Delft University of Technology

Thesis

A multi-omic approach to discovering methylation-based markers of colon cancer in cfDNA

Authors:
Ewoud Ruighaver

Supervisor:
Marcel Reinders
Stavros Makrodimitris

in the

Pattern Recognition and Bioinformatics Section
Intelligent Systems

November 19, 2021

DELFT UNIVERSITY OF TECHNOLOGY

Abstract

Electrical Engineering, Mathematics and Computer Science
Intelligent Systems

A multi-omic approach to discovering methylation-based markers of colon cancer in cfDNA

by
Ewoud Ruighaver

In recent years the advent of multi-omic techniques have shown great promise in the field of oncology. In light of these advancements, this thesis focuses on the use of multiple data types to find methylation markers around transcription start site regions for colorectal cancer in the cell-free DNA (cfDNA) domain. It combines several methods of finding these markers, based on publicly available data obtained from solid tissue biopsies. These methods are both based on a single data type, as well as integrating multiple different data types. The resulting selections of methylation markers are then tested for significance on two independent datasets of cfDNA samples. The selections produced are tested on these datasets for their significance in distinguishing colorectal cancer samples from healthy samples. On one of these datasets, the selections are also tested for being differentially methylated between a group of patients with recurring tumors versus non-recurring tumors. The results on these two different datasets vary, showing that the methods of selecting potential methylation markers are capable of doing so on one platform, but that these results cannot be validated on another.

Contents

Abstract **i**

1 Introduction **5**

1.1 Objectives . . . 6

2 Methods **8**

2.1 Data types and sources . . . 8

2.1.1 TCGA Pan-Cancer Atlas . . . 8

Methylation data . . . 8

Gene expression data . . . 8

2.1.2 Whole blood samples . . . 8

2.1.3 Liquid biopsy data . . . 9

2.2 Data preprocessing . . . 9

2.2.1 Pan-Cancer Atlas datasets . . . 9

2.2.2 Grouping individual CpG sites to transcription start sites . . . 9

Illumina platforms . . . 10

MeD-seq . . . 10

2.3 Multi-omic integration . . . 10

2.3.1 JIVE . . . 10

2.3.2 aJIVE . . . 11

2.4 Selection of differentially methylated TSSs . . . 12

2.4.1 Thresholding approach . . . 12

2.4.2 JIVE loadings . . . 13

aJIVE loadings . . . 14

2.4.3 Random forest . . . 14

2.4.4 Combining selections . . . 15

2.5 Differential analysis . . . 15

2.5.1 Liquid biopsy data . . . 15

2.5.2 MeD-seq data . . . 15

2.6 Evaluation . . . 16

3 Results **17**

3.1 Significance of found transcription start sites . . . 17

3.2 Validation on BeadChip liquid biopsy data . . . 18

3.3 Validation on MeD-seq data . . . 19

3.4 Selection of differentially methylated TSS regions in tissue . . . 21

3.5 Prognosis prediction . . . 22

4 Discussion **24**

4.1 Findings . . . 24

4.2 The benefit of integrating multiple data types . . . 25

4.3 Results on microarray data do not translate to MeD-seq . . . 25

4.4 Literature concerning *CTNND2* and *SLIT2* . . . 26

4.4.1 CTNND2 26
4.4.2 SLIT2 26
4.5 Concluding remarks 26

Appendix A JIVE example visualizations 28

Appendix B JIVE rank selection 29

Appendix C aJIVE example 30

Appendix D Volcano plots of PANCAN colon data with individual selections highlighted 32

Appendix E Hyperparameters for the Random Forest Classifier 34

Bibliography 35

List of Figures

1.1 Origins and range of alterations in cell-free DNA, from [5] 6

2.1 A summary of the method used to find a selection of TSS regions. . . 12
2.2 Overview of the different steps in the thresholding selection method. Example values are used for the thresholds denoted in red. 13
2.3 Factorized version of the JIVE model. From [18]. 14
2.4 Process by which the individual selections are combined. The weights shown here are to serve as an example and can be any real number. Letters are used to represent TSS regions. The final combined selection is denoted in bold. 15

3.1 Volcano plot for the combined selection of 100 transcription start sites. Selected TSS regions from the combined selection are marked in red. . 18
3.2 Volcano plot visualizing TSS regions in the GSE122126 data for the CRC patients versus healthy controls. Selected TSS regions from the combined selection are marked in red. 19
3.3 Volcano plot visualizing TSS regions in the MeD-seq data for the T0 CRC group versus healthy controls. Selected TSS regions from the combined selection are marked in red. 20
3.4 Volcano plot visualizing TSS regions in the MeD-seq data for the T0 CRC group versus healthy controls. Significant TSS regions from the solid tissue differential methylation analysis are marked in red. 22

A.1 JIVE example for two data blocks. From [18]. 28
A.2 JIVE estimates for the blocks from figure A.1. From [18]. 28

B.1 Rank selection for r. From [18] 29
B.2 Rank selection for r_j. From [18] 29

C.1 Dataset constructed from a joint and individual component. From [20] 30
C.2 aJIVE decomposition of data constructed in figure C.1. From [20] . . 31
C.3 JIVE decomposition of data constructed in figure C.1. From [20] . . . 31

D.1 Volcano plot for the thresholding selection of 100 transcription start sites. 32
D.2 Volcano plot for the JIVE joint selection of 100 transcription start sites. 32
D.3 Volcano plot for the JIVE individual selection of 100 transcription start sites. 33
D.4 Volcano plot for the aJIVE joint of 100 transcription start sites. 33
D.5 Volcano plot for the random forest selection of 100 transcription start sites. 33

List of Tables

2.1 Overview of the datasets used. 9

3.1 The number of significant ($p < 0.05$) selections for each type of selection in the PANCAN tissue differential analyses. 18
3.2 The number of significant ($p < 0.05$) selections for the differential methylation analysis on the Moss et al. cfDNA data. 19
3.3 Number of significant ($p < 0.05$) selections for the MeD-seq differential analysis. 20
3.4 Number of significant ($p < 0.05$) selections for the MeD-seq T0 vs T2 differential analysis. 22

E.1 Hyperparamaters for *scikit-learn*'s *RandomForestClassifier* 34

Chapter 1

Introduction

In the field of oncology, the paradigm of personalized medicine is becoming ever more important. The focus on the genomic information of the individual patient informs decisions in diagnostics, prognosis, treatment, and prevention of cancer[1]. However, this also brings with it several challenges.

The full scope of genomic information available from a patient is diverse, consisting of many different 'omic platforms. Not only this, but the scale of individual types of data is also expanding, with an increasing amount of datasets and increasingly more extensive datasets[2].

One way to exploit this wider spectrum of data is by using the data on multiple 'omics platforms to improve the selection of biomarkers for one of these platforms. The use of multi-omic data in the selection of biomarkers for one omic platform could potentially improve selections as opposed to finding biomarkers based on just a single type of data.

In analyzing this sort of multi-omic data, one might analyze the individual data types to combine their conclusions at a later point. However, this risks not considering the broader scope of the data[3] and drawing multiple conclusions that could be difficult to reconcile. The heterogeneous nature of the available data calls for different approaches that can form conclusions in a broader spectrum of 'omics data.

Not only is there great variety in the different types of genomic information used, but also in the way of sampling this data. In oncology, solid biopsies of, for instance, tumorous tissue remains an important method of collecting patient data. However, solid biopsies are a relatively invasive and intensive procedure.

Liquid biopsies are another way of obtaining data on a patient. The sampling of bodily fluids is significantly less invasive than solid tissue biopsies. Usually, this liquid is blood, but other liquids such as urine or cerebrospinal fluid are also used. Liquid biopsy samples hold a wealth of genetic information on the patient from which it is obtained. For example, liquid biopsy samples could also potentially carry more information than a solid tissue biopsy of a single metastatic site in metastatic patients[4]. For these reasons, this thesis will focus on data obtained from liquid biopsies rather than the classical solid biopsy domain.

Now that we have established some potential reasons for performing a liquid- rather than a solid biopsy, we should consider how a liquid like blood contains the data required to be useful in a clinical context in the same way solid biopsy samples are.

Products present in the blood carry genetic information. One of these carriers of genetic data is cell-free DNA (cfDNA), fragments of free-floating DNA. This DNA is released into the blood through a number of pathways. Both controlled and uncontrolled cell death, as well as secretion through extracellular vesicles, release cfDNA into the blood, as shown in figure 1.1. This figure also shows the types of data that can be obtained from cfDNA, for instance: changes in methylation, single nucleotide polymorphisms (SNPs), or copy number variations (CNVs). Both healthy and tumorous cells release cfDNA into the blood.

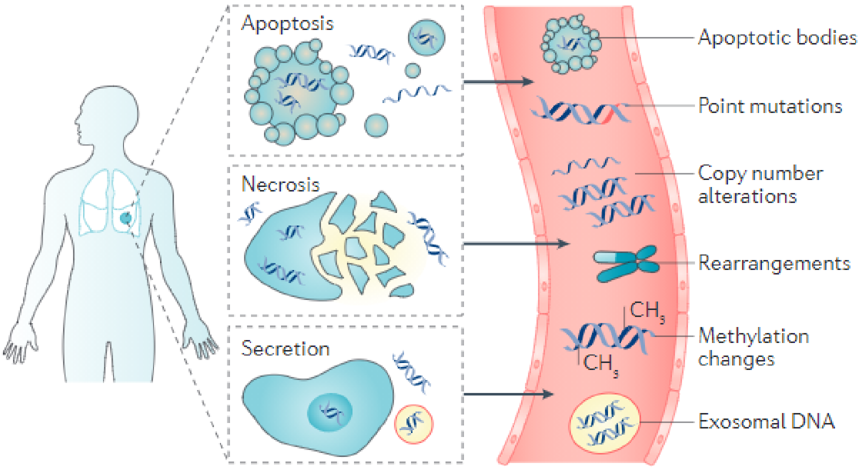


FIGURE 1.1: Origins and range of alterations in cell-free DNA, from [5]

Here, however, lies a potential challenge of translating results in a classical solid biopsy context to the domain of liquid biopsies, as there are significant differences between these two methods of sampling. For example, when sampling tumorous tissue in a solid biopsy, a large amount of the tissue that is sampled will be from tumorous tissue. In a liquid biopsy context, for instance, when working with cfDNA, the resulting signal mostly contains DNA from a healthy origin.

cfDNA is far from the only product that can be obtained from a liquid biopsy. Other products, for instance, circulating RNAs or circulating tumor cells (CTC) can also be found in blood[6]. In this thesis, however, the focus will be on cfDNA as data is more widely available for cfDNA than it is for example CTCs or RNA.

1.1 Objectives

Given the background above, we can lay out the objectives for this thesis.

