Causal Inference for the Relationship between DNA Methylation and Transcription in Breast Cancer

A Thesis

Presented in Partial Fulfilment of the Requirements for the

Degree of Master of Science

with a

Major in Statistical Science

in the

College of Graduate Studies

University of Idaho

by

Bandita Karki

Major Professor: Audrey Q. Fu, Ph.D.

Committee Members: Stephen M. Krone, Ph.D.; Christopher Williams, Ph.D.

Department Administrator: Hirotachi Abo, Ph. D.

May 2023

Abstract

DNA methylation, an epigenetic mechanism, plays an important role in transcription regulation and in complex diseases. Whereas methylation in gene promoters is known to be generally associated with silencing, the relationship between transcription and methylation in other parts of the gene is much less clear. Additionally, substantially different transcription and methylation profiles have been observed among breast cancer subtypes, but it is unclear whether and how these differences are influenced by different relationships between the two processes.

Here, we studied the relationships between transcription and methylation in estrogen receptor-positive (ER+) and negative (ER-) patients, using data from The Cancer Genome Atlas (TCGA) consortium and Genomics Data Common portal (GDC). We formulated trios, each consisting of the Copy Number Alteration (CNA) of a gene, expression (E) of this gene, and methylation (M) of a site located near or in the same gene. Since CNA is prevalent in cancer, it is a highly effective instrumental variable for this causal inference. In each subtype, we further derived principal components from genomewide expression and methylation data and identified those that are significantly associated with each trio as potential confounding variables.

We applied MRGN, a novel causal network inference method that accounts for many confounding variables under the principle of Mendelian randomization, to each of the 310,412 trios in each subtype. We further examined the features of methylation probes in mediation models as compared to the baseline models. Our analysis provides a first comprehensive picture of causal relationships between transcription and methylation in the two subtypes.

Acknowledgements

I would like to express my sincere gratitude to Dr. Audrey Fu, my major professor, for her guidance and support throughout my academic journey. Without her expertise and leadership, the extent of work I have accomplished would not have been possible. Additionally, I would like to thank Dr. Stephen Krone and Dr. Christopher Williams for serving as committee members. I will always be grateful for their contribution to my academic growth.

I would also like to extend my appreciation to Jarred Kvamme for providing the MRGN package used in the analysis and to Dr. Michelle Ward for her valuable insights and suggestions regarding the results. I would also like to acknowledge the Institute for Modeling, Collaboration, and Innovation (IMCI) for providing funding for my research project, and the Research Computing and Data Services (RCDS) for their provision of computing resources that were essential to the analysis process. Lastly, I am thankful to Jana Joyce, Melissa Gottschalk, and the rest of the staff within the Department of Mathematics and Statistical Science for their administrative and financial support, which have been instrumental to my success. Without the contributions of all these individuals and organizations, I would not be where I am today.

Dedication

I would like to dedicate this thesis to my parents and my sister, who have been my biggest cheerleaders and provided me with immense love and support. I would also like to dedicate this thesis to the rest of my family and friends who were there for me and inspired me throughout this journey. I am grateful for all of you and appreciate your role in shaping me into the person I am today.

Table of Contents

List of Tables

List of Figures

CHAPTER 1

Introduction

Gene regulatory networks have been an important area of research in understanding the biological processes involved in the formation of complex diseases. By examining the interactions between genes and their regulators, scientists can gain insight into how these networks influence cellular behavior and ultimately impact disease development.

One area of particular interest in gene regulatory network research is the study of causal relationships. Causal relationships are key in understanding the cause-and-effect relationships between an outcome and an exposure or risk factor. For instance, in the Framingham heart survey (1977) [22], researchers found that individuals with higher levels of HDL, commonly known as good cholesterol, had a lower risk of heart disease. This finding led to billions of dollars being invested in the development of drugs that targeted HDL to treat heart disease. However, these drugs ultimately proved to be ineffective. Although an inverse correlation between HDL and heart disease was established, it did not necessarily establish a causal relationship. As a result, it is essential to make causal inferences to determine whether a factor is the cause of an observed effect or merely associated with it [19].

The study of causal relationships also helps understand the regulation of genes in regard to other biological mechanisms. One such mechanism that has been the focus of much research is DNA methylation. DNA methylation plays a key role in regulating gene expression and is implicated in the development of disease. When a promoter region of DNA is methylated, it reduces expression. However, the relationship between methylation and gene expression in other parts of the gene is not as clear [8]. In **Figure 1.1**, we can observe the six major regions in a gene including TSS1500, TSS200, $5'$ UTR, 1st Exon, Body, and 3′ UTR (TSS is the Transcription Start Site and UTR is the Untranslated Region). When the CpG sites of the promoter region are unmethylated, the gene is expressed. However, methylation in those sites can lead to the suppression of the genes. We aim to understand these relationships in the remaining parts of the gene which have not been defined yet. Therefore, causal inference is essential to identify whether changes in one of these processes are causing a change in the other.

Additionally, DNA methylation and transcription profiles differ substantially between different breast cancer subtypes, such as estrogen receptor-positive (ER+) and estrogen receptor-negative (ER-) individuals [16, 20, 4, 3]. ER+ individuals have estrogen receptors in the cancer cells, and drugs that lower estrogen levels can be used to treat breast cancer, whereas ER- individuals do not have estrogen receptors. Therefore, the difference between these two subtypes can be an important factor to consider when investigating the causal effect between methylation and gene expression.

Mendelian Randomization (MR) is a widely used approach in causal inference [2]. It uses a genetic variant that has been randomized in humans through the Principal of Mendelian Randomization (PMR) as an instrumental variable (**Figure 1.2**). The genetic variant aids in studying the causal effect between the genetic variant and the outcome based on the treatment. Since the genetic variant remains unaffected by external factors, any relationship between the variant and the outcome can be explained by the exposure or the risk factor. In this case, CNA is the genetic variant denoted by C and the variables of interest are DNA methylation and gene expression denoted by M and E respectively.

To examine these relationships, we applied MRGN (Mendelian Randomization Graph Network), a novel causal network inference method, to each of the 310,412 trios in each subtype under PMR [13]. The trios were formed using the breast cancer data from The Cancer Genome Atlas (TCGA) and Genomics Data Common portal (GDC). MRGN uses the trios formed between the instrumental variable and the variables of interest to determine whether a causal relationship exists or not. It accounts for the confounding variables consisting of Principal Components (PCs) from the genome-wide methylation and gene expression data as well as the demographic information of individuals. The models are inferred based on conditional and/or marginal tests between the variables of interest. The trio data matrix contains the values from each of the three datasets for the common individuals between the datasets as well as one of the two cancer subtypes. Similarly, the confounder data matrix has a length similar to the trio data matrix depending on the number of individuals, however, it contains the highly correlated PCs in the column along with age and race of the selected individuals.

The inferred mediation models are further evaluated based on the three major regions in a gene: the TSS (including TSS1500 and TSS200), the gene body (including 1st Exon and Body), and the $5'/3'$ UTR as shown in **Figure 1.1**. Additionally, the features of the methylation probes, such as the CpG island location of the probe, its distance from a nearby island, and the length of the gene, give insights into how each of the mediation models is distributed [18]. This also helps identify the similarities and differences between the baseline and mediation models in the two breast cancer subtypes. Our approach provides refined results by accounting for confounding variables, improving on previous studies [6, 7, 9, 10] that have investigated the relationship between methylation and gene expression.

We have included these components in our data analysis pipeline, which we implemented in the R package MRTrios (discussed later). Our primary objective is to infer a causal network and identify the relationships between the variables of interest. To achieve this, we use mediation models to examine various probes' characteristics and highlight any noteworthy features. By leveraging these relationships, we aim to identify genes and gain insights that can help us develop targeted therapies for complex diseases. Our ultimate goal is to advance our understanding of the underlying biological mechanisms in complex diseases.

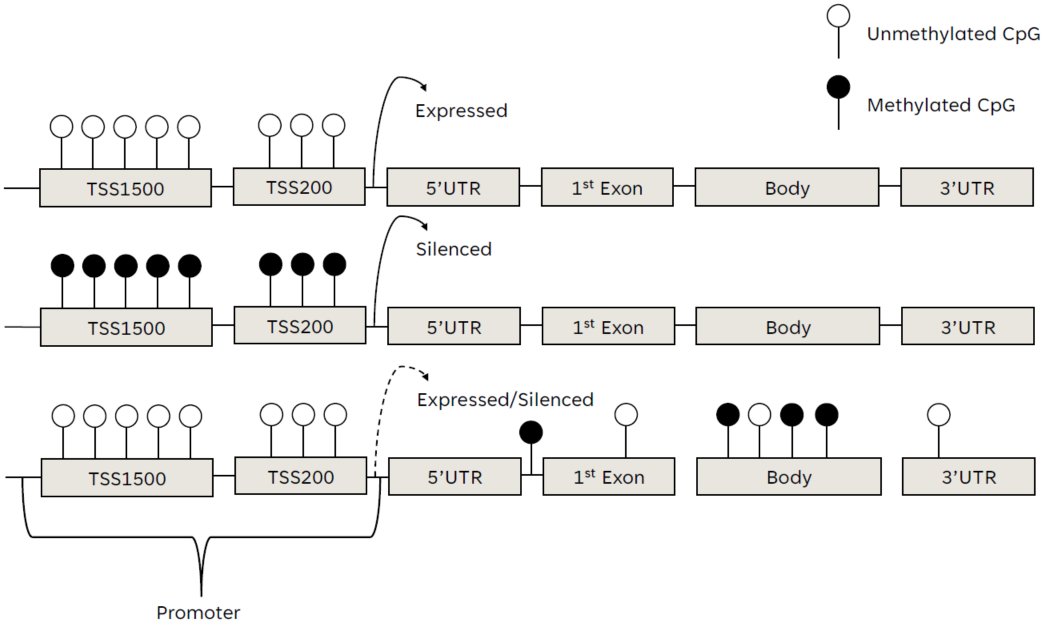

Figure 1.1: **DNA methylation and gene expression**. Methylation in the promoter of a gene can lead to silencing of the gene: the gene is not expressed or has much reduced expression. However, this is not always the case. The impact of methylation in other parts of a gene is even less clear.

