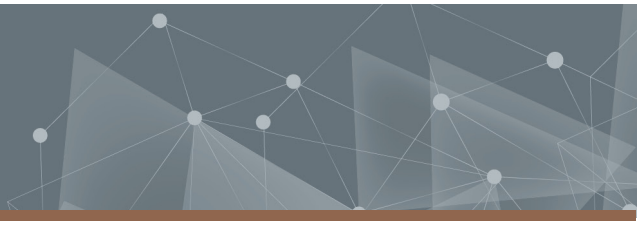
CHALMERS
UNIVERSITY OF TECHNOLOGY



Cross-tissue variance analysis of gene sets

Using statistical methods to quantitatively analyse the variance contribution in gene set enrichment scores

Mauritz Kööhler, MPENM
Oskar Thune, MPCAS

DEPARTMENT OF MATHEMATICAL SCIENCES

CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2023
www.chalmers.se

Master's thesis 2023

Cross-tissue variance analysis of gene signatures

Using statistical methods to quantitatively analyse the variance contribution in gene set enrichment scores

Mauritz Kööhler
Oskar Thune

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Mathematical Sciences
Chalmers University of Technology
Gothenburg, Sweden 2023

Cross-tissue variance analysis of gene signatures
Using statistical methods to quantitatively analyse the variance contribution in gene set enrichment scores
MAURITZ KÖÖHLER
OSKAR THUNE

Supervisor: Bastian Angermann, AstraZeneca
Examiner: Erik Kristiansen, Mathematical Sciences

Master's Thesis 2023
Department of Mathematical Sciences
Chalmers University of Technology
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Typeset in LaTeX
Printed by Chalmers Reproservice
Gothenburg, Sweden 2023

Cross-tissue variance analysis of gene signatures
Using statistical methods to quantitatively analyse the variance contribution in gene set enrichment scores
MAURITZ KÖÖHLER
OSKAR THUNE
Department of Mathematical Sciences
Chalmers University of Technology

Abstract

Gene set enrichment is used to investigate the differences between gene expression for genetic pathways in transcriptomic data. Gene set scoring methods like GSVA and singscore are used in gene set enrichment analysis to assess the enrichment of genes of interest, called gene sets. GSVA and singscore produces a score of how expressed a gene set is in relationship with a reference expression, a reference that is not always accessible.

In this work we apply variance decomposition to investigate the use of singscore and GSVA to create a baseline for RNA-seq data that lacks control samples and apply a VAE for prediction of gene set scores across tissues. To this end, variance decomposition was done on GTEx to assess the dataset's use as a baseline, and a VAE was trained on GTEx with the aim of predicting gene set scores across tissues.

Our results show that there is a limited use of using a reference dataset as a basis for RNA-seq data. The results are not conclusive enough to warrant usage in applications with the precision needed in pharmaceutical research. The VAE based prediction shows lacklustre results in predicting expression over tissues, and other machine learning methods should be investigated for this application.

Keywords: RNA-seq, GSVA, Transcriptomics, Bioinformatics, Variational autoencoder, GTEx, Variance decomposition.

Acknowledgements

We would like to thank our supervisor Bastian Angermann as well as Sue Monkley who have guided the work on this thesis, formulated the scope and always been happy to answer questions regarding both technical aspects and biology. We also want to give thanks to AstraZeneca and Daniel Muthas for the opportunity and the support we have received during our work. It has been amazing to have this chance to experience working in this environment.

Lastly, a big thank you to the entire Data science team, all the lunch talks about your work and the current internet trends have been lovely!

Mauritz Kööhler, Gothenburg, August 2023
Oskar Thune, Gothenburg, August 2023

List of Acronyms

Below is the list of acronyms that have been used throughout this thesis listed in alphabetical order:

AIC	Akaike Information Criterion
ANOVA	Analysis of Variance
CLT	Central Limit Theorem
GSEA	Gene Set Enrichment Analysis
GSVA	Gene Set Variation Analysis
GTEx	Genotype-Tissue Expression
IBD	Inflammatory Bowel Disease
MLP	Multi Layer Perceptron
MSE	Mean Square Error
MSigDB	Molecular Signatures Database
RNA-seq	RNA sequencing
SMD	Standardised Mean Difference
TPM	Transcripts Per Million
VAE	Variational Autoencoder

Contents

List of Figures

List of Tables

1
Introduction

This chapter will give a background to the project and give motivation for the work that has been pursued. Furthermore, aim, objective and demarcations will be introduced such that the goal of the project is clear. Finally, a thesis layout of the entire report will be introduced.

1.1 Background

Historically, diseases have been described by symptoms. A set of symptoms would suggest a person has a disease and the goal in producing a drug is to reduce the symptoms. In precision medicine, the aim is to understand the underlying cause of a disease and create drugs targeting specific biological pathways, a set of genes working in sequence. By analysing how present genes in a biological pathway are in a diseased population against a control, connection to diseases can be made. Because of the increase in available transcriptomic data [1, 2, 3], methods like Gene Set Enrichment Analysis (GSEA) [4] are used to investigate the impact a group of genes have related to regulatory pathways. By using gene sets rather than individual genes, the analysis can be done on pathways that are known to be connected to a given condition, or, help discover new pathways to a given condition. Gene set enrichment analysis gives results that are more trustworthy and reliable when coming to conclusions about the underlying biology in comparison to analysis on individual genes [5].

The research in this area has led to collections of gene sets that have been established to have a connection to some biological pathway [2], most notably The Molecular Signatures Database (MSigDB) [1] that is maintained by the Broad Institute of MIT and Harvard. MSigDB contains several collections of gene sets from both humans and mice, with a total of 31 322 gene sets on the human genome. Nota bene, gene signature and gene set are equivalent, and this study will favour the use of the word gene set.

Analysis on gene set enrichment analysis-methods always relies on the use of a baseline for what the gene expression could vary around. By comparing the difference in gene expression between a control group, the baseline, and a patient, one can gain information of the state of a patient. However, the data for the baseline is usually not available. The most apparent example is data sets that lack control subjects. It is therefore of interest to obtain these baselines. This study wants to investigate if one can construct a general baseline generated from two gene set enrichment scoring methods called singscore and GSVA. Both methods give a score for each gene set and sample.

Defining a baseline, however, is not an easy task. Genes, and biology in general, are complex and highly regulated [6]. Various factors can contribute to gene expressions that are seen in the transcriptomic data that exist today. These can be divided into biological, technical and human factors. Examples of factors are differences in gene expression between and within people, what methods were used to sequence the data and high gene expression is not equal to high protein expression, to name a few. Therefore, the meaning of baseline, or null, in this report will not be universal but tied to a given data set to minimise the effect of the mentioned factors. The distributions will be conditioned on gene signature and chosen meta factors e.g., sex and gender, in order to increase the biological significance of observations.

1.2 Aim & objectives

The report aims to design a method to derive an empirical null distribution of gene signature scores. This will be done for two different scoring methods, singscore and GSVA. Analysis of the produced method will be executed to verify how well the null distributions can be used as a replacement for control samples in a given data set. This will be done by investigating the relationship between the proposed null distribution with another data set. Additionally, the report aims to investigate the predictive power of a gene signature to create an empirical null distribution using a variational autoencoder (VAE). The VAE aims to investigate if a different tissue can predict the null distribution for another tissue of interest. Therefore, at the end of the project, the report aims to answer the following questions:

- Which factors describe a gene set's variability?
- Can singscore and GSVA be used to compare between different studies?
- Can a VAE model the relationship for gene sets between tissues?

1.3 Demarcations

To be able to answer the research questions stated in section 1.2, two main data sets will be investigated. Firstly, the Genotype-Tissue Expression (GTEx) project [7] will be used as reference data and the SPARC IBD cohort [8] will be used to compare GTEx with. GTEx will be compared to the SPARC IBD cohort and any conclusions drawn from the analysis in the report will be limited to the tissues and range of SPARC IBD cohort.

1.4 Thesis outline

The layout of the report aims to give a clear path to answer the research question stated in section 1.2.

The theory chapter gives an understanding of the methods that have been applied in the thesis. The method chapter covers the necessary steps needed to replicate the steps

done in this thesis and gives background to choices made in the work. The results chapter covers the findings made and the conclusion covers analysis of the results and recommendations for further work in this area.

2
Theory

The theory chapter focuses on giving an understanding of the methods that have been applied in the thesis. This chapter is aimed at individuals who have a basic understanding of the terminology and theory in statistics and machine learning, equivalent to an introductory course in each subject at university. The subchapters are recommended to read in order for ease of understanding.

2.1 RNA-seq & normalisation

Transcriptomic analysis, as done in this report, investigate a snapshot of the cells current RNA abundance [9, 6]. The flow of biological information is governed by the central dogma of molecular biology [10, 6]. The dogma states that information is stored in DNA, DNA can transport information into RNA and RNA can transport information to proteins. This is a helpful simplification of the complexity in molecular biology. By the central dogma, RNA is a transition state between the genomic data, stored in the DNA, to the more operational data in proteins. In theory, all cells have access to all the genomic data in the DNA, but different cells express genes differently, e.g., cells in the heart have a different gene expression than cells in the skin. This is true to an extent, but a distinction should be made on why it is not completely true. All of our DNA exists in all cells. However, one aspect that distinguishes differentiated, i.e. have a specified function, cells from each other are what genes are accessible for them to translate to RNA. Our DNA is a complex and long molecule that are stabilised by a protein called chromatin and depending on how the chromatin binds to the DNA, the accessibility to read differs between cells. The transcriptomic data from a cell, therefore, describes which genes that are expressed in each cell at a specific time and state, but also which genes that are accessible for that cell. Furthermore, the same cell can have a different expression at another time, in another state. A state could be if a cell is healthy or sick, dividing or not and dying or not.

The data sets investigated in this report was produced utilising a method called RNA-seq [9, 11]. Figure 2.1 summarise RNA-seq which start with an RNA-sample from multiple cells. This sample gets sequenced into millions of short strings of RNA, called reads, at random positions in the input RNA. These reads are then mapped onto a reference genome and the counts of aligned reads with a gene in the genome generates the abundance of that gene. In conclusion, the RNA-sample, which is a snapshot of how much genes are expressed, is read in multiple short strings and mapped back onto the genes in the DNA.

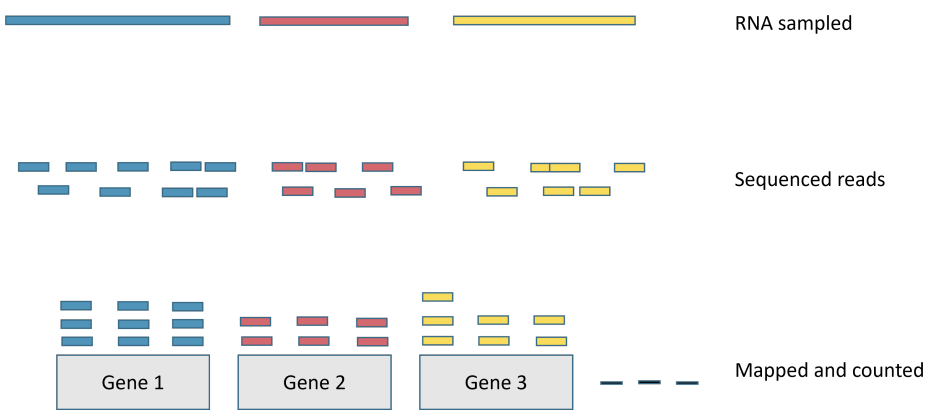


Figure 2.1: *A schematic view on RNA-seq where an RNA-sample is being mapped onto a reference genome.*

To enable comparison between samples, different normalisation methods, such as TPM [12], are utilised depending on the transcriptomic data at hand. TPM, transcripts per million, creates relative RNA abundant values within a sample by normalising over the length of the raw counts from, for example, RNA-seq. From figure 2.1, one can note that the first blue RNA strand is longer than the others. As a results, one could assume that this RNA should have more reads than the others since the reads are taken from random positions of the sample. TPM takes this into account and the normalisation over the length of the RNA results in the sum of all TPM-values in a sample being the same across all samples. TMP per gene is given by equation 2.1

$$TPM_i = \frac{q_i/l_i}{\sum_j q_j/l_j} \cdot 10^6 \tag{2.1}$$

where q_i is the raw counts from, for example, RNA-seq to a gene. l_i is length of the RNA that q_i comes from and $\sum_j q_j/l_j$ is the sum of normalised raw counts to RNA length over all samples.