We have already described that we will focus on multi-omic data in a liquid biopsy context and given our reasoning for these choices. Specifically, we wish to be able to specify biomarkers that appear in a cfDNA context that are indicative of the presence of cancer. Our hypothesis regarding this selection is that through the use of data on multiple 'omics platforms, we are able to find more biomarkers than what would be possible using only a single data type.

We have chosen to target changes in methylation in colorectal cancer specifically. These choices were made for a number of reasons. Solid biopsy data on colorectal cancer patients, especially methylation data was already widely available. On top of that, we had access to a dataset containing methylation data on colorectal cancer patients as well as healthy controls, by Erasmus University.

Besides these technical reasons colorectal cancer has one of the highest incidence rates worldwide, as well as having relatively high mortality compared to other types of cancer[7]. Detecting the presence of colorectal cancer early could significantly improve the chance of survival in such cases. Analysis of liquid biopsy samples could prove to be an essential method of early detection, potentially detecting tumors before they can be identified by current imaging techniques and sampled by classic solid tissue sampling.

This dataset contains data on metastatic colorectal cancer patients both before and after surgery, as well as information on whether and when a tumor recurred. Using this, we are not only able to analyze biomarkers for how well they differentiate between a colorectal cancer patient and a healthy individual. By combining the post-surgery data on a patient with the information on tumor recurrence, we can also analyze whether a biomarker appears to differentiate between a patient where a tumor will recur and a non-recurring patient. This latter analysis could point to biomarkers that indicate prognosis after surgery.

To summarize, our goals are as follows:

- This thesis centers around the primary goal of finding DNA methylation markers of colorectal cancer (CRC), with the focus being that these markers are also present in cfDNA samples.
- Additionally, this work selects biomarkers based on multiple data types and thus centers around the challenge of combining them in a meaningful way and comparing its results to a selection based on a single data type in isolation.
- Finally, another goal is to investigate whether potential DNA methylation markers are related to tumor recurrence.

Chapter 2

Methods

2.1 Data types and sources

This section will give an overview of the different data sources used in this thesis. The primary type of data used in the methods shown in this thesis consists of methylation data since our primary goal is to predict methylation markers for colorectal cancer. As we are focusing on finding these markers within blood, we also use data on normal blood samples.

Besides methylation data, this thesis also investigates the integration of other types of data in selecting methylation markers for colorectal cancer. Specifically, we also include gene expression data. Finally, in order to validate our findings we use two methylation datasets on cfDNA samples. An overview of the different datasets described here can also be found in table 2.1.

2.1.1 TCGA Pan-Cancer Atlas

Methylation data

Methylation data of solid tissue biopsies were obtained from the TCGA Pan-Cancer Atlas project[8] through the Xena browser[9]. The TCGA Pan-Cancer Atlas (PanCanAtlas or PANCAN) project contains methylation data on a great number of tissue samples for colorectal cancer, as well as a large variety of other cancers. This data encompasses 33 different types of tumors and also includes healthy tissue samples for a subset of the total patients ($n = 9639$) in the dataset. This data consists of Illumina HumanMethylation27, as well as Illumina HumanMethylation450[10] BeadChip though only the HumanMethylation450 data was used, targeting far more methylation sites than the HumanMethylation27 BeadChip.

Gene expression data

Besides methylation data, the TCGA Pan-Cancer Atlas also contains gene expression data on many of the samples from section 2.1.1.

The gene expression data used consisted of RNAseq data from the TCGA Pan-Cancer Atlas project, obtained through the Xena browser. This data was already processed to have batch effects removed, and the data were normalized, consisting of 20532 features across 11060 samples. This data was captured on the Illumina HiSeq platform.

2.1.2 Whole blood samples

Data on whole blood samples were obtained from studies by Hannum et al.[11] and Harris et al.[12]. These two datasets contain 656 and 20 samples of healthy subjects,

Name	Data type	Platform
TCGA Pan-Cancer Atlas[8]	DNA Methylation	Illumina HumanMethylation450 Illumina HumanMethylation27
TCGA Pan-Cancer Atlas[8]	RNA-seq	Illumina HiSeq
Hannum et al.[11]	DNA Methylation	Illumina HumanMethylation450
Harris et al.[12]	DNA Methylation	Illumina HumanMethylation450
Moss et al.[13]	DNA Methylation	Illumina MethylationEPIC
MeD-seq[14]	DNA Methylation	Methyl-dependent restriction enzyme based method

TABLE 2.1: Overview of the datasets used.

respectively. Both of these datasets were captured on the Illumina HumanMethylation450 BeadChip. These two datasets were combined to yield one dataset spanning 676 healthy blood samples.

2.1.3 Liquid biopsy data

A small, independent dataset of cfDNA samples originates from a study by Moss et al.[13]. This dataset contains 4 colorectal cancer (CRC) patients and 4 healthy controls as well as sepsis patients and lung- and breast cancer patients. These samples have been obtained on the Infinium MethylationEPIC BeadChip.

A separate dataset was obtained, which contains samples from 41 colorectal cancer patients at multiple points in time as well as 9 healthy controls, collected through Methylated DNA Sequencing (MeD-seq)[14], based on the use of a DNA methylation-dependent restriction enzyme.

The data on CRC patients were sampled at three points in time. *T0* samples were collected before the patients underwent surgery, *T1* samples were collected around one week after surgery and *T2* samples were collected at least two weeks after surgery.

2.2 Data preprocessing

2.2.1 Pan-Cancer Atlas datasets

For the data obtained from the Pan-Cancer Atlas, only samples that appeared in both the methylation data as well as the gene expression data were retained. This was done in order to be able to analyze this data in a multi-omic context, eliminating samples for which only one data type was available.

2.2.2 Grouping individual CpG sites to transcription start sites

Only CpG sites within a 2000 base pair region around the transcription start site of genes were included. The transcription start sites (TSS) were determined through the Ensembl human genome annotation.[15] In particular, version 104 was used. Based on this assembly, the groups of interest spanned 2kb, centered around the individual TSSs.

The function of this grouping is twofold, both to be able to directly link differential methylation of a group to a transcription start site as well as to reduce the number of features in our data.

With the use of the Ensembl human genome annotation, for each transcript, the start location on the genome was found. The region 1000kb upstream up to 1000kb downstream of this location was denoted as a single TSS region belonging to this transcript. This yielded a total of 206479 regions surrounding a transcription start site.

Illumina platforms

The datasets that were generated on the Illumina platforms use beta values as their measure of methylation. For a CpG site probe, a beta value is defined as the ratio of methylated intensity and overall intensity. This ratio results in a number between 0 and 1. A beta value of zero would indicate that a CpG site is completely unmethylated, while a value of 1 would mean that a CpG site is methylated in all measured copies of that site[16].

The datasets that use the Illumina HumanMethylation450 platform, as well as the Illumina MethylationEPIC platform, were transformed to this approach that groups individual values based on transcription start sites. Firstly, only CpG sites with values for $> 50\%$ of the total samples were retained. For each TSS, the beta value was taken to be the mean of the beta values of the CpG sites within the 2kb region centered around that TSS.

MeD-seq

In contrast to the Illumina platform, the MeD-seq sequencing technique does not yield the methylation level of a CpG site relative to the number of unmethylated copies that are read. MeD-seq sequencing yields a count of methylated copies. Thus, a value of 0 would, similarly to the beta value used by the Illumina platform also indicate that no methylated copies of a CpG site were read. However, this value is not constrained to be between 0 and 1, and for values greater than 0 defines only a measure of how many methylated CpG sites were read without comparing this count to any unmethylated copies of a CpG site.

In the case of the dataset based on the MeD-seq sequencing, the counts of the individual CpG sites are summed to yield a total count for each 2kb region centered on a TSS. Thus, the value of such a region represents all counts for sites within that region. Transforming the original data in this way yields 112609 TSS regions, as not all of the regions in section 2.2.2 are covered by the MeD-seq sequencing. Finally, trimmed mean of M values (TMM) normalization[17] was then applied on the MeD-seq dataset.

2.3 Multi-omic integration

2.3.1 JIVE

The integration of data across multiple types of genetic information is performed through JIVE[18]. JIVE was applied to the preprocessed TCGA Pan-Cancer Atlas methylation and gene expression data as two blocks across the same samples. This yielded a decomposition of the original data into:

- A joint term
- Individual terms
- Noise

This attempts to capture patterns present across different types of data, as well as signals that occur only in one individual data type. Figure A.1 in appendix A shows an example of data consisting of these three components. In particular, the R.JIVE implementation was used[19].

A formal description of the JIVE model is given as:

$$\begin{aligned} X_1 &= J_1 + A_1 + \epsilon_1 \\ X_k &= J_k + A_k + \epsilon_k, \end{aligned} \tag{2.1}$$

Where X_1 up to X_k are $n \times p_i$ matrices corresponding to k different datasets of p_i variables on n samples. Thus the number of variables in a dataset can differ from other datasets, though always over the same samples. J_1 up to J_k are the submatrices of a single joint matrix J corresponding to X_i. A_1 up to A_k represent the individual matrices for dataset X_i.

As in this thesis, only two different data types are considered for our purposes the JIVE model can be simplified to:

$$\begin{aligned} X_M &= J_M + A_M + \epsilon_M \\ X_{GE} &= J_{GE} + A_{GE} + \epsilon_{GE}, \end{aligned} \tag{2.2}$$

Where M denotes the methylation dataset and GE denotes the gene expression dataset.

The joint and individual terms in JIVE are constrained to be of a certain rank. This choice of these ranks is important as it functions as a measure of how much variance should be accounted for in the joint component and how much the dimensionality of the data is reduced. This rank constraint is applied to only those two terms. The dimensionality reduction that is performed by restricting both the joint and individual components to a lower rank than the input data is responsible for accounting for noise and separating it from the joint and individual effects.

The JIVE algorithm can be summarized as follows:

1. Take the complete data $X = [X_1, X_2, ..., X_k]$ as the joint term X_{joint}.
2. Find a rank r singular value decomposition of X_{joint} being $J = [J_1, J_2, ..., J_k]$
3. For each data type i, $A_i = X_i - J_i$
4. Take the rank r_i SVD of A_i as the new individual term $X_i^{individual}$
5. Update X_{joint} to be $X_i - A_i$
6. Select new rank constraints

This is then performed iteratively until it converges or a maximum number of iterations has been reached.

The algorithm that is used to estimate the ranks is included in appendix B

2.3.2 aJIVE

A number of other techniques have by now been developed that extend on JIVE, altering the process by which decompositions are generated. One of these techniques is aJIVE, or *angular JIVE*[20].

As one of the most impactful changes to the method, aJIVE foregoes the iterative algorithm that JIVE uses. Instead, aJIVE favors another approach where singular value decomposition is sequentially applied a fixed number of times to obtain the joint and individual components.

