Staging and Classification of Biliary Atresia:

An Analysis of Disease States

Network-Based Data Analysis Report

Thomas Sirchi
Student ID: 239007
Supervisor: Prof. Mario Lauria

July 2024

UNIVERSITÀ
DI TRENTO

Abstract

Biliary atresia (BA) is the most common cause of chronic cholestasis in infants and children, necessitating liver transplantation as the primary treatment option, about 270 cases are diagnosed each year in Europe. This analysis aimed to enhance the histological classification made by the authors of a dataset comprising 47 infants with BA by using a machine learning model trained on gene expression data and conducting enrichment analysis to gain deeper insights into the disease and its various states. Unsupervised learning revealed a division that nearly matched molecular groups through UMAP and PAM clustering. Various models were tested, with Random Forest providing better results. The model successfully classified previously unclassified samples, highlighting important genes such as SLC24A1, ARF6, and DUSP5, which have been previously linked to BA. This suggests the potential for identifying new targets and emphasizes the need for further exploration of other genes. Differential expression analysis identified numerous DEGs, indicating their relevance for feature selection and suggesting that different markers might be related to various stages of the disease. Functional enrichment analysis revealed differences between groups, with DAVID and gProfiler identifying terms related to collagen organization and inflammation, mirrored by distinct clusters in the STRING analysis. pathfindR highlighted terms related to viral infections and diseases, supporting the idea that BA might have a viral or external cause. Overall, this study underscores the value of molecular insights into BA and suggests that more comprehensive studies with advanced models and additional data could enhance our understanding and guide future treatment approaches.

Contents

1 Introduction

1.1 Background

Biliary atresia (BA) is a condition affecting infants and young children, characterized by a destructive inflammatory process that impacts both the intra- and extrahepatic bile ducts. This inflammation and subsequent fibrosis lead to the progressive narrowing and eventual obliteration of these ducts, impairing the flow of bile from the liver to the intestine. The consequence is the accumulation of bile within the liver, resulting in liver damage and ultimately cirrhosis, which can be life-threatening. This condition poses a significant burden as it is the most common cause of chronic cholestasis in infants and children, necessitating liver transplantation as the primary treatment option in the majority of cases. Unfortunately, the scarcity of size-matched liver donors, particularly for the young pediatric population, presents a considerable challenge in managing this disease [1]. Despite extensive research efforts, the precise etiology of BA remains elusive. However, one prevailing theory implicates a multifactorial process involving a primary insult, possibly a perinatal viral infection affecting the hepatobiliary system, followed by a secondary autoimmune-mediated injury to the bile ducts. This hypothesis is supported by evidence from both human studies and animal models, which have demonstrated the presence of viral elements and immune activation in affected individuals. Recent advancements in understanding the pathogenesis of BA have shed light on the intricate interplay between genetic predisposition, environmental factors, and immune dysregulation in driving disease progression [2].

1.2 State of the art

Efforts to delineate the pathways of immune and autoimmune-mediated bile duct injury within BA hold promise for the development of targeted therapeutic interventions. However, achieving this goal necessitates further research to unravel the complex mechanisms underlying BA pathogenesis and to translate these findings into clinically effective treatments [3].

1.3 Objectives

This further analysis points to the finding of gene correlation between the inflammatory process and the fibrosis completing the classification process done by the authors of the dataset. Furthermore utilizing bioinformatics and machine learning techniques to analyze potential interesting genes connected to the disease.

2 Methods

2.1 Dataset Description

In the dataset, the authors gathered and analyzed data from 47 infants diagnosed with biliary atresia. Detailed clinical and laboratory data were collected, including demographic information, disease severity indicators, and liver function tests. Concurrently, liver biopsies were obtained from each infant in the study. The liver biopsies underwent histological examination, where they were scored to assess the extent of inflammation and fibrosis present in the tissue samples. Gene expression profiles were used in classifying the molecular stages of inflammation or fibrosis within the entire cohort of infants but not enough to classify all the samples. Validation of the classifications was done with immunostaining methods to quantify specific hepatic inflammatory cell populations, providing a measure of the inflammatory response in the liver tissues. The clinical relevance of these molecular classifications was assessed by examining their associations with various clinical parameters. These included traditional markers of liver function, such as serum bilirubin levels and liver enzyme activities, indicators of disease complications, surgical response to portoenterostomy, and overall clinical outcomes, including survival rates.

2.2 Unsupervised Learning Methods

Of the numerous unsupervised learning techniques available, I have chosen to implement: Principal Component Analysis (PCA). Uniform Manifold Approximation and Projection (UMAP) a nonlinear dimensionality reduction technique designed to capture local and global structures in high-dimensional data [4]. Partition around Medoids (PAM) a robust clustering algorithm that employs medoids to partition a dataset into clusters based on dissimilarity measures. Unlike centroid-based

methods, PAM uses actual data points as medoids, this approach is particularly resilient to noise and outliers showing a clear advantage or more classic k-means approach. Lastly, Hierarchical clustering a method that constructs a cluster hierarchy by iteratively splitting or merging data points.

2.3 Supervised Learning Methods for Classification

The library used for the supervised learning is Caret (Classification And REgression Training) a comprehensive package in R that provides a unified interface for building and evaluating machine learning models[5]. The methods implemented are as follows: Random Forest is an ensemble learning technique that combines the predictions of multiple decision trees to achieve higher accuracy and robustness. The final prediction is then made by aggregating the predictions from all the individual trees[6]. LASSO (Least Absolute Shrinkage and Selection Operator) is a linear regression method that introduces L1 regularization to penalize the absolute values of the regression coefficients. This penalty encourages sparsity in the model, effectively performing feature selection by shrinking some coefficients to zero[7]. Elastic Net is a linear regression method that combines L1 and L2 regularization to penalize both the absolute values and the squared values of the regression coefficients[8]. K-nearest neighbours (KNN) is a non-parametric method in which the prediction for a new data point is determined by the majority class or average value of its 'k' nearest neighbours in the feature space[9]. SCUDO (Signature-based Clustering for Diagnostic Purposes) is a rank-based method for the analysis of gene expression profiles for diagnostic and classification purposes.[10]. All the models are used through caret and trained with three time repeated two-fold cross-validation with SMOTE (Synthetic Minority Over-sampling Technique)[11] correcting for class imbalances. The performance of each model has been evaluated using accuracy and F1-score.

2.4 Differential expression analysis and feature selection

Differentially Expressed Genes (DEGs) are genes whose expression levels change significantly under different experimental conditions. For this analysis, LIMMA (Linear Models for Microarray Data) has been used. LIMMA is a robust statistical method designed for gene expression data analysis, known for effectively handling data noise and complexities[12]. DEGs can be used when comparing two groups to apply feature selection and focus on the genes that are statistically more relevant to the single groups.

2.5 Functional Enrichment Analyses

Functional Enrichment Analysis identifies biological functions or pathways over-represented in a gene or protein list. It helps interpret large-scale genomic data by linking genes to known biological roles, providing insights into gene functions and disease associations. For this part, I choose two different tools. g:Profiler is a popular tool for functional enrichment analysis. It offers a user-friendly interface for analyzing gene lists and identifying enriched gene ontology terms, pathways, and protein domains[13]. DAVID is gene functional classification tool uses a novel fuzzy clustering algorithm to condense a list of genes or associated biological terms into organized classes of related genes or biology, called biological modules[14].

