Simo Inkala

DATA DRIVEN DRUG REPOSITIONING IN IDIOPATHIC PULMONARY FIBROSIS

Faculty of Medicine and
Health Technology
FHAIVE
Pro Gradu
June 21, 2023

Abstract

Simo Inkala: "Data driven drug repositioning in idiopathic pulmonary fibrosis"
Pro Gradu
Tampere University
Biomedical Technology, Master of Science
June 21, 2023

Idiopathic pulmonary fibrosis (IPF) is a progressive and chronic interstitial lung disease (ILD) that currently has few treatment options with limited efficacy and high cost. This study aims to shed light on the underlying mechanisms of IPF, identify potential biomarkers, and explore novel treatments using a data-driven approach. Additionally, the study evaluates the FAIRness of publicly available transcriptomics data repositories and integrates meta-analytical and network-based methods.

Microarray and RNA-seq datasets of both biopsies and different cell types of IPF patients and healthy controls were collected from GEO and ENA databases. The data were then curated and preprocessed using state-of-the-art methods. Gene co-expression networks were generated for each cell type (epithelial, macrophage, fibroblast, BAL). and biopsy. Subsequently, gene expression meta-analysis was conducted. The results indicate that potential treatments for IPF can be classified into five groups: collagenase enzymes, tyrosine kinase inhibitors, matrix metalloproteinase inhibitors, ion channel modulators and inhibitors, and proteins like monoclonal antibodies. Due to the complex pathogenesis of IPF, combination therapies may be more effective than monotherapies, and these five classes of drugs could be potential candidates. However, further research is necessary to determine the optimal dosages, administration routes, side effects and possible effects upon combination of these drugs.

In conclusion, the systems pharmacological approach used in this study is effective for the identification of new drug candidates for complex and poorly understood diseases like IPF. Combining network-based methods and meta-analytical approaches is an effective strategy, as they provide complementary perspectives. However, a challenge in using public repositories is ensuring the FAIRness of the data, which poses significant challenges despite the well-known principles of FAIR data.

Keywords: Bioinformatics, Gene co-expression networks, Collagenase enzymes, Data driven drug repositioning, Drug repurposing, ENA, FAIR data, GEO, Idiopathic pulmonary fibrosis, IPF, Ion channel modulators, Matrix metalloproteinase inhibitors, Meta-analysis, Monoclonal antibodies, Ocriplasmin, Public repositories, Systems biology, Sytstems pharmacology, Tyrosine kinase inhibitors

The originality of this thesis has been checked using the Turnitin Originality Check service.

Tiivistelmä

Simo Inkala: "Data driven drug repositioning in idiopathic pulmonary fibrosis"
Pro Gradu
Tampereen yliopisto
Bioteknologian ja biolääketieteen tekniikan maisteriohjelma
20.06.2023

Idiopaattinen keuhkofibroosi (IPF) on etenevä ja krooninen interstitiaalinen keuhkosairaus (ILD), jolla on tällä hetkellä vain muutama hoitovaihtoehto, joiden teho on rajallinen ja hinta korkea. Tämä tutkimus pyrkii valaisemaan IPF:n taustalla olevia mekanismeja, tunnistamaan mahdollisia biomarkkereita ja tutkimaan uusia hoitovaihtoehtoja datavetoisella lähestymistavalla. Lisäksi tutkimus arvioi julkisten tietorekisterien FAIR-periaatteiden noudattamista ja vertailee verkkopohjaisia ja differentiaalisen geeniekspression menetelmiä.

Mikrosiru- (microarray) ja RNAseq-aineistot kerättiin GEO- ja ENA-tietokannoista, ja ne käsiteltiin uusimmilla menetelmillä. Kullekin solutyypille (epiteelisolu, makrofagi, fibroblasti, BAL- ja biopsianäytteet) generoitiin yhteisesiintyvyysverkot, sekä suoritettiin differentiaalisen geeniekspression perusteella tehty meta-analyysi. Tulokset osoittavat, että mahdolliset IPF:n lääkehoidot voidaan luokitella viiteen ryhmään: kollagenaasientsyymit, tyrosiinikinaasin estäjät, matriksin metalloproteinaasin estäjät, ionikanavan säätelijät ja estäjät ja proteiinit, kuten monoklonaaliset vasta-aineet. IPF:n monimutkaisen patogeneesin vuoksi yhdistelmähoidot saattavat olla tehokkaampia kuin yksittäiset hoidot, ja nämä viisi lääkeryhmää voisivat olla potentiaalisia hoitovaihtoehtoja. On kuitenkin tarpeen tehdä lisätutkimuksia, jotta voidaan määrittää näiden lääkkeiden optimaaliset annokset, antotavat ja yhdistelmät.

Johtopäätöksenä tämän tutkimuksen käyttämä systeemifarmakologinen lähestymistapa on tehokas tapa löytää uusia lääkehoitoja monimutkaisiin ja huonosti tunnettuihin sairauksiin, kuten IPF:ään. Yhteisesiintyvyysverkkopohjaisten menetelmien ja differentiaalisen geeniekspression perusteella tehdyt menetelmät ovat tehokas tapa, koska ne tarjoavat toisiaan täydentäviä näkökulmia. Julkisten tietokantojen käytön haasteena on kuitenkin aineistojen FAIR-periaatteiden varmistaminen, mikä aiheuttaa merkittäviä haasteita siitä huolimatta, FAIR periaatteet ovat hyvin tunnettuja tiedeyhteisössä.

Avainsanat: Bioinformatiikka, Kollageenaasi-entsyymit, Dataohjautuva lääkeuudelleensijoitus, ENA, FAIR-data, GEO, Idiopaattinen keuhkofibroosi, IPF, Ionikanavasäätäjät, Lääkeuudelleenkäyttö, Okriplasmiini, Koekspressioverkostot, Monoklonaaliset vasta-aineet, Meta-analyysi, Matriksin metalloproteaasi-inhibiittorit, Julkiset tietokannat, Systeemibiologia, Systeemifarmakologia, Tyrosiinikinaasin estäjät

Preface

This thesis adheres to the Guide to Writing a Thesis in Technical Fields at Tampere University (2019) and has been created using the LaTeX environment due to its practicality and flexibility compared to other widely used text editor programs.

I would like to express my sincere gratitude to the amazing research group FHAIVE for providing me with the opportunity to undertake this study. I have been truly impressed by the team's enthusiasm, talent, and determined spirit. Special thanks to Antonio Federico for taking the time to answer my questions and offering invaluable tips. I also extend my appreciation to Dario Greco for being an unwavering and supportive boss. Your Italian attitude and approach are a valuable asset to this predominantly Finnish culture.

I would like to thank Emanuele for making it possible for me to participate in the Esperanto project, which was a priceless experience in software testing. I am grateful to Alisa, Angela, Anna, Giusy, Giorgia, Jack, Lena, Maria, Marcella, Maaret, Matias, Michele, Nicoletta, all the Lauras, Zeyad, as well as Aki, Heikki, and Maiju for providing peer support throughout this journey.

I would like to acknowledge Yliopiston Apteekki for the financial support that made this study possible. Additionally, I extend my gratitude to Gordon Carnegie for his help. To David Benton, thank you for your proofreading skills, belief in me, and encouragement to aim higher. And lastly, to my family, thank you for always being there for me.

Finally, I extend my appreciation to Tampere University for making this study and degree a reality.

Tampere, June 21, 2023

Simo Inkala

Contents

List of symbols and abbreviations

AE-IPF	Acute Exacerbation of Idiopathic Pulmonary Fibrosis
AT1	Alveolar epithelial cell type 1
AT2	Alveolar epithelial cell type 2
BAL	Bronchoalveolar lavage
BW	Burrow-Wheeler Transform
cGMP	Cyclic guanosine monophosphate
CLAD	Chronic lung allograft dysfunction
COP	Cryptogenic organizing pneumonia
CT	Computed tomography
DGE	Differential gene expression
ECM	Extracellular matrix
ESPERANTO	sEmi SuPERvised meta-dAta curatioN TOol
fHP	Fibrotic hypersensitivity pneumonitis
GFM	Graph-based FM index
GSEA	Gene Set Enrichment Analysis
HP	Hypersensitivity pneumonitis
ICER	Incremental cost-effectiveness ratio
ILD	Interstitial lung disease
IPF	Idiopathic pulmonary fibrosis
mAb	Monoclonal antibody
MCU	Mitochondrial calcium uniporter
MIP-1α	Macrophage inflammatory protein 1α
MMP	Matrix metalloproteinase
Mo-AM	Monocyte-derived alveolar macrophage
MWCNT	Multiwalled carbon nanotube
TR-AM	Tissue-resident alveolar macrophage
NAC	N-acetylcysteine
non-RTK	Cytoplasmic tyrosine kinase
NSIP	Non specific interstitial pneumonia
PCA	Principal Component Analysis
PDE	Phosphodiesterase
QALY	Quality-adjusted life year
QSP	Quantitative Systems Pharmacology
RB-ILD	Respiratory bronchiolitis-interstitial lung disease
RNA-seq	RNA sequencing
RTK	Receptor tyrosine kinase
sGC	Guanylate cyclase
SRA	Short Read Archive
SSc-ILD	Systemic sclerosis-associated interstitial lung disease
TIMP	Tissue inhibitor of a metalloproteinase
UF	Uncharacterized fibrosis
VMA	Vitreomacular adhesion
VRI	Vitreoretinal interface

Introduction

Lung functions and restrictive lung diseases

The primary function of the lung is to facilitate respiration, while also serving as a protective barrier between the body and the external environment. This barrier is essential in shielding the body from various harmful agents, including allergens, pollutants, chemicals, and pathogens that may threaten our health. Lungs have a vast surface area which means that they are potentially exposed to a plethora of toxins. In order to preserve homeostasis and protect itself from injury, lungs have evolved defense systems that guard it from these harmful entities (Suzuki et al. 2008).

Structural lung parenchyma cells are composed of various cell types such as epithelial cells, endothelial cells and fibroblasts. These cells are particularly susceptible to the damaging effects of harmful agents, with epithelial cells being the most affected. Epithelial cells are the structural barrier, and "muco-ciliary escalator" is a mechanical clearance system of the inhaled particles and microbes (Suzuki et al. 2008). There are numerous types of immune system cells in the distal parts of the lung that react against the unwanted agents. These cells contain for example macrophages, leukocytes, neutrophils, mast cells, dendritic cells and eosinophils. These cells are attracted by the cytokines and chemokines that modulate the inflammatory reaction and are extracted by a variety of cells when unwanted materials penetrate the lung (Suzuki et al. 2008).

Disruption of the homeostasis and the functional mechanisms of lungs can lead to variety of restrictive lung diseases such as idiopathic pulmonary fibrosis (IPF) which is a devastating lung disease that restricts lung function and reduces the ability to breathe. Studies have shown that chronic exposure to environmental pollutants, infections, and cigarette smoke can cause homeostasis disruption, which can contribute to the development of IPF (Krishna et al. 2022). Understanding the molecular and systemic mechanisms behind homeostasis disruption can help in the development of new therapeutic approaches for the treatment of IPF and other restrictive pulmonary diseases. In Figure 1 is illustrated a simplified disease progression mechanism of IPF.

Idiopathic pulmonary fibrosis

Idiopathic pulmonary fibrosis (IPF) is a progressive and chronic interstitial lung disease (ILD). It is a devastating, age-related lung disease that has unknown origin and only a few, and not very effective, treatment options (King et al. 2011). In IPF, progressive lung scarring occurs in the supporting interstitium of the lungs (Martinez et al. 2011). These events cause breathing to become increasingly difficult. IPF is irreversible, and usually leads to death (King et al. 2011). The main histopathological features of IPF are heterogeneous appearance of areas of subpleural and paraseptal fibrosis and honeycombing (fibrotic spaces lined by

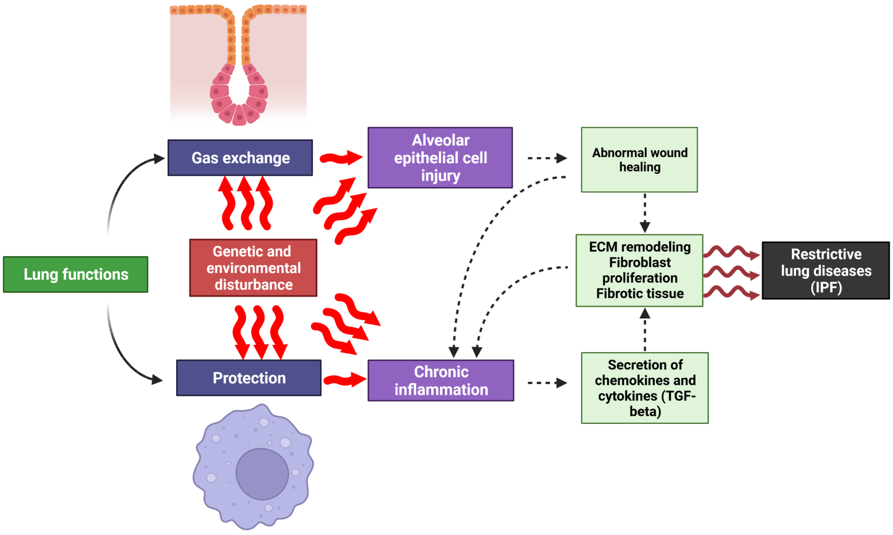
Gas exchange
Alveolar epithelial cell injury
Abnormal wound healing
Lung functions
Genetic and environmental disturbance
ECM remodeling
Fibroblast proliferation
Fibrotic tissue
Restrictive lung diseases (IPF)
Protection
Chronic inflammation
Secretion of chemokines and cytokines (TGF-beta)

Figure 1: Lung functions and disease progression in IPF.

bronchiolar epithelium and often filled by mucin and variable amounts of inflammatory cells) with areas of less affected or normal parenchyma (King et al. 2011).

One proposed pathogenesis of IPF is the connection to ageing-related susceptible lung which is targeted by repetitive alveolar injuries that are caused by, for example, inhaled cigarette smoke, microaspiration, nanomaterials, gastroesophageal reflux or viruses (King et al. 2011; Wuyts et al. 2013). These injuries can provoke type I and type II epithelial cell death. After microinjuries and epithelial cell apoptosis, increased vascular permeability to proteins like fibrinogen and fibronectin causes the formation of a wound clot. This process is followed by bronchiolar and alveolar epithelial cell migration and proliferation which is a frustrated effort of the lung to try to repair itself (King et al. 2011). Abnormally activated epithelial cells start to excrete different chemokines, cytokines and epidermal growth factors which attract fibroblasts and immune system cells like alveolar macrophages and monocytes that will differentiate into macrophages.

The cells also excrete TGF-β1 that promotes epithelial mesenchymal transition, extracellular matrix remodeling and the differentiation of fibroblasts to myofibroblasts. The heterogenous macrophage population also secrete chemokines and growth factors like TGF-β that induce the fibrotic tissue growth in the extracellular matrix. There are positive feedback loops that lead to progressive expansion of the fibrotic tissue (Misharin et al. 2017; Saarimäki et al. 2020; Sugeir et al. 2019).

Pathogenesis of IPF and the biological conditions used in this study

The pulmonary alveolar epithelium is essential for lung gas-exchange function and also represents an important barrier to protect our body from hazards. In response to acute injuries, pulmonary alveoli are usually able to quickly repair and regenerate new alveolar epithelial cells for restoring an intact epithelial layer. The alveolar epithelium is mainly composed of two types of epithelial cells: alveolar type I (AT1) and type II (AT2) cells. AT2 cells are smaller compared to AT1 cells. AT2 cells are cuboidal and they are best known for their functions in synthesizing and secreting pulmonary surfactant. In addition, AT2 cells function as alveolar stem cells and are able to differentiate into AT1 cells during alveolar homeostasis and post injury repair (Desai et al. 2014; Wang et al. 2018). AT1 cells are large squamous cells that cover 95 % of the alveolar surface area. They form the epithelial component of the air–blood barrier in alveoli. Both AT1 and AT2 cells differentiate at the late embryonic stage from alveolar progenitor cells and form distal epithelial saccules (Wang et al. 2018; Nikolić et al. 2017).

Following birth, the epithelial saccules undergo continuous subdivision, resulting in the formation of multiple smaller gas exchange units known as alveoli. This postnatal developmental process is called alveologenesis, which occurs with 90% of human alveoli. During alveologenesis, AT1 cells expand their surface area and flatten their cell body to accommodate postnatal lung growth. AT1 cells have been traditionally considered to be terminally differentiated cells. Although ATI cells were previously believed to be fully differentiated, recent studies have demonstrated that they possess cellular plasticity and can proliferate to generate AT2 cells during alveolar regeneration following post-pneumonectomy. However, the molecular genetics and fate specification of AT1 cells remain largely unknown due to limited knowledge about the development and heterogeneity of the adult AT1 cell population. Consequently, it is unclear whether all AT1 cells or only a subset can transdifferentiate into AT2 cells during alveolar regeneration. Furthermore, AT1 cell development during alveologenesis has not been thoroughly characterized at the transcriptome level due to the challenges associated with isolating these delicate cells (Wang et al. 2018).

AT2s are thought to play a critical role in the development of IPF. The pathogenesis of IPF is commonly believed to be initiated by damage to type AT2 cells, which leads to an epithelial-driven process that activates pro-fibrotic signaling mediated by TGF-$\beta1$. The activation of TGF-$\beta1$ causes a disruption in communication between fibroblasts and other cells, leading to the activation of myofibroblasts and an excessive buildup of extracellular matrix (ECM). TGF-β has three isoforms, which are TGF-$\beta1$, TGF-$\beta2$, and TGF-$\beta3$. In the pathogenesis of IPF, TGF-β leads to alveolar epithelial injury, fibroblast activation, myofibroblast transdifferentiation, excessive production of ECM, and inhibition of ECM degradation. The specific mediators that regulate this process are not entirely clear (Bueno et al. 2023).

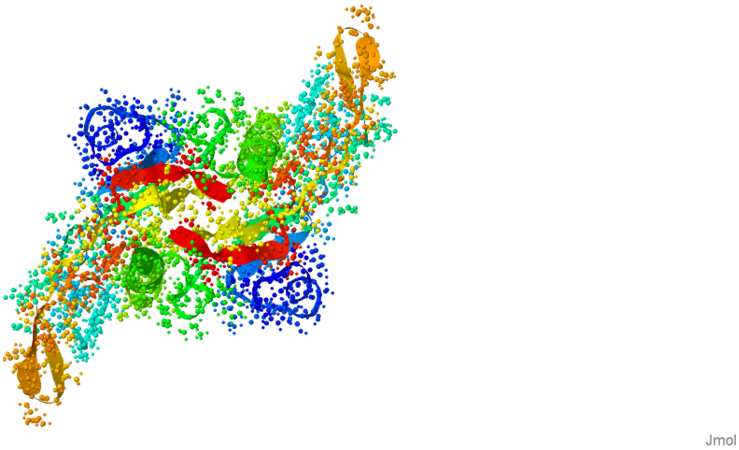

Figure 2: TGF-$\beta1$ structure.

Macrophages have a significant role in the pathogenesis of lung fibrosis, particularly in IPF (Novak et al. 2023; Misharin et al. 2017; Geng et al. 2021). Macrophages have the ability to release and respond to cytokines, which impacts their activation state and affects the functional behavior of surrounding cells. They play a vital role in coordinating the inflammatory response by releasing both pro-inflammatory cytokines (such as Il-1β, TNF-α, Il-6, and Il-8) and anti-inflammatory cytokines (such as Il-4, Il-10, Il-13, IFN-α, and TGF-β). During a typical wound healing response, macrophages are crucial in resolving fibrosis by taking up dead cells and excessive ECM, thereby degrading scar tissue and facilitating the resolution of the injury (Novak et al. 2023). Monocyte-derived alveolar macrophages (Mo-AMs) express higher levels of proinflammatory and profibrotic genes (ie. *ADAM8, ARG1, APOE, ITGA6, MFGE8, MMP12, MMP13, MMP14, and PDGFA*) than tissue-resident alveolar macrophages (TR-AMs). It has been revealed that the deletion of Mo-AMs after their recruitment to the lung markedly reduced the severity of fibrosis in common mouse models of lung fibrosis. Macrophage polarization, which refers to distinct sets of inflammatory or fibrotic genes that are expressed by macrophages in cell culture that are first induced toward differentiation and then treated with lipopolysaccharides and IFN-γ or Il-4, respectively (Misharin et al. 2017).

The majority of resident macrophages originate from progenitors in the bone marrow and relocate to various tissues where they acquire specific phenotypes through local environmental and signaling cues. Macrophage polarization leads to the formation of two distinctive phenotypes: the pro-inflammatory M1 subtype, which is induced by the Th1 cytokine interferon-γ, and the M2 phenotype, which is induced by the Th2 cytokines interleukin (Il)-4 or Il-13. The M2 phenotype plays a critical role in tissue remodeling and repair and is a crucial regulator of fibrogenesis in IPF. Following activation, M2 macrophages generate profibrotic mediators, such as TGF-β1, which activates fibroblasts and ECM deposition. The polarization of alveolar macrophages toward a profibrotic M2 phenotype is a contributing factor in the development of fibrosis (Geng et al. 2021). However, in a study by Misharin et al. 2017 the transcriptional data allowed to directly test this hypothesis by examining the expression of inflammatory and fibrotic genes during bleomycin-induced lung fibrosis. Interestingly, they found that both Mo-AMs and TR-AMs up-regulated inflammatory and fibrotic genes in response to bleomycin without a discernible shift in gene expression toward an inflammatory M1 or fibrotic M2 phenotype in either cell population (Misharin et al. 2017).

Fibroblasts comprise the predominant cell type in the connective tissues of the body and serve as the primary origin of the abundant ECM that characterizes these tissues (Kendall and Feghali-Bostwick 2014). Large amounts of profibrotic cytokines are produced by dysfunctional epithelial cells and polarized macrophages, which stimulate fibroblast differentiation into myofibroblasts. Depending on their activation state, macrophages can affect fibroblast gene expression by enhancing the expression of collagens, α-SMA, TGF-β, ECM synthesis, and fibroblast proliferation. Conversely, fibroblasts control the capacity of macrophages to produce pro-inflammatory cytokines like Il-6 and chemokines such as macrophage inflammatory protein 1α (MIP-1α) leading to a vicious loop (Novak et al. 2023). In IPF lungs, myofibroblasts display a pathological phenotype characterized by the excessive secretion of matrix within the lung parenchyma, leading to basement membrane disruption. Resident interstitial lung fibroblasts are the primary source of myofibroblasts in the lungs (Geng et al. 2021).

Additionally, fibroblasts/myofibroblasts exhibit distinct characteristics that contribute to lung fibrosis, such as increased proliferation, resistance to apoptosis, and invasive activity. While fibroblasts normally proliferate in response to tissue injury, they undergo apoptosis as the tissue returns to homeostasis. However, IPF fibroblasts resist apoptosis and show enhanced proliferation, possibly due to the abundance of PGF2α in IPF lungs that stimulates their proliferation and reduced PGE2 levels that contribute to their apoptotic resistance. Prostaglandins, such as PGE2, regulate many pathological features of lung fibroblasts and myofibroblasts, including proliferation, migration, collagen secretion, and TGF-β1-induced differentiation. IPF fibroblasts also invade the surrounding ECM similar to metastatic cancer cells, possibly due to de novo fatty acid synthesis or the upregulation of α-SMA-containing stress fibers through the C/EBP-β binding element in the α-SMA promoter (Geng et al. 2021).

In addition to alveolar epithelial cells, alveolar macrophages and fibroblasts, also broncoalveolar lavage (BAL) and biopsy samples were included in this study. BAL is a diagnostic method of the lower respiratory system in which a bronchoscope is passed through the mouth or nose into the airways in the lungs. A measured amount of fluid is introduced and then collected for examination. The BAL fluid of healthy individuals consists of macrophages (>80%). Normal in BAL may be 80–90 % alveolar macrophages, 5–15 % lymphocytes, 1–3 % polymorphonuclear neutrophils, 1% eosinophils, and <1 % mast cells (Stanzel 2012). BAL is known to contain an accumulation of extravasated inflammatory cells and cytokines that reflect the microenvironment around the alveoli. Several studies have confirmed that alterations in the alveolar microenvironment are associated with the advancement of idiopathic pulmonary fibrosis (Prasse et al. 2019b; He et al. 2022).

Lung biopsies are typically obtained from routine surgical resections, biopsy extractions during routine clinical care, or in a research setting involving healthy volunteers or lung transplantation programs. However, it can be difficult to obtain normal lung parenchyma. The most common sources for lung parenchyma are either uninvolved edges of lung resections (mostly for cancer) or tissue obtained from deceased donors through organ procurement organizations, which can be archived or processed fresh. Archived lung tissue is often obtained from large tissue banks and repositories established in many centers, often associated with academic, medical, or governmental organizations. To minimize tissue collection variability between different centers, standard principles for tissue processing and collection are being developed. Live tissue from bronchoscopic biopsies, organ procurement organizations, or surgical resections requires special attention after collection. Lung tissue biopsies can exhibit high cell-type heterogeneity depending on the precise location of surgery (Schiller et al. 2019).

Medication of IPF

Over the years, numerous medical therapies have been tested for treating patients with IPF, although they have mostly failed to demonstrate any benefits and have even caused harm (Dempsey et al. 2021; Raghu et al. 2015; Farrand et al. 2020; Biondini et al. 2020). Patients with IPF have a very poor response to anti-inflammatory medication such as corticosteroids (Jang et al. 2021; Farrand et al. 2020). Two antifibrotic drugs, pirfenidone and nintedanib have been approved for IPF treatment. Although these drugs can slow the progression of the disease in some patients, there is currently no cure for the disease (Cameli et al. 2022; Dempsey et al. 2021; Raghu et al. 2015; Sakamoto et al. 2013).

Studies have shown that patients treated with either pirfenidone or nintedanib have a lower risk of all-cause mortality and acute hospitalizations compared to untreated patients in a matched cohort. However, due to the rarity of the disease and its fatal nature, conducting a trial and ensuring appropriate follow-up to demonstrate a mortality effect is extemely challenging. Currently, pirfenidone and nintedanib treatments are the primary recommendations for IPF, as they have been shown to reduce the risk of all-cause mortality and acute hospitalizations when compared to untreated individuals (Dempsey et al. 2021; Raghu et al. 2015). Molecular structures of pirfenidone and nintedanib are illustrated in Figure 3.

Especially in the acute exacerbation of IPF (AE-IPF) the treatment remains undetermined. Studies have shown that the use of corticosteroids in these cases may lead to an increasing trend in in-hospital mortality, and the overall survival rate of patients receiving corticosteroid treatment may be significantly lower. There is no

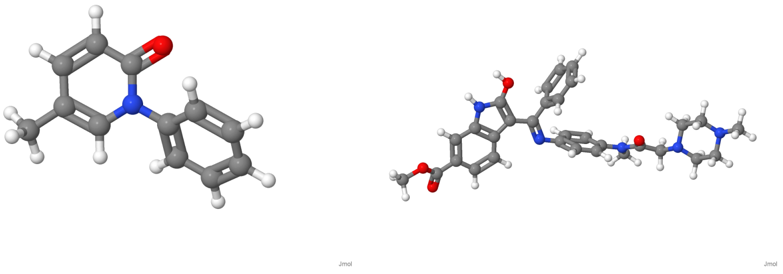

Figure 3: Pirfenidone (left) and nintedanib (right) molecular structures.

evidence that corticosteroid use improves outcomes in IPF patients admitted to the hospital with acute exacerbation (Farrand et al. 2020). Sometimes the corticosteroids are combined with immunosuppressants like cyclophosphamide. Efficacy of cyclophosphamide in AE-IPF is only suggested by small, retrospective non-randomized studies (Biondini et al. 2020).

Macrolides have been used to treat AE-IPF due to their anti-inflammatory and immunomodulatory properties, which may be beneficial beyond their antimicrobial effect. Additionally, macrolides have been suggested to facilitate alveolar epithelium regeneration following damage (Biondini et al. 2020; Guillot et al. 2011). In cases of AE-IPF, autoantibody production secondary to immune dysregulation has been suggested to contribute to the disease progression. To address this, critically ill patients with AE-IPF have been treated with therapeutic plasma exchange and rituximab, which has been supplemented in later cases with intravenous immunoglobulin. Also N-acetylcysteine (NAC) has been studied as a treatment in IPF. NAC is an antioxidant and mucolytic agent that has been investigated as a potential therapy for IPF. It has been suggested that combined therapy including NAC by oral administration may be more effective than monotherapy (Feng et al. 2019). Other treatments for AE-IPF are human recombinant thrombomodulin, hemoperfusion with polymyxin B-immobilized fibers and supportive treatments include substances like oxygen and opioids (Biondini et al. 2020).

Despite the benefits of pirfenidone and nintedanib treatments and the lack of alternative treatments, concerns remain regarding their cost-effectiveness. In the United States, these medications can cost over $100,000 per year per patient, which can add to the financial burden on IPF patients or society given the relatively poor effectiveness of these treatments (Dempsey et al. 2022). In Finland, the monthly cost of the original pirfenidone (Esbriet$^{\circledR}$) is 2272.25 €, while the cost of the generic product (Pirfenidone Ratiopharm$^{\circledR}$) 1175.78 € (801 mg twice a day) (Kansaneläkelaitos 2023). The monthly cost of nintedanib (Ofev$^{\circledR}$) ranges from 1380.56 to 2423.60 € depending on the dosage (Kansaneläkelaitos 2023). The price information provided by Kela was last updated on March 1st, 2023.

The cost-effectiveness of antifibrotic drugs in the USA has been shown to be suboptimal, with both pirfenidone and nintedanib having incremental cost-effectiveness ratios (ICERs) that are unreasonably high. In fact, nintedanib costs a staggering $1.6 million to gain just one additional quality-adjusted life year (QALY), a value 16 times higher than the commonly accepted willingness-to-pay threshold of $100,000. (Dempsey et al. 2022). On the other hand, a UK study found that nintedanib treatment resulted in fewer acute exacerbations and therefore fewer costs and more QALYs than pirfenidone. Based on the efficacy outcomes, over a patient's lifetime, nintedanib and pirfenidone gained 0.5 QALYs more than placebo. Given the high incremental cost difference between nintedanib, pirfenidone and placebo (NAC), the ICER was over £100,000 per QALY gained (Rinciog et al. 2017). These findings emphasize the need for more cost-effective treatments for IPF, and data-driven drug repositioning may be a promising approach for identifying such treatments.

Transcriptome investigation through omics techniques

RNA sequencing

Over the last decade, RNA sequencing (RNA-seq) has become a widely used tool in molecular biology, revolutionizing the understanding of genomic function. One of the most common applications of RNA-seq is differential gene expression (DGE) analysis, which follows a standard workflow. Starting with RNA extraction in the laboratory, the process involves mRNA enrichment or ribosomal RNA depletion, cDNA synthesis, and preparation of a sequencing library with adaptors. This library is then sequenced at a high read depth of 10-30 million reads per sample using a high-throughput platform, typically Illumina (Stark et al. 2019).

The final steps are computational, involving the alignment or assembly of sequencing reads to a transcriptome, quantification of reads that overlap with transcripts, filtering and normalization of data between samples, and statistical modeling of significant changes in gene and/or transcript expression levels between different sample groups. Despite advances in technology, the essential stages of the DGE assay have remained largely unchanged since its inception (Stark et al. 2019).

The Illumina short-read sequencing technology has been widely used to produce over 95 % of the published RNA-seq data found on the Short Read Archive (SRA). Since almost all of the available mRNA-seq data is obtained through short-read sequencing of cDNA, it is regarded as the standard RNA-seq technology, and its primary workflow and limitations are well known (Stark et al. 2019). The Ion Torrent protocol offers an alternative to Illumina technologies as it relies on pH measurements to read nucleotide sequences. Along with differences in sequencing technologies between these two platforms, there are also subtle variations in the data they generate. While Illumina data has uniformly-sized sequence reads in a single experiment, Ion Torrent reads have variable lengths (Lahens et al. 2017).

However, the emergence of long-read cDNA sequencing and dRNA-seq methods may soon replace short-read sequencing methods, as users look for methods that can provide better data on isoform levels. A DGE assay using the Illumina short-read sequencing platform involves several steps, including RNA extraction, cDNA synthesis, adaptor ligation, PCR amplification, sequencing, and analysis. The resulting cDNA fragments are typically less than 200 bp due to mRNA fragmentation and size selection during library purification. While RNA-seq is a robust technique, there are potential sources of imperfections and biases that can arise during both sample preparation and computational analysis. These limitations may impact the ability of the experiment to address specific biological questions, such as accurately identifying and quantifying which isoforms of a gene are expressed (Stark et al. 2019).

DNA Microarray

While hybridization-based approaches, including microarray technologies, have traditionally been used for expression profiling in toxicogenomics, they have gradually been replaced by RNA-seq and other next-generation sequencing methods (Nuwaysir et al. 1999; Rao et al. 2019; Zhao et al. 2014). Public repositories offer an abundance of microarray gene expression datasets that are readily accessible for research purposes (Taminau et al. 2012).

The genetic information encoded in DNA is expressed as proteins through the intermediate step of mRNA synthesis and subsequent translation. Gene expression can be indirectly assessed by analyzing the various mRNAs present in a sample. However, mRNA is relatively unstable and needs to be converted into a more stable form, such as cDNA, before further analysis. To label cDNA, fluorescent dyes such as Cy3 (green) and Cy5 (red) are used. Microarrays work based on the principle that complementary sequences will bind to each other. In this technique, unknown DNA molecules are fragmented using restriction endonucleases, and fluorescent markers are attached to these DNA fragments. The labeled fragments are then allowed to bind to probes on a DNA chip, and unbound fragments are washed away. The bound DNA fragments can be identified by their fluorescence emission upon excitation with a laser beam, and a computer records the pattern of fluorescence emission and DNA identification (Govindarajan et al. 2012).

RNA-Seq and microarrays differ primarily in that RNA-Seq enables complete sequencing of the entire transcriptome, whereas microarrays only allow for the profiling of predetermined transcripts/genes via hybridization (Rao et al. 2019). There are two primary microarray platforms commonly used: Affymetrix and Agilent. While there are similarities between the two, there are also several differences. One significant distinction is that Agilent may use a two-color detection method, whereas Affymetrix uses a single-color detection scheme. Another difference is that Agilent typically requires only one 60-mer per gene or transcript, while Affymetrix employs multiple 25-mers per transcript. Additionally, Agilent is considered to be highly reproducible and the most sensitive of the available array platforms, while Affymetrix is a well-established platform with an extensive array catalog, but requires a costly scanner upgrade for the higher density arrays. (Hardiman 2004).

FAIR data and batch effect

Effective data management should not be viewed as an end goal. Rather, it serves as a crucial pathway to uncovering new knowledge and fostering innovation. Moreover, it facilitates the integration and reuse of data and knowledge by the wider community, following the data publication process. Unfortunately, the current digital ecosystem that surrounds the publication of scholarly data hinders the ability to fully capitalize the research investments. As a partial response to this challenge, entities such as science funders, publishers, and government agencies are now demanding that researchers who receive public funding for their research projects must have a plan for managing and stewarding the data generated during their research. Effective data stewardship goes beyond mere collection, annotation, and archival, encompassing the idea of providing long-term care for valuable digital assets to ensure their discovery and reuse in future investigations, either on their own or in conjunction with newly generated data. There are four fundamental principles, Findability, Accessibility, Interoperability, and Reusability (FAIR), that guide the data producers and publishers to navigate these challenges. By following these principles, it is possible to optimize the benefits of modern, formal scholarly digital publishing (Wilkinson et al. 2016).

Public repositories contain growing amounts of omics data, which provides a diverse and valuable source of prior knowledge. This data can be particularly useful in fields such as chemical risk assessment and pharmacological research. Although standards for representing omics data have been clearly established, a significant portion of this data is not fully compliant with the FAIR principles. As a result, the integration and utilization of this data is limited (Wilkinson et al. 2016). Rigorous data curation is needed to meet this challenge (Odell et al. 2021). Although text mining and AI techniques have been suggested as potential solutions to automate the process, an experienced curator is still irreplaceable, despite the high accuracy of these methods. One aspect of this study was to test and evaluate a software named ESPERANTO, a R/Shiny (Chang et al. 2022) application, which aims to facilitate a simplified, semi-supervised curation process for omics metadata. This approach involves active user participation in harmonizing data within a standardized framework and improving data FAIRness, combining the benefits of both automated and manual curation methods. Regardless of the user's expertise, the interactive graphical interface guides the user through the whole data curation and integration pipeline. The purpose of ESPERANTO is to produce reproducible, and high-quality curated metadata (Di Lieto et al. 2023).

Non-biological variables can affect the data obtained from microarray or RNA-seq experiments. Batch effects may arise from several sources, including variations in ambient conditions during sample preparation, handling, amplification, labelling, hybridization protocol, use of different sites or laboratories, various chip or platform types, and varying scanners. These batch effects can negatively impact data quality and lead to inaccurate or erroneous conclusions (Federico et al. 2020b; Kupfer et al. 2012). Saarimäki and colleagues found that 35 datasets had to be excluded from data quality assessment due to problems with their overall usability, primarily stemming from experimental design issues. This suggests that many published toxicogenomics datasets may have significant design flaws that could compromise the validity of any findings derived from them. This highlights the importance of critically evaluating even FAIRified data. Proper study design and randomization are crucial in dealing with batch effects, as the most effective batch correction methods like ComBat may not be able to correct data with improper study design (Buhule et al. 2014; Federico et al. 2020b; Saarimäki et al. 2022).

For moderate batch effect there are correction methods. The SVA package includes the R ComBat function, which can be utilized to mitigate the effects of known batch variables and surrogate variables that are not associated with the variables of interest. ComBat implements an empirical Bayes approach to estimate systemic batch biases that affect large sets of genes. To perform batch correction using ComBat, the variable of interest, any biological covariates, and a set of known batches or surrogate variables are specified. However, it is important to note that each run of the ComBat function can only address the effect of one batch variable, and any additional variables that cause known batch effects can be directly added to the linear model used for differential expression analysis. It is essential to keep in mind that SVA can remove the effect of any biological information not addressed by known phenotype-related variables, such as phenotypic subgroups of interest. Therefore, an alternative linear modeling approach, such as the limma R package, can be used to investigate the effect of covariates included in the model. Principal Component Analysis (PCA) is a valuable tool that can help identify features that are affected by batch surrogates. PCA can capture both biological and technical variability and can quantify the effects of artifacts in the data, especially when estimated after accounting for the biological variables (Federico et al. 2020b).

Real-world networks and community detection

The world around us is filled with incredibly complicated systems. Take for instance, society, which necessitates collaboration among billions of individuals, or communication networks that integrate billions of cell phones with computers and satellites. Our ability to comprehend and make sense of the world requires the coherent activity of billions of neurons in our brains. Similarly, our biological existence relies on the smooth interactions between thousands of genes and metabolites within our cells. Collectively, these systems are referred to as complex systems, highlighting the challenge of predicting their collective behavior from a knowledge of their individual components. Given the crucial role these complex systems play in our daily lives, as well as in science and economy, comprehending, mathematically describing, predicting, and ultimately controlling them represents one of the major intellectual and scientific challenges of the 21st century (Barabási 2016).

To comprehend the intricate mechanisms of a system that involves hundreds to billions of interconnected elements, we require a comprehensive blueprint of the system's wiring diagram. In a social network, this would entail having an exact record of your friends, the friends of your friends, and so on. In the internet, this map shows us the interlinking of web pages. In a cell, the map corresponds to a detailed list of chemical reactions and binding interactions that involve genes, metabolites, and proteins (Barabási 2016). The basic parameters for mapping these complex networks are nodes (vertices) and edges (links). Nodes represent the components in the system and edges represent the interactions between the nodes. The degree of a node, which represents the number of connections it has to other nodes, is a fundamental characteristic of the node. The degree represents how many contacts a certain node has to other nodes. In undirected networks the total number of edges can be represented as the sum of node degrees where ki is the degree of an individual node. In an undirected network the total number of links can be expressed as, L:

$$L = \frac{1}{2} \sum_{i=1}^{N} k_i$$

A network can be directed or undirected. An example of an undirected network would be a social network, where friendships are established bidirectionally. If person A is friends with person B, then person B is also friends with person A. An example of a directed network would be a transportation network, where the flow of traffic moves in a specific direction, such as a highway or one-way street. Cars can only travel in one direction on these roads, creating a directed network. Network edges may also have a weight that are used to measure the strength of relationships between nodes. For example in social networks the weight of each edge may represent the strength or frequency of the relationship between two individuals (Barabási 2016).

Real-world networks are characterized by several properties that set them apart from random or regular networks. Scale-free topology is a feature of many real-world networks, such as the internet, social networks, and biological networks. In these networks, there are a few highly connected nodes (hubs) and many poorly connected nodes, resulting in a degree distribution that follows a power-law distribution. Another property is the small-world phenomenon, where most nodes can be reached from any other node by a relatively small number of steps, leading to a short average path length and a high clustering coefficient. Additionally, many real-world networks exhibit a community structure, where nodes form tightly connected groups or communities. Real-world networks are often robust to random failures but vulnerable to targeted attacks on highly connected nodes. Furthermore, many real-world networks evolve over time and undergo changes such as growth, rewiring, and deletion of nodes and edges. These properties have been observed in various real-world networks, including social networks and biological networks (Barabási 2016; Newman 2010).