2.2 Molecular Signatures Database

Gene sets are a collection of genes that are connected to some biological pathway [2]. The Broad Institute of MIT and Harvard has categorised 31 322 human gene sets into 9 different collections into a database called the Molecular Signatures Database (MSigDB) [1]. Each collection gathers gene sets with different biological connections, e.g., positional, immunological, computational. One of these collections, with specific importance to our study, is the hallmark collection where gene sets have been curated by the combination of an automatic approach together with experts [13]. Each hallmark gene set is derived from multiple other gene sets with an overlapping biological function or process, producing a coherent expression for the gene sets [13].

2.3 Gene Set Variation Analysis

Gene Set Variation Analysis (GSVA) is an unsupervised and nonparametric analytical method used to estimate variation of gene set expression in RNA-sequencing data [14]. GSVA enables analysis to be made focused on gene sets rather than individual genes, with the result that genetic pathways can more readily be analysed [5]. GSVA is commonly applied in research to estimate variance of pathway activity [15].

GSVA analysis of RNA-seq data requires two inputs, a normalised expression matrix $X = \{x_{ij}\}_{n\times p}$ for n samples and p genes; and a collection of gene sets $\Gamma = \{\gamma_1, \ldots, \gamma_m\}$. Here x_{ij} denotes the expression of gene i in sample j, and $|\gamma_k|$ is used to denote the number of genes in gene set k. GSVA initiates by evaluating the expression level of a gene i in relation to all samples of that gene. To that end, a Gaussian kernel

$$\hat{F}_{h_i}(x_{ij}) = \frac{1}{n}\sum_{k=1}^{n}\int_{-\infty}^{\frac{xij-xik}{h_i}} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\,\mathrm{d}t \tag{2.2}$$

is used and a bandwidth parameter $h_i = s_i/4$ where s_i is the sample standard deviation for the gene. The expression statistic is thereafter converted into ranks $\hat{F}_{h_i}(x_{(i)j})$ for each sample j and then normalised $r_{ij} = |\frac{p}{2} - \hat{F}_{h_i}(x_{(i)j})|$ in order to center the ranks on 0. The enrichment score is then assessed by a Kolmogorov-Smirnov statistic

$$\nu_{jk}\left(\ell\right) = \frac{\sum_{i=1}^{\ell} |r_{ij}| I\Big(g_{(i)} \in \gamma_k\Big)}{\sum_{i=1}^{p} |r_{ij}| I\Big(g_{(i)} \in \gamma_k\Big)} - \frac{\sum_{i=1}^{\ell} I\Big(g_{(i)} \notin \gamma_k\Big)}{p - |\gamma_k|} \tag{2.3}$$

with $I(g(i) \in \gamma_k)$ being the indicator function for gene i being part of gene set k. Thereafter the enrichment score, in equation 2.4, is calculated by taking the difference of the largest deviations in both the positive and negative directions from zero for equation 2.3.

$$ES_{jk}^{\text{diff}} = \left|ES_{jk}^{+} - ES_{jk}^{-}\right| = \max_{\ell=1}^{p}\Big(0, \nu_{jk}(\ell)\Big) - \min_{\ell=1}^{p}\Big(0, \nu_{jk}(\ell)\Big) \tag{2.4}$$

2.4 Singscore

Singscore is a rank-based single sample scoring method used to analyse RNA-sequencing data [16]. This differentiates singscore from GSVA as the score of a gene set for a sample will only depend on information from that sample, by ranking gene expressions correlated to a given gene set. This results in more stable scores for data sets where the number of samples is small [16] and more flexibility in how the scoring is done, as the score for samples are independent of each other.

Given k gene set γ_k and a normalised expression sample $X = \{x_{ij}\}_{n\times p}$ for n samples and p genes, the expressions x_{ij} are converted to ranks R_{ij} based on the expression level

of the genes within each sample i. A mean rank score per gene set γ_k is calculated from equation 2.5

$$S_{ki} = \frac{\sum_g R^g_{ki}}{N_{ki}} \tag{2.5}$$

where R^g_{ki} is the rank of the g^{th} gene in gene set k and sample i. N_{ki} is the number of genes in gene set k in sample i, considering missing genes in the expression data compared to gene set k. Furthermore, the mean rank score is then normalised, with relation to the theoretical maximum and minimum values, and centered around 0, as seen in equation 2.6.

$$\bar{S}_{ki} = \frac{\left(S_{ki} - S_{min,i}\right)}{S_{max,i} - S_{min,i}} \tag{2.6}$$

In equation 2.6, the maximum and minimum values are derived from an arithmetic sum of n numbers, starting at a with constant difference d, calculated as $(n/2)(2a+(n-1)d$. Setting $a = 1$(minimum), $a = (N_{total} - N_{ki})$(maximum), $d = 1$ and $n = N_{ki}$, as well as taking the average, one obtains:

$$S_{min,i} = \frac{N_{ki} + 1}{2} \tag{2.7}$$

$$S_{max,i} = \frac{2\,N_{total} - N_{ki} + 1}{2} \tag{2.8}$$

The theoretical scores from equation 2.6 is $-0,5$ to $0,5$ when only investigating gene sets that have one direction, either up regulated or down regulated.

2.5 Standardised Mean Difference

Standardised Mean Difference (SMD) is a unit free measure on how separated, in pooled standard deviation, two group means are from each other [17]. SMD is given by equation 2.9.

$$SMD = \frac{|\mu_1 - \mu_2|}{\sqrt{(s_1^2 + s_2^2)/2}} \tag{2.9}$$

where μ_1 and μ_2 are the means of the two distributions and the term $\sqrt{(s_1^2 + s_2^2)/2}$ is the pooled standard deviation of the two distributions. This study quantitatively distinguishes two distributions as significantly different if the SMD value is larger than 2. This measure only investigates how close the means are and does not take the tails of the two distributions into consideration.

2.6 Fisher's exact test

The Fisher exact test is, compared to standardised mean difference, investigating the difference between two distributions over the whole interval. Fisher's test is often used

when the restrictions of sample size in the χ^2-test are violated [18] as Fisher's test does not have any restrictions in that regard. Fisher's exact test exactly computes [19] the probability of finding an outcome as, or more, extreme as the one present, using counts. Using counts turns the problem into a combinatorial one which gives discrete outcomes. Given a discrete outcome, the only outcome with a lower p-value than the significance level of $\alpha = 0,05$ might be very close or very far from α, which makes the test conservative. The test was first proposed by Fisher in 1935 [19] when a lady claimed she could distinguish between if tea with milk had the tea or milk added first. Fisher constructed a random experiment for the lady to participate in and calculated how likely it was that the guesses from the lady was made by pure chance. Fisher calculated all possible outcomes from the experiment of the 8 cups, 4 with milk added first and 4 vice versa. If the probability of the outcome from the experiment is unlikely enough, one would reject the null hypothesis of the lady's choice being made by pure chance.

Fisher's exact test can be computed for continuous data by converting the data to counts over given intervals. Comparing the distribution of two data sets with continuous data, the amount of data points should be equal between the two data sets on the same interval if the two data sets are equally distributed. Therefore, the reference data gets divided into intervals such that equally many data points fall into each interval. Following this, the same intervals are applied on the comparing data and the expected count of the data is calculated, given by equation 2.10

$$E_{compare_i} = \frac{O_{ref_i}}{\sum_i O_{ref}} \sum_i O_{compare_i} \tag{2.10}$$

where O_{ref_i} is the observed counts for each interval for the reference data, $\sum_i O_{compare_i}$ is the sum over the observed counts for the compare data. Fisher's exact test is then applied on the observed and expected counts of the compared data set.

2.7 Variational Autoencoders

Variational Autoencoder (VAE) is an artificial neural network architecture used for unsupervised learning and generative modelling [20]. An overview of the architecture of a VAE can be seen in figure 2.2. VAEs consist of an encoder that maps input data **X** to a latent space **Z**, and a decoder that maps the latent space back to the original input space creating a recreated data **X'**.

The core aspect of VAEs lies in their probabilistic formulation. Instead of encoding the input into a fixed point in the latent space as an autoencoder would, VAEs model a set of distributions that the latent variables are sampled from. Specifically, the encoder generates a probability distribution of the latent space, modelled as a multivariate normal distribution with a diagonal covariance matrix and the latent variables are sampled from this distribution. This is achieved by parameterising the distribution with mean and variance vectors $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ obtained from the encoder's output giving the latent space distribution $\mathbf{Z} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma})$.

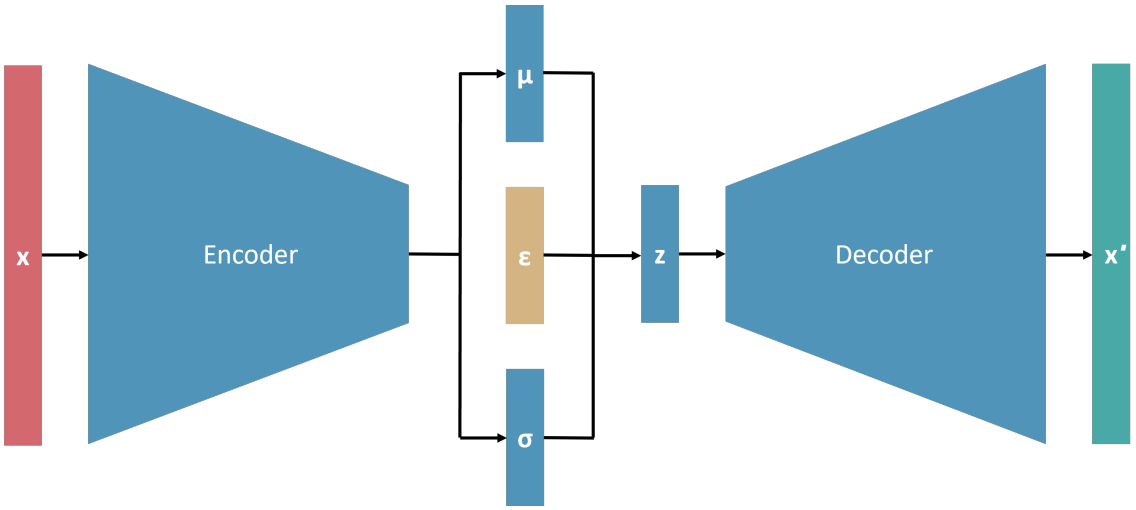


Figure 2.2: *Overview of general VAE architecture*

This random sampling introduces an issue in the backpropagation. Since the latent variables $\boldsymbol{Z}$ are sampled the backpropagation can not be performed as the sampling can not be differentiated [20]. To solve this, the distribution is parameterised with an auxiliary variable $\epsilon \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ and expressing the latent variables and enabling backpropagation to be implemented in the model according to standard procedure. The distributions for the latent variables can then be rewritten as:

$$\mathbf{Z} = \boldsymbol{\mu} + \boldsymbol{\sigma}\boldsymbol{\epsilon} \tag{2.11}$$

VAEs optimise a loss function during training that comprises of two terms [20]: the reconstruction and the regularisation. The reconstruction loss is based on the difference in the input and output of the model, rewarding accuracy in the reconstruction of the input. The second term is a regularisation term that promotes a smooth and continuous latent space.