2.6 Network analysis

Network analysis is a powerful approach used to study complex systems by representing them as networks of interconnected nodes and edges. pathfindR is an R package for enrichment analysis via active sub-networks. The package also offers functionality to cluster the enriched terms and identify representative terms in each cluster, score the enriched terms per sample, and visualize analysis results[15]. STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) is a biological database and web resource that provides information about protein-protein interactions (PPIs). STRING assigns confidence scores to interactions, helping researchers prioritize and interpret potential interactions between proteins and allowing for cluster analyses[16].

3 Results

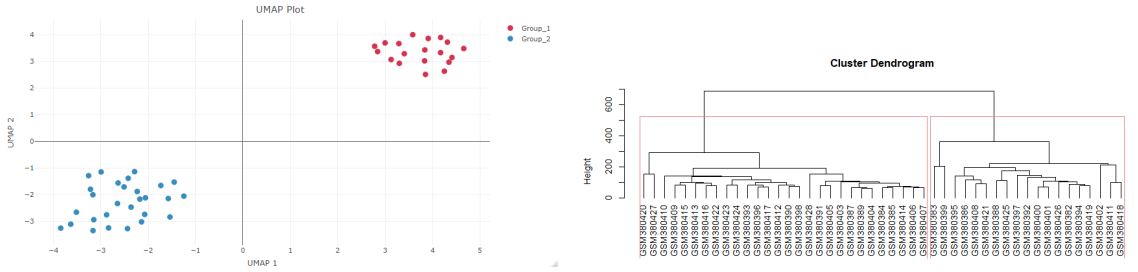
3.1 Unsupervised Learning Methods

The dataset has already undergone preprocessing and scaling by the authors. PCA and UMAP were conducted to explore any potential structure within the data, revealing a distinct division into two groups that didn't consider the molecular group of the sample. The UMAP plot (fig. 1b) shows a better separation. The distinction between the two groups became even clearer after I

(a) PCA Plot (b) UMAP Plot

Figure 1: Comparison of PCA and UMAP Plots

carried out PAM and hierarchical clustering. I decided on a value of k equal to 2 for PAM using the *silhouette method*, and based on consistent results, I used the same value of k for hierarchical clustering. As suspected the groups identified by the two methods overlap with the two clusters shown from the UMAP (fig. 2a). To investigate further, a differential expression analysis was



(a) UMAP plot with groups from PAM as labels. (b) Dendrogram plot.

Figure 2: UMAP and Dendrogram plots

performed and revealed six probes with an abnormal $\log_2$ fold change level as shown in (fig. 3). Those correspond to the gene XIST, which is involved in mammals' X chromosome inactivation to

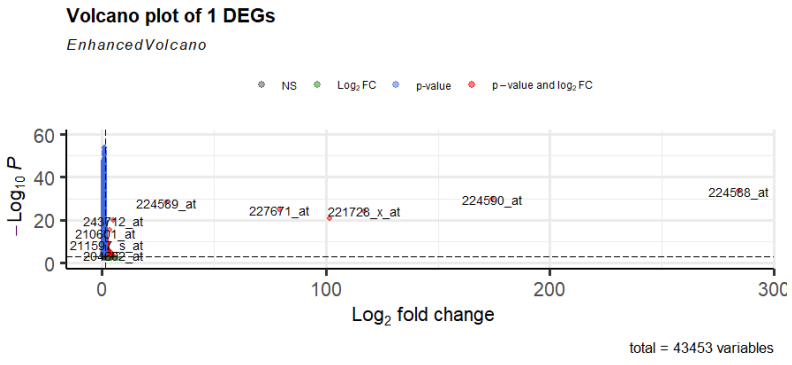


Figure 3: Volcano plot of the DEGs of group 1. It is possible to observe at least six probes with an abnormal expression level.

balance X-linked gene expression levels in XX females[17]. Those genes denote a difference between female and male samples which ha not been included by the authors in the metadata. I decided to remove the probes in question to avoid biases in further analysis. After removing the XIST probes I conducted a new UMAP analysis, where the samples were more evenly spread (fig. 4). The PAM

(a) UMAP plot with molecular groups as labels.

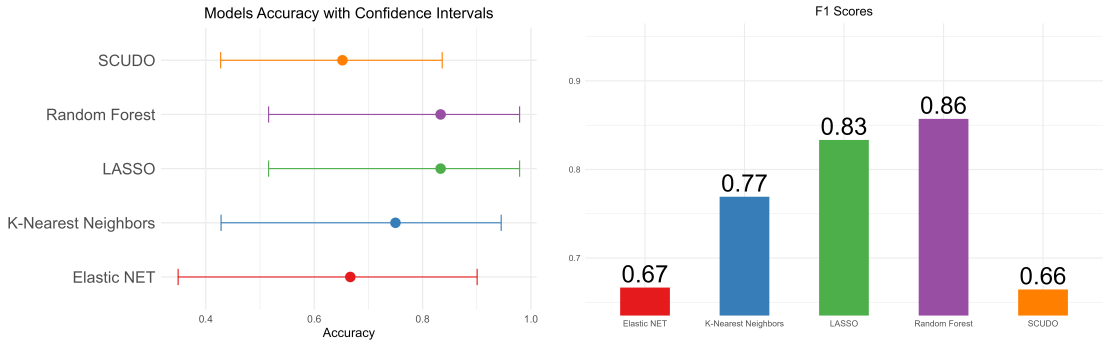
(b) UMAP plot with groups from PAM as labels.

Figure 4: UMAP with molecular labels and PAM clustering labels

groups now almost match the disease stages, indicating that the differences might be intrinsic to these stages. Therefore, even if some samples blend together, the nature of the disease suggests that a division between molecular groups is reasonable for further analysis.

3.2 Supervised Learning Methods for Classification

The models were evaluated by *accuracy* (5a) and *F1 score* (5b). Two models, LASSO and Random Forest, performed the best, both achieving an *accuracy* score of approximate 0.83. The choice of the best model for prediction was determined by the *F1 score*, favoring the Random Forest model with a score of 0.86. The random forest model, after extensive tuning, was used with parameters: $num.trees = 1000$, $importance =$ "impurity", $mtry = 143$, $splitrule =$ "gini" and $min.node.size = 8$. The results from the classification are shown in Table 1. For the next analysis, the new labels



(a) Dot-plot of the Accuracy of the models

(b) F1 scores of the models

Figure 5: Dot-plot and bar-plot of the metrics from caret models

for those samples will be used. The *ranger* library, which is used for the Random Forest model,

Sample ID	Molecular Group
GSM380425	Fibrosis
GSM380427	Fibrosis
GSM380426	Inflammation
GSM380428	Inflammation

Table 1: Sample IDs and predicted molecular groups.