Community detection in networks can be achieved by using various algorithms, two of which are the walktrap algorithm and the Louvain algorithm. The walktrap algorithm, a hierarchical clustering algorithm based on random walks, identifies communities in the network by clustering together vertices that are more likely to be visited by a random walk of a given length. The algorithm simulates random walks on the graph and creates a dendrogram based on the results, where initially each node is considered as a separate community. The algorithm then merges communities based on the similarity of their random walk trajectories, with the merging process repeated until a stopping criterion is reached. This can be a minimum number of communities or a maximum depth of the dendrogram. By cutting the dendrogram at a chosen level, a partition of the graph into communities can be obtained. The algorithm is efficient, handles large-scale networks, and works well with both directed and undirected sparse networks. Research has shown that the walktrap algorithm performs well in identifying communities in real-world networks and outperforms other community detection methods in terms of accuracy (Pons and Latapy 2005; Lancichinetti et al. 2009).

The Louvain algorithm is a widely used community detection algorithm for large-scale networks (Blondel et al. 2008). It is a bottom-up approach that aims to optimize a modularity score to identify communities in the network. Modularity is a measure of the density of edges within communities compared to the density of edges between communities. Initially, the algorithm assigns each node to a separate community and then iteratively merges communities to maximize the modularity score. At each iteration, the algorithm evaluates the modularity gain that would result from merging each pair of communities connected by at least one edge. The algorithm selects the pair that would result in the largest modularity gain and merges them into a new community. The process continues until no further modularity gain can be achieved. The Louvain algorithm has several advantages. It is fast, scalable, and can handle large networks with millions of nodes and edges. It is also relatively easy to implement and does not require extensive parameter tuning. Additionally, the algorithm can identify communities at multiple scales, revealing both smaller and larger communities within a network. The Louvain algorithm has been applied to a wide range of networks, including social networks, biological networks, and transportation networks, and has been shown to perform well compared to other community detection algorithms in terms of both speed and accuracy (Blondel et al. 2008; Held et al. 2016; Rahiminejad et al. 2018).

Systems pharmacology

Graph theory can be a useful tool for revealing significant gene-gene expression associations, both in normal physiological states and under pathological conditions. Currently, gene co-expression network analysis is used to explore the relationship between pairs of genes and to identify gene networks or modules that serve as a marker of dysfunctional biological processes in a disease (Federico et al. 2022). Graph theory tools such as clustering algorithms, centrality measures, and community detection algorithms can be used to identify functional modules within these networks, predict new interactions, and understand the overall topology and dynamics of the network. Systems biology, on the other hand, provides the biological context for understanding the function and behavior of these networks, by integrating experimental data from multiple sources and applying computational methods to model and simulate biological processes. Together, graph theory and systems biology provide a powerful framework for understanding the complex and dynamic behavior of biological systems at the molecular level (Federico et al. 2022; Gomez-Cabrero et al. 2014).

Systems pharmacology is an interdisciplinary field that combines the principles of systems biology, drug pharmacology, physiology, mathematics, and biochemistry to understand the mechanisms of drug action in a holistic and quantitative manner. The goal of systems pharmacology is to identify drug targets, optimize drug efficacy, and minimize drug toxicity by considering the complex interactions between drugs, targets, and biological networks. This involves integrating high-throughput experimental data, such as genomics, proteomics, and metabolomics, with computational and mathematical models to predict the effects of drugs on biological systems. Systems pharmacology has various applications in drug discovery, drug repurposing, and personalized medicine (Leil and Ermakov 2015).

Quantitative Systems Pharmacology (QSP) has the potential to revolutionize drug discovery and development for complex multi-factorial diseases, such as Alzheimer’s, multiple sclerosis, and IPF, which can have devastating impacts on patients and their loved ones. These diseases involve multiple physiological processes, can affect multiple organs, and often have limited treatment options. QSP can provide an integrated understanding of the pathology and the possible complex results of therapeutic interventions, offering an innovative approach for pharmaceutical research and development, particularly in diseases that are poorly translated from animal models (Leil and Bertz 2014).

Aims of the project

The primary objective of this project is to enhance the understanding of the molecular mechanisms underlying idiopathic pulmonary fibrosis (IPF) by identifying genes that could serve as potential biomarkers, and to investigate new medication possibilities for treating IPF. Specifically, data will be analyzed from multiple cell types to identify targets in pro-inflammatory and pro-fibrotic pathways, with the aim of finding effective combination therapies. Additionally, the current state of FAIRness (findable, accessible, interoperable, and reusable) of toxicogenomics data in public repositories will be evaluated. Another goal is to integrate network-based methods with meta-analytical approaches in the context of systems pharmacology and drug repositioning. Through these efforts, the aim is to advance our understanding of idiopathic pulmonary fibrosis and improve current therapeutic approaches.

Materials and methods

Tool versions

For downloading the metadata R version 4.2.1 was used on a local laptop with GEOquery_2.66.0 (Davis and Meltzer 2007). For metadata curation ESPERANTO (sEmi SuPERvised meta-dAta curatioN TOol) alpha testing versions were used (Di Lieto et al. 2023). For downloading the FASTQ- files of the RNA-seq data from European Nucleotide Archive (https://www.ebi.ac.uk/ena/browser/) the tool used was axel-2.15. FastQC v0.11.7 was used for the quality control of the FASTQ-files (Andrews 2010). For trimming and filtering the adapters of the reads cutadapt version 3.7. was used (Martin 2011b). For aligning the reads Hisat2-2.2.1 version was used with grch38 as an index (Kim et al. 2019; Kim et al. 2015; Pertea et al. 2016; Zhang et al. 2021). For filtering and sorting the uniquely mapped reads from the BAM -files samtools version 1.8-27-g0896262 was used (Danecek et al. 2021). Axel, FastQC, cutadapt, samtools and Hisat2 were used in Unix Bash.

For constructing the count matrices the featureCounts from Rsubread_1.34.4 package was used in R version 3.6.0 (2019-04-26) (RStudio in server). For differential expression analysis of the Microarray datasets eUTOPIA tool was used (Marwah et al. 2019). Inside eUTOPIA arrayQualityMetrics_3.54.0, sva_3.46.0, minfi_1.44.0 and limma_3.54.0 were used (Fortin et al. 2017; Kauffmann et al. 2009a; Leek et al. 2023; Ritchie et al. 2015a). For normalization of differential gene expression analysis for RNA-seq data DESeq2_1.24.0 was used (Love et al. 2014). In gene annotations gene symbols will be used. The annotation was performed with org.Hs.eg.db_3.8.2 and biomaRt_2.40.1 (Carlson 2019; Durinck et al. 2009; Durinck et al. 2005). biomaRt was used especially with microarray data where the genes were annotated as probe names.

The adjustment of the expressiondata from different sources was performed with pamr.batchadjust function from pamr_1.56.1 package in R (Trevor et al. 2022). The adjusted expressiondata was used for the network inference. In meta-analysis for calculating the effect-size gene rank e esc_0.5.1 package was used (Alessio et al. 2021). For p-value based rank metap_1.8 package was used (Mavridis and Higgins 2021). For computing the Borda count TopKLists_1.0.8 package was used (Schimek et al. 2015). The GSEA analysis was performed with fgsea_1.24.0 package (Korotkevich et al. 2019).

The networks were built with INfORM functions offered by FHAIVE in R (Marwah et al. 2018). The functions used for generating the ranked consensus matrix in order to build the networks is get_ranked_consensus_matrix, parse_edge_rank_matrix and get_iGraph. igraph_1.3.1 was used for making the graph objects (Csardi and Nepusz 2006). ComplexHeatmap_2.14.0 package was used for plotting the heatmaps (Gu et al. 2016). For the enrichment analysis of the networks clusterProfiler_3.12.0 and ReactomePA_1.28.0 packages were used (Wu et al. 2021; Yu et al. 2012).

Datasets

The microarray data provided in this study were collected from the NCBI Gene Expression Omnibus (GEO) public repository, while the RNA-Seq datasets were retrieved from the European Nucleotide Archive (ENA). In total, 1062 samples were collected, consisting of 634 disease samples and 428 healthy samples, across 25 datasets. The original datasets are represented in tables 1 and 2. In table 1 are represented the datasets that are derived from some specific cell type. In table 2 are represented the datasets that are from biopsy samples. In table 3 are represented the number of datasets RNA-seq and Microarray datasets and the number of disease and healthy samples in each cell type and protocol. In both tables 1 and 2, a number of datasets have a higher number of samples than biological replicates, causing a discrepancy between the total number of samples and the "disease/treatment + healthy" columns. It is worth noting that, for the disease group, only the IPF samples were included.

In the analysis of dataset GSE70866 (Prasse et al. 2019b), only the samples from Freiburgh were included (62 IPF and 20 healthy samples) due to a significant batch effect observed between samples collected from different cities (Supplementary Figure 57). In dataset GSE49072 (Yuanuan et al. 2013) There are familial IPF, spontaneous IPF (15), familial IPF (8), healthy volunteer (45) and healthy relative (16) samples. Only the spontaneous IPF and healthy relative samples were used in the analyses. The healthy relative samples are from the relatives of the familial IPF patients. The experimental design in the dataset GSE90010 (Marwick et al. 2018) is slightly different from the other datasets. There are two experimental sets. In the first experimental set there is 4 samples per group and 4 match paired groups: human monocyte-derived macrophages (MDM) co-cultured with or without apoptotic neutrophils and with or without lipopolysaccharides (1 ng/ml) for 9 hours. In the second experimental set (n=4 per group, 2 groups) alveolar macrophages from IPF and RB-ILD are patients isolated from bronchoalveolar lavage by cell sorting, RNA isolated using Qiagen RNeasy Kit.The decision to include this dataset is due to the lack of suitable datasets containing alveolar macrophages. Only the alveolar macrophage samples from IPF were used, using the monocyte derived macrophage control group as a control.

In dataset GSE185492 (the dataset does not provide a reference) the fibroblast samples have been collected from apical and basal regions of the lung. Both regions are in the healthy and in the disease samples and they all were used in this study. Each sample in dataset GSE185492 consisted of two single-end FASTQ read files, which were merged for downstream analyses. In dataset GSE40839 (Lindahl 2013) in the disease samples there are samples from systemic sclerosis-associated interstitial lung disease (SSc-ILD) patients (8) and IPF patients (3). Only the samples from IPF patients were used in the disease group. In GSE11196 (Larsson et al. 2008) from each biological replicate the total RNA and heavy ribosomial RNA have been sequenced. Only the total RNA samples were used in this study. In GSE11196 they also used contractile and non-contractile gel matrices as a growth medium and the samples in both of the growth mediums were used in this study. In the dataset GSE44723 (Peng et al. 2013) they compared bleomycin induced fibrosis in mice to the samples from fibrotic IPF samples from humans. In the study by Peng et al. 2013, they have samples from rapid (4) and slow (6) progressing IPF and 4 healthy samples. Both samples from rapid and slow patients were used in this study. The dataset GSE45686 (Parker et al. 2014) was decided to discard since the sequencing process was done with Illumina HumanHT-12 V4.0 microarray technique and eUTOPIA can only process Affymetrix, Agilent and Illumina methylation data in terms of microarray data (Marwah et al. 2019). The decision was based on two additional reasons: first, the time used for MSc thesis is based on 40 credits, and there was already a sufficient amount of data for MSc thesis work, and the workload is quite extensive in relation to the project. Second, there were already several datasets from fibroblasts.

The dataset GSE150910 (Furusawa et al. 2020) includes samples from patients with IPF (103) and hypersensitivity pneumonitis (HP) (82). Only the IPF (103) and healthy (103) samples were used. Dataset GSE213001 (the dataset does not provide a reference) contains samples from patients with IPF and healthy controls, and only these samples were included in the study. The ILD and CLAD (chronic lung allograft dysfunction) were excluded from this study. There are several samples from each replicate from different lung locations (apical & basal) and left and right lungs and the number of samples from each biological replicate differs within each biological replicate. In GSE124685 (McDonough et al. 2019) several samples were collected from each biological replicate. The number of samples from each biological replicate varied. In dataset GSE199949 (the dataset does not provide a reference) from each biological replicate a biopsy from central and peripheral lung regions from each IPF and healthy lungs were obtained. In dataset GSE199152 (the dataset does not provide a reference) there are 23 disease samples (20 UIP) and 3 (RA-ILD) and 4 healthy samples. The samples from patients with usual interstitial pneumonia (UIP) were included in the IPF group, as these two diseases are often used synonymously (Bois and K. 2007). This quote is from the GSE199152 page on the GEO website: "*The goal of this study is to compare transcriptome profiling (RNA-seq) of lung tissue biopsies derived from patients with idiopathic pulmonary fibrosis (IPF or IPF-UIP).*". The RA-ILD samples from GSE199152 dataset were excluded from this study.

In dataset GSE166036 (DePianto et al. 2021), there are 28 samples from patients with IPF and 8 samples from patients with SSc-ILD. However, only the IPF samples were used in this study. The dataset includes both BAL and biopsy samples. For network inference, the IPF BAL samples (8) and healthy BAL samples (4) were used separately. For the biopsy samples, there were 10 IPF samples and 3 healthy samples. The remaining samples in this dataset, which were from patients with SSc-ILD and cell digest samples, were excluded from the analysis. In dataset GSE1834316 (the dataset does not provide a reference), both IPF (10) and fibrotic hypersensitivity pneumonitis (fHP) samples are available, but only the IPF samples were included in this study. The fHP samples were excluded. In dataset GSE21369 (Cho et al. 2011), there were multiple diseases represented: 11 UIP, 5 non-specific interstitial pneumonia (NSIP), 2 cryptogenic organizing pneumonia (COP), 2 respiratory bronchiolitis-interstitial lung disease (RB-ILD), 2 HP, and 1 uncharacterized fibrosis (UF). However, only the UIP samples were used, along with 6 healthy samples, in this study.

For GSE173355 (Konigsberg et al. 2021) the raw sequencing data was not available, but the DESeq2 normalized data was usable since the same normalization method was used in this study as in GSE173355. In dataset GSE24206 (Meltzer et al. 2011), there are 1-2 samples from each biological replicate, with samples from upper and lower lobes of the lung and from advanced and early stage IPF. All the IPF samples were analyzed together with the healthy samples, but disease stage was used as a covariate in the analysis. Since dataset GSE76808 (Christmann et al. 2014) only contains samples from patients with SSc-ILD and no IPF samples, the whole dataset was excluded from this study. In dataset GSE72073 (Geng et al. 2015), the control samples were from normal tissue of primary spontaneous pneumothorax patients. In dataset GSE169500 (Brereton et al. 2022), the samples are grouped by disease and tissue source, and are bulk fixed paraffin embedded (FFPE) samples. There were healthy non-fibrotic lung tissue samples from alveolar septae (n=10), as well as UIP/IPF FFPE lung tissue myofibroblast foci (n=10) and adjacent non-affected alveolar septae (n=10). The comparison in this study was between the disease and healthy samples. The dataset GSE94060 (Megan et al. 2020) was excluded from the analysis because all the genes had an adjusted p-value of 1 in the differential gene expression analysis. The number of replicates in the data was low, and it was clear from the PCA plot (supplementary figures) that the IPF samples were clustered between the healthy samples, indicating that there was no statistical power in the analyses. In dataset GSE99621 (Luzina et al. 2018), there are three biological replicates in both disease and healthy groups. There are two to three samples from each replicate in the healthy group so that there are eight samples in the healthy group altogether. In the disease group, there are altogether 18 samples and 4-7 replicates per each biological replicate. The samples have been taken from scarred and normal tissue, and the scarring was used as a covariate in the analyses in this study.

Table 1: Datasets used for the study that contain other than biopsy as a cell type. Note that the contents of "GEO-dataset" and "Citation" columns are hyperlinks. The disease/treatment column numbers refer to biological replicates (if known) and number of samples are the samples in GEO.

GEO-dataset	Platform	Citation	Disease/treatment	Healthy	Number of samples	Cell type
1. GSE70866	Agilent-028004 SurePrint G3 8x60K	30141961	176 (IPF)	20	196	BAL
2. GSE151673	Illumina NextSeq 500	32605572	(IPF) 5	5	10	Epithelial from mesenchymal stem cells
3. GSE49072	Affymetrix U133A	23924348	15 (IPF) + 8 (FPF) = 23	61	84	Alveolar macrophage
4. GSE90010	Affymetrix 2.1 ST Array	29867198	4 (IPF) + 4 (RBILD) = 8	4	12	Alveolar macrophage
5. GSE185492	Illumina NovaSeq 6000	NA	12 (IPF)	12	24	Fibroblast
6. GSE40839	Affymetrix U133A	23915349	(SSc-ILD)8 + (IPF) 3 = 11	10	21	Fibroblasts
7. GSE11196	Affymetrix U133 Plus 2.0	187951029	(IPF) 6	6	24	Fibroblasts
8. GSE44723	Affymetrix U133 Plus 2.0	23565148	(IPF) 10	4	14	Fibroblasts
9. GSE45686	Illumina HumanHT-12 V4.0	24590289	(IPF) 5	5	40	Fibroblast

Table 2: Datasets used for the study that contain biopsy tissue as a cell type. Note that the contents of "GEO-dataset" and "Citation" columns are hyperlinks. The disease/treatment column numbers refer to biological replicates (if known) and number of samples are the samples in GEO.

GEO-dataset	Platform	Citation	Disease/treatment	Healthy	Number of samples	Cell type
10. GSE150910	Illumina NovaSeq 6000	32602730	(IPF) 103 + (HP) 82 = 185	103	288	Biopsy
11. GSE213001	Illumina HiSeq 3000	NA	(IPF) 20 + (ILD) 9 + (CLAD) 3 = 32	14	139	Biopsy
12. GSE124685	Ion Torrent Proton (Homo sapiens)	31600171	(IPF) 10	6	84	Biopsy
13. GSE199949	Illumina HiSeq 4000	NA	(IPF) 13	8	42	Biopsy
14. GSE92592	Illumina HiSeq 2000	28230051	(IPF) 20	19	39	Biopsy
15. GSE199152	Illumina HiSeq 2500	NA	(UIP) 20 + (RA-ILD) 3 = 23	4	27	Biopsy
16. GSE53845	Agilent-014850 4x44K G4112F	25217476	(IPF) 40	8	48	Biopsy
17. GSE166036	Illumina HiSeq 4000	33705361	(IPF) 28 + (Sc-ILD) 8 = 36	13	49	Biopsy, BAL
18. GSE184316	Ion Torrent Proton	NA	(fHP) 9 + (IPF) 10 = 19	6	100	Biopsy
19. GSE110147	Affymetrix 1.0 ST Array	30111332	(IPF) 37	11	48	Biopsy
20. GSE21369	Affymetrix U133 Plus 2.0	21241464	(UIP) 11 + 5 (NSIP) + 2 (COP) + 2 (RB-ILD) + 2 (HP) + 1 (UF)=23	6	29	Biopsy
21. GSE173355	Illumina NovaSeq 6000	34038697	(IPF) 24	14	37	Biopsy
22. GSE24206	Affymetrix U133 Plus 2.0 Array	21974901	(IPF) 11	6	23	Biopsy
23. GSE76808	Affymetrix U133A 2.0 Array	24574232	(SSc-ILD) 7	4	18	Biopsy
24. GSE72073	Affymetrix Transcriptome Array 2.0	26453058	(IPF) 5	3	8	Biopsy
25. GSE169500	Ion Torrent PGM	35188460	(IPF) 10	20	30	Biopsy
26. GSE94060	Affymetrix 1.0 ST Array	32533805	(IPF) 5	4	9	Biopsy
27. GSE99621	Illumina HiSeq 2500	29329637	(IPF) 3	3	26	Biopsy

Table 3: Datasets from different cell types. The number below RNA-seq represents the number of RNA-seq datasets and the number below Microarray represents the number of Microarray datasets. The number below **D** represents the number disease samples and number below **H** represents the healthy samples.

Biopsy				**Macrophage**				**Fibroblast**				**Epithelial**				**BAL**			
RNA-seq		**Microarray**		**RNA-seq**		**Microarray**		**RNA-seq**		**Microarray**		**RNA-seq**		**Microarray**		**RNA-seq**		**Microarray**	
11		5		0		2		1		3		1		0		1		1	
D	**H**	**D**	**H**	**D**	**H**	**D**	**H**	**D**	**H**	**D**	**H**	**D**	**H**	**D**	**H**	**D**	**H**	**D**	**H**
391	278	112	34	0	0	19	49	12	12	25	26	5	5	0	0	8	4	62	20

Disease	**Healthy**	**Disease**	**Healthy**	**Disease**	**Healthy**	**Disease**	**Healthy**	**Disease**	**Healthy**
503	312	19	49	37	38	5	5	70	24

Overall	
Disease	**Healthy**
634	428

Metadata curation

The metadata curation was executed with the ESPERANTO software. ESPERANTO is a R/Shiny (Chang et al. 2022) application performing a streamlined semi-supervised curation of transcriptomics metadata in compliance with GLP standards if needed. The user is actively involved in data harmonisation in a consistent framework. This approach enhances the data FAIRness, merging the advantages of both automated and manual curation approaches. The interactive graphical interface guides the user through the whole data curation and integration pipeline. ESPERANTO ensures a streamlined and standardized transcriptomics metadata harmonisation and the construction of a custom vocabulary. Both the metadata and vocabulary are essential inputs, or optionally, a previously saved curation session can be restored. During the curation process, the software assists the decision-making of the user on each modification before implementing it on the data and records the operation automatically (Di Lieto et al. 2023).

ESPERANTO performs the harmonisation of the uncurated transcriptomics metadata through the cross-comparison with a precompiled reference vocabulary. Essentially, the curation process takes place in three stages: 1) consistent renaming of table columns 2) deletion of column duplicates 3) potential editing of any remaining content. Any cross-comparison result is shown to be validated, or alternatively the user can edit all unique instances, storing the processed terms in a temporary-vocabulary. Once multiple datasets have been curated, the tool allows to evaluate their integration, highlighting potential inconsistencies among them and/or the vocabulary through a semaphore color system. The vocabulary is distinguished by 2 linked pairs of key:synonym(s) dictionary type objects (Lai 2022). Each feature in the metadata ties to all potential instances of the feature; both feature and instances are associated to their synonyms (i.e. "sex", "gender", "Male, Female", "M,F", "1,0"). Each round of curation offers the possibility to customise the vocabulary with new features and instances. Evaluation follows the color-coded mechanism established in the integration of multiple datasets (Di Lieto et al. 2023).

All the datasets underwent metadata curation and harmonisation. The curation process of the metadata consisted in the definition and usage of a common data model for all of the collected datasets. The dictionary, made with ESPERANTO curation process, to which the raw meta-data were mapped to, is reported in suppelmentary data. The data dictionary describes all the variables reported in the final metadata tables. For each variable, the description, type and allowed values are reported. This procedure significantly increases the FAIRness of publicly available IPF transcriptomics data and represents a valuable "ready-to-use" resource available to the scientific community (Di Lieto et al. 2023; Federico et al. 2020a; Wilkinson et al. 2016).

Preprocessing the RNA-seq data and differential gene expression analysis

The state of the art of a typical RNA-Seq preprocessing pipeline is shown in Figure 4. The process comprises the following steps: quality check, reads alignment, raw counts extraction, counts normalization and filtering and batch effect estimation and correction (Conesa et al. 2016; Federico et al. 2020b).

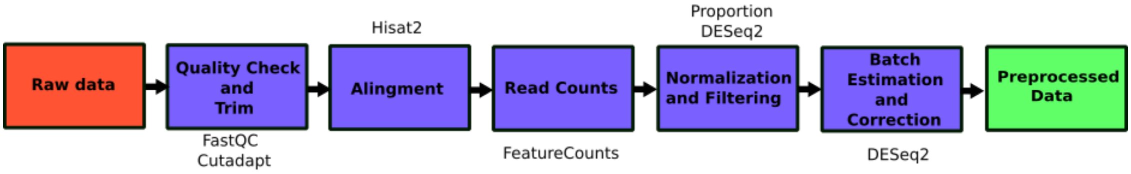

Figure 4: RNA-seq preprocessing pipeline.

Deep sequencing procedures may suffer from biases, which should be detected and corrected through an accurate quality check prior to upstream analyses. A pre-analytical check of the quality of the extracted RNA is necessary. There is not a current consensus to confirm whether a sample is unusable based on the levels of RNA degradation. The first step for an accurate analysis of sequencing-based transcriptomics data is the quality check of the raw reads. This step is essential in order to highlight biases and library contamination that might have occurred during the library preparation or sequencing procedure. One of the most used software to perform a quality check of the raw reads is FastQC and it was also used in this study. FastQC enables the analysis of multiple samples and offers a graphical interface. The Phred score, which quantifies quality, is determined using the following formula:

$$p = -10 * log_{10}(p)$$

where "p" represents the probability of an error in base-calling (Ewing and P. 1998; Federico et al. 2020b). In general, when the Phred score of the 3' bases of reads exceeds a certain threshold, it is advisable to "trim" the reads to achieve acceptable quality throughout the entire sequence. Cutadapt (Martin 2011a), TrimGalore and the FASTX are the most commonly used read trimmers. These software tools can remove residual adapters used in library construction as well as trim low-quality bases (Federico et al. 2020b). In this study Cutadapt was used for low quality read trimming. The settings for Cutadapt were a Phred quality score of 20 was selected, indicating a 1% chance of finding an incorrect base call among 100 bases. Additionally, a minimum read length of 60 bases was utilized.

The next step involves aligning the RNA-Seq reads to a reference genome to assign each sequenced read to a specific location on the genome or transcriptome. It is important to note that transcripts in eukaryotic genomes consist of non-contiguous regions (e.g., introns), and as a result, alignment tools for RNA-Seq reads should be able to handle spliced alignment with exceedingly large gaps. In this study, the HISAT2 algorithm was employed for read alignment. HISAT2 utilizes an innovative and efficient genome indexing method called Graph-based FM index

(GFM), which is an extension of the Burrow-Wheeler Transform (BWT). This method allows for the alignment of sequencing reads against both a genome representative of the general human population and a single reference genome. Additionally, HISAT2 utilizes a vast collection of local indexes covering genomic regions of 56 kilobases (Kim et al. 2019).

Assigning the mapped reads to specific genomic features such as genes or exons is critical after completing read mapping. This step is essential for quantifying gene/transcript expression and selecting the appropriate annotation is crucial. There are currently numerous annotations available, produced by different consortia. Quantifying gene expression accurately is crucial for transcriptome analysis in a toxico/pharmacogenomics setting as it aims to study the possible deregulation of gene expression upon exposure to a particular compound in a biological system. The aligned raw reads are summarized into a count matrix in this step, which can be utilized for differential expression analysis (Liao et al. 2019; Federico et al. 2020b). The count matrix in this study reports the genes in rows and samples in columns. The Bioconductor Rsubread package was utilized to construct the count matrices. This package contains functions for various steps of preprocessing NGS reads, as well as alignment and read summarization. Specifically, the featureCounts function was utilized for generating the matrices (Liao et al. 2019).

Preprocessing the microarray data and differential gene expression analysis

The preprocessing and differential expression analyses of the microarray datasets were performed with eUTOPIA software. eUTOPIA has been developed with R programming language, and it includes a graphical interface layer created with the R Shiny web development framework (Chang et al. 2022). eUTOPIA has the capability of processing data from four different microarray platforms: Agilent gene expression two-color microarray data (Samples specific to different colors channels), Agilent gene expression one-color microarray data, Affymetrix gene expression microarray data, and Illumina methylation microarray data (Marwah et al. 2019). In this particular study, data from all platforms, with the exception of methylation data, were analyzed. The microarray preprocessing pipeline is illustrated in the Figure 5.

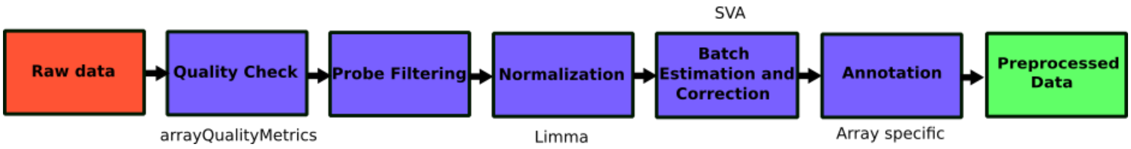

Figure 5: Microarray preprocessing pipeline.

There are seven main steps in the workflow: 1. data input, 2. quality control & filtering, normalization, batch correction & annotation and differential analysis eUTOPIA requires that the user provides a phenotype information file (metadata) with all biological and technical variables of the samples in the experiment. Depending on the microarray platform being used arrayQualityMetrics R package was utilized to generate a quality control report from the raw microarray data (Kauffmann et al. 2009b). Before normalizing the data, it is crucial to exclude any low-quality probes from the experimental dataset. To evaluate the quality of probes in gene expression platforms, their robustness against the background signal (measured by negative control probes) is analyzed. The probes are evaluated based on an expression value or p-value threshold (methylation data) obtained from the background signal, for a specified percentage of samples selected by the user. Probes that do not pass this evaluation are considered unreliable and are filtered out. Quantile normalization was used for normalizing the expression values. eUTOPIA uses methods from the limma package (Marwah et al. 2019; Ritchie et al. 2015b).

The sva function from the sva R package was applied in eUTOPIA for the batch correction. Batch effects in microarray data may arise from various known variables such as RNA quality, experiment date, dye, and other hidden sources of variation that cannot be explained by these known variables (Jeffrey et al. 2022). The metadata contains information on known biological variables like treatment, disease status, age, tissue, and technical variables such as dye and array. By utilizing the sva function, unknown sources of variation can be detected in the data. The differential analysis in eUTOPIA is performed with linear model application in the R package limma. Specifically, the lmFit function is utilized to fit gene-wise linear models to the microarray data, with the design of the model including the comparison between IPF and healthy samples as the biological variable of interest including the covariates (biological and technical batch variables from sva) in the model. The contrasts of interest are then specified to obtain contrast specific coefficients from the original coefficients of the linear model. The eBayes function is applied to assess differential expression by using the fitted model with the contrast coefficients. Final reporting of the differentially expressed genes is performed by using the toptable function where adjusted p-value for the multiple comparisons was obtained with Benjamini & Hochberg -method (Marwah et al. 2019).

Dataset integration

To address batch effects in the dataset, integration was performed separately for each cell type and for both disease and healthy samples. The integration was also done separately for microarray and RNA-seq datasets. Given that the gene expression datasets were obtained from various sources, it is essential to perform batch adjustment in order to accurately infer network relationships. Batch effects can result in inaccurate correlations between genes and interfere with the identification of biologically meaningful relationships (Federico et al. 2020b). Therefore, batch adjustment is necessary to remove systematic variation due to technical factors, enabling a more accurate identification of true biological relationships. The process of batch adjustment was carried out using the "multi_studies_adjust" function, which is a part of a meta-analysis pipeline that underwent validation as a part of this study. In this function the dataset integration is performed with "pamr.batchadjust" function from "pamr" (Trevor et al. 2022) package. pamr.batchadjust performs a genewise one-way ANOVA adjustment for expression values. Assuming that sample j is part of batch b and B represents the entire set of samples in that batch, $x(i, j)$ represents the expression level of gene i in sample j. The pamr.batchadjust -function adjusts the expression level $x(i, j)$ by subtracting the mean expression level of gene i across all samples in batch b, denoted by $mean[x(i, j)]$ (Trevor et al. 2022; Tibshirani et al. 2011)

When the distribution of samples across batches and groups is balanced, removing the average batch effect (zero-centering) will remove most of the variation due to batch differences, while preserving the variation due to differences between the groups. This can increase statistical power in downstream analyses. However, in unbalanced designs where batch differences are partially influenced by group differences, batch correction may also reduce genuine group differences, leading to reduced statistical power. In cases where there are multiple groups and the distribution of samples across batches is very uneven, batch correction may even induce spurious group differences (Nygaard et al. 2016). The benefits of this method is that it is easy to compute and not requiring essentially additional computation after the initial cross-validation. Studies indicate that it is reasonably accurate in general (Tibshirani et al. 2011; Nygaard et al. 2016). The PCA-plots of the dataset integration are illustrated in Figures 6-9.

Meta-analysis

The meta-analysis was performed individually for each cell type, as well as for all datasets combined and for all the datasets except the biopsy datasets. The first input used to perform the meta-analysis was a "metadata-frame" consisting of the adjusted p-values for each gene in each dataset (gene intersection between the datasets). The second input of the function is a vector called "class" that includes the class labels for each sample. In the case of a two-class unpaired design, the class label can be either 0 (representing the control group) or 1 (representing the case group). For one-class data, the label for each sample should be 1. In this study, "one-class data" was utilized for the meta-analysis since the group difference is already reflected in the adjusted p-values. The third input is a vector called "origin" that contained the origin labels for each sample.

The meta-analysis gene ranking function "run_ensembl_metanalysis" utilizes three methods to rank genes. Firstly, it employs the esc::effect_sizes() function (Alessio et al. 2021) to calculate the effect sizes for each gene. This method takes into account the mean adjusted p-value, sample sizes (as a vector representing the number of genes in each sample), and gene names. The fun argument is set to "chisq" to compute effect sizes using the chi-squared method. Secondly, a p-value

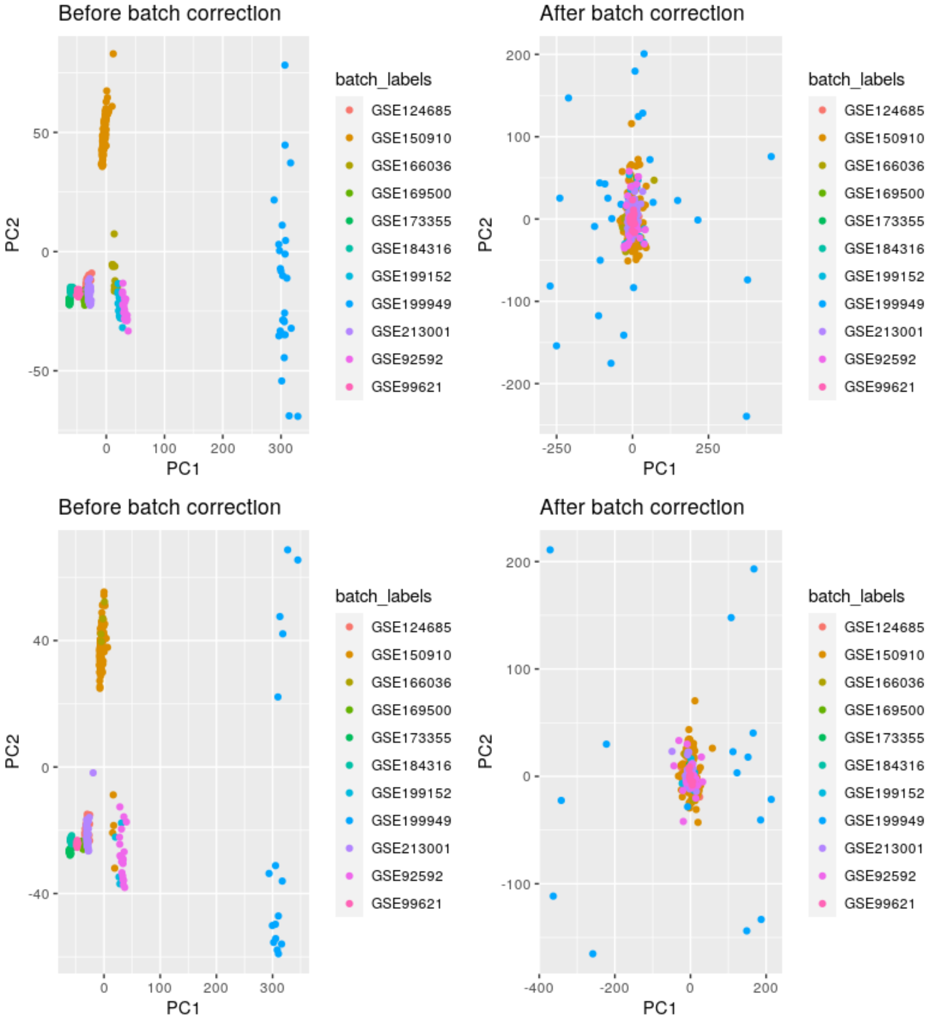

Figure 6: PCA plots of disease and healthy samples of RNA-seq biopsy samples before and after the batch adjustment (need to make again). Upper figure is the disease samples and lower figure the healthy samples.

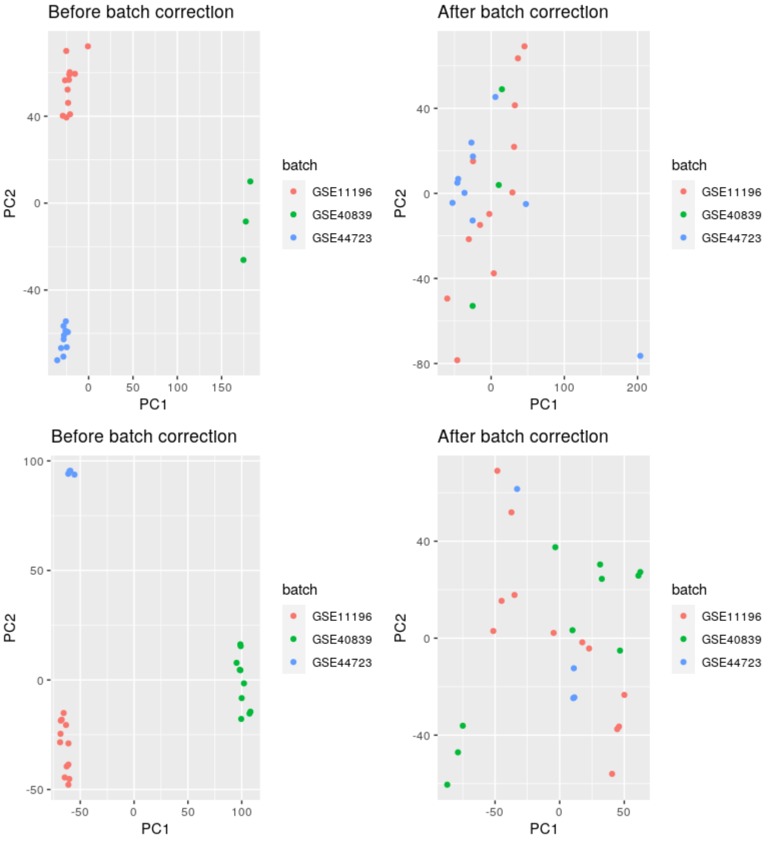

Figure 7: PCA plots of disease and healthy samples of fibroblast microarray samples before and after the batch adjustment. Upper figure is the disease samples and lower figure the healthy samples.

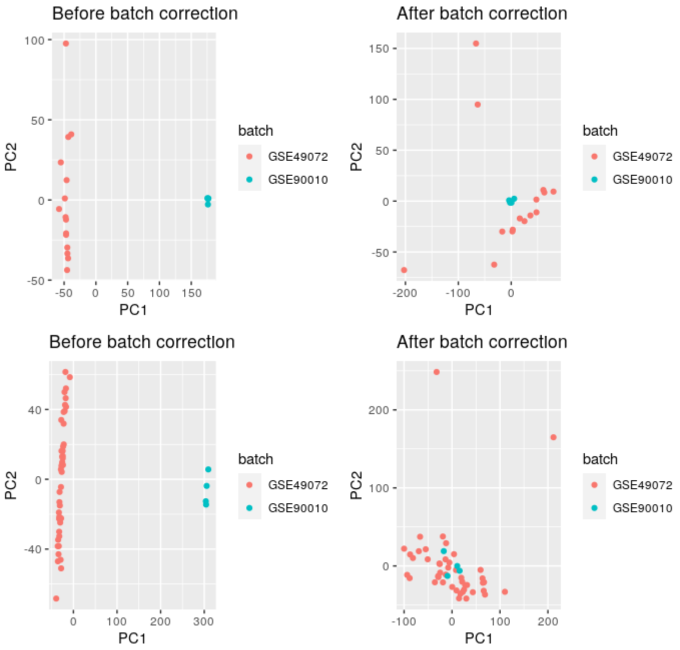

Figure 8: PCA plots of disease and healthy samples of macrophage microarray samples before and after the batch adjustment. Upper figure is the disease samples and lower figure the healthy samples.

