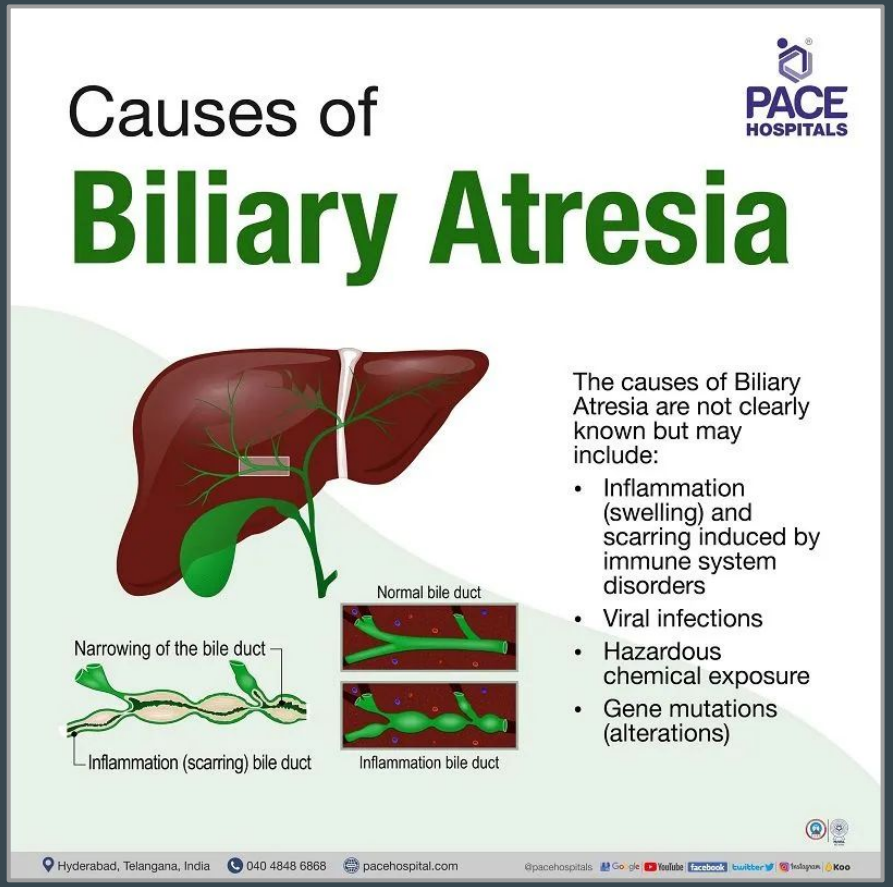
# Staging and classification of biliary atresia: an analysis of disease states

*Katie Moyer, Vivek Kaimal, Cristina Pacheco, Reena Mourya, Huan Xu, Pranavkumar Shivakumar, Ranajit Chakraborty, Marepalli Rao, John C Magee, Kevin Bove, Bruce J Aronow, Anil G Jegga & Jorge A Bezerra*

*Thomas Sirchi 239007*

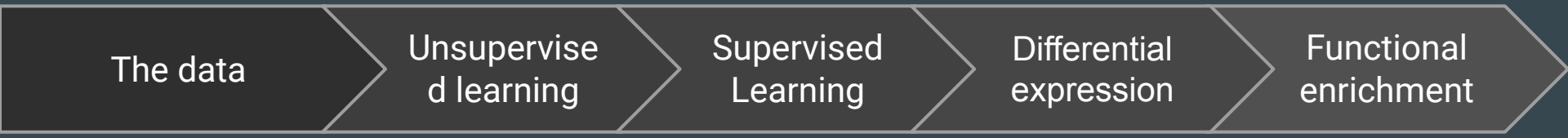
# The disease

- Biliary atresia (BA) affects infants and young children.
- Characterized by destructive inflammatory process in intra- and extrahepatic bile ducts.
- Leads or starts with fibrosis, progressive narrowing, and obliteration of bile ducts.
- About 270 cases are diagnosed each year in Europe




