Integrated Analysis of Methylation and RNA Sequencing Data to Uncover Potential Biomarkers for Endometriosis

Caroline Forsythe[1], Julia Lapucha[1], Astghik Sarukhanyan[1]

[1*]Department of Mathematics and Computer Science, Freie Universität, Arnimalle, Berlin, 14195, Germany.

Contributing authors: forsythec99@zedat.fu-berlin.de; lapuj92@zedat.fu-berlin.de; sarukhaa99@zedat.fu-berlin.de;

Abstract

Endometriosis is a severely understudied gynecological disorder affecting approximately 5%-10% of women [2]. As an all-female group of bioinformaticians with a special interest in women's health, a project was chosen focusing on endometriosis. RNA-seq (RNA sequencing) and MBD-seq (Methyl-binding domain sequencing) datasets provided by Akter et. al [5] were chosen to identify potential biomarkers. Differential gene expression and methylation analyses, combined with annotation, were employed to identify significantly expressed and methylated genes as well as regions. Functional enrichment analysis was performed to annotate significant results for biological functions, processes, components, pathways or diseases. Additionally, two ensemble classifiers were built to predict diseased versus healthy status in endometriosis patients and to identify important features/biomarkers for this classification. The first model was trained on the separate datasets then combined their probabilities to a meta model. The second model combines the datasets, trains a logistic regression, a decision tree and a support vector machine then combining their best predictions into a stacked classifier. Multiple genes and gene regions of interest were identified in the results, and these findings were compared with the current literature on endometriosis. While many findings could not be backed by literature, several genes that may serve as potential biomarkers were identified, such as IGF2 and the methylated region of CDCA2, which have been mentioned in endometriosis research and studies. This project showcases the need for more research regarding women's health, with a particular emphasis on endometriosis.

1 Background

Bioinformatics and data science are potent fields increasingly aiding research in various fields, especially in the field of medicine. By providing advanced computational and analytical methods these interdisciplinary fields can accomplish a variety of tasks by being able to analyze and interpret complex biological and clinical data. One such task is disease biomarker discovery through pattern recognition and statistical analysis. Discovering biomarkers can lead to better diagnostics as well as therapeutics.

One disease in current need of both better diagnostic methods and therapeutics is endometriosis, affecting approximately 190 million women and girls globally [37]. Endometriosis is characterized by the growth of tissue like that of which usually grows in the lining of the uterus, outside the uterus [37]. This can result in a patient experiencing symptoms like severe pelvic pain and even infertility. Unfortunately, the discomfort does not stop at the symptoms, but even the diagnosis involves an invasive surgical procedure called a laparoscopy, to examine the abdomen and pelvis [37]. As far as treatments, there is no set proven regiment but a combination of surgical procedures and medical management, such as pain relief and hormonal therapies, to aid in a patient's quality of life [37]. Aside from aiding diagnosis, biomarkers could advance personalized treatment approaches as well as ongoing monitoring.

One way to detect these biomarkers is to understand gene regulation and expression, in order to uncover specific genes that may be linked to endometriosis as biomarkers. In order to accomplish this kind of pattern detection, machine learning can be of great use. A common trope in machine learning problems is classification. By classifying patients as healthy and disease using biological data, features can be extracted and their importance analyzed. Previous work has been done using machine learning to understand patterns behind endometriosis. The paper in which this the data for this study was sourced from by Akter et al., used machine learning classifiers for endometriosis using transcriptomics and methylomics data [5]. Unlike the combination of logistic regression, support vector machine (SVM) and decision tree models in this study, Akter et al. solely used decision tree models and did not integrate the transcriptomic and methylation data [5]. Other studies such as that by Tan et al. used a single-cell analysis of endometriosis approach to aid the design of effective therapeutic strategies and or diagnostic biomarkers [42]. An additional study by Shih et al. also performed a single cell analysis, and was able to identify uterine natural killer (uNK) cells absent in the endometriosis patient population using machine learning clustering techniques [40].

This study aims to use both transcriptional and methylation data as input into machine learning methods to detect patterns, and find significant genes that may be relevant to endometriosis. Additional aspects of the workflow include differential gene and methylation analyses, as well as a functional enrichment analysis to tie to biological processes. The full outline of the workflow can be seen in Fig. 1.

2 Goal

The main goal of this project is to find potential biomarkers for endometriosis by combining information from gene expression (RNA-seq) and DNA methylation (MBD-seq)

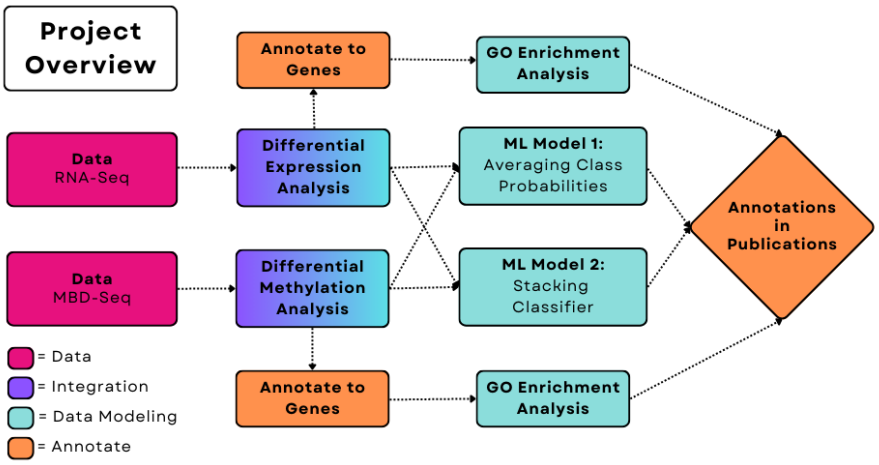


Figure 1: An outline of the project workflow.

data. We aim to identify specific genes and gene expression patterns that differ between endometriosis and healthy patients. Furthermore, machine learning techniques will be employed to predict disease based on significant results from differential gene expression and differential methylation analysis.

Endometriosis is a condition that currently requires invasive diagnostic and therapeutic procedures. As a predominately female group, we have a strong commitment to advancing women's health and are passionate about contributing to improvements in this field through modern bioinformatic workflows like differential gene expression and methylation analysis, GO enrichment analysis, and various machine learning approaches.

3 Data

This project takes a multi-omics approach, employing RNA-seq (RNA sequencing) and MBD-seq (Methyl-binding domain sequencing) data. The data is provided by [5], which also explores various machine learning classifiers for endometriosis. The data sets are from Homo sapiens, with samples consisting of both endometriosis patients and healthy controls.

3.1 RNA

The RNA-seq data was generated using the Illumina NextSeq 500 sequencing platform. The data can be obtained from the GEO database (GEO accession: GSE134056; Transcriptomics - LINK). The data set is available as a compressed count matrix file (`GSE134056_countdata_rnaseq.txt.gz`). The count matrix itself contains 58,050 rows, where each row corresponds to an Ensembl ID, and 38 columns representing the individual samples, including 22 control and 16 disease samples.

3.2 Methylation

The methylomics data contains a count matrix obtained using methyl-binding domain sequencing (MBD-seq). The MBD-seq workflow consists of DNA fragmentation, capturing of DNA with methylated CpGs using highly proteins with high affinity, elution and the generation of barcoded sequenceing libraries [4]. These libraries then get pooled, sequenced and aligned [4]. The Illumina HiSeq 2500 platform (GPL16791) was used. The dataset used in this project was obtained from the GEO database (GEO accession: GSE134052; Methylomics - LINK) provided by Akter et al. [5]. The dataset contains 77 samples, which are represented by the columns of the matrix, and 3088281 chromosomal regions, which form the rows of the matrix. Read counts are the number of aligned reads that uniquely map to the hg38 reference genome [5].

4 Methods

4.1 Differential Expression Analysis

For differential expression analysis and RNA-seq dataset annotation, the R programming language (`R version 4.3.2`) was utilized.

For data processing, the DESeq2 library was utilized for differential expression analysis of the RNA-seq data [33]. Sample information was manually added by creating a text file containing details such as sample name, ID, replicate number, and group classification (control vs. endometriosis).

Before proceeding with data preprocessing, we wanted to check for major differences in count distribution across all 38 samples, with a particular focus on comparing disease versus control samples (Fig. 10).

A filtering step was performed to retain only rows (transcripts), where the sum of counts was greater than or equal to 5. After filtering, there were 29,606 remaining Ensembl IDs out of the initial 58,050. To normalize for sequencing depth and RNA composition, we used DESeq2's median of ratios [6]. Following normalization, Log2 Fold Change (LFC) Shrinkage was applied to avoid poor ranking of genes by effect size and for better LFC estimates. Typically, genes with lower mean expression values exhibit greater variability in log2 fold change compared to those with higher mean expression values. LFC shrinkage borrows information across all genes to shrink the raw log2 fold change estimates towards zero. This typically results in more stable and reliable estimates, particularly for genes with low counts or high dispersion. Therefore, raw log2 fold change estimates might be noisy and unreliable for some genes.

The proposed shrinkage method, Approximate Posterior Estimation for generalized linear model, apeglm, is recommended by [47]. The LFC shrinkage has been visualized using DESeq2's default function to generate an MA-plot (M: log ratio; A: mean average) (Fig. 2). Data points with extreme values along the y-axis (M) indicate high gene expression levels [32]. This is visible by looking at data distribution before shrinkage was applied. This fanning effect was minimized after shrinkage (Fig. 2). By default, DESeq2's MA plot function colors data points with an adjusted p-value (padj) less than 0.01 in blue.