The changes that aJIVE makes to the JIVE technique both decrease runtime and are also aimed at improving the quality of the decompositions, especially when there exists correlation between individual components.

An example of this can be found in appendix C, where data is constructed consisting of a joint and individual component for blocks X and Y in figure C.1. Both an aJIVE and JIVE decomposition are given in figures C.2 and C.3 respectively. While the aJIVE decomposition manages to deconstruct the original signal accurately, the JIVE decomposition erroneously attributes underlying individual effects to the joint component.

2.4 Selection of differentially methylated TSSs

To select regions that are significantly differentially methylated in tumorous samples as opposed to healthy samples in cfDNA, five different selection methods were used:

- Thresholding
- JIVE loadings
- aJIVE loadings
- Random Forest
- A combined selection

Figure 2.1 gives a summarized overview of how these methods fit together and what type of data they use for their selection. These next sections will describe the details for each selection method.

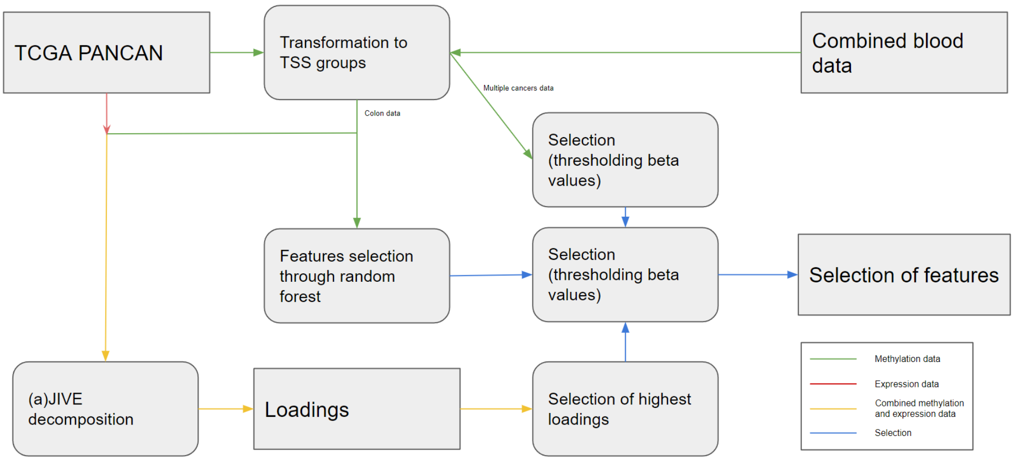

TCGA PANCAN
Transformation to TSS groups
Combined blood data
Multiple cancers data
Colon data
Selection (thresholding beta values)
Features selection through random forest
Selection (thresholding beta values)
Selection of features
(a)JIVE decomposition
Loadings
Selection of highest loadings
Methylation data
Expression data
Combined methylation and expression data
Selection

FIGURE 2.1: A summary of the method used to find a selection of TSS regions.

2.4.1 Thresholding approach

This approach is based, in part, on works by Cho et al.[21] and Lange et al.[22]. This method of selecting TSS regions compares tumorous versus normal tissue samples, as well as healthy tissue versus normal blood samples. The reasoning here is that the methylation level of selected regions should differ significantly in tumorous versus healthy tissue while not appearing to be methylated differently in healthy tissue versus blood. The hypothesis here is that this will select regions that will also appear differently methylated in cfDNA samples.

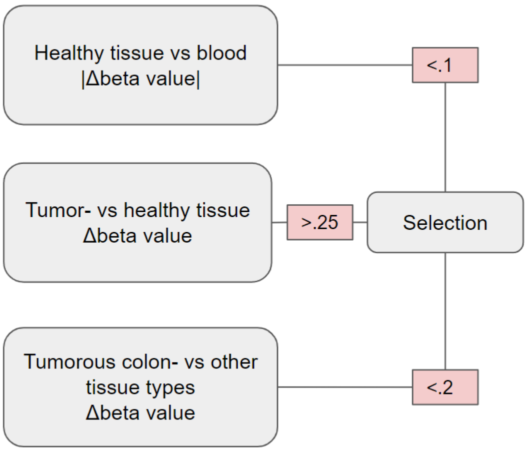


FIGURE 2.2: Overview of the different steps in the thresholding selection method. Example values are used for the thresholds denoted in red.

Δbeta values are calculated for each individual TSS across the set of tumorous colon tissue samples versus the set of healthy tissue samples by subtracting the mean beta value across the tumorous samples by the mean beta value across the healthy samples. This value represents the difference in beta value between these two sets. A threshold is then applied to these Δbeta values to select TSS regions that are significantly different in tumorous tissue as opposed to healthy tissue. This step corresponds to the top section of figure 2.2.

The next step is to filter the TSS regions selected in the previous step by their absolute Δbeta value in a healthy tissue versus whole blood sample comparison. This excludes regions with means that are significantly different between these two groups. In figure 2.2 this step is represented by the middle block.

Finally, another filtering step is then applied on the remaining TSS regions that calculates Δbeta values for tumorous colon tissue versus each other tumorous tissue type. TSSs that do not have an absolute delta-beta value larger than a certain threshold are then filtered out in order to select TSSs where the difference in mean beta value is unique to tumorous colon tissue. This corresponds to the lower part of figure 2.2.

2.4.2 JIVE loadings

Similar to principal component analysis, JIVE can yield scores and loadings. For each block X_i, this yields a scores and loadings matrix, U_i and S respectively for the joint component and W_i and S_i respectively for the individual component. A representation of the JIVE model including these scores and loadings is given in figure 2.3. These loadings were then used in forming a selection of TSSs.

$$\begin{aligned} X_1 &= U_1 S + W_1 S_1 + R_1 \\ &\vdots \\ X_k &= U_k S + W_k S_k + R_k . \end{aligned}$$

FIGURE 2.3: Factorized version of the JIVE model. From [18].

Once JIVE had obtained a decomposition as described in 2.3.1, the *jive.predict* function was used on these same two blocks to obtain loadings for the joint and individual components. The loadings for the joint component are given as a single matrix, which was split to match each individual data type. Both the joint and individual loadings were used to generate selections of transcription start sites.

Depending on the rank of the joint or individual component chosen by the JIVE algorithm, the number of columns in the loadings matrix varies. The top 200 highest absolute loadings were obtained for each column of the joint and individual spaces. The TSS regions corresponding to these were considered selected. It is to be noted that the individual component is more likely to be of higher rank, and thus selects a larger amount of TSS regions.

The idea behind this method of selecting potential methylation markers for colorectal cancer is that TSS regions with the highest loadings should be responsible for a relatively high amount of variance in the data. Since within a dataset of both normal tissue and tumorous tissue samples, it would seem likely that a large amount of variance would stem from the difference between these two groups. Through this, we expect TSS regions to be selected that will be significantly differentially methylated between normal and tumorous tissue.

aJIVE loadings

In a similar fashion to the application of JIVE detailed above, aJIVE[20] was also applied to generate a decomposition and loadings for the joint and individual components of this decomposition.

This was performed similarly to the steps in section 2.3.1 through the *ajive* library in R. Unlike the steps used in obtaining the loadings for JIVE, they could be directly obtained through the *get_block_loadings* method in the *ajive* package.

2.4.3 Random forest

A random forest classifier model was also used to obtain a selection of transcription start sites. The *RandomForestClassfier* from *scikit-learn*[23] was used for this. The training data consisted of a subset of the TCGA Pan-Cancer Atlas methylation dataset which contained samples of both normal colon tissue and tumorous colon tissue, with labels corresponding to their tumorous or non-tumorous origin. For this classifier, 200 trees were used. Other hyperparameters for the classifier were left to the *scikit-learn* defaults and are described in appendix E. The importance weights were then used to obtain a selection of features with the *SelectFromModel* method.

This method of selecting potential methylation markers is especially interesting in a comparison with the selection methods that use multiple data types. It serves primarily to highlight the performance of both the thresholding approach from section 2.4.1 as well as the multi-omic approaches in light of a relatively complex method that only takes a single data type into account.

2.4.4 Combining selections

While every selection type on its own is capable of providing meaningful selections of TSS regions, we hypothesize that a combination of these selections could yield a selection that contains more significantly differentially methylated regions proportional to the number of regions selected.

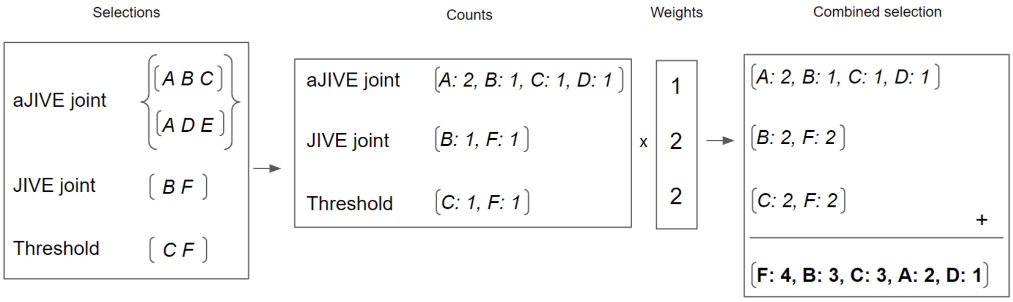


FIGURE 2.4: Process by which the individual selections are combined. The weights shown here are to serve as an example and can be any real number. Letters are used to represent TSS regions. The final combined selection is denoted in bold.

Through the steps above, which are also shown in figure 2.4, multiple selections can be combined into a single selection. Individual selections were combined by counting how often each TSS region occurred in each selection. The different types of selection were weighted differently with weights assigned to each type of selection. These weights were selected in order to balance the JIVE and aJIVE joint and individual components.

In practice, this means reducing the weight on the individual component. As noted earlier in section 2.4.2, the individual component is more likely to be of higher rank and thus select more TSS regions. Through this, the unweighted selection of the individual component could overpower the other selections. The reduced weight compensates for this.

The resulting, combined selection consists of scored transcription start site regions. Of these, the top 100 highest scoring regions are taken as the combined selection.

2.5 Differential analysis

2.5.1 Liquid biopsy data

To validate transcription start sites that are expected to have differential methylation, found by this method, a small independent dataset of cfDNA samples by Moss et al. described in section 2.1.3 is used. For each TSS that is selected, an independent two-sided t-test is performed between the group of healthy samples and the group of colorectal cancer patients to determine whether a TSS is indeed significantly differentially methylated in the tumor group. Similar to the Pan-Cancer methylation data, this is performed through the *limma* package in R. Bonferroni correction is then applied to account for multiple testing.