Biliary Atresia Symptoms, Causes, Types, Risk Factors & Complications

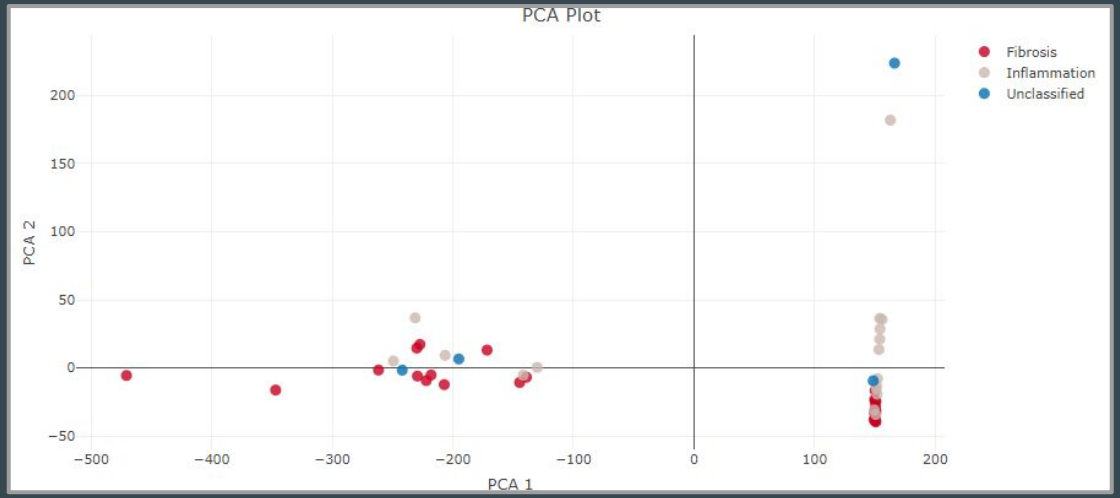
# The Pipeline

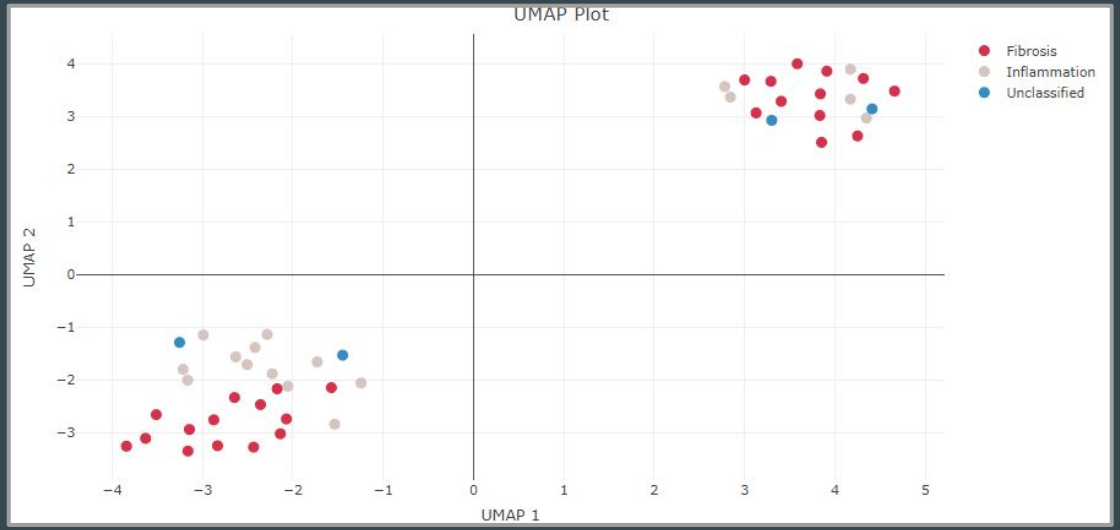



| The data | Unsupervised learning | Supervised Learning | Differential expression | Functional enrichment |
|---|---|---|---|---|
| Data retrieval<br>Metadata incorporation | PCA and UMAP<br>Clustering<br>Removal of biased probes | Random Forest<br>LASSO<br>Elastic NET<br>KNN<br>SCUDO | Feature selection<br>Heatmap most important features | David<br>Gprofiler<br>STRING<br>pathfindeR<br>Literature annotation |

