 and BRCA networks, respectively. The FOXA1 network did not have false positives that scored over 0.5 in multiple prior types. Some of these interactions were supported by recent literature. In the GATA3 network, there are 15 interactions that were supported by the results in at least two prior types. Among these, SERPINF1 →PPARG has been experimentally validated by Ho et al. (2007). This interaction is supported by 3 prior types.

Table 4.12. Top 15 scoring false positive interactions in GATA3 network computed by the DBN algorithm using various priors.

Regulator	Target	Two priors	Average prior	DNA Methylation	Copy number	Supporting Data
EDN1	SERPINF 1	0.77	0.69	0.74	0.55	
VGLL4	TMEM2	0.73	0.65	0.59	0.64	
SERPINF1	PPARG	0.77	0.56	0.99		Ho et al., 2007
EDN1	FOXA1		0.97	0.97	0.97	
RBMS1	MID2		0.60	0.55	0.50	
RBMS1	TFAP2A		0.60	0.53	0.53	
GATA3	ESR1		0.59	0.65	0.54	
RBMS1	GATA3		0.66		0.68	
RBMS1	APP		0.57		0.55	
RBMS1	ATOX1		0.51		0.60	
TFAP2A	STAT1		0.62			
SP1	VGLL4		0.55	0.55		
ESR1	AT2B1		0.50	0.83		
ATOX1	SP1		0.57	0.75		
TMEM	CDKN2A	0.74		0.74		
FOXA1	FOXP3	0.63		0.83		
MYC	CDKN2A	0.80				Zindy et al., 1998
VGLL4	MID2	0.79				
STAT1	BTG2	0.78				
CDKN2A	STAT4	0.72				
EDN1	STAT4	0.69				
APP	ATOX1	0.67				
SP1	STAT4	0.67				
SERPINF1	SP1	0.65				
PPARG	STAT1	0.64				Ricote et al., 1998
MYC	ATP2B1	0.63				
RBMS1	ESR1		0.53			
STX	CDKN2A			0.67		
APP	ATOX1			0.60		Martin et al., 2008
EDN1	STX3				0.52	
RBMS1	VGLL4				0.49	

Table 4.13. Top 15 scoring false positive interactions in BRCA network computed by the DBN algorithm using various priors.

Regulator	Target	Two priors	Average prior	DNA Methylation	Copy number	Supporting Data
PCNA	CHEK2	0.58	0.99	0.89	0.64	
PCNA	DDB2	0.85	0.96	0.89	0.89	
PCNA	JAK1	0.72	0.94	0.98	1	
PCNA	JAK2	0.72	0.49	0.66		
BRCA2	BRCA1		0.43	0.40	0.40	
ESR1	BRCA2		0.42	0.38	0.47	
BRCA1	BRCA2		0.41	0.37	0.40	Fan et al., 1998
IGF1R	STAT3		0.40	0.41	0.44	Zhang et al., 2006
IGF1R	BRCA2		0.39	0.38	0.39	
STAT3	BRCA1		0.42		0.42	
BRCA2	IGFR1		0.41		0.40	
ESR1	STAT3		0.40		0.39	Rokavec et al., 2012
BRCA2	STAT3		0.40		0.41	
STAT3	BRCA2		0.41	0.38		
STAT3	IGFR1		0.43	0.37		Scheidegger et al., 1999
IGF1R	BRCA1			0.40	0.39	
IGF1R	ESR1			0.36	0.45	Foulstone et al., 2013
DDB2	CHEK2	0.83				
FHIT	DDB2	0.82				
JAK1	ELK1	0.68				
GADD45A	MAP3K3	0.68				
ELK4	TAF10	0.65				
CHEK2	BCL2	0.59				
NEK2	JAK2	0.55				
BARD1	CHEK2	0.53				
PCNA	BCL2	0.5				
TAF10	ELK4	0.45				
NEK2	BCL2	0.44				
STAT3	TAF10			0.37		
ESR1	BRCA1				0.43	

It is worth to note that EDN1 → SERPINF1, and VGLL4 → TMEM2 were supported by all four prior types and could be novel interactions. Although no relationship is known, EDN1 is a vasoconstrictor and SERPINF1 induces apoptosis by inhibiting stromal vasculature (Doll et al., 2003). RBMS1 →MID2, RBMS1 → TFAP2A, and EDN1 → FOXA1 were supported by three prior types.

In the BRCA network, there are 17 interactions supported by our results in at least two prior types. IGF1R→STAT3 interaction, which was supported by three prior types has been experimentally validated (Zhang et al., 2006). Additionally, IGF1R→ESR1, BRCA1 → BRCA2, ESR1→STAT3, STAT3→IGFR1 have supporting evidence (Foulstone et al., 2013; Fan et al., 1998; Rokavec et al., 2012; Scheidegger et al., 1999). Although there is no validation yet, PCNA → CHEK2, PCNA→DDB2 and PCNA→JAK1 could be novel interactions as they were assigned high scores by all prior types.

4.4. Conclusions and future work

We developed an algorithm that scores regulatory interactions based on canonical correlation analysis (CCA) between various biological datasets and gene expression. We tested our algorithm on a breast cancer dataset that composed of DNA methylation, copy number and gene expression. We computed regulatory interactions in three gold standard networks, which were built with known interactions in HTRI and TRED databases. Our results showed that using DNA methylation and copy number data as regulator variables performed better than using expression data as regulator variables. This indicates that DNA methylation and copy number may establish directionality by distinguishing between regulator and target. These results also highlight the usefulness of epigenetic and structural information in GRN inference. Some of the CCA algorithm's top interactions were supported by literature although these interactions did not exist in the HTRI and TRED databases. These interactions might contain putative regulators controlled by DNA methylation or copy number changes, and their targets.

We used the results of our CCA algorithm as a prior for a dynamic Bayesian network (DBN) approach. We ran the DBN algorithm by utilizing DNA methylation- and copy number-based priors individually and simultaneously. We showed that additional improvements could be gleaned from using this method over the CCA alone. Like in the CCA algorithm results, some of the top interactions computed by the DBN algorithm did not exist in the HTRI and TRED databases, but supported by recent literature. This suggests that some of the other false positives that were supported by multiple DBN experiments could be novel interactions. In the absence of a gold standard, comparing false positives from different priors may reveal potential new interactions.

Due to the performance of DNA methylation in this study, future work should be geared towards using this data type to improve regulatory network inference in humans. If DNA methylation is highly correlated with a gene’s expression, it may be useful to use that information to detect downstream targets. Other sources of epigenetic or structural data should also be studied for this potential use.

CHAPTER 5

CARMMA: A computational pipeline to detect cancer-related miRNA-gene modules and associated disrupted biological processes

This chapter is a draft of a manuscript which will be submitted to a yet to be determined journal/conference. Some of the results are preliminary and future work is listed in Future Work section.

Abstract: microRNAs (miRNAs) regulate the expression of target genes by degradation of mRNA transcripts or repression of translation. Key miRNAs are dysregulated in cancer and can therefore disrupt important biological processes, such as cell cycle or apoptosis. Previous studies have proposed methods to uncover cancer-related miRNA-gene modules based on the relationship between miRNA expression and gene expression. In this study, we propose CARMMA, a computational pipeline to detect cancer-related miRNAs that are associated with target genes via disrupted biological processes and expression data. We applied CARMMA to luminal A breast and bladder cancer datasets from the TCGA Project. We found that the miRNA-gene modules formed by CARMMA are enriched in known interactions from the miRWalk database. Additionally, the miRNAs selected by CARMMA are enriched in known cancer-related miRNAs. We also examined the relationship between the expression of selected miRNAs and new tumor events after initial treatment. Overall, our results suggest that forming miRNA-gene modules based on biological processes can uncover important miRNAs in cancer.

5.1. Introduction

microRNAs (miRNAs) are short, non-coding molecules of RNA (Ambros, 2004). miRNAs decrease gene expression by destabilizing or cleaving mRNA transcripts or repressing translation (Lima et al., 2011). The expression of certain key miRNAs is known to be altered in cancer cells (Lu et al., 2005). Furthermore, miRNAs dysregulate many processes in cancer and are frequently located at fragile sites in genomic regions involved in cancer (Lima et al., 2011; Calin et al., 2004).

To elucidate the role of miRNAs in cancer, several tools have been developed to detect cancer related miRNAs and their associated target genes. Karim et al. outlined a methodology to infer miRNA-gene modules through collective group relationships (Karim et al., 2016). This methodology grouped miRNAs with similar targets and genes targeted by similar miRNAs and then established relationships between groups of miRNAs and groups of genes through canonical correlation analysis. Due to the heterogeneity of cancer, Jin and Lee proposed a biclustering approach to uncover gene-sample modules and then utilized a Bayesian network approach to connect candidate miRNAs to the genes in the gene-sample module (Jin and Lee, 2015).

However, these studies do not consider biological processes when building the modules of miRNAs and target genes. Certain biological processes are known to be dysregulated in cancer tumors via miRNAs, such as apoptosis (Lima et al., 2011) and cell cycle (Kim et al., 2009). Therefore, cancer-related miRNAs are more likely to target certain processes. Integrating information about disrupted processes could aid in elucidating the role of a particular miRNA in cancer.