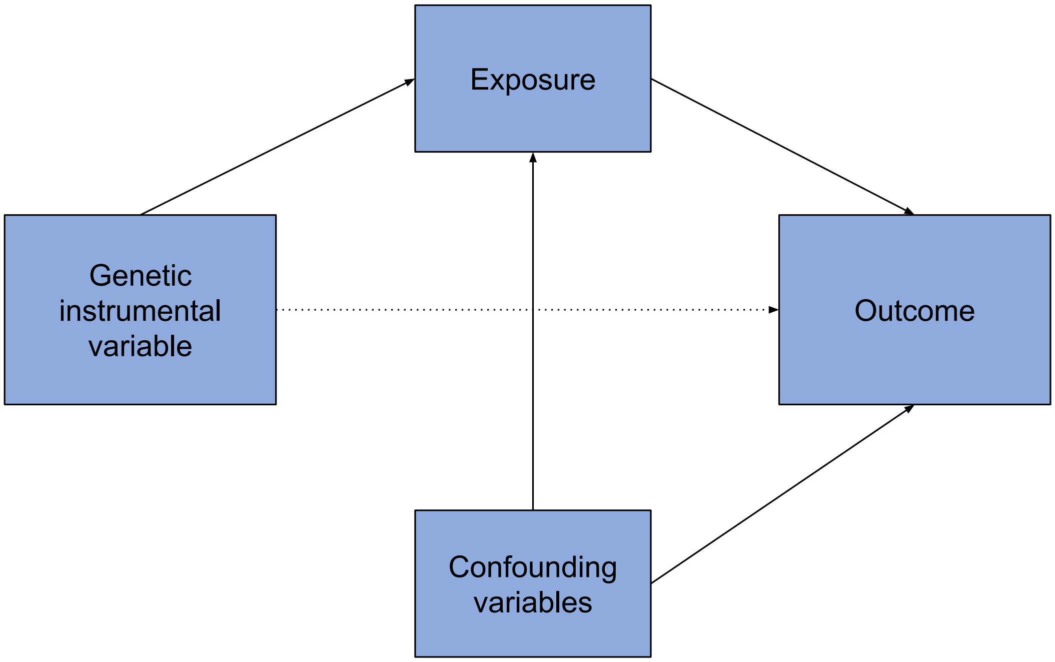

Figure 1.2: **Principle of Mendelian Randomization (PMR)**. The study of a causal relationship between a risk factor and outcome with the genetic variant as the instrumental variable. It helps classify the exposure into groups to identify the factor of influence.

CHAPTER 2

Materials and Methods

2.1 Overview of breast cancer data

We used The Cancer Genome Atlas (TCGA) breast cancer data downloaded from cbioportal [16] which consists of the gene expression, CNA, and clinical datasets. The methylation data is obtained from the Genomics Data Common portal (GDC) managed by the National Cancer Institute. The features of methylation probes are from Illumina and the features of the genes are from the Ensembl genome browser. The methylation and gene expression datasets consist of continuous values for methylation and expression in individuals whereas the CNA values range from -2 to 2 (integer) depending on the loss or gain of DNA sequence. The clinical datasets contain the individual information i.e. ID, age, and race of individuals.

The dimensions for the included datasets are as follows:

1. Methylation: 485577 probes and 895 individuals;
2. Gene Expression: 20531 genes and 1100 individuals;
3. Copy Number Alteration (CNA): 24776 genes and 1080 individuals;
4. Clinical data: 1053 individuals (814 ER+ and 239 ER-);
5. Methylation probe info: 486428 probes and 32 features;
6. Ensembl gene info: 63677 genes and 5 features.

We applied the logit transformation to the original methylation data which are proportions of methylation at each location.

2.2 Causal network inference for trios

Genes in the human body have been predisposed to MR which influences how they function. It allows researchers to group people based on their genetic code to identify potential factors and disease outcomes based on causal inference (**Figure 1.2**).

The causal inference models are aimed to be inferred by forming trios where the genetic variant is CNA (denoted as C) and the variables of interest are methylation and gene expression (denoted as M and E respectively). Among the models, M0.1 (C $\rightarrow$ E) and M0.2 (C $\rightarrow$ M) are null models, M1.1 (C $\rightarrow$ E $\rightarrow$ M) and M1.2 (C $\rightarrow$ M $\rightarrow$ E) are the mediation models, M2.1 (C $\rightarrow$ E $\leftarrow$ M) and M2.2 (C $\rightarrow$ M $\leftarrow$ E) are the v-structured models, M3 (M $\leftarrow$ C $\rightarrow$ E) is conditionally independent model, and M4 is fully connected model (**Figure 2.1**).

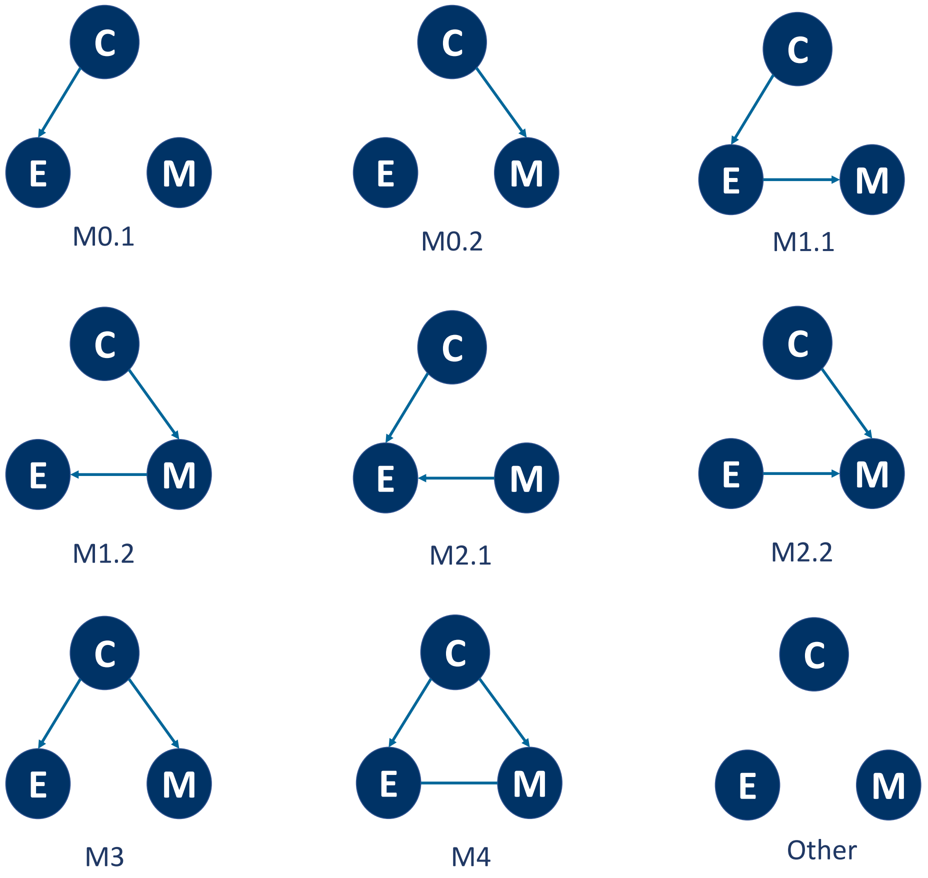

Figure 2.1: **Potential causal inference models**. The causal inference models are M0.1, M0.2, M1.1, M1.2, M2.1, M2.2, M3, and M4 model types. Any trios that are not classified in these model types are assigned as “Other”.

Each edge denotes that there exists a causal relationship between the nodes or the variables and the direction of the edge implies the direction of regulation. The assumption is that the genetic variant cannot be affected by the variables of interest. For example, in the case of the M1.1 model, we can say that there exists a causal relationship between C and M but it is mediated by E.

MRGN is a causal inference network method that uses a genetic variant as the instrumental variable to perform regression on the variables of interest under PMR. It uses regression on each of the variables of interest. Then based on the conditional and/or marginal tests on those variables (**Table 2.2**), it determines whether there is an edge or not and the direction of the edge.

The confounding variables are PC scores calculated using the Principal Component Analysis (PCA) that identifies the key variables in the data. We deduced the PC scores that are highly associated with the genes and methylation probes across the genome which are used as confounding variables. Each individual in the clinical data has been diagnosed with a certain cancer type: ER+ or ER-. We divided the individuals based on the category and extract their age and race information which are appended as confounding variables.

MRGN models the trio with regression of M (or E) on E (or M), C, and the confounding variables (denoted by the matrix U).

$$E = \beta_0 + \beta_{11}C + \beta_{12}M + \Gamma_1 U + \epsilon \tag{2.1}$$

$$M = \beta_0 + \beta_{21}C + \beta_{22}E + \Gamma_2 U + \epsilon \tag{2.2}$$

Table 2.1: **conditional and marginal tests**. This table shows the conditional and marginal tests used to infer the causal models. Conditional tests are used on all the models but additional marginal tests are used on M2 and M4 because they have the same conditional test.

Model	Sub model	Description	Conditional tests				Marginal tests	
			E $\perp\!\!\!\perp$ C $\mid$ (M, U)	E $\perp\!\!\!\perp$ M $\mid$ (C, U)	M $\perp\!\!\!\perp$ C $\mid$ (E, U)	M $\perp\!\!\!\perp$ E $\mid$ (C, U)	C $\perp\!\!\!\perp$ E	C $\perp\!\!\!\perp$ M
M0	M0.1	C $\rightarrow E; no$ relationship between E and M	$\neq 0$	$= 0$	$= 0$	$= 0$	-	-
	M0.2	C $\rightarrow M; no$ relationship between E and M	$= 0$	$= 0$	$\neq 0$	$= 0$	-	-
M1	M1.1	C $\rightarrow E \rightarrow M$	$\neq 0$	$\neq 0$	$= 0$	$\neq 0$	-	-
	M1.2	C → M → E	$= 0$	$\neq 0$	$\neq 0$	$\neq 0$	-	-
M2	M2.1	C → E ← M	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$	Yes	No
	M2.2	C → M ← E	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$	No	Yes
M3		E $\leftarrow C \rightarrow M$	$\neq 0$	$= 0$	$\neq 0$	$= 0$	-	-
M4		E $\leftarrow C \rightarrow M; E - M$	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$	No	No

The β coefficients in the equation are the potential pairwise conditional tests between C, E, and M. For the hypothesis test, $H_0 = \beta_{11} = 0$ and $H_A = \beta_{11} \neq 0$ shows the conditional test between E and C conditioned by M and the confounding variables. This can also be written as H_0 : E $\perp\!\!\!\perp C|$ (M, U) and H_A : E $\not\perp\!\!\!\perp C|$ (M, U). The conditional test using the β coefficients can be used to infer M0, M1, and M3 but marginal tests are required for M2 and M4 since they have the same conditional relationship (**Table 2.2**).