This regularisation loss is commonly the Kullback-Leibler divergence

$$D_{KL}(P||Q) = \int_{-\infty}^{\infty} p(x) \log\left(\frac{p(x)}{q(x)}\right) dx \tag{2.12}$$

where P and Q in the setting of a VAE are posterior distributions of the latent space and a chosen prior distribution, respectively. The Kullback-Leibler divergance is a measure of the difference of two distributions. In equation 2.12, $p(x)$ and $q(x)$ denotes the probability density functions of P and Q.

If the prior distribution is chosen as the multivariate standard normal distribution, equation 2.12 can be simplified to

$$D_{KL}(N(\boldsymbol{\mu}, \boldsymbol{\sigma})||N(0, I)) = -\frac{1}{2}\sum_{i} 1 + \log(\sigma_i^2) - \mu_i^2 - \sigma_i^2 \tag{2.13}$$

allowing for simplified computation [20].

3
Methods

The method section includes the necessary steps of data handling, preprocessing, scoring and further analysis where some methods are connected with the theory stated in section 2. In appendix A you can find a list of the versions used for programming languages and packages.

3.1 Data sets

Two different data sets were investigated in this study to help answer the objectives of the project stated in section 1.2. The Genotype-Tissue Expression (GTEx) project [7, 21], from Broad Institute of MIT and Harvard, was used as reference data due to its size and thorough meta data. GTEx was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. GTEx is composed of 17 384 samples from 948 donors across 54 tissues which covers a large amount of biologal information. The samples come from post-mortem donors with different causes of death. Due to the cross-tissue samples, GTEx is ideal as input data to a machine learning model that can compare inter-tissue relations. GTEx meta data also enables the use of ANOVA for variance analysis to see which factors have the largest contribution to the variance. To compare with GTEx, the SPARC IBD cohort [8] was analysed. The SPARC IBD cohort hereby mentioned as IBD Plexus was put in relationship with the results from the analysis of GTEx for verification of GTEx's use as a reference data set.

Links to the datasets and how to access them can be seen in appendix C

3.2 Gene & gene set preprocessing

Both of the studies used in this paper investigates a large number of genes, where a considerable amount are lowly expressed or not expressed at all. Including these in the analysis would negatively affect the quality of the results for both singscore and GSVA. The effect in singscore comes from the normalised mean ranks (equation 2.6), where all genes that have no expression, all get assigned rank 1. Naturally, this affect the theoretical min and max score (equation 2.7) used in normalising the ranks. For GSVA, the amount of lowly or none expressed genes, skew the estimated kernel density used in GSVA (equation 2.2), and therefore affect the results.

A cut off for which genes to include in the analysis is therefore paramount. Establishing a good cut off for gene set enrichment for multiple samples is not straightforward and interpretation of the results depends on the method used to calculate the cut off. This study aimed to investigate the mean rank for the genes in GTEx and decide on the cut off based on that investigation. The genes in GTEx were ranked within each sample and the mean rank for each gene was calculated and summarised in figure 3.1. The mean of the ranks should be close to normally distributed due to the central limit theorem, CLT. CLT states that independent and identically distributed random variable with mean μ and variance σ converges asymptotically to a normal distribution $\mathcal{N}(\mu, \sigma)$ when the samples size goes to infinity. The random variables in this instance are the mean ranks of the genes and are assumed to come from a common distribution and thus the means of the ranks should be asymptotically normally distributed for our samples size. However, due to the large number of non expressed genes, the mean rank is not normally distributed, as one can note from figure 3.1. Setting a cut off to remove all genes with any instance of no expression across all samples would be too conservative of an approach and the study therefore took away genes with a mean rank such that the distribution would be approximately normal. Based on this analysis of GTEx the cut off was set to 0.6722, meaning that 67,22% of the overall most expressed genes were kept, as this matched the above exclusion criteria. Since the exclusion was concluded from the mean rank, the genes that were excluded in the analysis were the ones with the overall lowest expression for a gene, across all samples in GTEx. This implicates that genes in samples from some tissues might be highly expressed but have no expression in other tissues. Another implication is that some genes will have a significant number of samples where that gene has no expression if that gene in general has a high expression across samples.

Both singscore and GSVA relies on assigning a score to a gene set. If the genes in a gene set are not available in the data, no gene set enrichment is possible. Therefore, a pipeline was produced to verify which gene sets in MSigDB have at least one gene in GTEx. The gene sets that were filtered out were not part of any further analysis and IBD Plexus's genes were filtered out from these gene sets.

3.3 Singscore

Singscore was implemented using Bioconductor, an open source collection of packages in R [16]. The singscore function ran with the parameter for unidirectional gene set, a gene set that either up or down regulates the genes included in the gene set.

Singscore was first applied on GTEx and the hallmark gene sets from MSigDB to implement a functioning pipeline for singscore. Following this, all MSigDB gene sets that had at least one gene in GTEx, were computed for GTEx and IBD Plexus. The gene sets that did not have any genes in GTEx were stored and used to filter out IBD Plexus for both singscore and GSVA.

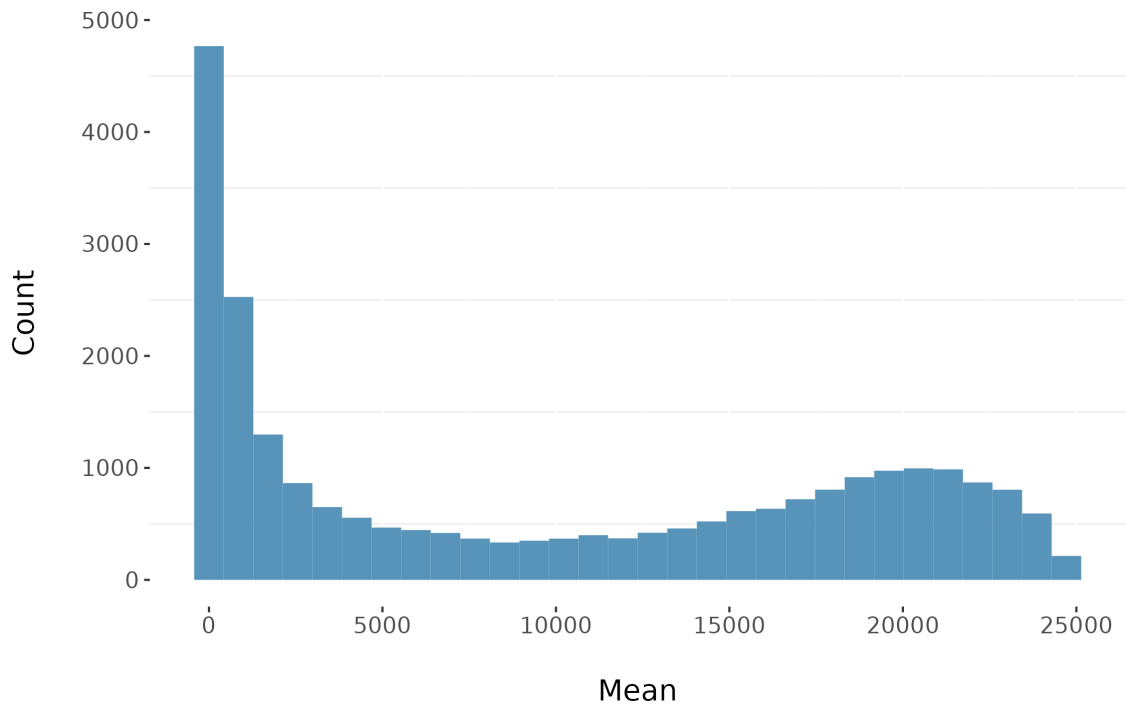


Figure 3.1: *A histogram of the mean rank of genes in GTEx.*

3.4 GSVA

GSVA was implemented using Bioconductor, an open source collection of packages in R that has a implementation of the method [14]. GSVA ran with the gaussian kernel computations.

The GSVA scores for GTEx were computed twice with different configurations. This approach was chosen due to time constraints for the computation. First the scores for the 50 Hallmark gene sets were computed for all samples in GTEx. Second, the scores for all gene sets in MSigDB were computed with the samples separated based on the tissue in order to reduce the computational time. The reduction in time comes from the quadratic scaling of computational time by number of samples in equation 2.2. Separating samples based on tissue was chosen since IBD Plexus contains samples of a few tissues. This separation on tissue for GTEx, should result in minimum introduced bias and give a relevant interpretation.

3.5 Variance decomposition

The obtained scores from both GSVA and singscore were initially investigated visually with the aim to intuitively get a better understanding of the variance behind the scores and investigate the research questions. Following this, to compare the distributions of the score for singscore and GSVA, several statistical methods were deployed. Firstly, to investigate which factors that contribute to the variance of the scores, AIC and ANOVA were applied. Furthermore, to compare distribution between studies, SMD and Fisher's

exact test were applied.

3.5.1 Model selection

To assess which factor contributed to the variance of the score from singscore and GSVA, an ANOVA analysis was performed on GTEx. Which factors to include in the analysis were decided by using AIC and performing an ANOVA analysis on the model with lowest AIC. 3 models were proposed and summarised in table 3.1, based on the meta data available for GTEx. The factors included in the full model were the ones that captured the most biological information, and implicitly have the most impact on the variability. The two smaller models excluded factors with decreasing prior thought of variability impact. Most factors are self-explanatory except for "HardyScale". Hardy scale is a four-point scale of how fast the donor died with the range going from violent and fast death to slow death. Including this in the analysis was to investigate if further analysis should be separated by the different levels of hardy scale in GTEx.

Table 3.1: *The three suggested linear models to use in ANOVA per scoring method.*

Large Model	Middle Model	Small Model
Tissue	Tissue	Tissue
GeneSet	GeneSet	GeneSet
Tissue:GeneSet	Tissue:GeneSet	Tissue:GeneSet
Age	Age	
Sex	Sex	
Sex:GeneSet	Sex:GeneSet	
HardyScale		
GeneSet:HardyScale		

This study uses AIC, Akaike information criterion, for model selection as AIC gives a measure of the quality of a model in comparison with other models given the same data [22]. The benefit of this approach is that AIC gives an easy, qualitative and comparable measure for which model has the highest quality given the same data. AIC is give by equation 3.1

$$AIC = 2k - 2\,ln(\hat{L}) \tag{3.1}$$

where k is the number of model parameters and $\hat{L}$ is the maximised likelihood for the model described by k parameters. A high-quality model produces a low AIC-value because only a small number of parameters was needed to produce a high maximised likelihood. AIC is relative to the data used in the calculations and is restricted to comparing models where the same data was used. Therefore, no comparison will be made between the AIC-values for singscore and GSVA.