In this study, we developed CARMMA, a tool that detects driver miRNAs, associated targets and disrupted biological processes. To our knowledge, CARMMA is the first tool that builds cancer-related miRNA-gene modules based on disrupted biological processes. The differentially expressed miRNAs are associated with biological process GO terms through the differentially expressed mRNAs. Modules are built based on the associated terms, and then refined by a LASSO-based method and binding site sequence information. We demonstrate that the modules found by CARMMA are enriched in interactions between the miRNAs and genes. Additionally, we show that the miRNAs detected by CARMMA are enriched for known cancer-related miRNAs. Finally, we examine the relationship between the expression of miRNAs detected by CARMMA and new tumor events after initial treatment using survival analysis. The survival analysis implicated two miRNAs selected by CARMMA, hsa-mir-185 and hsa-mir-141 as potential tumor suppressors in bladder cancer.

5.2. Methods

CARMMA is a tool to detect cancer-related miRNA-mRNA modules, along with associated disrupted biological processes. First, differentially expressed miRNAs are associated with GO biological process terms based on their relationship with differentially expressed genes. Next, preliminary modules of miRNAs and potential target genes are formed based on the associated biological processes. Finally, each module is refined to select candidate cancer-related miRNAs and target genes.

5.2.1. Data pre-processing and normalization

The scaling factors for the library sizes that minimize the log-fold changes between samples were computed using the trimmed-mean of M (TMM) values between each pair of samples in the edgeR package (Robinson et al., 2010). After accounting for the compositional biases between libraries with TMM normalization, the log counts per million (cpm) was used for expression in subsequent analysis, except for differential expression analysis.

DESeq2 was used to determine miRNAs and genes that are differentially expressed in normal versus cancer samples on miRNA-Seq and RNA-Seq raw counts, respectively (Love et al., 2014). A gene was considered differentially expressed if the adjusted p-value was < 0.001. A miRNA was considered differentially expressed if the adjusted p-value was < 0.0001.

5.2.2. GO term association step

In the following steps, we used the log(cpm) after TMM normalization for expression values. For each differentially expressed miRNA, we computed the correlation between the miRNA expression and the expression of each differentially expressed gene. This vector was used as a score for in a Kolmorgorov-Smirnov (KS) test to determine significant GO terms using the topGO package in R (Alexa and Rahnenfuhrer, 2016). Since miRNAs generally decrease gene expression, the KS test determined whether genes annotated with a particular GO term were more negatively associated with the miRNA. By this approach, differentially expressed miRNAs were associated with up to ten biological process GO terms through the differentially expressed genes.

We clustered semantically similar GO terms associated with the miRNAs as previously described for ProcessDriver (Section 3.2.1.2A). For each GO term cluster, we defined *GO term module* as the collection of differentially expressed miRNAs that were significantly associated with at least one GO term in that GO term cluster, and the differentially expressed genes that were annotated with at least one GO term in that GO term cluster.

5.2.3. GO term module refinement step

For each GO term module with *m* significantly associated miRNAs and *n* annotated genes, we computed a LASSO regression *n* times using the glmnet package in R (Friedman et al., 2010), each time with a different gene expression as a response and all of the *m* miRNAs in the GO term module as predictors (Algorithm 5.1). The value of λ that produced the sparest model in which the cross-validation error was within one standard error of the minimum error was used. For each miRNA, we computed *numTS*, the number of times the miRNA was selected in the *n* instances of LASSO. miRNAs that were in the top 80th percentile of all the *m numTS* values were considered further.

Algorithm 5.1. Module refinement step of CARMMA
Input:
For k GO term modules
Matrices of gene expression for genes in each module $Y_1, \dots, Y_k$
Matrices of miRNA expression for miRNAs in each module $X_1, \dots, X_k$

```
1. For i=1:k:
2.         $Y_i = \{y_1, \dots, y_n\}$ gene expression associated with module i with n genes
3.         $X_i = \{x_1, \dots, x_m\}$ miRNA expression associated with module i with m miRNAs
4.         $numTS = zeros(m)$
5.         For j=1:n
6.                 Compute LASSO regression with $y_j$ as a response and $X_i$ as predictors
7.                 For selected miRNA(s) $P$, $numTS_P = numTS_P + 1$
8.         End for
9.         Select s miRNAs in the top 80th percentile of numTS
10.        For q=1:s:
11.                Find binding sites in the 3' UTR of the genes miRNA q was selected for
12.                in steps 5-7
13.                Targets: genes miRNA q was selected for with binding sites for
14.                miRNA q in the 3' UTR
15.                if length(Targets) > 50:
16.                        Report miRNA q and Targets in module
17.                End if
18.        End for
19. End for
```

Next, for each of the *s* remaining miRNAs, we determined whether the genes that miRNA was selected for had binding sites in the 3’ UTR. We downloaded 3’ UTR sequences from BioMart (Dunrinck et al., 2009) and miRNA seed sequences from miRBase (Kozomara and Griffiths-Jones, 2014). We determined miRNA target sites in the 3’ UTRs using miRanda with an alignment threshold score of 130 (Enright et al., 2003). The number of targets associated with a miRNA is an important indicator of whether the miRNA is cancer-related (Jin and Lee, 2016). Therefore, if there were more than 50 genes that the miRNA was selected for in LASSO with a binding site in the 3’ UTR, the miRNA is part of the module and the more than 50 genes are its targets.

5.2.4. Data

We applied CARMMA to Illumina HiSeq 2000 RNA sequencing and miRNA sequencing data for 92 luminal A breast cancer samples and 118 bladder cancer samples from the TCGA database.

5.3. Results

5.3.1. GO term modules are enriched for known interactions

In bladder cancer, CARMMA created 21 GO term modules. After GO term module refinement step, four of the 21 modules did not select a miRNA, and were therefore discarded. In luminal A breast cancer, CARMMA created 20 modules and two of the 20 modules did not select a miRNA. Therefore, there were 17 and 18 modules for breast and bladder cancer, respectively (Tables 5.1 and 5.2).

We obtained experimentally validated interactions between differentially expressed miRNAs and differentially expressed genes from the miRWalk 2.0 database (Dweep et al., 2015). For each GO term module, we computed a p-value for the enrichment of interactions between the selected miRNAs and genes. This was determined by randomly resampling differentially expressed miRNAs and genes 1000 times, equal to the number of miRNA and genes that were selected in the module, and calculating the number of known interactions each time.

Table 5.1. Enrichment of interactions for GO term modules in bladder cancer. The number of genes and number of miRNAs are included for each module, as well as the number of experimentally validated interactions in the miRWalk database. For each module, the p-value associated with the number of interactions in the miRWalk database was computed by resampling the same number of miRNA and genes in the module 1000 times.

GO Terms	**Number of miRNAs**	**Number of genes**	**Number of interactions in miRWalk**	**p-value**
mitotic nuclear division, cell division, nuclear division, organelle fission, chromosome organization, organelle organization	7	386	63	0.050
cell cycle process, cell cycle, mitotic cell cycle	2	215	13	0.074
cell communication, signaling, signal transduction, cell surface receptor signaling pathway	18	674	349	0.001
single organism signaling	18	656	324	0
single-multicellular organism process	19	762	334	0.002
multicellular organismal process	19	771	334	0.004
system process, muscle system process, muscle contraction	8	195	38	0.044
regulation of multicellular organismal process	11	290	113	0.001
system development, anatomical structure morphogenesis, developmental process, multicellular organismal development, neuron differentiation, anatomical structure development, generation of neurons	7	618	121	0.014
response to stimulus, cellular response to stimulus	7	815	115	0.082
biological adhesion, cell adhesion	4	110	9	0.164
DNA replication, translational elongation, regulation of nucleobase-containing compound, DNA metabolic process	2	396	27	0.048
mitotic cell cycle process	2	134	9	0.058
biological regulation, regulation of cellular process, regulation of biological process	2	220	15	0.045
cell cycle phase transition	2	92	9	0.02
single-organism developmental process	1	76	0	1
muscle structure development	1	90	1	0.51

Table 5.2. Enrichment of interactions for GO term modules in luminal A breast cancer. The number of genes and number of miRNAs are included for each module, as well as the number of experimentally validated interactions in the miRWalk database. For each module, the p-value associated with the number of interactions in the miRWalk database was computed by resampling the same number of miRNA and genes in the module 1000 times.