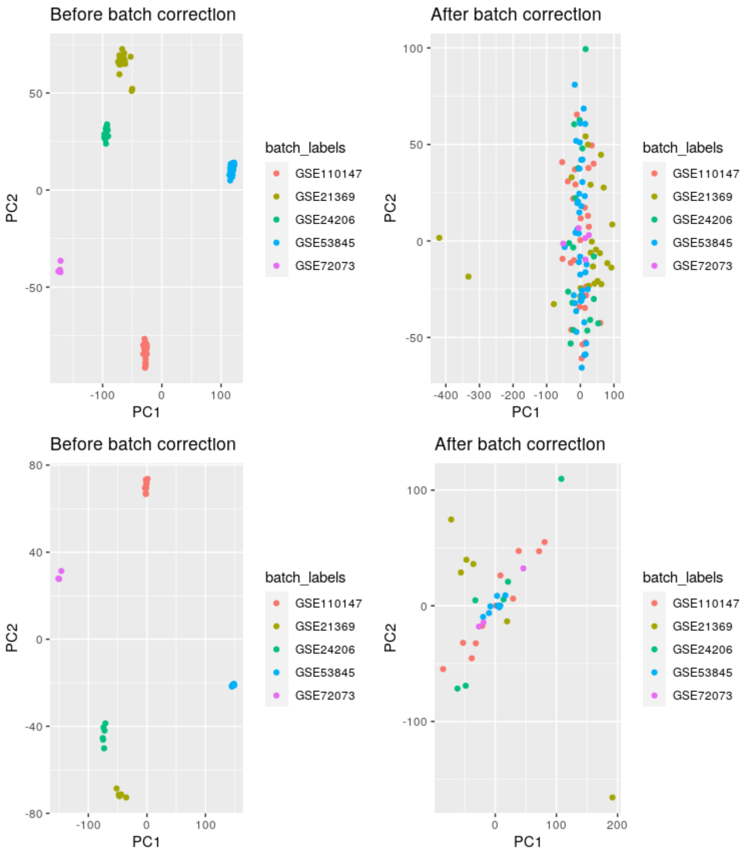

Figure 9: PCA plots of disease and healthy samples of biopsy microarray samples before and after the batch adjustment. Upper figure is the disease samples and lower figure the healthy samples.

based rank is calculated by aggregating the adjusted p-values for each gene using the sum of logs method. The metap::sumlog() function is used to aggregate the p-values for each gene (Mavridis and Higgins 2021). Thirdly, the function computes a rank product-based rank for the adjusted p-values. The RP.advance function (O'Quigley et al. 2020) is used to compute the rank product statistic and p-values for each gene across all samples, taking in the adjusted p-value data, class, and origin. In this study, class is set to 1 for all samples and origin is also set to 1 since the data is the adjusted p-values, not the gene expression matrix. Finally, the output of RP.advance is used to compute the Borda count ranks (De Borda 1781; Sokolova et al. 2017).

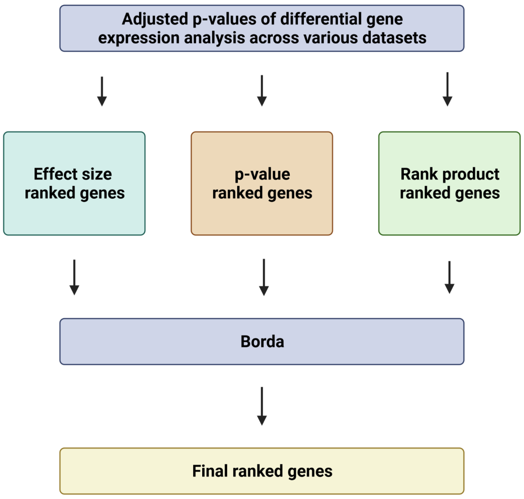
Adjusted p-values of differential gene expression analysis across various datasets
Effect size ranked genes
p-value ranked genes
Rank product ranked genes
Borda
Final ranked genes

Figure 10: Schematic illustration of the gene ranking pipeline. Figure created with BioRender.com

The meta-analysis gene ranking function utlilizes the TopKLists::Borda function to integrate the rankings and generates a final ranking. If the metric parameter is set to "median", the final ranking is computed based on the median Borda score. If it is set to "mean", the final ranking is computed based on the mean Borda score. Finally, the function returns the final ranked genes in a data frame format, either with the median or mean metric depending on the choice made. The meta-analysis gene ranking pipeline is illustrated in Figure 10.

Effect size measures the magnitude of differences, providing an estimate of the practical significance of the difference. Effect size is a standardized measure, making it easier to compare across different studies and data sets. (Sullivan and Feinn 2012; Cohen 1988). However, effect size can be influenced by sample size, so it can be sensitive to outliers or small number of replicates. Besides, effect size does not provide information about the statistical significance of the difference

(Cumming 2012; Lipsey and Wilson 1993). On the contrary, p-value provides a measure of statistical significance, indicating the probability of obtaining the observed results by chance alone. Sample size affects severely to the p-value, so it can be sensitive to outliers or small number of replicates. Furthermore, it is worth noting that the concept of p-value is becoming increasingly frustrating in the scientific community. Apart from being open to ambiguous interpretation, the p-value can be artificially reduced to any desired level by simply increasing the sample size (Demidenko 2016).

Rank product is a robust non-parametric method that can handle violations of assumptions about normality or equal variance. Moreover, rank product does not require a pre-determined threshold for statistical significance, making it appropriate for meta-analyses with diverse datasets. This method is based on biological reasoning related to the fold-change (FC) criterion. It identifies genes that consistently exhibit the strongest upregulation or downregulation across several replicate experiments. Additionally, it provides a solution to overcome the differences among multiple datasets, making it applicable for meta-analysis. Disadvantage of rank product ranking is that it assumes that the effects being measured are additive, so it may not be appropriate for data with interactions or non-linear relationships. Despite this limitation, the rank product method consistently displays high levels of reproducibility and specificity in both simulated and actual data applications, regardless of the degree of heterogeneity among datasets. This indicates that gene rankings generated through the rank product method are more resilient to noise and other underlying variables present in different datasets, in comparison to the effect size and p-value ranking methods (Breitling et al. 2004; Hong and Breitling 2008).

To integrate the different ranking methods, the Borda count method was utilized. The Borda count method is a non-parametric method, which means that it does not make assumptions about the underlying distribution of the data. This makes it a useful tool when working with data that does not conform to normal distribution. In addition, the Borda count method is considered robust as it is not significantly affected by outliers. This is particularly useful when working with large datasets where the presence of outliers can have a significant impact on the analysis. Overall, the Borda count method is a useful tool for ranking genes from multiple studies, particularly when working with large and complex datasets. It is a simple, non-parametric, and robust method that can be used with any type of data, making it a versatile tool (De Borda 1781; Qiu et al. 2016; Saari 1995; Sokolova et al. 2017).

Integrating rank information from the three methods mentioned offers several benefits. Using multiple methods can increase the statistical power of the analysis. By utilizing this approach, it is possible to discover additional genes that are genuinely associated to the phenotype of concern. Using multiple methods can help to ensure that the results are reproducible. If same genes are consistently highly ranked across all three methods, this increases confidence in their biological relevance. Overall, using multiple methods to rank genes can increase the robustness, comprehensiveness, statistical power, and reproducibility of the analysis (Perscheid et al. 2019; Richardson et al. 2016).

Gene set enrichment analysis

Gene set enrichment analysis (GSEA) was performed for the gene list that was obtained from the meta-analysis against a gene set provided through a GMT file. GSEA analysis was performed with the "fgsea" package in R, utilizing the "c2.cp.reactome.v2023.1.Hs.symbols.gmt" GMT file, which is part of the Molecular Signatures Database (MSigDB) (Liberzon et al. 2011). This particular GMT file contains 1654 gene sets, representing canonical pathways gene sets derived from the Reactome pathway database(Liberzon et al. 2011).

The function "compute_gsea" uses the gmtPathways function from the fgsea package to read the GMT file specified in the input argument and create a list of gene sets. The fgsea function is then used to perform the GSEA analysis by comparing the gene list from the meta-analysis against each gene set in the GMT file.The fgsea function takes several parameters including: pathways – a list of gene sets in GMT-file, stats – the ranked gene-list from the meta-analysis, minSize – the minimum size of a gene set to be considered in the analysis, and nperm – the number of permutations used in the calculation of the p-value. Additionally, the plotGseaTable function from the fgsea package is employed to visualize the GSEA results, where the topPathways argument specifies the number of top-scoring pathways to display in the plot. Finally, the function returns the fgsea results as a data-frame containing the enrichment p-value, Benjamini & Hochberg -adjusted p-value, enrichment score, normalized enrichment score, and the size of the pathway after removing genes that are not present in the gene list.

The function "compute_gsea_thresh" computes the GSEA threshold of the gene rank based on the enrichment score. It takes in three input parameters: geneList, which is a ranked gene list from the meta-analysis; fgsea_res, the GSEA result obtained from the "compute_gsea" function; and background, a whole gene set object derived from the fgsea::gmtPathways(gmt_file) function. First, the function sorts the fgsea_res object by p-value in ascending order and identifies the significant pathways using a p-value cutoff of 0.05. For each significant pathway, the function calculates the GSEA enrichment score using the calcGseaStat function from the fgsea package. Next, the function identifies the maximum enrichment score and corresponding genes in the geneList and stores the index of the corresponding genes in the max_vec vector. Finally, the function calculates the median of the maximum enrichment score indices from all the significant pathways.

Network inference

The network inference was performed on the batch-adjusted integrated expression matrices separately for both the disease and healthy samples, as well as for the microarray and RNA-seq samples. INfORM functions, offered by the research group FHAIVE, were used for the network inference (Marwah et al. 2018).The first step was to call the "get_ranked_consensus_matrix" function. This function computes a consensus matrix of gene expression data using the Context Likelihood of Relatedness method, with Pearson correlation as the estimator. The resulting consensus matrix represents the co-expression relationships between genes. The next step is to get a list of ranked edges for each method based on the calculated correlation matrix. Borda voting is then performed on the list of ranked edges for each method to create a consensus ranking. Finally, a consensus binary matrix is created by selecting the most significant ranked edges from the median rank of the Borda result (De Borda 1781).

The next step was to utilize the "parse_edge_rank_matrix" function, which takes the output of the "get_ranked_consensus_matrix" function as input and returns a list containing a binary inference matrix and a ranked edge list. The "default" edge selection strategy was employed, meaning that the function selects edges until it reaches a threshold of input genes. In other words, the function adds edges in descending order of rank until it has connected all the input genes with enough edges. After obtaining the parsed edge rank matrix, the "get_iGraph" function creates an igraph object from the input matrix. The function checks if the matrix is symmetric. If it is not, the function generates a symmetric matrix by taking the element-wise maximum of the original matrix and its transpose. Then, the function generates an undirected igraph object from the adjacency matrix, setting the weight to NULL. Following this, the function calculates several vertex-level attributes, including betweenness centrality, closeness centrality, degree, eigenvector centrality, and clustering coefficient, and assigns them to the vertices of the graph. Finally, the function assigns default visual attributes (color and highlight color) to the vertices and edges of the graph, and returns the resulting igraph object.

After obtaining the graphs, the "get_modules" function from the INfORM functions was utilized to determine the communities in the graph. This function takes an igraph object as input and a method argument specifying which community detection algorithm to use. The "walktrap" method was used (Pons and Latapy 2005). Firstly, the optimal number of steps for the walktrap algorithm was calculated. To achieve this, the modularity of the clustering was evaluated at different step sizes (ranging from 2 to 10) using the "cluster_walktrap" function from the igraph package. Then, the number of steps that resulted in the highest modularity value was selected. Finally, the walktrap algorithm was applied using the optimal step size, and the resulting community membership vector was returned.

The next step involved integrating the RNA-seq and microarray networks of disease and healthy networks for each cell type and biopsy. However, as the macrophage networks had only microarray datasets and the epithelial networks only one RNA-seq dataset, this step was not relevant for those networks. The Jaccard similarity index was then calculated for all the modules between the RNA-seq and microarray modules, and the 75th empirical quartile of the most similar modules were included in the upstream analysis. The genes from the original graphs were extracted, and it was most reasonable to take a union between the subgraphs since the intersection resulted in an extremely disconnected graph. This indicates large differences between the microarray and RNA-seq methods. Heatmaps illustrating the Jaccard similarity indexes of the microarray and RNA-seq modules are shown in Figures 11-16. The numbers of vetrices and edges for each network are represented in Tables 4-8.

$$Jaccard = \frac{length(intersect(A,B))}{length(union(A,B))}$$

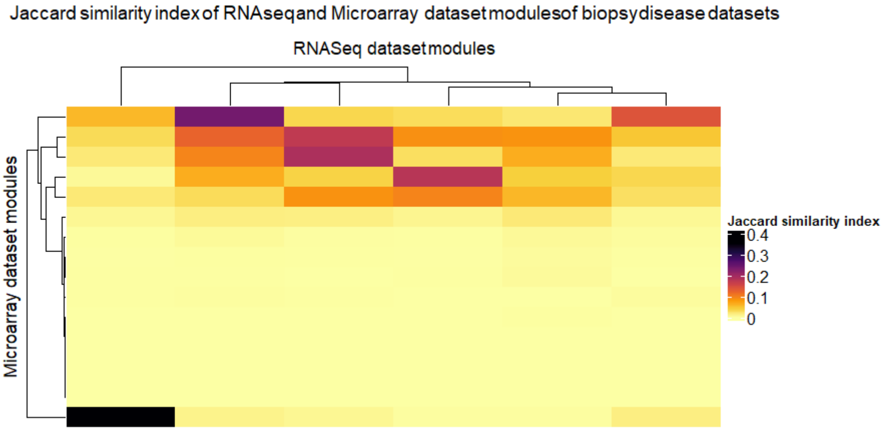

Figure 11: Heatmap of Jaccard similarity indexes in biopsy disease networks, with a 75 % quantile of 0.027 and a maximum value of 0.35. There are 9050 genes included.

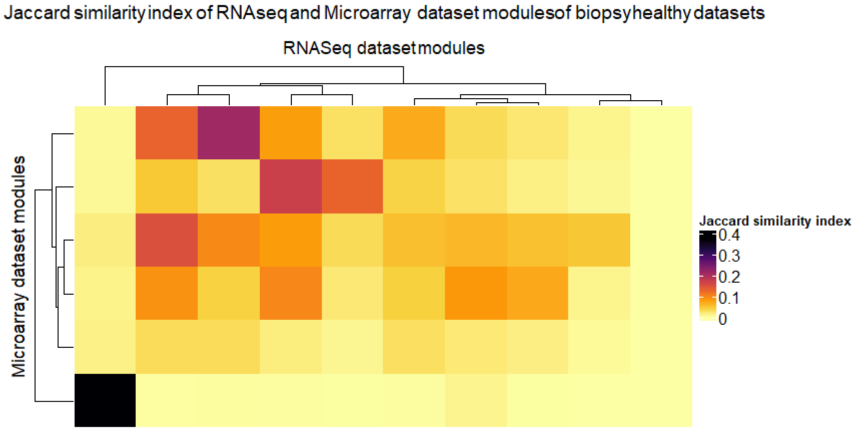

Figure 12: Heatmap of Jaccard similarity indexes in biopsy healthy networks with 75 % quantile of 0.055 and maximum value of 0.37. There are 6758 genes included.

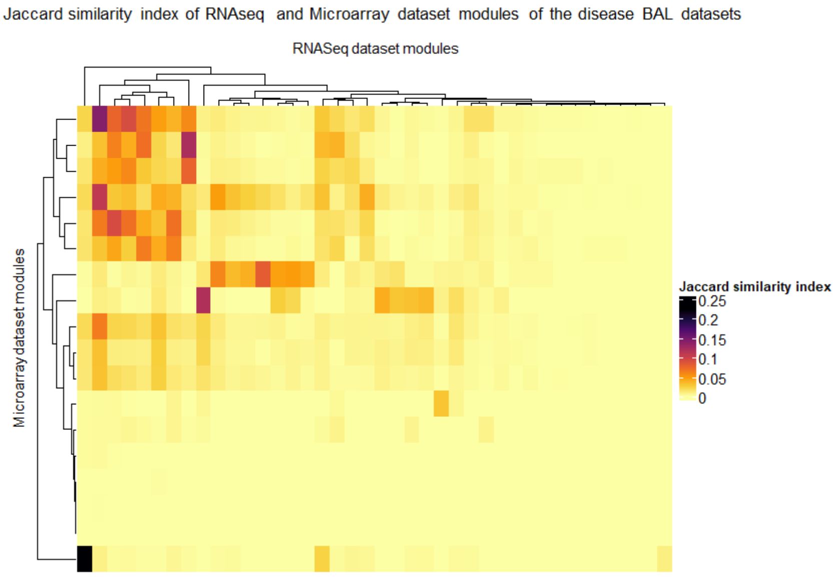

Figure 13: Heatmap of Jaccard similarity indexes in BAL disease networks with 75 % quantile of 0.0056 and maximum value of 0.22,. There are 12464 genes included.

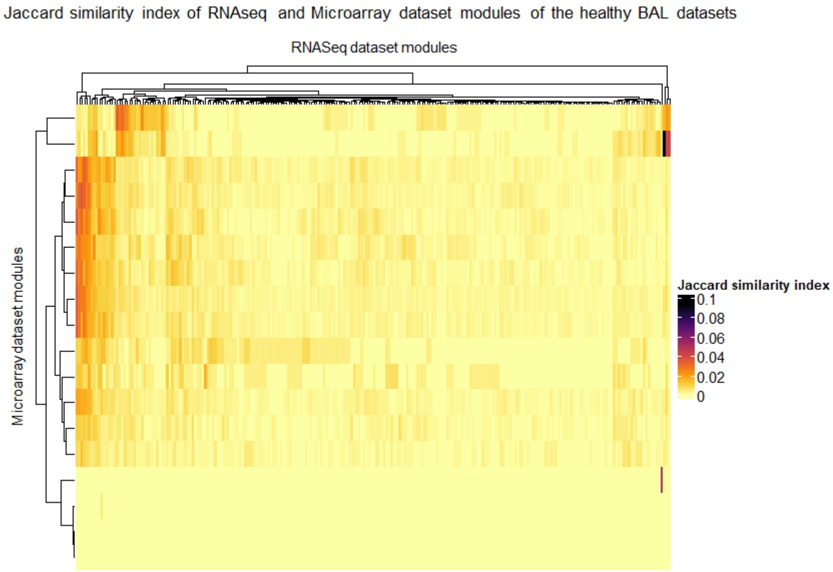

Figure 14: Heatmap of Jaccard similarity indexes in BAL healthy networks with 75 % quantile of 0.0026 and maximum value of 0.09. There are 10059 genes included.

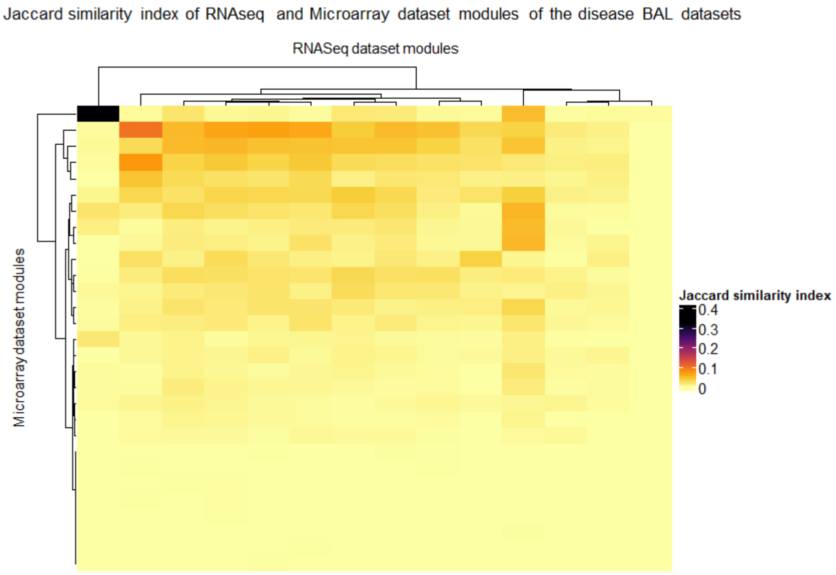

Figure 15: Heatmap of Jaccard similarity indexes in fibroblast disease networks with 75 % quantile of 0.014 and maximum value of 0.32. There are 7237 genes included.

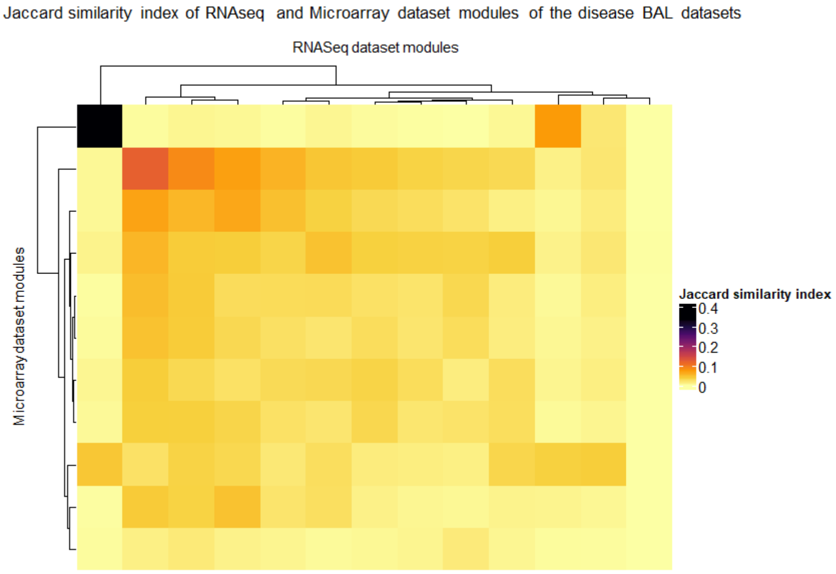

Figure 16: Heatmap of Jaccard similarity indexes in fibroblast healthy networks with 75 % quantile of 0.034 and maximum value of 0.34. There are 5647 genes included.

Table 4: Numbers of vertices and edges in biopsy networks. **D** represents the disease network **H** represents the healthy network. **micro** means microarray.

Biopsy D RNA-seq		Biopsy D micro		Biopsy H RNA-seq		Biopsy H micro	
Vertices	**Edges**	**Vertices**	**Edges**	**Vertices**	**Edges**	**Vertices**	**Edges**
10469	859270	15237	1281291	10469	1263699	15237	492850

Biopsy D union network		Biopsy H union network	
Vertices	**Edges**	**Vertices**	**Edges**
9050	1188046	6758	747684

Table 5: Numbers of vertices and edges in BAL networks. **D** represents the disease network **H** represents the healthy network. **micro** means microarray. The inconsistency between the number of vertices in BAL RNA-seq vertices between healthy and disease samples is because there were some vertices which Pearson correlation was zero in the adjusted expression matrix in BAL healthy RNA-seq samples.

BAL D RNA-seq		BAL D micro		BAL H RNA-seq		BAL H micro	
Vertices	**Edges**	**Vertices**	**Edges**	**Vertices**	**Edges**	**Vertices**	**Edges**
15236	195379	17322	622991	15106	133663	17322	453847

BAL D union network		BAL H union network	
Vertices	**Edges**	**Vertices**	**Edges**
12464	503480	10059	249450

Table 6: Numbers of vertices and edges in fibroblast networks. **D** represents the disease network **H** represents the healthy network. **micro** means microarray.

Fibroblast D RNA-seq		Fibroblast D micro		Fibroblast H RNA-seq		Fibroblast H micro	
Vertices	**Edges**	**Vertices**	**Edges**	**Vertices**	**Edges**	**Vertices**	**Edges**
12929	397608	11398	412629	12929	387267	11398	139737

BAL D union network		BAL H union network	
Vertices	**Edges**	**Vertices**	**Edges**
7237	324355	5647	137128

Table 7: Numbers of vertices and edges in macrophage networks. **D** represents the disease network **H** represents the healthy network. **micro** means microarray.

Macrophage D micro		Macrophage H micro	
Vertices	**Edges**	**Vertices**	**Edges**
11871	366291	11871	437531

Table 8: Numbers of vertices and edges in epithelial networks. **D** represents the disease network **H** represents the healthy network. **micro** means microarray.

Epithelial D RNA-seq		Epithelial H RNA-seq	
Vertices	**Edges**	**Vertices**	**Edges**
13527	121742	13527	194758

Network enrichment analysis, drug detection, and visualization methods

The Reactome pathway enrichment analysis was conducted on all modules of the union networks in biopsy, BAL, and fibroblast, as well as on the microarray macrophage and RNA-seq epithelial networks. To perform the network enrichment, the "get_reactome_from_modules" function was utilized. This function iterates over each module in the igraph object, extracts the gene names associated with that module, and converts them to Entrez IDs using the "bitr()" function from the clusterProfiler package. The ReactomePA::enrichPathway() function was then employed to perform a pathway enrichment analysis, which calculates the statistical significance of over-representation of each Reactome pathway among the set of genes in that module. A p-value cutoff of 0.05 was used to specify the threshold for statistical significance. If the pathway enrichment analysis identified any significantly enriched pathways (i.e., those with a p-value below the cutoff), the information was appended to a tab-separated text file that included the module name from which the enriched module came. The function "dotplot" function from the clusterProfiler package was used to generate a bubble plot visualization of the pathway enrichment results.

Next, the modules and their sizes from healthy and disease samples were retrieved. Modules containing less than ten genes were filtered out using the which() function. To determine the similarity between gene modules of healthy individuals and those with disease, a similarity matrix was constructed using the intersect() and union() functions to calculate the Jaccard similarity coefficient. The most dissimilar module between the two groups was then identified, and its genes were extracted. Next, the most significant pathway from the enrichment analysis of the most dissimilar module was retrieved, and the corresponding genes were extracted. The modules were sorted based on centrality, and a table was created with the sorted genes and their ranks.

Finally, the table containing the most enriched pathway genes from the most dissimilar module between healthy and disease samples was merged with the drug-target interaction data frame originating from open targets (Koscielny et al. 2017) parsed and offered by FHAIVE. The open targets data frame contains 2397 unique drugs but the dataframe has potential for further development. For example pirfenidone couldn't be found from the data frame. The reason for the absence of pirfenidone in the OpenTargets might be that the mechanism of action of pirfenidone has not been fully established (Grimminger et al. 2015). The drugs in the final output were sorted by the gene rank.

After extracting the top 50 ranked genes from the meta-analysis, the corresponding genes were retrieved from the open targets table that contained information on drugs, and sorted based on their gene rank. Also, the genes belonging to the most significant Reactome pathway, as indicated by the highest normalized enrichment score, were extracted, and the drugs targeting those genes were identified.

The visualization of the networks was carried out using the Gephi software version 0.10.1, where the Force Atlas algorithm was used for the layout. Only the modularity algorithm available in Gephi, i.e. Louvain, was used. For visualization purposes, a subset of the modules was selected, and the top 100 central genes from each module were used. In the epithelial networks, there were approximately 200 modules, and hence, only the top 50 genes from each module were used. Molecular structure visualizations were made with jmol-16.1.7 using data from RCBS database and PubChem. Illustrations were created with BioRender.com and Inkscape 1.1.

Results

Networks

Network inference of biopsy samples: disease modules, pathway enrichment, drugs and their targets

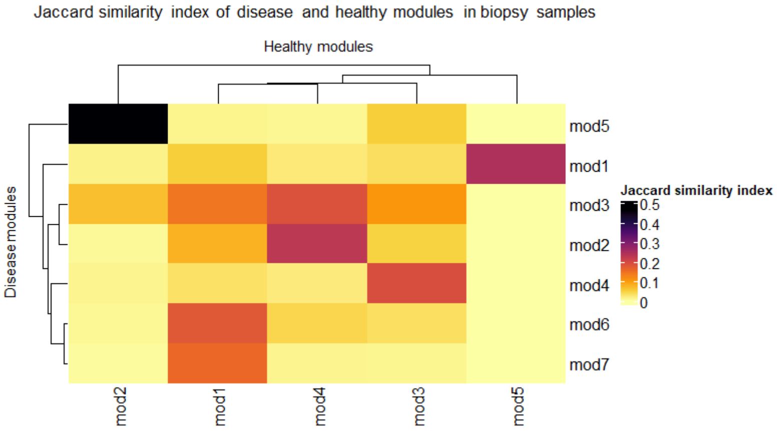

Figure 17: A heatmap indicating the Jaccard similarity indexes between the healthy and disease modules in biopsy samples. The disease modules are arranged in order from the least similar to the most similar, with the order being 7, 6, 3, 4, 2, 1, and 5.

Figure 17 shows the similarities between the healthy and disease modules in biopsy samples. The disease modules are ordered from the most dissimilar to the most similar, with the order being 7, 6, 3, 4, 2, 1, and 5. The Jaccard indexes for modules 1 through 7 are 0.25, 0.23, 0.19, 0.19, 0.46, 0.18, and 0.16, respectively. The sizes of the modules from 1 to 7 are 870, 1484, 3218, 778, 1212, 880, and 608 genes. The hub genes for modules 1 to 7 are *EHF, EFR3A, GNAI2, COTL1, FKBP8, FERMT2*, and *FOXC1*, respectively.

Figures 18 and 19 display the Reactome pathways enriched in each module. Figure 18 shows the pathways enriched in the disease network, while Figure 19 shows those enriched in the healthy network. In the disease network, the extracellular matrix organization is the most enriched pathway in module 7, which is also the most distinct module. This pathway was targeted for drugs, and Table 9 lists the drugs that were targeted for the genes involved. In Figure 20 are represented the visualizations of the biopsy disease network and the biopsy healthy network.

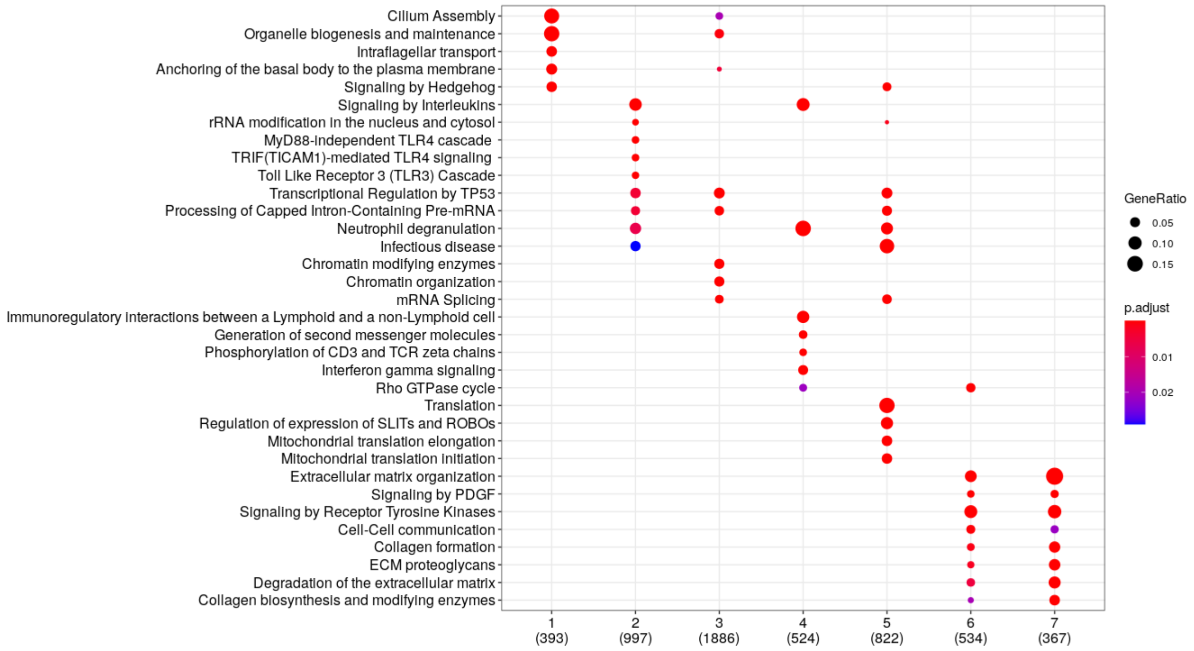

Figure 18: Enriched Reactome pathways in disease biopsy modules. The X-axis displays the module number, and the number in parentheses below indicates the number of genes that have been mapped to at least one Reactome pathway within that module.

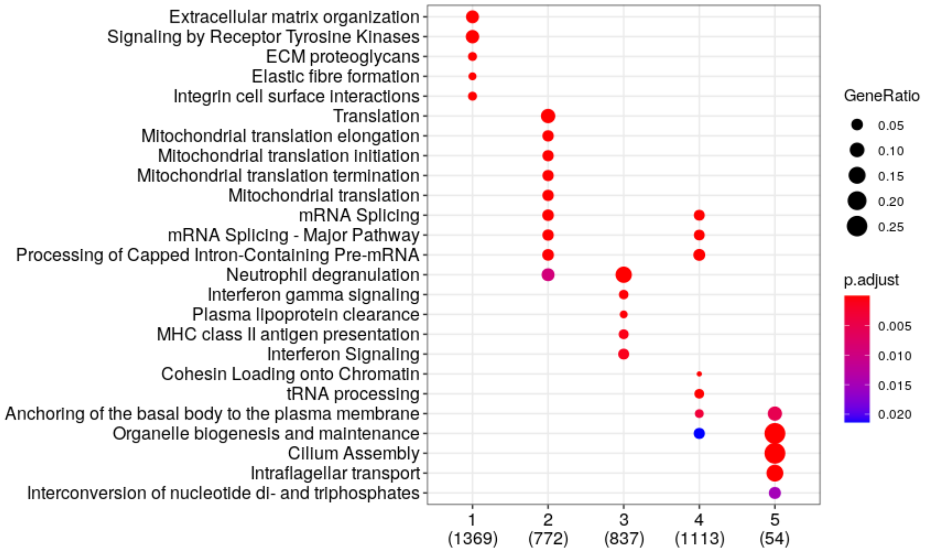

Figure 19: Enriched Reactome pathways in healthy biopsy modules. The X-axis displays the module number, and the number in parentheses below indicates the number of genes that have been mapped to at least one Reactome pathway within that module.

Table 9: Drugs targeting the genes in the biopsy disease network in the most dissimilar module between the biopsy disease network and the biopsy healthy network. The most enriched Reactome pathway in the disease module is the Extracellular Matrix Organization. Columns represent genes, drugs, the ranking of the gene in the module, target info, disease info, type of the molecule and the phase of the drug.

gene	drug	gene_module_rank	target_info	disease_info	molecule_type	drug_phase
DDR2	REGORAFENIB	3	discoidin domain receptor tyrosine kinase 2	neoplasm	Small molecule	Phase IV
COL6A2	COLLAGENASE CLOSTRIDIUM HISTOLYTICUM	6	collagen type VI alpha 2 chain	Dupuytren Contracture	Enzyme	Phase IV
COL6A2	OCRIPLASMIN	6	collagen type VI alpha 2 chain	macular holes	Enzyme	Phase IV
MMP2	MARIMASTAT	29	matrix metallopeptidase 2	lung carcinoma	Small molecule	Phase III
LOXL2	SIMTUZUMAB	43	lysyl oxidase like 2	primary myelofibrosis	Antibody	Phase II
ELN	VONAPANITASE	54	elastin	chronic kidney disease	Enzyme	Phase III
ITGB5	CILENGITIDE	216	integrin subunit beta 5	glioblastoma multiforme	Protein	Phase III
ITGB5	INTETUMUMAB	216	integrin subunit beta 5	metastatic colorectal cancer	Antibody	Phase II
ITGB5	ABITUZUMAB	216	integrin subunit beta 5	prostate adenocarcinoma	Antibody	Phase II
COL6A3	COLLAGENASE CLOSTRIDIUM HISTOLYTICUM	234	collagen type VI alpha 3 chain	Dupuytren Contracture	Enzyme	Phase IV
COL6A3	OCRIPLASMIN	234	collagen type VI alpha 3 chain	macular holes	Enzyme	Phase IV
TGFB3	FRESOLIMUMAB	285	transforming growth factor beta 3	malignant pleural mesothelioma	Antibody	Phase II
TNC	F16IL2	400	tenascin C	Merkel cell skin cancer	Antibody	Phase II
TNC	81C6 131I	400	tenascin C	neuroblastoma	Antibody	Phase II
TNC	F16SIP 131I	400	tenascin C	cancer	Antibody	Phase II
VWF	CAPLACIZUMAB	482	von Willebrand factor	autoimmune thrombocytopenic purpura	Antibody	Phase III

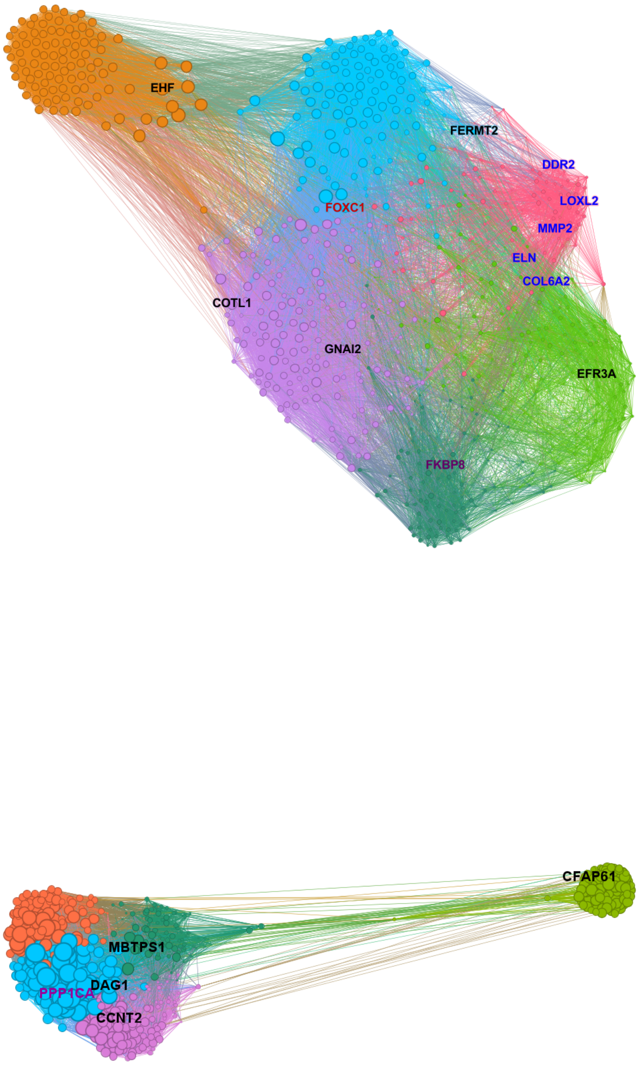

Figure 20: Graphs representing the biopsy samples' disease and healthy networks. The upper graph depicts the disease network, while the lower graph shows the healthy network. The blue labels in the disease network indicate the drug targets. Meanwhile, the black labels represent the central genes in the modules in both networks. The red label corresponds to the most central gene in the disease module, while the purple labels represent the most central genes in the modules that are most similar between the healthy and disease networks.

Network inference of BAL samples: disease modules, pathway enrichment, drugs and their targets

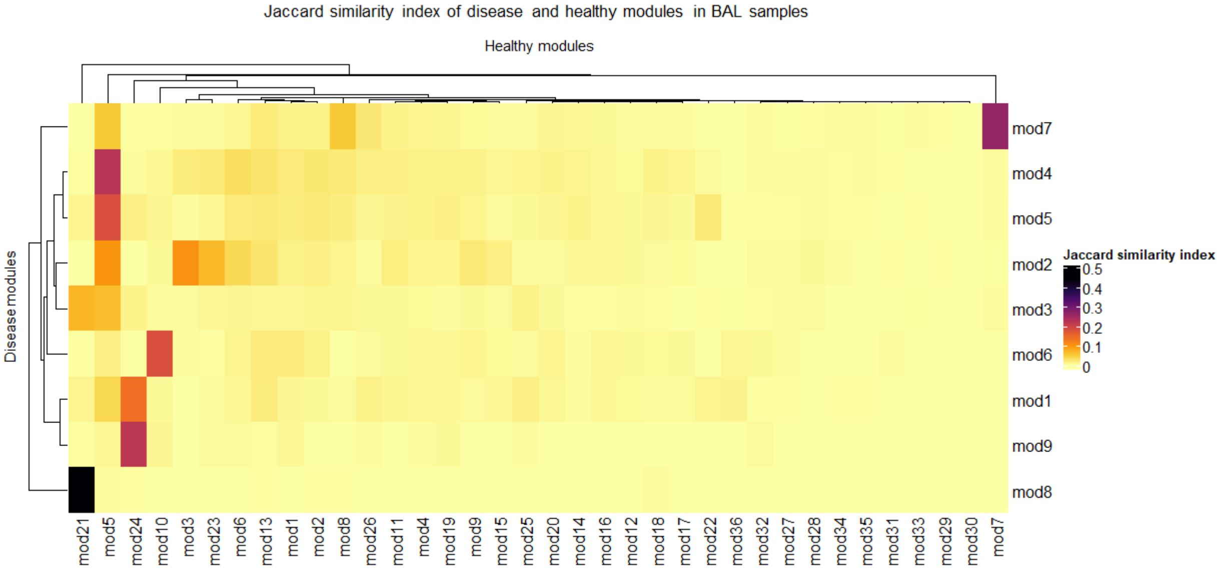

Figure 21: A heatmap indicating the Jaccard similarity indexes between the healthy and disease modules in BAL samples. The disease modules are arranged in order from the least similar to the most similar, with the order being 3, 2, 1, 6, 5, 9, 4, 7, and 8

Figure 21 shows the similarities between the healthy and disease modules in BAL samples. The disease modules are ordered from the most dissimilar to the most similar, with the order being 3, 2, 1, 6, 5, 9, 4, 7, and 8. The Jaccard indexes for modules 1 through 7 are 0.14, 0.11, 0.07, 0.23, 0.18, 0.18, 0.26, 0.44, and 0.22 , respectively. The sizes of the disease modules from 1 to 9 are 656 1815, 1117, 3995, 3110, 271, 1090, 240, and 170 genes. The hub genes for modules 1 to 9 are *EMP1, DLG1, CDCP1, DNAJC19, MSRA, C2CD4C, KXD1, ZAP70*, and *PRR15L*, respectively.