# Unsupervised learning: PCA & UMAP

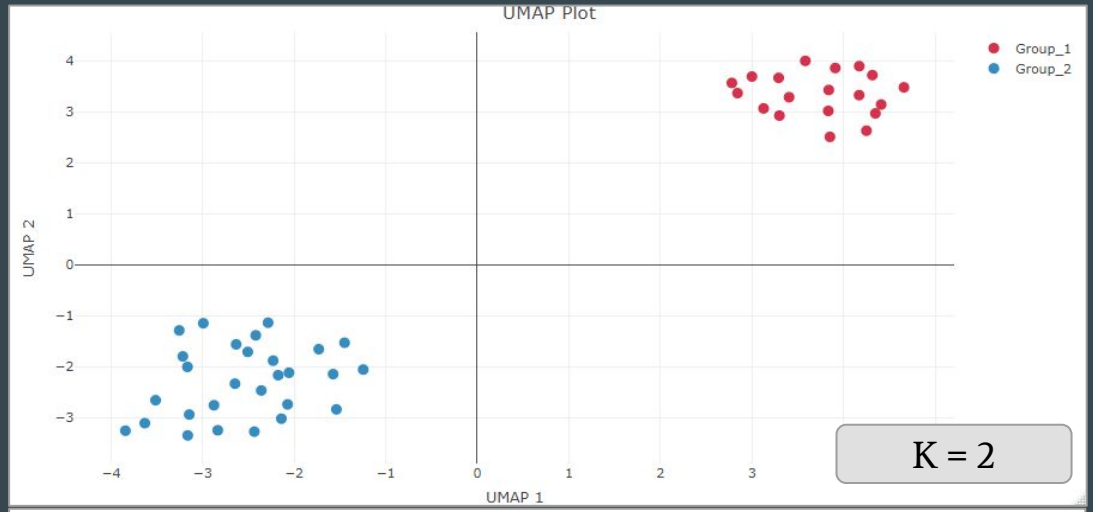
Clear division, not in the metadata

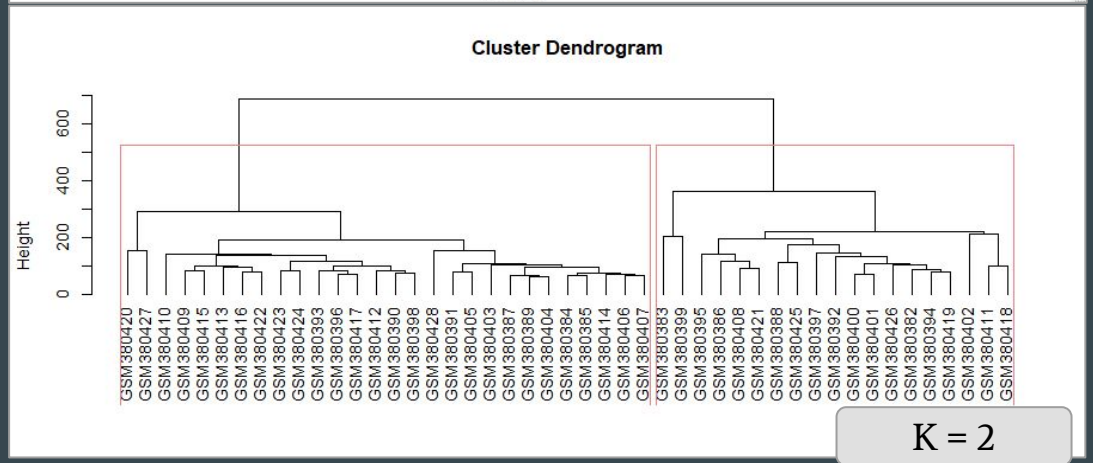


# Clustering

Clear division, not in the metadata

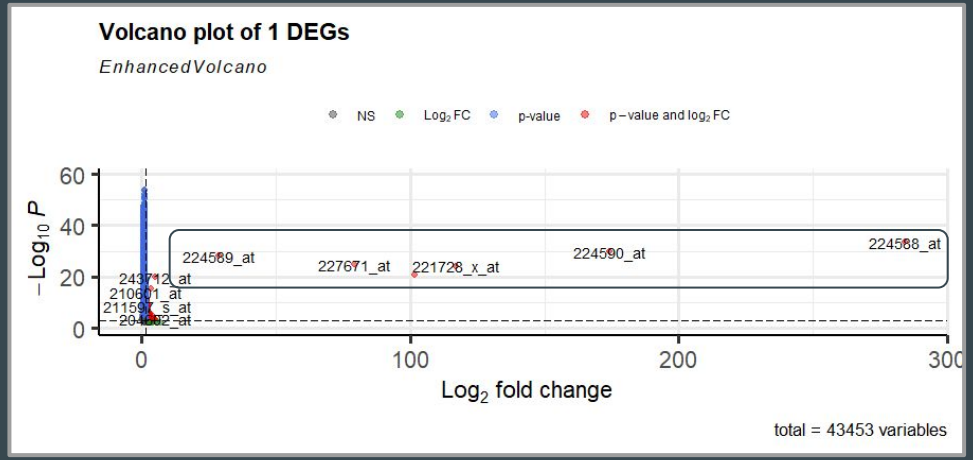


# X-inactive specific transcript (XIST)

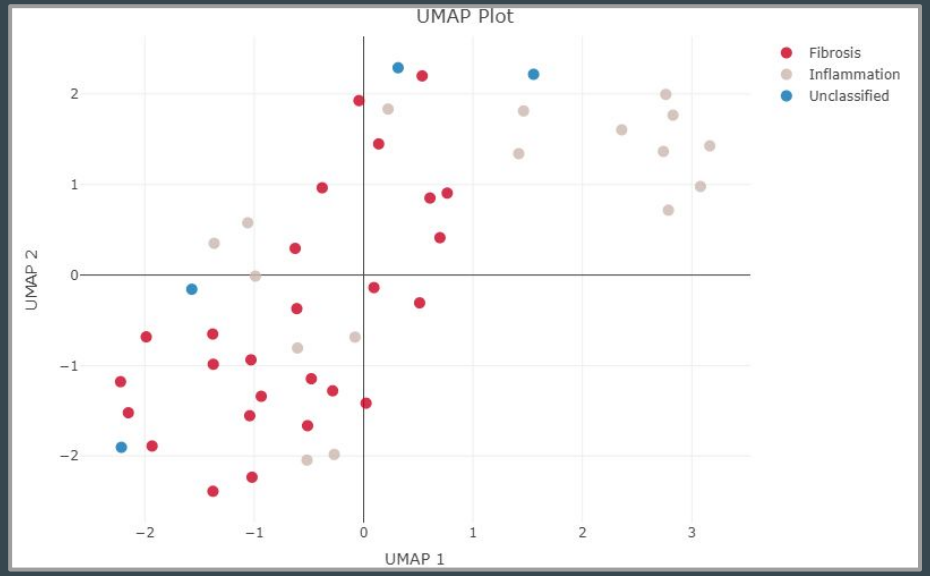
The probes shown on the volcano plot code for the XIST gene.