offers the capability to extract the most important features. The top 20 features are shown in the bar plot (fig. 6) divided by clustering and ordered by importance: High, Medium, and Low. The genes in the High category are the most important in the model and will be kept in consideration for the next analysis. These genes include: SLC35B4 (Solute Carrier Family 35 Member B4) and SLC24A1 (Solute Carrier Family 24 Member 1) encodes for member of the potassium-dependent sodium/calcium exchanger protein family, those have been take in consideration as targets for hepatobiliary transformation in animal models[18]. EIF1 (Eukaryotic Translation Initiation Factor

1) enables RNA binding activity and is involved in regulation of translational initiation [19]. RBMX (RNA Binding Motif Protein X-Linked) is a Protein Coding gene that belongs to the RBMY gene family [19]. ARF6 (ADP Ribosylation Factor 6) encodes a member of the human ARF gene family, which belongs to the RAS superfamily. SNPs associated with biliary atresia (BA) identify a susceptibility locus at chromosome 14q21.3, encompassing the ARF6 gene. Knockdown of ARF6 in zebrafish indicates early biliary dysgenesis as a basis for BA and suggests a role for EGFR/ARF6 signaling in BA pathogenesis [20]. DUSP5 (Dual Specificity Phosphatase 5) this gene is a member of the dual specificity protein phosphatase subfamily. They negatively regulate members of the mitogen-activated protein (MAP) kinase superfamily [19]. DUSP5 has been take in consideration as a marker to monitor perinatal exposure to an environmental toxin in the genesis of BA in animal model [21].

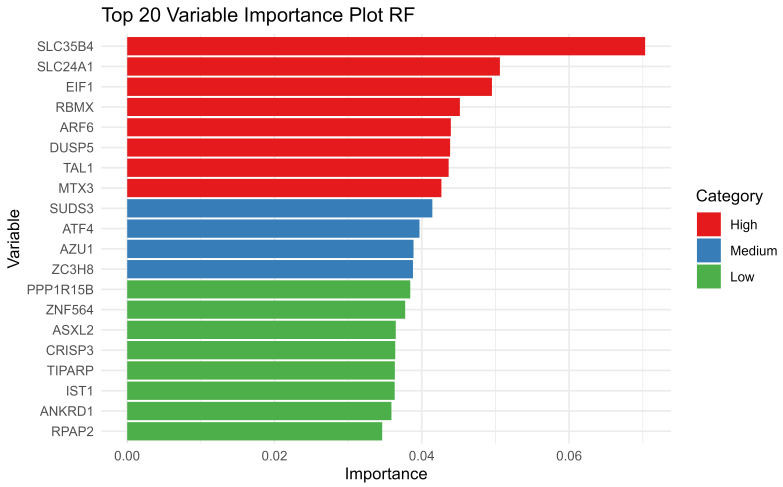
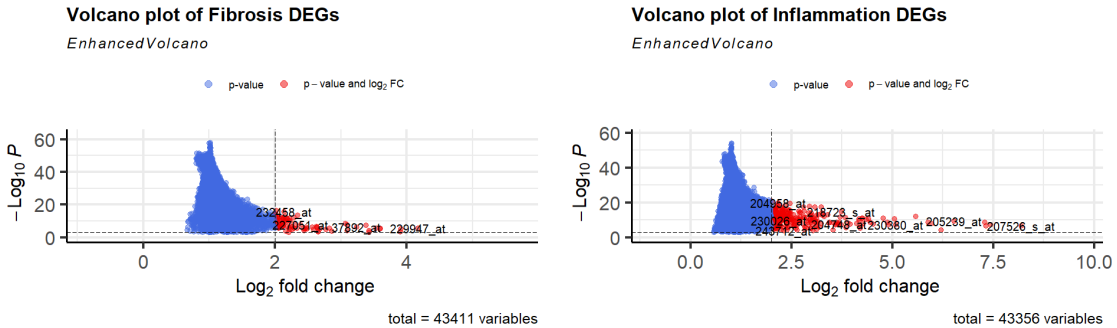


Figure 6: Bar-plot of the 20 most important genes, ranked by their importance

3.3 Differential expression analysis

Using LIMMA with a p-value threshold of 0.001 and a logFC threshold of 2.0, 206 genes have been identified as differentially expressed in inflammation (fig: 7b), and 74 genes have been identified as differentially expressed in fibrosis (fig: 7a). Those genes are going to be used in the following enrichment analysis. Comparing the DEGs with the features from the Random Forest model, we



(a) Vulcano plots for the fibrosis molecular group

(b) Vulcano plots for the inflammation molecular group

Figure 7: Volcano plots, p-valueCutoff = 0.001, FCcutoff = 2.0

observe that most of these genes are indeed overexpressed in one group compared to the other. Focusing on the most important features an heatmap (fig: 8) can show us a division, almost accordingly to the molecular group. The incorporation of a dendrogram alongside the rows where the genes are listed suggests that certain expression profiles might be linked to the stage or condition of BA.

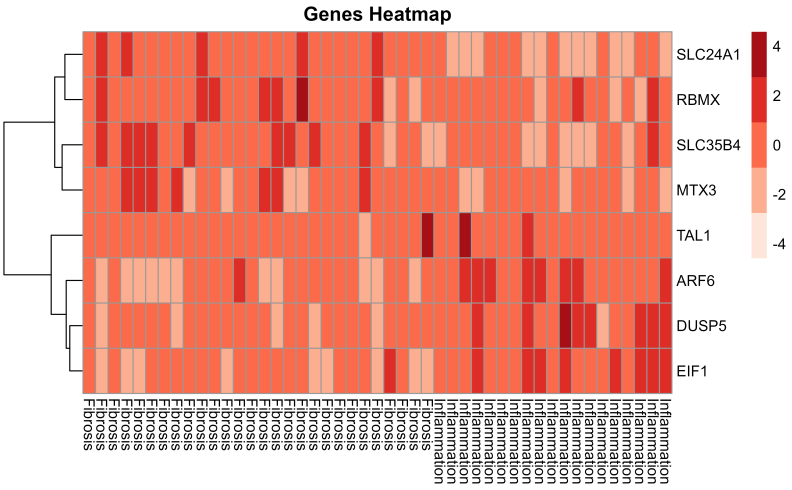


Figure 8: Heatmap of the most important features from the random forest model

3.4 Functional Enrichment Analyses

To explore the biological significance of the DEGs, I employed the DAVID and gprofiler tools for Over-Representation Analysis (ORA). I submitted separate queries for inflammation and fibrosis to identify enriched pathways and functional annotations related to each condition. From the DAVID analysis (fig. 9) we can see how the fibrosis list presents few terms related to inflammation and how the majority is about collagen. Multiple terms, from different categories address directly the organisation of extracellular matrix and the formation and trimming of collagen fibrils like: *extracellular matrix organization*,*Assembly of collagen fibrils and other multimeric structures* and *collagen fibril organization*. One term from KEGG particularly stands out, the *Viral protein interaction with cytokine and cytokine receptor* hinting at the possible viral nature of BA (fig. 9a). The inflammation terms, as expected, are mainly related to immune and inflammatory response. for the immune we can see terms such as: *IL-17 signaling pathway* and *TNF signaling pathway*. Regarding the inflammatory there are terms as:*Interleukin-10 signaling*,*Interleukin-4 and Interleukin-13 signaling* and *Cytokine Signaling in Immune system*(fig. 9b).