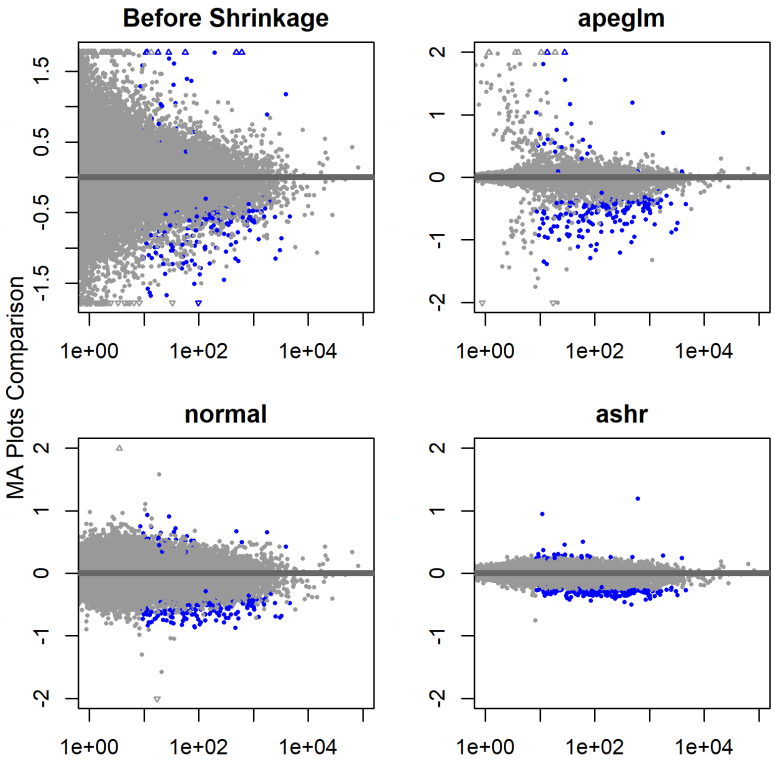


Figure 2: LFC Shrinkage stabilizes raw log fold change estimates, shrinking them towards zero. This results in more reliable estimates for genes with low counts or high variability. The MA-Plot visualizes log fold changes against mean expression levels, with significant genes highlighted in (p.adj <0.01). The apeglm method was utilized for further analysis

.

Multiple testing correction using the BH (Benjamini-Hochberg) method to control false discovery rate (FDR) is already pre-implemented in DESeq2's default workflow. Differentially expressed genes (DEGs) were identified by filtering results based on an adjusted p-value threshold of <0.05.

Annotation was performed employing Bioconductor packages org.Hs.eg.db [36] and biomaRt [8]. For the RNA-seq dataset, the timing of annotation did not make a significant difference. By utilizing biomaRt, we were able to identify a significantly larger number of gene symbols. This annotated results table was subsequently used for functional enrichment analysis and the machine learning pipeline.

4.2 Differential Methylation Analysis

For the annotation of the methylation data and the differential methylation analysis, we used the programming language R (`R version 4.3.2`). The methylation data was

first filtered to remove rows which had no counts across all samples. We build a data frame containing the chromosome name, start and end position in order to query the "hsapiens_gene_ensembl" dataset with the biomaRt package [8] and org.Hs.eg.db [36]. The resulting output contained the Ensembl IDs, external gene names, chromosome number, start and end position of the genes which are represented by the regions in the dataset. The mapping of the regions into the ensembls was done using two left joins, one for the starting position and one for the end of each gene in the query output chromosome wise in a for-loop. The starting and end positions of the query output all lied wholly inside at the starting and ending position of the mapped chromosome, so as a next step duplicate rows were removed, which contained the same regions, IDs and identical counts across all samples. Some IDs were found for multiple regions, therefore these specific cases were checked for an overlap between their start and end positions to check whether these regions are overlapping. If overlap had been found, this would not have been easily possible as to not count read counts twice. Fortunately, no overlap could be detected and in the rows containing identical IDs were merged to sum up their counts. This cut the dataset down to 62,819 rows/ensembls. In order to continue with the differential methylation analysis, an information file for the samples was collected and curated manually from the GEO database in Excel. The file included the link to the database IDs, sample names, group (diseased or control) and replicate numbers. A boxplot depiction of the expression of all 62,819 methylated (log transformed for better readability) regions for the 77 samples can be found in the appendix (Fig. 11).

After having mapped the chromosome regions and annotated the genes, the methylation data was prepped for the differential analysis workflow. Several packages exist for performing differential methylation analysis. Each package has pros and cons and may be tailored to the specific type of methylation data, such as array-based data versus bisulfite sequencing methods. A limitation with package selection was the input data format. Many packages such as the MEDIPs (Methylation Data Input Processing System) R package require raw sequencing reads [10]. As the methylation data used for this project was already preprocessed into a count matrix, this was difficult. The R package edgeR was decided upon for the workflow as it accepts count matrices as input similar to DESeq2. While edgeR was primarily developed for RNA-seq differential analysis, it has been documented to also be applicable to methylation data for differential analysis, such as in this case [12]. The source data paper also utilized edgeR for their differential methylation analysis [5]. The edgeR workflow was carried out by first creating a DGEList object storing the count data as well as the sample data, which had been previously extracted. The data was then normalized using the normalization method, trimmed mean of M values (TMM), in order to normalize read counts among different samples, which was also successfully performed in the Akter et al. source paper [5]. The edgeR workflow then estimated the dispersion, or measure of the variability of counts. In similar fashion to the Akter et al. paper, a generalized linear model (GLM) was applied followed by a likelihood ratio test [5]. From there, the top differentially methylated regions (which had been annotated to genes) could be extracted. In order to accurately interpret result significance, an adjusted p-value cutoff set at 5 percent using the Benjamini and Hochberg false discovery rate (FDR) method for multiple testing was implemented [7].

4.3 Functional Enrichment Analysis

The significant results from the previous differential expression analysis and differential methylation analysis were then used in various functional enrichment analysis (FEA) tools. Before beginning the queries, the Ensembl IDs were transformed to Entrez IDs. Unfortunately, some input gene IDs fail to map, leaving 34 DEGs and 18 DMRs to query ontologies and databases. Functional enrichment analysis associates a collection of experimentally gained gene lists to gene-set libraries, which organize accumulated knowledge about the function of groups of genes [11]. Using this association, inferences can be made to the molecular mechanisms and biological processes underlying the experimentally gained gene lists [21]. Usually a p-value, adjusted for multiple testing is computed, assuming the independence for the probability of any gene belonging to any set [3].

Over-represention analysis (ORA) methods were used in this project, which use Fisher's exact and hypergeometric distribution tests to determine significance of list overlap between this project's previous analysis of unranked gene identifiers and the query gene sets [1].

Commonly used Gene Set Enrichment Analysis (GSEA) tools include GO Term Enrichment Analysis and KEGG (Kyoto Encyclopedia of Genes and Genomes) Pathway enrichment [43]. KEGG (Kyoto Encyclopedia of Genes and Genomes) is a comprehensive database that includes genomic, biochemical, and phylogenetic information for systematic gene function analysis [43]. One key feature is the enrichment analysis using KEGG pathways, which are collections of manually curated pathway maps that illustrate molecular interactions, reactions, and relational networks [30].

Gene Ontology (GO) encompasses both an ontology, which defines terms and their relationships, and annotations linking genes to these terms. GO terms are categorized into Biological Processes (BP), Molecular Functions (MF), and Cellular Components (CC). GO enrichment analysis aims to identify enriched GO terms among differentially expressed genes [22].

The tools tested in this project's FEA are listed in Tab. 1. Except for g:Profiler and the GOstats package, all methods employ the Benjamini-Hochberg correction for multiple testing. For the GOstats package, which utilizes a hypergeometric test [26], the Bonferroni correction was applied for multiple testing. On the other hand, g:Profiler uses a custom approach called g:SCS, which corresponds to an experiment-wide threshold of a=0.05 [44]. The algorithm then calculates a threshold t, which is dependent on query list size:

> *"Given a fixed input query size, g:SCS analytically approximates a threshold t corresponding to the 5% upper quantile of randomly generated queries of that size. All actual p-values resulting from the query are transformed to corrected p-values by multiplying these to the ratio of the approximate threshold t and the initial experiment-wide threshold a=0.05."* [44].

Additionally, the GeneMANIA app from cytoscape [23] was tested, which was used by the paper this project sourced its data from. GeneMANIA can be used to predict gene functions based on gene interactions based on a database of genomics and proteomics data and shows available enriched Gene Ontology categories [20]. Unfortunately, results were unclear and difficult to interpret, therefore these results will not be included in this report.

Tool	Platform	Gene set
clusterProfiler[13]	R	all GO Terms[1], KEGG
GOstats[24]	R	GO [1]
Enrichr	Online	all (GO [1], curated versions of Reactome, KEGG, etc.)
g:Profiler	Online	GO [1], curated versions of Reactome, KEGG and WikiPathways
ReactomePA [2] [38]	R	Reactome pathway database
DOSE - Disease over-representation [18]	R	Disease Ontology

Table 1: Overview FEA tools

[1]GO Terms: Biological Process, Molecular Functions and Cellular Components [22].

[2]The ReactomePA package uses the clusterprofiler functions to connect to the Reactome database.