GO terms	**Number of miRNAs**	**Number of genes**	**Number of interactions in miRWalk**	**p-value**
mitotic cell cycle process	6	144	22	0.042
cell cycle process, cell cycle, mitotic cell cycle	11	267	70	0.032
mitotic nuclear division, organelle fission, organelle organization, nuclear division, cell division	14	471	120	0.074
DNA metabolic process	7	121	21	0.043
single-multicellular organism process	14	1113	240	0.115
multicellular organismal process	14	1135	221	0.119
system process, neurological system process	14	323	54	0.155
signaling, cell communication, signal transduction, cell surface receptor signaling pathway, G-protein coupled receptor signaling pathway	12	975	181	0.101
single organism signaling	12	968	185	0.084
response to stimulus, cellular response to stimulus	11	1275	219	0.092
developmental process, multicellular organismal development, anatomical structure development, system development, cardiovascular system development, circulatory system development, vasculature development, blood vessel development, tissue development, organ development	11	930	156	0.082
single-organism developmental process	7	895	96	0.077
cellular response to DNA damage stimulus	1	58	0	1
regulation of multicellular organismal process	6	410	43	0.061
regulation of developmental process	1	224	21	0.004
biological regulation	1	1094	0	1
single-organism organelle organization	1	209	2	0.438
mitotic cell cycle phase transition	1	61	0	1

5.3.2. CARMMA miRNAs are enriched for known cancer-related miRNAs

To assess whether the miRNAs selected in the GO term modules were more cancer-related than those that were not, we downloaded disease-associated miRNAs from the Human MicroRNA Disease Database (HMDD) and searched for miRNAs that were associated with neoplasms (Li et al., 2013). In bladder cancer, 33/43 (77%) of the selected miRNAs were associated with neoplasms as opposed to 58% (60/104) of the differentially expressed, but non-selected miRNAs. In breast cancer, 32 out of 36 (89%) of the selected miRNAs are associated with neoplasms in the HMDD database, as opposed to 43 out of 97 (42%) of the differentially expressed, non-selected miRNAs (Table 5.3). We also investigated more specific neoplasm-associations, breast and urinary bladder neoplasms. In luminal A breast cancer, there was a much higher percentage (24/36) of the miRNAs selected in GO term modules that were associated with breast neoplasms compared to differentially expressed miRNAs that were not selected (20/97). By randomly resampling 36 miRNAs from the 133 differentially expressed miRNAs 10,000 times, the p-value associated with selecting 24 or more breast neoplasm-associated miRNAs is 0. Using the same re-sampling technique, we calculated the p-value for neoplasm-associated miRNAs in breast and bladder cancer, and urinary bladder neoplasms in bladder cancer. The results are summarized in Table 5.3.

Table 5.3. Number of cancer-associated miRNAs that were selected in GO term modules by CARMMA versus not selected. For miRNAs that are selected in GO term modules, the percentage indicates the number of selected miRNAs that are associated with cancer in the respective database over the total number of miRNAs that were selected. A p-value for this percentage was determined by randomly resampling the number of selected miRNAs from the total number of differentially expressed miRNAs. For miRNAs that are not selected, the percentage indicates the total number of non-selected miRNAs associated with cancer in the respective database over the total number of non-selected miRNAs.

Bladder cancer				**Luminal A breast cancer**			
	Selected in modules with p-value		Not selected		Selected in modules with p-value		Not selected
Neoplasms (HMDD)	33/43 (77%)	0.02	60/104 (58%)	**Neoplasms (HMDD)**	32/36 (89%)	0	43/97 (42%)
Urinary bladder neoplasms (HMDD)	13/43 (30%)	0.135	21/104 (20%)	**Breast neoplasms (HMDD)**	24/36 (67%)	0	20/97 (21%)
miRCancer database	16/43 (37%)	0.17	29/104 (27%)	**miRCancer database**	16/36 (44%)	0.03	24/97 (25%)

We also downloaded cancer-related miRNAs from the miRCancer database (Xie et al., 2013). In bladder cancer, we found a higher percentage of miRNAs in the miRCancer database for the miRNAs that were selected in GO term modules versus the differentially expressed miRNAs that were not selected in both breast and bladder cancer (Table 5.3). A summary of the number of articles found in both the HMDD and miRCancer database for the selected miRNAs are presented in Tables 5.4 and 5.5.

Table 5.4. Number of articles in the HMDD and miRCancer database for selected miRNAs in bladder cancer. Number of unique articles for each miRNA that associate the miRNA with cancer in both of the databases.

miRNA	miRCancer	HMDD	miRNA	miRCancer	HMDD
hsa-mir-21	0	93	hsa-mir-185	0	2
hsa-mir-143	0	34	hsa-mir-147b	0	2
hsa-let-7c	20	23	hsa-mir-28	0	1
hsa-mir-17	0	23	hsa-mir-33a	0	1
hsa-mir-205	0	22	hsa-mir-455	4	1
hsa-mir-92a-1	0	18	hsa-mir-942	1	1
hsa-mir-210	0	17	hsa-mir-944	3	1
hsa-mir-18a	0	14	hsa-mir-345	0	1
hsa-mir-141	20	13	hsa-mir-33b	0	1
hsa-mir-183	0	10	hsa-mir-454	7	1
hsa-mir-19a	29	10	hsa-mir-584	1	1
hsa-mir-135b	15	10	hsa-mir-548ba	0	0
hsa-mir-218-1	0	9	hsa-mir-3934	0	0
hsa-mir-96	28	8	hsa-mir-4652	0	0
hsa-mir-106b	0	7	hsa-mir-4746	0	0
hsa-mir-204	0	7	hsa-mir-671	2	0
hsa-mir-195	0	6	hsa-mir-1307	0	0
hsa-mir-139	18	5	hsa-mir-940	5	0
hsa-mir-23b	0	5	hsa-mir-4664	0	0
hsa-mir-30a	27	4	hsa-mir-1247	0	0
hsa-mir-130b	0	3	hsa-mir-504	2	0
hsa-mir-301b	5	3			

Table 5.5. Number of articles in the HMDD and miRCancer database for selected miRNAs in breast cancer. Number of unique articles for each miRNA that associate the miRNA with cancer in both of the databases.

miRNA	miRCancer	HMDD	miRNA	miRCancer	HMDD
hsa-mir-21	0	93	hsa-mir-378a	0	5
hsa-mir-145	0	55	hsa-mir-342	0	5
hsa-mir-125b-1	0	26	hsa-mir-486	8	4
hsa-mir-200b	11	23	hsa-mir-497	32	4
hsa-mir-200a	21	21	hsa-mir-193a	0	3
hsa-mir-210	0	17	hsa-mir-129-1	0	3
hsa-mir-10b	0	15	hsa-mir-495	11	3
hsa-mir-141	20	13	hsa-mir-452	7	2
hsa-mir-182	0	12	hsa-mir-337	4	2
hsa-mir-29a	0	11	hsa-mir-148b	0	2
hsa-mir-183	0	10	hsa-mir-488	2	2
hsa-mir-429	8	9	hsa-mir-584	1	1
hsa-mir-218-2	0	8	hsa-mir-33b	0	1
hsa-mir-96	28	8	hsa-mir-381	0	1
hsa-mir-140	16	7	hsa-mir-190b	0	0
hsa-mir-204	0	7	hsa-mir-374c	0	0
hsa-mir-32	10	6	hsa-mir-592	4	0
hsa-mir-139	18	5	hsa-mir-203a	0	0

5.3.3. miRNAs selected by CARMMA are related to new tumor events after initial treatment

miRNAs that promote tumorigenesis may have metastatic properties (Pencheva and Tavazoie, 2013). We examined the relationship between the number of days to new tumor events after initial treatment and the expression of the selected miRNAs in bladder cancer. We discretized the miRNA expression into up/down expressed tumor samples using a log 1.2-fold-change over the median expression value. Baseline samples were removed from the analysis. We then created Kaplan-Meier (KM) plots and looked for significant differences in the number of days to new tumor events after initial treatment between up and down expressed samples. Ninety-five patients in the bladder cancer cohort experienced a new tumor event after initial treatment.

Only 2 of the 92 luminal A breast cancer patients experienced new tumor events, therefore we applied this analysis to bladder cancer data only.

We found that miRNA hsa-mir-185 had the most significant difference between the number of days to new tumor event of up and down expressed samples (p = 0.00013). hsa-mir-185 has been implicated as a tumor suppressor in multiple studies across multiple cancers, but not bladder cancer. Specifically, it is linked to cell cycle arrest in lung cancers and inhibition of proliferation in colorectal cells (Takahashi et al., 2009; Liu et al., 2011). Additionally, we found a significant difference in the number of days to new tumor events between up and down expressed levels of hsa-mir-141 (p = 0.01). This miRNA inhibits pancreatic cancer cell invasion and migration (Xu et al., 2014; Wang et al., 2015). These results suggest that these two miRNAs are good candidate tumor suppressor miRNAs in bladder cancer.

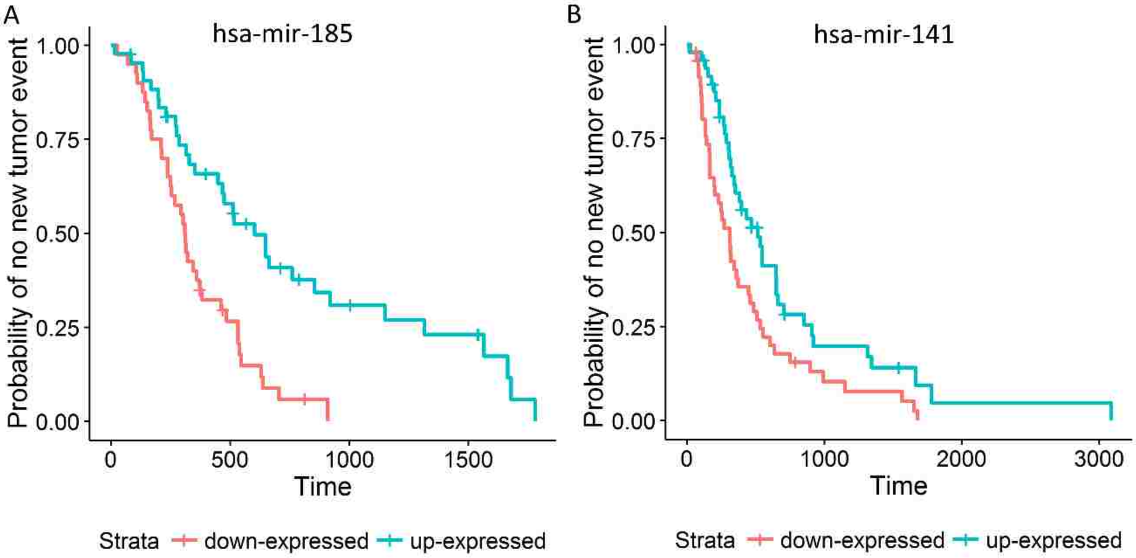


Figure 5.1. Kaplan-Meier plots for the number of days to new tumor events after initial treatment versus miRNA expression

5.4. Conclusions and future work

We designed and implemented CARMMA, a tool to detect cancer-associated miRNA, targets and disrupted biological processes. We applied CARMMA to luminal A breast cancer and bladder cancer datasets from the TCGA database. To our knowledge, CARMMA is the first algorithm to build miRNA-gene modules based on biological processes. We found that the GO term modules produced by CARMMA are enriched in known interactions between the miRNAs and genes. We also found that the miRNAs selected by CARMMA are enriched in known cancer-associated miRNAs.