2.5.2 MeD-seq data

Trimmed mean of M values, TMM normalization[17], was applied on the MeD-seq dataset consisting of 41 colorectal cancer patients and 9 healthy controls through the EdgeR package in R. Differential methylation analysis is performed through the EdgeR[24] package. For this *exactTest* is used. This performs the exact test by Robinson and Smyth[25], yielding two-sided p-values and fold change among other statistics.

The exact test was performed on two tasks. Firstly, the test was performed comparing the T0, before surgery, timepoint of the 41 colorectal cancer patients versus the 9 healthy controls. Secondly, the exact test was performed on data from the T2 timepoint, which was recorded >2 weeks after surgery as described in section 2.1.3. This data consisted of 13 patients in which the disease recurred within one year, versus 17 patients in which it did not recur within one year.

The exact test used by EdgeR does not automatically account for multiple testing. To account for this, multiple testing correction is applied to the p-values that *exactTest* returns. For this, *p.ajust* in the R stats package is used which is used with Bonferroni correction.

2.6 Evaluation

We want to test whether the selection of TSS regions produced by the different presented selection methods actually contains regions that are significantly differently methylated in tumorous tissue versus normal tissue.

In order to test this, p-values and fold changes were generated for each TSS region in the TCGA Pan-Cancer Atlas data. The R package *limma*[26] is used for this analysis. Bonferroni correction was applied afterward to the uncorrected p-values obtained from *limma* through the *p.ajust* function in the R stats package as multiple testing correction.

Two tests were performed here, one for the full TCGA dataset including all types of tissue, as well as one for a subset of the TCGA data including only colon samples. We denoted TSS regions with a p-value of < 0.05 as being significantly differently methylated in tumorous tissue versus normal tissue. For every selection method, the amount of TSS regions with a p-value of < 0.05 in these tests was calculated, as well as the percentage of significant regions out of the total amount of regions selected.

Chapter 3

Results

3.1 Significance of found transcription start sites

We wish to test whether transcription start site regions selected by the different methods presented actually are differentially methylated in tumorous tissue samples versus normal tissue samples in the TCGA Pan-Cancer data. In order to do so, the analysis described in 2.6 was performed.

These results of this analysis are given in table 3.1. This table lists each selection method, as well as the total amount (regardless of significance) of TSS regions selected by each method. For the JIVE and aJIVE methods, r denotes the rank of the joint and individual components. This rank determines the number of selected regions as noted in section 2.4.2.

A volcano plot was generated for the TCGA methylation data with the combined selection highlighted, visualized in figure 3.1. In this plot, the horizontal red line denotes a p-value of 0.05 and the 100 transcription start sites from the combined selection as performed in 2.4.4 are highlighted in red. Of the selection of 100 transcription start sites, 89 have p-values < 0.05. Volcano plots with highlighted selections for each individual selection method were also created. These individual selections can be found in appendix D.

The results in table 3.1 indicate that most of the methods for selecting transcription start site regions yield significant TSS regions for a large portion of the total regions selected. Another conclusion that can be drawn here is that, throughout each selection method except for the thresholding approach, the selections seem to contain more methylation markers that are significant in the subset of TGCA data from colon samples, thus highlighting that the techniques for selecting TSS regions are somewhat colon-specific. Finally, it can be noted that selections that are based on the JIVE and aJIVE methods select, by far, the greatest amount of TSS regions that are actually significant in the PANCAN bulk data.

For the JIVE individual component, due to its considerably higher rank than the other JIVE and aJIVE components, it produces a relatively large selection. This leads both to it having a high number of significant TSS regions, while also having a very low percentage of significant regions relative to the total number of selected regions.

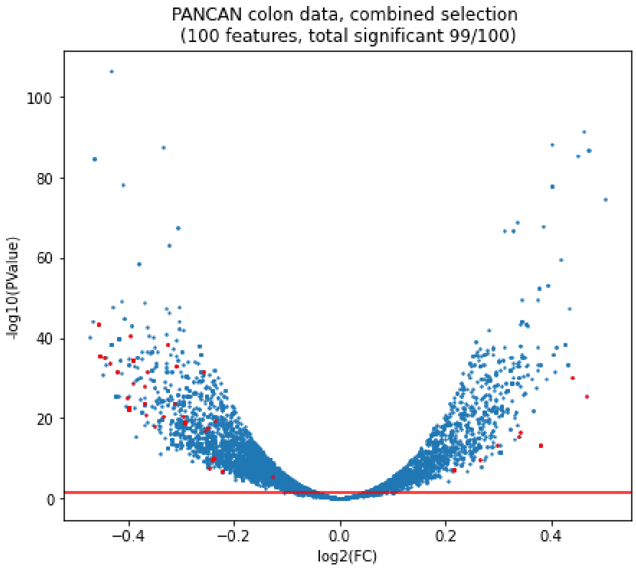


FIGURE 3.1: Volcano plot for the combined selection of 100 transcription start sites. Selected TSS regions from the combined selection are marked in red.

Selection	Total	PANCAN colon tissue tumor vs normal	PANCAN all tissue tumor vs normal
Thresholding	37	37 (100%)	37 (100%)
JIVE joint (r=1)	200	200 (100%)	185 (92.5%)
JIVE individual (r=12)	2400	561 (23.4%)	553 (23.0%)
aJIVE joint (r=2)	400	350 (87.5%)	314 (78.5%)
Random forest	325	184 (56.6%)	160 (49.2%)
Combined	100	99 (99%)	97 (97%)

TABLE 3.1: The number of significant ($p < 0.05$) selections for each type of selection in the PANCAN tissue differential analyses.

3.2 Validation on BeadChip liquid biopsy data

To validate the TSS regions selected by the different selection methods in a liquid biopsy context, we first turn to a relatively small ($n = 8$) dataset by Moss et al[27]. The choice for this dataset was made as it, like the PANCAN dataset, has been captured on the Illumina MethylationEPIC platform. This simplifies the comparison to the PANCAN data, as this is relatively similar to the Illumina HumanMethylation450 platform. Platforms yielding, for instance, different types of values could confound results.

Every selection method was tested on this dataset as described in methods section 2.5.1, the results of which are included in table 3.2. For the combined selection, the overlap between the set of TSS regions with statistically significant differential methylation ($p < 0.05$) and the selected TSS regions was determined, yielding 67 significant out of the 100 selected TSS regions.

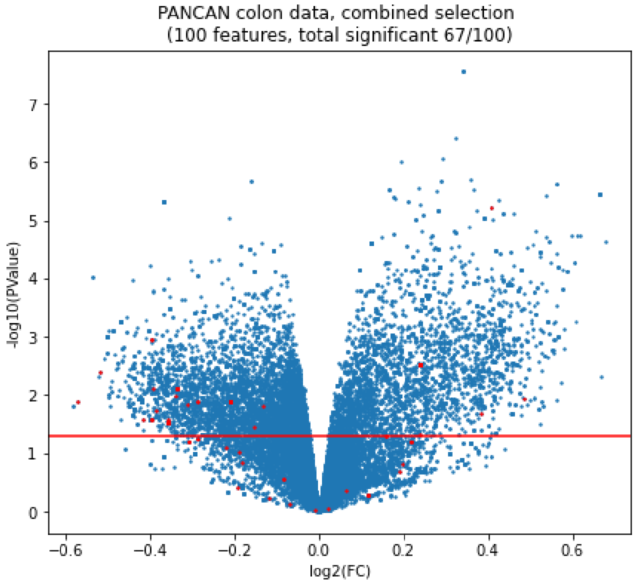


FIGURE 3.2: Volcano plot visualizing TSS regions in the GSE122126 data for the CRC patients versus healthy controls. Selected TSS regions from the combined selection are marked in red.

Selection	Total	Moss et al. cfDNA data
Thresholding	37	29 (78.4%)
JIVE joint (r=1)	200	124 (62%)
JIVE individual (r=12)	2400	272 (11.3%)
aJIVE joint (r=2)	400	174 (43.5%)
Random forest	325	145 (44.6%)
Combined	100	67 (67%)

TABLE 3.2: The number of significant ($p < 0.05$) selections for the differential methylation analysis on the Moss et al. cfDNA data.

These results show that the results in section 3.1 seem to also apply to an independent, albeit small, dataset on liquid biopsy data. Primarily that for every selection method a relatively large amount of TSS regions appear significant in the data by Moss et al.

Especially the selections based on the JIVE and aJIVE joint components select a large amount of significant TSS regions, while also having a high percentage of significant regions out of the total amount of regions selected. Although it is also to be noted that random forest performs similarly to the methods using multiple data types in this data.

3.3 Validation on MeD-seq data

After having validated the selected TSS regions on the dataset by Moss et al., we want to perform a similar validation on the data captured through the MeD-seq platform. The different selection methods were tested against a group of CRC patients at timepoint T0 versus a group of healthy controls.

The number of selected regions that appeared significant in this test are given in table 3.3. The combined selection method is highlighted in the visualization of figure 3.3. This yielded only one region associated with the transcript *CTNND2-201* that appeared in the combined selection of 100 regions that were found to be significant ($p < 0.05$) in the Bonferroni corrected MeD-seq p-values.

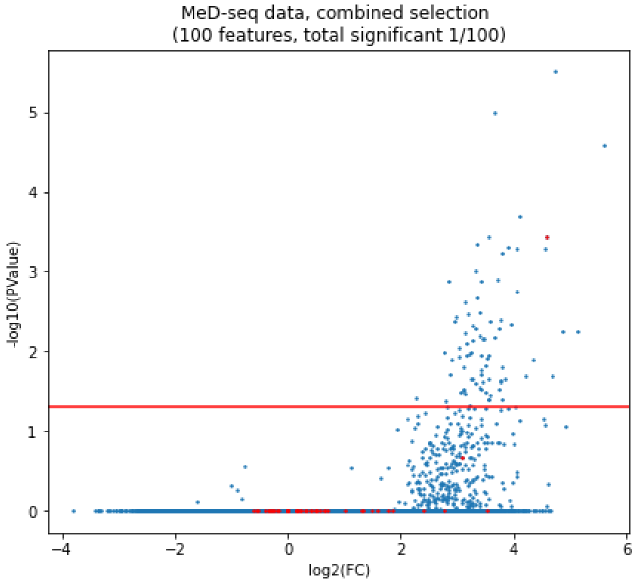


FIGURE 3.3: Volcano plot visualizing TSS regions in the MeD-seq data for the T0 CRC group versus healthy controls. Selected TSS regions from the combined selection are marked in red.

Selection	Total	MeD-seq T0 CRC vs Healthy
Thresholding	37	0 (0%)
JIVE joint (r=1)	200	0 (0%)
JIVE individual (r=12)	2400	1 (<1%)
aJIVE joint (r=2)	400	1 (<1%)
Random forest	325	0 (0%)
Combined	100	1 (1%)

TABLE 3.3: Number of significant ($p < 0.05$) selections for the MeD-seq differential analysis.