4.4 Machine Learning

The machine learning portion of this project was conducted using the Python programming language (`Python 3.11.5`). The first machine learning model was designed around the concept of averaging class probabilities from two separate estimators. These two estimators were both logistic regression models, one fitting on the RNA-seq data, and one on the methylation data. Both data sets consisted of count matrices of patients IDs vs. gene symbols, where only the most significant genes determined from the differentials were used. An original 70:30 train-test split was made on both datasets. The training portion of the splits were then passed to the initial estimators. The scikit-learn GridSearchCV function was implemented in each estimator respectively, performing a 5 fold cross validation and testing over different logistic regression hyperparameters [16]. These hyperparameters consisted of the solver ('lbfgs', 'liblinear'), penalty ('l1', 'l2'), and c values or penalty strength (0.01, 0.1, 1, 10, 100). In each cross validation test over the given hyperparameters, an accuracy score was calculated. The hyperparameters set with the highest accuracy were saved as the best estimator. For the RNA-seq logistic regression model best estimator, the hyperparameters solver: 'liblinear', penalty: 'l2', and C: 0.01 were used to achieve a training accuracy of 0.79. The methylation model used solver: 'lbfgs', penalty: 'l2', and C: 1, to achieve a training accuracy of 0.82.

Both estimators were fit on the training data. The thus far untouched test data, was then used to create predictions from both models, to ensure no training dependencies. The scikit function `predict_proba()` was used to extract the probability predictions from each model of classifying the test patient as '0' healthy or '1' diseases based on the input data [16]. Since the disease probabilities are of interest, only these probabilities were passed to the meta model. The meta model consisted of a simple function to calculate the average probabilities of the patient being classified as '1', or of the endometriosis disease state, from each of the individual RNA-seq and methylation logistic regression estimators. From this average probability it was determined that a patient was classified as healthy if the probability fell below 0.5, and diseased if greater than or equal to 0.5. Finally, metrics for the two individual models, as well as the meta model, were collected using scikit-learn built-in functions and simple calculations [15]. Final metrics included accuracy, specificity, recall, precision F1 score, and area under the curve, with confusion matrices and ROC curves.

For the second machine learning model, both the significant DEG and DMR datasets were combined after renaming the features (Ensemble IDs) according to the dataset it belonged t and filtering for overlapping samples. In order to make the datasets more comparable, the individual columns were scaled to a range between 0 and 1 individually using the MinMaxScaler() function from scikit-learn, since the RNA dataset had much higher counts than the methylation dataset. This transformation does not nullify outliers, but linearly scales the whole column so that the largest value is corresponding to 1 and the lowest to 0 [34]. Then the combined dataset was split into training and test data using a 70:30 split. Only the training data was used to train the models. Three individual classifiers were implemeted from the scikit-learn package: a logistic regression, a decision tree and a support vector machine (SVM). Their best estimators were combined by the stackingClassifier function from scikit-learn. For each classifier, a grid search was performed for hyperparameter optimization using cross validation of 10 splits. The grid search for the logistic regression consisted of the same hyperparameters as in model 1, beeing solver ('lbfgs', 'liblinear'), penalty ('l1', 'l2') and C (0.01, 0.1, 1, 10, 100). The best estimator consisted of C: 10, penalty: 'l1' and solver: 'liblinear', which led to a training accuracy of 0.97. For the decision tree the grid search included the number of features to consider when looking for the best split max_features ('auto', 'sqrt' or 'log2'), a parameter for Minimal Cost-Complexity Pruning ccp_alpha (0.1, .01 or .001), the maximum depth of the tree max_depth (5, 6, 7, 8, 9) and the function to measure the quality of a split criterion('gini', 'entropy') [14]. The best tree had a training accuracy of 0.68 with the hyperparameters of ccp_alpha: 0.1, criterion: 'entropy', max_depth: 5, max_features: 'auto'. For the SVM the grid search included the regularization parameter C (0.1, 1, 10, 100, 1000), the kernel ('rbf', 'sigmoid', 'poly', 'linear') and a kernel coefficient if the kernel is either coefficient for 'rbf', 'poly' or 'sigmoid' (1, 0.1, 0.01, 0.001, 0.0001) [41]. The estimator had a training accuracy of 0.85 with the hyperparameters of C: 10, gamma: 0.01, kernel: 'rbf'. The output of the best models established by the grid search were then combined into an ensemble model using the stackingClassifier function from scikit-learn. These outputs were then classified again by a logistic regression using 5 fold cross validation. Each model as well as the ensemble model were tested using the test data. We computed the test accuracy, specifity, recall, precision, F1-score and AUC scores for this model as well as confusion matrices and plotted a ROC curves. We included all plots in the appendix. The result table (Tab. 4 includes AUC values calculated with a different function, the auc function instead of the roc_auc_score function. The roc_auc_score function compute the AUC directly from true binary labels and predicted scores or probabilities [17], which was not possible for the SVM classifier as no probabilities are calculated. In order to keep the results comparable, the AUC values were recalculated using the auc function which worked for all classifiers, because it uses the false positive rates and true negative rates to calculate the AUC values [17].

5 Results

5.1 Differential Expression Analysis

From the differential expression analysis, we identified a total of 100 significant differentially expressed genes (DEGs) with an adjusted p-value threshold of padj < 0.05. Among these **100 DEGs** 83 genes exhibited downregulation in endometriosis samples compared to healthy controls (LFC <0). 17 genes were upregulated in endometriosis samples (LFC >0). Furthermore, during annotation using biomaRt, we found that 28 of the identified DEGs missed available gene symbols, suggesting they may correspond to novel transcripts.

The volcano plot illustrates the final results of our differential expression analysis (Fig. 3). The x-axis represents the log2 fold change (LFC) in gene expression between endometriosis samples and healthy controls. Positive values (LFC >0) indicate upregulation in endometriosis samples, while negative values (LFC <0) indicate downregulation. The y-axis represents the negative log10 of the adjusted p-value. Higher values on the y-axis correspond to lower adjusted p-values.

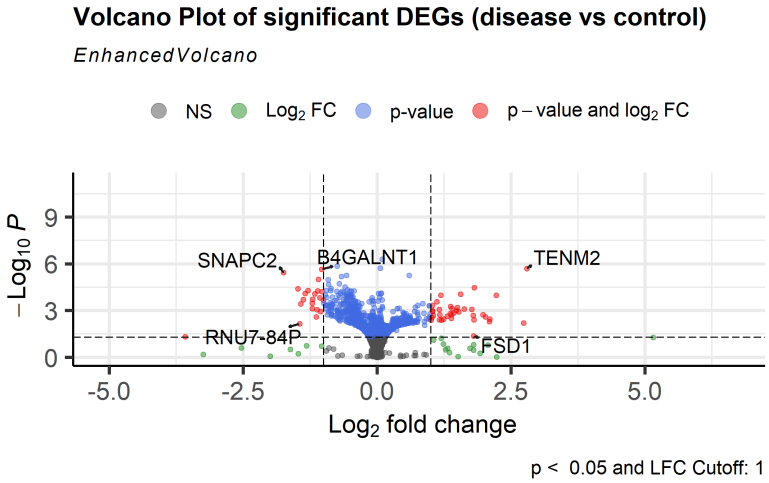


Figure 3: The volcano plot illustrates the final results of the differential expression analysis, where the x-axis represents log2 fold change (LFC) and the y-axis represents the negative log10 of the adjusted p-value. Gray dots indicate genes that are not significant with abs(LFC) <1, red dots highlight significant genes with abs(LFC) >1 and padj <0.05, blue dots show significant genes with abs(LFC) <1, and green dots represent non-significant genes with abs(LFC) >1

.

- Gray Dots: Represent genes that are not significant and have an absolute LFC <1. These genes do not show significant differential expression and therefore are not considered interesting.

- Red Dots: Represent genes with abs(LFC) >1 and padj <0.05. These genes are both statistically significant and have a larger change in gene expression. Therefore, they are of high interest due to their significant and large expression changes.
- Blue Dots: Represent genes with padj <0.05, but abs(LFC) <1. These genes are statistically significant but have smaller changes in expression.
- Green Dots: Represent genes that are not statistically significant but have an absolute LFC >1.

Additionally, we analyzed the top 10 genes based on the highest absolute log2 fold change (LFC) (Fig. 4). The boxplot provides a comparative look at count distributions between endometriosis and control samples for these genes.

Three out of these 10 genes are novel transcripts. All genes showed distinct differences in median count distribution between endometriosis and control samples. These genes include **COL1A1, FBRSL1, TENM2, IGF2, CYSRT1, LINC02593 and KLF2P1**. Genes such as TENM2, KLF2P1, and all identified novel transcripts showed upregulation, while the remaining genes showed downregulation in endometriosis samples compared to healthy controls. TENM2 showed the highest absolute LFC (2.79), followed by KLF2P1 (2.22).

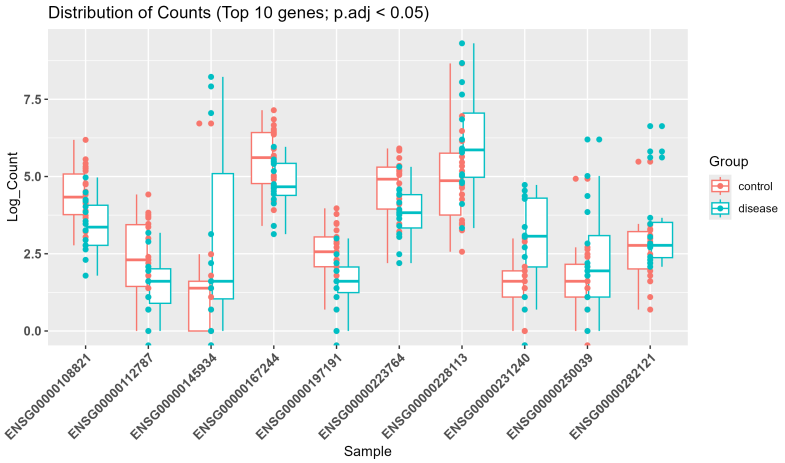


Figure 4: Boxplot illustrating the expression differences of the top 10 differentially expressed genes (DEGs) based on the highest absolute LFC. Three of these genes are novel transcripts. All genes show distinct differences in median count distribution between endometriosis vs control samples.