3.5.2 Factor contribution

ANOVA, Analysis of Variance, was performed on the model with the lowest AIC-value for each of the scoring methods. The ANOVA method investigates the contribution of

variance per factor and is well suited for the GTEx data. In ANOVA, the independent variables are categorical [23], which is true for the meta data in GTEx, and the response variable are continuous [23], likewise to the scores from singscore and GSVA. The number of combinations of factors for the large model, due to the large amount of expression data in GTEx; the amount of gene sets in MSigDB; and amount of factors in GTEx meta data, surpassed the computational power accessible for the project. Therefore the ANOVA analysis was focused on the hallmark gene sets in MSigDB (section 2.2). Described in section 3.4, GSVA was applied with two different configurations on GTEx: one separated on tissue applying all MSigDB gene sets and one where only the hallmark gene sets were applied using the whole GTEx data. Being able to interpret the ANOVA results in the same fashion for singscore and GSVA, tissue needs to have the same impact, thus, the GSVA configuration using only the hallmark gene sets was applied. Otherwise, the Gaussian kernel in equation 2.2 would normalise within each tissue and cancel out the impact of variance for that factor. For both GSVA and singscore, the score data was combined with the meta data with the factors included in the largest model in table 3.1.

3.5.3 Distribution comparison

Described in section 1.2, this project aims to answer how well singscore and GSVA could be used to compare between studies. This is enabled by quantitatively comparing GTEx against IBD Plexus using SMD and Fisher's exact test. The comparison will leverage that GTEx have a varied sample pool and IBD Plexus have patients in different stages of diagnosis related to the colon and therefore distinguish if GTEx can be used as a control data set and, by implication, if singscore and GSVA can be used to compare between data sets. The stated comparison is enabled when investigating the same tissue in GTEx as in IBD plexus and comparing the same gene sets within those tissues with each other. When comparing IBD Plexus with GTEx, three different combinations of samples were constructed depending on the meta data available in IBD Plexus. The three factors were: position, where the samples where taken; macroscopic appearance, the appearance of the tissue when taking the biopsy for the sample; and diagnosis, what diagnosis the patient had at the time of the sampling. These factors were investigated to see if GTEx have similar distributions depending on the factors used.

From equation 2.9, to calculate the SMD-value for a given gene set in a tissue, the standard deviation and the mean for that selection is needed. Thus, the mean and standard deviation was calculated for the scores from both singscore and GSVA for GTEx and IBD Plexus. The mean and standard deviation was calculated for the combination of tissue and gene set in GTEx and by the three factors stated previously and gene set for IBD Plexus. Next, equation 2.9 was applied on the computed mean and standard deviation and the tissue, gene set combination with a SMD-value larger than 2 (more than 2 pooled standard deviations away) were viewed as significant.

Fisher's exact test is the second, and final, method to analyse the score distributions and demands to manipulate the data into the format described in section 2.6. In this study, GTEx is used as the reference data set and IBD Plexus is used as the compare data set. The number of intervals to divide the data into was chosen as 5, due to the counts in

each interval being too low with more intervals. When comparing GTEx against IBD Plexus, the same separation on the meta factors were done as for SMD so that the outcome of Fisher's exact test can be compared with the outcome of the SMD analysis.

3.6 Variational autoencoder

A Variational autoencoder (VAE) was implemented in Python 3.7.2 with using Tensorflow 2.11 and Keras 2.11. Two models were created, one trained on GSVA results and one on singscore results.

The score data from the two methods was pre-processed before being used in the model. Firstly, in order to make the data easier to work with, the number of gene sets were reduced by choosing the 500 gene sets with the highest variance across all samples. Thereafter, the data was scaled uniformly between 0 and 1. The encoding for the tissues was done as one-hot style vector for both input and target tissue. 75% of this data was used for training and 15% for validation. The test data was created by taking the last 10% of samples and creating targets and corresponding target tissue vectors for all possible combinations of samples from the same donor as the input.

The reconstruction error was implemented as a Mean Squared Error (MSE) between the true and predicted expression. For the Kullback-Liebler divergence the prior distribution was chosen to be the multivariate standard normal distribution and was implemented according to equation 2.13. The total loss was calculated as

$$L_{total} = D_{KL}(N(\boldsymbol{\mu}, \boldsymbol{\sigma})||N(0, \boldsymbol{I})) + 500 * MSE(\boldsymbol{X'}, \boldsymbol{X}) \tag{3.2}$$

taking 500 from the number of variables in the input.

The architecture of the model can be seen in figure 3.2. Here the input $\boldsymbol{X}$ is mapped to the latent space $\boldsymbol{Z}$ together with the input tissue vector $\boldsymbol{S}$. The latent space is then decoded together with the target tissue vector $\boldsymbol{S'}$ to get the approximated input $\boldsymbol{X'}$. The implementation of the tissue vectors as complementary inputs is based on stVAE[24], that has been used to predict what tissue a given RNA sample was taken from in mice. For improved performance of the VAE, a naive parameter sweep was done. The parameters evaluated were the sizes of the hidden layers, the depth of the network, batch size and latent dimension. The parameters ran for all samples with the settings in table 3.2.

Table 3.2: *Values included in the parameter sweep.*

VAE parameter sweep	
Parameter	**Values**
Number of Layers	2,3
Layer sizes	512,364,256,128
Batch size	32,64,128
Latent dimension	50,100,150

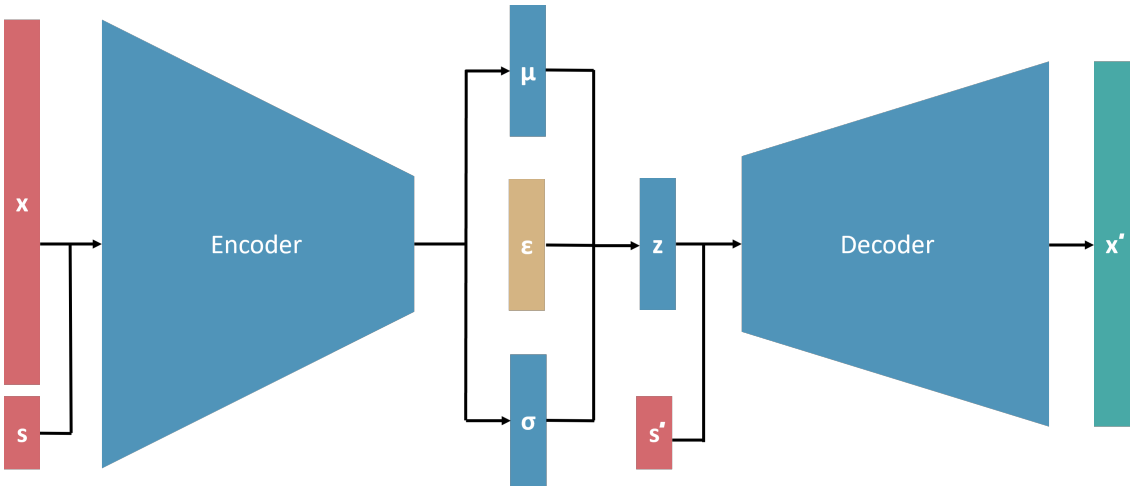


Figure 3.2: *An overview of our VAE models*

The size of the layers were only applied in a way so that the encoder had layers of descending size, and the decoder of ascending size. The layout of the encoder and decoder always mirrored each other. The model with the lowest MSE on the validation data after 10 epochs was chosen.

The final architecture of the model can be seen in table 3.3. The input is concatenated with the tissue vector and fed to the encoder that takes the form of an Multi Layer Perceptron, MLP, with 3 layers. Equation 2.11 is applied to get the latent variables which are concatenated with the target tissue vector and fed to the decoder which is a MLP with a structure mirroring the encoder. In order to speed up training, batch normalisation was implemented after the dense layers in both encoder and decoder. The model ran for 200 epochs with a learning rate of 10^{-5} and a batch size of 32.

Table 3.3: *Overview of the layers in the VAE models*

VAE Model		
Layer Name	**Layer size**	**Input**
Concatenate 1	530	X and S
Dense 1	364	Concatenate 1
Dense 2	256	Dense 1
Dense 3	128	Dense 2
Dense μ	100	Dense 3
Dense σ	100	Dense 3
Z Sampling function	100	Dense μ, Dense σ, ϵ
Concatenate 2	130	Z and S'
Dense 4	128	Concatenate 2
Dense 5	256	Dense 4
Dense 6	364	Dense 5

The evaluation of the model ran on the test data, attempting to predict the score in the

target tissue, based on the score in the original tissue. The predicted score values in the target tissue $\boldsymbol{Y}'$ were compared to the true values $\boldsymbol{Y}$ and the MSE between these vectors was calculated as a metric for the performance of the models. This was compared to the MSE between the score in the original tissue $\boldsymbol{X}$ and $\boldsymbol{Y}$, as well as the MSE between $\bar{\boldsymbol{X}}$ and the target $\boldsymbol{Y}$. Here $\bar{\boldsymbol{X}}$ denotes the mean value of the score for a gene set taken over all samples in the test data. These simple points of comparison were chosen to get a baseline reference point on the models' performance.

4
Results

This chapter is divided into two sections, variance decomposition and gene set prediction. Each section presents the results produced in relation to each subject.

4.1 Variance decomposition

The objective of the variance decomposition of the project was to investigate the use of GTEx as reference data and see which factors and subfactors that contributed to the variance in gene set scoring. The decomposition was performed for both the singscore and GSVA method by computing the ANOVA table first, where the meta factors' contribution to the variance of the score was visible. Following this, SMD and Fisher's exact test were applied to GTEx in comparison with the IBD Plexus data. This was done to investigate in which instances the distribution for GTEx was either similar or different to the other data set.

4.1.1 Visualising variance

Before any further analysis was made, a visual investigation was performed to get an intuitive feeling of the scores given by GSVA and singscore connected to the objectives of the report. The overall objective of the report is to investigate the variance of the scores, and the contributing factors for that variance. Therefore, first the mean and variance from the scores within each tissue in GTEx and gene set in MSigDB's C8 collection, a single-cell, tissue specific gene set collection, was summarised in figure 4.1. Each dot is the mean and variance for the combination of each gene set in the collection and tissue in GTEx. Figure 4.1a depicts the mean and variance for the scores given by singscore, while figure 4.1b depicts the same for GSVA. The general structure of the two plots are drastically different. Note that on the y-axis, the size of the variance differs in magnitude for the scoring methods where singscore have an overall lower variance than GSVA.

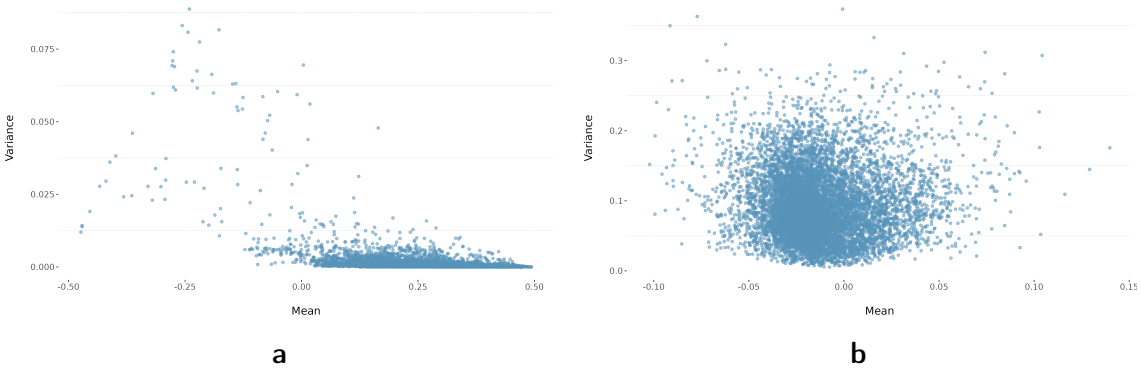


Figure 4.1: *Mean and variance of the score for the combination of gene set in the C8 collection from MSigDB and tissue from GTEx. Each point is the mean and variance for singscore in figure 4.1a and for GSVA in figure 4.1b.*

In figure 4.1a, the overall bulk of the scores have a small variance and a mean around 0 or higher, whereas some have a larger variance and a lower mean. The latter group with high variance, all comes from only three different gene set. The distribution of scores for these three gene sets are depicted in figure 4.2. Two things are worth noting from this figure. Firstly, all of the gene sets have a large spread of the score spectrum where FAN_EMBRYONIC_CTX_IN_5_INTERNEURON have scores over the whole range stated in section 2.4. Secondly, all three have a significant number of scores at $-0,5$, which indicates non-expressed genes in the gene set. Figure 4.1a in combination with figure 4.2 highlights that these gene sets are not representative of the overall behaviour of gene sets.