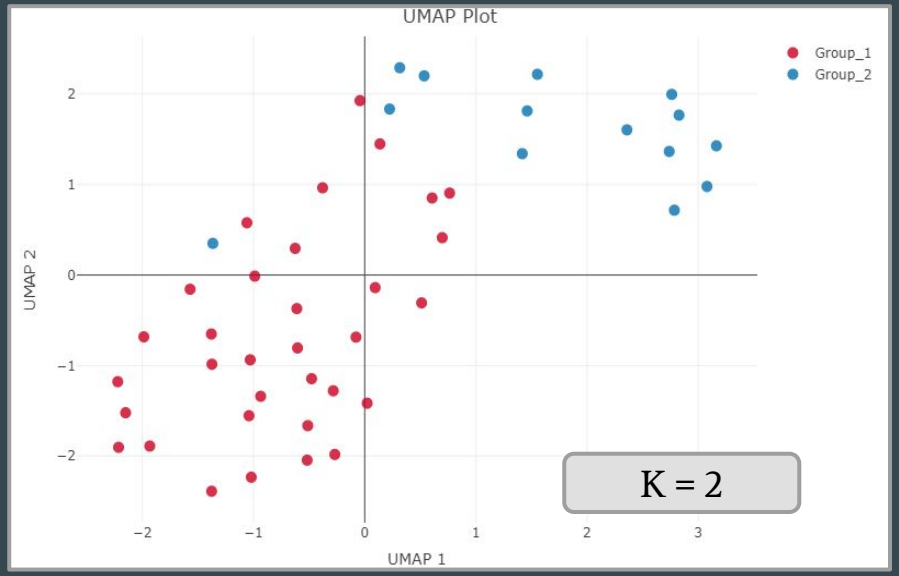
- Plays a crucial role in the process of X-chromosome inactivation in female mammals.
- This gene is located on the X chromosome and produces a long non-coding RNA (lncRNA) molecule