Our results based on survival analysis indicate potential miRNA biomarker for bladder cancer. Particularly, given its tumor suppressor activities in other cancers, hsa-mir-185 is a potential biomarker for bladder cancer. Additionally, hsa-mir-141 is another potentially novel biomarker for bladder cancer. Based on the relationship with the number of days to new tumor events, both miRNAs can be investigated for tumor suppressor activities in bladder cancer.

Overall, these preliminary results highlight the potential value of creating modules based on biological processes. This can aid in elucidating the potential process(es) that a miRNA disrupts in cancer. In the future, we plan to integrate DNA methylation and copy number data such that the relationship between miRNA and gene expression cannot be confounded by these factors. Additionally, we will compare CARMMA to previously published methods to determine the overall improvement of this method over other cancer-related miRNA-gene module detection methods.

CHAPTER 6

Conclusions and future work

This dissertation presented four tools that integrate multiple types of biological data to gain meaningful insights about the genes, miRNAs and interactions involved in cancer. Much of the focus has been on interactions between regulatory genes, miRNAs and associated downstream target genes. Previous work treated structural and epigenetic effects as confounding factors when examining relationships between regulatory genes, miRNAs and target genes. This dissertation examines the utility of structural and epigenetic information to aid in proposing candidate cancer driver genes and miRNAs, as well as their interactions. Particularly, Chapter 3 examined finding driver, regulatory genes within a copy number aberration and Chapter 4 examined leveraging structural and epigenetic states of regulators to establish relationships with targets.

6.1. Contributions of dissertation

The four tools developed in this dissertation may be useful towards future research. Our paper on selecting DNA methylation probes that are most predictive of gene expression can be useful to researchers who are working with 450K DNA methylation data (Chapter 2). Since there is an average of ~18 probes per gene, choosing the probe(s) that best represent the overall DNA methylation is important for downstream functional analysis. Some of the previous studies focused on only probes in upstream regions, which may ignore functionally important DNA methylation from the gene body (Farré et al, 2015; Rica et al., 2013). Furthermore, some studies did not consider gene expression when choosing which DNA methylation probes to study (Selamat et al., 2012; Noushmehr et al., 2010). In this work, we provided a comprehensive analysis of feature selection and classification methods for selecting the DNA methylation probe(s) that are most predictive of gene expression.

Since the algorithm was developed, the Infinium MethylationEPIC BeadChip microarray, covering 850,000 CpG methylation sites was developed (Moran et al., 2016). Since the sequential-forward selection with K-nearest neighbors algorithm is computationally inexpensive and the 850K array is similar to the 450K array, we hope that as data for the 850K arrays becomes available, this algorithm will be successful.

ProcessDriver was developed to compute candidate copy number based cancer driver genes, potential targets and associated biological processes (Chapter 3). We applied ProcessDriver to three cancer types and found that the drivers that were uncovered were enriched in known cancer genes. We also found that the drivers were associated with new tumor events using survival analysis. Aure et al. associated drivers with biological processes, but also used the correlation of copy number to cis gene expression to narrow down drivers (Aure et al., 2012). Other methods utilize the relationship of cis gene expression to trans gene expression to create modules of driver genes and associated targets, but do not use biological process information (Akavia et al., 2010). ProcessDriver's unique methodology builds modules based on biological processes and then narrows down drivers based on the relationship of cis gene expression to trans gene expression. In this work, we also demonstrated the value of utilizing the relationship between cis gene expression and trans gene expression to uncover drivers, as opposed to previous approaches that select the cis genes that are most correlated to their own copy number.

Our CCA/DBN algorithm demonstrated that structural and epigenetic aberrations can be leveraged to infer regulatory interactions, as opposed to being treated as a confounding factor (Chapter 4). In particular, leveraging DNA methylation states for regulators leads to an increased accuracy in the prediction of regulatory interactions, as opposed to using copy number or gene expression states for regulators. This methodology is unique because previous methodologies have mainly focused on the relationship between regulatory gene expression and target gene expression. In particular, the use of structural and epigenetic states for regulators should be examined in terms of its ability to establish directionality in a directed network.

In Chapter 5, we described our ongoing work to develop CARMMA, a tool to build miRNA-gene modules based on biological processes, allowing for additional insight into the processes a miRNA can disrupt. Our preliminary results propose some candidate cancer-associated miRNAs in bladder cancer, based on previous literature from other cancers and survival analysis. These and other miRNAs uncovered by CARMMA may be valuable to future research and could be therapeutic targets for future cancer drugs.

6.2. Future directions

As technology in biology develops, so does the need for computational tools that can integrate multiple types of biological data and aid in making meaningful insights. Due to the scale of the data available, genome-wide approaches are becoming more useful in prioritizing regulatory miRNA and genes of interest to cancer. Both of the most recent tools in this dissertation, ProcessDriver and CARMMA, take a genome-wide approach. However, the nuance of pin-pointing exact interactions between miRNAs or genes, as in the DBN approach, gets lost in these genome-wide approaches. Therefore, much of the future work should be geared towards also prioritizing cancer-related interactions genome-wide that have a greater chance of being accurate when experimentally validated.

It was noted in the dissertation that genes that are predicted well by DNA methylation are enriched in regulatory biological processes (Chapter 2). Additionally, utilizing DNA methylation for potential regulators achieved high accuracy in the CCA/DBN approach (Chapter 4). This suggests that genes controlled by DNA methylation may regulate many target genes in cancer. Therefore, the concept of epigenetic drivers should be examined further.

Most of the data used in this work was from the TCGA repository. These samples are composed of heterogeneous cell populations, and therefore may have various hidden confounding factors. In fact, many of these technologies rely on bulk RNA and DNA which only provides information about the average state of the cells present (Navin and Hicks, 2011). Solid tumors contain non-cancerous cells such as fibroblasts, lymphocytes and macrophages (Navin and Hicks, 2011). Furthermore, solid tumors could contain multiple clonal subtypes which implies that if multiple clones are present in a tumor, the data obtained is the average or more representative of the more dominant clone, which also may not be the most malignant (Navin and Hicks, 2011).

These multiple clones could confound analysis. Therefore, one exciting new area of research is in single cell sequencing. DNA sequencing technology has advanced to the point where little DNA is required, making it more feasible to analyze the DNA of single cells (Shapiro et al., 2013). Due to this technology, new studies have come out investigating intra tumor genetic heterogeneity in cancer development and response (Gawad et al., 2016). While the single-cell sequencing has mostly been limited to probing the transcriptome, a new method, scM&T, has been developed that performs simultaneous genome-wide sequencing of the transcriptome and methylome (Koch, 2016).

Single cell sequencing has promise in the field of personalized medicine. For example, only 5% to 10% of patients with ductal carcinoma *in situ* (DCIS) progress to an invasive carcinoma, and studies of DCIS suggests heterogeneity is present in the early stages (Navin and Hicks, 2011). If it is possible to ascertain the tumor heterogeneity in an individual clinical sample of a DCIS patient, it is possible to predict if that tumor is likely to become invasive (Navin and Hicks, 2011).

Additionally, applying copy number driver detection methods to samples of a single cell type within a tumor type may allow for more targeted therapies. More intensive focus could be placed on identifying therapeutic target genes (e.g. driver genes) and miRNAs within the more malignant clonal subtypes, as opposed to using the “averaged-out” bulk RNA and DNA data used currently. When single cell approaches to uncovering copy number and DNA methylation become more feasible and widely-adopted, identifying drivers disrupted by a copy number or DNA methylation change within a particular clonal subtype in a tumor can become possible.

BIBLIOGRAPHY

Adorjan, P., Distler, J., Lipscher, E., Model, F., Muller, J., Pelet, C., et al. (2002). Tumour class prediction and discovery by microarray-based DNA methylation analysis. *Nucleic Acids Research, 30*(5), e21-e21.

Akavia, U. D., Litvin, O., Kim, J., Sanchez-Garcia, F., Kotliar, D., Causton, H. C., et al. (2010). An integrated approach to uncover drivers of cancer. *Cell, 143*(6), 1005-1017.