Figures 22 and 23 display the Reactome pathways enriched in each module. Figure 22 shows the pathways enriched in the BAL disease network, while Figure 23 shows those enriched in the healthy network. In the disease network, the neutrophil degranulation is the most enriched pathway in module 3, which is also the most distinct module. This pathway was targeted for drugs, and Table 10 lists the drugs that were targeted for the genes involved. In Figure 24 are represented the visualizations of the BAL disease network and the BAL healthy network.

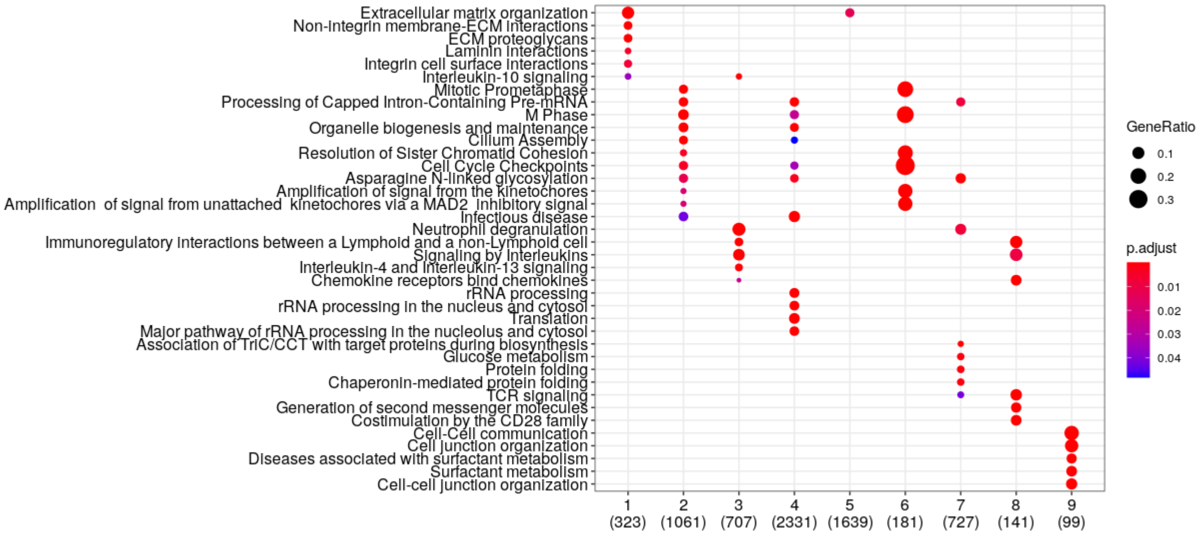

Figure 22: Enriched Reactome pathways in disease BAL modules. The X-axis displays the module number, and the number in parentheses below it indicates the number of genes that have been mapped to at least one Reactome pathway within that module.

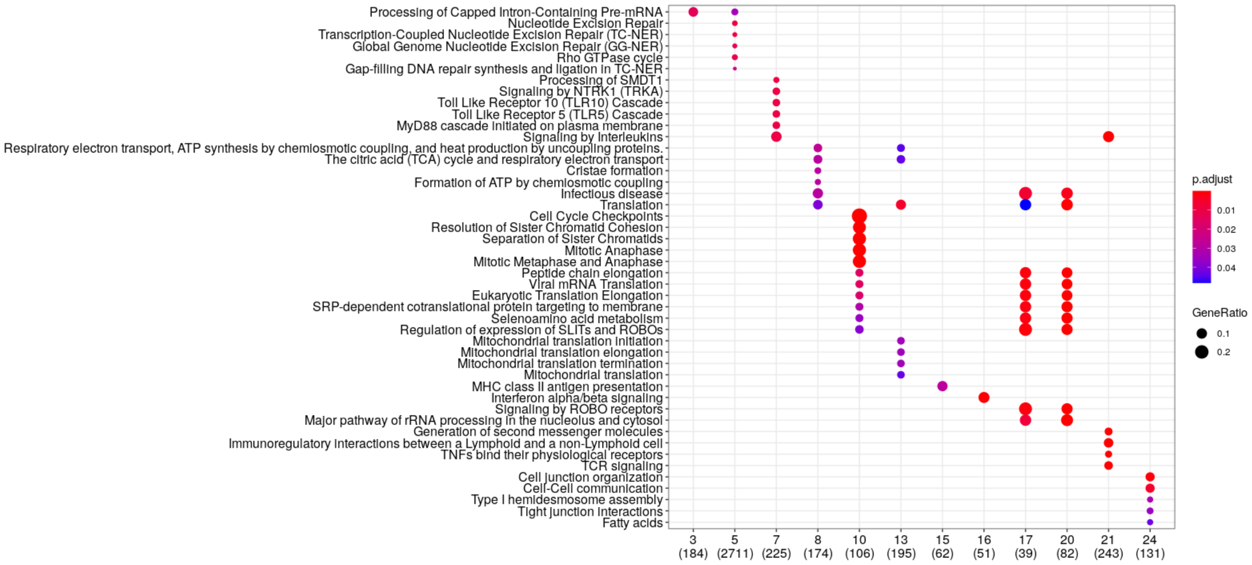

Figure 23: Enriched Reactome pathways in healthy BAL modules. The X-axis displays the module number, and the number in parentheses below it indicates the number of genes that have been mapped to at least one Reactome pathway within that module.

Table 10: Drugs targeting genes in the BAL disease network in the most dissimilar module between the BAL disease network and the BAL healthy network. The most enriched Reactome pathway in the disease module is the Neutrophil degranulation Pathway. Columns represent genes, drugs, the ranking of the gene in the module, target info, disease info, type of the molecule and the phase of the drug.

gene	drug	gene_module_rank	target_info	disease_info	molecule_type	drug_phase
MGAM	ACARBOSE	165	maltase-glucoamylase	type II diabetes mellitus	Small molecule	Phase IV
MGAM	VOGLIBOSE	165	maltase-glucoamylase	type II diabetes mellitus	Small molecule	Phase III
MGAM	MIGLITOL	165	maltase-glucoamylase	diabetes mellitus	Small molecule	Phase IV
CXCR1	REPARIXIN	169	C-X-C motif chemokine receptor 1	type I diabetes mellitus	Small molecule	Phase III
MMP9	MARIMASTAT	181	matrix metallopeptidase 9	lung carcinoma	Small molecule	Phase III
MMP9	ANDECALIXIMAB	181	matrix metallopeptidase 9	gastric adenocarcinoma	Antibody	Phase III
CD14	IC14	279	CD14 molecule	amyotrophic lateral sclerosis	Antibody	Phase II
CXCR2	DANIRIXIN	293	C-X-C motif chemokine receptor 2	chronic obstructive pulmonary disease	Small molecule	Phase II
CXCR2	ELUBRIXIN	293	C-X-C motif chemokine receptor 2	ulcerative colitis	Small molecule	Phase II
CXCR2	NAVARIXIN	293	C-X-C motif chemokine receptor 2	asthma	Small molecule	Phase II
CXCR2	REPARIXIN	293	C-X-C motif chemokine receptor 3	breast carcinoma	Small molecule	Phase III
CXCR2	LADARIXIN	293	C-X-C motif chemokine receptor 2	bullous pemphigoid	Small molecule	Phase II
SELL	RIVIPANSEL	303	selectin L	Sickle cell anemia	Small molecule	Phase III
SELL	BIMOSIAMOSE	303	selectin L	chronic obstructive pulmonary disease	Small molecule	Phase II
MME	SACUBITRIL	486	membrane metalloendopeptidase	heart failure	Small molecule	Phase IV
HPSE	MUPARFOSTAT	637	heparanase	hepatocellular carcinoma	Oligosaccharide	Phase III
HBB	EFAPROXIRAL	985	hemoglobin subunit beta	cancer	Small molecule	Phase III

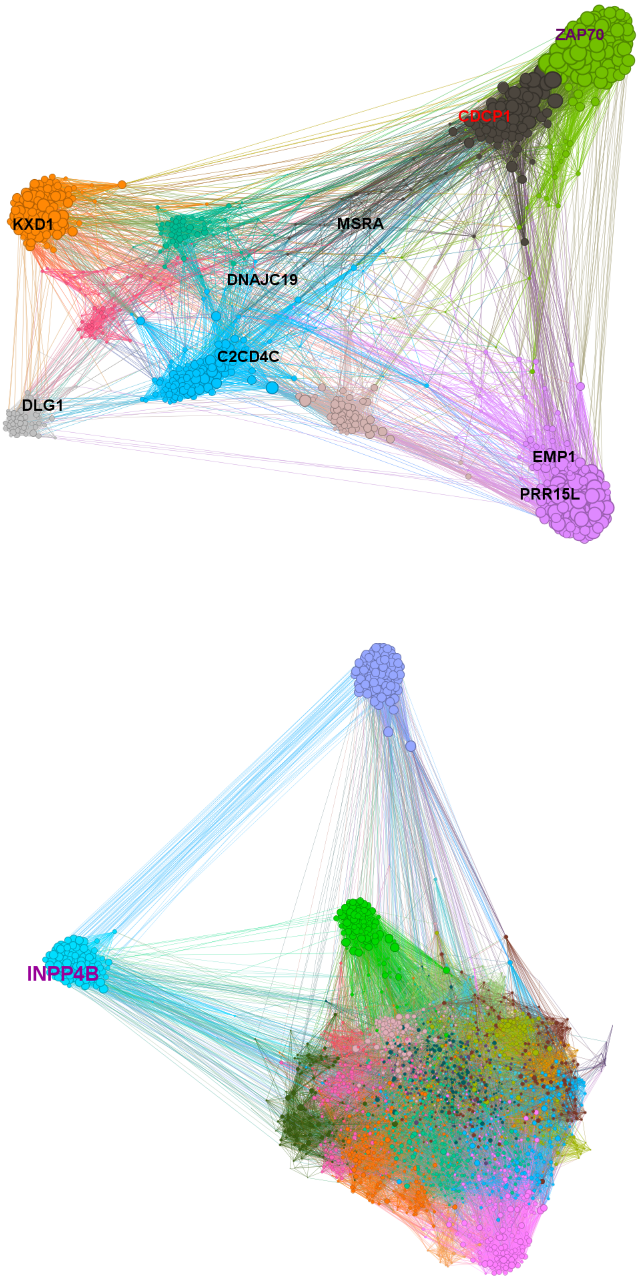

Figure 24: Graphs representing the BAL samples' disease and healthy networks. The upper graph depicts the disease network, while the lower graph shows the healthy network. There was no drug targets in the disease module ranked so high that they are visible in this illustration. The black labels represent the central genes in the modules in both networks. The red label corresponds to the most central gene in the disease module, while the purple labels represent the most central genes in the modules that are most similar between the healthy and disease networks.

Network inference of fibroblast samples: disease modules, pathway enrichment, drugs and their targets

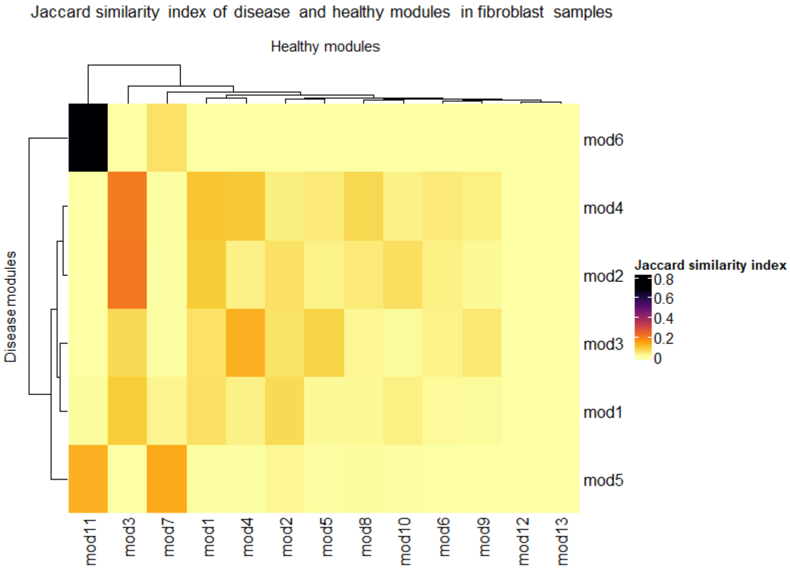

Figure 25: A heatmap indicating the Jaccard similarity indexes between the healthy and disease modules in fibroblast samples. The disease modules are arranged in order from the least similar to the most similar, with the order being 1, 3, 5, 4, 2, and 6.

Figure 25 shows the similarities between the healthy and disease modules in fibroblast samples. The disease modules are ordered from the most dissimilar to the most similar, with the order being 1, 3, 5, 4, 2, and 6. The Jaccard indexes for modules 1 through 6 are 0.08, 0.20, 0.12, 0.20, 0.13, and 0.68, respectively. The sizes of the disease modules from 1 to 6 are 876, 2800, 838, 2416, 152, and 155 genes. The hub genes for modules 1 to 6 are *EDNAJB9, KCNK1, FYN, FLOT1, ZW10,* and *NCAPH*, respectively.

Figures 26 and 27 display the Reactome pathways enriched in each module. Figure 26 shows the pathways enriched in the fibroblast disease network, while Figure 27 shows those enriched in the healthy network. In the disease network, the processing of capped intron-containing pre-mRNA is the most enriched pathway in module 1, which is also the one showing the highest modularity. Table 11 lists the drugs that target genes involved in this pathway. In addition, for fibroblasts, the drugs were targeted for module 3 pathways: extracellular matrix organization and signaling by tyrosine kinases. This decision was based on the fact that module 1 and module 3 are closely related, as can be seen from the dendrogram in Figure 25 and the network representations in Figure 28. The drugs that target the extracellular matrix organization are listed in Table 12, while those targeting signaling by tyrosine kinases are listed in Tables 13-15. In Figure 28 are represented the visualizations of the fibroblast disease network and the fibroblast healthy network.

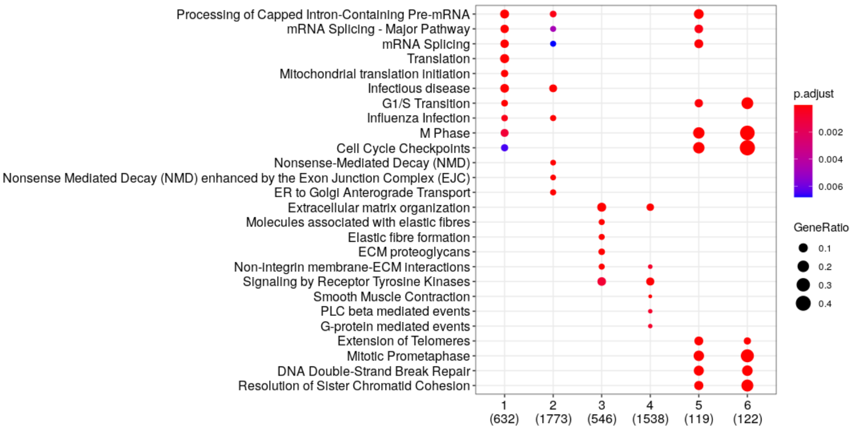

Figure 26: Enriched Reactome pathways in disease fibroblast modules. The X-axis displays the module number, and the number in parentheses below it indicates the number of genes that have been mapped to at least one Reactome pathway within that module.

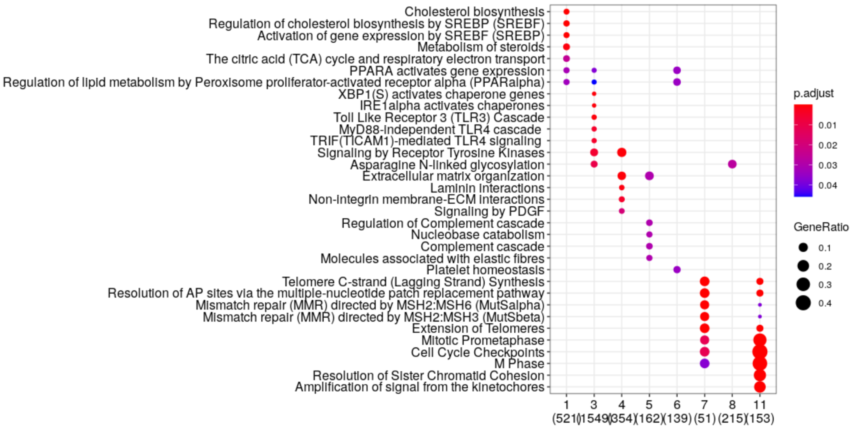

Figure 27: Enriched Reactome pathways in healthy fibroblast modules. The X-axis displays the module number, and the number in parentheses below it indicates the number of genes that have been mapped to at least one Reactome pathway within that module.

Table 11: Drugs targeting genes in the fibroblast disease network in the most dissimilar module between the fibroblast disease network and the fibroblast healthy network. The most enriched Reactome pathway in the disease module is the Processing of Capped Intron-Containing Pre-mRNA. Columns represent genes, drugs, the ranking of the gene in the module, target info, disease info, type of the molecule and the phase of the drug.

gene	drug	gene_module_rank	target_info	disease_info	molecule_type	drug_phase
HSPA8	FORIGERIMOD	648	heat shock protein family A (Hsp70) member 8	systemic lupus erythematosus	Protein	Phase III

Table 12: Drugs targeting genes in the fibroblast disease network in the second most dissimilar module between the fibroblast disease network and the fibroblast healthy network. The most enriched Reactome pathway in the disease module is the Extracellular matrix organization. Columns represent genes, drugs, the ranking of the gene in the module, target info, disease info, type of the molecule and the phase of the drug.

gene	drug	gene_module_rank	target_info	disease_info	molecule_type	drug_phase
COL4A4	COLLAGENASE CLOSTRIDIUM HISTOLYTICUM	190	collagen type IV alpha 4 chain	diabetic foot	Enzyme	Phase IV
COL4A4	OCRIPLASMIN	190	collagen type IV alpha 4 chain	macular holes	Enzyme	Phase IV
ITGAV	ABCIXIMAB	195	integrin subunit alpha V	acute coronary syndrome	Antibody	Phase IV
ITGAV	INTETUMUMAB	195	integrin subunit alpha V	prostate adenocarcinoma	Antibody	Phase II
ITGAV	ABITUZUMAB	195	integrin subunit alpha V	metastatic colorectal cancer	Antibody	Phase II
ITGAV	CILENGITIDE	195	integrin subunit alpha V	glioblastoma multiforme	Protein	Phase III
ITGAV	STX-100	195	integrin subunit alpha V	idiopathic pulmonary fibrosis	Antibody	Phase II
ITGAV	ETARACIZUMAB	195	integrin subunit alpha V	rheumatoid arthritis	Antibody	Phase II
COL15A1	COLLAGENASE CLOSTRIDIUM HISTOLYTICUM	259	collagen type XV alpha 1 chain	diabetic foot	Enzyme	Phase IV
COL15A1	OCRIPLASMIN	259	collagen type XV alpha 1 chain	macular holes	Enzyme	Phase IV
APP	BAPINEUZUMAB	663	amyloid beta precursor protein	Alzheimer's disease	Antibody	Phase III
APP	CRENEZUMAB	663	amyloid beta precursor protein	Alzheimer's disease	Antibody	Phase III
APP	SOLANEZUMAB	663	amyloid beta precursor protein	Alzheimer's disease	Antibody	Phase III
APP	GANTENERUMAB	663	amyloid beta precursor protein	Alzheimer's disease	Antibody	Phase III
APP	GSK933776	663	amyloid beta precursor protein	atrophic macular degeneration	Antibody	Phase II
APP	ADUCANUMAB	663	amyloid beta precursor protein	Alzheimer's disease	Antibody	Phase III
APP	PONEZUMAB	663	amyloid beta precursor protein	cerebral amyloid angiopathy	Antibody	Phase II
APP	BAN2401	663	amyloid beta precursor protein	Alzheimer's disease	Antibody	Phase II
TGFB3	FRESOLIMUMAB	794	transforming growth factor beta 3	malignant pleural mesothelioma	Antibody	Phase II

Table 13: Drugs targeting genes in the fibroblast disease network in the second most dissimilar module between the fibroblast disease network and the fibroblast healthy network. The second most enriched Reactome pathway in the disease module is the Signaling by receptor tyrosine kinases. Columns represent genes, drugs, the ranking of the gene in the module, target info, disease info, type of the molecule and the phase of the drug.

gene	drug	gene_module_rank	target_info	disease_info	molecule_type	drug_phase
FYN	JNJ-26483327	1	FYN proto-oncogene, Src family tyrosine kinase	cancer	Small molecule	Phase I
FYN	XL-228	1	FYN proto-oncogene, Src family tyrosine kinase	acute lymphoblastic leukemia	Small molecule	Phase I
FYN	DASATINIB	1	FYN proto-oncogene, Src family tyrosine kinase	chronic myelogenous leukemia	Small molecule	Phase IV
COL3A1	COLLAGENASE CLOSTRIDIUM HISTOLYTICUM	9	collagen type III alpha 1 chain	diabetic foot	Enzyme	Phase IV
COL3A1	OCRIPLASMIN	9	collagen type III alpha 1 chain	macular holes	Enzyme	Phase IV
HGF	RILOTUMUMAB	10	hepatocyte growth factor	gastric carcinoma	Antibody	Phase III
HGF	FICLATUZUMAB	10	hepatocyte growth factor	non-small cell lung carcinoma	Antibody	Phase II
FGFR1	ARQ-087	65	fibroblast growth factor receptor 1	intrahepatic cholangiocarcinoma	Small molecule	Phase II
FGFR1	SULFATINIB	65	fibroblast growth factor receptor 1	neoplasm	Small molecule	Phase III
FGFR1	RG-1530	65	fibroblast growth factor receptor 1	neoplasm	Small molecule	Phase I
FGFR1	AZD-4547	65	fibroblast growth factor receptor 1	squamous cell carcinoma	Small molecule	Phase II
FGFR1	ORANTINIB	65	fibroblast growth factor receptor 1	hepatocellular carcinoma	Small molecule	Phase III
FGFR1	NINTEDANIB	65	fibroblast growth factor receptor 1	idiopathic pulmonary fibrosis	Small molecule	Phase IV
FGFR1	LUCITANIB	65	fibroblast growth factor receptor 1	cancer	Small molecule	Phase II
COL4A4	COLLAGENASE CLOSTRIDIUM HISTOLYTICUM	190	collagen type IV alpha 4 chain	diabetic foot	Enzyme	Phase IV
COL4A4	OCRIPLASMIN	190	collagen type IV alpha 4 chain	macular holes	Enzyme	Phase IV
ITGAV	ABITUZUMAB	195	integrin subunit alpha V	metastatic colorectal cancer	Antibody	Phase II
ITGAV	CILENGITIDE	195	integrin subunit alpha V	glioblastoma multiforme	Protein	Phase III
ITGAV	INTETUMUMAB	195	integrin subunit alpha V	prostate adenocarcinoma	Antibody	Phase II
ITGAV	ETARACIZUMAB	195	integrin subunit alpha V	rheumatoid arthritis	Antibody	Phase II
ITGAV	STX-100	195	integrin subunit alpha V	idiopathic pulmonary fibrosis	Antibody	Phase II
ITGAV	ABCIXIMAB	195	integrin subunit alpha V	acute coronary syndrome	Antibody	Phase IV
PGF	AFLIBERCEPT	201	placental growth factor	diabetic macular edema	Protein	Phase IV
PGF	TB-403	201	placental growth factor	neoplasm	Antibody	Phase I
PGF	CONBERCEPT	201	placental growth factor	vitreous hemorrhage	Protein	Phase III

Table 14: Previous table continues. Drugs targeting genes in the fibroblast disease network in the second most dissimilar module between the fibroblast disease network and the fibroblast healthy network. The second most enriched Reactome pathway in the disease module is the Signaling by receptor tyrosine kinases. Columns represent genes, drugs, the ranking of the gene in the module, target info, disease info, type of the molecule and the phase of the drug.

gene	drug	gene_module_rank	target_info	disease_info	molecule_type	drug_phase
FGFR4	FGF401	223	fibroblast growth factor receptor 4	hepatocellular carcinoma	Small molecule	Phase II
FGFR4	LY-2874455	223	fibroblast growth factor receptor 4	adult acute myeloid leukemia	Small molecule	Phase I
FGFR4	BAY-1163877	223	fibroblast growth factor receptor 4	bladder transitional cell carcinoma	Small molecule	Phase II
FGFR4	BRIVANIB ALANINATE	223	fibroblast growth factor receptor 4	colorectal carcinoma	Small molecule	Phase III
FGFR4	BRIVANIB	223	fibroblast growth factor receptor 4	carcinoma	Small molecule	Phase III
FGFR4	ENMD-981693	223	fibroblast growth factor receptor 4	colorectal carcinoma	Small molecule	Phase II
IGF1	DUSIGITUMAB	300	insulin like growth factor 1	breast carcinoma	Antibody	Phase II
PDGFRA	TOVETUMAB	356	platelet derived growth factor receptor alpha	glioblastoma multiforme	Antibody	Phase II
PDGFRA	CEDIRANIB	356	platelet derived growth factor receptor alpha	neoplasm	Small molecule	Phase III
PDGFRA	MASITINIB	356	platelet derived growth factor receptor alpha	neoplasm	Small molecule	Phase III
PDGFRA	REGORAFENIB	356	platelet derived growth factor receptor alpha	neoplasm	Small molecule	Phase IV
PDGFRA	ILORASERTIB	356	platelet derived growth factor receptor alpha	cancer	Small molecule	Phase II
PDGFRA	PAZOPANIB	356	platelet derived growth factor receptor alpha	renal cell carcinoma	Small molecule	Phase IV
PDGFRA	CRENOLANIB	356	platelet derived growth factor receptor alpha	acute myeloid leukemia	Small molecule	Phase III
PDGFRA	MOTESANIB	356	platelet derived growth factor receptor alpha	non-small cell lung carcinoma	Small molecule	Phase III
PDGFRA	DOVITINIB	356	platelet derived growth factor receptor alpha	renal cell carcinoma	Small molecule	Phase III
PDGFRA	OLARATUMAB	356	platelet derived growth factor receptor alpha	neoplasm	Antibody	Phase IV
PDGFRA	MIDOSTAURIN	356	platelet derived growth factor receptor alpha	neoplasm	Small molecule	Phase IV
PDGFRA	XL-820	356	platelet derived growth factor receptor alpha	Gastrointestinal stromal tumor	Small molecule	Phase II
PDGFRA	FORETINIB	356	platelet derived growth factor receptor alpha	breast carcinoma	Small molecule	Phase II
PDGFRA	BECAPLERMIN	356	platelet derived growth factor receptor alpha	skin wound	Protein	Phase IV
PDGFRA	AMUVATINIB	356	platelet derived growth factor receptor alpha	small cell lung carcinoma	Small molecule	Phase II
PDGFRA	SUNITINIB	356	platelet derived growth factor receptor alpha	renal cell carcinoma	Small molecule	Phase IV
PDGFRA	VATALANIB	356	platelet derived growth factor receptor alpha	neoplasm	Small molecule	Phase III
PDGFRA	QUIZARTINIB	356	platelet derived growth factor receptor alpha	acute myeloid leukemia	Small molecule	Phase III
PDGFRA	SU-014813	356	platelet derived growth factor receptor alpha	breast neoplasm	Small molecule	Phase II
PDGFRA	XL-999	356	platelet derived growth factor receptor alpha	non-small cell lung carcinoma	Small molecule	Phase II
PDGFRA	TAK-593	356	platelet derived growth factor receptor alpha	neoplasm	Small molecule	Phase I
PDGFRA	LINIFANIB	356	platelet derived growth factor receptor alpha	non-small cell lung carcinoma	Small molecule	Phase III
PDGFRA	FAMITINIB	356	platelet derived growth factor receptor alpha	metastatic colorectal cancer	Small molecule	Phase III
PDGFRA	X-82	356	platelet derived growth factor receptor alpha	retinopathy	Small molecule	Phase II

Table 15: Previous table continues. Drugs targeting genes in the fibroblast disease network in the second most dissimilar module between the fibroblast disease network and the fibroblast healthy network. The second most enriched Reactome pathway in the disease module is the Signaling by receptor tyrosine kinases. Columns represent genes, drugs, the ranking of the gene in the module, target info, disease info, type of the molecule and the phase of the drug.

gene	drug	gene_module_rank	target_info	disease_info	molecule_type	drug_phase
MAPKAPK2	AT-13148	485	mitogen-activated protein kinase-activated protein kinase 2	neoplasm	Small molecule	Phase I
PTK2B	DEFACTINIB	546	protein tyrosine kinase 2 beta	non-small cell lung carcinoma	Small molecule	Phase II
INSR	INSULIN DETEMIR	558	insulin receptor	type I diabetes mellitus	Protein	Phase IV
INSR	INSULIN HUMAN	558	insulin receptor	type II diabetes mellitus	Protein	Phase IV
INSR	INSULIN PORK	558	insulin receptor	type II diabetes mellitus	Protein	Phase IV
INSR	INSULIN GLARGINE	558	insulin receptor	Hyperglycemia	Protein	Phase IV
INSR	INSULIN SUSP ISOPHANE BEEF	558	insulin receptor	type II diabetes mellitus	Protein	Phase IV
INSR	INSULIN SUSP ISOPHANE RECOMBINANT HUMAN	558	insulin receptor	type II diabetes mellitus	Protein	Phase IV
INSR	INSULIN GLULISINE	558	insulin receptor	type I diabetes mellitus	Protein	Phase IV
INSR	INSULIN ASPART	558	insulin receptor	type II diabetes mellitus	Protein	Phase IV
INSR	LINSITINIB	558	insulin receptor	adrenal cortex carcinoma	Small molecule	Phase III
INSR	INSULIN DEGLUDEC	558	insulin receptor	diabetes mellitus	Protein	Phase IV
INSR	INSULIN PURIFIED PORK	558	insulin receptor	diabetes mellitus	Protein	Phase IV
INSR	INSULIN SUSP ISOPHANE SEMISYNTHETIC PURIFIED HUMAN	558	insulin receptor	diabetes mellitus	Protein	Phase IV
INSR	INSULIN LISPRO	558	insulin receptor	type II diabetes mellitus	Protein	Phase IV
INSR	BMS-754807	558	insulin receptor	breast carcinoma	Small molecule	Phase II
INSR	INSULIN PURIFIED BEEF	558	insulin receptor	diabetes mellitus	Protein	Phase IV
INSR	INSULIN ASPART PROTAMINE RECOMBINANT	558	insulin receptor	type II diabetes mellitus	Protein	Phase IV
INSR	INSULIN ZINC SUSP RECOMBINANT HUMAN	558	insulin receptor	diabetes mellitus	Protein	Phase IV
INSR	INSULIN PEGLISPRO	558	insulin receptor	type II diabetes mellitus	Protein	Phase III
INSR	INSULIN LISPRO PROTAMINE RECOMBINANT	558	insulin receptor	type II diabetes mellitus	Protein	Phase IV
INSR	KW-2450	558	insulin receptor	breast carcinoma	Small molecule	Phase II
PDE3B	PENTOXIFYLLINE	750	phosphodiesterase 3B	obesity	Small molecule	Phase IV
PDE3B	ANAGRELIDE	750	phosphodiesterase 3B	essential thrombocythemia	Small molecule	Phase IV
PDE3B	DIPYRIDAMOLE	750	phosphodiesterase 3B	Recurrent thrombophlebitis	Small molecule	Phase IV
PDE3B	THEOPHYLLINE	750	phosphodiesterase 3B	kidney failure	Small molecule	Phase IV
PDE3B	LEVOSIMENDAN	750	phosphodiesterase 3B	heart failure	Small molecule	Phase III

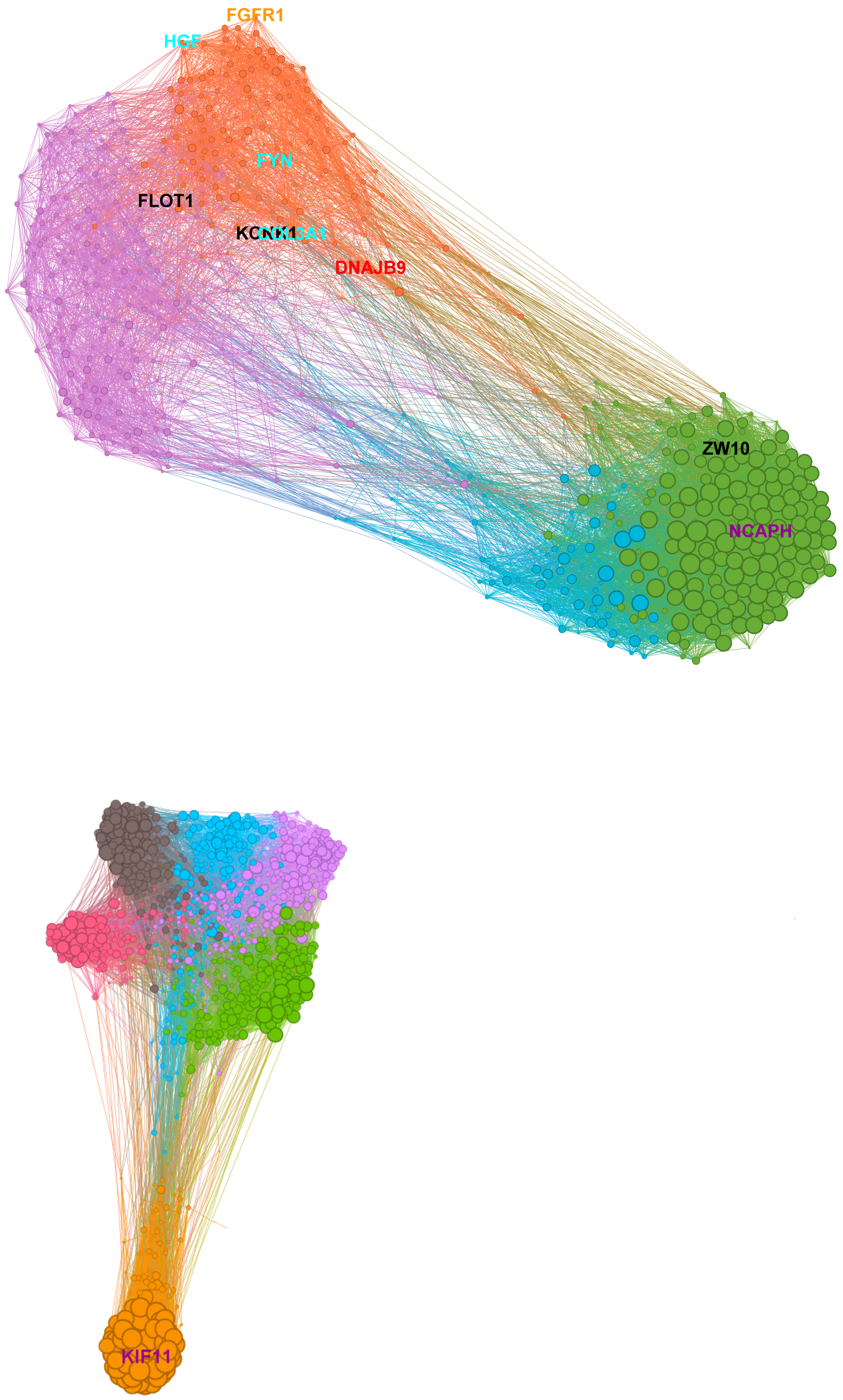

Figure 28: Graphs representing IPF fibroblasts disease network and the healthy counterpart. The upper graph depicts the disease network, while the lower graph shows the healthy network. The turqoise labels in the disease network indicate the drug targets (no drug targets visible in the most dissimilar module). Meanwhile, the black labels represent the central genes in the modules in both networks. The red label corresponds to the most central gene in the disease module, while the purple labels represent the most central genes in the modules that are most similar between the healthy and disease networks. The orange label is the target of nintedanib.

Network inference of macrophage samples: disease modules, pathway enrichment, drugs and their targets

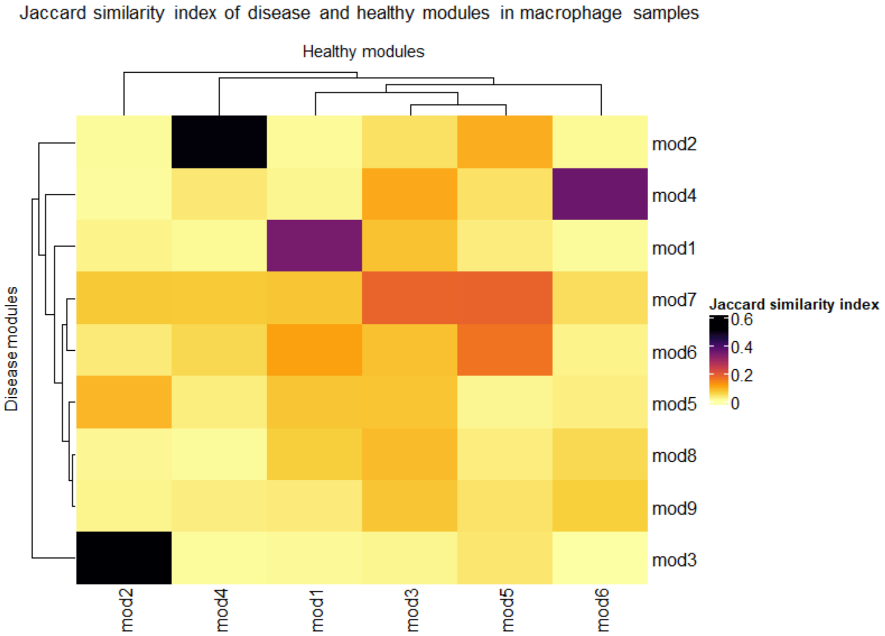

Figure 29: A heatmap indicating the Jaccard similarity indexes between the healthy and disease modules in macrophage datasets. Module size threshold is 10 genes. The disease modules are arranged in order from the least similar to the most similar, with the order being 9, 8, 5, 6, 7, 1, 4, 2, and 3.

Figure 29 depicts the degree of similarity between the healthy and disease modules in macrophage networks. The disease modules are arranged in descending order of dissimilarity: 12, 11, 10, 9, 8, 5, 6, 7, 1, 4, 2, and 3. Modules 12, 11, and 10 had only one gene, resulting in no enriched pathways and thus excluded from the analysis. Although module 9 did not contain any enriched Reactome pathways, module 8 showed significant similarity to module 9 (as shown in Figure 29) and had a Reactome pathway enriched in muscle contraction. The Jaccard indexes for modules 1 through 9 were 0.34, 0.51, 0.51, 0.36, 0.087, 0.16, 0.18, 0.083, and 0.070, respectively. The sizes of the disease modules from 1 to 12 are 1316, 2225, 1195, 1289, 756, 1425, 2495, 592, 575, 1, 1, and 1 genes, respectively. The hub genes for modules 1 to 12 are *RRP9 HLA-J, MTREX, PLGLA, PCDHB8, GPR31 PYHIN1, TCAP, PCDH17, BRAP, GABRA3,* and *LAMP1*, respectively.

Figures 30 and 31 display the Reactome pathways enriched in each module. Figure 30 shows the pathways enriched in the macrophage disease network, while Figure 31 shows those enriched in the healthy network. In the disease network, the Muscle contraction is the most enriched pathway in module 8, which is also the second most distinct module (no enriched Reactome pathways in module 9). This pathway was targeted for drugs, and Table 16 lists the drugs that were targeted for the genes involved. In Figure 32 are represented the visualizations of the macrophage disease network and the macrophage healthy network.

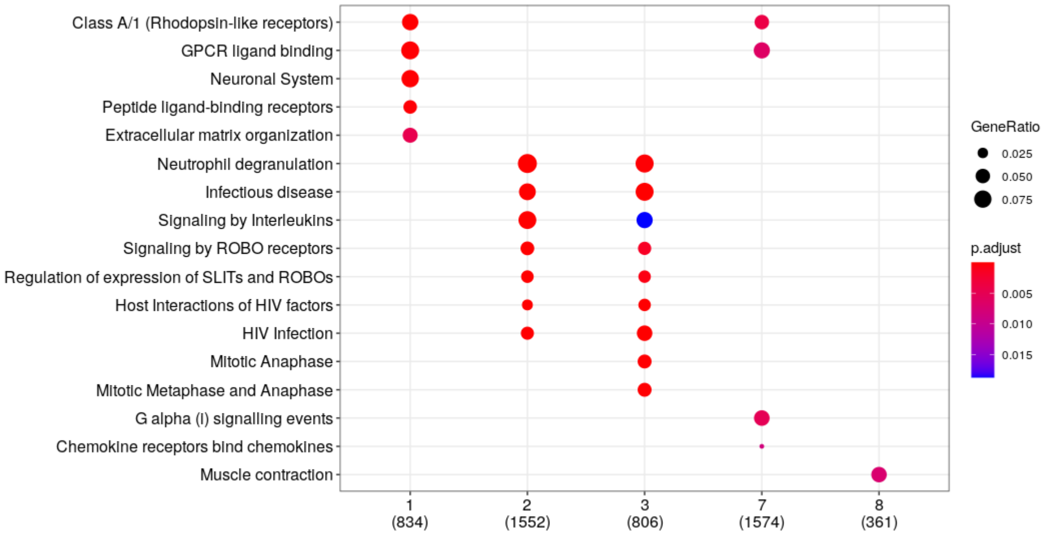

Figure 30: Enriched Reactome pathways in disease macrophage modules. The X-axis displays the module number, and the number in parentheses below it indicates the number of genes that have been mapped to at least one Reactome pathway within that module.