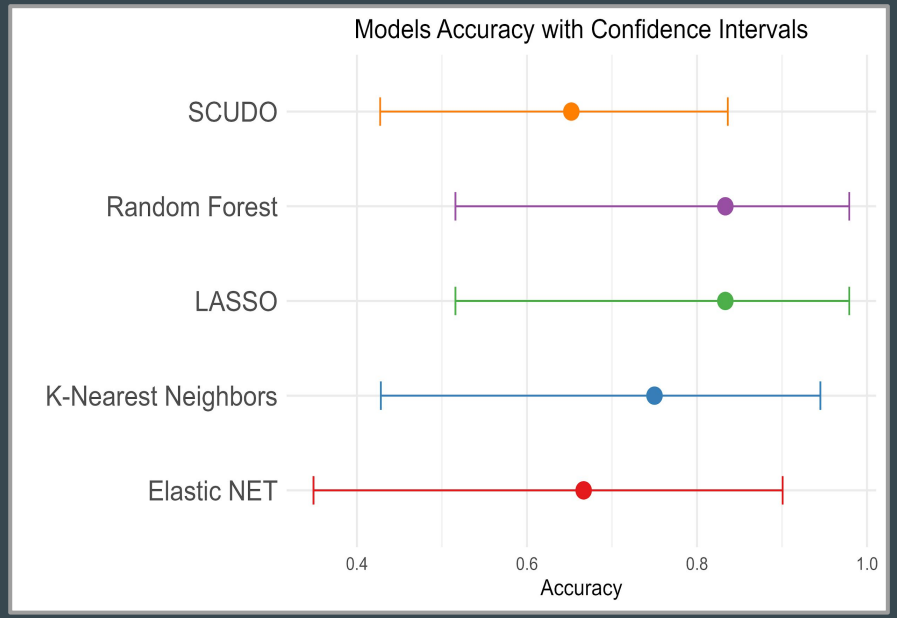

Removal of those probes assures to mitigate Male - Female differences

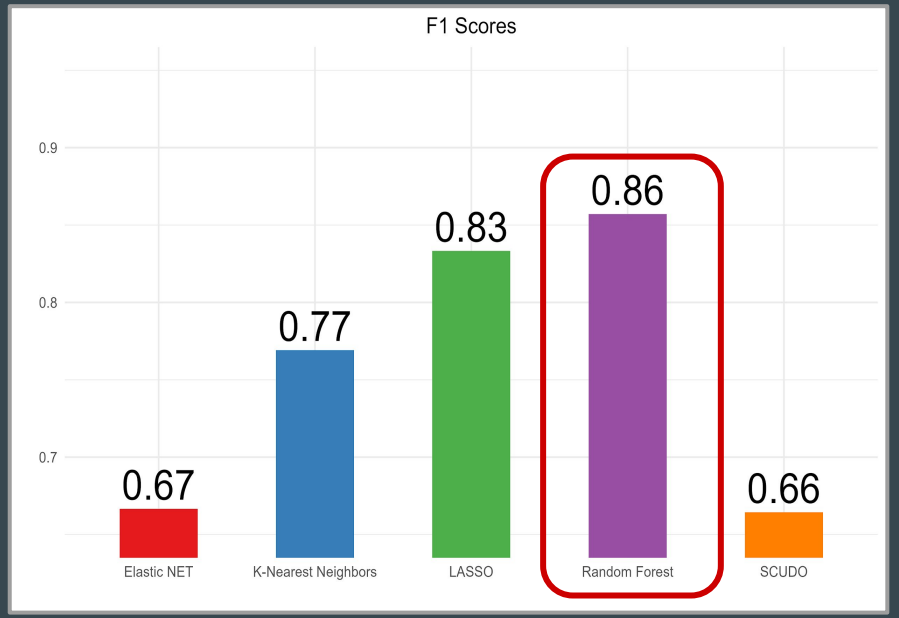
# New UMAP and PAM clustering



# Supervised Learning : Models performance and predictions

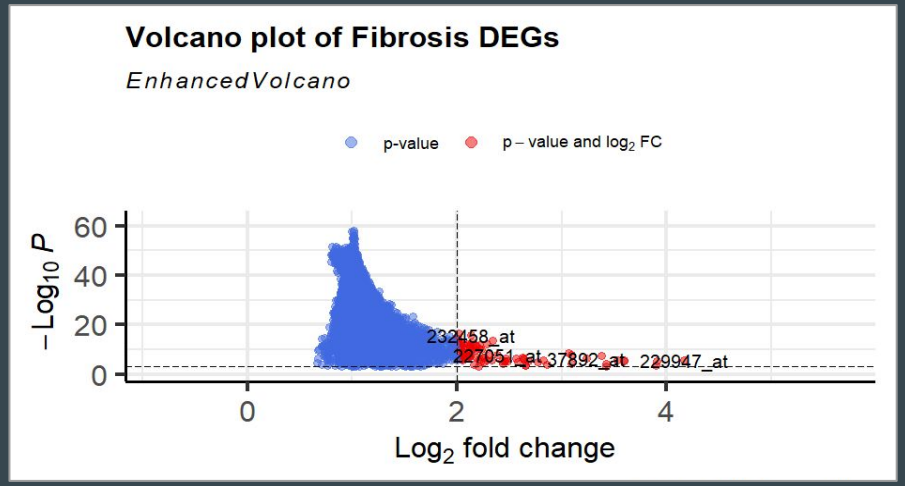


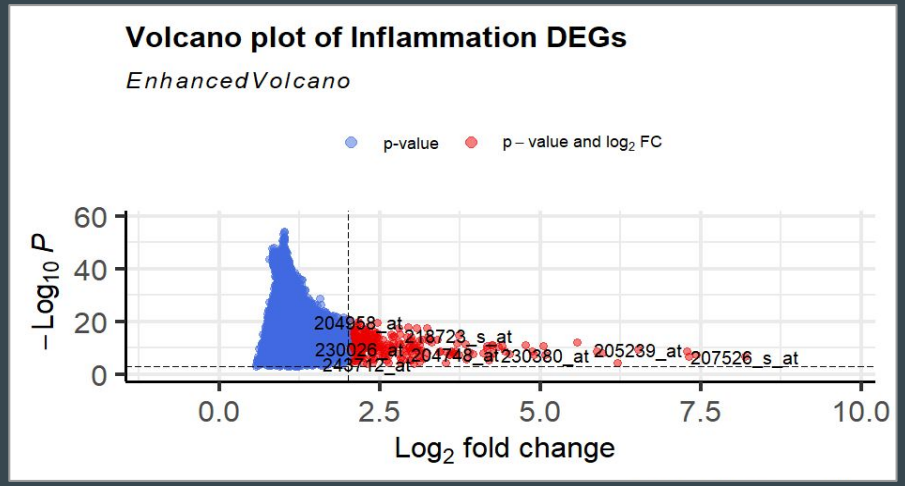




$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

# Differential expression analysis

LIMMA with a p-value threshold of 0.001 and a logFC threshold of 2.0



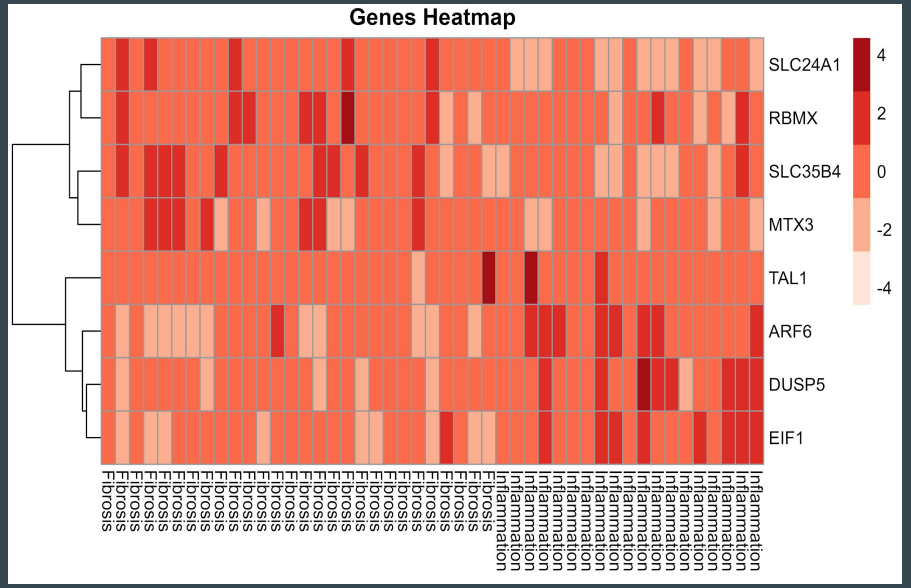




206 genes have been identified as differentially expressed in inflammation

74 genes have been identified as differentially expressed in fibrosis

# DEGs and Importance

- DEGs and Random forest importance
- Pheatmap library
- Two groups, diving almost as molecular groups




| | |
|---|---|
| *SLC35B4/24A1* (Solute Carrier Family) | Targets for hepatobiliary transformation in animal models |
| *EIF1* (Eukaryotic Translation Initiation Factor 1) | Involved in regulation of translational initiation |
| *ARF6* (ADP Ribosylation Factor 6) | Susceptibility locus at chromosome 14q21.3 |
| *DUSP5* (Dual Specificity Phosphatase 5) | Marker to monitor perinatal exposure to environmental toxin |

# David Fibrosis

| 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 |