Akhavan-Niaki, H., & Samadani, A. A. (2013). DNA methylation and cancer development: Molecular mechanism. *Cell Biochemistry and Biophysics, 67*(2), 501-513.

Ambatipudi, S., Gerstung, M., Pandey, M., Samant, T., Patil, A., Kane, S., et al. (2011). Genome wide expression and copy number analysis identifies driver genes in gingivobuccal cancers. *Genes, Chromosomes & Cancer, 51*(2), 161-173.

Ambros, V. (2004). The functions of animal microRNAs. *Nature, 431*(7006), 350-355.

Babatunde, O., Armstrong, L. J., Leng, J., & Dean, D. (2014). A genetic algorithm-based feature selection. *International Journal of Electronics Communication and Computer Engineering*, *5*(4), 889-905.

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., et al. (2012). NCBI GEO: Archive for functional genomics data sets update. *Nucleic Acids Research, 41*, D991-D995.

Baur, B., & Bozdag, S. (2015). A canonical correlation analysis-based dynamic bayesian network prior to infer gene regulatory networks from multiple types of biological data. *Journal of Computational Biology*, *22*(4), 289-299.

Baur, B., & Bozdag, S. (2016). A feature selection algorithm to compute gene centric methylation from probe level methylation data. *PloS One*, *11*(2), e0148977.

Bernard, A., & Hartemink, A. J. (2005). Informative structure priors: Joint learning of dynamic regulatory networks from multiple types of data. *Pacific Symposium on Biocomputing.Pacific Symposium on Biocomputing*, 459-470.

Bernardo, G. M., Lozada, K. L., Miedler, J. D., Harburg, G., Hewitt, S. C., Mosley, J. D., et al. (2010). FOXA1 is an essential determinant of ER-alpha expression and mammary ductal morphogenesis. *Development (Cambridge, England), 137*(12), 2045-2054.

Beroukhim, R., Getz, G., Nghiemphu, L., Barretina, J., Hsueh, T., Linhart, D., et al. (2007). Assessing the significance of chromosomal aberrations in cancer: Methodology and application to glioma. *Proceedings of the National Academy of Sciences of the United States of America, 104*(50), 20007-20012.

Beroukhim, R., Zhang, X., & Meyerson, M. (2017). Copy number alterations unmasked as enhancer hijackers. *Nature Genetics, 49*(1), 5-6.

Bibikova, M., Barnes, B., Tsan, C., Ho, V., Klotzle, B., Le, J. M., et al. (2011). High density DNA methylation array with single CpG site resolution. *Genomics; New Genomic Technologies and Applications, 98*(4), 288-295.

Bird, A. (2002). DNA methylation patterns and epigenetic memory. *Genes & Development, 16*(1), 6-21.

Bovolenta, L. A., Acencio, M. L., & Lemke, N. (2012). HTRIdb: An open-access database for experimentally verified human transcriptional regulation interactions. *BMC Genomics, 13*(1), 405.

Bozdag, S., Li, A., Wuchty, S., & Fine, H. A. (2010). FastMEDUSA: A parallelized tool to infer gene regulatory networks. *Bioinformatics, 26*(14), 1792-1793.

Braig, S., & Bosserhoff, A. (2013). Death inducer-obliterator 1 (Dido1) is a BMP target gene and promotes BMP-induced melanoma progression. *Oncogene, 32*(7), 837-848.

Brenet, F., Moh, M., Funk, P., Feierstein, E., Viale, A. J., Socci, N. D., et al. (2011). DNA methylation of the first exon is tightly linked to transcriptional silencing. *PloS One, 6*(1), e14524.

Calin, G. A., Sevignani, C., Dumitru, C. D., Hyslop, T., Noch, E., Yendamuri, S., et al. (2004). Human microRNA genes are frequently located at fragile sites and genomic regions Involved in cancers. *Proceedings of the National Academy of Sciences of the United States of America, 101*(9), 2999-3004.

Cavalli, L. R., Riggins, R. B., Wang, A., Clarke, R., & Haddad, B. R. (2010). Frequent loss of heterozygosity at the interferon regulatory factor-1 gene locus in breast cancer. *Breast Cancer Research and Treatment, 121*(1), 227-231.

Chatr-aryamontri, A., Breitkreutz, B., Oughtred, R., Boucher, L., Heinicke, S., Chen, D., et al. (2014). The BioGRID interaction database: 2015 update. *Nucleic Acids Research,* 43, D470-478

Chatterjee, A., Stockwell, P. A., Rodger, E. J., & Morison, I. M. (2012). Comparison of alignment software for genome-wide bisulphite sequence data. *Nucleic Acids Research,* 40(10), e79

Chavez, L., Bais, A. S., Vingron, M., Lehrach, H., Adjaye, J., & Herwig, R. (2009). In silico identification of a core regulatory network of OCT4 in human embryonic stem cells using an integrated approach. *BMC Genomics, 10*(1), 314.

Chen, H., Maduranga, D. A. K., Mundra, P. A., & Zheng, J. (2013). Integrating epigenetic prior in dynamic bayesian network for gene regulatory network inference. *Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2013 IEEE Symposium on,* pp. 76-82.

Chen, L., Li, Y., Lin, C. H., Chan, T. H. M., Chow, R. K. K., Song, Y., et al. (2013). Recoding RNA editing of AZIN1 predisposes to hepatocellular carcinoma. *Nature Medicine, 19*(2), 209-216.

Choi, M. R., An, C. H., Yoo, N. J., & Lee, S. H. (2016). Frameshift mutations of CAB39L, an activator of LKB1 tumor suppressor, in gastric and colorectal cancers. *Pathology & Oncology Research, 22*(1), 225-226.

Choong, L. Y., Lim, S., Chong, P. K., Wong, C. Y., Shah, N., & Lim, Y. P. (2010). Proteome wide profiling of the MCF10AT breast cancer progression model. *Plos One, 5*(6), e11030.

Coccia, E. M., Del Russo, N., Stellacci, E., Orsatti, R., Benedetti, E., Marziali, G., et al. (1999). Activation and repression of the 2-5A synthetase and p21 gene promoters by IRF-1 and IRF-2. *Oncogene, 18*(12), 2129-2137.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273-297.

Dawson, M., & Kouzarides, T. Cancer epigenetics: From mechanism to therapy. *Cell, 150*(1), 12-27.

Di Leva, G., Garofalo, M., & Croce, C. M. (2014). MicroRNAs in cancer. *Annual Review of Pathology, 9*, 287-314.

Doll, J. A., Stellmach, V. M., Bouck, N. P., Bergh, A. R., Lee, C., Abramson, L. P., et al. (2003). Pigment epithelium-derived factor regulates the vasculature and mass of the prostate and pancreas. *Nature Medicine, 9*(6), 774-780.

Dong, S., Pang, J. C., Hu, J., Zhou, L. F., & Ng, H. K. (2002). Transcriptional inactivation of TP73 expression in oligodendroglial tumors. *International Journal of Cancer, 98*(3), 370 375.

Down, T. A., Rakyan, V. K., Turner, D. J., Flicek, P., Li, H., Kulesha, E., et al. (2008). A bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. *Nat Biotech, 26*(7), 779-785.

Durinck, S., Spellman, P. T., Birney, E., & Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nature Protocols, 4*(8), 1184-1191.

Dweep, H., & Gretz, N. (2015). miRWalk2.0: A comprehensive atlas of microRNA-target interactions. *Nature Methods, 12*(8), 697.

Eden, E., Navon, R., Steinfeld, I., Lipson, D., & Yakhini, Z. (2009). GOrilla: A tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics, 10*, 48 2105-10-48.

Enright, A. J., John, B., Gaul, U., Tuschl, T., Sander, C., & Marks, D. S. (2003). MicroRNA targets in drosophila. *Genome Biology, 5*(1), R1.

Ettahar, A., Ferrigno, O., Zhang, M., Ohnishi, M., Ferrand, N., Prunier, C., et al. (2013). Identification of PHRF1 as a tumor suppressor that promotes the TGF-Î² cytostatic program through selective release of TGIF-driven PML inactivation. *Cell Reports, 4*(3), 530-541.

Evan, G. I., & Vousden, K. H. (2001). Proliferation, cell cycle and apoptosis in cancer. *Nature, 411*(6835), 342-348.

Fan, B., Dachrut, S., Coral, H., Yuen, S. T., Chu, K. M., Law, S., et al. (2012). Integration of DNA copy number alterations and transcriptional expression analysis in human gastric cancer. *Plos One, 7*(4), e29824.

Fan, S., Wang, J. A., Yuan, R. Q., Ma, Y. X., Meng, Q., Erdos, M. R., et al. (1998). BRCA1 as a potential human prostate tumor suppressor: Modulation of proliferation, damage responses and expression of cell regulatory proteins. *Oncogene, 16*(23), 3069-3082.

Farré, P., Jones, M. J., Meaney, M. J., Emberly, E., Turecki, G., & Kobor, M. S. (2015). Concordant and discordant DNA methylation signatures of aging in human blood and brain. *Epigenetics & Chromatin, 8,* 10.1186/s13072-015-0011-y.

Feuk, L., Carson, A. R., & Scherer, S. W. (2006). Structural variation in the human genome. *Nature Reviews. Genetics, 7*(2), 85-97.