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
	GOTERM_BP_DIRECT	immune response	RT		12	16.2	9.7E-7	6.6E-4
	GOTERM_BP_DIRECT	extracellular matrix organization	RT		7	9.5	2.7E-5	9.0E-3
	GOTERM_MF_DIRECT	chemokine activity	RT		5	6.8	3.1E-5	3.7E-3
	KEGG_PATHWAY	Rheumatoid arthritis	RT		6	8.1	3.5E-5	4.3E-3
	GOTERM_MF_DIRECT	extracellular matrix structural constituent conferring tensile strength	RT		5	6.8	3.9E-5	3.7E-3
	REACTOME_PATHWAY	Chemokine receptors bind chemokines	RT		5	6.8	4.7E-5	2.8E-3
	REACTOME_PATHWAY	Assembly of collagen fibrils and other multimeric structures	RT		5	6.8	5.3E-5	2.8E-3
	GOTERM_MF_DIRECT	extracellular matrix structural constituent	RT		6	8.1	5.8E-5	3.7E-3
	REACTOME_PATHWAY	Extracellular matrix organization	RT		8	10.8	6.1E-5	2.8E-3
	REACTOME_PATHWAY	Collagen degradation	RT		5	6.8	6.4E-5	2.8E-3
	GOTERM_BP_DIRECT	collagen fibril organization	RT		5	6.8	7.0E-5	1.6E-2
	GOTERM_BP_DIRECT	adaptive immune response	RT		9	12.2	1.4E-4	2.4E-2
	GOTERM_BP_DIRECT	neutrophil chemotaxis	RT		5	6.8	1.8E-4	2.5E-2
	REACTOME_PATHWAY	Collagen formation	RT		5	6.8	2.4E-4	8.4E-3
	GOTERM_BP_DIRECT	positive regulation of immune response	RT		4	5.4	2.7E-4	3.0E-2
	REACTOME_PATHWAY	Collagen chain trimerization	RT		4	5.4	4.5E-4	1.3E-2
	GOTERM_BP_DIRECT	antigen processing and presentation	RT		4	5.4	4.8E-4	4.6E-2
	REACTOME_PATHWAY	Peptide ligand-binding receptors	RT		6	8.1	5.7E-4	1.4E-2
	KEGG_PATHWAY	Viral protein interaction with cytokine and cytokine receptor	RT		5	6.8	6.9E-4	3.5E-2
	GOTERM_BP_DIRECT	chemotaxis	RT		5	6.8	7.7E-4	6.5E-2

(a) Top 20 terms from DAVID on fibrosis

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
	GOTERM_BP_DIRECT	inflammatory response	RT		31	15.0	1.8E-17	3.2E-14
	GOTERM_BP_DIRECT	neutrophil chemotaxis	RT		15	7.3	7.7E-14	6.6E-11
	REACTOME_PATHWAY	Interleukin-10 signaling	RT		12	5.8	3.9E-11	2.5E-8
	GOTERM_BP_DIRECT	negative regulation of apoptotic process	RT		25	12.1	2.3E-10	1.3E-7
	GOTERM_MF_DIRECT	protein binding	RT		169	82.0	4.0E-9	1.7E-6
	KEGG_PATHWAY	IL-17 signaling pathway	RT		13	6.3	1.0E-8	1.6E-6
	GOTERM_BP_DIRECT	cellular response to tumor necrosis factor	RT		13	6.3	1.1E-8	4.6E-6
	KEGG_PATHWAY	TNF signaling pathway	RT		14	6.8	1.4E-8	1.6E-6
	GOTERM_BP_DIRECT	positive regulation of apoptotic process	RT		18	8.7	2.3E-8	7.9E-6
	GOTERM_BP_DIRECT	positive regulation of cell migration	RT		17	8.3	2.8E-8	8.0E-6
	GOTERM_BP_DIRECT	chemotaxis	RT		12	5.8	3.3E-8	8.0E-6
	GOTERM_BP_DIRECT	cellular response to interleukin-1	RT		10	4.9	9.5E-8	2.0E-5
	REACTOME_PATHWAY	Immune System	RT		58	28.2	1.2E-7	3.6E-5
	GOTERM_BP_DIRECT	immune response	RT		21	10.2	1.5E-7	2.7E-5
	GOTERM_BP_DIRECT	positive regulation of cell population proliferation	RT		21	10.2	1.6E-7	2.7E-5
	REACTOME_PATHWAY	Cytokine Signaling in Immune system	RT		32	15.5	1.7E-7	3.6E-5
	REACTOME_PATHWAY	Neutrophil degranulation	RT		24	11.7	2.9E-7	4.7E-5
	GOTERM_BP_DIRECT	negative regulation of cell population proliferation	RT		19	9.2	3.3E-7	5.2E-5
	REACTOME_PATHWAY	Interleukin-4 and Interleukin-13 signaling	RT		12	5.8	3.6E-7	4.7E-5
	GOTERM_BP_DIRECT	monocyte chemotaxis	RT		8	3.9	3.7E-7	5.3E-5

(b) Top 20 terms from DAVID on inflammation

Figure 9: DAVID list of enriched terms

Those two identical lists were also submitted to gProfiler. This tool possesses an automatic algorithm to highlight the most important terms. The results obtained were concordant, confirming almost all the previously found terms (fig. 10). For the fibrosis list, gProfiler confirmed the findings of DAVID and presented even more terms related to collagen (fig. 10a). From the inflammation terms we see a confirmation of the immune and inflammatory response terms with some addiction. Two terms, in particular, are worth noting: the *positive regulation of smooth muscle cell proliferation*, which has been connected with BA before. The lower expression of genes related to this term is associated with a better outcome of the Kasai portoenterostomy surgery, making it a marker for the success of this procedure [22]. The *MAPK cascade* term that incorporates one of the important genes and possible marker, DUSP5.

ID	Source	Term ID	Term Name	p_{adj} (query_1)
1	GO:MF	GO:0005201	extracellular matrix structural constituent	4.408×10^{-5}
2	GO:MF	GO:0008009	chemokine activity	3.462×10^{-4}
3	GO:BP	GO:0030198	extracellular matrix organization	1.394×10^{-4}
4	GO:BP	GO:0009888	tissue development	5.385×10^{-3}
5	GO:BP	GO:0042127	regulation of cell population proliferation	2.950×10^{-2}
6	GO:BP	GO:0045785	positive regulation of cell adhesion	3.573×10^{-2}
7	GO:BP	GO:0030593	neutrophil chemotaxis	3.864×10^{-2}
8	REAC	REAC:R-HSA-2022090	Assembly of collagen fibrils and other multimeric structures	8.207×10^{-4}
9	REAC	REAC:R-HSA-1650814	Collagen biosynthesis and modifying enzymes	3.500×10^{-2}
10	REAC	REAC:R-HSA-1474290	Collagen formation	5.781×10^{-3}
11	REAC	REAC:R-HSA-8948216	Collagen chain trimerization	6.578×10^{-3}
12	GO:MF	GO:0045236	CXCR chemokine receptor binding	1.419×10^{-2}