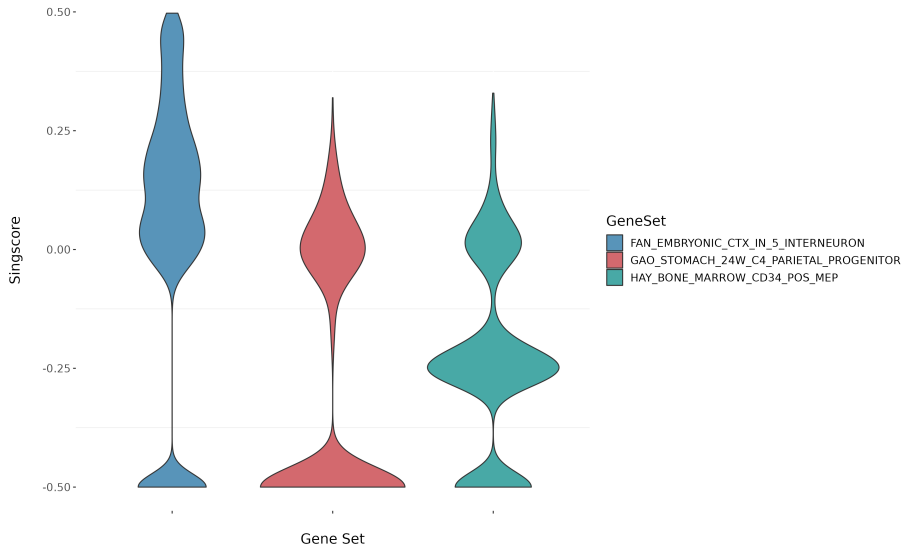


Figure 4.2: *The distribution of the singscore results for the three gene sets in MSigDB's C8 collection in GTEx with the highest variance.*

Additionally, an investigation of how the scores in GTEx behaved for a tissue specific

gene set, within or without that specific tissue was made. It was chosen to investigate three gene sets in MSigDB that was curated from liver cells and compared with samples in GTEx. The results can be found in figure 4.3 and two things can be noted. Firstly, in figure 4.3a, all of the scores for the ones in the liver is more dense and have slightly larger mean than the ones not in liver. Secondly, in figure 4.3b, for the samples in the liver, all the gene sets are more dense and have a higher mean than the ones not in liver. Figure 4.3b depicts how singscore separate the scores as anticipated per gene set in liver. The gene sets chosen for the investigation comes from analysis of single-cell sequencing from cells in the liver but the hepoatocytes cells (the blue group in figure 4.3 are more homogeneous cells than the two others. When comparing the gene sets expression with the liver samples in GTEx, where the samples are a result from biopsies which gather a variety of tissue cells per sample, a more specific and homogeneous gene set should be better expressed in those samples. This is clearly seen how singscore separates the three gene sets into two groups.

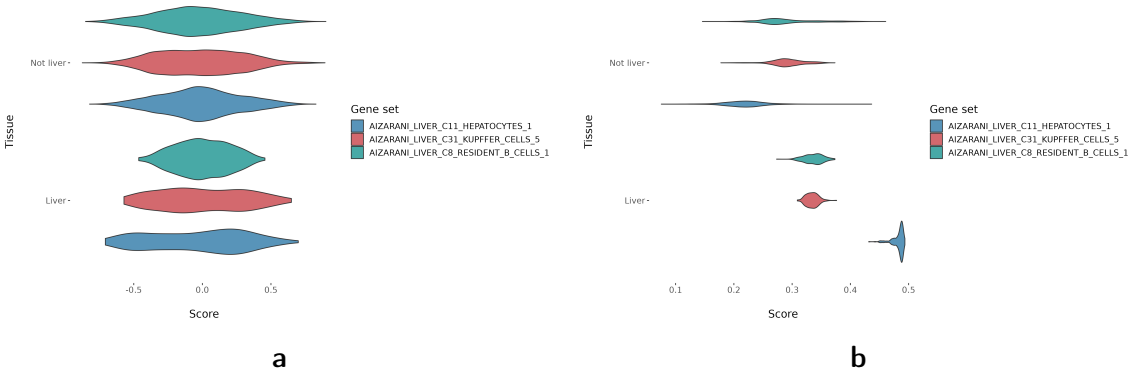
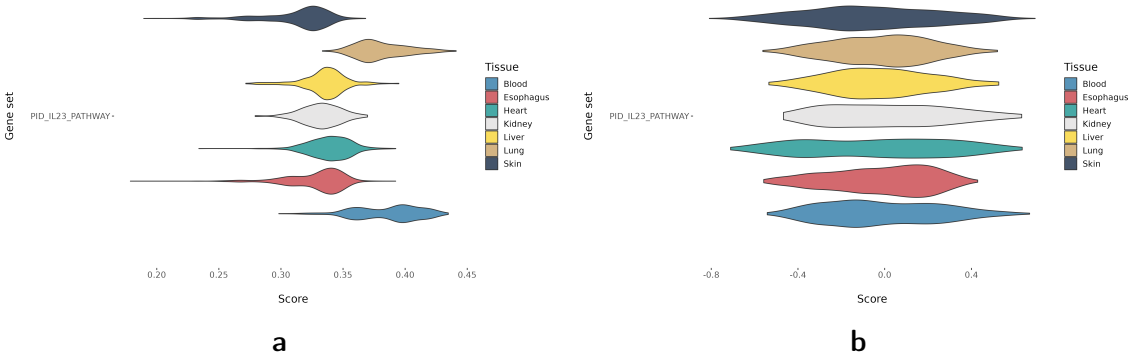


Figure 4.3: *The distributions of GSVA scores (figure 4.3a) and singscore (figure 4.3b) for three liver specific gene sets in GTEx, comparing between the samples in the liver and outside of the liver.*

To see the behaviour of a single gene set in multiple tissues, figure 4.4 was constructed. This figure looks at the distribution of scores from singscore in figure 4.4a and GSVA in figure 4.4b and how the scores vary between tissues. The used gene set, PID_IL23_PATHWAY, comes from the hallmark collection and was curated from genes activated during inflammation.

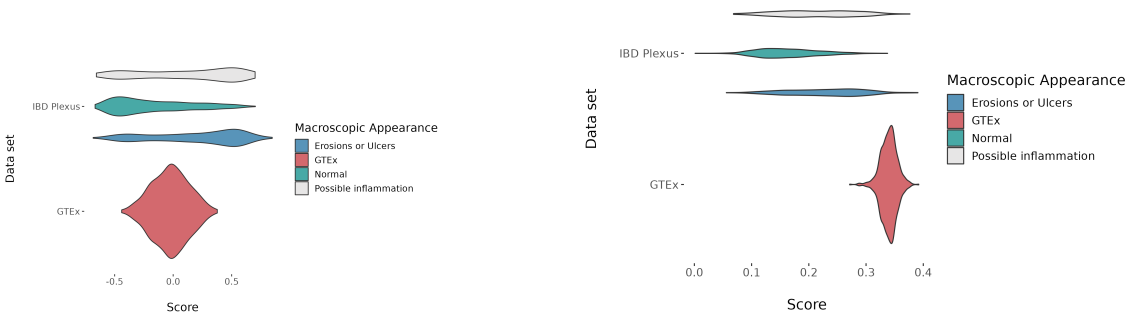
Stated in section 1.2, another objective was to investigate if singscore and GSVA can be used to compare between data sets. Visually, this was investigated by looking at the distribution of GTEx compared to another data set for a gene set that is tissue specific. This investigation is summarised for both GSVA and singscore in figure 4.5 where GTEx is compared to the scores in IBD Plexus for the meta factor macroscopic appearance. As IBD plexus contains samples from patients from the colon where some have possible or actual inflammation in the colon, a comparison can be made for a gene set, PID_IL23_PATHWAY, which is up-regulated during inflammation compared to GTEx. One can note in figure 4.5 that GTEx is more dense. For GSVA in figure 4.5a the group with possible inflammation and erosion or ulcers have a group with higher scores. In



a b

Figure 4.4: *Violin plots showcasing the distributions of the gene set PID_IL23_PATHWAY across some tissues in GTEx for singscore (figure 4.4a) and GSVA (figure 4.4b).*

figure 4.5b, GTEx have a dense and high mean compared to any of the subfactors.



a b

Figure 4.5: *Violin plots comparing the distribution of the GSVA scores (figure 4.5a) and Singscore (figure 4.5b) between IBD Plexus's factor macroscopic appearance with the colon samples in GTEx for the gene set PID_IL23_PATHWAY.*

4.1.2 ANOVA and AIC

To assess which factor contributed to the variance of the Singscore and GSVA, an ANOVA analysis was performed on the GTEx data. What factors to include in the analysis was deduced by using AIC. Three models were proposed (table 3.1) based on the meta data available for GTEx and the AIC for each model was calculated and summarised in table 4.1 - 4.2.

Table 4.1: *AIC table for the three models with the scores from singscore of the Hallmark gene sets.*

	df	AIC
Small model	1501	-4 510 102
Middle model	1556	-4 511 830
Large model	1756	-4 495 799

Table 4.2: *AIC table for the three models with the score from GSVA of the Hallmark gene sets.*

	df	AIC
Small model	1501	-495 278
Middle model	1556	-496 794
Large model	1756	-525 415

One can note in table 4.1 and table 4.2 that all of the AIC values are very negative. The model with the lowest AIC-value was selected for the ANOVA analysis, middle for singscore and the large model for GSVA. The ANOVA table for singscore can be seen in table 4.3 and for GSVA in table 4.4.

Table 4.3: *The ANOVA table from the model with lowest AIC score with scores from singscore of the Hallmarks as gene sets.*

	df	SS	MS	F-value	Pr(>F)
Tissue	29	111,0	3,82	11 750	>2e-16
GeneSet	49	4233	86,38	265 600	>2e-16
Age	5	0	0,02000	65,86	>2e-16
Sex	1	0	0,1300	403,1	>2e-16
Tissue:GeneSet	1421	449	0,32	972,7	>2e-16
GeneSet:Sex	49	0,0000	0,01	22,54	>2e-16
Residual	867 545	282,0	0,00		

Table 4.4: *The ANOVA table from the model with lowest AIC score with scors from GSVA of the Hallmark gene sets.*

	df	SS	MS	F-value	Pr(>F)
Tissue	29	12 925	445,7	14 015,92	>2e-16
GeneSet	49	700	14,3	449,27	>2e-16
Age	5	30	6,0	189,43	>2e-16
Sex	1	3	3,4	107,27	>2e-16
HardyScale	4	233	58,2	1 831,79	>2e-16
Tissue:GeneSet	1421	26 983	19,0	597,17	>2e-16
GeneSet:Sex	49	20	0,4	12,77	>2e-16
GeneSet:HardyScale	196	704	3,6	112,89	>2e-16
Residual	861 795	27 404	0,0		

An observation that can be drawn from both table 4.3 and 4.4 is that all factors are highly significant. However, this is due to the large amount of data points which reduces the residual and inflates the F-values. Looking at the MS values can give an indication of the overall weight of that factor in relationship to the other factors.