Foulstone, E. J., Zeng, L., Perks, C. M., & Holly, J. M. (2013). Insulin-like growth factor binding protein 2 (IGFBP-2) promotes growth and survival of breast epithelial cells: Novel regulation of the estrogen receptor. *Endocrinology, 154*(5), 1780-1793.

Friedman, J. H., Bentley, J. L., & Finkel, R. A. (1977). An algorithm for finding best matches in logarithmic expected time. *ACM Trans.Math.Softw., 3*(3), 209-226.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software, 33*(1), 1-22.

Frohlich, H., Speer, N., Poustka, A., & Beissbarth, T. (2007). GOSim--an R-package for computation of information theoretic GO similarities between terms and gene products. *BMC Bioinformatics, 8*, 166.

Frontini, M., Vijayakumar, M., Garvin, A., & Clarke, N. (2009). A ChIP-chip approach reveals a novel role for transcription factor IRF1 in the DNA damage response. *Nucleic Acids Research, 37*(4), 1073-1085.

Furic, L., Rong, L., Larsson, O., Koumakpayi, I. H., Yoshida, K., Brueschke, A., et al. (2010). eIF4E phosphorylation promotes tumorigenesis and is associated with prostate cancer progression. *Proceedings of the National Academy of Sciences of the United States of America, 107*(32), 14134-14139.

Gawad, C., Koh, W., & Quake, S. R. (2016). Single-cell genome sequencing: Current state of the science. *Nature Reviews. Genetics, 17*(3), 175-188.

Gevaert, O., Tibshirani, R., & Plevritis, S. K. (2015). Pancancer analysis of DNA methylation driven genes using MethylMix. *Genome Biology, 16*(1), 17.

Gilli, S. C., Salles, T. S., & Saad, S. T. (2004). Regulation of the GATA3 promoter by human T cell lymphotropic virus type I tax protein. *Journal of Cellular Biochemistry, 93*(6), 1178 1187.

Gu, H., Smith, Z. D., Bock, C., Boyle, P., Gnirke, A., & Meissner, A. (2011). Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling. *Nature Protocols, 6*(4), 468-481.

Gundem, G., Perez-Llamas, C., Jene-Sanz, A., Kedzierska, A., Islam, A., Deu-Pons, J., et al. (2010). IntOGen: Integration and data mining of multidimensional oncogenomic data. *Nat Meth, 7*(2), 92-93.

Haigermoser, C., Fujimoto, M., Iguchi-Ariga, S. M. M., & Ariga, H. (1996). Cloning and characterization of the genomic DNA of the human MSSP genes. *Nucleic Acids Research, 24*(19), 3846-3857.

Hanahan, D., & Weinberg, R. A. (2000). The hallmarks of cancer. *Cell, 100*(1), 57.

Hanahan, D., & Weinberg, R. (2011). Hallmarks of cancer: The next generation. *Cell, 144*(5), 646-674.

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika, 57*(1), 97-109.

Hecker, M., Lambeck, S., Toepfer, S., van Someren, E., & Guthke, R. (2009). Gene regulatory network inference: Data integration in dynamic models"A review. *Biosystems, 96*(1), 86-103.

Henrichsen, C. N., Chaignat, E., & Reymond, A. (2009). Copy number variants, diseases and gene expression. *Human Molecular Genetics, 18*(R1), R1-R8.

Henrichsen, C. N., Vinckenbosch, N., Zollner, S., Chaignat, E., Pradervand, S., Schutz, F., et al. (2009). Segmental copy number variation shapes tissue transcriptomes. *Nature Genetics, 41*(4), 424-429.

Ho, T. C., Chen, S. L., Yang, Y. C., Liao, C. L., Cheng, H. C., & Tsao, Y. P. (2007). PEDF induces p53-mediated apoptosis through PPAR gamma signaling in human umbilical vein endothelial cells. *Cardiovascular Research, 76*(2), 213-223.

Hocking, R. R. (1976). A biometrics invited paper. the analysis and selection of variables in linear regression. *Biometrics,* 1-49.

Honore, B., Baandrup, U., Nielsen, S., & Vorum, H. (2002). Endonuclein is a cell cycle regulated WD-repeat protein that is up-regulated in adenocarcinoma of the pancreas. *Oncogene, 21*(7), 1123-1129.

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika, 28*(3-4), 321-377.

Huang, D. W., Sherman, B. T., Tan, Q., Collins, J. R., Alvord, W. G., Roayaei, J., et al. (2007). The DAVID gene functional classification tool: A novel biological module-centric algorithm to functionally analyze large gene lists. *Genome Biology, 8*(9), R183-2007-8-9 r183.

Huret, J. L., Ahmad, M., Arsaban, M., Bernheim, A., Cigna, J., Desangles, F., et al. (2013). Atlas of genetics and cytogenetics in oncology and haematology in 2013. *Nucleic Acids Research, 41*(Database issue), D920-4.

Husmeier, D. (2003). Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic bayesian networks. *Bioinformatics, 19*(17), 2271 2282.

Imoto, S., Higuchi, T., Goto, T., Tashiro, K., Kuhara, S., & Miyano, S. (2004). Combining microarrays and biological knowledge for estimating gene networks via Bayesian networks. *Journal of Bioinformatics and Computational Biology, 02*(01), 77-98.

Jensen, S. A., Calvert, A. E., Volpert, G., Kouri, F. M., Hurley, L. A., Luciano, J. P., et al. (2014). Bcl2L13 is a ceramide synthase inhibitor in glioblastoma. *Proceedings of the National Academy of Sciences, 111*(15), 5682-5687.

Jiang, X., Tan, J., Li, J., Kivimäe, S., Yang, X., Zhuang, L., et al. (2008). DACT3 is an epigenetic regulator of wnt/beta-catenin signaling in colorectal cancer and is a therapeutic target of histone modifications. *Cancer Cell, 13*(6), 529-541.

Jin, D., & Lee, H. (2015). A computational approach to identifying gene-microRNA modules in cancer. *PLOS Computational Biology, 11*(1), e1004042.

Jin, D., & Lee, H. (2016). Prioritizing cancer-related microRNAs by integrating microRNA and mRNA datasets. *Scientific Reports, 6*, 10.1038/srep35350.

John, G. H., & Langley, P. (1995). Estimating continuous distributions in bayesian classifiers. *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, Montreal, Quebec, Canada. pp. 338-345.

Jones, P. A. (2012). Functions of DNA methylation: Islands, start sites, gene bodies and beyond. *Nature Reviews. Genetics, 13*(7), 484-492.

Jones, P. A., & Baylin, S. B. (2007). The epigenomics of cancer. *Cell, 128*(4), 683-692.

Karolchik, D., Barber, G. P., Casper, J., Clawson, H., Cline, M. S., Diekhans, M., et al. (2014). The UCSC genome browser database: 2014 update. *Nucleic Acids Research, 42*(Database issue), D764-70.

Kim, Y., Yu, J., Han, T. S., Park, S., Namkoong, B., Kim, D. H., et al. (2009). Functional links between clustered microRNAs: Suppression of cell-cycle inhibitors by microRNA clusters in gastric cancer. *Nucleic Acids Research, 37*(5), 1672-1681.

Kloten, V., Becker, B., Winner, K., Schrauder, M. G., Fasching, P. A., Anzeneder, T., et al. (2013). Promoter hypermethylation of the tumor-suppressor genes ITIH5, DKK3, and RASSF1A as novel biomarkers for blood-based breast cancer screening. *Breast Cancer Research, 15*(1), R4.

Koch, L. (2016). Epigenomics: Parallel single-cell sequencing. *Nature Reviews. Genetics, 17*(3), 125-125.

Kozomara, A., & Griffiths-Jones, S. (2014). miRBase: Annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Research, 42*(Database issue), D68-73.

Kuijjer, M. L., Rydbeck, H., Kresse, S. H., Buddingh, E. P., Lid, A. B., Roelofs, H., et al. (2012). Identification of osteosarcoma driver genes by integrative analysis of copy number and gene expression data. *Genes, Chromosomes and Cancer*, *51*(7), 696-706.

Kumar, M. S., Pester, R. E., Chen, C. Y., Lane, K., Chin, C., Lu, J., et al. (2009). Dicer1 functions as a haploinsufficient tumor suppressor. *Genes & Development, 23*(23), 2700 2704.

Lambertz, I., Nittner, D., Mestdagh, P., Denecker, G., Vandesompele, J., Dyer, M. A., et al. (2010). Monoallelic but not biallelic loss of Dicer1 promotes tumorigenesis in vivo. *Cell Death and Differentiation, 17*(4), 633-641.

Lapik, Y. R., Fernandes, C. J., Lau, L. F., & Pestov, D. G. (2004). Physical and functional interaction between Pes1 and Bop1 in mammalian ribosome biogenesis. *Molecular Cell*, *15*(1), 17.

Lee, H., Dang, T. C., Lee, H., & Park, J. C. (2014). OncoSearch: Cancer gene search engine with literature evidence. *Nucleic Acids Research, 42,* W416-W421.

Li, H., Chiappinelli, K. B., Guzzetta, A. A., Easwaran, H., Yen, R. W., Vatapalli, R., et al. (2014). Immune regulation by low doses of the DNA methyltransferase inhibitor 5-azacitidine in common human epithelial cancers. *Oncotarget, 5*(3), 587-598.