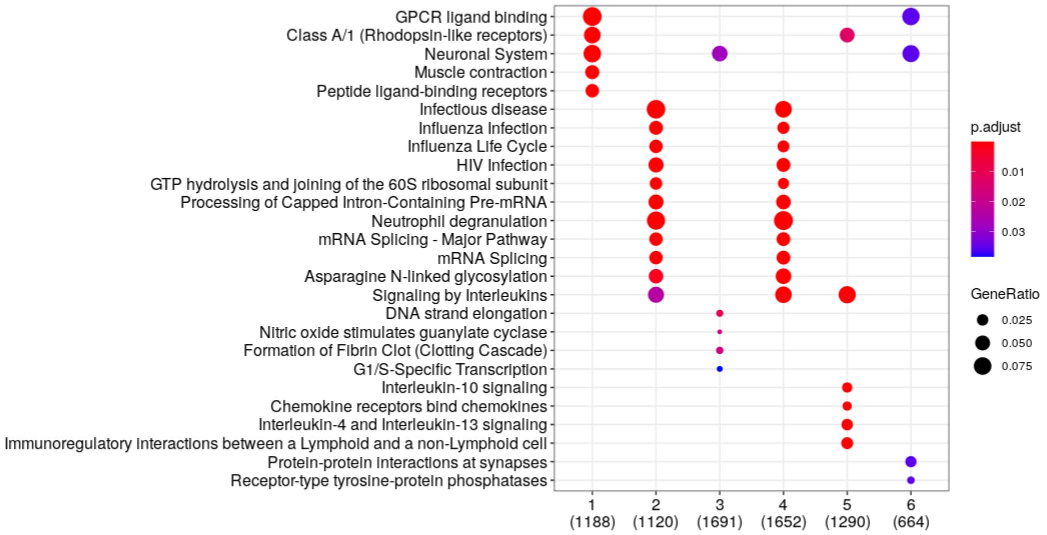

Figure 31: Enriched Reactome pathways in healthy macrophage modules. The X-axis displays the module number, and the number in parentheses below it indicates the number of genes that have been mapped to at least one Reactome pathway within that module.

Table 16: Drugs targeting genes in the macrophage disease network in the second most dissimilar module between the macrophage disease network and the macrophage healthy network. The most enriched Reactome pathway in the disease module is the Muscle contraction. Columns represent genes, drugs, the ranking of the gene in the module, target info, disease info, type of the molecule and the phase of the drug.

gene	drug	gene_module_rank	target_info	disease_info	molecule_type	drug_phase
GUCY1A2	ISOSORBIDE MONONITRATE	18	guanylate cyclase 1 soluble subunit alpha 2	coronary artery disease	Small molecule	Phase IV
GUCY1A2	SODIUM NITROPRUSSIDE	18	guanylate cyclase 1 soluble subunit alpha 2	myocardial infarction	Small molecule	Phase IV
GUCY1A2	NITROGLYCERIN	18	guanylate cyclase 1 soluble subunit alpha 2	coronary artery disease	Small molecule	Phase IV
GUCY1A2	RIOCIGUAT	18	guanylate cyclase 1 soluble subunit alpha 2	Idiopathic and/or familial pulmonary arterial hypertension	Small molecule	Phase IV
GUCY1A2	ISOSORBIDE DINITRATE	18	guanylate cyclase 1 soluble subunit alpha 2	cardiovascular disease	Small molecule	Phase IV
GUCY1A2	NITRIC OXIDE	18	guanylate cyclase 1 soluble subunit alpha 2	asthma	Small molecule	Phase IV
CACNG3	GABAPENTIN ENACARBIL	51	calcium voltage-gated channel auxiliary subunit gamma 3	Reunion Island's Larsen syndrome	Small molecule	Phase IV
CACNG3	ATAGABALIN	51	calcium voltage-gated channel auxiliary subunit gamma 3	insomnia	Small molecule	Phase II
TNNC2	TIRASEMTIV	139	troponin C2, fast skeletal type	amyotrophic lateral sclerosis	Small molecule	Phase III
CACNG4	SULOCTIDIL	262	calcium voltage-gated channel auxiliary subunit gamma 4	cardiovascular disease	Small molecule	Phase IV
CACNG4	BEPRIDIL	262	calcium voltage-gated channel auxiliary subunit gamma 4	cardiovascular disease	Small molecule	Phase IV
CACNG4	IMAGABALIN	262	calcium voltage-gated channel auxiliary subunit gamma 4	generalized anxiety disorder	Small molecule	Phase III
CACNA2D1	PREGABALIN	264	calcium voltage-gated channel auxiliary subunit alpha2delta 1	Seizures	Small molecule	Phase IV
CACNA2D1	GABAPENTIN	264	calcium voltage-gated channel auxiliary subunit alpha2delta 1	epilepsy	Small molecule	Phase IV
CACNA2D1	MIROGABALIN	264	calcium voltage-gated channel auxiliary subunit alpha2delta 1	fibromyalgia	Small molecule	Phase III
ATP1B2	DESLANOSIDE	377	ATPase Na+/K+ transporting subunit beta 2	cardiovascular disease	Small molecule	Phase IV
ATP1B2	DIGITOXIN	377	ATPase Na+/K+ transporting subunit beta 2	cardiovascular disease	Small molecule	Phase IV
ATP1B2	ACETYLDIGITOXIN	377	ATPase Na+/K+ transporting subunit beta 2	cardiovascular disease	Small molecule	Phase IV
ATP1B2	DIGOXIN	377	ATPase Na+/K+ transporting subunit beta 2	atrial fibrillation	Small molecule	Phase IV
MYL2	OMECAMTIV MECARBIL	495	myosin light chain 2	heart failure	Small molecule	Phase III

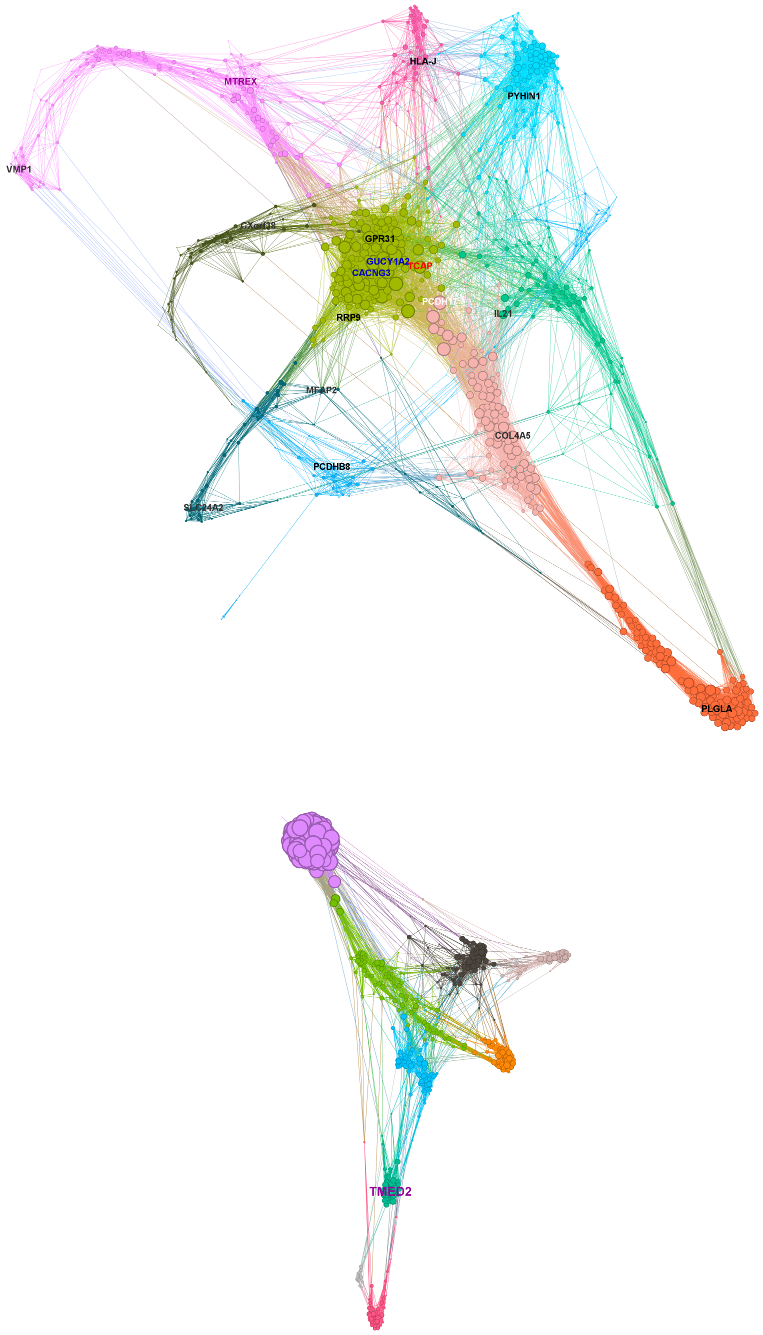

Figure 32: Graphs representing the macrophage disease and healthy networks. The upper graph depicts the disease network, while the lower graph shows the healthy network. The blue labels in the disease network indicate the drug targets. Meanwhile, the black labels represent the central genes in the modules in both networks. The red label corresponds to the most central gene in the disease module, while the purple labels represent the most central genes in the modules that are most similar between the healthy and disease networks. The white label is the most central gene in disease network in the most dissimilar module between the disease network and healthy network.

Network inference of epithelial samples: disease modules, pathway enrichment, drugs and their targets

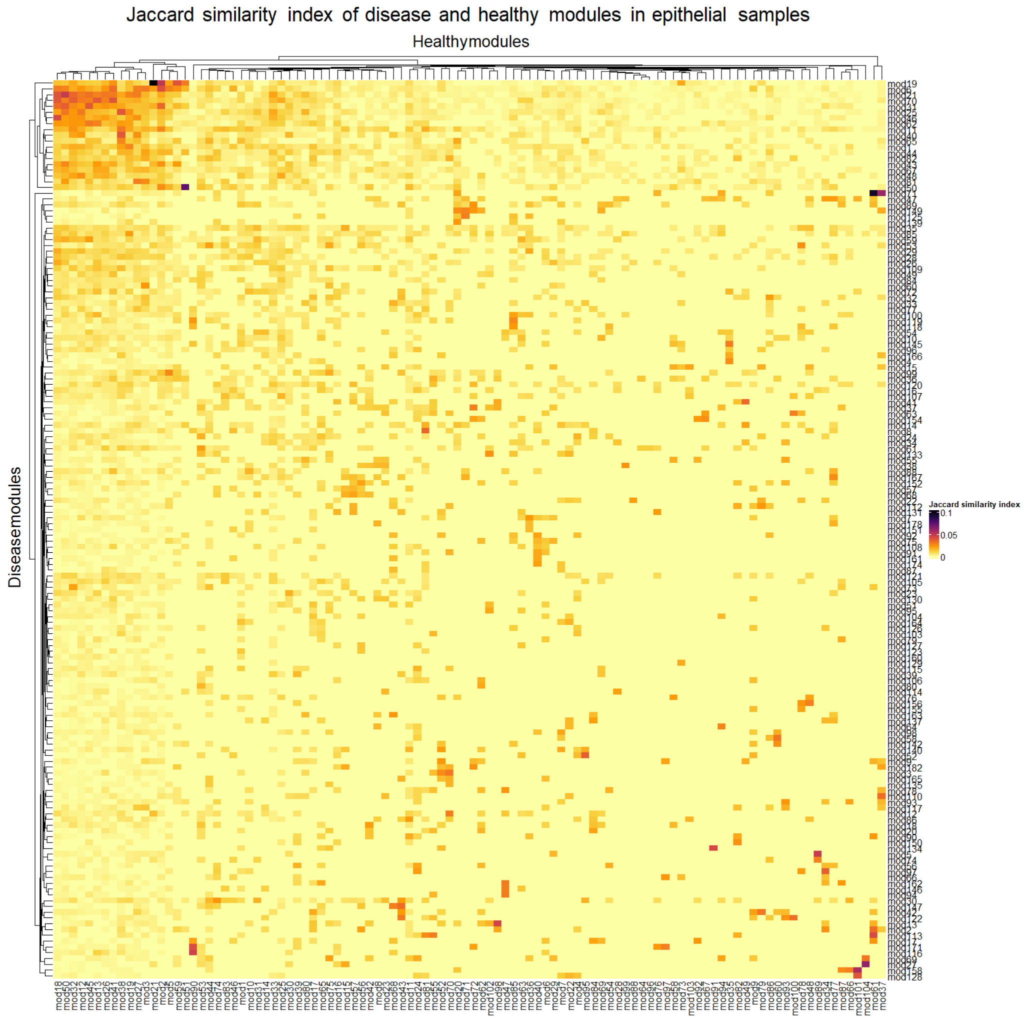

Figure 33: A heatmap indicating the Jaccard similarity indexes between the healthy and disease modules in the epithelial dataset. Module size threshold is 10 genes.

Figure 33 displays the degree of similarity between healthy and disease modules in epithelial networks. The figure indicates that there are significantly more modules in the epithelial networks in this study compared to other cell types. The epithelial disease network consists of 183 modules, while the healthy epithelial network comprises 104 modules. Among these, 155 disease epithelial modules and 98 healthy epithelial modules have more than 10 genes. The module that stands out as the most dissimilar in the disease network compared to the healthy modules is module 160, which contains 19 genes. The genes sorted by centrality in this pathway are *KRT222, ZNF121, CTDSP2, SYNJ2BP, ABHD3, RIOK2, STAT1, TM4SF1, CTSC, ENPP4, ZNF320, PQLC2L, OAS2, ZNF790, FRK, HIST1H2BC, HIST2H2BE, TLCD2,* and *ZNF235*. Figure 35 shows that the enriched reactome pathways in this module are related to neutrophil degranulation and cellular senescence. Although the network is scattered, the neutrophil degranulation pathway in module 160 has one gene (*FRK*) that has drugs targeting it, which can be seen in Table 17. The gene is Fyn Related Src Family Tyrosine Kinase.

The disease module pathways in epithelial networks are presented in Figures 35 and 36, while the healthy module pathways are displayed in figures 37 and 38. Additionally, the epithelial pathway graphs are visible in Figure 42. The top 10 most dissimilar modules in the disease network compared to the healthy network are 160, 84, 121, 16, 105, 123, 126, 77, 130, and 127.

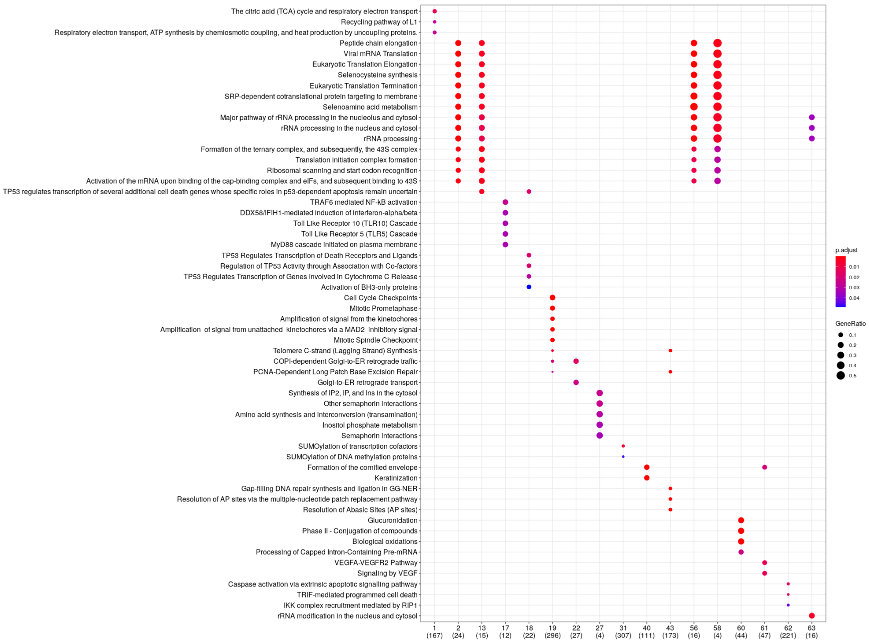

Figure 34: Enriched Reactome pathways in disease epithelial modules (1). The X-axis displays the module number, and the number in parentheses below it indicates the number of genes that have been mapped to at least one Reactome pathway within that module.

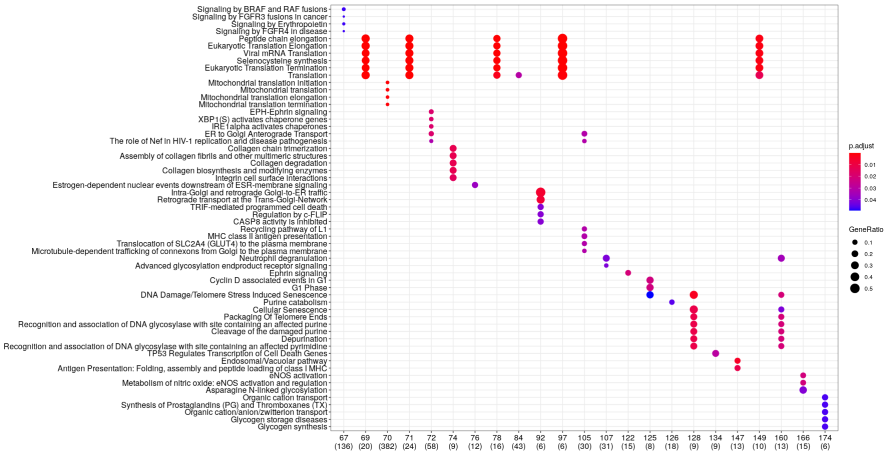

Figure 35: Enriched Reactome pathways in disease epithelial modules (2). The X-axis displays the module number, and the number in parentheses below it indicates the number of genes that have been mapped to at least one Reactome pathway within that module.

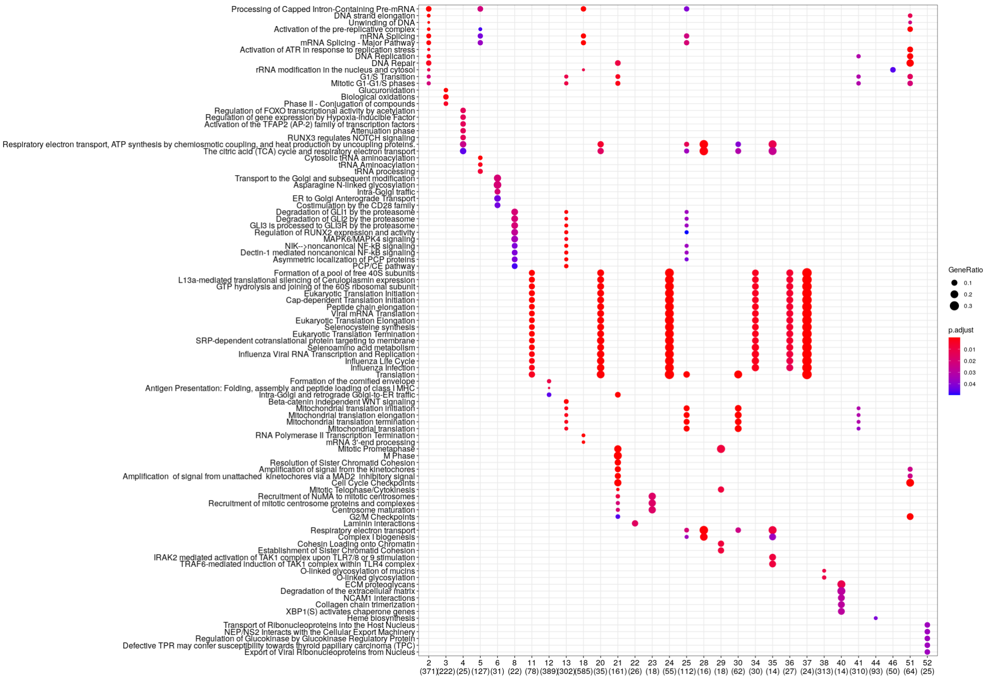

Figure 36: Enriched Reactome pathways in healthy epithelial modules (1). The X-axis displays the module number, and the number in parentheses below it indicates the number of genes that have been mapped to at least one Reactome pathway within that module.

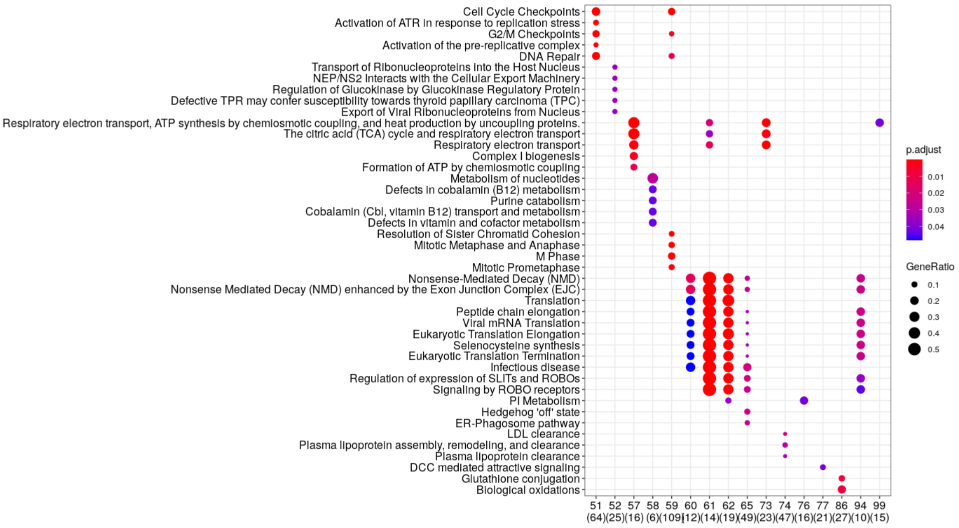

Figure 37: Enriched Reactome pathways in healthy epithelial modules (2). The X-axis displays the module number, and the number in parentheses below it indicates the number of genes that have been mapped to at least one Reactome pathway within that module.

Table 17: Drugs targeting genes in the epithelial disease network in the most dissimilar module between the epithelial disease network and the epithelial healthy network. The most enriched Reactome pathway in the disease module is the Neutrophil degranulation pathway. Columns represent genes, drugs, the ranking of the gene in the module, target info, disease info, type of the molecule and the phase of the drug.

gene	drug	gene_module_rank	target_info	disease_info	molecule_type	drug_phase
FRK	REGORAFENIB	15	fyn related Src family tyrosine kinase	neoplasm	Small molecule	Phase IV
FRK	DASATINIB	15	fyn related Src family tyrosine kinase	acute lymphoblastic leukemia	Small molecule	Phase IV
FRK	ENMD-981693	15	fyn related Src family tyrosine kinase	pancreatic carcinoma	Small molecule	Phase II
FRK	ILORASERTIB	15	fyn related Src family tyrosine kinase	cancer	Small molecule	Phase II
FRK	XL-228	15	fyn related Src family tyrosine kinase	chronic myelogenous leukemia	Small molecule	Phase I

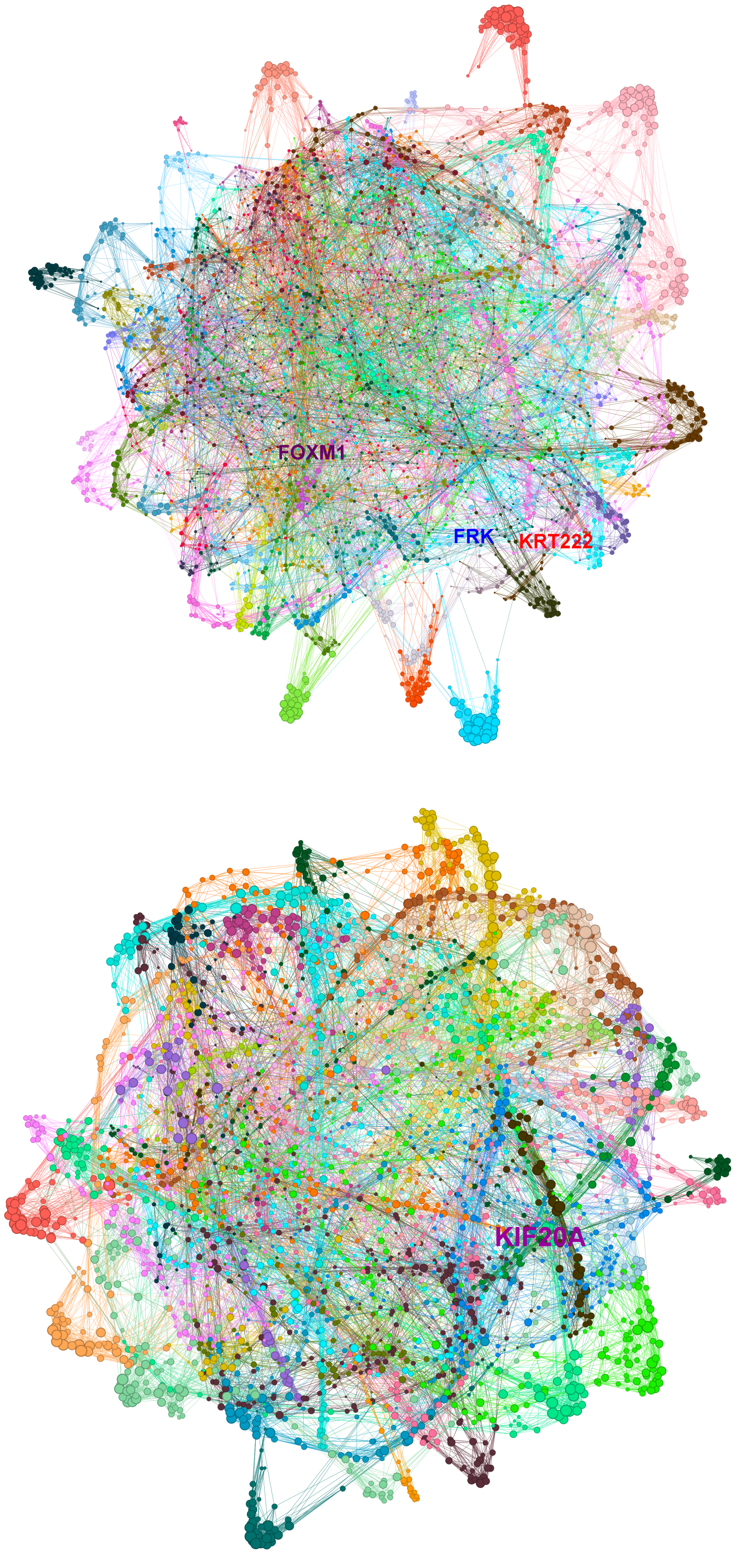

Figure 38: Graphs representing the epithelial IPF and healthy networks. The upper graph depicts the disease network, while the lower graph shows the healthy network. The blue label in the disease network indicates the drug target. . The red label corresponds to the most central gene in the disease module, while the purple labels represent the most central genes in the modules that are most similar between the healthy and disease networks.

Meta-analysis results: pathway enrichment, drugs and their targets

Table 18 displays the Reactome pathways with the highest enrichment scores in the GSEA analysis, covering all the datasets. The table also includes the top 15 genes ranked in the meta-analysis. There are in total 6187 genes included in the analysis. The Collagen degradation pathway has the highest normalized enrichment score among the most enriched pathways. In Table 19 is illustrated a compilation of drugs that target genes within the Collagen degradation pathway.

Within the top 50 ranked genes in the meta-analysis, *COL1A1* ranked 2nd. *COL1A1* is targeted by collagenase clostridium histolyticum and ocriplasmin. *KCNN4* (potassium calcium-activated channel subfamily N member 4), ranking 7th, is targeted by chlorzoxazone and senicapoc. Odanacatib targets *CTSK* (cathepsin K), which holds the 8th rank in the meta-analysis. ADRB2 (adenoreceptor beta 2), positioned at 12th, is the target of various drugs including corticosteroids (salmeterol, formoterol), beta-blockers (propranolol, carvedilol), and epinephrine (adrenaline). TNC (tenacin C), ranking 28th, is targeted by several phase II cancer drugs such as F16IL2, 81C6 131I, and F16SIP 131I. *COL15A1*, ranked 29th, is targeted by collagenase clostridium histolyticum and ocriplasmin. *RAMP1* (receptor activity modifying protein 1), positioned at 45th, is targeted by Pramlintide and Davalintide, both employed in the treatment of diabetes and obesity. Finally, *EPHB2* (EPH receptor B2), holding the 49th rank, is targeted by Vandetanib, a drug used in the treatment of thyroid carcinoma (Koscielny et al. 2017).

Table 20 shows the Reactome pathways that are most enriched in the GSEA analysis, which includes all datasets except for the biopsy samples. This table is formatted similarly to Table 17. There are in total 7522 genes included in the analysis. The pathway with the highest normalized enrichment score among the most enriched pathways is interferon alpha beta signaling degradation, and Table 21 lists drugs that target genes in this pathway. Chlorzoxazone and senicapoc are found to target the *KCNN4* gene, which ranks 7th in the analysis among the top 50 ranked genes. Ribonucleotide Reductase Regulatory Subunit M2 (*RRM2*), which ranks 22nd, is targeted by various drugs, including cancer medicines such as gemcitabine that are used in a variety of cancers (Koscielny et al. 2017). *CDK1* (cyclin dependent kinase 1) is ranked 46th in the meta-analysis of all datasets except for the biopsy samples, and it is also targeted by a variety of drugs used for treating various cancers, including acute non-small cell lung carcinoma. One example of a drug in this group is seliciclib, which is also used for the treatment of cystic fibrosis (Koscielny et al. 2017).

Table 18: 15 pathways that are most significantly enriched in a meta-analysis that included all datasets. The rightmost column of the table shows the top 15 ranked genes in the meta-analysis.

pathway	pval	padj	ES	NES	Size	top 15 genes
HEMOSTASIS	9.999E-05	0.009333333	0.375022826	1.39405301	235	FHL2
COLLAGEN_DEGRADATION	0.00010102	0.009333333	0.76818329	2.295157776	21	COL1A1
DEGRADATION_OF_THE_EXTRACELLULAR_MATRIX	9.999E-05	0.009333333	0.538333492	1.798165148	51	IGFBP7
EXTRACELLULAR_MATRIX_ORGANIZATION	9.999E-05	0.009333333	0.486527213	1.734781966	113	ADAM12
COLLAGEN_FORMATION	0.00010007	0.009333333	0.560476391	1.821378442	39	GOLM1
COLLAGEN_BIOSYNTHESIS_AND_MODIFYING_ENZYMES	0.000100371	0.009333333	0.619738966	1.922810356	27	KCNN4
ASSEMBLY_OF_COLLAGEN_FIBRILS_AND_OTHER_MULTIMERIC_STRUCTURES	0.000100553	0.009333333	0.642644961	1.972091265	25	CTSK
INTEGRIN_CELL_SURFACE_INTERACTIONS	0.000100371	0.009333333	0.611372614	1.896852801	27	CRABP2
ECM_PROTEOGLYCANS	0.00010019	0.009333333	0.585795549	1.850325838	31	STX3
COLLAGEN_CHAIN_TRIMERIZATION	0.000104987	0.009333333	0.836761134	2.252644313	12	CLDN1
ELASTIC_FIBRE_FORMATION	0.000203604	0.01292884	0.661542827	1.927578259	18	DNAJC22
CELL_SURFACE_INTERACTIONS_AT_THE_VASCULAR_WALL	0.00019998	0.01292884	0.504527131	1.687490502	52	TMEM45A
REGULATION_OF_INSULIN_LIKE_GROWTH_FACTOR_IGF_TRANSPORT_AND_UPTAKE_BY_INSULIN_LIKE_GROWTH_FACTOR_BINDING_PROTEINS_IGFBPS	0.00020008	0.01292884	0.521418615	1.702522378	41	ADRB2
INTERLEUKIN_4_AND_INTERLEUKIN_13_SIGNALING	0.00019998	0.01292884	0.517824039	1.703818947	44	EPB41L5
CELL_JUNCTION_ORGANIZATION	0.00030024	0.017228682	0.532790607	1.715640948	36	UNC13B

Table 19: The drugs that target genes in the Reactome collagen degradation pathway in a meta-analysis that included all datasets.

gene	drug	gene_meta_analysis_rank	target_info	disease_info	molecule_type	drug_phase
COL1A1	COLLAGENASE CLOSTRIDIUM HISTOLYTICUM, OCRIPLASMIN	2	collagen type I alpha 1 chain	Abnormality of connective tissue	Enzyme	Phase IV
CTSK	ODANACATIB	8	cathepsin K	prostate carcinoma	Small molecule	Phase III
COL15A1	COLLAGENASE CLOSTRIDIUM HISTOLYTICUM, OCRIPLASMIN	29	collagen type XV alpha 1 chain	diabetic foot	Enzyme	Phase IV
MMP2	MARIMASTAT	65	matrix metallopeptidase 2	lung carcinoma	Small molecule	Phase III
COL5A1	COLLAGENASE CLOSTRIDIUM HISTOLYTICUM, OCRIPLASMIN	222	collagen type V alpha 1 chain	Skin ulcer	Enzyme	Phase IV
COL6A2	COLLAGENASE CLOSTRIDIUM HISTOLYTICUM, OCRIPLASMIN	233	collagen type VI alpha 2 chain	Dupuytren Contracture	Enzyme	Phase IV
COL6A1	COLLAGENASE CLOSTRIDIUM HISTOLYTICUM, OCRIPLASMIN	265	collagen type VI alpha 1 chain	Dupuytren Contracture	Enzyme	Phase IV
COL4A1	COLLAGENASE CLOSTRIDIUM HISTOLYTICUM, OCRIPLASMIN	360	collagen type IV alpha 1 chain	diabetic foot	Enzyme	Phase IV
COL18A1	COLLAGENASE CLOSTRIDIUM HISTOLYTICUM, OCRIPLASMIN	720	collagen type XVIII alpha 1 chain	diabetic foot	Enzyme	Phase IV
COL4A2	COLLAGENASE CLOSTRIDIUM HISTOLYTICUM, OCRIPLASMIN	1020	collagen type IV alpha 2 chain	diabetic foot	Enzyme	Phase IV

Table 20: 15 pathways that are most significantly enriched in a meta-analysis that included all except biopsy datasets. The rightmost column of the table shows the top 15 ranked genes in the meta-analysis.

pathway	pval	padj	ES	NES	Size	top 15 genes
HEMOSTASIS	9.999E-05	0.019523905	0.388563424	1.456221185	295	AMIGO2
CYTOKINE_SIGNALING_IN_IMMUNE_SYSTEM	9.999E-05	0.019523905	0.352617582	1.336593	409	EMP1
CELL_SURFACE_INTERACTIONS_AT_THE_VASCULAR_WALL	9.999E-05	0.019523905	0.505146544	1.715448256	61	TFPI
REGULATION_OF_INSULIN_LIKE_GROWTH_FACTOR_IGF_TRANSPORT_AND_UPTAKE_BY_INSULIN_LIKE_GROWTH_FACTOR_BINDING_PROTEINS_IGFBPS	9.999E-05	0.019523905	0.536664507	1.794704389	52	TNFRSF12A
INTERFERON_ALPHA_BETA_SIGNALING	0.00010002	0.019523905	0.583965128	1.915467935	43	CSF1
NON_INTEGRIN_MEMBRANE_ECM_INTERACTIONS	0.0002002	0.032565899	0.551476011	1.764861906	35	CCNA2
CHOLESTEROL_BIOSYNTHESIS	0.000303613	0.037446286	0.605112483	1.788128805	20	KCNN4
MET_ACTIVATES_PTK2_SIGNALING	0.000306937	0.037446286	0.665877754	1.891589226	16	MMP10
GPCR_LIGAND_BINDING	0.00049995	0.0443592	0.474584335	1.603071444	58	CLDN1
METABOLISM_OF_LIPIDS	0.00049995	0.0443592	0.330529971	1.253384246	415	CEP55
INTERLEUKIN_10_SIGNALING	0.000415973	0.0443592	0.731655674	1.991657029	13	KIF20A
DEPOLYMERISATION_OF_THE_NUCLEAR_LAMINA	0.000748023	0.051419841	0.729924763	1.871225724	10	SERPING1
CELL_JUNCTION_ORGANIZATION	0.000900901	0.051419841	0.50109129	1.614093541	37	NT5DC2
CYCLIN_A_B1_B2_ASSOCIATED_EVENTS_DURING_G2_M_TRANSITION	0.000706001	0.051419841	0.571662731	1.725782575	23	SLC1A4
INTERFERON_GAMMA_SIGNALING	0.0008	0.051419841	0.482430908	1.605128383	49	CHST15

Table 21: The drugs that target genes in the Reactome interferon alpha beta signaling pathway in a meta-analysis that included all except biopsy datasets.

gene	drug	gene_meta_analysis_rank	target_info	disease_info	molecule_type	drug_phase
PSMB8	CARFILZOMIB, BORTEZOMIB, IXAZOMIB CITRATE, MARIZOMIB, OPROZOMIB	1639	proteasome subunit beta 8	neoplasm	Protein	Phase IV

Table 22 displays the most enriched Reactome pathways in the GSEA analysis of the biopsy datasets. There are in total 9645 genes included in the analysis. These enriched pathways have many similarities with the pathways in the analysis that includes all samples, most likely because the biopsy datasets are the most abundant. The Syndecan interactions pathway has the highest normalized enrichment score of the most enriched pathways in the biopsy datasets, and the drugs targeting genes on this pathway are listed in Table 23.

Among the highest-ranked genes in the meta-analysis of the biopsy datasets, *CDH3* (cadherin 3) is targeted by PF-03732010, which is currently in clinical trials for neoplasm. *CDH3* ranked 2nd in the biopsy meta-analysis. *MMP7*, ranked 6th in the meta-analysis, is targeted by marimastat and doxycycline. *COL1A1*, ranked 7th, is targeted by collagenase clostridium histolyticum and ocriplasmin. *IL13RA2*, ranked 8th, is the target of Cintredekin besudotox (phase I), which is used in central nervous system cancer. Insulin-like growth factor 1 (*IGF1*), ranked 30th, is targeted by dusigitumab, used in various cancers. *FAP* (fibroblast activation protein alpha), ranked 36th, is targeted by sibrotuzumab, currently undergoing clinical trials for non-small cell lung carcinoma. *CD27* (CD27 molecule), ranked 40th, is targeted by varlilumab, used for melanomas and lymphomas (Koscielny et al. 2017).

SMO (smoothened, frizzled class receptor), ranked 44th in the meta-analysis of biopsy samples, is the target of several drugs primarily used in various cancers. Vismodegib is one of these drugs, currently being studied for IPF treatment alongside pirfenidone. *CHRM3* (cholinergic receptor muscarinic 3), ranked 45th, is the target of several cholinergic and anticholinergic medicines, including acetylcholine and ipratropium (Koscielny et al. 2017).

Table 24 displays the most enriched Reactome pathways in the GSEA analysis of the BAL datasets. A total of 13301 genes were included in the analysis, but none of the Reactome pathways were found to be significantly enriched. However, the ECM organization pathway was the closest. The drugs targeting genes on this pathway are listed in Table 25. In the meta-analysis of the BAL samples, there are 5 genes among the top 50 ranked that have drug targets. *CFTR* (cystic fibrosis transmembrane conductance regulator) is ranked 4th and is targeted by ivacaftor, tezacaftor, and lumacaftor, which are used to treat cystic fibrosis. *SNCA* (synuclein-α) is ranked 13th and is targeted by Parkinson's drug BIIB054, which is in phase II trials. *SCNN1A* (sodium channel epithelial 1 α subunit) is ranked 20th and is targeted by two hypertension medicines, amiloride and triamterene. *IL1B* (interleukin 1-β) is ranked 39th and is targeted by rilonacept, canakinumab, and gevokizumab, which are proteins used for several autoimmune diseases like arthritis and gout, but also for non-small cell lung carcinoma (canakinumab). *MST1R* (macrophage stimulating 1 receptor) is ranked 39th and is targeted by several cancer medicines, such as narnatumab. All of these drugs are currently in phase I or phase II of clinical trials (Koscielny et al. 2017).