For instance, after applying the regression in equations **2.1** and **2.2**, we use the significant conditional tests from the Wald test and significant marginal tests from the Pearson correlations test to infer which model each trio belongs to. If a trio has a significant test for E $\perp\!\!\!\perp C|$ (M, U) but the other tests are not significant, we identify that trio as M0.1 (**Table 2.2**) (**Figure 2.1**). Once the trios have been allocated into their respective model, we summarize them to study the distribution and investigate further to review the features of trios in the mediation models. The causal network inference is the core of using MRGN on the datasets. This section has been implemented in Step 4 of the **Section 2.3**.

2.3 The R package MRTrios: a data analysis pipeline for trios

In this pipeline, we formed trios between a genetic variant (CNA) and the molecular phenotypes (methylation and gene expression) based on gene names and their Entrez IDs. Then we applied MR to the trios along with the confounders and inferred a genetic regulatory network. The data analysis pipeline has been implemented in the R package MRTrios available on GitHub [11]

Step 1: **Generating trios:**

(a) Splitting gene names: We split the gene names accessed from the methylation data for each probe, as one probe may be associated with multiple genes. We then identified unique gene names and found a total of 21,054 unique genes in the dataset;

(b) Skipping missing values: We checked for missing values across all individuals for each probe or gene. If a probe or gene had missing values for all individuals, we skipped that gene and its trios;

(c) Duplicates (genes with the same Entrez ID for gene expression and CNA): In the CNA and Gene Expression data, there are duplicated rows that have a different gene name but the same Entrez ID and values for all the individuals introducing repetition in the data. Within the datasets, we found that there are 70 genes (duplicates) in the CNA data and 8 genes (duplicates) in the gene expression data. (Note: these are the number of duplicates and not the appearances.);

(d) Matching gene names: We matched each of the unique gene names from (a) between all three datasets. Then we extracted the row numbers where the genes were found in the corresponding datasets and saved them to a file with 4 columns (gene name in the first column and the remaining 3 columns for row numbers in each of the 3 datasets);

(e) Entrez ID matching: We used the rows in the trio dataset which was matched based on gene names and extracted rows that have missing values in either the CNA column or gene expression column. Then we picked one out of the two columns mentioned earlier without a missing value and extracted an Entrez ID in their respective dataset. The Entrez ID was used to find a match in the dataset with a missing value in the trio dataset. For instance, if we had missing values in the CNA column, we picked the row number in the gene expression column and extracted the Entrez ID for the specific gene in the gene expression dataset. Then we selected that Entrez ID to find a match in the CNA dataset. If a match was found, we replaced the missing value in the CNA column in the trio dataset with the row number of the new match that was found in the CNA dataset;

(f) Fitting in missing Entrez IDs: Following the steps above, there were still Entrez IDs missing in some of the rows of the trio dataset in either the CNA column or gene expression column. Therefore, we used an external package "org.Hs.eg.db" from Bioconductor consisting of gene names and their Entrez IDs. We matched the gene names in the CNA and gene expression data to the gene names in the library "org.Hs.eg.db" to extract the Entrez ID and replaced it with the missing values in the respective dataset if a match was found. Since the "org.Hs.eg.db" package consists of multiple Entrez IDs for a particular gene in some cases, we saved the additional match to a different file and merge the datasets together in the end. We also made sure that the newly matched Entrez ID from the package and its associated row(s) is not one of the duplicated rows that we found earlier. If it was, we skipped over that particular row.

Step 2: **Filtering:**

(a) We selected the trios of the protein-coding genes and lncRNAs. We also selected the common individuals between the datasets (CNA, methylation, gene expression, and clinical) and extracted their ages and race. We observed 292534 trios that were further used for analysis.

Step 3: **Principal component:**

(a) We used PCA to calculate and identify principal components which are the key variables in the data. Then we derived the PCs that are highly associated with the genes and methylation probes across the genome. The PCs along with age and race were used as confounders (potential variables that could influence the causal inference).

Step 4: **Trio analysis with MRGN:**

(a) We applied MRGN to the trio data along with the confounders to infer the causal models (**Figure 2.1**) based on their conditional or/and marginal tests as discussed in **Section 2.2**.

Step 5: **Post hoc filtering:**

Since MRGN applies conditional tests for all the models but only applies marginal tests for M2 and M4, we do post hoc filtering to make sure the results are reliable.

(a) Since all the models are supposed to have at least one edge, we employed the marginal tests of C and M or C and E to make sure there exists a strong correlation. Any trio without a significant p-value is reassigned as “Other”;

(b) The M0 and M2 trios have only one edge either between C and M or C and E according to the conditional tests performed using MRGN. However, if the marginal tests are highly significant for both edges in M0 or M2 models, it creates a disparity with the results from MRGN’s conditional tests. Therefore, we reassigned the trios with significant marginal tests for both edges as “Other”;

(c) For M1.1 trios, since E is mediating the network, there should be a strong correlation between C and M. We reassigned any trio without a significant marginal p-value as “Other”. We repeated the process for M2.1 where M is the mediator and we verified the correlation between C and E.

Step 6: **Examining features of methylation probes in mediation trios:**

The features of methylation probes are generated to further investigate the probes assigned in the mediation models. However, we only evaluated the unique probes since a probe can be associated with multiple genes.

(a) The MRGN output data consists of row numbers in trios and model types for each trio. The trio data consists of previously mentioned row numbers and the row numbers of individual trios in methylation data. Therefore, we matched the trio row numbers from MRGN output data and trio data to get the row number in methylation data;

(b) Since the methylation data consists of probe IDs, we used the row numbers obtained from part (a) to extract the probe IDs of the individual trios;

(c) Then we matched the probe IDs in the methylation data with the probe IDs in the human methylation data to extract any probe information necessary;

(d) We identified the midpoint of the nearby CpG island for every probe. Then we calculated the distance between that midpoint and the location of the CpG in the methylation probe;

(e) We extracted the location of the CpG in the methylation probe and calculated its distance from the start site of a gene;

(f) We calculated the length of the gene using the start and end site of a gene;

(g) We summarized the location of the CpG in the methylation probe relative to the CpG island;

(h) We extracted the methylation levels for every probe in a trio and calculated the mean for every common individual among the 3 datasets (methylation, CNA, and gene expression);

(i) We extracted the Guanine-Cytosine (GC) content for each probe in a trio.

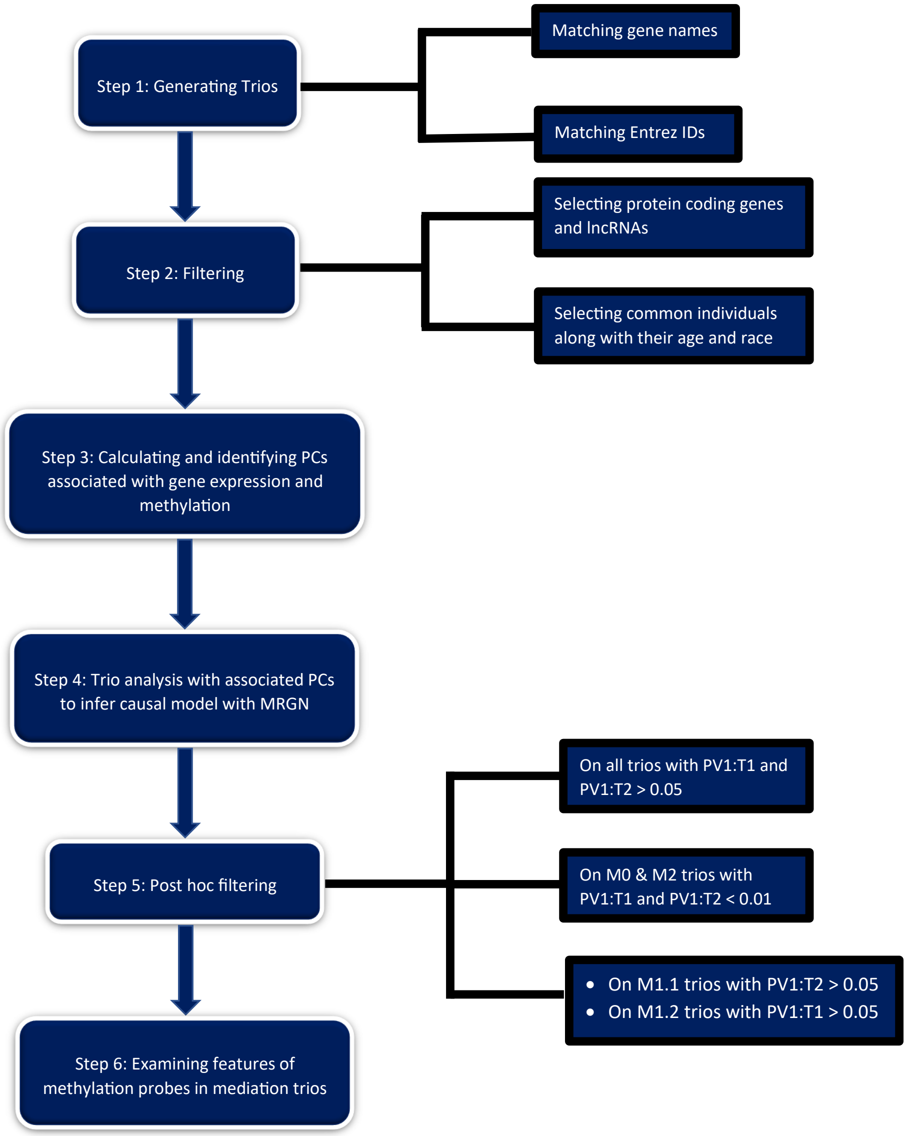
Step 1: Generating Trios
Matching gene names
Matching Entrez IDs
Step 2: Filtering
Selecting protein coding genes and lncRNAs
Selecting common individuals along with their age and race
Step 3: Calculating and identifying PCs associated with gene expression and methylation
Step 4: Trio analysis with associated PCs to infer causal model with MRGN
Step 5: Post hoc filtering
On all trios with PV1:T1 and PV1:T2 > 0.05
On M0 & M2 trios with PV1:T1 and PV1:T2 < 0.01
On M1.1 trios with PV1:T2 > 0.05
On M1.2 trios with PV1:T1 > 0.05
Step 6: Examining features of methylation probes in mediation trios

Figure 2.2: **Flow chart of the data analysis pipeline**. The figure gives an overview of each of the main steps in the data analysis pipeline which has been included in the MRTRios package.

CHAPTER 3

Results

We applied the causal inference network inference method to all trios involving protein-coding genes and lncRNAs and we summarized the results in this chapter.