(a) Most enriched terms identified by gProfiler on fibrosis

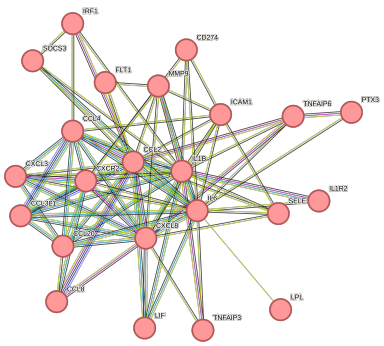
ID	Source	Term ID	Term Name	p_{adj} (query_1)
1	GO:MF	GO:0140375	immune receptor activity	6.944×10^{-3}
2	GO:BP	GO:2000351	regulation of endothelial cell apoptotic process	1.511×10^{-2}
3	GO:CC	GO:0005615	extracellular space	3.411×10^{-10}
4	GO:MF	GO:0005125	cytokine activity	9.204×10^{-7}
5	GO:MF	GO:0008009	chemokine activity	2.836×10^{-4}
6	REAC	REAC:R-HSA-6783783	Interleukin-10 signaling	4.532×10^{-10}
7	REAC	REAC:R-HSA-6785807	Interleukin-4 and Interleukin-13 signaling	2.469×10^{-5}
8	REAC	REAC:R-HSA-168256	Immune System	8.389×10^{-5}
9	KEGG	KEGG:04657	IL-17 signaling pathway	7.025×10^{-8}
10	KEGG	KEGG:04933	AGE-RAGE signaling pathway in diabetic complica...	1.493×10^{-3}
11	GO:MF	GO:0050786	RAGE receptor binding	9.332×10^{-4}
12	GO:BP	GO:0097529	myeloid leukocyte migration	1.176×10^{-11}
13	GO:BP	GO:0048661	positive regulation of smooth muscle cell prolifera...	1.673×10^{-4}
14	GO:BP	GO:0000165	MAPK cascade	1.087×10^{-5}
15	GO:BP	GO:0002685	regulation of leukocyte migration	9.032×10^{-8}

(b) Most enriched terms identified by gProfiler on inflammation

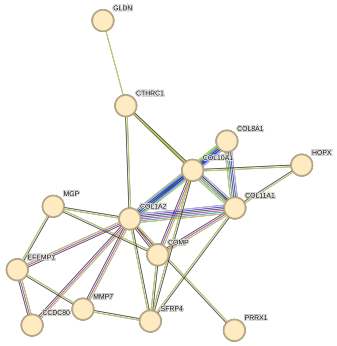
Figure 10: gProfiler tables of results from both condition

3.5 Network Analyses

The STRING tool reveals that both lists exhibit networks with significantly more interactions than expected. Even after adjusting the confidence level to 0.7, the networks continued to show a high number of edges, with a *PPI enrichment p-value* of 1.0e-16. Due to the high number of edges I applied DBSCAN algorithm, implemented directly by the STRING website, to gather information about potential clusters in the networks. From the fibrosis network, a primary cluster emerged, focusing on collagen and the extracellular matrix. All genes in the fibrosis main cluster are involved in collagen-related processes. Notably, four genes: COL11A1, COL10A1, COL8A1, and COL1A2 exhibit more edges than the others (fig. 11b). COL1A2, in particular, has been identified as a potential marker for BA[23]. The overexpression of COL1A2 and its associated genes, along with their presence in the importance list in the Random Forest model, suggests their significance in BA. The inflammation cluster is more densely populated, largely due to the larger gene list (fig. 11b). The network reveals two genes centrally connected to most nodes: IL-6, known in the BA context as a potential biomarker of deficient nutritional status[24] and CXCL8 (C-X-C Motif Chemokine Ligand 8), commonly known as IL-8 (interleukin-8), CXCL8 is identified as a potential marker for BA[23] and it's product overexpression, IL-8, positively correlates with inflammation and liver fibrosis[25]. pathfindeR was used with the KEGG database. The results mostly confirmed what was already shown by DAVID and gProfiler, supporting the validity of those findings. Many terms were related to viral infections and diseases, which supports the idea that BA might have a viral origin (fig. 12). This consistent pattern across different analysis tools highlights the potential connection between viral factors and the development of BA. The pathfindeR library offers the possibility to cluster the terms and then to compare the two analysis, forming a network of the common terms with genes from both molecular group. The network can be visualized as a graph

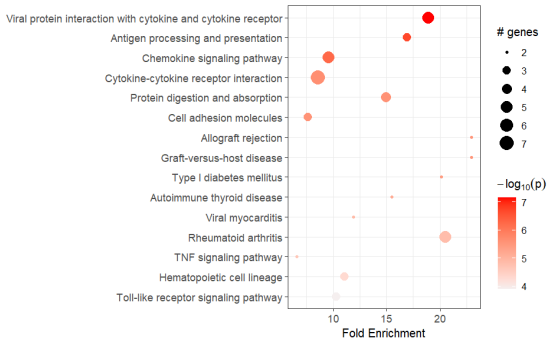


(a) Main cluster from the inflammation network obtained through the DBSCAN algorithm (confidence of 0.7), mainly related to inflammation as expected.

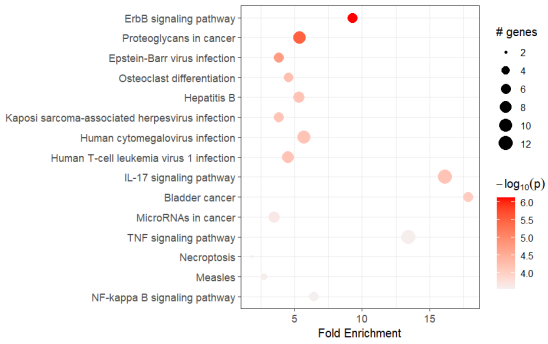


(b) Clusters from the fibrosis network obtained through the DBSCAN algorithm (confidence of 0.7), mainly related to collagen and fibrillar collagen.

Figure 11: Main clusters from STRING (confidence of 0.7)



(a) Most enriched terms for fibrosis.



(b) Most enriched terms for inflammation.

Figure 12: 15 most enriched terms from pathfindR in KEGG.

(fig. 13), where the genes are represented as green nodes. These green nodes connect to yellow nodes, which represent common terms. The genes can be unique or shared between molecular groups. In table 2, we can find the IDs and names of each of the six terms. The first two terms, which were identified in earlier analyses, are associated with viral interactions and include genes such as CXCL8, IL-6, and IL1B. These genes are central to the network, connecting most of the other terms. The terms *hsa04657* and *hsa04659* , related to IL-17 signaling pathway and Th17 cell differentiation, have been found to be up-regulated in liver tissues from patients with BA and considered potential markers for better diagnosis[26]. Furthermore, mice that were developing BA and given antibodies against IL-17 had lower levels of liver inflammation, making IL-17 a possible therapeutic target[27]. *hsa04010*, from the MAPK signaling pathway, highlights the presence of the gene DUSP5, This gene, part of the MAPK pathway, may also be linked to the action of ARF6. The canonical MAPK pathways and functional networks involved in cellular development and proliferation further support the role of EGFR/ARF6 signaling in zebrafish BA[20]. Lastly the unusual term *hsa04970*, realted to Salivary secretion, contains the gene ADRB2 up-regulated in the inflammatory group. The ARDB2 gene has been associated with inflammatory status, particularly in relation to IL1B, also present in the network. Single-cell studies have shown that NAC treatment improves bilirubin metabolism and bile acid flow in patients with BA. Additionally, NAC treatment downregulates innate and adaptive proinflammatory responses by also decreasing the activation of the IL1B/ADRB2 interaction[28].