From these results, it seems that there is very little overlap between regions that appear significant in the MeD-seq data and the selections that are significant in the PANCAN colon methylation data (table 3.1) as well as the cfDNA microarray dataset by Moss et al. (table 3.2). As is apparent from the results presented above, nearly none of the selected TSS regions across every selection method are significant in the MeD-seq data. Only one of these selections, transcript *CTNND2-201* appeared significant in both analyses.

This particular gene might be notable, however, as this same TSS region did also appear significant in both the PANCAN colon methylation data as well the independent cfDNA dataset by Moss et al as detailed in section 3.2.

Another observation that can be made here is that only the JIVE- and aJIVE based methods find this singular TSS region. The combined selection also includes this region, but it is a combination of the individual methods. This also serves to underline that the use of multiple data types increases the ability to find differentially methylated regions in this case.

3.4 Selection of differentially methylated TSS regions in tissue

We have now reflected on how many of the regions selected by the methods presented are significant in a number of different datasets. Although until now we have not considered how these methods actually compare to simply performing differential methylation analysis on the TCGA tissue samples and using the regions that appear significant as a selection.

The reasoning behind performing this additional analysis is simple. If testing the tumorous versus normal groups in the TCGA data and using regions with p-values under a certain threshold as a selection would yield better performance to the selection methods presented, there would be no benefit to using the presented selection methods.

This selection is defined as regions that appear significant ($p < 0.05$) in the differential analysis as performed with the R *limma* package as described in 2.6. This was compared to the EdgeR analysis of the MeD-seq data. More specifically, the number of regions that overlap between the TSS regions with low p-values ($p < 0.05$) in the MeD-seq data and the TCGA tissue data was calculated.

A volcano plot for the TSS regions in the MeD-seq data is given in figure 3.4, with the selections based on the significance of TSS regions in the test on the TCGA tissue data for tumorous tissue versus normal tissue.

This yielded three significant TSS regions, including the same region found to be significantly differently methylated in section 3.3, *CTNND2-201*.

It is to be noted that the amount of TSS regions that appeared significant from the differential analysis of the tissue data (8955) was far greater than the number of selected TSS regions in the combined selection of 3.3 (100). Thus, it does not appear that simply testing each region in the tumorous versus normal TCGA tissue samples and selecting regions with low p-values performs better than the selection methods presented. When only selecting the TSS regions with the 1000 lowest p-values from the differential methylation analysis of the tissue data, none appear significant in the MeD-seq data.

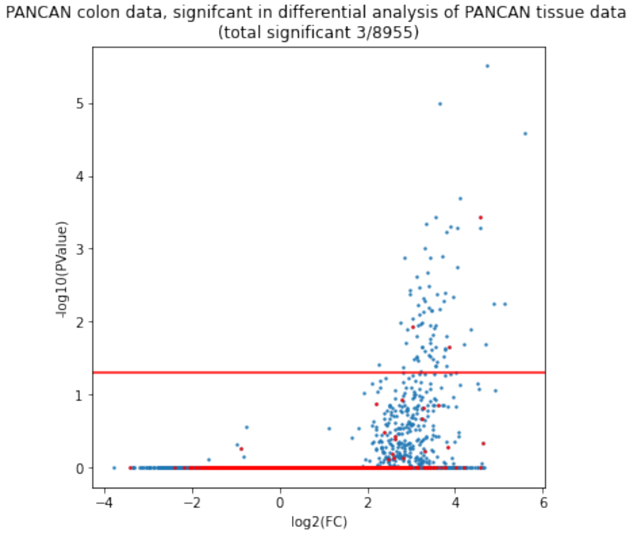


FIGURE 3.4: Volcano plot visualizing TSS regions in the MeD-seq data for the T0 CRC group versus healthy controls. Significant TSS regions from the solid tissue differential methylation analysis are marked in red.

3.5 Prognosis prediction

The MeD-seq data is recorded over several points in time for each patient, as described in section 2.1.3. This allows, not only, for finding markers that differentiate between healthy controls and colorectal cancer patients, but also for comparing patients with and without recurrence. The different selections of genes were tested against a group of patients with less than 1 year of disease-free survivial versus a group of patients with greater than one year of DFS in the MeD-seq data. The combined selection yielded one transcription start site region that appeared significant in the MeD-seq data across a combined selection as performed in 2.4.4 of 100 TSS regions. This TSS corresponded to the *SLIT2-202* transcript.

Selection	Total	MeD-seq <1y DFS vs >1y DFS
Thresholding	37	0 (0%)
JIVE joint (r=1)	200	0 (0%)
JIVE individual (r=12)	2400	3 (<1%)
aJIVE joint (r=2)	400	1 (<1%)
Random forest	325	3 (<1%)
Combined	100	1 (1%)
PANCAN Colon differential methylation	8955	37 (<1%)

TABLE 3.4: Number of significant ($p < 0.05$) selections for the MeD-seq T0 vs T2 differential analysis.

The results for every individual selection is given in table 3.4. Here, the selection of transcripts that appear significant ($p < 0.05$) in the differential methylation analysis on the PANCAN colon data as performed in 2.6 were also considered, denoted as 'PANCAN Colon differential methylation' in the table above. 37 of the transcription start site regions that appeared significant in the PANCAN tumorous versus normal tissue analysis were also significant in the MeD-seq analysis when comparing patients with less than 1 year of disease-free survival versus the group of patients with greater than one year of DFS in the MeD-seq data.

Chapter 4

Discussion

4.1 Findings

In this thesis, we have presented methods for selecting regions of the genome associated with transcription start sites that could serve as methylation-based biomarkers for colorectal cancer. The JIVE- and aJIVE based methods take into account gene expression data, as well as methylation data to produce these markers.

Firstly, the results in section 3.1 show that a large part of the TSS regions selected by the presented methods is differentially methylated in the TCGA data. These results are primarily presented in table 3.1. Along with this, the results also show that the JIVE and aJIVE methods generate the most TSS regions that appear significant in the TCGA data in an absolute sense. The thresholding method produces a lower number of selected TSS regions, but all of them appear significant. It also demonstrated the ability of the combined method to integrate the different individual methods into a single selection of TSS regions with 99% of selected regions being significant in the TCGA data.

Secondly, the individual selections were validated on an independent dataset of methylation samples obtained through liquid biopsies. The data resulting from this was obtained on the Illumina MethylationEPIC platform, a similar platform to the platform on which the TCGA tissue data was obtained. Here, primarily in table 3.2, we see that the conclusions that can be drawn from section 3.1 seem to correspond to these results. Again, the selection methods presented seem to select TSS regions that are differentially methylated, also in this dataset. Here we can also note that the JIVE- and aJIVE based selection methods, especially when looking at the joint components, produce a large number of significant selections, both absolute as well as relative to the total number of regions selected.

The results from sections 3.1 and 2.5.1 do not seem to translate to the MeD-seq data. From table 3.3 we see that only one of the selected regions throughout all of the different selection methods appears to be significantly differentially methylated between CRC patients and healthy controls in this data.

A simple alternative method for selecting TSS regions based on TSS regions that appear significantly differentially methylated in the TCGA tissue data when comparing tumorous tissue versus normal tissue was also considered in section 3.4. This produced three TSS regions that appeared differentially methylated in the analysis on the MeD-seq data. However, the large amount of TSS regions that appear significant in the TCGA dataset makes this a comparatively poor method of selecting potential methylation markers for colorectal cancer.

Besides the selection of biomarkers for the presence of colorectal cancer, the MeD-seq data also allows us to validate whether TSS regions are differentially methylated in a group of patients with less than 1 year of disease-free survival versus a group of patients with greater than one year of DFS. From the findings, primarily in table

3.4, we can note that like the validation of biomarkers for the presence of CRC, the different selections presented only find one TSS region that appears significantly differentially methylated.

4.2 The benefit of integrating multiple data types

This thesis is primarily focused on the use of multiple data types in finding methylation markers for colorectal cancer. The methods for doing so presented in this thesis appear to assist in finding significant methylation markers. This conclusion stems from the results in section 3.1 as well as from section 3.3. Selecting the largest amount of significant regions in the former and selecting the only significant region in the latter.

One thing to note is that the JIVE methods, as well as derivatives of this method (like aJIVE), are capable of integrating far more different types than just the two types shown in this thesis. Thus, research into integrating more than just the two data types shown in this thesis might be able to show how changes in the amount of different data types or the particular types used might influence the discovery of biomarkers for one particular data type.

The selection methods presented here only focus on combining two particular types of data, methylation and gene expression data. The reasoning for this is that these two data types are available for a relatively large number of samples within The Cancer Genome Atlas project, adding more data types to the analysis would have decreased the number of samples for which all of these types are available.

4.3 Results on microarray data do not translate to MeD-seq

From the results presented in section 3.3 and 3.2 it is clear that there is a significant difference between the amount of selected TSS regions that appear significant in the independent microarray cfDNA dataset compared to the MeD-seq dataset.

A large amount of TSS regions that were found in the tissue data through the different selection methods could be verified to also be significantly differentially methylated in the cfDNA microarray data. This would support the idea that the different methods of selecting potential methylation markers for CRC in tissue presented here are capable of finding potential biomarkers for CRC in the cfDNA domain. This leaves the question of why the data captured in the cfDNA domain on one platform, Illumina HumanMethyalation450 in this case, does not translate to data captured through MeD-seq.

One potential explanation can be found when comparing the volcano plots in figures 3.2 and 3.3. One thing that is apparent here is the far greater number of significant TSS regions that appear differentially methylated between the CRC patients and the healthy controls in the microarray dataset. Because of this, the chance of (part of) a selection appearing in the set of significant TSS regions also increases.

Still, it remains a somewhat puzzling find and potentially warrants further research into how methods based on one platform can be translated to the greater scope of available data. As well as if results stemming from data obtained on a single platform are a good representation of the real-world processes that this data describes.

4.4 Literature concerning *CTNND2* and *SLIT2*

Following the results from section 3.3 and section 3.5, an informal search for literature concerning these two genes, in particular, was performed.

4.4.1 CTNND2

This particular gene has been shown to be related to cancer formation as found in the RefSeq sequence database[28], although not specifically said to be related to colorectal cancer. Another work details that the protein coded for by this gene "to be overexpressed in various types of cancers"[29].

While not being directly connected to colorectal cancer, it is clear that this gene and proteins coded by this gene seem to play a role in the formation of cancer. The study by Huang et al. cited above concludes that -Catenin, a protein coded for by *CTNND2*, promotes the malignancy of human lung cancer. Next to this, the authors also note its potential as a biomarker for various types of cancers.