In Section 3.1, we examined the distribution of the selected confounders and the inferred causal models in the two subtypes following the application of MRGN on the genomic trios. This helps us further understand the proportion of mediation models and identify genes or probes subsequently involved in the biological mechanisms underlying the relationship.

To gain a deeper understanding of the relationship between gene expression and methylation, we evaluated the mediation models, wherein either methylation or gene expression acts as a mediator. In this approach, the mediator regulates the outcome, while the instrumental variable has no direct causal relationship, although a high correlation exists between them. By applying this technique, we can assess the relationship between the mediator and the outcome. This enables us to examine how the outcome is regulated by the mediator and gain insights into the causal relationship between these variables.

In Section 3.2, we discuss the features of the probes categorized into the mediation models as compared to the baseline models as the relationship between our variables of interest depends on the features of the genes in each trio [1]. This provides an insight into the similarities and differences between these models as well as their effect on the overall structure of the models. The features are where a methylation probe is relative to the nearby CpG island, how far it is from the start of a gene, its average methylation level, and the length and the GC content of the gene.

3.1 Confounder Selection

Before application to trios, we identified confounders from the PCs generated using all the gene expression and methylation data. A large number of PCs were identified for those trios, especially for ER+ patients. The range of PC count for ER+ individuals is between 6 and 44 while the range of PC count for ER- individuals is between 1 and 19. The average PC count for ER+ individuals is 24 whereas the average PC count for ER- individuals is 9 (**Figure 3.1**). This variation could be due to a larger sample size for ER+ individuals. These PCs are identified from the PC score matrix generated for each dataset i.e. methylation and gene expression. They are used to find significantly associated PCs that are important in the analysis which helps reduce bias in the causal inference. The PC scores that were highly correlated with each of the probes or each of the genes in the genome-wide data were selected to be used in the analysis. This helps identify variables that are potentially associated with the outcome and exposure that could introduce bias in the results.

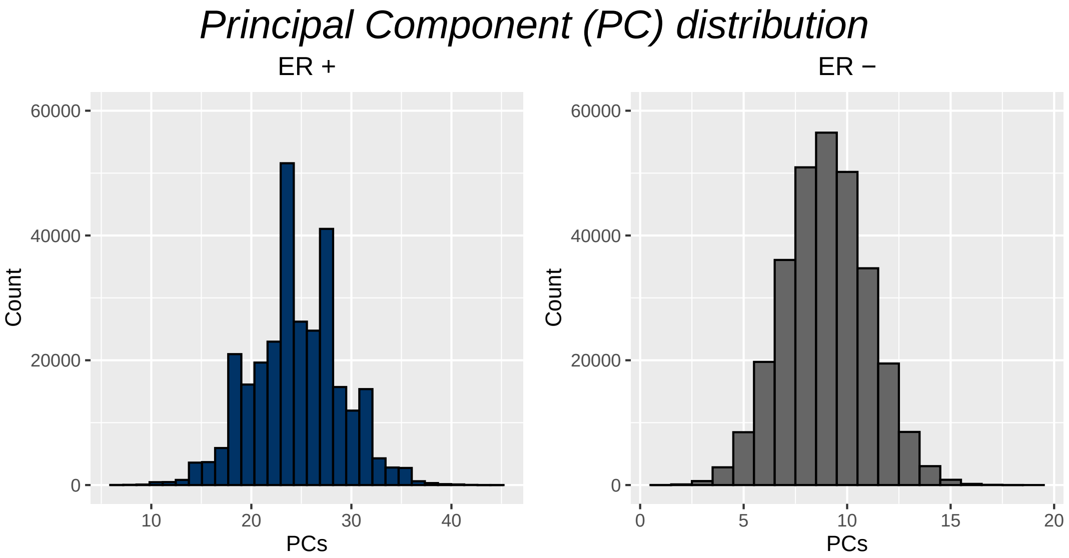

Figure 3.1: **Principle Component (PC) distribution for ER+ and ER- individuals**. The figure shows the distribution of PC counts in ER+ and - sub-types. For ER+, the mean is 26 and the median is 24. For ER-, the mean is 9 and the median is 9.

3.2 Inferred causal models

We incorporated the confounders and applied MRGN on the 292534 trios. These trios were categorized in each of the potential causal inference models. For the ER+ patients, we identified that 32% of all trios are Null models (M0), 6% are mediation (M1), 4% the v-structure (M2), 16% conditionally independent (M3), and 6% fully connected (M4) among the three nodes and the remaining are “Other”. For the ER- patients, we identified that 42% of all trios are Null models (M0), 3% are mediation (M1), 2% v-structure (M2), 9% conditionally independent (M3), and 1% fully connected (M4) among the three nodes and the remaining are “Other” (**Table 3.2**).

The mediation (M1), v-structure (M2), conditional independence (M3), and fully connected (M4) models are more meaningful because there exists a regulatory edge between the variables of interest. These results are obtained after a stringent filtering process to make them reliable. We observed that about 32% of the trios have an interesting causal model after rigorous filtering. Among them, M3 was the most common followed by M1, M4, and M2 for ER+ cancer type. We observed a similar trend for ER- cancer type except we found more M2 than M4. Among the two mediation models, M1.1 is more common which means on average the mediation models have gene expression as the mediator compared to M1.2 (**Figure 3.2**).

After conducting our analysis, we observed that there is a higher proportion of M1.1 models than M1.2 models. While it is widely recognized that DNA methylation in the promoter region can suppress gene expression, it remains unclear whether this is a causal factor or a consequence of gene expression. Although studies have revealed varying correlations between methylation and gene expression, with both positive and negative outcomes, the underlying mechanism for this relationship is still a topic of ongoing discussion [14, 17]. However, it is worth considering that a higher number of trios have been inferred as M1.1 instead of M1.2, despite the uncertainty surrounding the mechanism involved.

Table 3.1: **Inferred causal models by MRGN for ER+ and ER- individuals**. The table summarizes the counts of the trios inferred for each of the models and their percentages.

ER type	M0	M1.1	M1.2	M2	M3	M4	Other
Pos	92331	14557	3239	11228	45512	17620	107978
Pos %	0.316	0.05	0.011	0.038	0.156	0.06	0.369
Neg	123404	7758	1044	4579	27208	2937	125401
Neg %	0.422	0.026	0.004	0.016	0.093	0.01	0.429

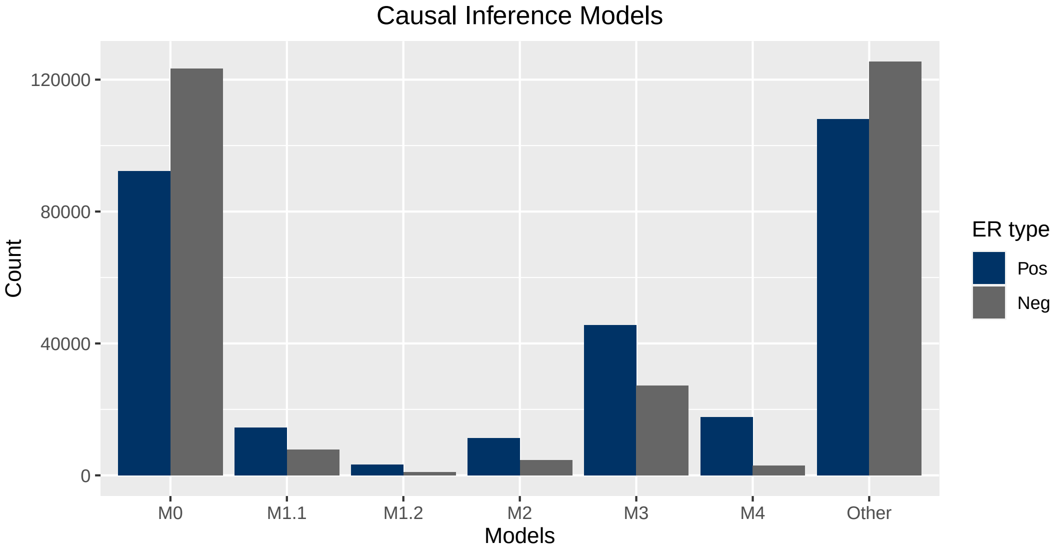

Figure 3.2: **Counts of the inferred causal models by MRGN for ER+ and ER- individuals**. The figure demonstrates the counts of trios classified in each of the model types for both ER+ and - individuals.

3.3 Features of methylation probes in mediation models

The mediation models are further evaluated because either methylation or gene expression acts as a mediator and the genetic variant acts as an instrumental variable. This results in the mediator regulating the outcome whereas the instrumental variable has no causal relationship although a high correlation exists between them. Since our goal is to understand the relationship between gene expression and methylation, the mediation models help us determine the effect given an independent genetic variant.

We investigated the M1.1 and M1.2 mediation models to determine the location of probes in genes where gene expression and methylation act as mediators, respectively. For ER+ patients in M1.1 model, 16.8% of the probes are in TSS1500, 12.8% are in TSS200, 16.7% are in $5'$ UTR, 7.6% are in 1stExon, 41.6% are in the gene body, and 4.5% are in $3'$ UTR. For ER- patients in M1.1 model, 18.3% of the probes are in TSS1500, 13.6% are in TSS200, 17.8% are in $5'$ UTR, 8.1% are in 1stExon, 38.2% are in the gene body, and 4% are in $3'$ UTR (**Table 3.2**).

For ER+ patients in M1.2 model, 12.7% of the probes are in TSS1500, 8.5% are in TSS200, 14.2% are in $5'$ UTR, 7.4% are in 1stExon, 52% are in the gene body, and 5.2% are in $3'$ UTR. For ER- patients in M1.2 model, 9.2% of the probes are in TSS1500, 7.9% are in TSS200, 9.5% are in $5'$ UTR, 7.9% are in 1stExon, 62.4% are in the gene body, and 3.1% are in $3'$ UTR (**Table 3.4**).

Additionally, we summarized the results above by merging locations with similar regions in the gene. Specifically, we merged TSS1500 and TSS200 as TSS, 1stExon and Body as Body, and $5'$ UTR and $3'$ UTR as $5'/3'$ UTR (**Table 3.3 and 3.5**). This provides a general overview of the distribution of probes in the three major regions of the gene. Following the merger, we observed that the majority of methylation probes are in the gene body in both mediation models (M1.1 and M1.2) between the two cancer types followed by TSS and $5'/3'$ UTR. In M1.1, TSS is higher, Body is lower, and $5'/3'$ UTR is similar as compared to M1.2 for ER+. We see a similar trend between M1.1 and M1.2 for ER- except $5'/3'$ UTR is much higher in M1.1 than M1.2 (**Figure 3.3**).