# David Inflammation

| 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 |

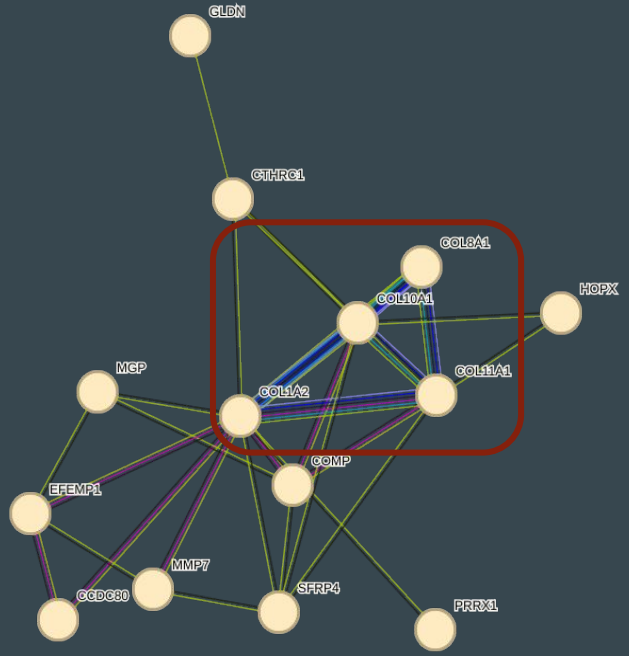
# Gprofiler Fibrosis

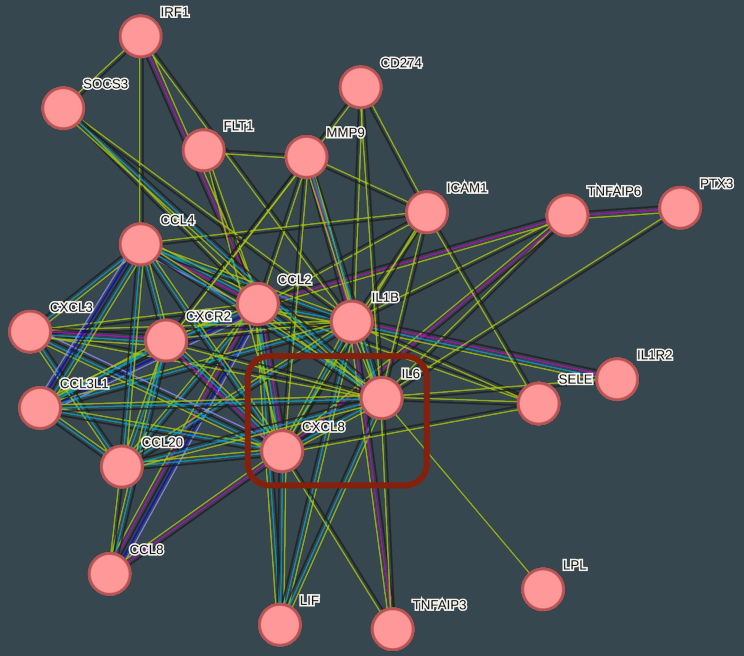
| ID | Source | Term ID | Term Name | $p_{adj}$ (query_1) |
|---|---|---|---|---|
| 1 | GO:MF | GO:0005201 | extracellular matrix structural constituent | $4.408\times10^{-5}$ |
| 2 | GO:MF | GO:0008009 | chemokine activity | $3.462\times10^{-4}$ |
| 3 | GO:BP | GO:0030198 | extracellular matrix organization | $1.394\times10^{-4}$ |
| 4 | GO:BP | GO:0009888 | tissue development | $5.385\times10^{-3}$ |
| 5 | GO:BP | GO:0042127 | regulation of cell population proliferation | $2.950\times10^{-2}$ |
| 6 | GO:BP | GO:0045785 | positive regulation of cell adhesion | $3.573\times10^{-2}$ |
| 7 | GO:BP | GO:0030593 | neutrophil chemotaxis | $3.864\times10^{-2}$ |
| 8 | REAC | REAC:R-HSA-2022090 | Assembly of collagen fibrils and other multimeric structures | $8.207\times10^{-4}$ |
| 9 | REAC | REAC:R-HSA-1650814 | Collagen biosynthesis and modifying enzymes | $3.500\times10^{-2}$ |
| 10 | REAC | REAC:R-HSA-1474290 | Collagen formation | $5.781\times10^{-3}$ |
| 11 | REAC | REAC:R-HSA-8948216 | Collagen chain trimerization | $6.578\times10^{-3}$ |
| 12 | GO:MF | GO:0045236 | CXCR chemokine receptor binding | $1.419\times10^{-2}$ |