4.1.3 Standardised mean difference

The first method aimed to investigate the similarities or differences between the distributions of the score from GTEx compared to IBD Plexus, was SMD, standardised mean. For IBD Plexus, as stated in section 3.5.3, three different meta factors were analysed per scoring method are summarised in figure 4.6 and 4.7. The figures show how many of the combinations of gene set and sub factor combinations are significantly different from each other. Noticeable is that GSVA (figure 4.6) have a majority of none significant sub factor and gene set combinations while singscore have a majority of significant combinations (figure 4.7).

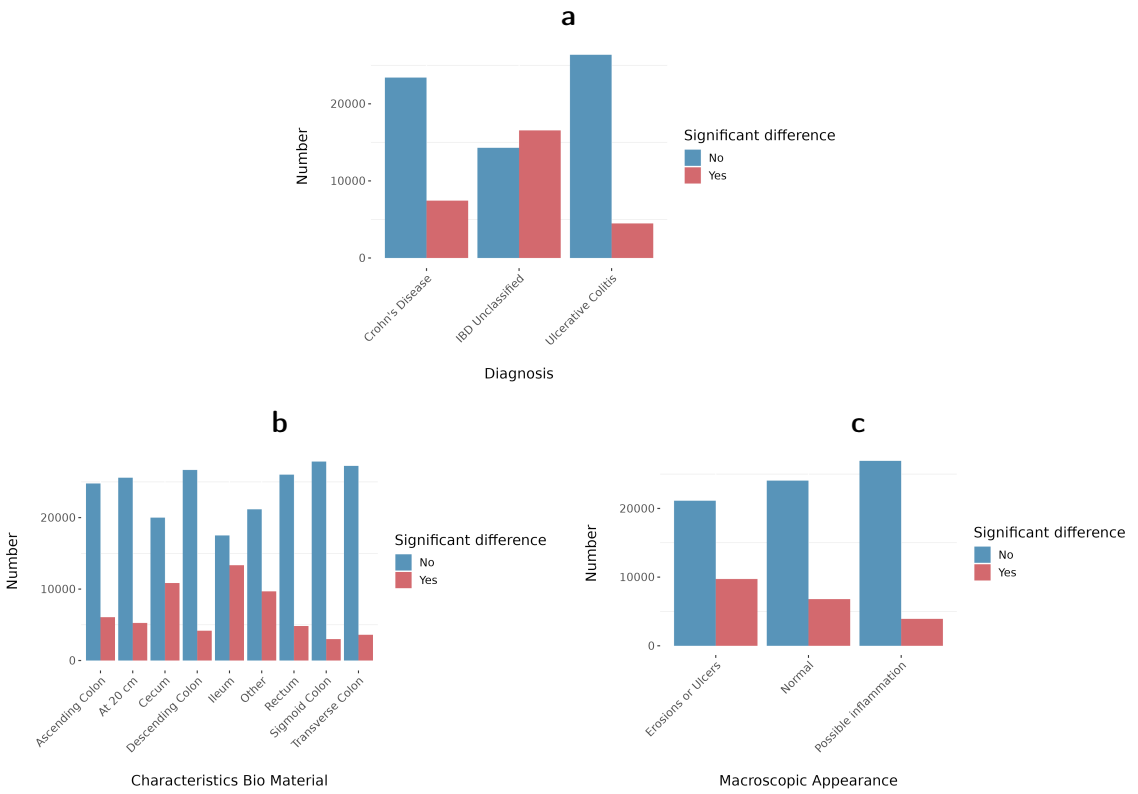


Figure 4.6: *Histograms of the SMD results for GSVA applied on IBD Plexus in comparison with the colon samples in GTEx using the entire MSigBD gene set collection.*

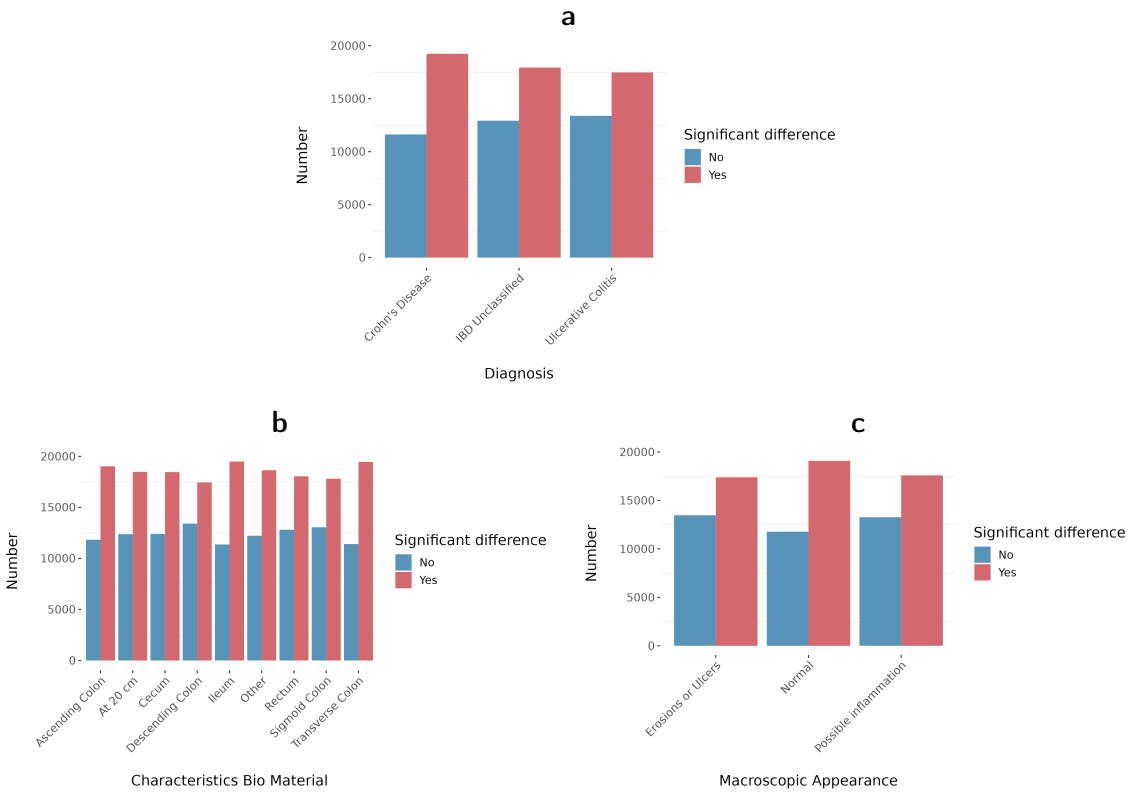


Figure 4.7: *Histograms of the SMD results for singscore applied on IBD Plexus in comparison with the colon samples in GTEx using the entire MSigBD gene set collection.*

To see if GSVA and singscore distinguish the same combinations as significant, table 4.5 was constructed which summarises the overlap of significant combinations between GSVA and singscore. One can note the the actual overlap of combinations of sub factors and gene set vary drastically from the lowest of only 13,2% to the highest of 62,9%. This means that GSVA and singscore distinguish 13.2% of the same gene set for samples taken from the Sigmoid colon as significantly different while 62,9% for samples with IBD unclassified.

Table 4.5: *The overlap of significant combinations, according to the SMD results, of gene set and meta factors between GSVA and singscore for IBD Plexus.*

Position	Percent overlap	Macroscopic appearance	Percent overlap	Diagnosis	Percent overlap
Ascending colon	0,264	Erosions or Ulcers	0,432	Crohn's Disease	0,324
At 20 cm	0,221	Normal	0,295	IBD unclassified	0,629
Cecum	0,458	Possible inflammation	0,174	Ulcerative colitis	0,197
Descending colon	0,192				
Ileum	0,522				
Other	0,400				
Rectum	0,199				
Sigmoid colon	0,132				
Transverse colon	0,145				

The analysis in the previous figures and tables for SMD was made across all gene sets in MSigDB. However, a real-life application would be to investigate if GTEx can be used as a reference for a specific tissue and gene set. Therefore, in table 4.6-4.7, one can see the SMD scores for GSVA and singscore between IBD Plexus and GTEx for a specific gene set, PID_IL23_PATHWAY, that up-regulates genes during an inflammation.

Table 4.6: *SMD values for the GSVA scores compared between GTEx and IBD Plexus for the gene set PID_IL23_PATHWAY.*

Position	SMD	Significant difference	Macroscopic appearance	SMD	Significant difference	Diagnosis	SMD	Significant difference
Ascending Colon	1,35	No	Erosions or Ulcers	0,541	No	Crohn's Disease	1,45	No
At 20 cm	1,74	No	Normal	2,18	Yes	IBD unclassified	1,67	No
Cecum	1,96	No	Possible inflammation	0,781	No	Ulcerative colitis	1,53	No
Descending colon	1,08	No						
Ileum	1,16	No						
Other	0,817	No						
Rectum	1,02	No						
Sigmoid colon	1,38	No						
Transverse colon	1,35	No						

Table 4.7: *SMD values for the scores from singscore compared between GTEx and IBD Plexus for the gene set PID_IL23_PATHWAY.*

Position	SMD	Significant difference	Macroscopic appearance	SMD	Significant difference	Diagnosis	SMD	Significant difference
Ascending Colon	3,16	Yes	Erosions or Ulcers	2,42	Yes	Crohn's Disease	3,56	Yes
At 20 cm	3,63	Yes	Normal	4,66	Yes	IBD unclassified	2,95	Yes
Cecum	3,90	Yes	Possible inflammation	2,66	Yes	Ulcerative colitis	3,32	Yes
Descending colon	2,82	Yes						
Ileum	3,53	Yes						
Other	3,06	Yes						
Rectum	2,68	Yes						
Sigmoid colon	2,76	Yes						
Transverse colon	3,28	Yes						

4.1.4 Fisher's exact test

Fisher's exact test was used to further investigate if the score distributions from GTEx compared to another data set was different or not. The same procedure was used as for SMD and the results for IBD Plexus can be found for GSVA in figure 4.8 and for singscore in figure 4.9. One can note that we find a majority of significant differences between the sub factors and gene set combinations, but GSVA have a lower rate.