We performed a Chi-square test to compare the trio distribution in three locations (TSS, Body, and $5'/3'$ UTR) between two groups: M1.1 and M1.2 models in ER+ individuals versus Overall model (including both ER+ and ER- individuals) as presented in (**Table 3.3**). The results indicate that the counts for the two groups are significantly different ($p < 0.05$) in both M1.1 ER + and M1.2 ER+ versus Overall comparisons. Analyzing the percentages of M1.1 ER+ and Overall in each location, we found that TSS regions had similar proportions in both groups while the Body and $5'/3'$ UTR regions showed some disparity. Therefore, the observed differences in the Chi-square test results could be attributed primarily to the differences in the Body and $5'/3'$ UTR regions. Similarly, for M1.2 ER+ and Overall comparison, we observed that $5'/3'$ UTR regions had similar proportions in both groups while the Body and TSS regions showed some disparity. Hence, the differences observed in the Chi-square test could be primarily due to the differences in the Body and TSS regions.

CpG islands are regions in a genome that contain a high concentration of Cytosine-Guanine pairs. The CpG content at a promoter could affect how that gene is regulated as high CpG content in promoter regions can enhance transcriptional activity [21, 15, 5]. They have been known as a crucial part of a gene because methylated CpG sites found in cancer tumors are generally responsible for silencing the expression of tumor suppressor genes [18]. The CpG location of methylation probes and their distance from the nearby CpG island or start site of the gene in our mediation models helps us study the variations between the former models as compared to the baseline across the cancer subtypes. In addition to this, we also inspected the location of the CpG relative in a methylation probe relative to the nearby CpG island, the length of the corresponding gene, the distribution of methylation values in each probe, and its GC content.

For ER+ patients, on average the probe tends to be 2.524 Kb, 2.694 Kb, 2.642 Kb, and 2.743 Kb away from the nearby CpG island for M0.1, M0.2, M1.1, and M1.2 models respectively. For ER- patients, on average the probe tends to be 2.531 Kb, 2.675 Kb, 2.639 Kb, and 2.772 Kb away from the nearby CpG island for M0.1, M0.2, M1.1, and M1.2 models respectively (**Figure 3.4**). The distribution of the distance appears to be consistent for all the models with a mode between 2 and 3 Kb, however, the M1.2 model for the ER- patients has a comparably higher frequency.

The distribution of the distance from a CpG in methylation probe to the start site of the gene tends to be bimodal for both ER+ and ER- patients. The modes are between 2 and 3 and 4 and 5 in increasing order of frequencies. Although the distribution appears similar for all the models across the two types, the frequencies for M0.2 and M1.2 models for ER- patients seem to be more peaked. (**Figure 3.5**).

For ER+ patients, the average length of the gene is 4.492 Kb, 4.433 Kb, 4.468 Kb, and 4.523 Kb for M0.1, M0.2, M1.1, and M1.2 models respectively. For ER- patients, the average length of the gene is 4.493 Kb, 4.482 Kb, 4.539 Kb, and 4.633 Kb for M0.1, M0.2, M1.1, and M1.2 models respectively (**Figure 3.6**). Although we observe a similar trend between the respective baseline models and the mediation models across the ER+ and ER- patients, the shape of the distribution for the M0.1 and M1.1 models is more peaked as compared to the M0.2 and M1.2 models.

The area 2 kb upstream and downstream of CpG islands are referred to as CpG shores, and the area 2 kb upstream and downstream of CpG shores as CpG shelves. For the M0.1 model of the ER+ patients, 46% of the probes were found in CpG island followed by 14% in N Shore, 12% S Shore, 3% in N Shelf, 3% S Shelf, and the remaining 22% no relation. Similarly, for M0.2, the order is Island (28%), N Shore (11%), S Shore (9%), N Shelf (4%), S Shelf (3%), and remaining 45% no relation. Additionally, for M1.1, the order is Island (36%), N Shore (16%), S Shore (13%), N Shelf (5%), S Shelf (4%), and remaining 26% no relation and for M1.2, Island (25%), N Shore (13%), S Shore (11%), N Shelf (5%), S Shelf (4%), and remaining 42% no relation.

For the M0 model of the ER- patients, 45% of the probes were found in CpG island followed by 13% in N Shore, 11% S Shore, 4% in N Shelf, 3% S Shelf, and the remaining 24% no relation. Similarly, for M0.2, the order is Island (32%), N Shore (12%), S Shore (9%), N Shelf (4%), S Shelf (4%), and remaining 39% no relation. Additionally, for M1.1, the order is Island (38%), N Shore (18%), S Shore (14%), N Shelf (4%), S Shelf (3%), and remaining 23% no relation and for M1.2, Island (26%), N Shore (17%), S Shore (11%), N Shelf (5%), S Shelf (3%), and remaining 38% no relation (**Figure 3.7**).

Among the two subtypes, the proportion of genes in the various locations seems to be more or less similar. However, there is a higher proportion of genes in the CpG island of the M0.1 and M1.1 models compared to the M0.2 and M1.2 models followed by N Shore, S Shore, N Shelf, and S Shelf. This again shows the similarities between the respective baseline and mediation models.

The distribution of methylation means is obtained from the log transformation of the original distribution by assuming a normal distribution on the transformed data for any single gene (**Figure 3.8**). We can observe that the distributions for both the baseline models have a large difference although they have a bimodal distribution. For ER+ patients, the bimodal distribution has modes between -3 and -2.5 which has a higher frequency and between 2 and 3 which has a lower frequency in M0.1 and M1.1 models. Although M0.1 is similar to M1.1 there is a noticeable difference in the mode. For the first peak in the distribution, M0.1 has a higher frequency in comparison to M1.1. The pattern in M0.2 and M1.2 is entirely different with 3 modes between -3 and -2.5, -0.5 and 0.5, and 2 and 2.5 with increasing order of frequency. For ER- patients, we observe a similar trend in the modes, however, the frequency peaks between M0.1 and M1.1 seem to be more alike in this case. While the second peak is the lowest in M0.2, it tends to be the highest in M1.2 followed by the first peak and the third peak in both models.

To elaborate, GC content is the percentage of nucleotide bases that are either Guanine (G) or Cytosine (C) out of the total bases in a DNA or RNA molecule. Previous research has found a strong positive correlation between GC content and gene expression levels, underscoring the importance of this factor in our analysis [12]. For ER+ patients, the mean GC content for the probe is approximately 46.793%, 47.284%, $47.339, and\%$ 47.393% for M0.1, M0.2, M1.1, and M1.2 models respectively. For ER- patients, the mean GC content for the probe is approximately 46.625%, 47.6%, 46.966%, and 47.466% for M0.1, M0.2, M1.1, and M1.2 models respectively (**Figure 3.9**) The distribution of the GC content appears to be alike for all the models with a mode around 40%, however, the M1.2 model for the ER- patients has a comparably higher frequency.

Among all of the features we studied, the M0.1 model serves as a baseline for the M1.1 models whereas the M0.2 model serves as a baseline for M1.2 because of their respective edge between C and M or C and E.

Table 3.2: **M1.1 model inferred by MRGN for ER+ and ER- individuals based on probe location**. The table summarizes the counts for the locations of the M1.1 trios and their percentages.

ER type	TSS1500	TSS200	5′ UTR	1stExon	Body	3′ UTR
Pos	5105	3897	5067	2315	12654	1348
Pos %	0.168	0.128	0.167	0.076	0.416	0.045
Neg	3022	2254	2943	1336	6300	659
Neg %	0.183	0.136	0.178	0.081	0.382	0.04

Table 3.3: **Counts and percentage of the M1.1 model by MRGN for ER+ and ER- individuals merged into 3 locations**. The table summarizes the counts for the locations of the M1.1 trios and their percentages.

ER type	TSS	Body	$5'/3'$ UTR
Pos	9002	14969	6415
Pos %	0.296	0.493	0.211
Neg	5276	7636	3602
Neg %	0.32	0.462	0.218
Overall	382330	724208	243254
Overall %	0.283	0.537	0.18

Table 3.4: **Counts and percentage of the M1.2 model by MRGN for ER+ and ER- individuals**. The table summarizes the counts for the locations of the M1.2 trios and their percentages.

ER type	TSS1500	TSS200	$5'$ UTR	1stExon	Body	$3'$ UTR
Pos	887	597	996	516	3636	365
Pos %	0.127	0.085	0.142	0.074	0.52	0.052
Neg	305	263	313	263	2072	103
Neg %	0.092	0.079	0.095	0.079	0.624	0.031

Table 3.5: **Counts and percentage of the M1.2 model by MRGN for ER+ and ER- individuals merged into 3 locations**. The table summarizes the counts for the locations of the M1.2 trios and their percentages.

ER type	TSS	Body	$5'/3'$ UTR
Pos	1484	4152	1361
Pos %	0.212	0.593	0.195
Neg	568	2335	416
Neg %	0.171	0.704	0.125
Overall	382330	724208	243254
Overall %	0.283	0.537	0.18

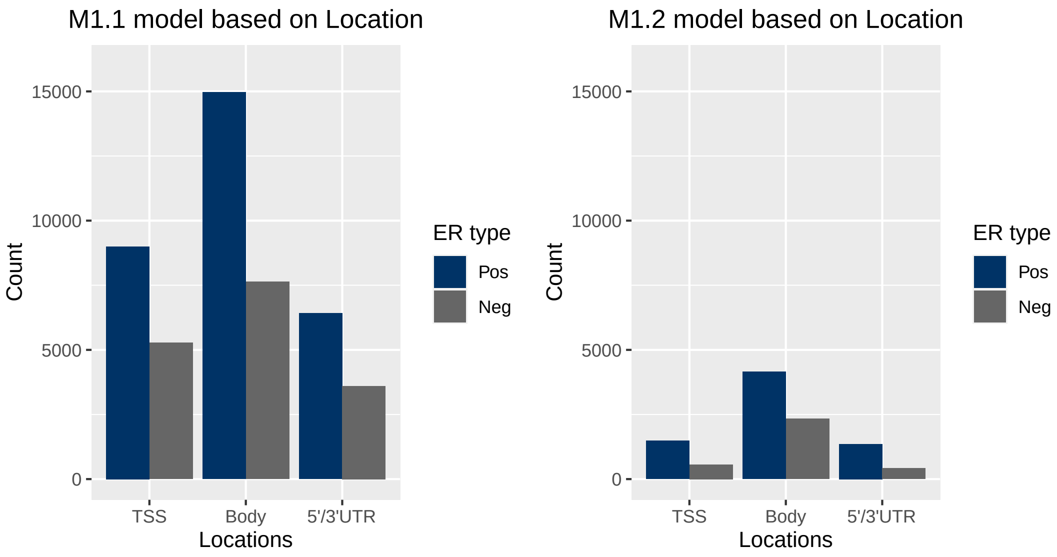

Figure 3.3: **Counts of the mediation models (M1.1 and M1.2) based on their locations for ER+ and ER- individuals**. The figure shows the counts for the locations in the gene where the mediation trios are located for both ER+ and - individuals. Transcription Start Site (TSS) is in the gene's promoter region, Body is the body of the gene, and $5'/3'$ UTR is in the end region of the gene.