# Gprofiler Inflammation

| ID | Source | Term ID | Term Name | $p_{adj}$ (query_1) |
|---|---|---|---|---|
| 1 | GO:MF | GO:0140375 | immune receptor activity | $6.944\times10^{-3}$ |
| 2 | GO:BP | GO:2000351 | regulation of endothelial cell apoptotic process | $1.511\times10^{-2}$ |
| 3 | GO:CC | GO:0005615 | extracellular space | $3.411\times10^{-10}$ |
| 4 | GO:MF | GO:0005125 | cytokine activity | $9.204\times10^{-7}$ |
| 5 | GO:MF | GO:0008009 | chemokine activity | $2.836\times10^{-4}$ |
| 6 | REAC | REAC:R-HSA-6783783 | Interleukin-10 signaling | $4.532\times10^{-10}$ |
| 7 | REAC | REAC:R-HSA-6785807 | Interleukin-4 and Interleukin-13 signaling | $2.469\times10^{-5}$ |
| 8 | REAC | REAC:R-HSA-168256 | Immune System | $8.389\times10^{-5}$ |
| 9 | KEGG | KEGG:04657 | IL-17 signaling pathway | $7.025\times10^{-8}$ |
| 10 | KEGG | KEGG:04933 | AGE-RAGE signaling pathway in diabetic complica... | $1.493\times10^{-3}$ |
| 11 | GO:MF | GO:0050786 | RAGE receptor binding | $9.332\times10^{-4}$ |
| 12 | GO:BP | GO:0097529 | myeloid leukocyte migration | $1.176\times10^{-11}$ |
| 13 | GO:BP | GO:0048661 | positive regulation of smooth muscle cell prolifera... | $1.673\times10^{-4}$ |
| 14 | GO:BP | GO:0000165 | MAPK cascade | $1.087\times10^{-5}$ |
| 15 | GO:BP | GO:0002685 | regulation of leukocyte migration | $9.032\times10^{-8}$ |

# STRING (confidence = 0.7), networks show significantly more edges



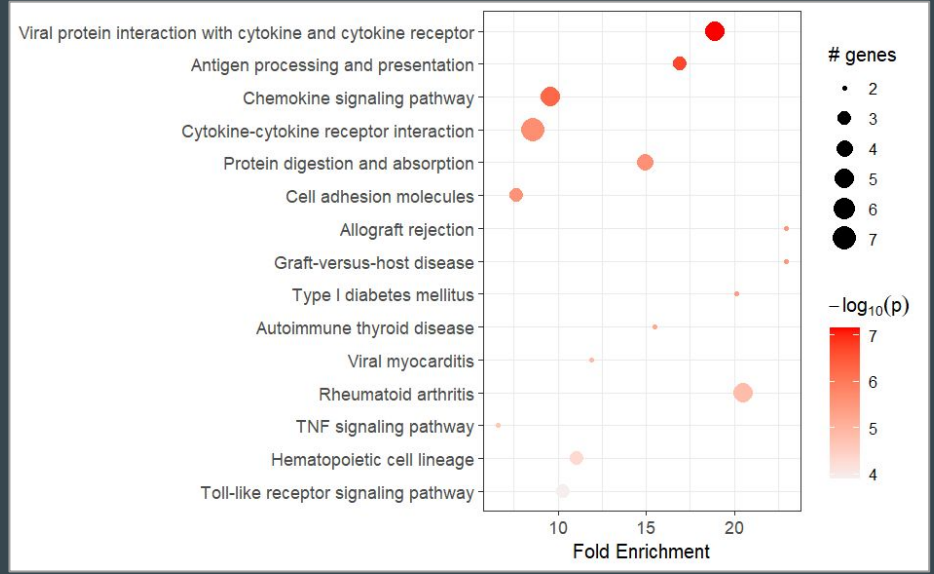




Fibrosis cluster about collagen and fibrillar collagen: COL1A2.