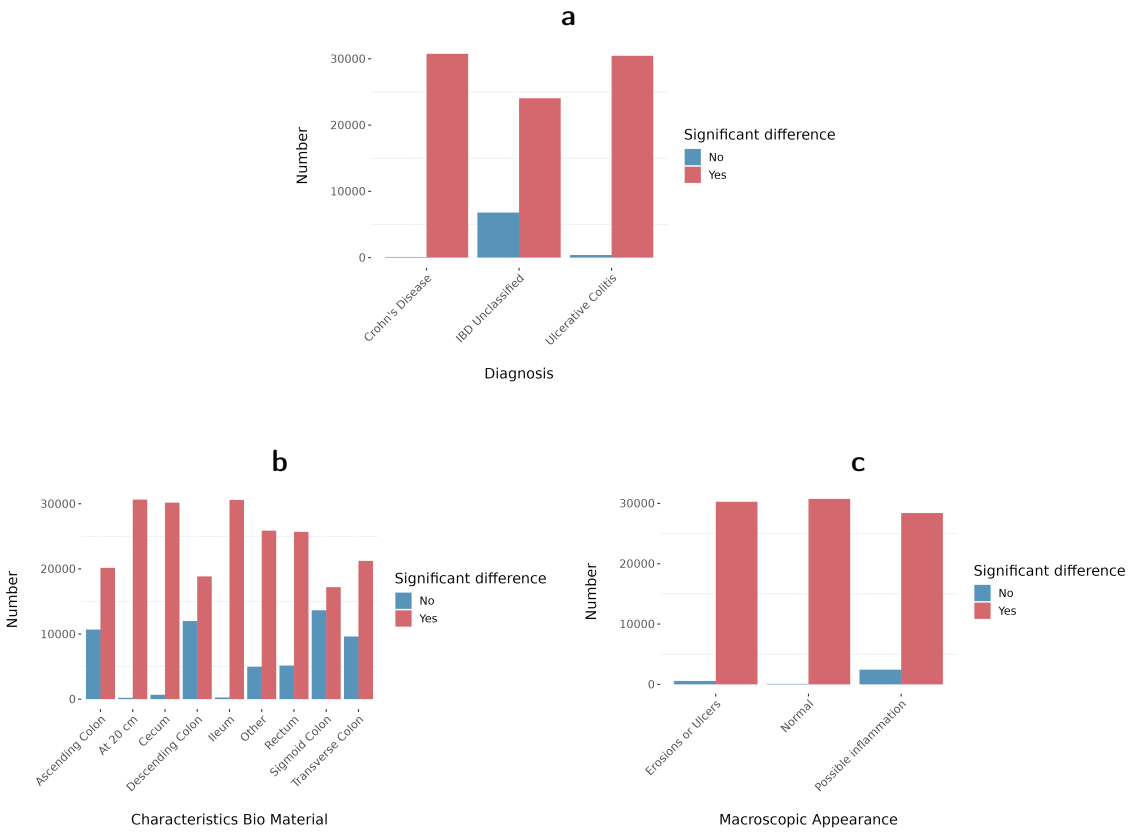


Figure 4.8: *Histograms of the Fisher's exact test results for GSVA applied on IBD Plexus in comparison with the lung samples in GTEx using the entire MSigBD gene set collection.*

Due to the high amount of significant combinations of sub factors and gene set, an investigation of the p-values was performed. Almost all of the combinations had the lowest p-value possible for the analysis. Thus, no further analysis was made as making conclusions where everything is significantly different is not relevant.

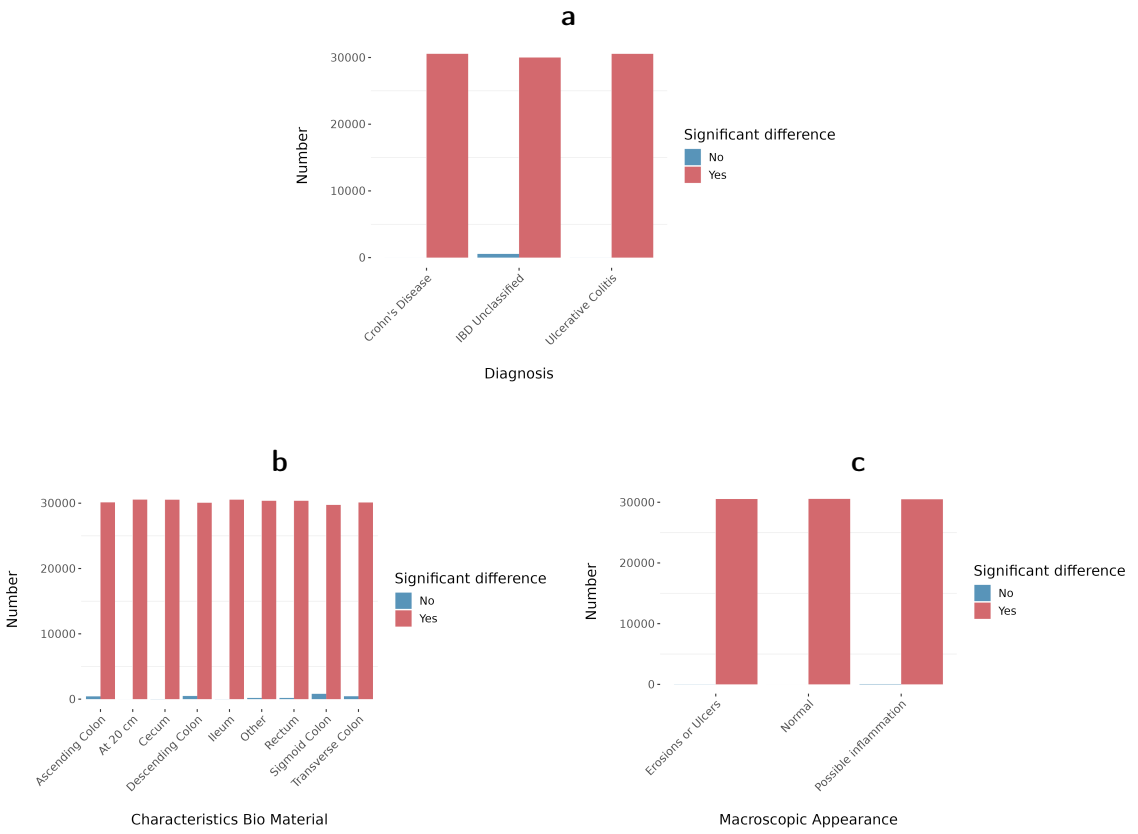


Figure 4.9: *Histograms of the Fisher's exact test results for scores from singscore applied on IBD Plexus in comparison with the lung samples in GTEx using the entire MSigBD gene set collection.*

4.2 Variational autoencoder

The objective of the VAE analysis was to qualitatively state if a VAE was able to model the expression of genes in a different tissue based on the expression in the current tissue. To this end the comparison of Mean Square Error laid out in chapter 3.6 was done. The results of the two models when compared to the naive comparisons of taking the input directly as well as the mean of the variables can be seen in table 4.8.

Table 4.8: *The results of the VAEs, $MSE(Y', Y)$ compared to taking the input directly, $MSE(X, Y)$ and taking the mean, $MSE(\bar{X}, Y)$.*

MSE results of expression prediction			
	$MSE(Y', Y)$	$MSE(X, Y)$	$MSE(\bar{X}, Y)$
GSVA Model	0,060	0,269	0,119
singscore Model	0,036	0,269	0,090

5
Discussion

The discussion chapter goes over the results presented in chapter 4 and ties them to the research questions posed in chapter 1.2. Further work related to the research questions are discussed in each chapter.

5.1 Variation analysis

In section 1.2 the objective and research question of this report was stated. This section aims to answer which factors that contributes to the variance of the scores given by GSVA and singscore by discussing the results from the visual investigation and the ANOVA analysis.

To investigate the factors which contributed to the variance of the scores, an ANOVA analysis was implemented on GTEx using the hallmark gene sets in MSigDB. From table 4.1 and 4.2 one can note that the two scoring methods have different models which describes the data best. This is further highlighted in table 4.3 and 4.4 where the most contributing factors differ for the two scoring methods. However, one should note that the order of the factors in the models matter. ANOVA calculates the total variance and for each factor ask how much of a contribution they add to that variance, given the factors that already have been added to the analysis. E.g., when investigating the impact of the factor "GeneSet" in 4.3, the variance contribution is in relationship with the factor "Tissue" already being in the model. Therefore, this highlights the contribution of variance that the factor "HardyScale" have in GSVA in table 4.4, when the MS-value is so high even though "HardyScale" is added so late in the analysis. The conclusion drawn from "HardyScale" being significant is that the factor introduces a lot of variability, and the results could benefit from separating the samples in GTEx into the levels of "HardyScale". This was, however, not done due to the inbalanced sample distribution of the levels in "HardyScale". Furthermore, correlated factors will have a shared contribution to the variance and depending on which order they are added will have different significance. However, no correlation analysis was made. For both scoring methods, gene set and tissue have a large contribution to the variance. That being said, as the analysis was only made on the hallmark gene sets, which include only 50 gene sets, the effect of gene set might be exaggerated.

From the visual analysis in section 4.1.1, one can note that the two scoring methods singscore and GSVA, have two very different behaviours, as seen in figure 4.1. The figure is only for one of the 9 gene set collections from MSigDB but they highlight

a fundamental difference between the two methods that will be further discussed in section 5.2. Additionally, figure 4.4 showcase how the underlying distribution of the scores varies when investigating a single gene set in different tissues. This highlights the use of knowing the pathway that a gene set is connected to and know the relevance of analysing a gene set across tissues. Figure 4.3 adds to the conclusion of figure 4.4 as this shows how gene sets have different behaviours if they are analysed in the context from where they were curated. Lastly, figure 4.3 also connects with how well singscore and GSVA are at comparing between and within data sets, where in this case singscore prominently show to distinguish the correct behaviour as expected.

5.2 GSVA and singscore

The thesis set out to analyse the viability of using singscore and GSVA for broad analysis of many gene sets on large amounts of data. This with the aim to answer whether analysis across different studies is viable for each of the scoring methods.

From the statistical analysis in section 4.1 information was gathered on the feasibility of cross-study evaluation. In figures 4.5 - 4.9 it can be observed that for all cases looked at, GSVA has significantly more similar distributions than singscore. A likely reason for this difference is the more aggressive normalisation in the GSVA method, which in this broad case appears to remove away some information. Another point to highlight is that in figure 4.5 it can be seen that GSVA has a distribution most similar to the cases that have no indication of disease , while singscore is closest to the samples labelled as having diseased appearance. Lastly, looking to table 4.6 and 4.7 we can note that the GSVA results are noticeably more similar when using SMD to compare between the two datasets. But a worrisome detail is that the only significant difference in GSVA is for the *Normal* macroscopic appearance. This is the opposite of what is wanted for using GTEx as a reference dataset. Looking to the results of Fisher's exact test in figure 4.8 and 4.9 the significance of the difference is even more prominent. Therefore, for both SMD and Fisher's exact test the results suggest a significantly different distribution across all combinations. This implies that using GTEx as a reference for other datasets is not possible.

Taken together this mixed result gives no clear answer on which method is to prefer, but the results do not support the usage of either method for cross-study analysis. Further analysis on this application for the scoring methods should focus on more specific tissues and gene sets that have well researched connections. After the behaviour of the models has been analysed for these specific methods, more broad analyses, like the one in this report, should be better supported and viable. More work on batch correction methods would also need to be used in order to align the distributions of the different datasets better.

Underlying all the results, are the expressed genes in the data sets. In figure 4.2, one can note that a large portion of the scores in all three groups have samples with genes with no expression as they have received the score of $-0,5$. Section 3.2 discusses the problem of choosing a cut off which includes the genes that you want to analyse and

exclude data that has low quality and would lower the quality of the results. Due to the method in this report being rather broad, a generous cut off was chosen. However, this means keeping genes in some samples with no expression, which affects the results. Comparing distributions where samples have genes with no expression with samples that have an expression will naturally produce significant differences. For future work, a more in-depth analysis on which cut off to use could improve the overall quality of the results.

5.3 VAE modelling

The objective of applying the VAE was that it would pick up on the nonlinear behaviours of the underlying data and be able to give robust prediction of expression over samples. The results in table 4.8 show that while the model does not fail altogether, the level of accuracy needed in GSEA is not achieved in these tests.

A factor that is likely to have played in heavily in the results is the fact that the gene sets with the highest variance were selected. Picking out features with high variance is a common method in machine learning to do a rough optimisation of the amount of information left for the model while reducing its size. But it has been shown that the high variance gene sets for GTEx can occur as a result of contamination and other defects [25]. This could have affected the model severely and would need to be further analysed for future models. It would also be prudent to trial other more robust measures of feature selection.

All in all, the VAE model doesn't show promise for this application. The result in table 4.8 show that some intricacies are modelled but the results are not significant enough to warrant enthusiasm. The underwhelming result is not fully surprising, as direct prediction of high dimensional data is not a common application of VAEs, rather categorical prediction is more commonly tried.