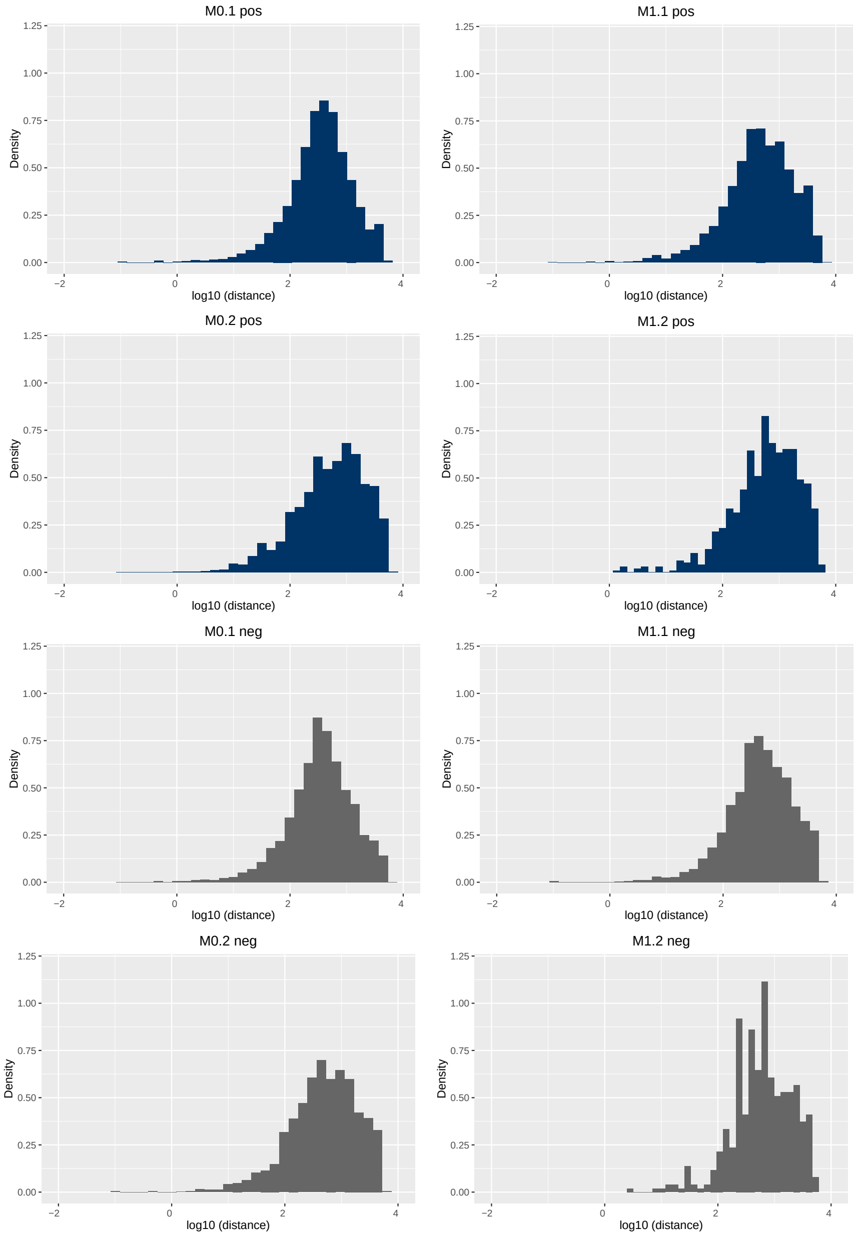

Figure 3.4: **Distribution of the log10 (distance) between the nearby CpG island and the CpG in the methylation probe.** The figure uses the midpoint of the nearby CpG island to calculate the difference and shows how far the CpG in a probe is from a repeated CG pair sequence nearby in the baseline and mediation models across the cancer subtypes.

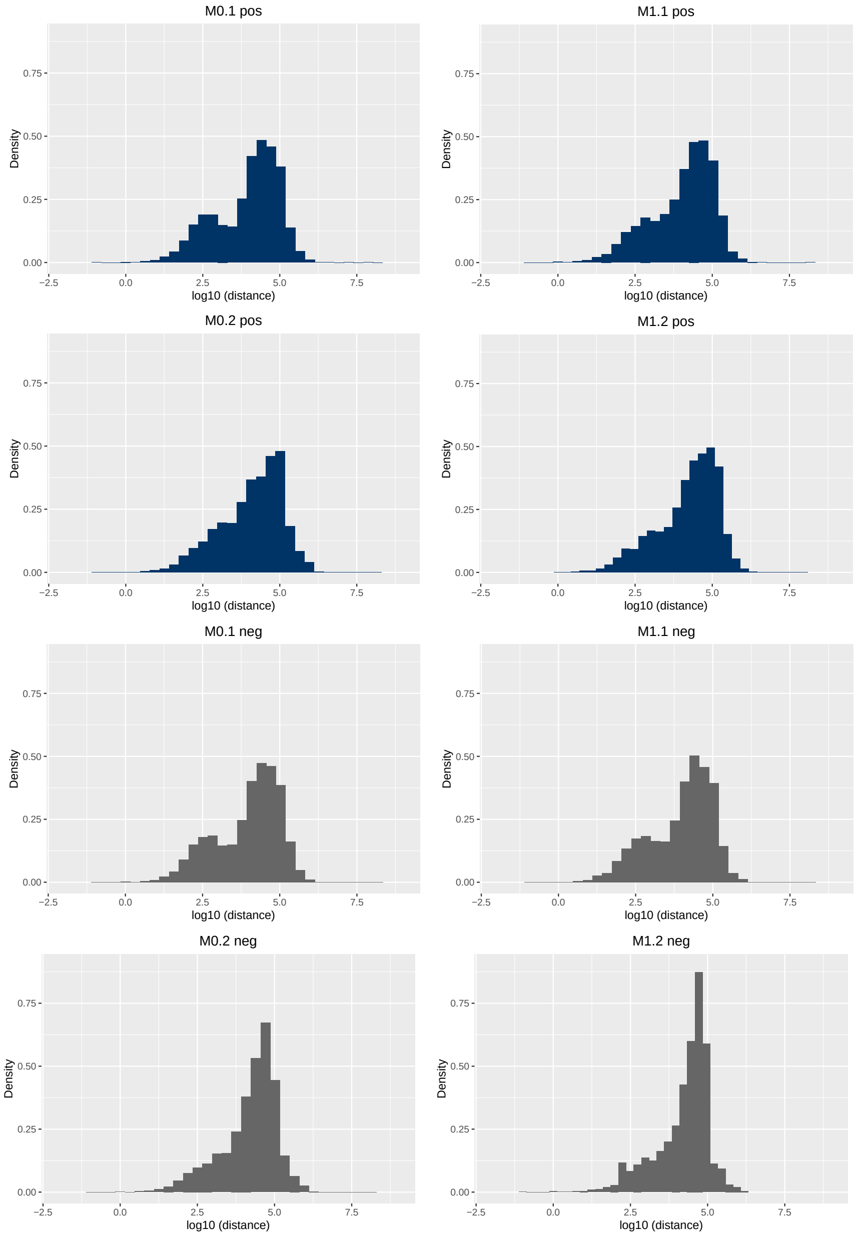

Figure 3.5: **Distribution of the log 10 (distance) between the CpG in the methylation probe and the start position of a gene**. The figure shows how far the CpG site in a probe is from the start site of the corresponding gene in the baseline and mediation models across the cancer subtypes.

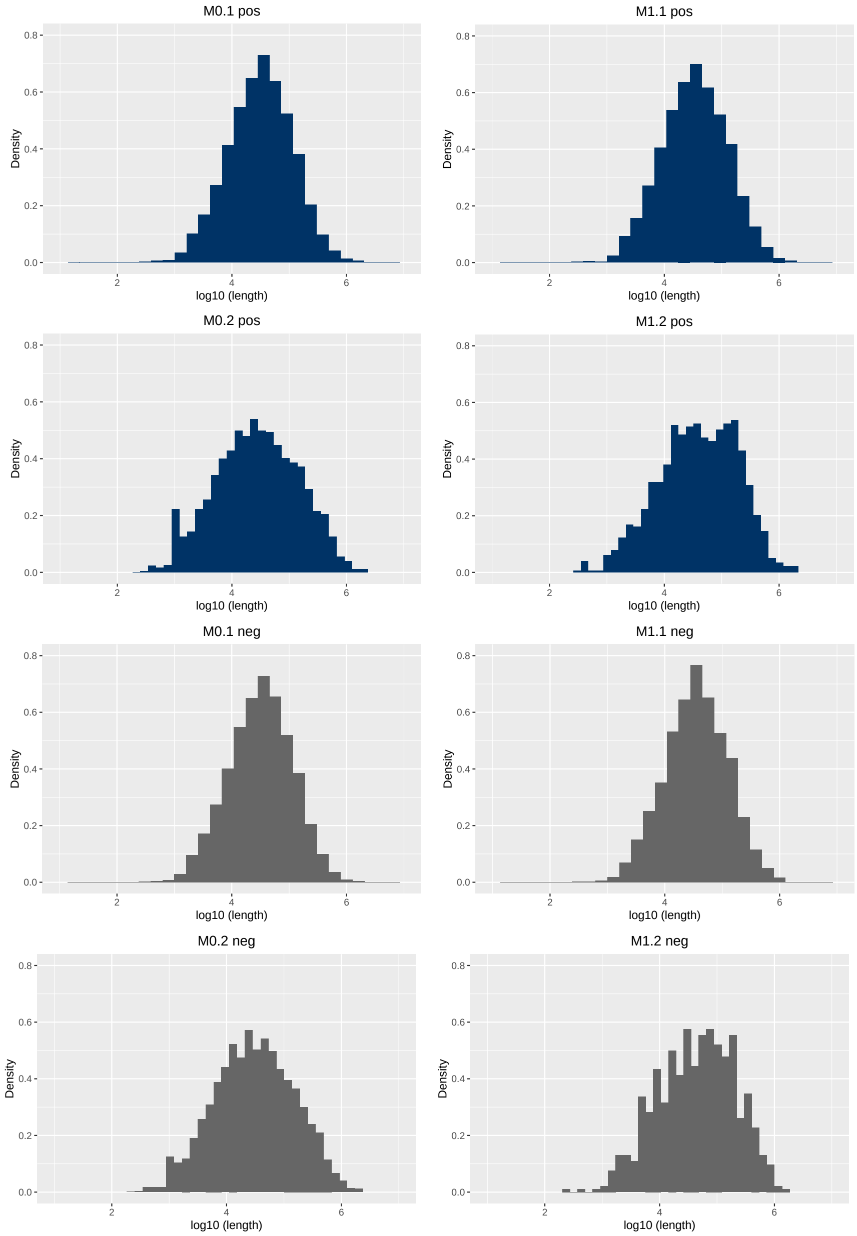

Figure 3.6: **Distribution of the log 10 (length) between the start position and the end position of a gene**. This figure shows the average length of the gene in the baseline and mediation models across the cancer subtypes.