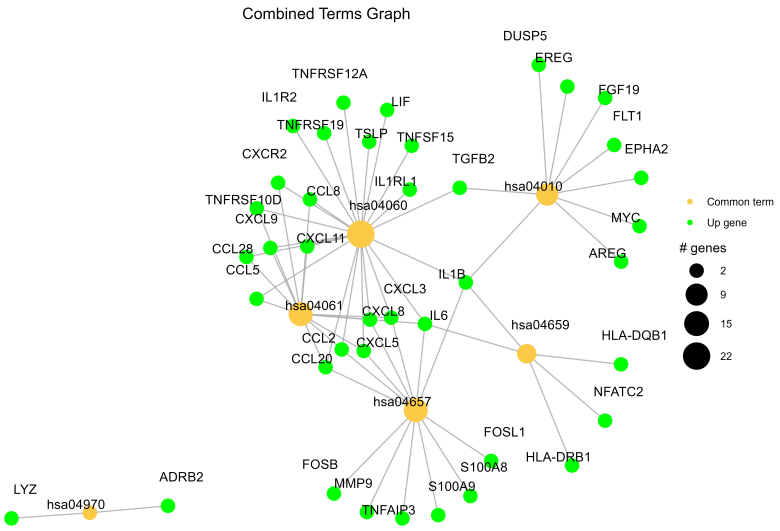


Figure 13: Network of the combined terms from both the molecular groups.

Pathway ID	Pathway Name
hsa04060	Cytokine-cytokine receptor interaction
hsa04061	Viral protein interaction with cytokine and cytokine receptor
hsa04657	IL-17 signaling pathway
hsa04659	Th17 cell differentiation
hsa04010	MAPK signaling pathway
hsa04970	Salivary secretion

Table 2: Pathways ID and term names from the combined network.

4 Discussion

This analysis aimed to complement the authors' histological classification of the dataset, composed of 47 infants diagnosed with biliary atresia, by using a machine learning model trained on the gene expression data. Additionally, enrichment analysis was conducted to gather more insight into the disease and different states. In recent years, attention towards BA has shifted to molecular profiling of the disease. Researchers are now focusing on understanding the mechanisms behind BA by examining the molecular and genetic aspects. From the unsupervised learning analysis, after filtering out the biased sex-related probes, an approximate division that nearly matches the molecular groups was revealed through UMAP and PAM clustering. To complement the classification various models have been used showing a large array of performances. The tie between LASSO and Random Forest was resolved by a higher F1 score achieved by the latter. Four unclassified samples were successfully classified by the model. In the process, the library *ranger* provided the ability to extract the most important features. Among these features, several genes such as SLC24A1, ARF6, and DUSP5, which have been previously identified in the literature for BA, were found as high importance. This supports the idea that these models can be useful not only for classification, but also for identifying new potential targets. It also suggests that further investigation into the other genes could be beneficial. The differential expression analysis reveled an high amount of DEGs in both the molecular groups, creating an ideal opportunity to use these DEGs as statistically and biologically significant features for selection. Filtering the DEGs by the most important features from the model revealed a subtle pattern in the expression of these genes between the molecular groups. This observation supports the possibility that different markers may be related to distinct stages of the disease. Functional enrichment analysis uncovered a great difference of terms between the groups. DAVID and gProfiler identified mostly common terms related to collagen organization and inflammatory response. This was also reflected in the STRING analysis, which showed distinct clusters in each group. The pathfindR results highlighted a significant presence of terms related to viral infections and diseases, supporting the hypothesis that BA might have a viral or external cause. Lastly, by combining the pathfindR terms from both groups, three terms related to inflammatory response, two related to viral interactions, including the MAPK pathway

involving DUSP5, and one related to salivary secretion, which includes ADRB2, were identified. DUSP5 could be an indicator of external influence in BA and ADRB2 could potentially serve as a marker for NAC experimental treatment in the early stages of inflammation. This study highlights the potential benefits of gaining molecular insights into BA for understanding the disease. A more complex study with a refined model, additional data, and novel platforms might uncover further details about BA. This study underscores the value of molecular insights for understanding BA. The findings suggest that employing models with additional data and exploring the inflammatory state could yield further details about BA's development and inform treatment approaches. Future research should focus on these areas to uncover more about the disease's mechanisms and potential therapeutic targets.

References

[1] William F Balistreri, Richard Grand, Jay H Hoofnagle, Frederick J Suchy, Frederick C Ryckman, David H Perlmutter, and Ronald J Sokol. Biliary atresia: current concepts and research directions. summary of a symposium. *Hepatology*, pages 1682–1692, 1996.

[2] Cara L Mack. The pathogenesis of biliary atresia: evidence for a virus-induced autoimmune disease. In *Seminars in liver disease*, pages 233–242. Copyright© 2007 by Thieme Medical Publishers, Inc., 333 Seventh Avenue, New . . . , 2007.

[3] Katie Moyer, Vivek Kaimal, Cristina Pacheco, Reena Mourya, Huan Xu, Pranavkumar Shivakumar, Ranajit Chakraborty, Marepalli Rao, John C Magee, Kevin Bove, et al. Staging of biliary atresia at diagnosis by molecular profiling of the liver. *Genome Medicine*, pages 1–13, 2010.

[4] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Grossberger. Umap: Uniform manifold approximation and projection. *The Journal of Open Source Software*, 3(29):861, 2018.

[5] Kuhn and Max. Building predictive models in r using the caret package. *Journal of Statistical Software*, 28:1–26, 2008.

[6] Marvin N Wright and Andreas Ziegler. ranger: A fast implementation of random forests for high dimensional data in c++ and r. *arXiv preprint arXiv:1508.04409*, 2015.

[7] J. Kenneth Tay, Balasubramanian Narasimhan, and Trevor Hastie. Elastic net regularization paths for all generalized linear models. *Journal of Statistical Software*, 106(1):1–31, 2023.

[8] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[9] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002. ISBN 0-387-95457-0.

[10] Matteo Ciciani, Thomas Cantore, and Mario Lauria. rscudo: an r package for classification of molecular profiles using rank-based signatures. *Bioinformatics*, 36(13):4095–4096, 2020.

[11] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[12] Matthew E Ritchie, Belinda Phipson, DI Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic acids research*, 43(7):e47–e47, 2015.

[13] Liis Kolberg, Uku Raudvere, Ivan Kuzmin, Jaak Vilo, and Hedi Peterson. gprofiler2– an r package for gene list functional enrichment analysis and namespace conversion toolset g:profiler. *F1000Research*, 9 (ELIXIR)(709), 2020. R package version 0.2.3.

[14] Da Wei Huang, Brad T Sherman, Qina Tan, Jack R Collins, W Gregory Alvord, Jean Roayaei, Robert Stephens, Michael W Baseler, H Clifford Lane, and Richard A Lempicki. The david gene functional classification tool: a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome biology*, 8:1–16, 2007.