5.2 Differential Methylation Analysis

From the differential methylation analysis, there were 27 significant differentially methylated regions (DMRs) with an adjusted p-value threshold of padj < 0.05 identified in total. Amongst the **27 DMRs**, 11 regions exhibited hypomethylation in

endometriosis samples compared to healthy controls (LFC <0), while 16 regions were hypermethylated in endometriosis samples (LFC >0). Of these 27 identified DMRs, only 10 had gene symbols extracted from the previous annotation. Many of the unannotated regions corresponded to novel transcripts, such as long non-coding RNAs (lncRNAs). Unfortunately, none of these novel transcripts were found to be associated with endometriosis through literature searches.

From these DMRs a volcano plot was generated (Fig. 5).

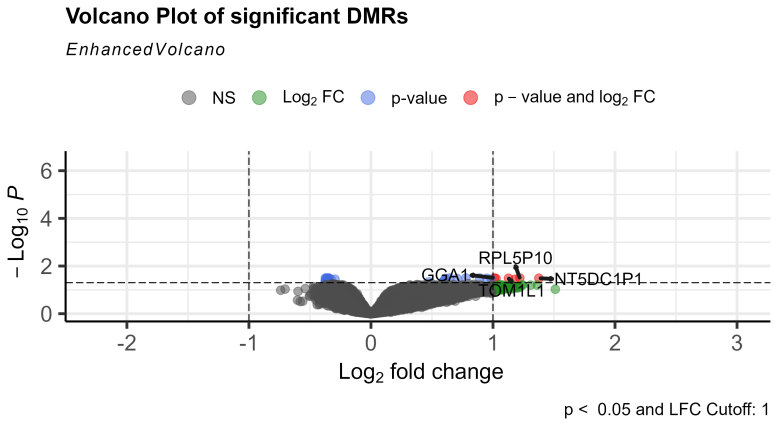


Figure 5: Volcano plot illustrating the final results of the differential methylation analysis. The x-axis represents log2 fold change (LFC) and the y-axis represents the negative log10 of the adjusted p-value. Gray dots indicate genes that are not significant with abs(LFC) <1, red dots highlight significant genes with abs(LFC) >1 and padj. <0.05, blue dots show significant genes with abs(LFC) <1, and green dots represent non-significant genes with abs(LFC) >1

.

The plot represents the result table accurately as most of the DMRs are grey, not passing the cut-off of abs(LFC) >1 and p-Value <0.05. Blue DMRs pass the p-value cutoff only and green the LFC cut off respectively. Red DMRs pass both cut offs, which are **GGA1, TOM1L1, RPL5P10 and NT5DC1P1**.

For the DMR's the top 10 genes were also analyzed based on the highest absolute log2 fold change (LFC) (Fig.6) to compare the read count distribution between the endometrios and control samples. The read counts were log transformed for better readability of the plot. All DMRs show clear differences in median count distribution, with the endometriosis samples having higher counts, therefore showcasing hypermethylation in these gene regions. The top 10 regions include two unnamed novel transcripts, two unnamed pseudogenes as well as the gene regions of **GGA1, TOM1L1, RN7SKP54, RPL5P10, NT5DC1P1 and KRTAP4-16**. Out of these regions, NT5DC1P1 has the highest absolute LFC (1.38) followed by RPL5P10 (1.22).

A literature search for the associated genes did not bring forward relevant literature about endometriosis or their methylation status regarding the disease.

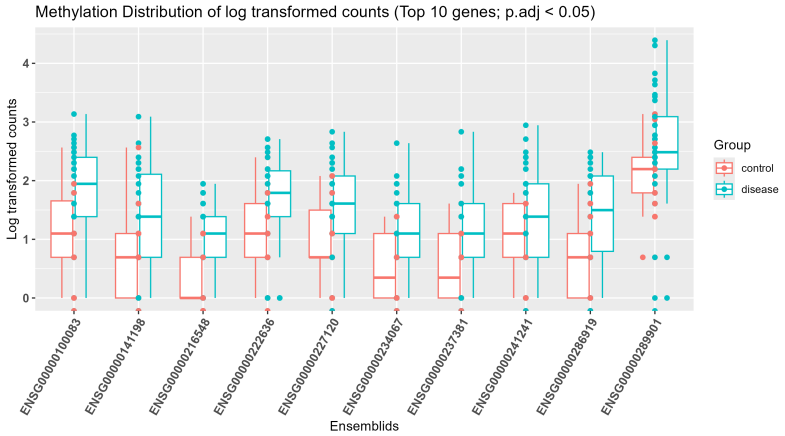


Figure 6: Boxplot illustrating the log transformed expression differences of the top 10 differentially methylated regions (DMRs) based on the highest absolute LFC. Two of these regions belong to novel transcripts and two are pseudogenes. All regions show higher median count distribution for the endometriosis samples when comparing the diseased vs control group.

.

5.3 Functional Enrichment Analysis

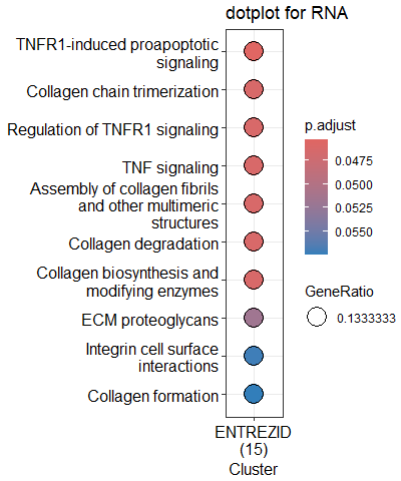
Despite testing multiple packages and libraries for functional enrichment analysis, initial findings were disappointing. After applying multiple test corrections, nearly all results remained statistically non-significant.

As previously noted in section 4.3, widely employed Gene Set Enrichment Analysis (GSEA) tools encompass GO Term Enrichment Analysis and KEGG Pathway enrichment.

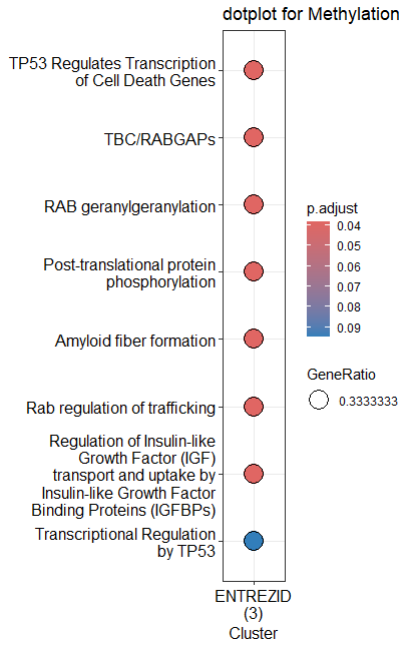
The KEGG analysis showed only one significant result (hsa04974 - protein digestion and absorption), while the GO enrichment analysis did not produce any significant findings, with the lowest adjusted p-value being 0.1561283 (GO:0000101 - sulfur amino acid transport).

The Enrichr and g:Profiler online tools unfortunately did not yield any biologically significant results for any of its libraries.

However, despite initial disappointment the Bioconductor package ReactomePA for reactome pathway analysis was extremely helpful [45], the results can be seen in Fig. 7a and Fig. 7b. The compareCluster() function clusters the genes of interest, which is represented by the bracketed number at the bottom of the plot. On the left site pathways are listed where overlap between this project's experimental lists and the pathway gene lists could be found. The color gradient from red to blue of the dots represents the significance of the overlap, where red marking overlap of a smaller adjusted p-value. The dot size depicts the extent of gene overlap between the query list and the reactome pathway lists (GeneRatio).



(a) Dot plot of the reactome FEA for the significant DEGs. The dot coloring represents the overlap significance of the adjusted p-value between the DEGs and the pathway gene lists. The dot size corresponds to the ratio of overlapping genes between the lists.



(b) Dot plot of the Reactome FEA for the significant DMRs. The dot coloring represents the overlap significance of the adjusted p-value between the DMRs and the pathway gene lists. The dot size corresponds to the ratio of overlapping genes between the lists.

28 of the 34 queried DEGs could be mapped by the Ractome package, which is shown in Fig. 7a at the bottom. We found multiple significant results for collagen pathways which have an overlap of 2-3 genes, mainly COL1A1 and COL6A2. Another interesting gene listed is TNRF1. In the literature, there is a possible connection between TNRF1 and endometriosis mentioned by Salmeri et al. [39]. A downregulation of TNRF1 can be associated with early stages of endometriosis. Unfortunately, TNRF1 itself was not one of the DEGs. The genes from this project's data were involved in that pathway were RNF31 and MIB 2. They are downregulated in the methylation dataset, but no literature further linking these genes to endometriosis was found. A gene of interest for the mentioned collagen pathways is COL1A1. Zheng et al. [46] found overexpression of COL1A1 in perilesional (tissue around endometriosis lesions) and lesional tissues. The authors found that the hyperacetylation of the COL1A1 promotors in lesions resulted in activated gene expression and appearance of de novo progressive fibrotic scarring. Unfortunately, in the differential expression results, COL1A1 was a downregulated DEG in patients from the diseased group compared to the control group.