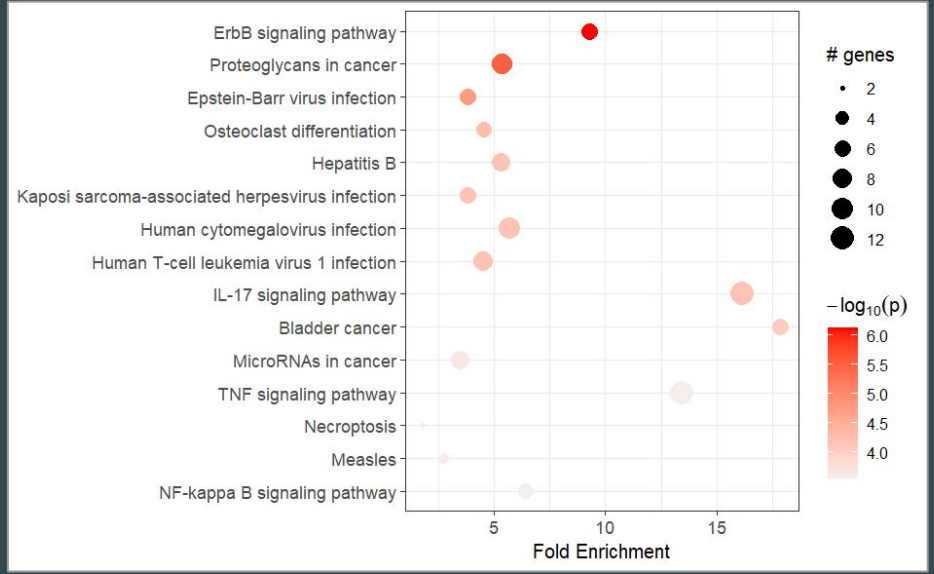
Inflammation cluster about chemokines and interleukins: IL6, CXCL8.

# pathfindeR with KEGG

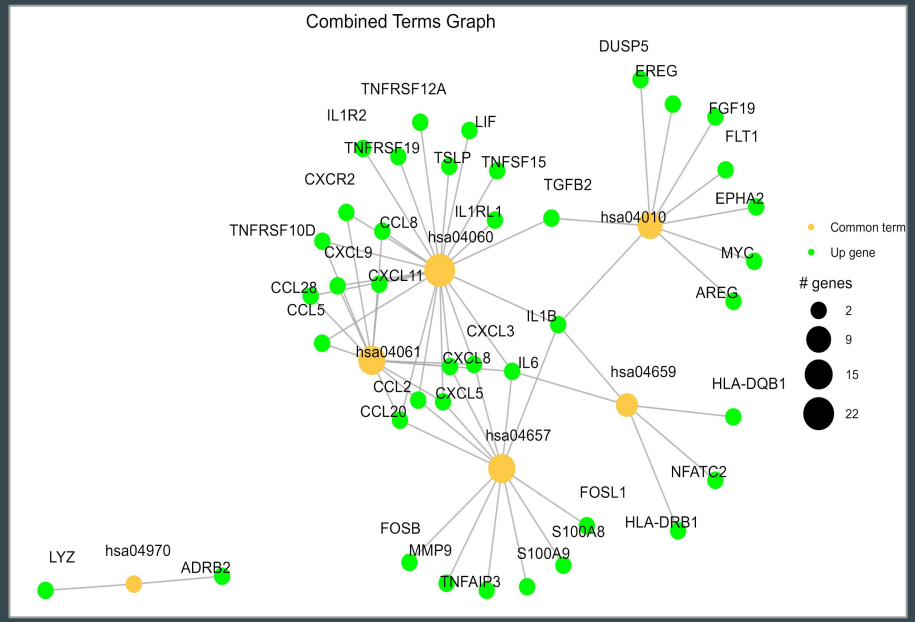
## Fibrosis




## Inflammation



# pathfindeR Combined




| 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 |

# Conclusions

Machine Learning Integration and Classification:

- Enhanced histological classification with machine learning models.
- Achieved good performance with the Random Forest model.
- Successfully classified four samples that were previously unclassified.

Molecular Insights and Feature Importance:

- Identified key genes (SLC24A1, ARF6, DUSP5) of high importance.
- Confirmed relevance of known genes and potential for new targets.
- DEG filtering revealed patterns linked to disease stages.
- Machine learning provided insights into potential therapeutic targets.

# Conclusions

Functional Enrichment and Pathway Analysis:

- DAVID and gProfiler identified terms related to collagen organization and inflammation.
- STRING analysis revealed distinct gene clusters per molecular group.
- pathfindR highlighted significant terms related to viral infections and inflammation.
- Identified key pathways such as MAPK (DUSP5), ARF6, and salivary secretion (ADRB2).

Future Directions and Implications:

- Molecular insights are valuable for understanding BA.
- Future research should involve more complex models, additional data, and novel platforms.
- Exploring inflammatory states and specific gene interactions may reveal more about BA mechanisms and inform treatment strategies.

# Thank you