4.4.2 SLIT2

The *SLIT2* has been quoted by various sources to be related to colorectal cancer. One study reports overexpression of the *SLIT2* gene in intestinal tumors[30]. Other studies also report methylation of regions of the DNA associated with the SLIT2 gene in cancer-affected patients[31][32].

One of these latter studies by Dallol et al.[31] suggests that the *SLIT2* gene acts as a tumor suppressant and is silenced through methylation, not only in colorectal cancer patients but also in lung and breast tumors.

4.5 Concluding remarks

This thesis has detailed a number of methods of finding methylation markers for colorectal cancer, as well as a combination of these methods. These selections were then evaluated for how many of their selections were significant in multiple datasets, both in tissue as well as in a liquid context. These tests compared tumor samples versus normal samples, as well as patients with tumor recurrence as well as non-recurring patients.

In these tests, results were given as an amount of transcription start site regions that appeared significant in the tested data, both as an absolute number as well as a percentage of the total amount of regions selected.

There is, however, a certain difficultly in judging the performance of a method as well as comparing performance between different methods. When both the absolute amount of regions with significant differential methylation as well as the percentage relative to the total amount of regions selected is much higher for one method than another, this comparison is relatively clear. However, when two methods are closer or one method produces far more selections than another, thus obtaining a high absolute amount of significant selections but a low percentage compared to the total amount of selections, the comparison becomes blurry.

In an ideal scenario, we wish to evaluate selections by how well they select TSS regions associated with genes that actually play a role in colorectal cancer. For two of our selected regions, we have found literature supporting the hypothesis that they play a role in cancer in section 4.4 of this discussion. But we have also demonstrated in section 4.3 that our findings on one platform do not necessarily translate to another platform, or that they are indicators of real-world processes.

Finally, in light of this, the methods of selecting markers for colorectal cancer presented in this thesis have selected some TSS regions corresponding to genes associated with (colorectal) cancer, but that more research is needed in investigating how well methods like these are actually capable of identifying markers that can be confirmed in other platforms and how specific their selections truly are.

Appendix A

JIVE example visualizations

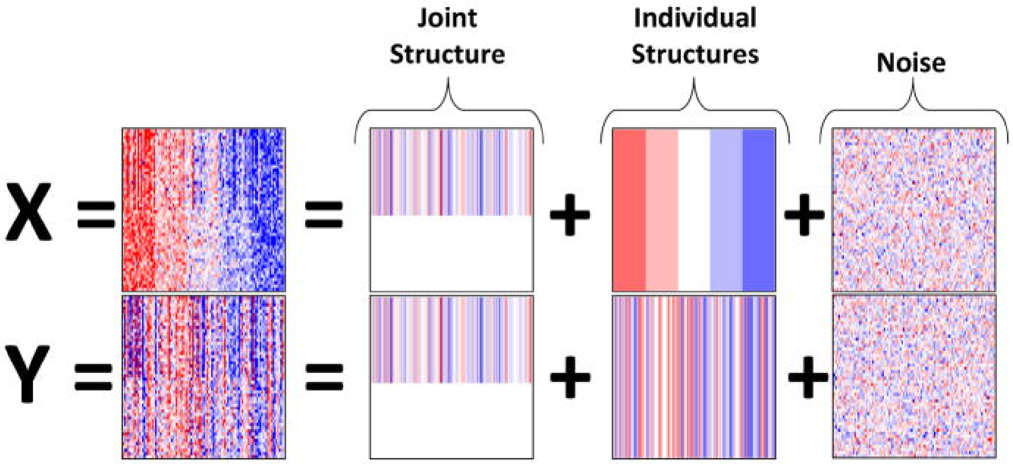


FIGURE A.1: JIVE example for two data blocks. From [18].

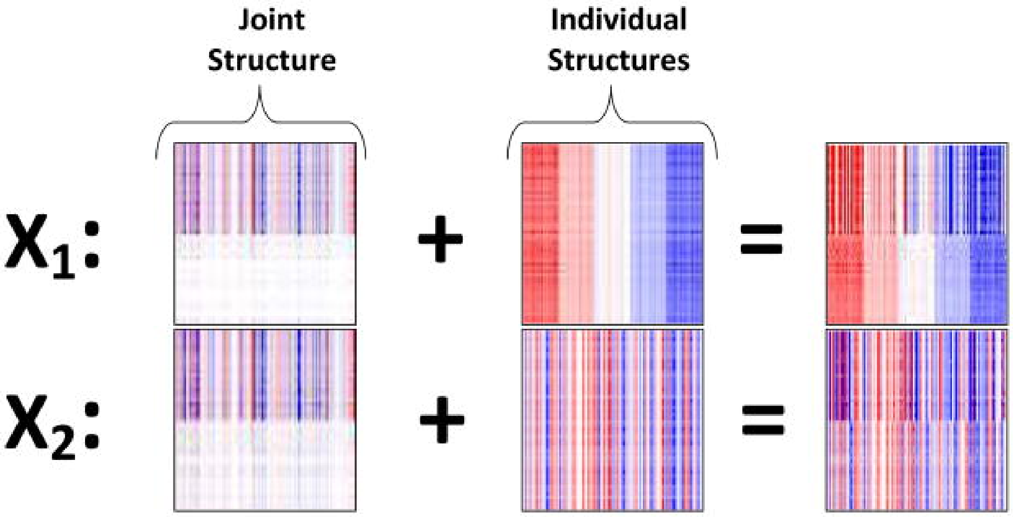
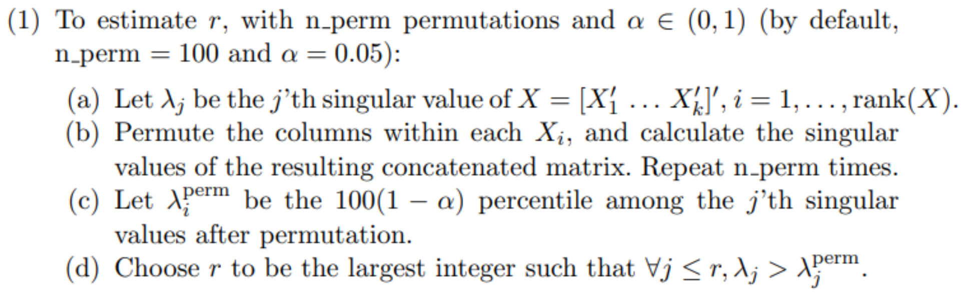


FIGURE A.2: JIVE estimates for the blocks from figure A.1. From [18].

Appendix B

JIVE rank selection

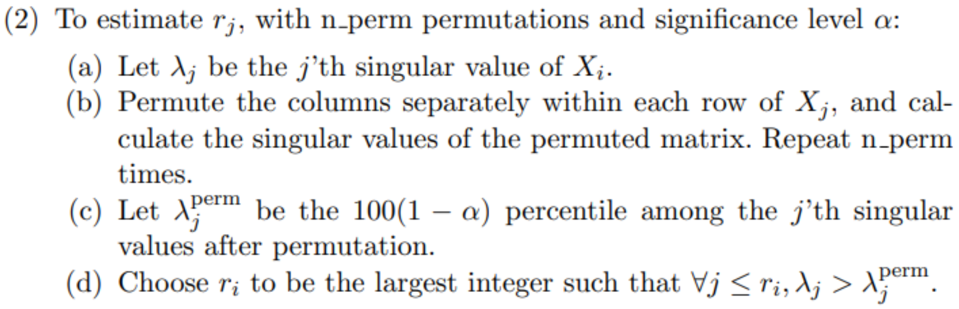
On both the joint and the individual matrices, the ranks are constrained to be r for the joint term and r_i for $i = 1, ..., k$ individual terms corresponding to k datasets. Below are given the algorithms through which the initial rank constraints are found.



(1) To estimate r, with n_perm permutations and $\alpha \in (0, 1)$ (by default, n_perm $= 100$ and $\alpha = 0.05$):

(a) Let λ_j be the j'th singular value of $X = [X_1' \dots X_k']'$, $i = 1, \dots, \text{rank}(X)$.
(b) Permute the columns within each X_i, and calculate the singular values of the resulting concatenated matrix. Repeat n_perm times.
(c) Let λ_i^{perm} be the $100(1 - \alpha)$ percentile among the j'th singular values after permutation.
(d) Choose r to be the largest integer such that $\forall j \leq r, \lambda_j > \lambda_j^{\text{perm}}$.

Figure B.1: Rank selection for r. From [18]



(2) To estimate r_j, with n_perm permutations and significance level α:

(a) Let λ_j be the j'th singular value of X_i.
(b) Permute the columns separately within each row of X_j, and calculate the singular values of the permuted matrix. Repeat n_perm times.
(c) Let λ_j^{perm} be the $100(1 - \alpha)$ percentile among the j'th singular values after permutation.
(d) Choose r_i to be the largest integer such that $\forall j \leq r_i, \lambda_j > \lambda_j^{\text{perm}}$.

Figure B.2: Rank selection for r_j. From [18]

Appendix C

aJIVE example

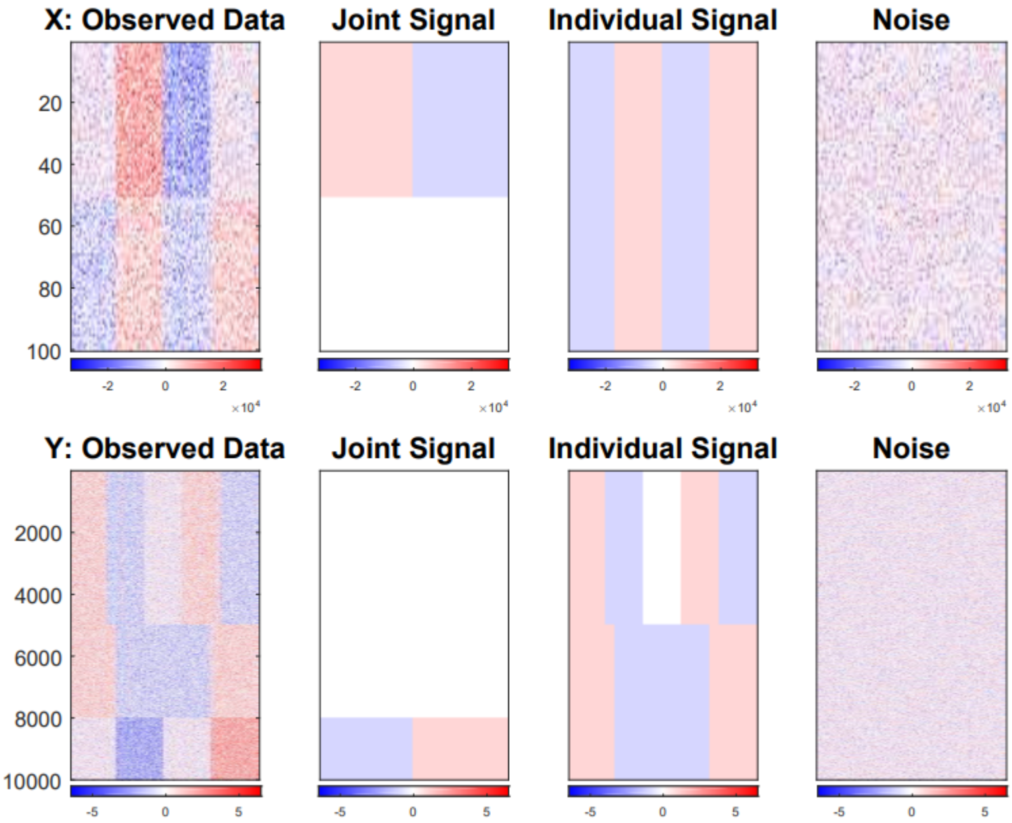


Figure C.1: Dataset constructed from a joint and individual component. From [20]