[15] Ege Ulgen, Ozan Ozisik, and Osman Ugur Sezerman. pathfindr: an r package for comprehensive identification of enriched pathways in omics data through active subnetworks. *Frontiers in genetics*, 10:858, 2019.

[16] Damian Szklarczyk, Annika L Gable, David Lyon, Alexander Junge, Stefan Wyder, Jaime Huerta-Cepas, Milan Simonovic, Nadezhda T Doncheva, John H Morris, Peer Bork, et al. String v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic acids research*, 47(D1):D607–D613, 2019.

[17] Neil Brockdorff, Joseph S Bowness, and Guifeng Wei. Progress toward understanding chromosome silencing by xist rna. *Genes & development*, 34(11-12):733–744, 2020.

[18] Yu-Wen Chung-Davidson, Chu-Yin Yeh, Ugo Bussy, Ke Li, Peter J Davidson, Kaben G Nanlohy, C Titus Brown, Steven Whyard, and Weiming Li. Hsp90 and hepatobiliary transformation during sea lamprey metamorphosis. *BMC developmental biology*, 15:1–15, 2015.

[19] Marilyn Safran, Naomi Rosen, Michal Twik, Ruth BarShir, Tsippi Iny Stein, Dvir Dahary, Simon Fishilevich, and Doron Lancet. The genecards suite. *Practical guide to life science databases*, pages 27–56, 2021.

[20] Mylarappa Ningappa, Juhoon So, Joseph Glessner, Chethan Ashokkumar, Sarangarajan Ranganathan, Jun Min, Brandon W Higgs, Qing Sun, Kimberly Haberman, Lori Schmitt, et al. The role of arf6 in biliary atresia. *PloS one*, 10(9):e0138381, 2015.

[21] Kristin Lorent, Weilong Gong, Kyung A Koo, Orith Waisbourd-Zinman, Sara Karjoo, Xiao Zhao, Ian Sealy, Ross N Kettleborough, Derek L Stemple, Peter A Windsor, et al. Identification of a plant isoflavonoid that causes biliary atresia. *Science translational medicine*, 7(286):286ra67–286ra67, 2015.

[22] Priya Ramachandran, Ashitha K Unny, Mukul Vij, Mohamed Safwan, Muthukrishnan Saravana Balaji, and Mohamed Rela. α-smooth muscle actin expression predicts the outcome of kasai portoenterostomy in biliary atresia. *Saudi Journal of Gastroenterology*, 25(2):101–105, 2019.

[23] Hui Li, Lei Cao, and Hong Li. Col3a1, cxcl8, vcan, thbs2, and col1a2 are correlated with the onset of biliary atresia. *Medicine*, 102(11):e33299, 2023.

[24] Maria Ines de Albuquerque Wilasco, Carolina Uribe-Cruz, Daniele Santetti, Gabriel Rodrigo Fries, Cristina Toscani Leal Dornelles, and Themis Reverbel Da Silveira. Il-6, tnf-α, il-10, and nutritional status in pediatric patients with biliary atresia. *Jornal de Pediatria (Versão em Português)*, 93(5):517–524, 2017.

[25] Rui Dong and Shan Zheng. Interleukin-8: a critical chemokine in biliary atresia. *Journal of gastroenterology and hepatology*, 30(6):970–976, 2015.

[26] Peisong Chen, Zhihai Zhong, Hong Jiang, Huadong Chen, Junjian Lyu, and Luyao Zhou. Th17-associated cytokines multiplex testing indicates the potential of macrophage inflammatory protein-3 alpha in the diagnosis of biliary atresia. *Cytokine*, 116:21–26, 2019.

[27] Christian Klemann, Arne Schröder, Anika Dreier, Nora Möhn, Stephanie Dippel, Thomas Winterberg, Anne Wilde, Yi Yu, Anja Thorenz, Faikah Gueler, et al. Interleukin 17, produced by $\gamma\delta$ t cells, contributes to hepatic inflammation in a mouse model of biliary atresia and is increased in livers of patients. *Gastroenterology*, 150(1):229–241, 2016.

[28] Rongchen Ye, Sige Ma, Yan Chen, Jiarou Shan, Ledong Tan, Liang Su, Yanlu Tong, Ziyang Zhao, Hongjiao Chen, Ming Fu, et al. Single cell rna-sequencing analysis reveals that n-acetylcysteine partially reverses hepatic immune dysfunction in biliary atresia. *JHEP Reports*, 5(11):100908, 2023.