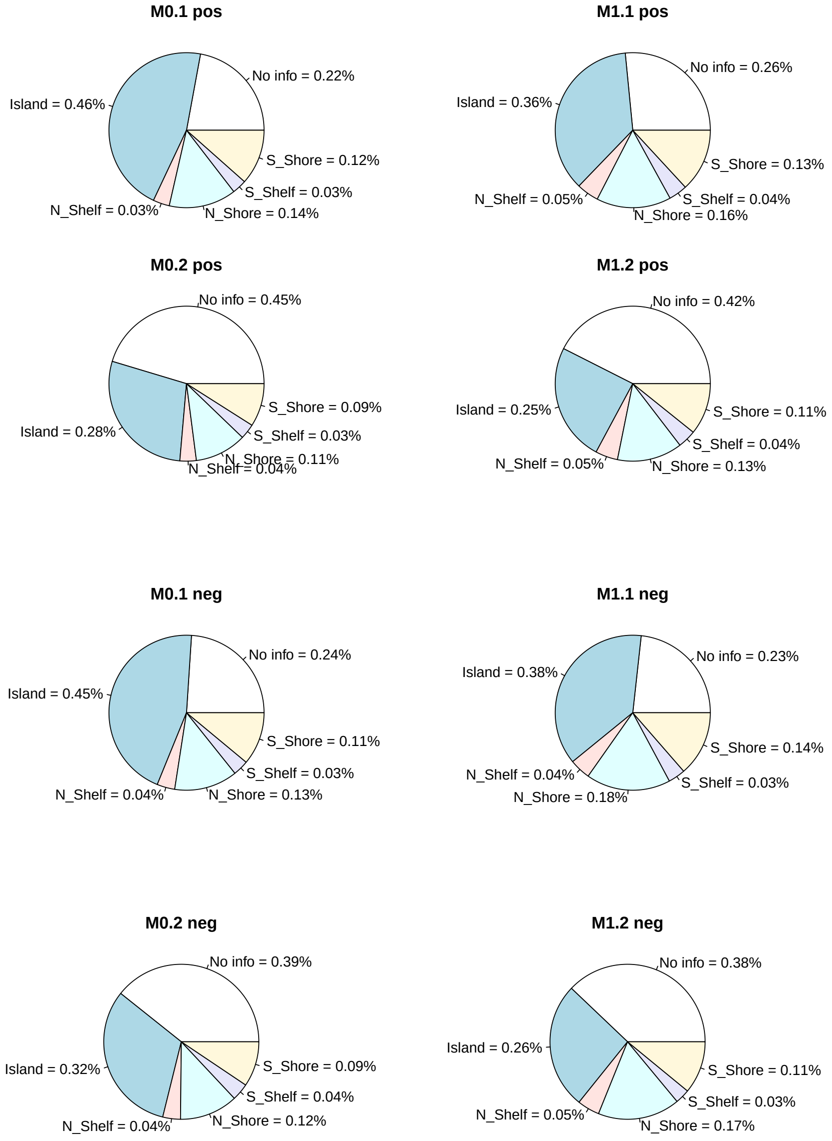

Figure 3.7: **Distribution of the relation to CpG island**. The location relative to CpG island is divided into five sections: Island, N shore, S shore, N shelf, S shelf, and No Info which shows how close or far they are from the island in the baseline and mediation models across the cancer subtypes. The “No Info” indicates that no nearby CpG island was found.

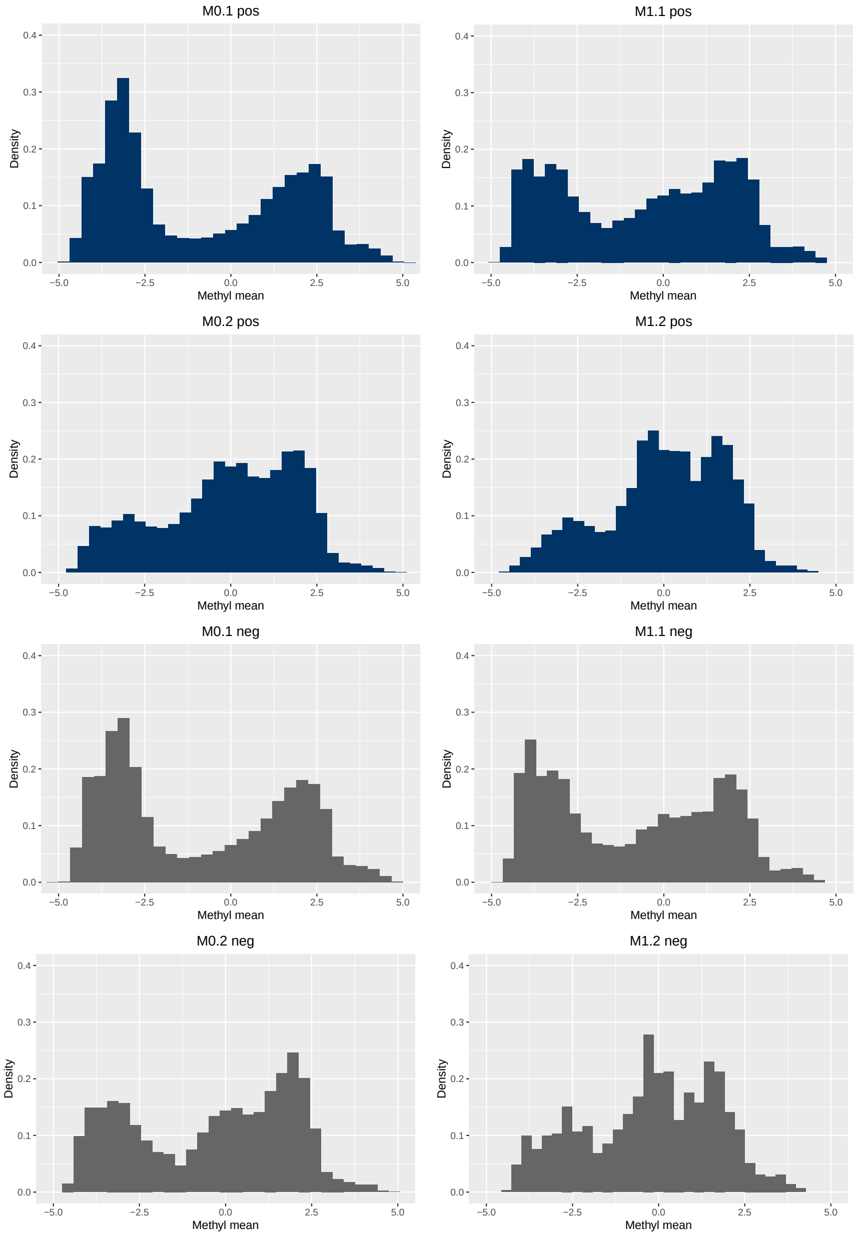

Figure 3.8: **Distribution of methylation values**. The figure shows the distribution of methylation values in the baseline and mediation models across the cancer subtypes. The genes selected are protein-coding or lncRNAs and we selected common individuals between the 3 datasets.

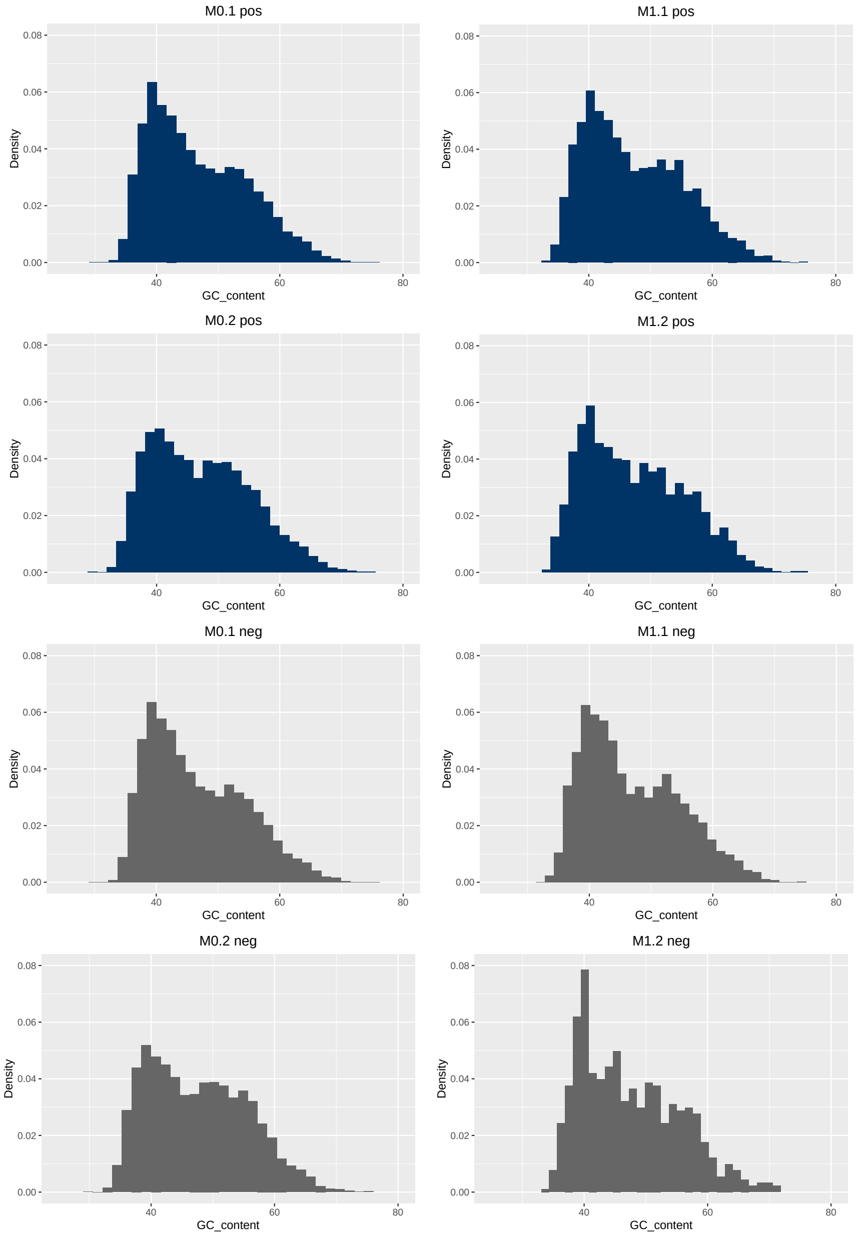

Figure 3.9: **Distribution of the GC content**. The figure shows the distribution of the GC content which is the percentage of C and G among the four nucleotide base pairs (A, T, C, and G) in the baseline and mediation models across the cancer subtypes.