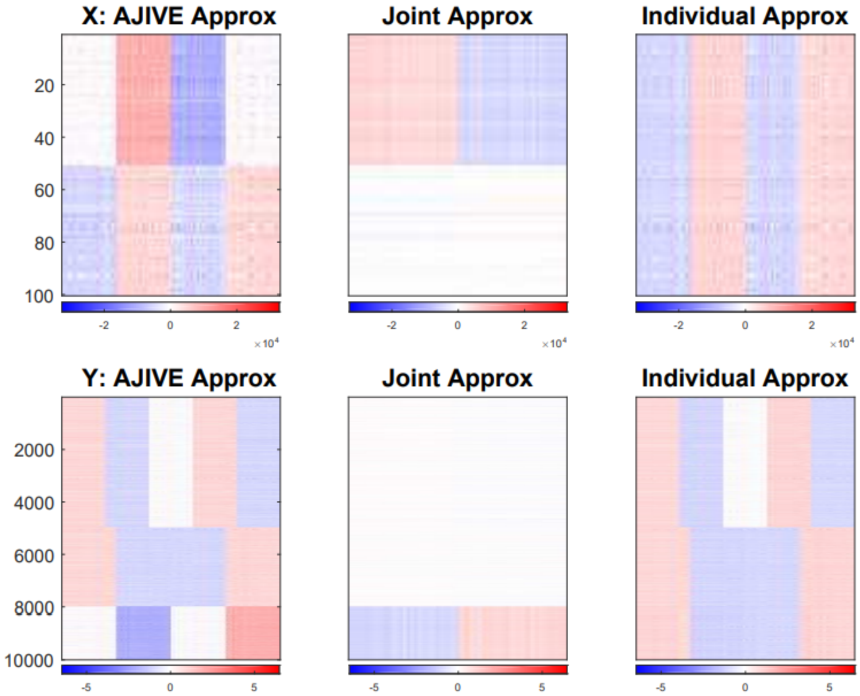


Figure C.2: aJIVE decomposition of data constructed in figure C.1. From [20]

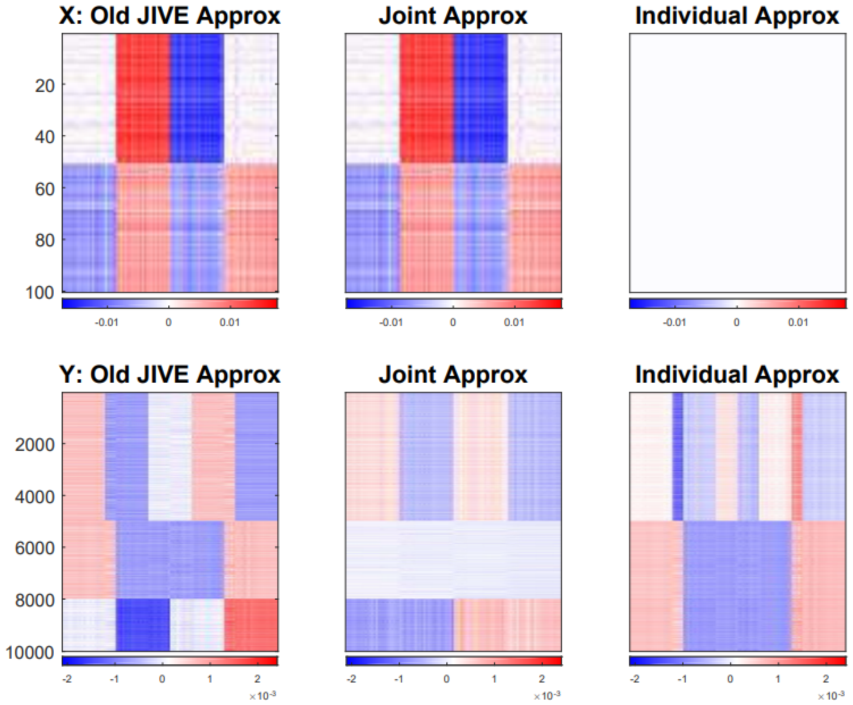


Figure C.3: JIVE decomposition of data constructed in figure C.1. From [20]

Appendix D

Volcano plots of PANCAN colon data with individual selections highlighted

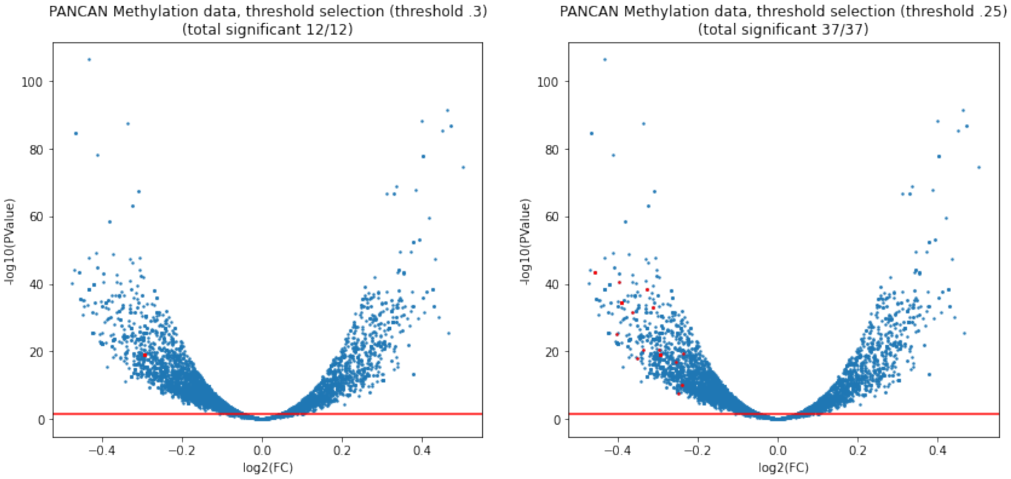


Figure D.1: Volcano plot for the thresholding selection of 100 transcription start sites.

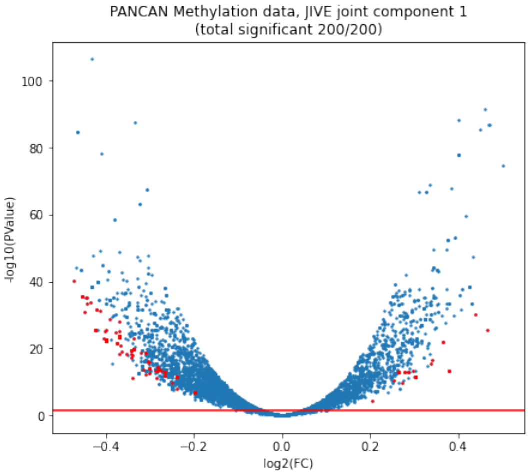


Figure D.2: Volcano plot for the JIVE joint selection of 100 transcription start sites.

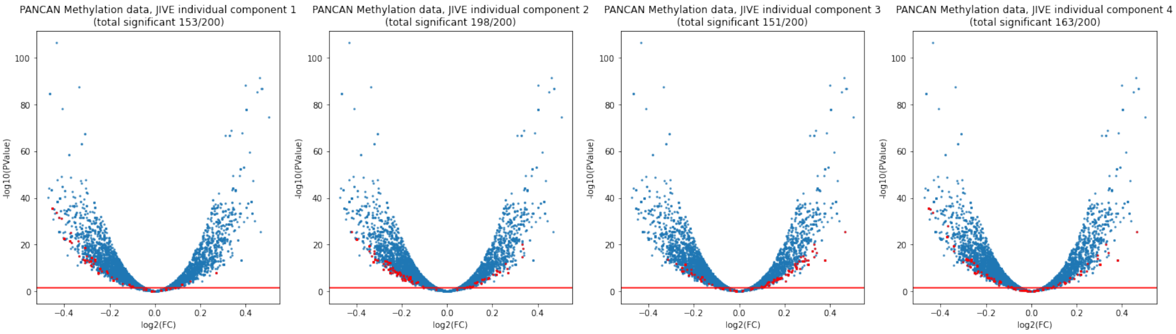


Figure D.3: Volcano plot for the JIVE individual selection of 100 transcription start sites.

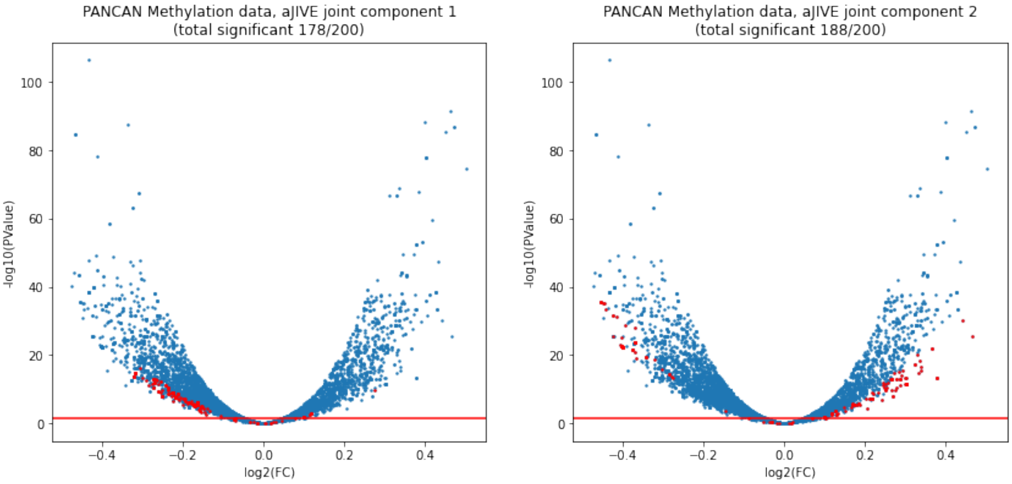


Figure D.4: Volcano plot for the aJIVE joint of 100 transcription start sites.