> " The data from these allograft experiments suggests that lesion-specific diminished KLF11 levels and transcriptional dysregulation of Col1a1 were critical for disease progression." [46]

This might link to the downregulation TNRF1 and possibly RNF31 as well as MIB2, but no further literature could be found.

For the DMRs Insulin-like Growth Factors (IGF) were found and the gene IGFBP3 for this pathway. For IGFBP3 the work of Kai et al. found a possible link in it's upregulation to the pathophysiology of endometriosis and its possible involvement in the ectopic growth of endometriotic lesions [29]. Further information on IGFBP3s methylation status however was not available. GGA1 is another gene found in 3 out of 8 significant pathways, but no literature could be found regarding possible involvements in endometriosis. The last gene found in reactome was RAB40C. Unfortunately further literature for this FEA results could not be found as well.

The Disease Ontology Semantic and Enrichment Analysis did not find any significant disease ontologies for the DEGs after adjusting for multiple testing (adj. p ¿ 0.05). Out of the 18 DMRs that could be mapped to Entrez Ids two were enriched in the pathways listed in Tab. 6.

Description	GeneRatio	Adjusted p-Value	Gene Names
Focal Segmental Glomerulosclerosis	2/5	0.031	IGFBP3, HR
Glomerulosclerosis	2/5	0.031	IGFBP3, HR
Acromegaly	1/5	0.059	IGFBP3
Epidermolysis Bullosa Dystrophica	1/5	0.059	IGFBP3
Congenital Disorder of Glycosylation	1/5	0.059	IGFBP3
Sotos Syndrome	1/5	0.059	IGFBP3

Table 2: Results from the Disease Ontology Semantic and Enrichment Analysis results including ontologies, ratio of overlapping genes (overlap between the DMRs represented by the ontology and all DMRs in the query), adjusted p-value and external gene names.

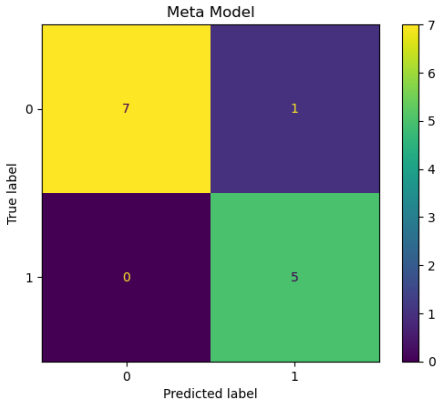
We conducted a literature research for the results and could not find a direct link between glomerulosclerosis and endometriosis. "Focal segmental glomerulosclerosis" is used to describe either a disease caused by damage to kidney podocytes or it is used to describe lesions occuring in any chronic kidney disease (CKD) [19]. We could find literature suggesting that endometriosis is inversely associated with CKD in women in Taiwan, which does mention glomerulosclerosis [27]. The described effect was however mediated by menopause [27]. HR and IGFBP3 are the DMRs involved in the found glomerulosclerosis onthologies from this project's dataset. For HR literature linking it to endometriosis was discovered. For IGFBP3, as mentioned in the result section of the functional enrichment analysis, Kai et al. found a possible link of its upregualtion to the pathophysiology of endometriosis and its possible involvement in the ectopic growth of endometriotic lesions [29]. Information on its methylation status however was not available.

5.4 Machine Learning

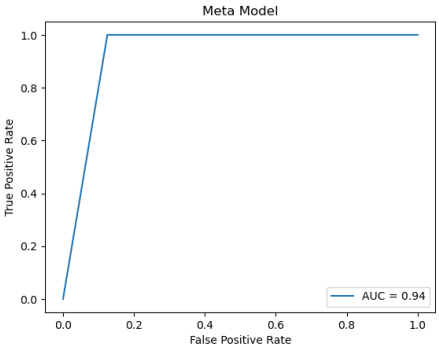
As previously mentioned, metrics and scores were calculated to quantify the quality of predictions made by the machine learning models to predict the disease state of a patient. These metrics have been summarized in Tab. 3. Referencing the table, it can be seen that the meta model outperformed the individual RNA and methylation logistic regression estimators. Note that for these metrics, a score of 1.0 is most ideal. The meta model scored higher in terms of both testing accuracy and area under the curve (AUC), measuring the model's ability to discriminate between positive and negative classes. It ties in performance with the RNA model for specificity, and the methylation for recall, while outperforming both in its precision scoring. In balance of precision and recall with the F1 score, the meta model also predominated. These results are better visualized in Fig. 8 (Note visualizations for the individual estimators can be found in Fig. 12 and Fig. 13 in the appendix). In the confusion matrix (Fig. 8a) the dark colors indicating low counts for false positives and false negatives show good discrimination, while the light colors indicating higher accounts for true positives and true negatives show good performance. In the Receiver Operating Characteristic (ROC) curve (Fig. 8b) an almost perfect "L" shape is seen indicating the model distinguishes well between positive and negative instances, backed-up by the high AUC calculation at a 0.94 out of an ideal 1.0 [25].

Classifier	Test Accuracy	Specifity	Recall	Precision	F1-score	AUC
RNA Logistic Regression	0.62	0.38	1.0	0.50	0.67	0.69
Methylation Logistic Regression	0.77	0.88	0.60	0.75	0.67	0.74
Meta Model	0.92	0.88	1.0	0.83	0.91	0.94

Table 3: Test accuracy, specifity, recall, precision, F1-score and AUC scores for model 1. The used classifiers include an RNA logistic regression, a methylation logistic regression, and the combined "meta model".



(a) Confusion matrix for the metamodel where '0' defines a healthy patient and '1' defines disease state i.e. endometriosis.



(b) ROC curve projecting the false positive rate vs. true positive rate for the metamodel. Here, an area under the curve (AUC) of 0.94 is shown.

Figure 8: Confusion Matrix and ROC Curve for Model 1: Meta Model

To further analyze the machine learning model and what features or genes played an importance in prediction, the beta values of the individual logistic regression models were considered. Beta values are the coefficient values in the logistic regression equation and represent the relationship between the dependent and independent values [9]. A stronger or more important relationship is therefore signified by a larger beta value. For this reason, the top ten absolute beta values for the RNA and methylation logistic regression models respectively were extracted and their corresponding annotated genes were researched to look for pre-existing links to these genes and endometriosis. A few interesting correlations were found.

The first gene with a high beta value from the RNA logistic regression model of interest was KLF2P1. KLF2P1 was found to be more prevalent in the disease group during initial data exploration, possibly indicating an association with the disease state. In correlation with this indication of KLF2P1's significance in endometriosis prediction, the source paper from which the data was originally taken, also used KLF2P1 in three of the decision tree models [5]. Unfortunately, beyond this no literature thus far has suggested a link between KLF2P1 and endometriosis. A second interesting find was SCAF1, which has been noted as a part of mitochondrial super complex assembly expressed abundantly in clinical breast and endometrial cancers. While this is of interest as some symptoms of endometrial cancer and endometriosis coincide, it should be noted the two remain distinct diseases [28]. The third gene of significance, IGF2 or insulin growth factor 2, had the opposite trend of KLFP21 and was found to be less prevalent in the disease group during data exploration. Previous research has shown that dysregulation of IGF2 may facilitate endometriosis predisposition [31].

Two genes of interest were found through literature searches of the top methylation logistic regression beta values. The first CDCA2, showed a positive log fold change i.e. hypermethylation in the endometriosis group in a study by Naqvi et al., consistent with

this study's findings [35]. The second, pseudogene RPL37AP1, showed an opposite trend found to be hypomethylated in methylated regions of interest (MROI) in the source paper by Akter et al., also consistent with this study's outcomes. Similarly to KLF2P1, this pseudogene was a feature extracted from the source study's methylation decision tree models [5].

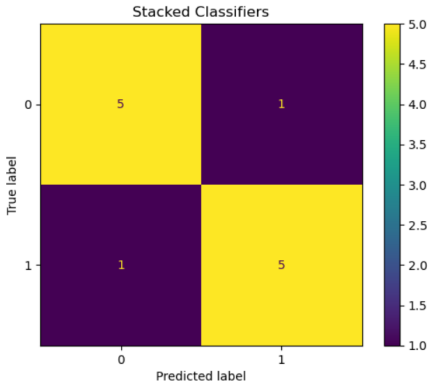
For the second approach, the stacked model or model 2, the results, including test accuracy, specifity, recall, precision, F1-score and AUC scores, for all classifiers are shown in Tab. 4 (the confusion matrices and ROC curves of the individual classifiers can be found in the appendix).

Classifier	Test Accuracy	Specifity	Recall	Precision	F1-score	AUC
Logistic Regression	0.75	0.67	0.83	0.71	0.77	0.75
Decision Tree	0.42	0.67	0.17	0.33	0.22	0.42
Support Vector Machine	0.75	0.83	0.67	0.80	0.73	0.75
Stacked Classifier	0.83	0.83	0.83	0.83	0.83	0.83

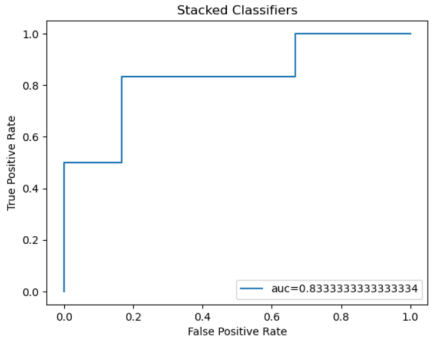
Table 4: Test accuracy, specifity, recall, precision, F1-score and AUC scores for the stacked model. The used classifiers include a logistic regression, a decision tree and a support vector machine which were then also stacked to a forth ensemble classifier named "stacked classifier".