CHAPTER 4

Conclusions and Discussions

In this thesis, we conducted an extensive analysis to explore the relationship between methylation and gene expression and establish a causal inference network. To infer a causal framework, we utilized both conditional and marginal relationships between the genetic variant and the variable of interest. Our post-filtering process further tests for disparities between these tests, which significantly improves the inference method. One of the primary challenges in causal inference is bias, as the analysis is susceptible to external confounders. To account for this, our analysis takes into consideration the effect of confounding variables in genome-wide associations. By identifying and implementing confounders, we streamlined the analysis and eliminate potential bias introduced by these variables.

Our analysis revealed that approximately 32% of the trios exhibited an interesting causal model. ER+ patients exhibited the most common model as M3, followed by M1, M4, and M2. Meanwhile, ER- patients demonstrated the most common model as M3, followed by M1, M2, and M4. However, some models remained as M0 or "Other" indicating a higher percentage of single-edged or no-edged models. We acknowledge that one of the reasons for a higher percentage of single-edged or no-edged models was the demanding filtering procedure. However, we also recognize that other unknown factors may have contributed to this outcome. It is possible that these models accurately represent the true biological mechanism, or it may be due to our model's inability to capture the actual case. Therefore, we plan to continue evaluating our methodology to make it better and more reliable for the analysis.

We aimed to gain a deeper understanding of the mediation models (M1.1 and M1.2) by performing a detailed analysis of their features. We focused on several key features, including the proximity of the methylation probe to the nearest CpG island, its distance from the start of the gene, its average methylation level, and the gene's length and GC content. Our findings revealed similarities between the baseline models and the mediation models. Specifically, we observed that the M0.1 model closely resembled the M1.1 model, while the M0.2 model was similar to the M1.2 model. This could suggest that the mediation models build upon the baseline models by incorporating additional information, rather than introducing entirely new features. We noticed comparable trends in the distribution of these features, indicating that they are similar across both subtypes. However, we did observe noticeable differences in the frequency peaks, suggesting that there may be some subtype-specific differences in the relationship between methylation and gene expression.

An essential factor during the analysis process is the trio formation step. It comprises of matches between datasets (e.g., methylation, gene expression, CNA) based on gene name or their unique identifiers. This can result in data loss for genes when matches aren't found. Additionally, implementing genomic data can be challenging, as much of it tends to be incomplete or ever-changing. Therefore, it is necessary to consider these factors to ensure that the results aren't affected by data availability or lack thereof.

To make the analysis more effective, we focused on studying the features of protein-coding genes and lncRNA, as variations in expression from these genes lead to the development or continuation of complex diseases. Although this method of analysis is only applicable to molecular phenotypes like methylation and gene expression in the presence of a genetic variant due to the PMR assumption, it can be applied across various trios in the omics discipline. Hence, we aim to expand and examine the causal relationship between various biological processes in other cancer types. This can help locate genetic factors responsible for the growth of the disease and establish targeted therapy or treatment with improved accuracy.

While we have meticulously processed and filtered our analysis, it is important to acknowledge that the results may be influenced by the current sample size. It may not provide a complete representation of the relationships between the variables of interest in the two subtypes. Additionally, the results could also be affected by the unequal sample sizes between the ER+ and ER- groups. Notably, the ER+ group contains a significantly higher number of samples than the ER- group. To address this issue, we will perform additional analyses by downsampling the ER+ group. This will enable us to assess whether our findings are robust across different sample sizes or whether they are driven by unequal group sizes.

Overall, our method for inferring causal networks provides valuable insights into the underlying mechanisms of diseases, enabling the identification of potential therapeutic targets and the development of targeted therapies. While challenges such as bias and data availability must be addressed, our analysis can help pave the way for understanding the biological mechanisms of cancers in the future.

Bibliography

[1] Dafni Anastasiadi, Anna Esteve-Codina, and Francesc Piferrer. Consistent inverse correlation between DNA methylation of the first intron and gene expression across tissues and species. *Epigenetics & chromatin*, 11(1):1–17, 2018.

[2] Md Bahadur Badsha, Evan A Martin, and Audrey Qiuyan Fu. MRPC: An R package for inference of causal graphs. *Frontiers in Genetics*, 12:651812, 2021.

[3] Rajbir Nath Batra, Aviezer Lifshitz, Ana Tufegdzic Vidakovic, Suet-Feung Chin, Ankita Sati-Batra, Stephen-John Sammut, Elena Provenzano, H Raza Ali, Ali Dariush, Alejandra Bruna, et al. DNA methylation landscapes of 1538 breast cancers reveal a replication-linked clock, epigenomic instability and cis-regulation. *Nature communications*, 12(1):5406, 2021.

[4] Elizaveta V Benevolenskaya, Abul BMMK Islam, Habibul Ahsan, Muhammad G Kibriya, Farzana Jasmine, Ben Wolff, Umaima Al-Alem, Elizabeth Wiley, Andre Kajdacsy-Balla, Virgilia Macias, et al. DNA methylation and hormone receptor status in breast cancer. *Clinical epigenetics*, 8:1–10, 2016.

[5] Leandros Boukas, Hans T Bjornsson, and Kasper D Hansen. Promoter CpG density predicts downstream gene loss-of-function intolerance. *The American Journal of Human Genetics*, 107(3):487–498, 2020.

[6] Ethan Cerami, Jianjiong Gao, Ugur Dogrusoz, Benjamin E Gross, Selcuk Onur Sumer, Bülent Arman Aksoy, Anders Jacobsen, Caitlin J Byrne, Michael L Heuer, Erik Larsson, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer discovery*, 2(5):401–404, 2012.

[7] James Y Dai, Xiaoyu Wang, Bo Wang, Wei Sun, Kristina M Jordahl, Suzanne Kolb, Yaw A Nyame, Jonathan L Wright, Elaine A Ostrander, Ziding Feng, et al. DNA methylation and cis-regulation of gene expression by prostate cancer risk SNPs. *PLoS Genetics*, 16(3):e1008667, 2020.

[8] Maxim VC Greenberg and Deborah Bourc'his. The diverse roles of DNA methylation in mammalian development and disease. *Nature reviews Molecular cell biology*, 20(10):590–607, 2019.

[9] Maria Gutierrez-Arcelus, Tuuli Lappalainen, Stephen B Montgomery, Alfonso Buil, Halit Ongen, Alisa Yurovsky, Julien Bryois, Thomas Giger, Luciana Romano, Alexandra Planchon, et al. Passive and active DNA methylation and the interplay with genetic variation in gene regulation. *elife*, 2:e00523, 2013.

[10] Gibran Hemani, Kate Tilling, and George Davey Smith. Orienting the causal relationship between imprecisely measured traits using GWAS summary data. *PLoS genetics*, 13(11):e1007081, 2017.

[11] Bandita Karki and Audrey Qiuyan Fu. The R package MRTrios: a data analysis pipeline for trios. `https://github.com/audreyqyfu/MRTrios`, 2023.

[12] Grzegorz Kudla, Leszek Lipinski, Fanny Caffin, Aleksandra Helwak, and Maciej Zylicz. High guanine and cytosine content increases mrna levels in mammalian cells. *PLoS biology*, 4(6):e180, 2006.

[13] Jarred Kvamme and Audrey Qiuyan Fu. An accurate and efficient causal gene network inference method that handles many confounding variables. *In preparation.*

[14] Yulia A Medvedeva, Abdullah M Khamis, Ivan V Kulakovskiy, Wail Ba-Alawi, Md Shariful I Bhuyan, Hideya Kawaji, Timo Lassmann, Matthias Harbers, Alistair RR Forrest, and Vladimir B Bajic. Effects of cytosine methylation on transcription factor binding sites. *BMC genomics*, 15(1):1–12, 2014.

[15] Michael D Morgan and John C Marioni. CpG island composition differences are a source of gene expression noise indicative of promoter responsiveness. *Genome biology*, 19(1):1–13, 2018.

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

[17] Ieva Rauluseviciute, Finn Drabløs, and Morten Beck Rye. DNA hypermethylation associated with upregulated gene expression in prostate cancer demonstrates the diversity of epigenetic regulation. *BMC medical genomics*, 13(1):1–15, 2020.

[18] Wei Sun, Paul Bunn, Chong Jin, Paul Little, Vasyl Zhabotynsky, Charles M Perou, David Neil Hayes, Mengjie Chen, and Dan-Yu Lin. The association between copy number aberration, DNA methylation and gene expression in tumor samples. *Nucleic acids research*, 46(6):3009–3018, 2018.

[19] Gary Taubes. Researchers find a way to mimic clinical trials using genetics. *MIT Technology Review*, Apr 2020.

[20] Matthew Ung, Xiaotu Ma, Kevin C Johnson, Brock C Christensen, and Chao Cheng. Effect of estrogen receptor α binding on functional DNA methylation in breast cancer. *Epigenetics*, 9(4):523–532, 2014.

[21] Tanya Vavouri and Ben Lehner. Human genes with CpG island promoters have a distinct transcription-associated chromatin organization. *Genome biology*, 13(11):1–12, 2012.

[22] PW Wilson, Robert D Abbott, and William P Castelli. High density lipoprotein cholesterol and mortality. The Framingham Heart Study. *Arteriosclerosis: An Official Journal of the American Heart Association, Inc.*, 8(6):737–741, 1988.