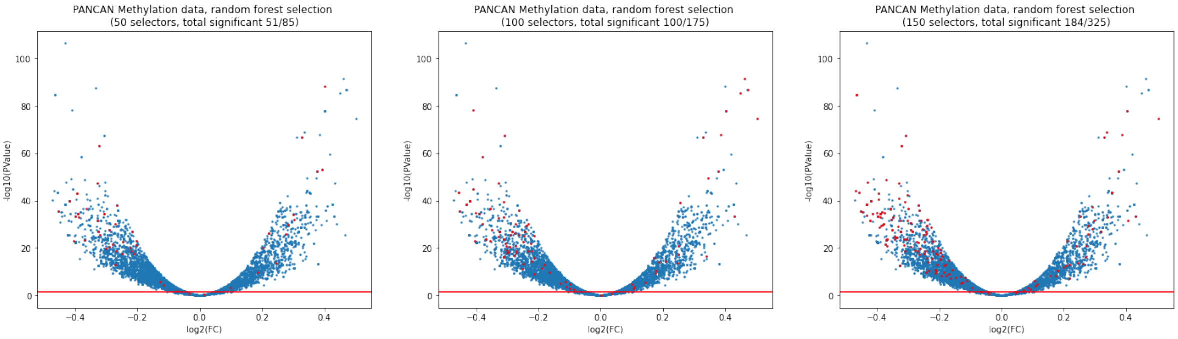


Figure D.5: Volcano plot for the random forest selection of 100 transcription start sites.

Appendix E

Hyperparameters for the Random Forest Classifier

Parameter	**Setting**
criterion	gini
max_depth	None
min_samples_split	2
min_samples_leaf	1
min_weight_fraction_leaf	0.0
max_features	auto
max_leaf_nodes	None
min_impurity_decrease	0.0
bootstrap	True
oob_score	False

TABLE E.1: Hyperparamaters for *scikit-learn*'s *RandomForestClassifier*

Bibliography

[1] F. Ciardiello, D. Arnold, P. Casali, *et al.*, “Delivering precision medicine in oncology today and in futurethe promise and challenges of personalised cancer medicine: A position paper by the european society for medical oncology (esmo),” *Annals of Oncology*, vol. 25, no. 9, pp. 16731678, 2014. DOI: `10.1093/annonc/mdu217`.

[2] J. Gauthier, A. T. Vincent, S. J. Charette, and N. Derome, “A brief history of bioinformatics,” *Briefings in Bioinformatics*, vol. 20, no. 6, pp. 19811996, 2018. DOI: `10.1093/bib/bby063`.

[3] A. Singh, C. Shannon, B. Gautier, *et al.*, “DIABLO: from multi-omics assays to biomarker discovery, an integrative approach,” *bioRxiv*, p. 067 611, 2016. DOI: `10.1101/067611`.

[4] M. Hegemann, A. Stenzl, J. Bedke, K. N. Chi, P. C. Black, and T. Todenhöfer, *Liquid biopsy: ready to guide therapy in advanced prostate cancer?* Dec. 2016. DOI: `10.1111/bju.13586`.

[5] J. C. Wan, C. Massie, J. Garcia-Corbacho, *et al.*, *Liquid biopsies come of age: Towards implementation of circulating tumour DNA*, Apr. 2017. DOI: `10.1038/nrc.2017.7`.

[6] S. Bach, N. R. Sluiter, J. J. Beagan, *et al.*, *Circulating tumor DNA analysis: Clinical implications for colorectal cancer patients. A systematic review*, Sep. 2019. DOI: `10.1093/jncics/pkz042`.

[7] L. A. Torre, F. Bray, R. L. Siegel, J. Ferlay, J. Lortet-Tieulent, and A. Jemal, “Global cancer statistics, 2012,” *CA: A Cancer Journal for Clinicians*, vol. 65, no. 2, pp. 87108, 2015. DOI: `10.3322/caac.21262`.

[8] J. N. Weinstein, E. A. Collisson, G. B. Mills, *et al.*, “The cancer genome atlas pan-cancer analysis project,” *Nature Genetics*, vol. 45, no. 10, pp. 11131120, 2013. DOI: `10.1038/ng.2764`.

[9] M. J. Goldman, B. Craft, M. Hastie, *et al.*, “Visualizing and interpreting cancer genomics data via the xena platform,” *Nature Biotechnology*, vol. 38, no. 6, pp. 675678, 2020. DOI: `10.1038/s41587-020-0546-8`.

[10] M. Bibikova, B. Barnes, C. Tsan, *et al.*, “High density dna methylation array with single cpg site resolution,” *Genomics*, vol. 98, no. 4, pp. 288295, 2011. DOI: `10.1016/j.ygeno.2011.07.007`.

[11] G. Hannum, J. Guinney, L. Zhao, *et al.*, “Genome-wide methylation profiles reveal quantitative views of human aging rates,” *Molecular Cell*, vol. 49, no. 2, pp. 359367, 2013. DOI: `10.1016/j.molcel.2012.10.016`.

[12] A. R. Harris, D. Nagy-Szakal, N. Pedersen, *et al.*, “Genome-wide peripheral blood leukocyte dna methylation microarrays identified a single association with inflammatory bowel diseases,” *Inflammatory Bowel Diseases*, vol. 18, no. 12, pp. 23342341, 2012. DOI: `10.1002/ibd.22956`.

[13] J. Moss, J. Magenheim, D. Neiman, *et al.*, “Comprehensive human cell-type methylation atlas reveals origins of circulating cell-free dna in health and disease,” *Nature Communications*, vol. 9, no. 1, 2018. DOI: 10.1038/s41467-018-07466-6.

[14] R. Boers, J. Boers, B. de Hoon, *et al.*, “Genome-wide dna methylation profiling using the methylation-dependent restriction enzyme lpnpi,” *Genome Research*, vol. 28, no. 1, pp. 8899, 2017. DOI: 10.1101/gr.222885.117.

[15] K. L. Howe, P. Achuthan, J. Allen, *et al.*, “Ensembl 2021,” *Nucleic Acids Research*, vol. 49, no. D1, 2020. DOI: 10.1093/nar/gkaa942.

[16] P. Du, X. Zhang, C.-C. Huang, *et al.*, “Comparison of beta-value and m-value methods for quantifying methylation levels by microarray analysis,” *BMC Bioinformatics*, vol. 11, no. 1, 2010. DOI: 10.1186/1471-2105-11-587.

[17] M. D. Robinson and A. Oshlack, “A scaling normalization method for differential expression analysis of rna-seq data,” *Genome Biology*, vol. 11, no. 3, 2010. DOI: 10.1186/gb-2010-11-3-r25.

[18] E. F. Lock, K. A. Hoadley, J. S. Marron, and A. B. Nobel, “Joint and individual variation explained (jive) for integrated analysis of multiple data types,” *The Annals of Applied Statistics*, vol. 7, no. 1, 2013. DOI: 10.1214/12-aoas597.

[19] M. J. OConnell and E. F. Lock, “R.jive for exploration of multi-source molecular data,” *Bioinformatics*, vol. 32, no. 18, pp. 28772879, 2016. DOI: 10.1093/bioinformatics/btw324.

[20] Q. Feng, M. Jiang, J. Hannig, and J. Marron, “Angle-based joint and individual variation explained,” *Journal of Multivariate Analysis*, vol. 166, pp. 241265, 2018. DOI: 10.1016/j.jmva.2018.03.008.

[21] N.-Y. Cho, J.-W. Park, X. Wen, *et al.*, “Blood-based detection of colorectal cancer using cancer-specific dna methylation markers,” *Diagnostics*, vol. 11, no. 1, p. 51, 2020. DOI: 10.3390/diagnostics11010051.

[22] C. P. Lange, M. Campan, T. Hinoue, *et al.*, “Genome-scale discovery of dna-methylation biomarkers for blood-based detection of colorectal cancer,” *PLoS ONE*, vol. 7, no. 11, 2012. DOI: 10.1371/journal.pone.0050266.

[23] F. Pedregosa, G. Varoquaux, A. Gramfort, *et al.*, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[24] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, “Edger: A bioconductor package for differential expression analysis of digital gene expression data,” *Bioinformatics*, vol. 26, no. 1, pp. 139140, 2009. DOI: 10.1093/bioinformatics/btp616.

[25] M. D. Robinson and G. K. Smyth, “Small-sample estimation of negative binomial dispersion, with applications to sage data,” *Biostatistics*, vol. 9, no. 2, pp. 321332, 2007. DOI: 10.1093/biostatistics/kxm030.

[26] M. Alhamdoosh, C. W. Law, L. Tian, J. M. Sheridan, M. Ng, and M. E. Ritchie, “Easy and efficient ensemble gene set testing with egsea,” *F1000Research*, vol. 6, p. 2010, 2017. DOI: 10.12688/f1000research.12544.1.

[27] J. Moss, J. Magenheim, D. Neiman, *et al.*, “Comprehensive human cell-type methylation atlas reveals origins of circulating cell-free dna in health and disease,” *Nature Communications*, vol. 9, no. 1, 2018. DOI: 10.1038/s41467-018-07466-6.

[28] N. A. O'Leary, M. W. Wright, J. R. Brister, *et al.*, "Reference sequence (refseq) database at ncbi: Current status, taxonomic expansion, and functional annotation," *Nucleic Acids Research*, vol. 44, no. D1, 2015. DOI: 10.1093/nar/gkv1189.

[29] F. Huang, J. Chen, Z. Wang, R. Lan, L. Fu, and L. Zhang, "-catenin promotes tumorigenesis and metastasis of lung adenocarcinoma," *Oncology Reports*, 2017. DOI: 10.3892/or.2017.6140.

[30] Q.-Q. Zhang, D.-l. Zhou, Y. Lei, *et al.*, "Slit2/robo1 signaling promotes intestinal tumorigenesis through src-mediated activation of the wnt/-catenin pathway," *Oncotarget*, vol. 6, no. 5, pp. 31233135, 2014. DOI: 10.18632/oncotarget.3060.

[31] A. Dallol, D. Morton, E. Maher, and F. Latif, "Slit2 axon guidance molecule is frequently inactivated in colorectal cancer and suppresses growth of colorectal carcinoma cells," *American Association for Cancer Research*, vol. 63, no. 5, pp. 10541058, 2003.

[32] F. J. Carmona, D. Azuara, A. Berenguer-Llergo, *et al.*, "Dna methylation biomarkers for noninvasive diagnosis of colorectal cancer," *Cancer Prevention Research*, vol. 6, no. 7, pp. 656665, 2013. DOI: 10.1158/1940-6207.capr-12-0501.