Further work in this area could take the approach of using VAE architectures made for data augmentation and treat the input as faulty data that should be corrected to fit the given tissue instead of predicting directly. Another interesting experiment is applying this sort of direct prediction model we have done and applying to TPM normalised expression data and comparing how well that can be modelled.

Bibliography

[1] Arthur Liberzon et al. "Molecular signatures database (MSigDB) 3.0". In: *Bioinformatics* 27.12 (June 2011), pp. 1739–1740. ISSN: 13674803. DOI: 10.1093/bioinformatics/btr260.

[2] Christiaan A. De Leeuw et al. "The statistical properties of gene-set analysis". In: *Nature Reviews Genetics* 17.6 (June 2016), pp. 353–364. ISSN: 14710064. DOI: 10.1038/nrg.2016.29.

[3] Farhad Maleki et al. *Gene Set Analysis: Challenges, Opportunities, and Future Research*. June 2020. DOI: 10.3389/fgene.2020.00654.

[4] Aravind Subramanian et al. "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles". In: *Proceedings of the National Academy of Sciences* 102.43 (Oct. 2005), pp. 15545–15550. ISSN: 0027-8424. DOI: 10.1073/pnas.0506580102.

[5] D. Nam and S.-Y. Kim. "Gene-set approach for expression pattern analysis". In: *Briefings in Bioinformatics* 9.3 (Jan. 2008), pp. 189–197. ISSN: 1467-5463. DOI: 10.1093/bib/bbn001.

[6] Bruce Alberts et al. *Molecular Biology of the Cell*. 6th ed. Garland Science, 2014, pp. 1–1465. ISBN: 9780815344322.

[7] John Lonsdale et al. *The Genotype-Tissue Expression (GTEx) project*. June 2013. DOI: 10.1038/ng.2653.

[8] Laura E. Raffals et al. "The Development and Initial Findings of A Study of a Prospective Adult Research Cohort with Inflammatory Bowel Disease (SPARC IBD)". In: *Inflammatory Bowel Diseases* 28.2 (Feb. 2022), pp. 192–199. ISSN: 15364844. DOI: 10.1093/ibd/izab071.

[9] F. Finotello and B. Di Camillo. "Measuring differential gene expression with RNA-seq: challenges and strategies for data analysis". In: *Briefings in Functional Genomics* 14.2 (Mar. 2015), pp. 130–142. ISSN: 2041-2649. DOI: 10.1093/bfgp/elu035.

[10] FRANCIS CRICK. "Central Dogma of Molecular Biology". In: *Nature* 227.5258 (Aug. 1970), pp. 561–563. ISSN: 0028-0836. DOI: 10.1038/227561a0.

[11] Ana Conesa et al. "A survey of best practices for RNA-seq data analysis". In: *Genome Biology* 17.1 (Dec. 2016), p. 13. ISSN: 1474-760X. DOI: 10.1186/s13059-016-0881-8.

[12] Yingdong Zhao et al. "TPM, FPKM, or Normalized Counts? A Comparative Study of Quantification Measures for the Analysis of RNA-seq Data from the NCI Patient-Derived Models Repository". In: *Journal of Translational Medicine* 19.1 (Dec. 2021), p. 269. ISSN: 1479-5876. DOI: 10.1186/s12967-021-02936-w.

[13] Arthur Liberzon et al. "The Molecular Signatures Database Hallmark Gene Set Collection". In: *Cell Systems* 1.6 (Dec. 2015), pp. 417–425. ISSN: 24054712. DOI: 10.1016/j.cels.2015.12.004.

[14] Sonja Hänzelmann, Robert Castelo, and Justin Guinney. "GSVA: gene set variation analysis for microarray and RNA-Seq data". In: *BMC Bioinformatics* 14.1 (Dec. 2013), p. 7. ISSN: 1471-2105. DOI: 10.1186/1471-2105-14-7.

[15] Shanshan Yu et al. "Seven-Gene Signature Based on Glycolysis Is Closely Related to the Prognosis and Tumor Immune Infiltration of Patients With Gastric Cancer". In: *Frontiers in Oncology* 10 (Sept. 2020). ISSN: 2234943X. DOI: 10.3389/fonc.2020.01778.

[16] Momeneh Foroutan et al. "Single sample scoring of molecular phenotypes". In: *BMC Bioinformatics* 19.1 (Nov. 2018). ISSN: 14712105. DOI: 10.1186/s12859-018-2435-4.

[17] Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Academic press, 1988, pp. 19–66.

[18] Murray J. Fisher, Andrea P. Marshall, and Marion Mitchell. "Testing differences in proportions". In: *Australian Critical Care* 24.2 (May 2011), pp. 133–138. ISSN: 10367314. DOI: 10.1016/j.aucc.2011.01.005.

[19] Ronald Aylmer Fisher. *Design of Experiment*. 1935, pp. 13–29.

[20] Diederik P Kingma and Max Welling. "Auto-Encoding Variational Bayes". In: (Dec. 2013). URL: http://arxiv.org/abs/1312.6114.

[21] Broad Institute of MIT and Harvard. *GTEx Portal*. URL: https://gtexportal.org/home/.

[22] Jan De Leeuw. *INFORMATION THEORY AND AN EXTENSION OF THE MAXIMUM LIKELIHOOD PRINCIPLE BY HIROTOGU AKAIKE*. Tech. rep.

[23] Ruth G. Shaw and Thomas Mitchell-Olds. "Anova for Unbalanced Data: An Overview". In: *Ecology* 74.6 (Sept. 1993), pp. 1638–1645. ISSN: 00129658. DOI: 10.2307/1939922.

[24] Nikolai Russkikh et al. "Style transfer with variational autoencoders is a promising approach to RNA-Seq data harmonization and analysis". In: *Bioinformatics* 36.20 (Dec. 2020), pp. 5076–5085. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btaa624.

[25] Tim O. Nieuwenhuis et al. "Consistent RNA sequencing contamination in GTEx and other data sets". In: *Nature Communications* 11.1 (Apr. 2020), p. 1933. ISSN: 2041-1723. DOI: 10.1038/s41467-020-15821-9.

A
Appendix 1

Table A.1 shows the different versions used in the project for program language, packages and MSigDB.

Table A.1: *The different version for the languages, packages and MSigDB that was used in the project*

What	Version
Python	3.7.2
R	4.1.0
Tensorflow	2.11
Keras	2.11
Singscore	1.14.0
GSVA	1.42.0
MSigDB	v7.2.hs.SYM
GSEABase	1.56.0

B

Appendix 2

The full list of packages used in R

```
R version 4.1.0 (2021-05-18)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: CentOS Linux 7 (Core)

Matrix products: default
BLAS/LAPACK: /opt/scp/software/OpenBLAS/
0.3.5-GCC-8.2.0-2.31.1/lib/libopenblas_haswellp-r0.3.5.so

locale:
 [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
LC_TIME=en_US.UTF-8        LC_COLLATE=en_US.UTF-8

 [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
  LC_PAPER=en_US.UTF-8       LC_NAME=C

 [9] LC_ADDRESS=C               LC_TELEPHONE=C
LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C

attached base packages:
[1] parallel stats4    stats     graphics  grDevices datasets
utils     methods   base

other attached packages:
 [1] limma_3.50.3                biomaRt_2.50.3
rlang_1.1.0                 rapiclient_0.1.3
 [5] GSEABase_1.56.0             graph_1.72.0
annotate_1.72.0             XML_3.99-0.14
 [9] AnnotationDbi_1.56.2        singscore_1.14.0
BiocManager_1.30.20         slurmR_0.5-3
[13] lubridate_1.9.2             forcats_1.0.0
stringr_1.5.0               dplyr_1.1.1
[17] purrr_1.0.1                 readr_2.1.4
tidyr_1.3.0                 tibble_3.2.1
[21] ggplot2_3.4.2               tidyverse_2.0.0
msigdb_1.2.0                SummarizedExperiment_1.24.0
```

```
[25] Biobase_2.54.0                  GenomicRanges_1.46.1
GenomeInfoDb_1.30.1                IRanges_2.28.0
[29] S4Vectors_0.32.4                MatrixGenerics_1.6.0
matrixStats_0.63.0                 ExperimentHub_2.2.1
[33] AnnotationHub_3.2.2             BiocFileCache_2.2.1
dbplyr_2.3.2                       BiocGenerics_0.40.0
[37] GSVA_1.42.0                     CePa_0.8.0

loaded via a namespace (and not attached):
 [1] colorspace_2.1-0                  ellipsis_0.3.2
XVector_0.34.0                       rstudioapi_0.14
 [5] bit64_4.0.5                      interactiveDisplayBase_1.32.0 fansi_1.
xml2_1.3.3
 [9] sparseMatrixStats_1.6.0          cachem_1.0.7
knitr_1.42                         png_0.1-8
[13] shiny_1.7.4                     HDF5Array_1.22.1
compiler_4.1.0                     httr_1.4.5
[17] Matrix_1.3-4                    fastmap_1.1.1
cli_3.6.1                          later_1.3.0
[21] BiocSingular_1.10.0             prettyunits_1.1.1
htmltools_0.5.5                    tools_4.1.0
[25] rsvd_1.0.5                      igraph_1.4.2
gtable_0.3.3                       glue_1.6.2
[29] GenomeInfoDbData_1.2.7          reshape2_1.4.4
rappdirs_0.3.3                     Rcpp_1.0.10
[33] vctrs_0.6.1                     Biostrings_2.62.0
rhdf5filters_1.6.0                 DelayedMatrixStats_1.16.0
[37] xfun_0.38                       beachmat_2.10.0
timechange_0.2.0                   mime_0.12
[41] lifecycle_1.0.3                 irlba_2.3.5.1
renv_0.17.3                        edgeR_3.36.0
[45] zlibbioc_1.40.0                 scales_1.2.1
hms_1.1.3                          promises_1.2.0.1
[49] rhdf5_2.38.1                    SingleCellExperiment_1.16.0
yaml_2.3.7                         curl_5.0.0
[53] memoise_2.0.1                   stringi_1.7.12
RSQLite_2.3.1                      BiocVersion_3.14.0
[57] ScaledMatrix_1.2.0              filelock_1.0.2
BiocParallel_1.28.3                pkgconfig_2.0.3
[61] bitops_1.0-7                    evaluate_0.20
lattice_0.20-44                    Rhdf5lib_1.16.0
[65] bit_4.0.5                       tidyselect_1.2.0
plyr_1.8.8                         magrittr_2.0.3
[69] R6_2.5.1                        generics_0.1.3
DelayedArray_0.20.0                DBI_1.1.3
```

```
[73] pillar_1.9.0
KEGGREST_1.34.0
[77] crayon_1.5.2
tzdb_0.4.0
[81] progress_1.2.2
grid_4.1.0
[85] Rgraphviz_2.38.0
xtable_1.8-4
[89] munsell_0.5.0
     withr_2.5.0
RCurl_1.98-1.12
     utf8_1.2.3
rmarkdown_2.21
     locfit_1.5-9.7
blob_1.2.4
     digest_0.6.31
httpuv_1.6.9
```

C
Appendix 3

Below there are links to where more information can be found on how we can get access to the data used in this thesis.

MSigDB: https://www.gsea-msigdb.org/gsea/msigdb/index.jsp

GTEx: GTEx was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS and can be found at https://gtexportal.org/home/datasets

IBD Plexus: The SPARC IBD data are available upon approved application to Crohn's & Colitis Foundation IBD Plexus https://www.crohnscolitisfoundation.org/research/grants-fellowships/ibd-plexus

DEPARTMENT OF MATHEMATICAL SCIENCES
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden
www.chalmers.se

CHALMERS
UNIVERSITY OF TECHNOLOGY