Table 22: 15 pathways that are most significantly enriched in a meta-analysis that included biopsy datasets. The rightmost column of the table shows the top 15 ranked genes in the meta-analysis.

pathway	pval	padj	ES	NES	Size	top 15 genes
COLLAGEN_DEGRADATION	0.00010005	0.0104154	0.591191464	1.926519078	41	FNDC1
DEGRADATION_OF_THE_EXTRACELLULAR_MATRIX	9.999E-05	0.0104154	0.479355881	1.674025144	84	CDH3
EXTRACELLULAR_MATRIX_ORGANIZATION	9.999E-05	0.0104154	0.450714954	1.656661892	193	ASPN
ASSEMBLY_OF_COLLAGEN_FIBRILS_AND_OTHER_MULTIMERIC_STRUCTURES	0.00010002	0.0104154	0.544372362	1.801952712	47	PTGFRN
MOLECULES_ASSOCIATED_WITH_ELASTIC_FIBRES	0.000100563	0.0104154	0.619653983	1.897442187	25	CTHRC1
INTEGRIN_CELL_SURFACE_INTERACTIONS	9.999E-05	0.0104154	0.567079882	1.918363833	58	MMP7
SYNDECAN_INTERACTIONS	0.000101317	0.0104154	0.70721551	2.095441706	20	COL1A1
ECM_PROTEOGLYCANS	0.0001	0.0104154	0.542409552	1.820781215	54	IL13RA2
REGULATION_OF_INSULIN_LIKE_GROWTH_FACTOR_IGF_TRANSPORT _AND_UPTAKE_BY_INSULIN_LIKE_GROWTH_FACTOR_BINDING_PROTEINS_IGFBPS	9.999E-05	0.0104154	0.50089747	1.712147352	65	COL14A1
COLLAGEN_CHAIN_TRIMERIZATION	0.000100321	0.0104154	0.626909569	1.967049241	30	CXCL14
COLLAGEN_FORMATION	0.00019998	0.015543728	0.464534249	1.592513021	67	DIO2
ACTIVATION_OF_MATRIX_METALLOPROTEINASES	0.000211685	0.015543728	0.73632275	1.943679247	11	COL15A1
COLLAGEN_BIOSYNTHESIS_AND_MODIFYING_ENZYMES	0.00020004	0.015543728	0.512546093	1.699595024	48	FHL2
BINDING_AND_UPTAKE_OF_LIGANDS_BY_SCAVENGER_RECEPTORS	0.000200965	0.015543728	0.580325952	1.803891181	28	FAM167A
HEMOSTASIS	0.00029997	0.01819029	0.337519893	1.27351481	361	COL10A1

Table 23: The drugs that target genes in the Reactome syndecan interactions pathway in a meta-analysis that included biopsy datasets.

gene	drug	gene_meta_analysis_rank	target_info	disease_info	molecule_type	drug_phase
COL1A1	COLLAGENASE CLOSTRIDIUM HISTOLYTICUM OCRIPLASMIN	7	collagen type I alpha 1 chain	Abnormality of connective tissue	Enzyme	Phase IV
COL3A1	COLLAGENASE CLOSTRIDIUM HISTOLYTICUM OCRIPLASMIN	24	collagen type III alpha 1 chain	diabetic foot	Enzyme	Phase IV
COL1A2	COLLAGENASE CLOSTRIDIUM HISTOLYTICUM OCRIPLASMIN	35	collagen type I alpha 2 chain	Skin ulcer	Enzyme	Phase IV
COL5A1	COLLAGENASE CLOSTRIDIUM HISTOLYTICUM OCRIPLASMIN	74	collagen type V alpha 1 chain	Skin ulcer	Enzyme	Phase IV
TNC	F16IL2, 81C6 131I, F16SIP 131I	106	tenascin C	Merkel cell skin cancer	Antibody	Phase II
COL5A2	COLLAGENASE CLOSTRIDIUM HISTOLYTICUM OCRIPLASMIN	194	collagen type V alpha 2 chain	diabetic foot	Enzyme	Phase IV
FN1	OCRIPLASMIN, L19IL2, L19TNFA, L19SIP 131I	497	fibronectin 1	macular holes	Enzyme	Phase IV
ITGAV	ABCIXIMAB, CILENGITID, ETARACIZUMAB STX-100, INTETUMUMAB, ABITUZUMAB	621	integrin subunit alpha V	acute coronary syndrome	Antibody	Phase IV
ITGB5	CILENGITIDE, INTETUMUMAB, ABITUZUMAB	849	integrin subunit beta 5	glioblastoma multiforme	Protein	Phase III
ITGB1	NATALIZUMAB, VOLOCIXIMAB, INTETUMUMAB FIRATEGRAST, ABITUZUMAB	1291	integrin subunit beta 1	multiple sclerosis	Antibody	Phase IV

Table 24: 15 pathways that are most significantly enriched in a meta-analysis that included BAL datasets. The rightmost column of the table shows the top 15 ranked genes in the meta-analysis.

pathway	**pval**	**padj**	**ES**	**NES**	**Size**	**top 15 genes**
EXTRACELLULAR_MATRIX_ORGANIZATION	9.999E-05	0.115988401	0.363094305	1.340513702	207	FAH
METABOLISM_OF_NUCLEOTIDES	0.00049995	0.144985501	0.427786309	1.495323353	84	PERP
REGULATION_OF_INSULIN_LIKE_GROWTH_FACTOR_IGF_TRANSPORT _AND_UPTAKE_BY_INSULIN_LIKE_GROWTH_FACTOR_BINDING_PROTEINS_IGFBPS	0.00049995	0.144985501	0.431792407	1.495863822	75	CFTR
MUSCLE_CONTRACTION	0.00039996	0.144985501	0.397728393	1.423035117	118	CD207
GASTRULATION	0.001001402	0.232325255	0.522498067	1.663697515	33	ENPP3
PHASE_0_RAPID_DEPOLARISATION	0.001343114	0.251132985	0.632768344	1.758809173	14	LY75
SURFACTANT_METABOLISM	0.001515458	0.251132985	0.565874301	1.689738885	21	PABPC4
DEGRADATION_OF_THE_EXTRACELLULAR_MATRIX	0.0029997	0.267826221	0.395310546	1.390381828	91	CREB3L1
SLC_MEDIATED_TRANSMEMBRANE_TRANSPORT	0.00249975	0.267826221	0.364999519	1.323552191	146	FAM83E
ACTIVATION_OF_ANTERIOR_HOX_GENES_IN_HINDBRAIN_DEVELOPMENT_DURING_EARLY_EMBRYOGENESIS	0.0030003	0.267826221	0.4429041	1.481842447	51	FCGBP
O_LINKED_GLYCOSYLATION_OF_MUCINS	0.003001501	0.267826221	0.478272406	1.543546213	37	ARHGAP44
ALK_MUTANTS_BIND_TKIS	0.002325581	0.267826221	-0.499510874	-1.841084506	12	SNCA
KINESINS	0.00190076	0.267826221	0.471991108	1.541752665	41	MYO1A
INTERLEUKIN_4_AND_INTERLEUKIN_13_SIGNALING	0.00379962	0.31482566	0.392871632	1.380460173	90	TCEA3
DEFECTIVE_C1GALT1C1_CAUSES_TNPS	0.004599422	0.355688665	0.658968611	1.701547237	10	KRT19

Table 25: The drugs that target genes in the Reactome extracellular matrix organization pathway in a meta-analysis that included BAL datasets.

gene	drug	gene_meta_analysis_rank	target_info	disease_info	molecule_type	drug_phase
ITGAL	EFALIZUMAB, LIFITEGRAST	94	integrin subunit alpha L	psoriasis	Antibody	Phase IV
CD44	BIVATUZUMAB	140	CD44 molecule (Indian blood group)	breast neoplasm	Antibody	Phase I
COL4A1	COLLAGENASE CLOSTRIDIUM HISTOLYTICUM, OCRIPLASMIN	268	collagen type IV alpha 1 chain	diabetic foot	Enzyme	Phase IV
P4HB	LOMITAPIDE	439	prolyl 4-hydroxylase subunit beta	cardiovascular disease	Small molecule	Phase IV
COL28A1	COLLAGENASE CLOSTRIDIUM HISTOLYTICUM, OCRIPLASMIN	693	collagen type XXVIII alpha 1 chain	Dupuytren Contracture, macular holes	Enzyme	Phase IV
PRKCA	MIDOSTAURIN, UCN-01, SOTRASTAURIN, GSK-690693	993	protein kinase C alpha	neoplasm	Small molecule	Phase IV
ICAM1	BI-505	1021	intercellular adhesion molecule 1	multiple myeloma	Antibody	Phase II
COL18A1	COLLAGENASE CLOSTRIDIUM HISTOLYTICUM, OCRIPLASMIN	1122	collagen type XVIII alpha 1 chain	diabetic foot, macular holes	Enzyme	Phase IV
MMP9	MARIMASTAT, ANDECALIXIMAB	1206	matrix metallopeptidase 9	lung carcinoma	Small molecule	Phase III
ITGA5	VOLOCIXIMAB, PF-04605412	1258	integrin subunit alpha 5	lung carcinoma	Antibody	Phase III
CASP3	EMRICASAN	1377	caspase 3	non-alcoholic steatohepatitis	Small molecule	Phase II
LAMA5	OCRIPLASMIN	1893	laminin subunit alpha 5	macular holes	Enzyme	Phase IV
CTSK	ODANACATIB	1961	cathepsin K	prostate carcinoma	Small molecule	Phase III
COL27A1	COLLAGENASE CLOSTRIDIUM HISTOLYTICUM, OCRIPLASMIN	2050	collagen type XXVII alpha 1 chain	Dupuytren Contracture, macular holes	Enzyme	Phase IV
TNC	F16IL2, 81C6 131I, F16SIP 131I	2094	tenascin C	Merkel cell skin cancer	Antibody	Phase II
MMP12	MARIMASTAT	2213	matrix metallopeptidase 12	lung carcinoma	Small molecule	Phase III
FGG	FIBRINOLYSIN, HUMAN	2234	fibrinogen gamma chain	Recurrent thrombophlebitis	Unknown	Phase IV
ITGB7	VEDOLIZUMAB, NATALIZUMAB, ETROLIZUMAB, FIRATEGRAST, ABRILUMAB	2335	integrin subunit beta 7	Crohn's disease	Antibody	Phase IV
COL6A2	COLLAGENASE CLOSTRIDIUM HISTOLYTICUM, OCRIPLASMIN	2920	collagen type VI alpha 2 chain	Dupuytren Contracture	Enzyme	Phase IV
LAMC2	OCRIPLASMIN	2994	laminin subunit gamma 2	macular holes	Enzyme	Phase IV
ITGB3	TIROFIBAN, ABCIXIMAB, EPTIFIBATIDE, CILENGITIDE, INTETUMUMAB, ETARACIZUMAB, ABITUZUMAB	3098	integrin subunit beta 3	Non-ST Elevation Myocardial Infarction	Small molecule	Phase IV
MMP7	DOXYCYCLINE, MARIMASTAT	3463	matrix metallopeptidase 7	periodontitis	Small molecule	Phase IV
LAMB1	OCRIPLASMIN	3469	laminin subunit beta 1	macular holes	Enzyme	Phase IV
COL1A1	COLLAGENASE CLOSTRIDIUM HISTOLYTICUM, OCRIPLASMIN	3609	collagen type I alpha 1 chain	Abnormality of connective tissue, macular holes	Enzyme	Phase IV
VWF	CAPLACIZUMAB	3691	von Willebrand factor	autoimmune thrombocytopenic purpura	Antibody	Phase III
LAMB3	OCRIPLASMIN	3692	laminin subunit beta 3	macular holes	Enzyme	Phase IV
LAMB2	OCRIPLASMIN	3852	laminin subunit beta 2	macular holes	Enzyme	Phase IV
SDC1	INDATUXIMAB RAVTANSINE	4054	syndecan 1	multiple myeloma	Antibody	Phase I
ITGB5	CILENGITIDE, INTETUMUMAB, ABITUZUMAB	4082	integrin subunit beta 5	glioblastoma multiforme	Protein	Phase III
SERPINE1	ALEPLASININ	4855	serpin family E member 1	Alzheimer's disease	Small molecule	Phase I
MMP1	DOXYCYCLINE, MARIMASTAT	4998	matrix metallopeptidase 1	acne	Small molecule	Phase IV
APP	BAPINEUZUMAB, SOLANEZUMAB, GANTENERUMAB, CRENEZUMAB, ADUCANUMAB, PONEZUMAB, GSK933776, BAN2401	5098	amyloid beta precursor protein	Alzheimer's disease	Antibody	Phase III
NCSTN	TARENFLURBIL, SEMAGACESTAT, AVAGACESTAT, BEGACESTAT	5150	nicastrin	Alzheimer's disease	Small molecule	Phase III
LAMA3	OCRIPLASMIN	5194	laminin subunit alpha 3	macular holes	Enzyme	Phase IV
LAMC1	OCRIPLASMIN	5449	laminin subunit gamma 1	macular holes	Enzyme	Phase IV
COL4A4	COLLAGENASE CLOSTRIDIUM HISTOLYTICUM, OCRIPLASMIN	5617	collagen type IV alpha 4 chain	Dupuytren Contracture, macular holes	Enzyme	Phase IV

Table 26: 15 pathways that are most significantly enriched in a meta-analysis that included fibroblast datasets. The rightmost column of the table shows the top 15 ranked genes in the meta-analysis.

pathway	pval	padj	ES	NES	Size	top 15 genes
EXTRACELLULAR_MATRIX_ORGANIZATION	9.999E-05	0.015605493	0.404771006	1.487902326	192	CADM1
CHEMOKINE_RECEPTORS_BIND_CHEMOKINES	0.000104037	0.015605493	0.743172913	2.033178799	13	IFI44L
COOPERATION_OF_PREFOLDIN_AND_TRIC_CCT_IN_ACTIN_AND_TUBULIN_FOLDING	0.00010098	0.015605493	0.613012416	1.839443808	22	SGCE
INTERLEUKIN_10_SIGNALING	0.000102522	0.015605493	0.727547578	2.069384909	16	RTP4
INTERFERON_GAMMA_SIGNALING	9.999E-05	0.015605493	0.553842696	1.863684641	55	DDX60
INTERFERON_ALPHA_BETA_SIGNALING	0.0001	0.015605493	0.680334439	2.237663703	44	EIF2AK2
INTERFERON_SIGNALING	9.999E-05	0.015605493	0.451991269	1.629284676	134	PKIG
POST_CHAPERONIN_TUBULIN_FOLDING_PATHWAY	0.00021015	0.027582221	0.712801018	1.912908074	12	ASB1
HEMOSTASIS	0.00039996	0.046662	0.339404304	1.279213132	351	TNFRSF10D
CYTOKINE_SIGNALING_IN_IMMUNE_SYSTEM	0.00049995	0.048716212	0.332807437	1.263831656	443	IFI35
FORMATION_OF_TUBULIN_FOLDING_INTERMEDIATES_BY_CCT_TRIC	0.00051036	0.048716212	0.633127765	1.820484539	17	PSMB9
ELASTIC_FIBRE_FORMATION	0.001001201	0.075090108	0.496885233	1.589798456	35	SLC19A2
CLASS_A_1_RHODOPSIN_LIKE_RECEPTORS	0.0009999	0.075090108	0.427891237	1.479876632	75	ARHGEF6
GPCR_LIGAND_BINDING	0.00089991	0.075090108	0.405125352	1.440441806	108	GAL
NEUTROPHIL_DEGRANULATION	0.00189981	0.132986701	0.339262847	1.268544654	282	PPL

Table 27: The drugs that target genes in the Reactome interferon alpha beta signaling pathway in a meta-analysis that included fibroblast datasets.

gene	drug	gene_meta_analysis_rank	target_info	disease_info	molecule_type	drug_phase
PSMB8	CARFILZOMIB, BORTEZOMIB, IXAZOMIB CITRATE, MARIZOMIB, OPROZOMIB	1055	proteasome subunit beta 8	neoplasm	Protein	Phase IV

Table 26 displays the most enriched Reactome pathways in the GSEA analysis of the fibroblast datasets. There are in total 8512 genes included in the analysis. These enriched pathways have many similarities with the pathways in the analysis that includes all except the biopsy samples, most likely because the fibroblast datasets are the most abundant in the all except the biopsy analysis. The interferon alpha beta signaling pathway has the highest normalized enrichment score of the most enriched pathways in the biopsy datasets, and the drugs targeting genes on this pathway are listed in Table 27. In table 27 the target gene is *PSMB8* (proteasome subunit-β 8) which is ranked 1055th in the meta-analysis. *PSMB9* rank in the meta-analysis is 9 and is targeted by the same drugs which are used for treatment of several cancers. In addition to this *INHBA* (inhibin subunit-β A) was targeted by sotatercept which is in clinical trials for the treatment of pulmonary hypertension (Koscielny et al. 2017).

Table 28 shows enriched Reactome pathways in macrophage datasets with the highest normalized enrichment score. TNFs bind their receptor pathway is the second most enriched pathway and drugs targeting it are listed in Table 29. Only tarexutumab targets *NOTCH3*, ranked 23rd among top 50 genes in analysis, used in various cancers (Koscielny et al. 2017). There are 11871 genes in the analysis, and the *SEMA4D* induced cell migration and growth-cone collapse pathway lacks drug targets.

Table 30 shows the Reactome pathways that have the highest level of enrichment in the epithelial datasets. There are in total 13527 genes included in the analysis. The keratinization pathway has the highest normalized enrichment score among the most enriched pathways in the epithelial datasets, but unfortunately, it doesn't have any known drug targets. On the other hand, the DNA strand elongation pathway has the highest normalized enrichment score and has drugs targeting it listed in Table 31. Additionally, Table 32 lists the drugs targeting the genes in the extracellular matrix organization Reactome pathway in the epithelial meta-analysis GSEA, which was also significantly enriched.

In the epithelial dataset, *CHRNA1* (cholinergic receptor nicotinic alpha 1 subunit) is ranked sixth in the meta-analysis and is the target of various muscle relaxants, such as rocuronium and suxamethonium. *CACNA1H* (calcium voltage-gated channel subunit alpha1 H) is ranked seventh in the meta-analysis and is a target of calcium channel modulators, such as pregabalin and gabapentin, which are used to treat fibromyalgia, epilepsy, and pain. *SCNN1A* (sodium channel epithelial 1 alpha subunit) is ranked twelfth in the meta-analysis and is the target of amiloride and triamterene, which have been used to treat hypertension. The same gene ranked twentieth in the BAL meta-analysis. *CSF2* (colony-stimulating factor 2) ranked thirtieth in the epithelial meta-analysis and is the target of drugs in phase I or phase II clinical trials for the treatment of psoriasis, rheumatoid arthritis, and multiple sclerosis, such as namilumab, lenzilumab, MOR-103, and gimsilumab. (Koscielny et al. 2017)

Table 28: 15 pathways that are most significantly enriched in a meta-analysis that included macrophage datasets. The rightmost column of the table shows the top 15 ranked genes in the meta-analysis.

pathway	pval	padj	ES	NES	Size	top 15 genes
CYTOKINE_SIGNALING_IN_IMMUNE_SYSTEM	0.00149985	0.141698362	0.308687978	1.18347844	625	TRHDE
ADAPTIVE_IMMUNE_SYSTEM	0.00039996	0.141698362	0.315646531	1.208798144	596	GPC4
INNATE_IMMUNE_SYSTEM	0.00069993	0.141698362	0.305464338	1.180198172	855	PROS1
IMMUNOREGULATORY_INTERACTIONS_BETWEEN_A_LYMPHOID_AND_A_NON_LYMPHOID_CELL	0.00049995	0.141698362	0.417975334	1.475503524	97	GPR85
GLUTAMATE_NEUROTRANSMITTER_RELEASE_CYCLE	0.001513623	0.141698362	0.561487703	1.68927985	22	TBKBP1
SEMA4D_INDUCED_CELL_MIGRATION_AND_GROWTH_CONE_COLLAPSE	0.000203128	0.141698362	0.63768204	1.885856568	20	RAI14
TNFS_BIND_THEIR_PHYSIOLOGICAL_RECEPTORS	0.00050317	0.141698362	0.580610406	1.770097275	24	F3
RESPIRATORY_ELECTRON_TRANSPORT	0.00109989	0.141698362	0.431274309	1.479187454	67	DSP
NEUTROPHIL_DEGRANULATION	0.00149985	0.141698362	0.322838614	1.222210328	399	SMAD7
RHO_GTPASE_CYCLE	0.00139986	0.141698362	0.329709808	1.24259779	352	ACSM3
RAC1_GTPASE_CYCLE	0.00069993	0.141698362	0.388783498	1.401212302	136	ASAP2
TRANSCRIPTIONAL_REGULATION_BY_NPAS4	0.001106417	0.141698362	0.556642933	1.707443655	25	GASK1B
SIGNALING_BY_RHO_GTPASES_MIRO_GTPASES_AND_RHOBTB3	0.0009999	0.141698362	0.316299366	1.207953964	534	TSC22D3
SEMA4D_IN_SEMAPHORIN_SIGNALING	0.001811412	0.157463448	0.547604097	1.669471493	24	SLC7A5
PLASMA_LIPOPROTEIN_CLEARANCE	0.002305302	0.187036851	0.500721831	1.586519989	32	RMDN3

Table 29: The drugs that target genes in the Reactome TNFs bind their physiological receptor pathway in a meta-analysis that included macrophage datasets. SEMA4D pathways did not include drug targets.

gene	drug	gene_meta_analysis_rank	target_info	disease_info	molecule_type	drug_phase
TNFRSF9	UTOMILUMAB, URELUMAB	227	TNF receptor superfamily member 9	diffuse large B-cell lymphoma	Antibody	Phase III
TNFRSF1A	GSK-1995057	822	TNF receptor superfamily member 1A	respiratory system disease	Antibody	Phase I
TNFSF4	OXELUMAB	1405	TNF superfamily member 4	asthma	Antibody	Phase II
LTA	BAMINERCEPT, PATECLIZUMAB	1565	lymphotoxin alpha	rheumatoid arthritis	Protein	Phase II

Table 30: 15 pathways that are most significantly enriched in a meta-analysis that included epithelial datasets. The rightmost column of the table shows the top 15 ranked genes in the meta-analysis.

pathway	pval	padj	ES	NES	Size	top 15 genes
EXTRACELLULAR_MATRIX_ORGANIZATION	9.999E-05	0.010574004	0.369169412	1.360040134	205	CHST6
CELL_CYCLE	9.999E-05	0.010574004	0.334399625	1.281847676	599	C2CD4A
RHO_GTPASE_EFFECTORS	9.999E-05	0.010574004	0.372478012	1.381903063	240	PLPPR1
SLC_MEDIATED_TRANSMEMBRANE_TRANSPORT	9.999E-05	0.010574004	0.413252641	1.495729275	151	CHRNA1
KERATINIZATION	9.999E-05	0.010574004	0.533606135	1.855052851	83	CACNA1H
FORMATION_OF_THE_CORNIFIED_ENVELOPE	9.999E-05	0.010574004	0.528244332	1.834863926	82	TMPRSS11D
MITOTIC_PROMETAPHASE	9.999E-05	0.010574004	0.39324437	1.443427064	191	C15orf48
DNA_STRAND_ELONGATION	0.000100271	0.010574004	0.593148499	1.875074984	32	KRT7
CELL_CYCLE_MITOTIC	9.999E-05	0.010574004	0.357247772	1.361380781	483	SCNN1A
CELL_CYCLE_CHECKPOINTS	9.999E-05	0.010574004	0.382033157	1.420366866	252	AMY1A
O_LINKED_GLYCOSYLATION_OF_MUCINS	0.00010006	0.010574004	0.529460056	1.73342222	43	BPGM
ACTIVATION_OF_ATR_IN_RESPONSE_TO_REPLICATION_STRESS	0.0002003	0.014549468	0.533628936	1.710225842	36	MMP10
RESOLUTION_OF_SISTER_CHROMATID_COHESION	0.00019998	0.014549468	0.412009928	1.467175773	116	ASB2
INTRA_GOLGI_AND_RETROGRADE_GOLGI_TO_ER_TRAFFIC	0.00019998	0.014549468	0.369007893	1.350720666	182	TMPRSS11B
SIGNALING_BY_RHO_GTPASES_MIRO_GTPASES_AND_RHOBTB3	0.00019998	0.014549468	0.323037155	1.238516909	604	KRT75

Table 31: The drugs that target genes in the Reactome DNA strand elongation pathway in a meta-analysis that included epithelial datasets. Keratinization pathway did not include drug targets.

gene	drug	gene_meta_analysis_rank	target_info	disease_info	molecule_type	drug_phase
POLD1	CYTARABINE, GEMCITABINE, FLUDARABINE PHOSPHATE, CLOFARABINE, TROXACITABINE	613	DNA polymerase delta 1, catalytic subunit	acute lymphoblastic leukemia	Small molecule	Phase IV
PRIM1	GEMCITABINE, FLUDARABINE PHOSPHATE, CYTARABINE, CLOFARABINE, TROXACITABINE	749	DNA primase subunit 1	pancreatic adenocarcinoma	Small molecule	Phase IV
POLD3	FLUDARABINE PHOSPHATE, CYTARABINE, GEMCITABINE, CLOFARABINE, TROXACITABINE	930	DNA polymerase delta 3, accessory subunit	chronic lymphocytic leukemia	Small molecule	Phase IV
POLA2	CYTARABINE, GEMCITABINE, FLUDARABINE PHOSPHATE, CLOFARABINE, TROXACITABINE	2629	DNA polymerase alpha 2, accessory subunit	acute lymphoblastic leukemia	Small molecule	Phase IV
POLD2	CYTARABINE, GEMCITABINE, FLUDARABINE PHOSPHATE, CLOFARABINE, TROXACITABINE	2810	DNA polymerase delta 2, accessory subunit	acute lymphoblastic leukemia	Small molecule	Phase IV

Table 32: The drugs that target genes in the extracellular matrix organization pathway in a meta-analysis that included epithelial datasets.

gene	drug	gene_meta_analysis_rank	target_info	disease_info	molecule_type	drug_phase
ITGB7	VEDOLIZUMAB, NATALIZUMAB, ETROLIZUMAB, FIRATEGRAST, ABRILUMAB	125	integrin subunit beta 7	Crohn's disease	Antibody	Phase IV
ICAM1	BI-505	161	intercellular adhesion molecule 1	multiple myeloma	Antibody	Phase II
APP	BAPINEUZUMAB, SOLANEZUMAB, GANTENERUMAB, CRENEZUMAB, ADUCANUMAB, PONEZUMAB, GSK933776, BAN2401	267	amyloid beta precursor protein	Alzheimer's disease	Antibody	Phase III
VWF	CAPLACIZUMAB	508	von Willebrand factor	autoimmune thrombocytopenic purpura	Antibody	Phase III
MMP3	MARIMASTAT	1077	matrix metallopeptidase 3	lung carcinoma	Small molecule	Phase III
ITGB2	EFALIZUMAB, LIFITEGRAST, AME-133V	1131	integrin subunit beta 2	psoriasis	Antibody	Phase IV
PSEN1	SEMAGACESTAT, TARENFLURBIL, AVAGACESTAT, BEGACESTAT	1155	presenilin 1	Alzheimer's disease	Small molecule	Phase III
PDGFB	PEGPLERANIB SODIUM, RINUCUMAB	1331	platelet derived growth factor subunit B	age-related macular degeneration	Unknown	Phase III
FN1	OCRIPLASMIN, L19IL2, L19TNFA, L19SIP 131I	1452	fibronectin 1	macular holes	Enzyme	Phase IV
ITGA5	VOLOCIXIMAB, PF-04605412	1602	integrin subunit alpha 5	lung carcinoma	Antibody	Phase III
ITGB1	NATALIZUMAB, VOLOCIXIMAB, INTETUMUMAB, FIRATEGRAST, ABITUZUMAB	1671	integrin subunit beta 1	multiple sclerosis	Antibody	Phase IV
NCSTN	TARENFLURBIL, SEMAGACESTAT, AVAGACESTAT, BEGACESTAT	1706	nicastrin	Alzheimer's disease	Small molecule	Phase III
ITGA4	VEDOLIZUMAB, NATALIZUMAB, ABRILUMAB, FIRATEGRAST	1772	integrin subunit alpha 4	Crohn's disease	Antibody	Phase IV
CD44	BIVATUZUMAB	1803	CD44 molecule (Indian blood group)	breast neoplasm	Antibody	Phase I
LAMC1	OCRIPLASMIN	1857	laminin subunit gamma 1	macular holes	Enzyme	Phase IV
SERPINE1	ALEPLASININ	1931	serpin family E member 1	Alzheimer's disease	Small molecule	Phase I
LAMB1	OCRIPLASMIN	2133	laminin subunit beta 1	macular holes	Enzyme	Phase IV
COL4A1	COLLAGENASE CLOSTRIDIUM HISTOLYTICUM, OCRIPLASMIN	2273	collagen type IV alpha 1 chain	diabetic foot	Enzyme	Phase IV
ITGAV	ABCIXIMAB, CILENGITIDE, ETARACIZUMAB, STX-100, INTETUMUMAB, ABITUZUMAB	2769	integrin subunit alpha V	acute coronary syndrome	Antibody	Phase IV
ITGA2	VATELIZUMAB	2925	integrin subunit alpha 2	ulcerative colitis	Antibody	Phase II
LAMA3	OCRIPLASMIN	3470	laminin subunit alpha 3	macular holes	Enzyme	Phase IV
COL5A2	COLLAGENASE CLOSTRIDIUM HISTOLYTICUM, OCRIPLASMIN	3531	collagen type V alpha 2 chain	diabetic foot	Enzyme	Phase IV
LAMC2	OCRIPLASMIN	3547	laminin subunit gamma 2	macular holes	Enzyme	Phase IV
COL1A1	COLLAGENASE CLOSTRIDIUM HISTOLYTICUM, OCRIPLASMIN	3660	collagen type I alpha 1 chain	Abnormality of connective tissue	Enzyme	Phase IV
LAMB3	OCRIPLASMIN	3704	laminin subunit beta 3	macular holes	Enzyme	Phase IV
COL5A1	COLLAGENASE CLOSTRIDIUM HISTOLYTICUM, OCRIPLASMIN	3754	collagen type V alpha 1 chain	Skin ulcer	Enzyme	Phase IV
ITGB6	STX-100, INTETUMUMAB, ABITUZUMAB	3764	integrin subunit beta 6	idiopathic pulmonary fibrosis	Antibody	Phase II
KLKB1	ECALLANTIDE, APROTININ, LANADELUMAB	3954	kallikrein B1	Hereditary angioedema	Protein	Phase IV
BSG	GAVILIMOMAB	3965	basigin (Ok blood group)	acute graft vs. host disease	Antibody	Phase III

Discussion

IPF druggability based on the network inference and meta-analysis results

Table 32: The noteworthy drug categories determined from the analysis of networks and meta-analysis outcomes.

Enzymes	Tyrosine kinase inhibitors	Matrix metalloproteinase inhibitors	Ion channel modulators /inbibitors	Monoclonal antibodies
Ocriplasmin	Regorafenib	Marimastat	Isosorbide mononitrate	Simtuzumab
Collagenase clostridium histolyticum	Dasatinib	Doxycycline	Gabapentin	Canakinumab
	Ilorasertib	Andecaliximab	Pregabalin	Andecaliximab
	Seliciclib		Amiloride	Narnatumab
	Nintedanib		Nitroglycerin	Sibrotuzumab
			Isosorbide dinitrate	Fresolimumab
			Senicapoc	STX-100

Several promising drugs for treating IPF were identified through network and meta-analysis. Table 32 summarizes the drug categories that emerged from the analysis, along with specific drugs within each category. Notable categories include enzymes, tyrosine kinase inhibitors, matrix metalloproteinase inhibitors, ion channel modulators and inhibitors, and biological drugs. These drug classes will be further discussed in the following sections.

The drugs that had been associated to IPF in the OpenTargets (Koscielny et al. 2017) table were **nintedanib**, prednisolone phosphoric acid, **doxycycline**, ambrisentan, bosentan, interferon gamma-1b, fentanyl, sildenafil, thalidomide, prednisolone, warfarin, macitentan, azathioprine, gefapixant, rituximab, tanzisertib, tralokinumab, imatinib, **simtuzumab**, **stx-100**, **vismodegib**, pamrevlumab, omeprazole, qax-576, lebrikizumab, losartan, beclomethasone dipropionate, zileuton, prednisone, octreotide, **dasatinib**, omipalisib, **fresolimumab** and albuterol. The bolded drugs are the ones that were found in this study.

All the drugs associated to IPF in the OpenTargets (Koscielny et al. 2017) which were found in the analyses of this study are listed in the Table 31 except vismodegib. In the analysis vismodegib was found in the biopsy meta-analysis where it's target gene SMO (smoothened, frizzled class receptor) was ranked 44th. Vismodegib was the first pharmacologic agent approved by the FDA in 2012 to target the Hedgehog signaling pathway, particularly Sonic Hedgehog (SHH), which is associated with many basal cell carcinomas. Basal cell carcinoma (BCC) is a prevalent form of nonmelanoma skin cancer (NMSC) and accounts for more than 50 % of all NMSC cases (Zito et al. 2023). Vismodegib has been in clinical trials for IPF used with pirfenidone (Prasse et al. 2019a).

Enzymes

According to the analysis, Ocriplasmin and Collagenase clostridium histolyticum were the most frequently identified drugs. In the biopsy network analysis, they targeted the 6th highest central gene (*COL6A2*) in the disease module in the extracellular matrix organization pathway (Table 9, Figure 20). In the fibroblast network analysis, they were also found to target genes involved in extracellular matrix organization and signaling by tyrosine kinases, including *COL4A4, COL15A1*, and *COL3A1* (Table 12, Table 13). Additionally, the meta-analysis revealed that both ocriplasmin and collagenase clostridium histolyticum were present in all datasets - meta-analysis, biopsies, BAL, and epithelial datasets (Tables 19, 22, 25, and 32). Ocriplasmin targets a variety of genes, COL4A4, COL11A2, *LAMA1, COL2A1, COL6A3, COL4A6, LAMB2, COL6A1, COL27A1, COL4A3, LAMB4, COL24A1, COL1A1, COL6A2, COL3A1, LAMC3, LAMB1, COL4A2, COL1A2, COL4A1, COL5A2, FN1, LAMC1, COL11A1, COL5A1, COL6A6, COL4A5, LAMC2, COL18A1, LAMA4, COL15A1, LAMB3, LAMA3, LAMA2, COL6A5, LAMA5, COL28A1*, and *COL5A3* (Koscielny et al. 2017). Enrichment analysis of these genes with in R indicates that the Extracellular matrix organization Reactome pathway is significantly enriched (adjusted p-value of 3.682671e-57). Other related pathways, such as Non-integrin membrane-ECM interactions, Assembly of collagen fibrils and other multimeric structures, and Collagen chain trimerization, are also highly enriched (adjusted p-values less than 1e-50).

Collagenase clostridium histolyticum targets genes *COL11A2, COL6A2, COL15A1, COL2A1, COL6A1, COL4A5, COL1A2, COL6A5, COL4A1, COL24A1, COL5A3, COL3A1, COL27A1, COL11A1, COL4A3, COL18A1, COL4A6, COL1A1, COL4A4, COL6A6, COL5A1, COL6A3, COL28A1, COL4A2,* and *COL5A2* (Koscielny et al. 2017). The genes targeted by collagenase clostridium histolyticum are very similar to the genes targeted by ocriplasmin, except that collagenase clostridium histolyticum only targets collagen genes, while ocriplasmin also targets laminin and fibronectin genes such as *LAMB2*, *LAMA1A* and fibronectin genes (*FN1*). Enrichment analysis of the collagenase clostridium histolyticum genes revealed that the most enriched Reactome pathway is collagen chain trimerization (adjusted p-value of 1.903166e-62), while the Extracellular matrix organization pathway is also enriched (adjusted p-value of 4.796064e-39). The enrichment analysis of the ocriplasmin and collagenase clostridium histolyticum genes reveals similar Reactome pathways, albeit with a different order. It can be inferred that both enzymes affect collagen assembly and extracellular matrix organization, with ocriplasmin having a broader-spectrum effect compared to collagenase clostridium histolyticum. The molecular structure of collagenase clostridium histolyticum is represented in Figure 39. Structure of ocriplasmin is not available.

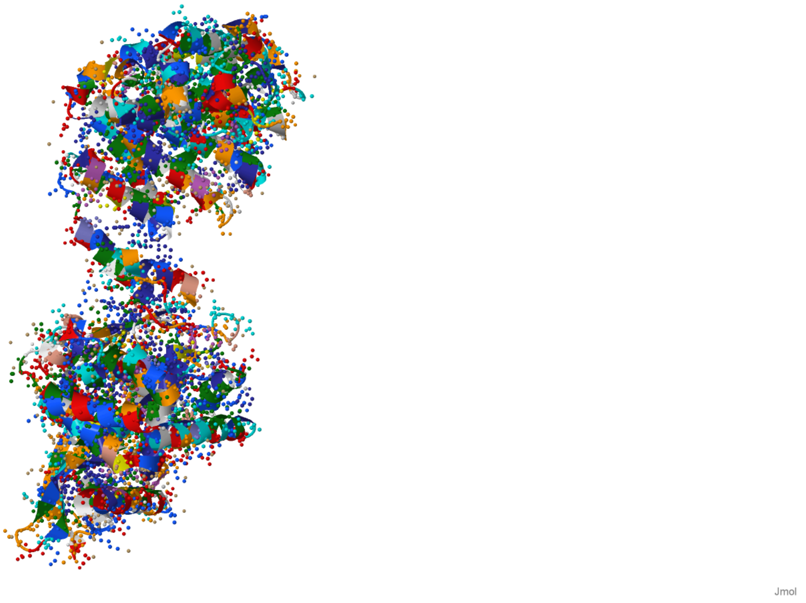

Figure 39: Molecular structure of collagenase clostridium histolyticum. Structure of ocriplasmin is not available.

Ocriplasmin is administered through intravitreal injection and is indicated for the treatment of vitreomacular traction and closed macular holes. It has been demonstrated that ocriplasmin effectively resolves vitreomacular traction, and is more effective than placebo (Stalmans et al. 2012). Vitreomacular traction syndrome is a relatively rare condition that occurs when there is incomplete separation of the posterior vitreous from the macula, leading to persistent attachment. This condition has been found to be associated with several other macular disorders (Shao and Wei 2014). Ocriplasmin has proteolytic activity against the vitreous body and the vitreoretinal interface (VRI), such as laminin, fibronectin, and collagen. This proteolytic activity leads to the dissolution of the protein matrix that causes vitreomacular adhesion (VMA). The adverse effects of ocriplasmin may include decreased Vision, intravitreal Injection Procedure Associated Effects, potential for lens subluxation, retinal breaks, and dyschromatopsia (Novartis Pharmaceuticals Corporation 2016).

Collagenases are enzymes that can break down collagen molecules in their natural triple helical structure under normal physiological conditions. Common adverse effects of collagenase clostridium histolyticum include peripheral edema (swelling at the injection site), contusion, injection site hemorrhage, injection site reaction, and pain in the injected area (BioSpecifics Technologies Corporation 2022). Collagenase clostridium histolyticum has been used for Dupuytren's disease. Dupuytren's disease is a prevalent hand disorder that primarily affects the palmar fascia. The condition initially presents as either skin thickening or pitting on the palm (Karbowiak et al. 2016). Other diseases that collagenase clostridium histolyticum has been used are diabetic foot, decubitus ulcer, skin ulcer, skin wound, abnormality of connective tissue, frozen shoulder, contracture, lipoma, and tendinopathy (Koscielny et al. 2017). Collagenase clostridium histolyticum is administrated as an injection (Hurst et al. 2009; BioSpecifics Technologies Corporation 2022).

Some preclinical studies have investigated the use of collagenase nanocapsules as a potential approach for treating fibrosis (Villegas et al. 2018). The preclinical results appear to be encouraging, suggesting that this approach could be a viable option for fibrosis treatment in the future. However, there is no literature indicating that ocriplasmin has been studied as a treatment for pulmonary fibrosis. Comparing the gene targets of ocriplasmin and collagenase clostridium histolyticum, it seems that ocriplasmin might be even better option for the IPF treatment when looking just at the target genes. Remodeling of the extracellular matrix is a common feature in lung diseases such as IPF (Åhrman et al. 2018; Qian et al. 2019). In ocriplasmin genes the extracellular matrix organization reactome pathway was more significant compared to collagenase clostridium histolyticum genes since ocriplasmin genes contain also laminin and fibronectin genes in addition to collagen genes (Koscielny et al. 2017).

Administering enzymes into the lung can be challenging, and it is important to limit their administration to the fibrotic foci to prevent any unnecessary adverse effects on healthy tissue. One potential approach for treating fibrosis is the use of collagenase nanocapsules, as discussed earlier (Villegas et al. 2018). Ocriplasmin may also be able to be delivered in these nanocapsules, but further studies are necessary. Another option, or option combined with the nanocapsule formulation, is intrapulmonary injection aided by imaging techniques. CT-guided (computed tomography) intrapulmonary injection has been studied extensively, and it has been used to safely and effectively dye pulmonary nodules (Ko et al. 2019; Wicky et al. 1994). In a study of fifteen patients with active inoperable pulmonary aspergilloma, a therapeutic paste of glycerin and amphotericin B was percutaneously injected under CT guidance (Giron et al. 1993). Another study involving CT-guided cyanoacrylate injections for 113 patients with pulmonary lesions reported a success rate of 100 % with no severe complications (Huang et al. 2019). However, intrapulmonary injection carries the risk of adverse effects such as pneumothorax, bloody sputum, intravascular air, pneumonia, and cerebral infarction (Ito et al. 2020). Both ocriplasmin and collagenase clostridium histolyticum show promising potential for treating IPF. The potential drug administration techniques, as well as the safety and efficacy of the treatments, require further investigation.