Out of the individual classifiers before stacking, one can see that the decision tree performs the worst across all metrics while the logistic regression and the SVM perform solidly over all. The stacked classifier outperforms all individual "weak learners", scoring a 0.83 over all metrics, evening out the individual models weaknesses such as the recall of the SVM and the specificity of the logistic regression. However, a better result might have been possible, as the decision tree results hinder a better classification result. The classification results of the stacked model are depicted in Fig. 9, on the right-hand side the confusion matrix and on the left the ROC.

The confusion matrix shows a good ratio for true positives and true negatives, as 10 from 12 patients were correctly classified. The good discrimination between the classes is also visible in the coloring of the tiles, which are very distinct from each other.The ROC curve of the stacked classifier has a stepped appearance, showing a good trade-off between the true positive and true negative rate. Overall the model does perform well taking all metrics into account, but is outperformed by the approach in the metamodel of the first model where the datasets are first classified separately. As already mentioned, a better result might have been possible for the second model if a more robust classifier had been used instead of the decision tree. Tho further analyze this approaches results, the beta values of the initial logistic regression of model 2 were inspected similarly as in the analysis of model 1. When combining the datasets in the logistic regression, more features are penalized, leading to much less beta values than model one. All 9 are listed in Tab. 5. At first glance, familiar gene names from previous analyses such as GGA1, KLF2P1 and IGFBP3 appear in the lower half of the table. Therefore, while these genes/regions contribute to the classification result, other features were more involved. Although GGA1 was found in almost all of the



(a) Confusion matrix for the stacked model where '0' defines a healthy patient and '1' defines disease state i.e. endometriosis.



(b) ROC curve projecting the false positive rate vs. true positive rate for the stacked model. Here an area under the curve (AUC) of 0.83 is shown.

Figure 9: Confusion Matrix and ROC Curve for Model 2: Stacked Model

projects conducted analyses, no literature linking the two endometriosis was available. KLF2P1 was also a relevant beta value in model one and in the reference paper [5]. For this gene, there also is no literature regarding endometriosis, meaning it can not be linked to the disease. For IGFBP3, as per the FEA results, literature was found linking IGFBP3 upregulation to the pathophysiology of endometriosis and its possible involvement in the ectopic growth of endometriosis lesions [29]. Further information on its methylation status, however was not available, which makes it impossible to link this result to endometriosis as well.

Feature	External Genename	Beta Value	Dataset
ENSG00000214655	ZSWIM8	-5.810	DEGs
ENSG00000210140	MT-TC	4.316	DEGs
ENSG00000116221	MRPL37	-3.753	DEGs
ENSG00000172725	CORO1B	-3.646	DEGs
ENSG00000216548	MT-TA	2.695	DMRs
ENSG00000231240	KLF2P1	2.678	DEGs
ENSG00000100083	GGA1	1.819	DMRs
ENSG00000146674	IGFBP3	1.544	DMRs
ENSG00000234409	CCDC188	-1.223	DEGs

Table 5: Beta values of the logistic regression used before stacking in model 2 for the combined dataset of significant DEGs and DMRs in descending order of absolute values.

Although other ZSWIM8, MT-TC, MRPL37, CORO1B and MT-TA have higher absolute beta values than the reoccurring DEGs and DMRs. As endometriosis is a severly understudied topic, further literature on these features was not available.

6 Discussion

This study had several limitations that if addressed could lead to improvements for future work. One of the motivations for this study was the lack of research and attention given to this debilitating disease in women, the downside to this being there is a limited amount of data and previous work available. This lack of research affects a few aspects of the study, the first of which being the sample size. While the overall sample size is sufficient, the overlap of available samples in this data set with both RNA-seq and methylation data available is relatively small. With an increased initial dataset, improvements could be made such as in the machine learning models helping to reduce any learned training dependencies and give a larger test split with a larger initial dataset. With more data and time, model validation on external datasets could have taken place to evaluate model robustness.

Additionally, higher sample size could potentially also reveal more functional enrichment results and more interesting pathways that could help to understand disease genesis or related pathways and biological processes in the disease development. The lack of research already available makes it more difficult to uncover findings, and amongst those findings, a lot of novel transcripts and genes are not well described in literature.

Bootstrapping techniques could help handle issues arising from the limited sample size by generating multiple resampled datasets. Further improvements for future work could include repeating the study with methylation arrays to look out for hyper/hypomethylated regions, and maybe additionally consider the epigenetic role in endometriosis development. Better techniques for clustering the data such as Uniform Manifold Approximation and Projection (UMAP) could better reveal potential dependencies in samples. Further integration of the different datasets could take place to better relate methylation state to gene expression, although this is a difficult task. Perhaps an additional study on different endometriosis lesions could be of interest. For example, are there tissue or lesion specific gene expression differences?

Further research could also focus on validating our findings in larger, independent cohorts or exploring the functional roles of the identified genes as potential biomarkers. Genes like KLF2P1, SCAF1, CDCA2, RPL37AP1 or IGF2 could be particularly interesting. It is still too early to claim that definitive biomarkers for endometriosis have been identified, even though genes like IGF2 and CDCA2 have been linked to endometriosis before.

Integrating other omics data, such as proteomics and metabolomics, could provide a more extensive understanding of endometriosis. It was observed that many of the 100 differentially expressed genes were downregulated. It would be interesting to investigate whether endometriosis is generally associated with gene expression suppression.

Finally, in terms of the machine learning architecture, relatively simple models were used, and perhaps better learning could be achieved with deeper learning models. While this project lays an interesting base for the exploration of endometriosis and its potential genetic link, much improvement can be made and much more research is needed to better understand this complex disease and lead to better diagnostics and therapeutics.

7 Conclusion

This project aims to reveal significant insights into the molecular mechanisms underlying endometriosis and to identify potential biomarkers for the disease. Differential gene expression and methylation analyses were performed, complemented by annotation. Functional enrichment analysis was conducted to interpret significant results in terms of biological functions, processes, components, pathways, and diseases.

Additionally, two ensemble classifiers were developed to predict the diseased versus non-diseased status in endometriosis patients and to identify important features or biomarkers for this classification. The first model was trained on separate datasets and then their probabilities were combined into a metamodel. The second model merged the datasets, trained a logistic regression, a decision tree, and a support vector machine, and then combined their best predictions into a stacked classifier.

The analysis identified multiple genes and gene regions of interest, which were compared with current literature on endometriosis. Although most of the findings were not corroborated by existing studies, the project identified possible biomarkers such as the gene IGF2 and the methylated region of CDCA2, which have been previously mentioned in literature regarding endometriosis.

This project faced several limitations, including a small sample size and limited existing research on endometriosis, particularly in the identification of potential biomarkers. Future work could benefit from a larger sample size, the application of bootstrapping techniques, and the use of more complex models, including deep learning approaches. This project highlights the critical need for further research in women's health, particularly endometriosis.

8 Author contributions

Section	Author Contributions
Abstract	Julia drafted and wrote the abstract.
Background	Caroline drafted and wrote the background of the project.
Goal	Astghik drafted and wrote the project goal.
Data	Astghik handled the RNA-seq data section; Julia wrote the MBD-seq data section.
Methods	Astghik contributed to Differential Expression Analysis and Functional Enrichment Analysis. Julia and Caroline drafted and wrote the Differential Methylation Analysis and Machine Learning sections, with Julia also contributing to Functional Enrichment Analysis.
Results	Similar task allocation as Methods was applied to the results section.
Discussion	Astghik and Caroline finalized the discussion section.
Conclusion	The conclusion section was collectively written.
Editing, Corrections, Finetuning, etc.	All authors equally contributed to reviewing, editing, correcting, and fine-tuning all report sections.
Coding	ChatGPT (for the purpose of understanding functions and packages)

Table 6: Distribution of author contributions to different sections of the report.

9 Appendix

9.1 Boxplots of count distribution (endometriosis vs control samples)

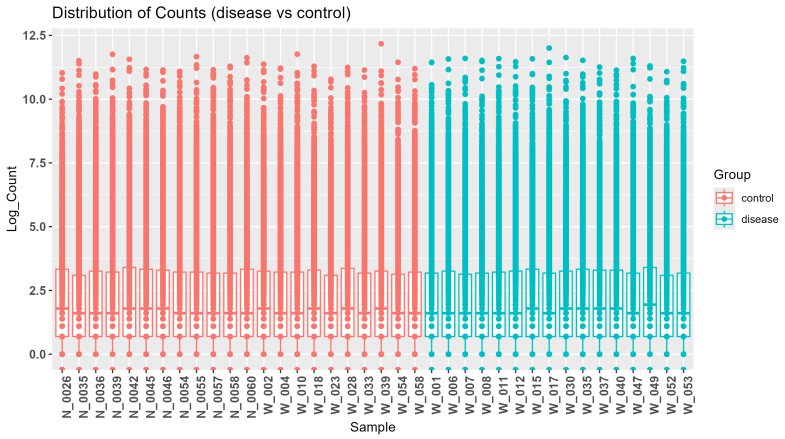


Figure 10: Boxplot illustrating the distribution of RNA-seq counts between endometriosis and control samples, showing log-transformed counts across all 38 samples. The boxplot includes a total of 29,606 genes.