Li, J., Ching, T., Huang, S., & Garmire, L. X. (2015). Using epigenomics data to predict gene expression in lung cancer. *BMC Bioinformatics, 16 Suppl 5,* S10-2105-16-S5-S10. Epub 2015 Mar 18.

Li, W., Lee, A., & Gregersen, P. K. (2009). Copy-number-variation and copy-number-alteration region detection by cumulative plots. *BMC Bioinformatics, 10*(1), S67.

Li, Y., Qiu, C., Tu, J., Geng, B., Yang, J., Jiang, T., et al. (2014). HMDD v2.0: A database for experimentally supported human microRNA and disease associations. *Nucleic Acids Research, 42*(Database issue), D1070-4.

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P., & Mesirov, J. P. (2011). Molecular signatures database (MSigDB) 3.0. *Bioinformatics, 27*(12), 1739-1740.

Lima, R. T., Busacca, S., Almeida, G. M., Gaudino, G., Fennell, D. A., & Vasconcelos, M. H. (2011). MicroRNA regulation of core apoptosis pathways in cancer. *European Journal of Cancer, 47*(2), 163.

Liu, M., Lang, N., Chen, X., Tang, Q., Liu, S., Huang, J., et al. (2011). miR-185 targets RhoA and Cdc42 expression and inhibits the proliferation potential of human colorectal cells. *Cancer Letters, 301*(2), 151-160.

Liu, X., Zhou, J., Zhou, N., Zhu, J., Feng, Y., & Miao, X. (2016). SYNJ2BP inhibits tumor growth and metastasis by activating DLL4 pathway in hepatocellular carcinoma. *Journal of Experimental & Clinical Cancer Research, 35*(1), 115.

Liu, Y., Xia, J., Sun, J., & Zhao, M. (2015). OCGene: A database of experimentally verified ovarian cancer-related genes with precomputed regulation information. *Cell Death & Disease, 6*(12), e2036.

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion or RNA-seq data with DESeq2. *Genome Biology, 15*(12), 1-21.

Lu, T., Lai, L., Tsai, M., Chen, P., Hsu, C., Lee, J., Hsiao, C., & Chuang, E.Y. (2011). Integrated analyses of copy number variations and gene expression in lung adenocarcinoma. *Plos One, 6*(9), 1-11.

Lu, J., Getz, G., Miska, E. A., Alvarez-Saavedra, E., Lamb, J., Peck, D., et al. (2005). MicroRNA expression profiles classify human cancers. *Nature, 435*(7043), 834-838.

Maeda, O., Ando, T., Ohmiya, N., Ishiguro, K., Watanabe, O., Miyahara, R., et al. (2014). Alteration of gene expression and DNA methylation in drug-resistant gastric cancer. *Oncology Reports, 31*(4), 1883-1890.

Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla Favera, R., et al. (2006). ARACNE: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics, 7 Suppl 1*, S7.

Masud Karim, S. M., Liu, L., Le, T. D., & Li, J. (2016). Identification of miRNA-mRNA regulatory modules by exploring collective group relationships. *BMC Genomics, 17*(1), 7.

Meissner, A., Mikkelsen, T. S., Gu, H., Wernig, M., Hanna, J., Sivachenko, A., et al. (2008). Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature, 454*(7205), 766-770.

Mermel, C. H., Schumacher, S. E., Hill, B., Meyerson, M. L., Beroukhim, R., & Getz, G. (2011). GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biology, 12*(4), R41-2011-12-4-r41. Epub 2011 Apr 28.

Nagashima, T., Shimodaira, H., Ide, K., Nakakuki, T., Tani, Y., Takahashi, K., et al. (2007). Quantitative transcriptional control of ErbB receptor signaling undergoes graded to biphasic response for cell differentiation. *The Journal of Biological Chemistry, 282*(6), 4045-4056.

Navin, N., & Hicks, J. (2011). Future medical applications of single-cell sequencing in cancer. *Genome Medicine, 3*(5), 31.

Ni, M., Chen, Y., Fei, T., Li, D., Lim, E., Liu, X. S., et al. (2013). Amplitude modulation of androgen signaling by c-MYC. *Genes & Development*, *27*(7), 734-748.

Noushmehr, H., Weisenberger, D. J., Diefes, K., Phillips, H. S., Pujara, K., Berman, B. P., et al. (2010). Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma. *Cancer Cell, 17*(5), 510-522.

Pickering, C. R., Zhang, J., Yoo, S. Y., Bengtsson, L., Moorthy, S., Neskey, D. M., et al. (2013). Integrative genomic characterization of oral squamous cell carcinoma identifies frequent somatic drivers. *Cancer Discovery, 3*(7), 770-781.

Quinlan, J. R. (1999). Simplifying decision trees. *International Journal of Human-Computer Studies, 51*(2), 497.

Rauhala, H. E., Teppo, S., Niemela, S., & Kallioniemi, A. (2013). Silencing of the ARP2/3 complex disturbs pancreatic cancer cell migration. *Anticancer Research, 33*(1), 45-52.

Rhee, J. K., Kim, K., Chae, H., Evans, J., Yan, P., Zhang, B. T., et al. (2013). Integrated analysis of genome-wide DNA methylation and gene expression profiles in molecular subtypes of breast cancer. *Nucleic Acids Research, 41*(18), 8464-8474.

Rica, L. d. l., Urquiza, J. M., Gómez-Cabrero, D., Islam, A. B. M. M. K., López-Bigas, N., Tegnér, J., et al. (2013). Identification of novel markers in rheumatoid arthritis through integrated analysis of DNA methylation and microRNA expression. *Journal of Autoimmunity, 41*, 6.

Ricote, M., Li, A. C., Willson, T. M., Kelly, C. J., & Glass, C. K. (1998). The peroxisome proliferator-activated receptor-gamma is a negative regulator of macrophage activation. *Nature, 391*(6662), 79-82.

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: A bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics, 26*(1), 139-140.

Rodriguez, J. A., Schuchner, S., Au, W. W., Fabbro, M., & Henderson, B. R. (2004). Nuclear cytoplasmic shuttling of BARD1 contributes to its proapoptotic activity and is regulated by dimerization with BRCA1. *Oncogene, 23*(10), 1809-1820.

Rokavec, M., Wu, W., & Luo, J. L. (2012). IL6-mediated suppression of miR-200c directs constitutive activation of inflammatory signaling circuit driving transformation and tumorigenesis. *Molecular Cell, 45*(6), 777-789.

Saeki, N., Usui, T., Aoyagi, K., Kim, D. H., Sato, M., Mabuchi, T., et al. (2009). Distinctive expression and function of four GSDM family genes (GSDMA-D) in normal and malignant upper gastrointestinal epithelium. *Genes, Chromosomes and Cancer, 48*(3),261-271.

Sanceau, J., Hiscott, J., Delattre, O., & Wietzerbin, J. (2000). IFN-beta induces serine phosphorylation of stat-1 in ewing's sarcoma cells and mediates apoptosis via induction of IRF-1 and activation of caspase-7. *Oncogene, 19*(30), 3372-3383.

Sanchez-Garcia, F., Akavia, U. D., Mozes, E., & Pe'er, D. (2010). JISTIC: Identification of significant targets in cancer. *BMC Bioinformatics, 11*, 189-2105-11-189.

Sato, N., Maeda, M., Sugiyama, M., Ito, S., Hyodo, T., Masuda, A., et al. (2015). Inhibition of SNW1 association with spliceosomal proteins promotes apoptosis in breast cancer cells. *Cancer Medicine, 4*(2), 268-277.

Scheidegger, K. J., Du, J., & Delafontaine, P. (1999). Distinct and common pathways in the regulation of insulin-like growth factor-1 receptor gene expression by angiotensin II and basic fibroblast growth factor. *The Journal of Biological Chemistry, 274*(6), 3522-3530.

Scholl, C., Fröhling, S., Dunn, I. F., Schinzel, A. C., Barbie, D. A., Kim, S. Y., et al. (2009). Synthetic lethal interaction between oncogenic KRAS dependency and STK33 suppression in human cancer cells. *Cell, 137*(5), 821-834.

Schuebel, K. E., Chen, W., Cope, L., Glockner, S. C., Suzuki, H., Yi, J. M., et al. (2007). Comparing the DNA hypermethylome with gene mutations in human colorectal cancer. *PLoS Genetics, 3*(9), 1709-1723.

Schubeler, D. (2015). Function and information content of DNA methylation. *Nature, 517*(7534), 321-326.

Scutari, M. (2010). Learning bayesian networks with the bnlearn R package. *Journal of Statistical Software, 35*(1), 1-22.

Selamat, S. A., Chung, B. S., Girard, L., Zhang, W., Zhang, Y., Campan, M., et al. (2012). Genome-scale analysis of DNA methylation in lung adenocarcinoma and integration with mRNA expression. *Genome Research, 22*(7), 1197-1211.

Setty, M., Helmy, K., Khan, A. A., Silber, J., Arvey, A., Neezen, F., et al. (2012). Inferring transcriptional and microRNA-mediated regulatory programs in glioblastoma. *Molecular Systems Biology, 8*, 605.