Tyrosine kinase inhibitors

Tyrosine kinase signaling is essential in various cellular processes, and extensive evidence from both in vitro studies and animal models suggests that certain tyrosine kinases play a crucial role in the development of pulmonary fibrosis. Tyrosine kinases can be classified into two categories: receptor tyrosine kinases (RTKs) and non-receptor cytoplasmic tyrosine kinases (non-RTKs). RTKs are cell membrane receptors that initiate intracellular signaling pathways by binding to growth factors on their extracellular domains. In contrast, non-RTKs do not have extracellular or transmembrane domains and instead regulate signaling pathways within the cytoplasm. Several RTKs such as *PDGF, FGF, VEGF*, and *EGF* are believed to contribute to the development of pulmonary fibrosis. In addition, the SRC family of non-RTKs, which includes *FYN, YES, FGR, LYN, HCK, LCK*, and *BLK*, have been found to be necessary for the epithelial-mesenchymal transition following TGF-β1 signaling in alveolar epithelial cells (Grimminger et al. 2015).

The process of protein phosphorylation, which involves the addition of phosphate groups by kinase enzymes, is a critical mechanism of signal transduction in eukaryotic cells. Protein kinases regulate various cellular processes, such as cell proliferation, cell cycle progression, metabolic homeostasis, transcriptional activation, differentiation and development, and apoptosis. The human kinome, which refers to the entire set of protein kinases encoded by the genome, comprises 90 protein tyrosine kinases. Certain RTKs have the ability to transactivate one another, such as the PDGF receptor (PDGFR) being able to transactivate the epidermal growth factor receptor (EGFR). Transactivation of RTKs has been strongly associated with inflammation and the process of tissue healing. Similar to RTKs, the activation of non-RTKs also involves phosphorylation and autophosphorylation. The effects of tyrosine kinase activation on cells are complex and are dependent on various factors such as the cell type and the specific signal transduction pathway that is triggered (Grimminger et al. 2015).

In this study, various drugs targeted several tyrosine kinases. Among them, *DDR2* (Discoidin Domain Receptor Tyrosine Kinase 2) was identified as the third most central gene in the biopsy disease module (Table 9, Figure 20). Regorafenib, a tyrosine kinase inhibitor, targets *DDR2* and was also found to target *FRK* (fyn related Src family tyrosine kinase) in epithelial network analysis (Table 16, Figure 38). In turn, dasatinib targets *FYN* (FYN proto-oncogene, Src family tyrosine kinase), which was the most central gene in module 3 (second most dissimilar module to healthy network) in fibroblast network analysis (Table 13, Figure 28). Other tyrosine kinase inhibitors were also identified in the study, such as ilorasertib, seliciclib, and nintedanib (Table 14, Table 16, Table 32). According to DisGeNET, the disease specificity index for *FYN* is 0.556, while that for *DDR2* is 0.541. The disease specificity index is a metric that ranges between 0 and 1 and measures the degree to which a gene is associated with a specific disease (Janet et al. 2019).

Previous studies have reported on the interaction between FYN-kinase and caveolin-1 in the alveolar epithelium of various bleomycin (BLM)/TGF-β damage models. In wild type mouse lung tissues, strong signals for FYN-kinase were found in alveolar epithelial type I cells, whereas in caveolin-1 knock out animals, expression was observed to shift to alveolar epithelial type II cells. FYN-kinase has been identified to play a profibrotic role by phosphorylating the TβRII, thereby activating the TGF-β signaling pathway and initiating profibrotic processes such as the epithelial-to-mesenchymal transition. (Menzel et al. 2022).

Fibrillar collagen I is a critical component of the lung extracellular matrix (ECM) that exists in both normal and fibrotic lungs. Besides providing structural support, collagen I facilitates cellular signaling by interacting with ECM receptors. Integrins like integrin $\alpha 2\beta 1$, and discoidin domain receptors (DDRs) 1 and 2 are among the collagen I signaling receptors in the lung. DDR1 and DDR2 proteins are receptor tyrosine kinases that bind and get activated by collagens. DDRs are widely expressed, but their expression is cell-type specific; *DDR1* is mainly expressed on epithelial cells, whereas *DDR2* is present on fibroblasts and cells of mesenchymal origin. The evidence indicates that disease progression may be influenced by the upregulation and activation of these receptors, which occurs in response to tissue damage or as the disease advances. Studies on *DDR1* knockout mice propose that this receptor promotes fibrosis and inflammation in kidneys and lungs by modulating inflammatory responses and ECM synthesis/deposition. The role of *DDR2* in organ fibrosis is still uncertain and controversial (Borza et al. 2018).

While *DDR2*-null mice showed increased liver fibrosis following chronic liver injury, *DDR2* deficiency or downregulation reduced bleomycin-induced lung fibrosis, and deleting *DDR2* in cardiac fibroblasts decreased angiotensin-induced collagen I expression (George et al. 2016; Zhao et al. 2016). These studies indicate that DDR2-mediated functions are cell-type dependent. In this context, Zhao et al. (Zhao et al. 2016) demonstrated that blocking DDR2 kinase activity or reducing *DDR2* expression protected against bleomycin-induced lung fibrosis. Although the authors used dasatinib, which is not a selective *DDR2* inhibitor, more selective *DDR2* inhibitors have been reported. Nevertheless, these inhibitors still affect *DDR1* due to the high homology in the kinase domain between the two receptor tyrosine kinases (Borza et al. 2018; Grither and Longmore 2018).

Regorafenib is an orally active, multi-kinase inhibitor used to treat colorectal cancer and gastrointestinal stromal tumors. Its molecular structure and targets has significant similarities to nintedanib. The molecular structures of regorafenib and nintedanib are illustrated in Figure 40. In vivo experiments have shown that regorafenib can suppress collagen accumulation and myofibroblast activation by inhibiting the TGF-β1/Smad and non-Smad signaling pathways, thereby reducing extracellular matrix production and myofibroblast migration. Additionally, regorafenib has been suggested to promote apoptosis in myofibroblasts and augment autophagy by suppressing TGF-β1/mTOR signaling (mechanistic target of

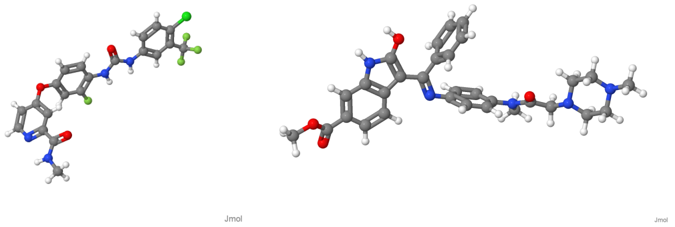

Figure 40: Regorafenib (left) and nintedanib (right) molecular structures.

rapamycin) (Li et al. 2021). A three-month supply of regorafenib (Stivarga®), which includes 3x28 tablets, is priced at 2782.84 € in Finland, as of April 15th, 2023 (Kansaneläkelaitos 2023).

Matrix metalloproteinase inhibitors

Matrix metalloproteinases (MMPs), also known as matrixins, are a family of 23 known zinc-dependent proteases that function in the extracellular environment of cells to degrade both matrix and non-matrix proteins. They play important roles in morphogenesis, wound healing, tissue repair, and remodeling in response to injury, such as after myocardial infarction, and in the progression of diseases such as atheroma, arthritis, cancer, and chronic tissue ulcers. Their activities are regulated by tissue inhibitors of metalloproteinases (TIMPs) (Nagase et al. 2006). In the pathogenesis of fibrosis, MMPs play a crucial role in regulating ECM turnover, chemokine metabolism, cell migration, and mediator activation (Mahalanobish et al. 2020).

In patients with IPF, the expression of MMPs is disrupted, leading to significant architectural remodeling in the lung microenvironment. MMP1, which breaks down fibrillar collagens, is found to be highly expressed in IPF lung tissue and BAL fluid, despite being expressed at low levels in healthy tissue (Mahalanobish et al. 2020). Various MMPs were observed to be highly expressed or central in the analysis of this study as seen in the results (Table 9, Table 10, Table 19, Table 20, Table 22, Table 30). MMPs not only play a role in ECM turnover, but also have multiple components that affect abnormal tissue repair and the behavior of epithelial and mesenchymal cells. The uncoordinated regulation and expression of multiple MMPs can contribute to severe architectural remodeling in the lungs of IPF patients (Hambly et al. 2015).

In the biopsy disease module, *MMP2* was ranked as the 29th most central gene, whereas in the meta-analysis that considered all datasets, it was ranked 69th (as shown in Table 9, Figure 20, and Table 19). *MMP9* was found to be present in both the BAL network analysis and the meta-analysis (Table 10 and Table 25). *MMP2* and *MMP9*, which are classified as gelatinases, possess three fibronectin type II domain repeats in the catalytic domain that enables their interaction with gelatin substrates. These domains are also crucial for MMP9 to degrade types V and XI collagens. Moreover, these MMPs have the ability to activate several other MMPs. (Mahalanobish et al. 2020).

The drugs that inhibit matrix metalloproteinases that were found in this analysis are marimastat, doxycycline and andecaliximab (Table 32). The genes that marimastat targets are: *MMP7, MMP2, MMP3, MMP1, MMP12*, and *MMP9*. The doxycyxline target genes are *MMP8, MMP1, MMP13*, and *MMP7*. Andecaliximab is a specific MMP9 inhibitor (Koscielny et al. 2017). The highest DisGenNET disease specificity index score for these MMPs is for *MMP12* being 0.484 (Janet et al. 2019).
Marimastat was the first actual MMP inhibitor to be tested in clinical trials and was found to be effective in inhibiting tumor progression in mice with cancer. Following this, it was tested in humans with various types of cancer, including pancreatic, lung, breast, colorectal, and gastric adenocarcinoma, and was found to have a favorable pharmacokinetic profile when administered orally (Evans et al. 2001; Pijet et al. 2020). Short-term treatment with marimastat was generally well-tolerated by patients. However, long-term or chronic use of the drug was associated with side effects related to musculoskeletal toxicity that ultimately made it unsuitable for use in cancer treatment (Pijet et al. 2020; Sparano et al. 2004). Molecular structure of marimastat is illustrated in Figure 41.

However, a study in mice investigated the role of MMPs in the serum-borne bioactivity and endothelial cell dysfunction induced by multiwalled carbon nanotubes (MWCNT), using marimastat. The study found that MMP inhibition did not reduce the severity of pulmonary inflammation, indicating that the systemic circulatory and vascular effects of MWCNT may be secondary due to macrophage activation and pulmonary influx of polymorphonuclear neutrophils (Young et al. 2021). On the other hand another study suggests that *MMP7* knockout mice resist bleomycin-induced fibrosis due at least in part to the decrease in sFasL levels in their blood (Nareznoi et al. 2020). Also Todd et al. suggest that certain MMPs or TIMPs could serve as biomarkers for IPF in patients. However, further research is needed to confirm these results, as well as to investigate the relationship between MMP/TIMP expression and clinical outcomes (Todd et al. 2020).

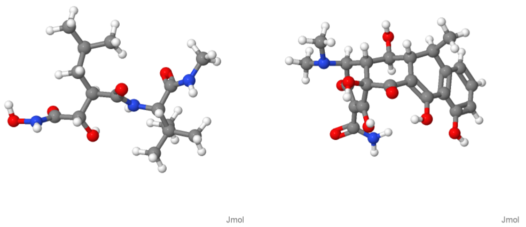

Figure 41: Marimastat (left) and doxycycline (right) molecular structures.

Doxycycline is a tetracycline antibiotic which possesses potent antimicrobial activity, has excellent tissue penetration, and exhibits high oral bioavailability. By binding to the decoding center of the small ribosomal subunit of the bacteria, it obstructs protein synthesis (Zhang et al. 2019). Pfizer Inc. of New York, NY developed and clinically tested Doxycycline in the early 1960s, and it was subsequently marketed under the brand name Vibramycin®. In 1967, the FDA approved Vibramycin®, making it Pfizer's first broad-spectrum antibiotic to be taken once a day (Tan et al. 2011). Molecular structure of doxycycline is illustrated in Figure 41.

Doxycycline has been shown to attenuate bleomycin-induced pulmonary fibrosis, as well as the production of TGF-β1-induced mediators that contribute to its progression in alveolar epithelial cells in vitro. According to Fujita H.'s study, doxycycline inhibited the TGF-β1-induced mRNA expression of *PDGFA, CTGF, MMP2,* and *MMP9*, as well as the protein production of *PGDF-AA* and *MMP2* in alveolar epithelial cells in vitro. Doxycycline affected the production of MMP2 and MMP9, which are involved in cell proliferation, adhesion, migration, and differentiation through proteolytic effects on ECM components and basement membranes that contribute to fibrosis progression. *MMP2* and *MMP9* gene and protein expression have been reported to be elevated in tissues and bronchoalveolar lavage fluid from patients with IPF. Therefore, doxycycline appears to attenuate pulmonary fibrosis by inhibiting growth factors and MMP production in alveolar epithelial cells, as demonstrated by Fujita H. et al. (Fujita et al. 2011). Similar results were also found in a study by Amartya, M. et al. (Amartya et al. 2011). As of April 15th, 2023, 100 tablets of doxycycline cost 46.32 € in Finland (Kansaneläkelaitos 2023).

Andecaliximab is a monoclonal antibody that inhibits *MMP9*, a protein involved in matrix remodeling, tumor growth, and metastasis. A phase I and Ib study of modified oxaliplatin, leucovorin, and fluorouracil with andecaliximab showed promising antitumor activity in patients with gastric or gastroesophageal junction adenocarcinoma (Shah et al. 2021). In a study by Espindola et al., it was demonstrated that *MMP9* expression was increased in airway basal cell-like cells from lungs of patients with IPF compared to those from normal lungs. Blocking *MMP9* activity with an anti-MMP9 antibody, such as andecaliximab, inhibited TGF-β1-induced Smad2 phosphorylation. However, in a subset of cells from patients with IPF, TGF-β1 activation in their airway basal cell-like cells was unaffected or even enhanced by *MMP9* blockade, indicating a lack of response to the treatment (i.e., non-responders) (Espindola et al. 2021).

Ion channel modulators and inhibitors

Ion channel modulators and inhibitors were identified through network analysis of macrophage datasets (Table 16 and Figure 32). In the macrophage network analysis (Table 16 and Figure 32), *GUCY1A2*, which was the 18th central gene in the disease module, is targeted by various nitrates, such as isosorbide mononitrate and nitroglycerin. Also, a meta-analysis of all datasets showed that *KCNN4* (calcium-activated channel subfamily N member 4 gene) was ranked 8th in this analysis. Additionally, the *SCNN1A* gene, which encodes the alpha subunit of the sodium channel epithelial 1, was ranked 20th in the BAL meta-analysis and is the target of two hypertension medications, namely amiloride and triamterene, both of which act as sodium channel blockers (Koscielny et al. 2017; Nesterov et al. 2012).

Nitrates exert their pharmacological effects by releasing nitric oxide (NO), which is an endothelium-derived relaxing factor (EDRF). Endogenous NO is produced in the endothelium to facilitate blood vessel dilation. Nitrates activate the enzyme soluble guanylate cyclase in vascular smooth muscles, leading to increased levels of intracellular cGMP and associated protein kinases, such as cGMP-dependent protein kinases (cGK-I). The cGMP then activates the myosin light chain phosphatase (MLCP), which causes dephosphorylation of the myosin light chain. Moreover, cGMP-cGK-I inhibits the inositol-1,4,5-trisphosphate (IP3)-dependent calcium release, leading to decreased intracellular calcium levels. Ultimately, the decrease in intracellular calcium levels results in smooth muscle relaxation, leading to vasodilation (Balasubramanian and Chowdhury 2022).

Relatively recent studies have demonstrated that in animal models of various fibrotic diseases, the use of soluble guanylate cyclase (sGC) stimulators has resulted in a reduction of fibrotic events (Beyer et al. 2012; Lambers et al. 2019). Also, many preclinical studies have demonstrated anti-fibrotic efficacy of sGC-cGMP activation in various experimental fibrosis models but the molecular basis of the efficacy in these models are not well understood (Kim et al. 2020). One possible mechanism is due to the inhibition of TGF-β1 induced *ERK1/2* signalling in human lung fibroblasts, leading to reduced de novo synthesis of collagen type I. The findings of Lambers et al. suggest that the sGC activator BAY 41-2272 represents a promising therapeutic option for treating IPF. In addition, combining an sGC activator with a cAMP activator such as forskolin may enhance the antifibrotic potential (Lambers et al. 2019).

A study by Blanco et al., which was conducted on a selected group of patients with IPF, suggests that signaling molecules produced by the endothelium may play a role in regulating pulmonary vascular tone during exercise. The study also found that inhaled nitric oxide (NO) can reduce pulmonary vascular resistance both at rest and during exercise, without affecting gas exchange. In IPF patients, during exercise, the limitation of alveolar-to-capillary oxygen diffusion becomes a significant factor contributing to the decrease in oxygen partial pressure (Blanco et al. 2011). For over a century, nitrates have been used as a vasodilator to treat pulmonary hypertension. However, exogenous NO donors have limitations due to increased oxidative stress and tolerance. As a result, current treatment strategies aim to inhibit the degradation of cyclic guanosine monophosphate (cGMP) by targeting phosphodiesterases, specifically PDE5 using medications such as sildenafil. Another approach is to increase the enzymatic activity of soluble guanylate cyclase, which mainly involves the use of sGC activators and stimulators.

These drugs enhance the activity of both oxidized and reduced forms of the sGC enzyme, respectively (Kim et al. 2020).

In a network analysis of macrophages, *CACNG3, CACNG4,* and *CACNA2D1* were identified as potential drug targets (Table 16 and Figure 32). Additionally, a meta-analysis of all datasets showed *KCNN4* (calcium-activated channel subfamily N member 4 gene) was ranked 8th in this analysis. Mitochondrial oxidative stress and turnover in alveolar macrophages are directly linked to pulmonary fibrosis, but the exact molecular mechanisms regulating mitochondrial dynamics remain unknown. The mitochondrial calcium uniporter (MCU) is a selective ion channel that transports Ca2+ into the mitochondrial matrix, affecting cellular metabolism. After exposure to asbestos, the MCU has been shown to polarize macrophages towards a profibrotic phenotype by regulating ATP production (Zhang et al. 2018).

It has also been suggested that there is a central role for Ca2+ in pro-fibrotic fibroblast function and fibrosis-related diseases. Growth factors evoke rhythmic Ca2+ oscillations through a complex interplay between Ca2+ release and entry mechanisms in fibroblasts resulting in pro-fibrotic activities. Therefore, therapeutic strategies that target Ca2+ homeostatic mechanisms or Ca2+-mediated signalling events could prove successful. There is evidence, consistent across many organ systems, demonstrates a fundamental role for *KCNN4* activation in the generation of pathological fibrosis due to its ability to regulate Ca2+ influx. This is not surprising as ion channels regulate all cellular processes and many effective pharmacological agents in use today work through the modulation of ion channels (Roach and Bradding 2019).

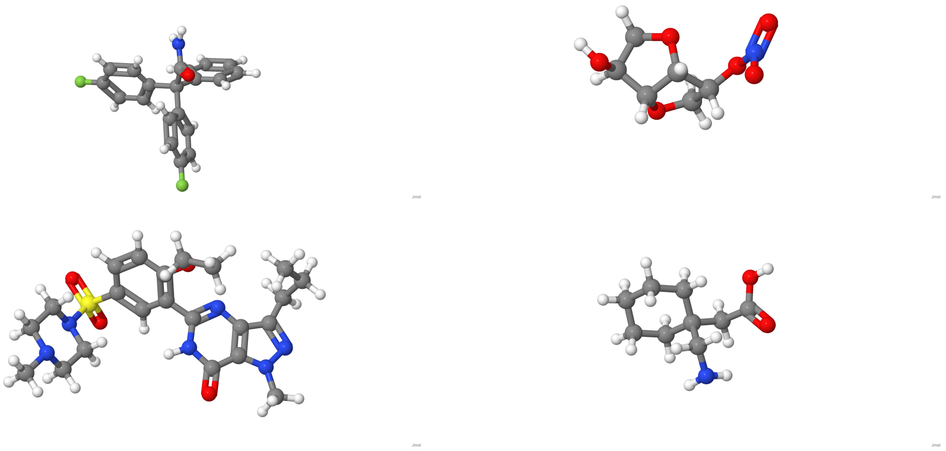

Figure 42: Senicapoc (top left), isosorbide mononitrate (top right), sildenafil (bottom left) and gabapentin (bottom right) molecular structures.

In a study by Roach and Bradding, the *KCNN4* selective blocker senicapoc was found to be well-tolerated in humans for up to 12 months. Additionally, *KCNN4* inhibition in human airway epithelial cells appeared to attenuate the development of TGF-β1-dependent epithelial-mesenchymal transition, suggesting that *KCNN4* may play a role in promoting fibrosis. Furthermore, in the same study, senicapoc treatment in sheep reduced non-specific airway hyperresponsiveness to carbachol following allergen challenge and also reduced eosinophil numbers in BAL samples collected 48 hours post-allergen challenge.

Network analysis of macrophages showed that gabapentin, a neuromodulator used for chronic pain and as an antispasmodic agent among other indications, targeted the gene *CACNG3* and was ranked 51st most central in the disease module. Similarly, pregabalin, with indications quite similar to gabapentin, targets the calcium channel *CACNA2D1*, which was ranked 264th most central in the disease module. Notably, gabapentin has been found to be particularly effective in treating cough in patients with IPF, and there are several randomized controlled trials that support the use of neuromodulator drugs, such as gabapentin and pregabalin, for unexplained chronic cough (Birring et al. 2018; Ryan et al. 2012; Vertiganm et al. 2016). Moreover, it is believed that gabapentin and pregabalin possess anti-inflammatory properties by altering the response of the neurokinin-1 receptor, which is a substance P receptor. It has been showed that gabapentin and pregabalin were able to reduce the production of Il-6 and Il-8 in glioblastoma cells when induced by substance P (Yamaguchi et al. 2017).

According to the latest available information, senicapoc is currently in a trial phase and not yet widely used (Ali et al. 2020). On the other hand, drugs like sildenafil, nitrates, pregabalin, and gabapentin, which have also been discussed as ion channel modulators, are already widely used and are relatively inexpensive (Kansaneläkelaitos 2023). Molecular stuctures of senicapoc, isosorbide mononitrate, sildenafil and gabapentin are illustrated in Figure 42.

Connection between IPF, cancer and autoimmune diseases and monoclonal antibody therapies

The altered process of "wound healing" in IPF is driven by various pathogenic events that are commonly observed in other degenerative/fibrotic diseases and even in cancer. Cancer has been described by some authors as a "wound that does not heal" (Dvorak 1986; Vancheri 2015). With an often unknown etiology, cancer shares some of the risk factors that are associated with IPF, and the presence of a specific genetic background is considered important for the occurrence of both diseases. Like cancer, IPF affects susceptible individuals and shares common risk factors, such as smoking, environmental or professional exposure, viral infections, and chronic tissue injury (Vancheri 2015). Based on these similarities, poor response to medical treatment, and prognosis, which is often worse than many cancers, IPF has frequently, although vaguely, been compared to a type of malignant disease (Vancheri et al. 2010; Vancheri 2015).

Tissue homeostasis is maintained through cell-to-cell communication facilitated by connexins, junctional channels that coordinate cell functions. In contrast, cancer cells often exhibit impaired intercellular communication, indicating the need for isolation from surrounding normal cells to proliferate. The decreased expression of connexin 43, a connexin found in many cancers including lung and gastric cancer, has been associated with cancer cell proliferation. Similarly, IPF myofibroblasts have also been shown to have a reduced ability to express connexin 43, suggesting that the loss of proliferative control in IPF cells may be due to altered fibroblast-to-fibroblast communication caused by decreased connexin 43 expression (Mori et al. 2006; Vancheri 2015). The origin of myofibroblasts is similar in both IPF and cancer. Various signal transduction pathways are involved in the pathogenesis of cancer and IPF, including the Wnt/β-catenin signaling pathway. This pathway regulates the expression of molecules involved in cancer progression and tissue infiltration and is strongly activated in IPF lung tissue, as evidenced by extensive nuclear accumulation of β-catenin in various involved sites, such as bronchiolar lesions, damaged alveolar structures, and fibroblast foci (Mazières et al. 2005; Vancheri 2015). The PI3K/AKT signaling pathway, which regulates cell growth, proliferation, and cell protection from apoptosis, is also strongly involved in the pathogenesis of cancer and IPF (Vancheri 2015).

In addtion to cancer, IPF and autoimmune diseases share some common pathogenic mechanisms at the biochemical level, such as the activation of inflammatory cells, immune dysregulation, oxidative stress, and abnormal wound healing leading to excessive extracellular matrix deposition, particularly collagen. However, there are also notable differences, such as the specific immune cells involved in the disease process, the type of inflammation, and the distribution of fibrosis within the lung tissue. For instance, in autoimmune diseases, immune cells such as T cells and B cells play a key role in the initiation and progression of the disease. On the other hand, in IPF, immune cells such as macrophages and fibroblasts play a more prominent role in the disease process. Another difference is that in autoimmune diseases, the inflammation is often organized into granulomas or lymphoid aggregates, whereas in IPF, the inflammation is more diffuse and can be found throughout the lung tissue. Biochemical markers such as cytokines, chemokines, and growth factors have also been implicated in both IPF and autoimmune diseases. For example, TGF-β has been found to be elevated in both IPF and systemic sclerosis, while Il-6 has been associated with IPF and rheumatoid arthritis (Popper et al. 2022).

Monoclonal antibodies (mAbs) have become an essential component of pharmacotherapy for a wide range of medical conditions over the past two decades, particularly for autoimmune disorders and cancers in addition to other indications. As of 2017, more than 40 mAbs have been authorized by the US Food and Drug Administration, with several dozen more in clinical development. Structurally, mAbs are similar to IgG, as they are large heterodimeric protein molecules with a molecular weight of approximately 150 kDa. Due to their production process in living cells, mAbs are defined by their production process rather than their chemical structure, and the batch-to-batch variability in the resulting product needs to be tightly controlled through carefully established and controlled conditions during the cell culturing, product processing, and purification steps. Monoclonal antibodies are a unique class of therapeutics with pharmacokinetics determined and controlled by the specific mechanisms and processes involved in their disposition. Although there are substantial differences in the pharmacokinetics of individual mAbs, their general behavior can be considered a class property as it is driven by and similar to their endogenous counterpart IgG (Ryman and Meibohm 2017). A molecular structure of a monoclonal antibody infliximab is illustrated in Figure 43.

Figure 43: An example of a structure of a monoclonal antibody infliximab.

Monoclonal antibody -based orphan drugs have gained popularity over the past decade for treating especially cancers and autoimmune diseases due to their ability to selectively target specific molecules and regulate signaling pathways, leading to improved therapeutic outcomes. However, the expensive production and acquisition costs of these drugs impose a significant financial burden on both patients and society. Therefore, evaluating their cost-effectiveness is crucial to determine if their clinical benefits justify their high costs. According to Park et al., a study of nine mAbs revealed that four orphan drugs (cetuximab, ipilimumab, rituximab, and trastuzumab) were cost-effective, while one drug (bevacizumab) was not cost-effective in cost-utility analysis studies. The cost-effectiveness results of infliximab were inconsistent across the studies (Park et al. 2015). One potential solution to the high cost of mAbs is the use of biosimilars, which could provide some relief for the societal burden of the expensive medicines (Kvien et al. 2022).

Monoclonal antibodies and other biologics were found accross the analysis in different cell types and in both network- and meta-analysis. Fresolimumab (Table 9 and Table 12) was found in the network based analysis in biopsy and fibroblast samples. In the biopsy analysis fresolimumab is targeting TGFβ3 which is the 285th central gene in the disease module. In fibroblasts fresolimumab appeared in the secondmost dissimilar module between disease and healthy samples when extracting the extracellular matrix organization related genes. It targeted the same gene TGFβ3 which was 794th central gene in that particular module.

Fresolimumab (GC1008) is a recombinant human monoclonal antibody that inhibits all three isoforms of transforming growth factor-beta (TGFβ), which plays a central role in the pathogenesis of IPF. TGFβ1 is the most implicated isoform in perpetuating the fibrotic process in IPF. TGF-β is produced by various cell types in the lungs and secreted in an inactive form. Its further activation is mediated by factors such as matrix metalloproteinases, integrins, and reactive oxygen species, which cleave the latent TGFβ1–binding protein complex. Downregulating the TGF-β pathway is believed to be crucial for halting the fibrosing process. In 2005, a phase I open-label, non-randomized, multicenter, single-dose, dose-escalating study was conducted to investigate the safety, tolerability, and pharmacokinetics of fresolimumab in five dose groups of patients with IPF. However, this study was completed without any reported results, and fresolimumab has not been investigated for IPF since then (Sgalla et al. 2020).

Simtuzumab was found in network analysis of biopsies where it targets *LOXL2* gene, which is ranked 43rd most central gene in the particular analysis. Simtuzumab is a type of human monoclonal antibody that targets *LOXL2*, a group of enzymes that help stabilize the ECM by facilitating the cross-linking of collagen molecules. In fibrotic diseases, the increased cross-linking of matrix proteins can lead to pathologically increased matrix stiffness. Elevated levels of *LOXL2* contribute to myofibroblast differentiation and matrix production, further driving fibrosis progression. *LOXL2* is highly expressed in fibrotic regions of IPF lung, and serum levels are increased in patients with progressive IPF. Inhibiting *LOXL2* can interrupt the vicious loop in which a stiff matrix delivers signals to fibroblasts, which further stiffen the matrix via producing more collagen. By inhibiting *LOXL2*, there is a reduction in activated fibroblasts, decreased production of growth factors, downregulation of TGF-β, and its pathway signaling in human fibroblasts. A clinical trial (NCT01769196) investigating the use of simtuzumab in patients with IPF failed to show any benefit over placebo for the primary and secondary endpoints (Sgalla et al. 2020).

STX-100 was observed in various parts of the study analysis (Tables 12, 13, and 32). It targets the *ITGAV* gene in the fibroblasts network analysis in the extracellular matrix and tyrosine kinase signaling pathways. The gene was ranked 195th in the module. Additionally, STX-100 was found in the meta-analysis of epithelial cells where it targeted *ITGB6*, which was ranked 3954th in that particular analysis.

STX-100 is a first-of-its-kind humanized anti-$\alpha_v\beta_6$ IgG1 monoclonal antibody that hinders the binding of $\alpha_v\beta_6$ to the latent form of TGFβ, thus restraining TGF-β activation. This small phase-IIa study found that once-weekly subcutaneous administration of STX-100 at doses lower than 1.0 mg/kg was generally well-tolerated by IPF patients. However, acute IPF exacerbation was observed among patients receiving higher doses of STX-100. Furthermore, the study met the predefined stopping criteria at the 3.0 mg/kg dose, indicating a potential risk of respiratory status decline with higher doses of STX-100. As the study had a small number of patients and an exploratory nature, the results should be interpreted with caution. Larger studies are necessary to verify the potential dose-response effect of STX-100 on TGF-β suppression and to define clinical effectiveness (Raghu et al. 2022).

Over the past decade, numerous monoclonal antibodies targeting different molecular targets have been investigated in the treatment of IPF, often concurrently with the approval of nintedanib and pirfenidone. However, many randomized controlled trials with robust rationales for targeting key fibrotic pathways, such as the $\alpha_v\beta_6$ integrin and *LOXL2*, have yielded disappointing results. Although the Il-13 and Il-4 blockade pathway has been extensively investigated, consistently negative results suggest this therapeutic strategy may not be effective for IPF. Recent phase 2 study of the anti-CTGF antibody pamrevlumab showed promising results, sparking hope for new effective therapies. However, due to the complex and multifactorial pathogenesis of IPF, targeting a single pathway may not be the ideal approach, and combination therapy may prove to be more efficient (Sgalla et al. 2020; Wells 2015) Additionally, the lack of a priori cohort enrichment with patients likely to respond to experimental therapies, along with the exclusion of patients receiving background therapy with nintedanib or pirfenidone, have limited the success of many randomized controlled trials. Therefore, more robust mechanistic evidence and reliable predictive biomarkers are needed for the development of personalized therapeutic options for IPF (Sgalla et al. 2020).

Combination therapies and personalized medicine for the treatment of IPF

Due to the involvement of multiple coactivated pathways in the pathogenesis of IPF, it is unlikely that targeted therapies will be effective in isolation. Instead, the evolution towards combination therapy has become the norm in other respiratory diseases such as lung cancer, COPD, and asthma. In interstitial lung diseases other than IPF, combination therapy is attractive in principle, as it can address diagnostic uncertainty and target both pro-inflammatory and profibrotic pathways (Wuyts et al. 2014). One of the major challenges ahead is the question of which compounds to combine and how to evaluate combination therapies in clinical trials. The drugs most likely to provide additive efficacy when used in combination with one of the approved therapies are those with alternative, complementary, or synergistic mechanisms of action. Drugs with overlapping adverse event profiles are less likely to make good combination partners (Kolb et al. 2017).

Possible combination regimens for definite or probable IPF could involve pirfenidone in combination with nintedanib, which has been suggested to have a manageable safety profile and potentially higher efficacy than monotherapy (Huh et al. 2021; Wuyts et al. 2014). Other suggested combinations include pirfenidone or nintedanib with a novel antifibrotic therapy, antifibrotic therapy combined with regular antireflux treatment, antifibrotic therapy in combination with microbiome treatment (such as co-trimoxazole or another antibiotic like doxycycline which was found in this study), antifibrotic therapy with targeted therapy for pulmonary hypertension, and immunomodulation with antifibrotic therapy for possible or probable IPF cases (Wuyts et al. 2014).

Based on this data-driven drug repositioning study, potential combination therapies for the treatment of IPF could involve drug classes such as collagenase enzymes like ocriplasmin or collagenase clostridium histolyticum, tyrosine kinase inhibitors like regorafenib or dasatinib, matrix metalloproteinase inhibitors like doxycycline, marimastat or andecaliximab, ion channel modulators like nitrates, PDE5 inhibitors and/or gabapentin or pregabalin, and possibly some monoclonal antibodies. However, it is not advisable to combine drugs with similar mechanisms of action as it may lead to adverse effects (Cascorbi 2012; Kolb et al. 2017). Therefore, it is crucial to ensure that the drugs used in combination therapies work well together. For instance, in vitro studies suggest that tetracycline antibiotics like doxycycline interfere with the degradation of collagen by metalloproteinases (collagenases) by inhibiting their activity. However, this evidence is limited to in vitro studies, and its relevance to humans is unknown. It is important to note that a reduction in collagenase activity may potentially decrease the therapeutic efficacy of collagenase histolyticum (Watt and Hentz 2011).

Combination therapies can effectively use different drug administration routes. Inhalation is an efficient administration route for drugs targeted at the apical side of the lung, such as epithelial cells and bronchoalveolar lavage. An example of an inhaled dry powder galectin-3 inhibitor is TD139 (Galecto Biotech/Bristol-Myers Squibb), which regulates the expression of TGF-β receptors on the surface of alveolar epithelial cells and is a mediator of TGFβ-induced lung fibrosis. Studies on bleomycin-treated galectin-3 knockout mice have shown reduced lung fibrosis and collagen levels. Additionally, epithelial cells and fibroblasts from galectin-3 knockout mice have reduced (myo)fibroblast activation, epithelial mesenchymal transition, and collagen I production in response to TGF-β (Kolb et al. 2017; MacKinnon et al. 2012). Single doses of TD139 in healthy subjects were found to be well-tolerated, with mild adverse events, including headache, cough, and dose-related paraguesia (Kolb et al. 2017). For the basal side of the lung, which includes extracellular matrix, macrophages, and fibroblasts, systematic administration through oral form or injections (such as ocriplasmin and collagenase clostridium histolyticum) can be considered, as discussed earlier.

The use of personalized medicine holds promise for the treatment of many malignant diseases (Hoeben et al. 2021). Personalized medicine is based on the concept that, even in seemingly uniform diseases, differences in treatment response can be predicted from specific biomarkers, which indicate a predominant pathogenic pathway in individual patients. While the idea of selecting monotherapies based on personalized biomarker signals is attractive, it is challenging to implement in the routine management of IPF due to the complex nature of its pathogenesis (Thannickal and Antony 2018; Wuyts et al. 2014). IPF involves the co-activation of multiple pathways, and selective inhibition of a key mediator may lead to the rapid enhancement of alternative pathways (Wuyts et al. 2014). Nonetheless, the identification of biomarkers linked to specific disease endotypes, including epithelial cell dysfunction, impaired host defense, T-cell exhaustion, fibroblast activation, oxidative stress, and senescence/aging, may aid in the development of targeted therapies with greater efficacy and tolerability for IPF patients in the future (Thannickal and Antony 2018).

Personalized medication, combined with the principles of combination therapy, is an intriguing concept for treating IPF, a disease characterized by clinical variability that can be attributed to distinct and overlapping pathobiological mechanisms. Molecular imaging, system pharmacological, and pharmacogenomic approaches could aid in this endeavor. As newer drug targets are identified and biomarkers discovered, personalized medicine could become a reality for IPF patients, resulting in more effective and tolerable targeted therapies. However, the complex nature of IPF's pathogenesis makes personalized medicine challenging in routine management in the short-term future (Thannickal and Antony 2018).

Comparison of meta-analysis and gene co-expression network analysis

Differential gene expression (DGE) analysis is a frequently used method to unveil the altered molecular mechanisms of complex diseases. However, conventional DGE analysis, such as the t-test or rank sum test, tests individual genes independently and overlooks the interactions between them. As a result, the top-ranked differentially regulated genes prioritized by the analysis may not necessarily correspond to the coherent molecular changes that underlie complex diseases. Joint analyses of gene co-expression and DGE have been utilized to identify the disrupted molecular modules that underlie complex diseases (Wu et al. 2013). The analysis of gene co-expression networks is a crucial method in deciphering gene function and association based on genome-wide expression data. This approach enables the detection of co-expression modules comprising highly related genes, as well as modules associated with clinical features, providing valuable insights into the function of co-expressed genes and identifying key genes involved in human diseases (Bai et al. 2020).

The emerging field of network medicine has revolutionized the way human diseases are defined and analyzed. Rather than considering a disease or particular genes as an isolated entity, network medicine recognizes the interplay of multiple molecular processes in causing disease. It proposes a holistic approach, emphasizing the importance of understanding disease complexity at the cellular and molecular levels and studying the relationships between different pathophenotypes. Network medicine aims to identify and characterize potential network modules that can be targeted for clinical intervention to gain a better understanding of how perturbations propagate through the system. Despite being a relatively new field, network medicine has seen rapid growth in scientific research, with numerous methods being developed to investigate disease etiology, model genetic and molecular interactions, identify potential biomarkers, and design therapeutic interventions, including drug discovery and drug repurposing (Barabási et al. 2011; Goh et al. 2007; Paci et al. 2021).

Both the meta-analysis and the network-based approach yielded both similarities and differences in their results. For instance, upon comparing the biopsy dataset network-based analysis and meta-analysis (Tables 9 and 22), it is evident that collagenase clostridium histolyticum, ocriplasmin, and marimastat were present in both analyses. It is worth noting that the reactome pathway from which the genes were extracted differed. The meta-analysis did not identify tyrosine kinase inhibitors in the biopsy analysis (Table 22). Based on speculation without any statistical inference, network-based methods appear to provide more coherent results concerning data-driven drug repositioning. This is is logical since the network approach focuses on genes belonging to the disease module instead of differentially expressed genes from the whole gene pool. On the other hand, the network-based approach identified e.g. vismodegib, which was not present in the meta-analysis. Combining both approaches would be beneficial since they provide different perspectives, and the information can be integrated to see which results support each other.

FAIR data

In order to optimize the benefits of formal digital publishing, data producers and publishers can be guided by the four fundamental principles of FAIR - Findability, Accessibility, Interoperability, and Reusability. Following these principles would be ideal for navigating the challenges of modern scholarly digital publishing (Wilkinson et al. 2016). Saarimäki et al. recently published a paper in which they assessed the quality of datasets used in their study. They found that out of the datasets published in peer-reviewed articles, 35 were excluded due to problems in overall usability, rather than reusability. The issues were related to experimental design, indicating that some toxicogenomics datasets published in peer-reviewed articles had substantial design flaws that could compromise the validity of any results obtained from them. The study emphasizes the importance of critically evaluating data, even when they have been FAIRified. The authors suggest that although rigorous reporting of data is important, it does not automatically ensure quality. Instead, efforts to ensure quality should be addressed in the early phases of experimental design (Saarimäki et al. 2022).