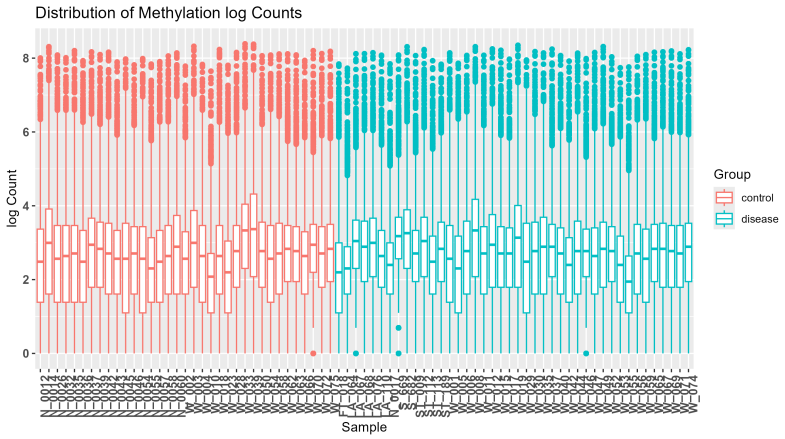
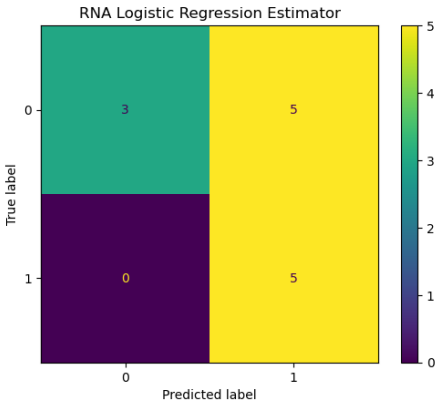


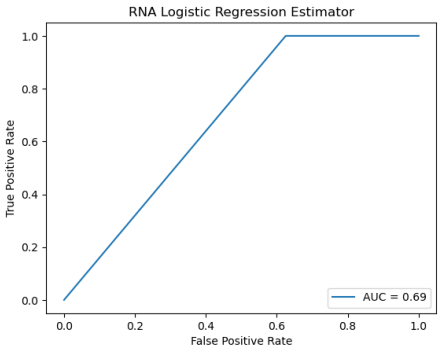
Figure 11: Boxplot illustrating the distribution of log transformed MBD-seq counts between endometriosis and control samples, showing log-transformed counts across all 77 samples (35 control vs 42 disease). The boxplot includes a total of 62819 genes.

9.2 Plots for Averaging Class Probabilities Machine Learning Model

The following figures Fig. 12 and Fig. 13 show the respective confusion matrices and ROC curves for the individual RNA and methylation logistic regression models used to construct the meta model.



(a) Confusion matrix for the RNA logistic regression model where '0' defines a healthy patient and '1' defines disease state i.e. endometriosis.

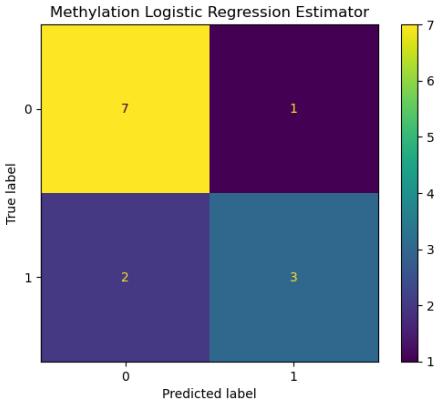


(b) ROC curve projecting the false positive rate vs. true positive rate for the RNA logistic regression model. Here an area under the curve (AUC) of 0.69 is shown.

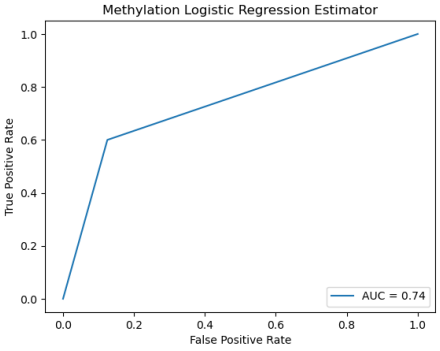
Figure 12: Confusion Matrix and ROC Curve for Model 1: RNA Logistic Regression

9.3 Plots for Stacked Machine Learning Model

The AUC values displayed in the plots of this section differ from the values in the result Tab. 4, because different functions were used to calculate them. For the SVM the roc_auc_score function does not work because the SVM does not calculate probabilities, as it is often used in tutorials and directly calculates the AUC from the true labels and predicted probabilities [17]. In order to be able to properly compare the results, the AUC was recalculated using the auc function, which requires a false positive rate and true negative rate calculation [17] and can lead to different values.

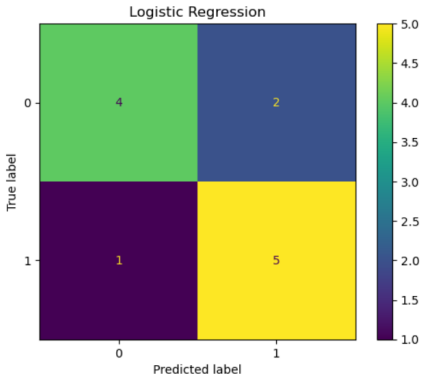


(a) Confusion matrix for the methylation logistic regression model where '0' defines a healthy patient and '1' defines disease state i.e. endometriosis.

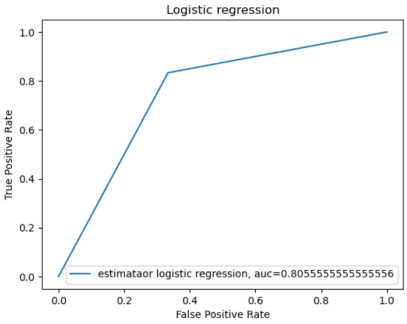


(b) ROC curve projecting the false positive rate vs. true positive rate for the methylation logistic regression model. Here an area under the curve (AUC) of 0.74 is shown.

Figure 13: Confusion Matrix and ROC Curve for Model 1: Methylation Logistic Regression

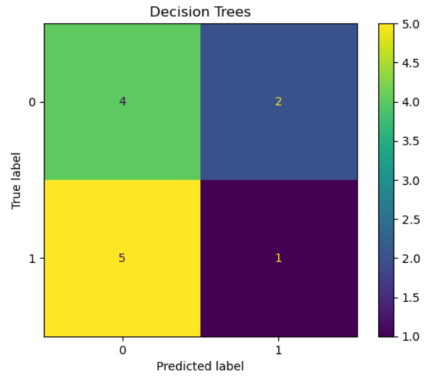


(a) Confusion matrix for the Logistic Regression. Out of the 12 patients in the test data 9 were correctly classified as healthy or diseased while 2 patients were wrongly classified as healthy and one patient wrongly classified as sick.

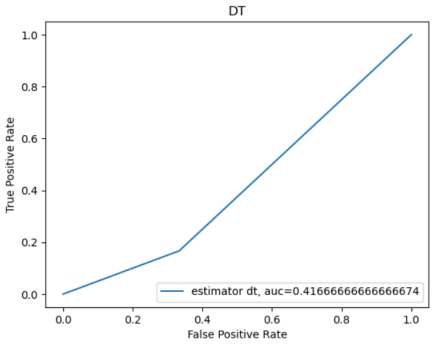


(b) ROC curve projecting the false positive rate vs. true positive rate for the Logistic Regression. The AUC for this classifier is shown as 0.81, but the calculated result of 0.75 was used in the analysis.

Figure 14: Confusion Matrix and ROC Curve for the Logistic Regression of the stacked classifier.

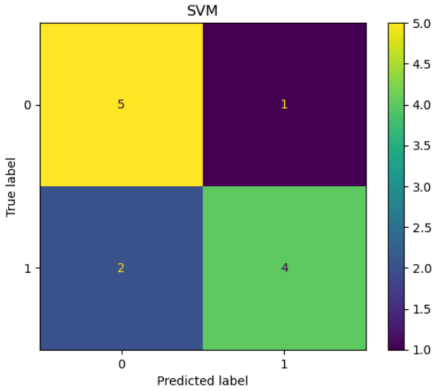


(a) Confusion matrix for the Decision Tree. Out of the 12 patients in the test data 5 were correctly classified as healthy or diseased while 5 patients were wrongly classified as healthy and 2 patient wrongly classified as sick.

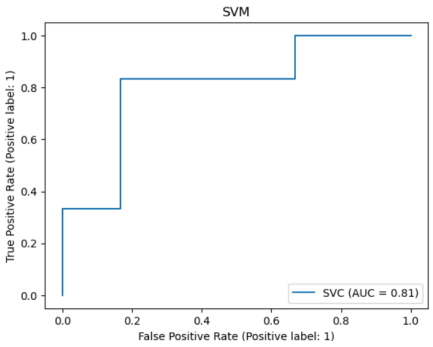


(b) ROC curve projecting the false positive rate vs. true positive rate for the Decision Tree, displaying an AUC of 0.42.

Figure 15: Confusion Matrix and ROC Curve for the Decision Tree used in the stacked classifier.



(a) Confusion matrix for the SVM. Out of the 12 patients in the test data, 9 were correctly classified as healthy or diseased while 2 patients were wrongly classified as healthy and 1 patient wrongly classified as sick.



(b) ROC curve projecting the false positive rate vs. true positive rate for the SVM with an AUC of 0.81, but 0.75 was used in this projects analysis.

Figure 16: Confusion Matrix and ROC Curve for the SVM used in the stacked classifier.

References

[1] [Online; accessed 17. Jul. 2024]. May 2024. URL: https://girke.bioinformatics.ucr.edu/GEN242/tutorials/rfea/rfea/.

[2] [Online; accessed 19. Jul. 2024]. URL: https://www.yalemedicine.org/conditions/endometriosis.

[3] URL: https://maayanlab.cloud/Enrichr/help#background&q=3.