Appendix

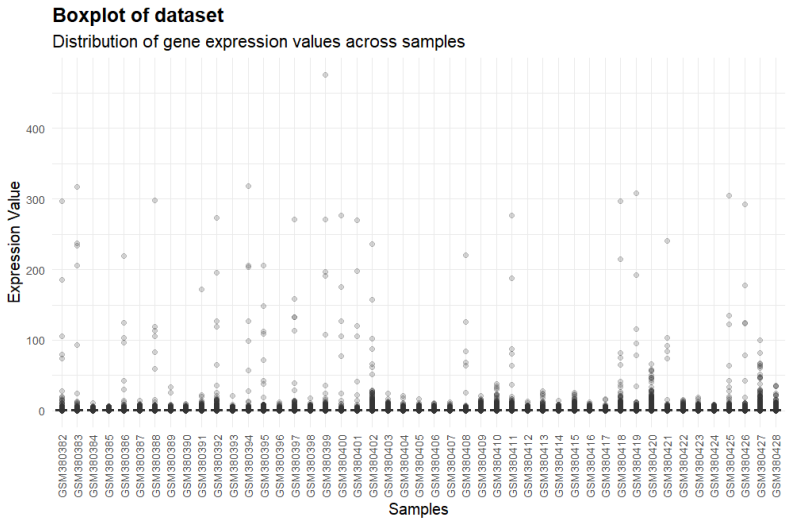


Figure 1: Boxplot of the dataset

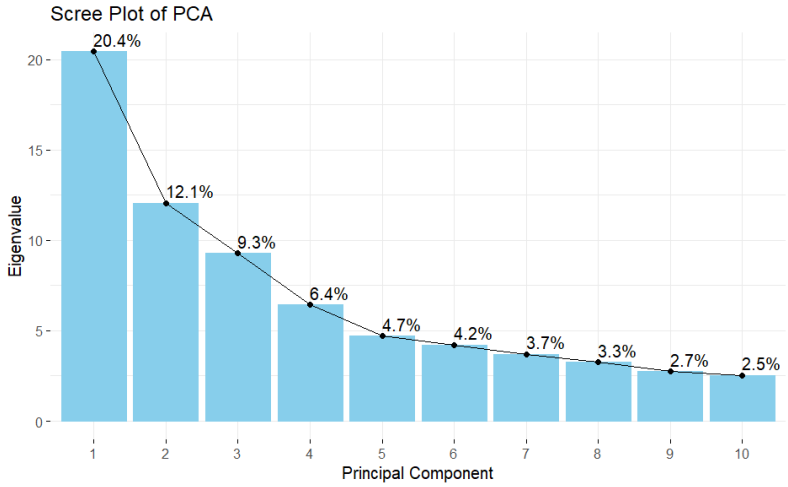


Figure 2: scree plot after removing the XIST probes

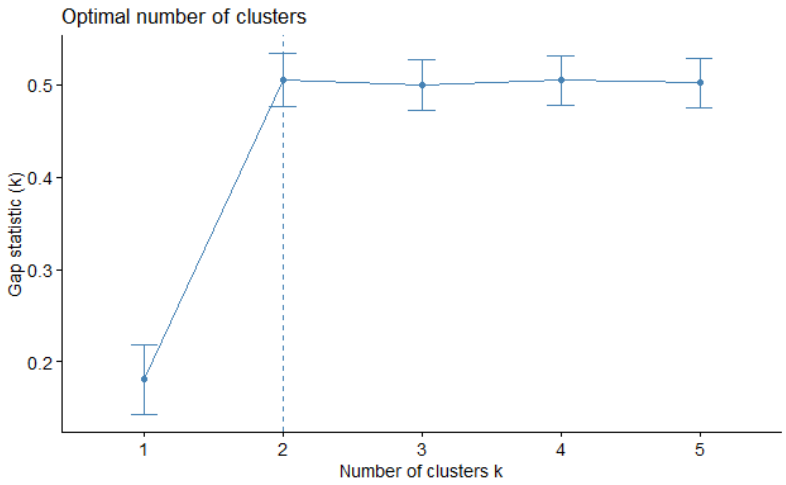


Figure 3: Best value of K for PAM and hierarchical clustering, after removing the XIST probes

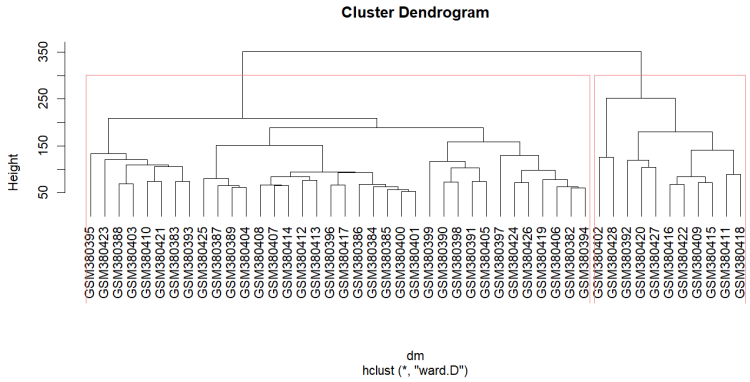


Figure 4: hierarchical clustering, after removing the XIST probes

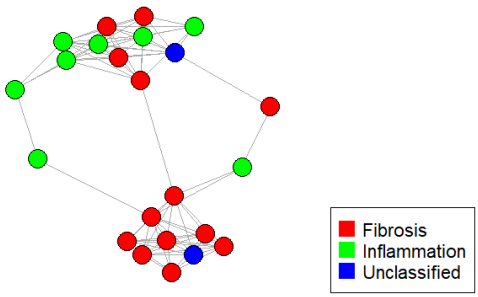


Figure 5: SCUDO network

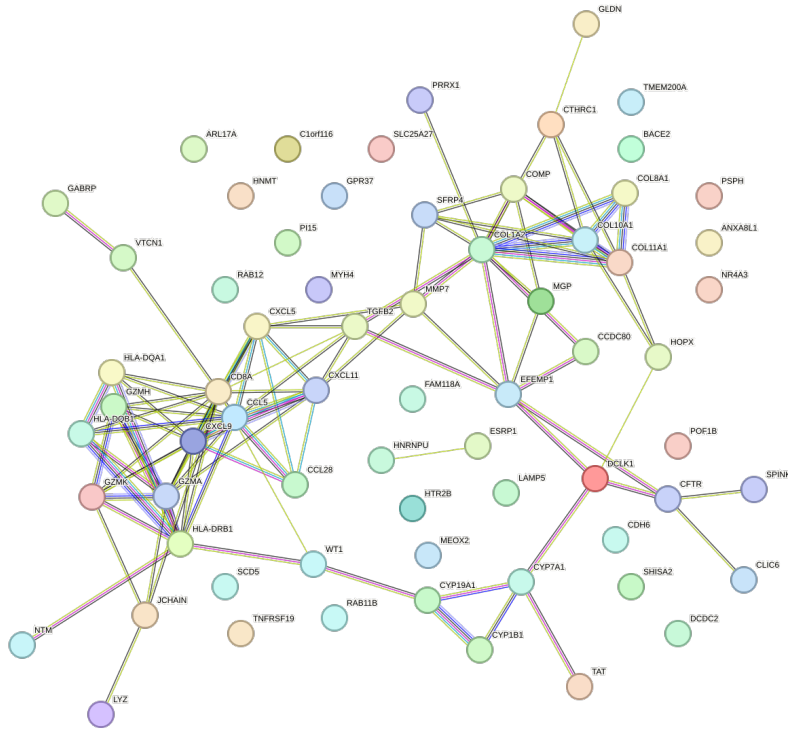


Figure 6: Full fibrosis STRING network (confidence = 0.7)

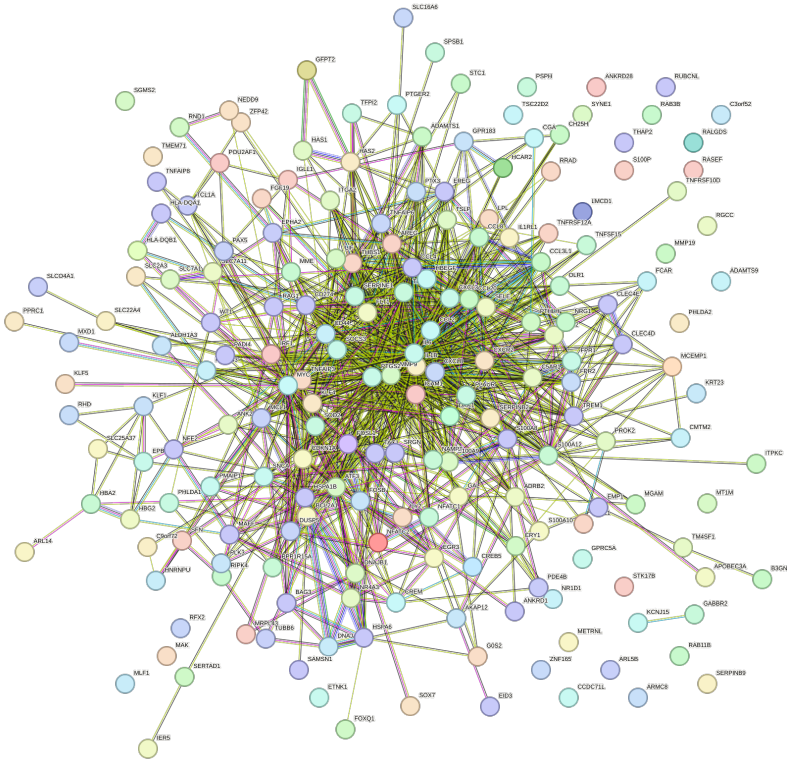


Figure 7: Full inflammation STRING network (confidence = 0.7)