Shapiro, E., Biezuner, T., & Linnarsson, S. (2013). Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nature Reviews. Genetics, 14*(9), 618-630.

Shen, H., Cai, M., Zhao, S., Wang, H., Li, M., Yao, S., et al. (2014). Overexpression of RFC3 is correlated with ovarian tumor development and poor prognosis. *Tumor Biology, 35*(10), 10259-10266.

Siegel, R. L., Miller, K. D., & Jemal, A. (2016). Cancer statistics, 2016. CA: A Cancer Journal for Clinicians, 66(1), 7-30.

Smith, G. K. (2005). Limma: Linear models for microarray data. In V. C. Gentleman, & S. Dudoit (Eds.), *Bioinformatics and computational biology solutions using R and bioconductor* (pp. 397-420) Springer.

Stockinger, B., & Veldhoen, M. (2007). Differentiation and function of Th17 T cells. *Current Opinion in Immunology, 19*(3), 281-286.

Stone, A., Cowley, M. J., Valdes-Mora, F., McCloy, R. A., Sergio, C. M., Gallego-Ortega, D., et al. (2013). BCL-2 hypermethylation is a potential biomarker of sensitivity to antimitotic chemotherapy in endocrine-resistant breast cancer. *Molecular Cancer Therapeutics, 12*(9), 1874-1885.

Stratton, M. R. (2011). Exploring the genomes of cancer cells: Progress and promise. *Science, 331*(6024), 1553-1558.

Subauste, M. C., Sansom, O. J., Porecha, N., Raich, N., Du, L., & Maher, J. F. (2010). Fem1b, a proapoptotic protein, mediates proteasome inhibitor-induced apoptosis of human colon cancer cells. *Molecular Carcinogenesis, 49*(2), 105-113.

Takahashi, Y., Forrest, A. R. R., Maeno, E., Hashimoto, T., Daub, C. O., & Yasuda, J. (2009). MiR-107 and MiR-185 can induce cell cycle arrest in human non small cell lung cancer cell lines. *Plos One, 4*(8), e6677. doi:10.1371/journal.pone.0006677.

Tamborero, D., Lopez-Bigas, N., & Gonzalez-Perez, A. (2013). Oncodrive-CIS: A method to reveal likely driver genes based on the impact of their copy number changes on expression. *Plos One*, *8*(2), e55489.

The Cancer Genome Atlas Research Network. (2011). Integrated genomic analyses of ovarian carcinoma. *Nature*, 474, 609-615.

The Cancer Genome Atlas Research Network. (2012). Comprehensive molecular portraits of human breast tumours. *Nature, 490,* 61-70.

The Cancer Genome Atlas Research Network. (2014). Comprehensive molecular characterization of urothelial bladder carcinoma. *Nature, 507*(7492), 315-322.

Tsamardinos, I., Brown, L. E., & Aliferis, C. F. (2006). The max-min hill-climbing bayesian network structure learning algorithm. *Machine Learning, 65*(1), 31-78.

Tsarovina, K., Reiff, T., Stubbusch, J., Kurek, D., Grosveld, F. G., Parlato, R., et al. (2010). The Gata3 transcription factor is required for the survival of embryonic and adult sympathetic neurons. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience, 30*(32), 10833-10843.

Tucker, K. L. (2001). Methylated cytosine and the brain: A new base for neuroscience. *Neuron,* 30 (3), 649-652.

Varley, K. E., Gertz, J., Bowling, K. M., Parker, S. L., Reddy, T. E., Pauli-Behn, F., et al. (2013). Dynamic DNA methylation across diverse human cell lines and tissues. *Genome Research, 23*(3), 555-567.

Vincent, K., Pichler, M., Lee, G. W., & Ling, H. (2014). MicroRNAs, genomic instability and cancer. *International Journal of Molecular Sciences, 15*(8), 14475-14491.

von Arnim, C. A., Verstege, E., Etrich, S. M., & Riepe, M. W. (2006). Mechanisms of hypoxic tolerance in presymptomatic APP23 transgenic mice. *Mechanisms of Ageing and Development, 127*(2), 109-114.

Waaijenborg, S., Verselewel de Witt Hamer, P. C., & Zwinderman, A. H. (2008). Quantifying the association between gene expressions and DNA-markers by penalized canonical correlation analysis. *Statistical Applications in Genetics and Molecular Biology, 7*(1), 3-6115.1329. Epub 2008 Jan 23.

Wang, Q., Williamson, M., Bott, S., Brookman-Amissah, N., Freeman, A., Nariculam, J., et al. (2007). Hypomethylation of WNT5A, CRIP1 and S100P in prostate cancer. *Oncogene, 26*(45), 6560-6565.

Wang, X. L., Xie, H. Y., Zhu, C. D., Zhu, X. F., Cao, G. X., Chen, X. H., et al. (2015). Increased miR-141 expression is associated with diagnosis and favorable prognosis of patients with bladder cancer. *Tumour Biology: The Journal of the International Society for Oncodevelopmental Biology and Medicine, 36*(2), 877-883.

Wang, X., Wang, R. H., Li, W., Xu, X., Hollander, M. C., Fornace, A. J.,Jr, et al. (2004). Genetic interactions between Brca1 and Gadd45a in centrosome duplication, genetic stability, and neural tube closure. *The Journal of Biological Chemistry, 279*(28), 29606-29614.

Wang, Y., Su, M. A., & Wan, Y. Y. (2011). An essential role of the transcription factor GATA-3 for the function of regulatory T cells. *Immunity, 35*(3), 337-348.

Werhli Adriano, V., & Dirk, H. (2007). *Reconstructing gene regulatory networks with Bayesian networks by combining expression data with multiple sources of prior knowledge*

Witten, D. M., Robert Tibshirani, & Trevor Hastie. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics, 10*(3), 515-534.

Xie, B., Ding, Q., Han, H., & Wu, D. (2013). miRCancer: A microRNA-cancer association database constructed by text mining on literature. *Bioinformatics, 29*(5), 638-644.

Xu, C., Li, H., Zhang, L., Jia, T., Duan, L., & Lu, C. (0926). *MicroRNA19153p prevents the apoptosis of lung cancer cells by downregulating DRG2 and PBX2*

Xu, L., Li, Q., Xu, D., Wang, Q., An, Y., Du, Q., et al. (2014). Hsa-miR-141 downregulates TM4SF1 to inhibit pancreatic cancer cell invasion and migration. *International Journal of Oncology*, 44(2), 459-466.

Yamaguchi, K., Yamaguchi, R., Takahashi, N., Ikenoue, T., Fujii, T., Shinozaki, M., et al. (2014). Overexpression of cohesion establishment factor DSCC1 through E2F in colorectal cancer. *Plos One, 9*(1), e85750.

Yamashita, M., Toyota, M., Suzuki, H., Nojima, M., Yamamoto, E., Kamimae, S., et al. (2010). DNA methylation of interferon regulatory factors in gastric cancer and noncancerous gastric mucosae. *Cancer Science, 101*(7), 1708-1716.

Yan, K., & Zhang, D. (2015). Feature selection and analysis on correlated gas sensor data with recursive feature elimination. *Sensors and Actuators B: Chemical, 212*, 353-363.

Yang, S., Zhou, L., Reilly, P.T., Shen, S., He, P., Zhu, X., et al. (2016). ANP32B deficiency impairs proliferation and suppresses tumor progression by regulating AKT phosphorylation. *Cell Death Dis, 7*, e2082.

Yang, X., Han, H., De Carvalho, D., Lay, F., Jones, P., & Liang, G. (2014). Gene body methylation can alter gene expression and is a therapeutic target in cancer. Cancer Cell, 26(4), 577-590.

Zhang, W., Zong, C. S., Hermanto, U., Lopez-Bergami, P., Ronai, Z., & Wang, L. (2005). RACK1 recruits STAT3 specifically to insulin and insulin-like growth factor 1 receptors for activation, which is important for regulating anchorage-independent growth. *Molecular and Cellular Biology, 26*(2), 413-424.

Zhao, F., Xuan, Z., Liu, L., & Zhang, M. Q. (2005). TRED: A transcriptional regulatory element database and a platform for in silico gene regulation studies. *Nucleic Acids Research, 33*, D103-D107.

Zhao, X. M., Liu, K. Q., Zhu, G., He, F., Duval, B., Richer, J. M., et al. (2015). Identifying cancer-related microRNAs based on gene expression data. *Bioinformatics, 31*(8), 1226-1234.

Zheng, J., Chaturvedi, I., & Rajapakse, J. C. (2011). Integration of epigenetic data in bayesian network modeling of gene regulatory network. In M. Loog, L. Wessels, M. J. T. Reinders & D. de Ridder (Eds.), *Pattern recognition in bioinformatics: 6th IAPR international conference, PRIB 2011, delft, the netherlands, november 2-4, 2011. proceedings* (pp. 87-96). Berlin, Heidelberg: Springer Berlin Heidelberg.

Zindy, F., Eischen, C. M., Randle, D. H., Kamijo, T., Cleveland, J. L., Sherr, C. J., et al. (1998). Myc signaling via the ARF tumor suppressor regulates p53-dependent apoptosis and immortalization. *Genes & Development, 12*(15), 2424-2433.

Zou, M., & Conzen, S. D. (2005). A new dynamic bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data. *Bioinformatics, 21*(1), 71-79.