[4] Karolina A Aberg, Robin F Chan, and Edwin J C G van den Oord. "MBD-seq - realities of a misunderstood method for high-quality methylome-wide association studies". In: *Epigenetics* (Apr. 2020). URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7153543/.

[5] Sadia Akter et al. "Machine Learning Classifiers for Endometriosis Using Transcriptomics and Methylomics Data". In: *Front. Genet.* 10 (Sept. 2019), p. 466838. ISSN: 1664-8021. DOI: 10.3389/fgene.2019.00766.

[6] Simon Anders and Wolfgang Huber. "Differential expression analysis for sequence count data". In: *Genome Biol.* 11.10 (Oct. 2010), pp. 1–12. ISSN: 1474-760X. DOI: 10.1186/gb-2010-11-10-r106.

[7] Yoav Benjamini and Yosef Hochberg. "Controlling the false discovery rate: A practical and powerful approach to multiple testing". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 57.1 (Jan. 1995), pp. 289–300. DOI: 10.1111/j.2517-6161.1995.tb02031.x.

[8] *biomaRt*. [Online; accessed 17. Juliy 2024]. May 2024. URL: https://bioconductor.org/packages/release/bioc/html/biomaRt.html.

[9] Jason Brownlee. *Logistic regression for machine learning*. Dec. 2023. URL: https://machinelearningmastery.com/logistic-regression-for-machine-learning/.

[10] Lukas Chavez. *MEDIPS*. 2014. URL: https://bioconductor.org/packages/release/bioc/html/MEDIPS.html.

[11] Edward Y Chen et al. "ENRICHR: Interactive and collaborative HTML5 Gene List Enrichment Analysis Tool". In: *BMC Bioinformatics* 14.1 (2013). DOI: 10.1186/1471-2105-14-128.

[12] Yunshun Chen. *Edger: Differential analysis of sequence read count data ...* Apr. 2024. URL: https://www.bioconductor.org/packages/devel/bioc/vignettes/edgeR/inst/doc/edgeRUsersGuide.pdf.

[13] *clusterProfiler*. [Online; accessed 17. Jul. 2024]. July 2024. URL: https://bioconductor.org/packages/release/bioc/html/clusterProfiler.html.

[14] *DecisionTreeClassifier*. [Online; accessed 19. Jul. 2024]. URL: https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html.

[15] scikit-learn developers. *3.4. metrics and scoring: Quantifying the quality of predictions*. 2012. URL: https://scikit-learn.org/stable/modules/model_evaluation.html#classification-report.

[16] scikit-learn developers. *GRIDSEARCHCV*. June 2012. URL: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html.

[17] *Different results with $roc_auc_score()$ and $AUC()$*. [Online; accessed 19. Jul. 2024]. URL: https://www.geeksforgeeks.org/different-results-with-rocaucscore-and-auc/.

[18] *DOSE*. [Online; accessed 17. Jul. 2024]. July 2024. URL: https://bioconductor.org/packages/release/bioc/html/DOSE.html.

[19] Agnes B. Fogo. "Causes and pathogenesis of focal segmental glomerulosclerosis". In: *Nature Reviews Nephrology* 11.2 (Dec. 2014), pp. 76–87. DOI: 10.1038/nrneph.2014.216.

[20] Max Franz et al. "GeneMANIA update 2018". In: *Nucleic Acids Research* (June 2018). ISSN: 0305-1048. DOI: 10.1093/nar/gky311. eprint: https://academic.oup.com/nar/article-pdf/46/W1/W60/25110212/gky311.pdf. URL: https://doi.org/10.1093/nar/gky311.

[21] Adrian Garcia-Moreno et al. "Functional Enrichment Analysis of regulatory elements". In: *Biomedicines* 10.3 (2022), p. 590. DOI: 10.3390/biomedicines10030590.

[22] *Gene Ontology*. [Online; accessed 17. Jul. 2024]. July 2024. URL: https://geneontology.org/docs/ontology-documentation/.

[23] *GeneMania*. [Online; accessed 19. Jul. 2024]. URL: https://genemania.org/.

[24] *GOstats*. [Online; accessed 17. Jul. 2024]. July 2024. URL: https://bioconductor.org/packages/release/bioc/html/GOstats.html.

[25] Karimollah Hajian-Tilaki. *Receiver operating characteristic (ROC) curve analysis for Medical Diagnostic Test Evaluation*. 2013. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3755824/.

[26] *How To Use GOstats Testing Gene Lists for GO Term Association*. [Online; accessed 18. Jul. 2024]. May 2024. URL: https://www.bioconductor.org/packages/release/bioc/vignettes/GOstats/inst/doc/GOstatsHyperG.html.

[27] Ben-Shian Huang et al. "Endometriosis might be inversely associated with developing chronic kidney disease: A population-based cohort study in Taiwan". In: *International Journal of Molecular Sciences* 17.7 (July 2016), p. 1079. DOI: 10.3390/ijms17071079.

[28] Kazuhiro Ikeda et al. "Mitochondrial Supercomplex Assembly promotes breast and endometrial tumorigenesis by metabolic alterations and enhanced hypoxia tolerance". In: *Nature Communications* 10.1 (Sept. 2019). DOI: 10.1038/s41467-019-12124-6.

[29] Kentaro Kai et al. "MicroRNA-210-3p Regulates Endometriotic Lesion Development by Targeting IGFBP3 in Baboons and Women with Endometriosis". In: *Reproductive Sciences* 30.10 (May 2023), pp. 2932–2944. DOI: 10.1007/s43032-023-01253-5.

[30] *KEGG PATHWAY Database*. [Online; accessed 18. Jul. 2024]. July 2024. URL: https://www.genome.jp/kegg/pathway.html.

[31] Hiroshi Kobayashi et al. "Genes downregulated in endometriosis are located near the known imprinting genes". In: *Reproductive Sciences* 21.8 (Aug. 2014), pp. 966–972. DOI: 10.1177/1933719114526473.

[32] Adam McDermaid et al. "Interpretation of differential gene expression results of RNA-seq data: review and integration". In: *Briefings Bioinf.* 20.6 (Nov. 2019), p. 2044. DOI: 10.1093/bib/bby067.

[33] Simon Anders Michael I. Love. *Analyzing RNA-seq data with DESeq2*. [Online; accessed 17. Jul. 2024]. Apr. 2024. URL: https://www.bioconductor.org/packages/release/bioc/vignettes/DESeq2/inst/doc/DESeq2.html.

[34] *MinMaxScaler*. [Online; accessed 19. Jul. 2024]. URL: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html.

[35] Hanyia Naqvi et al. "Altered genome-wide methylation in endometriosis". In: *Reproductive Sciences* 21.10 (Oct. 2014), pp. 1237–1243. DOI: 10.1177/1933719114532841.

[36] *org.Hs.eg.db*. [Online; accessed 17. Jul. 2024]. July 2024. URL: https://bioconductor.org/packages/release/data/annotation/html/org.Hs.eg.db.html.

[37] World Health Organization. *Endometriosis*. Mar. 2023. URL: https://www.who.int/news-room/fact-sheets/detail/endometriosis#:~:text=Overview,period%20and%20last%20until%20menopause..

[38] *ReactomePA*. [Online; accessed 17. Jul. 2024]. July 2024. URL: https://bioconductor.org/packages/release/bioc/html/ReactomePA.html.

[39] Francesca M. Salmeri et al. "Behavior of tumor necrosis factor- and tumor necrosis factor receptor 1/tumor necrosis factor receptor 2 system in mononuclear cells recovered from peritoneal fluid of women with endometriosis at different stages". In: *Reproductive Sciences* 22.2 (Feb. 2015), pp. 165–172. DOI: 10.1177/1933719114536472.

[40] AJ Shih. *Single-cell analysis of menstrual endometrial tissues defines phenotypes associated with endometriosis*. Sept. 2022. URL: https://pubmed.ncbi.nlm.nih.gov/36104692/.

[41] *SVC*. [Online; accessed 19. Jul. 2024]. URL: https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html.

[42] Yuliana Tan et al. *Single-cell analysis of endometriosis reveals a coordinated transcriptional programme driving immunotolerance and angiogenesis across eutopic and ectopic tissues*. July 2022. URL: https://www.nature.com/articles/s41556-022-00961-5.

[43] *Tools (GO & KEGG) for Gene Set Enrichment Analysis (GSEA)*. [Online; accessed 18. Jul. 2024]. July 2024. URL: https://www.novogene.com/eu-en/resources/blog/tools-go-kegg-for-gene-set-enrichment-analysis-gsea.

[44] Tartu Ülikool. *g:Profiler help - g:SCS algorithm*. [Online; accessed 17. Jul. 2024]. URL: https://biit.cs.ut.ee/gprofiler_archive2/r1760_e93_eg40/web/help.cgi?help_id=5.

[45] Guangchuang Yu and Qing-Yu He. "ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization". In: *Mol. Biosyst.* 12.2 (Jan. 2016), pp. 477–479. ISSN: 1742-206X. DOI: 10.1039/C5MB00663E.

[46] Ye Zheng et al. "Epigenetic Modulation of Collagen 1A1: Therapeutic Implications in Fibrosis and Endometriosis". In: *Biology of Reproduction* 94.4 (Apr. 2016). DOI: 10.1095/biolreprod.115.138115.

[47] Anqi Zhu, Joseph G. Ibrahim, and Michael I. Love. "Heavy-tailed prior distributions for sequence count data: removing the noise and preserving large differences". In: *Bioinformatics* 35.12 (June 2019), pp. 2084–2092. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bty895.