The encountered lack of data quality during the curation process was notable. For instance, considering dataset GSE70866 "BAL cell gene expression is indicative of outcome and airway basal cell involvement in idiopathic pulmonary fibrosis," which has received 80 citations on Scopus as of May 1st, 2023 (Prasse et al. 2019b). The MDS plot of the normalized data for this study is included in supplementary Figure 57, left figure. The data was collected from three different cities, and the only control samples available are from Freiburgh. Unfortunately, there is no batch correction method that could account for such an artificial batch effect, which led me to exclude the Sienna and Leuven samples from the analysis. If these samples had been included, the enormous batch effect would have made nearly all genes differentially expressed. The lack of randomization in this dataset is quite problematic, and unfortunately, it is not the only example. Figure 61 also exhibits a similar lack of randomization. In their recent paper, Saarimäki et al. have also highlighted this issue, indicating that the minimum standards for toxicogenomics data often result in poor usability due to incomplete characterization of the experimental design and execution, as well as the lack of description regarding potential systematic effects caused by reagents, microarrays, and other factors (Saarimäki et al. 2022).

During the data preprocessing phase, another issue was encountered where the metadata files did not contain the sample names for the corresponding data, and the order of the metadata samples was completely unrelated to that of the actual data. This presented a significant challenge, as even with sophisticated data curation processes, it is difficult to reconcile the metadata with the actual data. It can be a very time-consuming and error-prone task to try to determine which sample name corresponds to each sample in the metadata. This issue was prevalent e.g. in dataset GSE199152. Despite efforts to address this issue, it remains a persistent problem when the metadata and actual data are not properly aligned.

In conclusion, the challenges of modern scholarly digital publishing can be addressed by adhering to the four fundamental principles of FAIR, which guide data producers and publishers to optimize the benefits of formal digital publishing. However, even when data are FAIRified, there can be issues related to experimental design and execution that compromise the validity of any results obtained from them. This highlights the importance of critically evaluating data quality and addressing issues in the early phases of experimental design. Furthermore, issues related to metadata alignment and lack of randomization can present significant challenges during data preprocessing, and efforts should be made to mitigate these issues.

Conclusions

In this study, potential biomarkers were identified as novel drug targets, and the drugs discovered were classified into five categories: enzymes, matrix metalloproteinase inhibitors, tyrosine kinase inhibitors, ion channel inhibitors and modulators, and biologics, such as monoclonal antibodies. Of particular interest was the enzyme ocriplasmin, which has the potential to treat IPF by breaking down the extracellular matrix. However, the administration of ocriplasmin in the lungs remains challenging, and further research is needed to determine its safety and efficacy. One possibility for administration is through the use of nanocapsules via oral or parenteral routes. Studies in mice have shown that nanocapsules can deliver collagenase clostridium histolyticum (which targets similar areas as ocriplasmin) in a sustained and controlled manner, protecting the enzyme's activity for up to 10 days (Villegas et al. 2018).

Due to the complex nature of IPF pathogenesis, multiple pathways are coactivated in the disease, making monotherapies less effective in treating IPF. One potential approach for IPF treatment is to target both pro-inflammatory and pro-fibrotic pathways. To achieve this, multiple cell types, including epithelial cells, macrophages, fibroblasts, BAL, and biopsies, have been analyzed. Tyrosine kinase inhibitors (such as regorafenib, dasatinib, and nintedanib) and matrix metalloproteinase inhibitors (such as doxycycline and marimastat) have been found to affect several cellular processes that are crucial for IPF pathogenesis. Many of these compounds have been used in a variety of cancers, but further research is needed to determine their effectiveness in IPF treatment. The goal is to find medications with specific targets that can inhibit IPF-related pathways with an effective dose and as few adverse effects as possible.

In the analysis of macrophages, it was discovered that nitrates and neuromodulators like gabapentin could assist in managing pulmonary hypertension and cough. Additionally, gabapentin and pregabalin have been demonstrated to possess anti-inflammatory qualities (Yamaguchi et al. 2017). While nitrates might have some potential downsides, such as oxidative stress and tolerance, PDE5 inhibitors may provide a better alternative. Nevertheless, these inexpensive and widely used neuromodulators, nitrates, and PDE5 inhibitors are intriguing treatment options that could reduce the cost of IPF treatment for both patients and society. Analyses also revealed many monoclonal antibodies, but clinical studies of numerous promising biologics for IPF treatment have been disappointing. Nonetheless, recent clinical studies have indicated some promise for effective monoclonal antibody treatment in IPF, such as the recent anti-CTGF antibody pamrevlumab.

Recently, methods that combines co-expression and differential gene expression analyses has been utilized to identify functional characteristics that underlie complex diseases. This approach enables the detection of gene co-expression networks and disease modules consisting of highly related genes that are associated with clinical features. This provides valuable insights into gene functions and helps to identify crucial genes involved in human diseases. Network medicine aims to identify and characterize potential network modules that can be targeted for clinical intervention, to gain a better understanding of how perturbations propagate through the holistic system. On the other hand, differential gene expression analyses provide information about individual genes that behave differently in diseased and healthy individuals. Although there are similarities and differences in the results obtained from the meta-analysis and the network-based approach, combining both methods would be advantageous since they offer different perspectives, and the information can be integrated to determine which results support each other.

Ensuring high-quality datasets is crucial for maximizing the benefits of formal digital publishing. Following the FAIR principles (Findability, Accessibility, Interoperability, and Reusability) can assist data producers and publishers in navigating the challenges of contemporary scholarly digital publishing. However, it's important to note that thorough reporting of data doesn't automatically guarantee quality, and measures to ensure quality should be taken during the early phases of experimental design. Even though the FAIR principles are employed, toxicogenomic datasets frequently encounter usability issues due to inadequate documentation of experimental design and execution. Additionally, poor quality and systematic effects caused by the absence of randomization in the study design also contribute to these issues.

References

Åhrman, E. et al. (2018). “Quantitative proteomic characterization of the lung extracellular matrix in chronic obstructive pulmonary disease and idiopathic pulmonary fibrosis”. In: *Journal of Proteomics* 189, pp. 23–33.

Alessio, C. et al. (2021). *esc: Effect Size Computation for Meta Analysis*. R package version 2.3.0. URL: https://CRAN.R-project.org/package=esc.

Ali, M. A. et al. (2020). “Efficacy and safety of recently approved drugs for sickle cell disease: a review of clinical trials”. In: *Clinical and applied thrombosis/hemostasis* 26.7.

Amartya, M. et al. (2011). “An alternative therapy for idiopathic pulmonary fibrosis by doxycycline through matrix metalloproteinase inhibition”. In: *Lung India* 28, pp. 174–179. DOI: 10.4103/0970-2113.83972.

Andrews, Simon (2010). *FastQC: A Quality Control Tool for High Throughput Sequence Data*. Online. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Bai, Q. et al. (Dec. 2020). “Identification of Hub Genes Associated With Development and Microenvironment of Hepatocellular Carcinoma by Weighted Gene Co-expression Network Analysis and Differential Gene Expression Analysis”. In: *Frontiers in Genetics* 11, p. 615308. DOI: 10.3389/fgene.2020.615308.

Balasubramanian, S. and Y. S. Chowdhury (July 2022). *Isosorbide*. Accessed on 26 Apr. 2023. URL: https://www.ncbi.nlm.nih.gov/books/NBK557839/.

Barabási, A. et al. (2011). “Network medicine: a network-based approach to human disease”. In: *Nature Reviews Genetics* 12.1, pp. 56–68.

Barabási, A. (2016). *Network Science*. Cambridge University Press. URL: http://networksciencebook.com.

Beyer, C. et al. (2012). “Stimulation of soluble guanylate cyclase reduces experimental dermal fibrosis”. In: *Annals of the rheumatic diseases* 71.6, pp. 1019–1026.

Biondini, D. et al. (2020). “Acute exacerbations of idiopathic pulmonary fibrosis (AE-IPF): an overview of current and future therapeutic strategies”. In: *Expert Review of Respiratory Medicine* 14.7, pp. 697–710.

BioSpecifics Technologies Corporation (2022). “XIAFLEX® (collagenase clostridium histolyticum) for injection, for intralesional use”. In: *Full Prescribing Information*. Initial U.S. Approval: 2010. URL: https://www.accessdata.fda.gov/drugsatfda_docs/label/2022/125338s129lbl.pdf.

Birring, S. S. et al. (2018). “Treatment of interstitial lung disease associated cough: CHEST guideline and expert panel report”. In: *Chest* 154.4, pp. 904–917.

Blanco, I. et al. (2011). “Effects of inhaled nitric oxide at rest and during exercise in idiopathic pulmonary fibrosis”. In: *Journal of applied physiology* 110.3, pp. 638–645.

Bois, R. du and Talmadge E. K. (2007). “Challenges in Pulmonary Fibrosis · 5: The NSIP/UIP Debate”. In: *Thorax* 62.11, pp. 1008–1012.

Borza, C. M. et al. (2018). “Discoidin Domain Receptor 2, a Potential Therapeutic Target in Lung Fibrosis”. In: *American Journal of Respiratory Cell and Molecular Biology* 59.3, pp. 277–278. DOI: 10.1165/rcmb.2018-0161ED.

Breitling, R. et al. (2004). “Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments”. In: *FEBS Letters* 573 (1-3), pp. 83–92.

Bueno, M. et al. (2023). “CYB5R3 in type II alveolar epithelial cells protects against lung fibrosis by suppressing TGF-β1 signaling”. In: *JCI Insight* 8.5, e161487.

Buhule, O. et al. (2014). “Stratified randomization controls better for batch effects in 450K methylation analysis: a cautionary tale”. In: *Frontiers in Genetics* 5, p. 354. DOI: 10.3389/fgene.2014.00354.

Cameli, P. et al. (2022). “The Effectiveness of Nintedanib in Patients with Idiopathic Pulmonary Fibrosis, Familial Pulmonary Fibrosis and Progressive Fibrosing Interstitial Lung Diseases: A Real-World Study”. In: *Biomedicines* 10.8, pp. 1973–.

Carlson, M. (2019). *org.Hs.eg.db: Genome wide annotation for Human*. R package version 3.8.2.

Cascorbi, I. (2012). “Drug interactions–principles, examples and clinical consequences”. In: *Deutsches Ärzteblatt International* 109.33-34, pp. 546–556. DOI: 10.3238/arztebl.2012.0546.

Chang, W. et al. (2022). *Shiny: Web Application Framework for R*. URL: https://CRAN.R-project.org/package=shiny.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Routledge.

Conesa, A. et al. (2016). “A Survey of Best Practices for RNA-Seq Data Analysis”. In: *Genome Biology* 17.1, p. 13.

Cross, A. L. et al. (Jan. 2023). “Pregabalin”. In: *StatPearls [Internet]*. [Updated 2022 Nov 14]. URL: https://www.ncbi.nlm.nih.gov/books/NBK470341/.

Csardi, G. and T. Nepusz (2006). “The igraph software package for complex network research”. In: *InterJournal* Complex Systems, p. 1695. URL: https://igraph.org.

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Routledge.

Danecek, Petr et al. (2021). “Twelve years of SAMtools and BCFtools”. In: *GigaScience* 10.2. ISSN: 2047-217X. DOI: 10.1093/gigascience/giab008. URL: https://doi.org/10.1093/gigascience/giab008.

Davis, Sean and Paul Meltzer (2007). “GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor”. In: *Bioinformatics* 14.14, pp. 1846–1847.

De Borda, J. (1781). “Mémoire sur les élections au scrutin”. In: *Histoire de L’Académie Royale des Sciences* 102, pp. 657–665.

Demidenko, E. (2016). “The p-Value You Can’t Buy”. In: *The American Statistician* 70 (1), pp. 33–38.

Dempsey, T. M. et al. (2021). “Clinical effectiveness of antifibrotic medications for idiopathic pulmonary fibrosis”. In: *Chest* 160.5, pp. 1819–1831.

— (2022). “Cost-effectiveness of the anti-fibrotics for the treatment of idiopathic pulmonary fibrosis in the United States”. In: *BMC pulmonary medicine* 22.1, p. 18.

Desai, T. J. et al. (2014). “Alveolar Progenitor and Stem Cells in Lung Development, Renewal and Cancer”. In: *Nature (London)* 507.7491, pp. 190–194.

Di Lieto, E. et al. (Apr. 2023). “ESPERANTO: a GLP-fied sEmi-SuPERvised toxicogenomics meta-dAta curatioN Tool”. Manuscript ID: BIOINF-2023-0552 SUBMITTED ARTICLE.

Durinck, Steffen et al. (2005). “BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis”. In: *Bioinformatics* 21, pp. 3439–3440.

— (2009). “Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt”. In: *Nature Protocols* 4, pp. 1184–1191.

Dvorak, H. F. (1986). “Tumors: wounds that do not heal. Similarities between tumor stroma generation and wound healing”. In: *New England Journal of Medicine* 315.26, pp. 1650–1659. DOI: 10.1056/NEJM198612253152606.

Espindola, M. S. et al. (2021). “Differential Responses to Targeting Matrix Metalloproteinase 9 in Idiopathic Pulmonary Fibrosis”. In: *Am J Respir Crit Care Med.* 203, pp. 458–470. DOI: 10.1164/rccm.201910-1977OC.

Evans, J.D. et al. (2001). “A phase II trial of marimastat in advanced pancreatic cancer”. In: *British Journal of Cancer* 85.12, pp. 1865–1870. DOI: 10.1054/bjoc.2001.2168.

Ewing, B. and Green P. (1998). “Base-Calling of Automated Sequencer Traces Using Phred. II. Error Probabilities”. In: *Genome research* 8.3, pp. 186–194.

Farrand, E. A. et al. (2020). “Corticosteroid use is not associated with improved outcomes in acute exacerbation of idiopathic pulmonary fibrosis”. In: *Respirology* 25 (7), pp. 629–635. DOI: https://doi.org/10.1111/resp.13753.

Federico, A. et al. (2020a). “Manually Curated and Harmonised Transcriptomics Datasets of Psoriasis and Atopic Dermatitis Patients”. In: *Scientific data* 7.1, p. 343.

Federico, A. et al. (2020b). “Transcriptomics in Toxicogenomics, Part II: Preprocessing and Differential Expression Analysis for High Quality Data”. In: *Nanomaterials (Basel, Switzerland)* 10.5, pp. 903–.

— (2022). “The integration of large-scale public data and network analysis uncovers molecular characteristics of psoriasis”. In: *Human Genomics* 16.1, p. 62.

Feng, F. et al. (2019). “Efficacy and safety of N-acetylcysteine therapy for idiopathic pulmonary fibrosis: An updated systematic review and meta-analysis”. In: *Experimental and Therapeutic Medicine* 18.1, pp. 802–816.

Fortin, J. et al. (2017). “Preprocessing, normalization and integration of the Illumina HumanMethylationEPIC array with minfi”. In: *Bioinformatics* 33.4. DOI: 10.1093/bioinformatics/btw691.

Fujita, H. et al. (2011). “Effects of Doxycycline on Production of Growth Factors and Matrix Metalloproteinases in Pulmonary Fibrosis”. In: *Respiration* 81, pp. 420–430. DOI: 10.1159/000324080.

Geng, J. et al. (2021). “Fatty Acid Metabolism and Idiopathic Pulmonary Fibrosis”. In: *Frontiers in Physiology* 12.

George, M. et al. (2016). “Molecular basis and functional significance of angiotensin II-induced increase in discoidin domain receptor 2 gene expression in cardiac fibroblasts”. In: *Journal of molecular and cellular cardiology* 90, pp. 59–69.

Giron, J.M. et al. (1993). “Inoperable pulmonary aspergilloma: percutaneous CT-guided injection with glycerin and amphotericin B paste in 15 cases”. In: *Radiology* 188.3, pp. 825–827.

Goh, K.I. et al. (2007). “The human disease network”. In: *Proceedings of the National Academy of Sciences* 104.21, pp. 8685–8690.

Gomez-Cabrero, D. et al. (2014). “Data integration in the era of omics: current and future challenges”. In: *BMC systems biology* 8.Suppl 2, p. I1.

Govindarajan, R. et al. (2012). “Microarray and its applications”. In: *Journal of pharmacy & bioallied sciences* 4.Suppl 2, S310.

Grimminger, F. et al. (2015). “The role of tyrosine kinases in the pathogenesis of idiopathic pulmonary fibrosis”. In: *European Respiratory Journal* 45.5, pp. 1426–1433.

Grither, W. R. and G. D. Longmore (2018). “Inhibition of tumor–microenvironment interaction and tumor invasion by small-molecule allosteric inhibitor of DDR2 extracellular domain”. In: *Proceedings of the National Academy of Sciences* 115.43, E9977–E9986.

Gu, Zuguang et al. (2016). “Complex heatmaps reveal patterns and correlations in multidimensional genomic data”. In: *Bioinformatics* 32.18, pp. 2847–2849. DOI: 10.1093/bioinformatics/btw313.

Guillot, L. et al. (2011). “Macrolides: new therapeutic perspectives in lung diseases”. In: *Int J Biochem Cell Biol* 43, pp. 1241–1246.

Hambly, N. et al. (2015). “Molecular classification of idiopathic pulmonary fibrosis: Personalized medicine, genetics and biomarkers”. In: *Respirology* 20.5, pp. 610–620.

Hardiman, Gary (2004). “Microarray platforms–comparisons and contrasts”. In: *Pharmacogenomics* 5.5, pp. 487–502. DOI: 10.1517/14622416.5.5.487.

He, Y. et al. (2022). “An 8-Ferroptosis-Related Genes Signature from Bronchoalveolar Lavage Fluid for Prognosis in Patients with Idiopathic Pulmonary Fibrosis”. In: *BMC pulmonary medicine* 22.1.

Held, P. et al. (2016). “Dynamic Clustering in Social Networks using Louvain and Infomap Method”. In: *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, pp. 875–882.

Hoeben, A. et al. (2021). “Personalized Medicine: Recent Progress in Cancer Therapy”. In: *Cancers* 13.2, p. 242. DOI: 10.3390/cancers13020242.

Hong, F. and R. Breitling (2008). “Gene expression: a comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments”. In: *Bioinformatics* 24 (3), pp. 374–382.

Huang, B.Y. et al. (2019). “Clinical analysis of percutaneous computed tomography-guided injection of cyanoacrylate for localization of 115 small pulmonary lesions in 113 asymptomatic patients”. In: *Journal of Cancer Research and Therapeutics* 15.6, pp. 1257–1261.

Huh, J. Y. et al. (2021). “Efficacy and safety of combined use of pirfenidone and nintedanib in patients with idiopathic pulmonary fibrosis”. In: *European Respiratory Journal* 58, PA468. DOI: 10.1183/13993003.congress-2021.PA468. URL: https://erj.ersjournals.com/content/58/suppl_63/PA468.

Hurst, L. C. et al. (2009). “Injectable collagenase clostridium histolyticum for Dupuytren’s contracture”. In: *The New England journal of medicine* 361.10, pp. 968–979.

Ito, K. et al. (2020). “Safety and reliability of computed tomography-guided lipiodol marking for undetectable pulmonary lesions”. In: *Interactive CardioVascular and Thoracic Surgery* 30.4, pp. 546–551.

Janet, P. et al. (2019). *The DisGeNET knowledge platform for disease genomics*. DOI: 10.1093/nar/gkz1021.

Jang, H. J. et al. (2021). “Corticosteroid Responsiveness in Patients with Acute Exacerbation of Interstitial Lung Disease Admitted to the Emergency Department”. In: *Scientific reports* 11.1, pp. 5762–5762.

Jeffrey, T. L. et al. (2022). *sva: Surrogate Variable Analysis*. R package version 3.46.0.

Kansaneläkelaitos (Mar. 2023). *Lääkehakau Kela*. URL: https://www.kela.fi/laakkeet_laakehaku.

Karbowiak, M. et al. (2016). “Dupuytren’s disease”. In: *CLINICAL UPDATES* 18, pp. 1–6.

Kauffmann, A. et al. (2009a). “arrayQualityMetrics–a Bioconductor package for quality assessment of microarray data”. In: *Bioinformatics* 25.3, pp. 415–416.

— (2009b). “arrayQualityMetrics–a bioconductor package for quality assessment of microarray data”. In: *Bioinformatics* 25 (3), pp. 415–416.

Kendall, R. T. and C. A. Feghali-Bostwick (2014). “Fibroblasts in fibrosis: novel roles and mediators”. In: *Frontiers in pharmacology* 5, p. 123. DOI: 10.3389/fphar.2014.00123.

Kim, C. K. et al. (2000). “Bronchoalveolar Lavage Cellular Composition in Acute Asthma and Acute Bronchiolitis”. In: *The Journal of pediatrics* 137.4, pp. 517–522.

Kim, D. et al. (2015). “HISAT: a fast spliced aligner with low memory requirements”. In: *Nature Methods*.

— (2019). “Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype”. In: *Nature biotechnology* 37.8, pp. 907–915.

Kim, S. et al. (2020). “Identification of the Molecular Basis of Anti-Fibrotic Effects of Soluble Guanylate Cyclase Activator Using the Human Lung Fibroblast Phosphoproteome”. In: *bioRxiv*.

King, T. E. et al. (2011). “Idiopathic Pulmonary Fibrosis”. In: *The Lancet* 378.9807.6, pp. 1949–1961.

Ko, K.H. et al. (2019). “A Simple and Efficient Method to Perform Preoperative Pulmonary Nodule Localization: CT-Guided Patent Blue Dye Injection”. In: *Clinical imaging* 58, pp. 74–79. DOI: 10.1016/j.clinimag.2019.03.001. URL: https://doi.org/10.1016/j.clinimag.2019.03.001.

Kolb, M. et al. (2017). “Therapeutic targets in idiopathic pulmonary fibrosis”. In: *Respiratory medicine* 131, pp. 49–57. DOI: 10.1016/j.rmed.2017.06.016. URL: https://www.resmedjournal.com/article/S0954-6111(17)30211-4/fulltext.

Korotkevich, G. et al. (2019). “Fast gene set enrichment analysis”. In: *bioRxiv*. DOI: 10.1101/060012. URL: http://biorxiv.org/content/early/2016/06/20/060012.

Koscielny, G. et al. (2017). “Open Targets: a platform for therapeutic target identification and validation”. In: *Nucleic Acids Research* 45.D1, pp. D985–D994. DOI: 10.1093/nar/gkw1055.

Krishna, R. et al. (Aug. 2022). *Idiopathic Pulmonary Fibrosis*. Treasure Island (FL). URL: https://www.ncbi.nlm.nih.gov/books/NBK448162/.

Kupfer, P. et al. (2012). “Batch correction of microarray data substantially improves the identification of genes differentially expressed in rheumatoid arthritis and osteoarthritis”. In: *BMC medical genomics* 5.1.

Kvien, T. K. et al. (2022). “The cost savings of biosimilars can help increase patient access and lift the financial burden of health care systems”. In: *Seminars in Arthritis and Rheumatism* 52.1, p. 151939.

Lahens, N. et al. (2017). “A comparison of Illumina and Ion Torrent sequencing platforms in the context of differential gene expression”. In: *BMC genomics* 18.1, p. 602.

Lai, R. (2022). *Collections: High Performance Container Data Types*. URL: https://CRAN.R-project.org/package=collections.

Lambers, C. et al. (2019). “Combined activation of guanylate cyclase and cyclic AMP in lung fibroblasts as a novel therapeutic concept for lung fibrosis”. In: *BioMed Research Internationa* 2019. DOI: https://doi.org/10.1155/2019/1345402.

Lancichinetti, A. et al. (2009). “Detecting the overlapping and hierarchical community structure in complex networks”. In: *New Journal of Physics* 11.3, p. 033015.

Leek, J. T. et al. (2023). *sva: Surrogate Variable Analysis*. R package version 3.48.0.

Leil, T. A. and R. Bertz (2014). “Quantitative Systems Pharmacology can reduce attrition and improve productivity in pharmaceutical research and development”. In: *Frontiers in pharmacology* 5, p. 247.

Leil, T. A. and S. Ermakov (2015). “Editorial: The emerging discipline of quantitative systems pharmacology”. In: *Frontiers in pharmacology* 6, p. 129. DOI: 10.3389/fphar.2015.00129.

Li, X. et al. (2021). “Regorafenib-Attenuated, Bleomycin-Induced Pulmonary Fibrosis by Inhibiting the TGF-β1 Signaling Pathway”. In: *International journal of molecular sciences* 22.4, pp. 1985–.

Liao, Y. et al. (2019). “The R Package Rsubread Is Easier, Faster, Cheaper and Better for Alignment and Quantification of RNA Sequencing Reads”. In: *Nucleic acids research* 47.8, e47–.

Liberzon, A. et al. (Jan. 2011). “The Molecular Signatures Database (MSigDB) hallmark gene set collection”. In: *Nucleic Acids Research* 39.suppl_1, pp. D152–D160. DOI: 10.1093/nar/gkq1189. URL: https://doi.org/10.1093/nar/gkq1189.

Lipsey, M. W. and D. B. Wilson (1993). “The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis”. In: *American Psychologist* 48 (12), pp. 1181–1209.

Love, M. I. et al. (2014). “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2”. In: *Genome Biology* 15, p. 550. DOI: 10.1186/s13059-014-0550-8.

MacKinnon, A. C. et al. (2012). “Regulation of transforming growth factor-β1-driven lung fibrosis by galectin-3”. In: *American Journal of Respiratory and Critical Care Medicine* 185.5, pp. 467–593. DOI: 10.1164/rccm.201106-0965OC.

Martin, M. (2011a). “Cutadapt Removes Adapter Sequences from High-Throughput Sequencing Reads”. In: *EMBnet.journal* 17.1, pp. 10–.

Martin, Marcel (May 2011b). “Cutadapt removes adapter sequences from high-throughput sequencing reads”. In: *EMBnet.journal* 17.1, pp. 10–12. DOI: http://dx.doi.org/10.14806/ej.17.1.200. URL: http://journal.embnet.org/index.php/embnetjournal/article/view/200.

Martinez, F. J. et al. (2011). “Idiopathic Pulmonary Fibrosis”. In: *Nature reviews. Disease primers* 3.1, pp. 17074–17074.

Marwah, V. S. et al. (2018). “INfORM: Inference of NetwOrk Response Modules”. In: *Bioinformatics* 34 (12), pp. 2136–2138.

— (2019). “eUTOPIA: solUTion for Omics data PreprocessIng and Analysis”. In: *Source code for biology and medicine* 14.1.

Mavridis, G.A.D and J.P.T Higgins (2021). *metap: Meta-Analysis Package for R*. URL: https://CRAN.R-project.org/package=metap.

Mazières, J. et al. (2005). “Wnt signaling in lung cancer”. In: *Cancer letters* 222, pp. 1–10. DOI: 10.1016/j.canlet.2004.08.040.

Menzel, V. et al. (2022). “Fyn-kinase and caveolin-1 in the alveolar epithelial junctional adherence complex contribute to the early stages of pulmonary fibrosis”. In: *European Journal of Pharmaceutical Sciences* 175. Open Access under CC BY 4.0 license, p. 106236. DOI: 10.1016/j.ejps.2022.106236. URL: https://www.sciencedirect.com/science/article/pii/S0928098722003137.

Misharin, A. V. et al. (2017). “Monocyte-Derived Alveolar Macrophages Drive Lung Fibrosis and Persist in the Lung over the Life Span”. In: *The Journal of experimental medicine* 214.8, pp. 2387–2404.

Mori, R. et al. (2006). “Acute downregulation of connexin43 at wound sites leads to a reduced inflammatory response, enhanced keratinocyte proliferation and wound fibroblast migration”. In: *Journal of cell science* 119.24, pp. 5193–5203.

Nagase, H. et al. (2006). “Structure and function of matrix metalloproteinases and TIMPs”. In: *Cardiovascular research* 69.3, pp. 562–573. DOI: 10.1016/j.cardiores.2005.12.002.

Nareznoi, D. et al. (2020). “Matrix Metalloproteinases Retain Soluble FasL-mediated Resistance to Cell Death in Fibrotic-Lung Myofibroblasts”. In: *Cells* 9.2, p. 411. DOI: 10.3390/cells9020411.

Nesterov, V. et al. (Nov. 2012). “Aldosterone-dependent and -independent regulation of the epithelial sodium channel (ENaC) in mouse distal nephron”. In:

American Journal of Physiology-Renal Physiology 303.9, F1289–F1299. DOI: 10.1152/ajprenal.00247.2012.

Newman, M.E.J. (2010). *Networks: an introduction*. Oxford University Press.

Nikolić, M. Z. et al. (2017). “Human Embryonic Lung Epithelial Tips Are Multipotent Progenitors That Can Be Expanded in Vitro as Long-Term Self-Renewing Organoids”. In: *eLife* 6.

Novak, C. M. et al. (2023). “Alveolar Macrophages Drive Lung Fibroblast Function During Idiopathic Pulmonary Fibrosis”. In: *American journal of physiology. Lung cellular and molecular physiology*.

Novartis Pharmaceuticals Corporation (2016). “JETREA® (ocriplasmin) Intravitreal Injection, 2.5 mg/mL”. In: *Full Prescribing Information*. Initial U.S. Approval: 2012. Revised: November 2016. URL: https://www.accessdata.fda.gov/drugsatfda_docs/label/2016/125422s037lbl.pdf.

Nuwaysir, E. et al. (1999). “Microarrays and toxicology: The advent of toxicogenomics”. In: *Molecular carcinogenesis* 24.3, pp. 153–159. DOI: 10.1002/(SICI)1098-2744(199903)24:3<153::AID-MC1>3.0.CO;2-P.

Nygaard, V. et al. (2016). “Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses”. In: *Biostatistics* 17 (1), pp. 29–39. DOI: 10.1093/biostatistics/kxv027.

O’Quigley, J.C. et al. (2020). *RP.advance: An Implementation of Wang and Tsiatis (1987) for Escalation with Overdose Control*. URL: https://CRAN.R-project.org/package=RP.advance.

Odell, S. G. et al. (2021). “The art of curation at a biological database: Principles and application”. In: *Current Plant Biology* 28.

Paci, P. et al. (2021). “Gene co-expression in the interactome: moving from correlation toward causation via an integrated approach to disease module discovery”. In: *npj Systems Biology and Applications* 7.1, pp. 1–15.

Park, T. et al. (2015). “Cost effectiveness of monoclonal antibody therapy for rare diseases: a systematic review”. In: *BioDrugs* 29.4, pp. 259–274.

Perscheid, C. et al. (2019). “Integrative gene selection on gene expression data: providing biological context to traditional approaches”. In: *Journal of integrative bioinformatics* 16.

Pertea, M. et al. (2016). “Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown”. In: *Nature Protocols*.

Pijet, B. et al. (2020). “The matrix metalloproteinase inhibitor marimastat inhibits seizures in a model of kainic acid-induced status epilepticus”. In: *Scientific Reports* 10.1.

Pons, P. and M. Latapy (2005). “Computing communities in large networks using random walks”. In: *International Symposium on Computer and Information Sciences*. Springer, pp. 284–293.

Popper, H. et al. (2022). “Lung fibrosis in autoimmune diseases and hypersensitivity: how to separate these from idiopathic pulmonary fibrosis”. In: *Rheumatology International* 42.7, pp. 1321–1330. DOI: 10.1007/s00296-021-04858-1.

Prasse, A. et al. (Dec. 2019a). “A Phase 1b Study of Vismodegib with Pirfenidone in Patients with Idiopathic Pulmonary Fibrosis”. In: *Pulmonary Therapy* 5.2, pp. 151–163. DOI: 10.1007/s41030-019-0096-8.

— (2019b). “BAL Cell Gene Expression Is Indicative of Outcome and Airway Basal Cell Involvement in Idiopathic Pulmonary Fibrosis”. In: *American journal of respiratory and critical care medicine* 199.5, pp. 622–630.

Qian, W. et al. (2019). “Complex Involvement of the Extracellular Matrix, Immune Effect, and Lipid Metabolism in the Development of Idiopathic Pulmonary Fibrosis”. In: *Frontiers in Physiology* 10, p. 1422.

Qiu, Z. et al. (2016). “Choosing between alternative placement strategies for conservation buffers using Borda count”. In: *Landscape and Urban Planning* 153, pp. 66–73.

Raghu, G. et al. (2015). “An Official ATS/ERS/JRS/ALAT Clinical Practice Guideline: Treatment of Idiopathic Pulmonary Fibrosis An Update of the 2011 Clinical Practice Guideline”. In: *American Journal of Respiratory and Critical Care Medicine* 192.2, e3–e19.

— (2022). “Randomized Phase IIa Clinical Study of an Anti-αvβ6 Monoclonal Antibody in Idiopathic Pulmonary Fibrosis”. In: *American Journal of Respiratory and Critical Care Medicine* 206.9, pp. 1166–1168.

Rahiminejad, S. et al. (2018). “Topological and functional comparison of community detection algorithms in biological networks”. In: *BMC bioinformatics* 19.1, p. 524.

Rao, M. S. et al. (2019). “Comparison of RNA-Seq and Microarray Gene Expression Platforms for the Toxicogenomic Evaluation of Liver From Short-Term Rat Toxicity Studies”. In: *Frontiers in genetics* 9, p. 636. DOI: 10.3389/fgene.2018.00636.

Richardson, S. et al. (2016). “Statistical Methods in Integrative Genomics”. In: *Annual Review of Statistics and Its Application* 3, pp. 181–209. DOI: 10.1146/annurev-statistics-041715-033506.

Rinciog, C. et al. (2017). “A Cost-Effectiveness Analysis of Nintedanib in Idiopathic Pulmonary Fibrosis in the UK”. In: *PharmacoEconomics* 35.5, pp. 479–491. DOI: 10.1007/s40273-017-0486-4.

Ritchie, M. E. et al. (2015a). “limma powers differential expression analyses for RNA-sequencing and microarray studies”. In: *Nucleic Acids Research* 43.7. DOI: 10.1093/nar/gkv007.

Ritchie, ME. et al. (2015b). “limma powers differential expression analyses for RNA-sequencing and microarray studies”. In: *Nucleic Acids Research* (7), e47. URL: https://doi.org/10.1093/nar/gkv007.

Roach, K.M. and P. Bradding (2019). “Ca2+ signalling in fibroblasts and the therapeutic potential of KCa3.1 channel blockers in fibrotic diseases”. In: *British Journal of Pharmacology* 176.22, pp. 4252–4272. DOI: 10.1111/bph.14939. URL: https://bpspubs.onlinelibrary.wiley.com/doi/full/10.1111/bph.14939.

Ryan, N. M. et al. (2012). “Gabapentin for refractory chronic cough: a randomised, double-blind, placebo-controlled trial”. In: *The Lancet* 380.9853, pp. 1583–1589.

Ryman, J. T. and B. Meibohm (2017). “Pharmacokinetics of Monoclonal Antibodies”. In: *CPT: pharmacometrics & systems pharmacology* 6.9, pp. 576–588. DOI: 10.1002/psp4.12224.

Saari, D. G. (1995). *Basic Geometry of Voting*. Vol. 48. Springer Science & Business Media.

Saarimäki, L. A. et al. (2020). “Toxicogenomics Analysis of Dynamic Dose-Response in Macrophages Highlights Molecular Alterations Relevant for Multi-Walled Carbon Nanotube-Induced Lung Fibrosis”. In: *NanoImpact* 20, pp. 100274–.

— (2022). “Prospects and challenges for FAIR toxicogenomics data”. In: *Nature Nanotechnology* 17. Matters Arising to this article was published on 23 December 2021, pp. 17–18. DOI: 10.1038/s41565-021-00990-1.

Sakamoto, S. et al. (2013). “Efficacy of Pirfenidone in Patients with Advanced-Stage Idiopathic Pulmonary Fibrosis”. In: *Internal medicine* 52.22, pp. 2495–2501.

Schiller, H. B. et al. (2019). “The Human Lung Cell Atlas: A High-Resolution Reference Map of the Human Lung in Health and Disease”. In: *American Journal of Respiratory Cell and Molecular Biology* 61.1, pp. 31–41.

Schimek, Michael G. et al. (2015). “TopKLists: a comprehensive R package for statistical inference, stochastic aggregation, and visualization of multiple omics ranked lists”. In: *Stat Appl Genet Mol Biol* 14.3, pp. 311–316. DOI: 10.1515/sagmb-2014-0093.

Sgalla, G. et al. (2020). “Antibody-based therapies for idiopathic pulmonary fibrosis”. In: *Expert Opinion on Biological Therapy* 20.5, pp. 463–470. DOI: 10.1080/14712598.2020.1735344.

Shah, M. A. et al. (2021). “Phase III Study to Evaluate Efficacy and Safety of Andecaliximab With mFOLFOX6 as First-Line Treatment in Patients With Advanced Gastric or GEJ Adenocarcinoma (GAMMA-1)”. In: *J Clin Oncol* 20, pp. 990–1000. DOI: 10.1200/JCO.20.02755.

Shao, L. and W. Wei (2014). “Vitreomacular traction syndrome”. In: *Chinese medical journal* 127.8, pp. 1566–1571.

Sokolova, M. et al. (2017). *TopKLists: Inference, Aggregation and Visualization for Top-K Ranked Lists.* URL: https://CRAN.R-project.org/package=TopKLists.

Sparano, J. A. et al. (2004). “Randomized Phase III Trial of Marimastat Versus Placebo in Patients With Metastatic Breast Cancer Who Have Responding or Stable Disease After First-Line Chemotherapy: Eastern Cooperative Oncology Group Trial E2196”. In: *Journal of Clinical Oncology* 22.12, pp. 2289–2303. DOI: 10.1200/JCO.2004.08.054.

Stalmans, P. et al. (2012). “Enzymatic vitreolysis with ocriplasmin for vitreomacular traction and macular holes”. In: *The New England journal of medicine* 367.7, p. 606.

Stanzel, F. (2012). “Bronchoalveolar Lavage”. In: *Principles and Practice of Interventional Pulmonology. New York, NY: Springer New York*, pp. 165–176.

Stark, R. et al. (2019). “RNA sequencing: the teenage years”. In: *Nature Reviews Genetics* 20.10, pp. 631–656.

Sugeir, S. et al. (2019). “Bronchoscopy in the Intensive Care Unit”. In: *Cham: Springer International Publishing*, pp. 49–55.

Sullivan, G. M. and R. Feinn (2012). “Using effect size–or why the p value is not enough”. In: *Journal of graduate medical education* 4 (3), pp. 279–282. DOI: 10.4300/JGME-D-12-00156.1.

Suzuki, T. et al. (2008). “Role of Innate Immune Cells and Their Products in Lung Immunopathology”. In: *The international journal of biochemistry & cell biology* 40.6, pp. 1348–1361.

Taminau, J. et al. (2012). “Unlocking the potential of publicly available microarray data using inSilicoDb and inSilicoMerging R/Bioconductor packages”. In: *BMC bioinformatics* 13.1, p. 335.

Tan, K. R. et al. (2011). “Doxycycline for Malaria Chemoprophylaxis and Treatment: Report from the CDC Expert Meeting on Malaria Chemoprophylaxis”. In: *The American journal of tropical medicine and hygiene* 84.4, pp. 517–531. DOI: 10.4269/ajtmh.2011.10-0285.

Thannickal, V. J. and V. B. Antony (2018). “Is personalized medicine a realistic goal in idiopathic pulmonary fibrosis?” In: *Expert review of respiratory medicine* 12.6. PMCID: PMC6157605, pp. 441–443. DOI: 10.1080/17476348.2018.1464913.

Tibshirani, R. et al. (2011). “A bias correction for the minimum error rate in cross-validation”. In: *The Annals of Applied Statistics* 5 (4), pp. 2339–2362.

Todd, J. L. et al. (2020). “Circulating matrix metalloproteinases and tissue metalloproteinase inhibitors in patients with idiopathic pulmonary fibrosis in the multicenter IPF-PRO Registry cohort”. In: *BMC Pulmonary Medicine* 20.1, pp. 1–8.

Trevor, H. et al. (2022). *pam: prediction analysis for microarrays*. R package version 1.9.7. URL: https://cran.r-project.org/package=pamr.

Vancheri, C. et al. (2010). “Idiopathic pulmonary fibrosis: a disease with similarities and links to cancer biology”. In: *European Respiratory Journal* 35, pp. 496–504.

Vancheri, C. (2015). “Idiopathic pulmonary fibrosis and cancer: do they really look similar?” In: *BMC medicine* 13.1, p. 220. DOI: 10.1186/s12916-015-0478-1.

Vertiganm, A. E. et al. (2016). “Pregabalin and Speech Pathology Combination Therapy for Refractory Chronic Cough: A Randomized Controlled Trial”. In: *Chest* 149.3, pp. 639–648.

Villegas, M. R. et al. (2018). “Collagenase Nanocapsules: A Approach for Fibrosis Treatment”. In: *Journal of the American Chemical Society* 140.32, pp. 10239–10247.

Wang, Y. et al. (2018). “Pulmonary Alveolar Type I Cell Population Consists of Two Distinct Subtypes That Differ in Cell Fate”. In: *Proceedings of the National Academy of Sciences - PNAS* 115.10, pp. 2407–2412.

Watt, A. J. and V. R. Hentz (Apr. 2011). “Collagenase clostridium histolyticum: a novel nonoperative treatment for Dupuytren’s disease”. In: *International Journal of Clinical Rheumatology* 6.2, pp. 123–133. DOI: 10.2217/ijr.11.4. URL: https://doi.org/10.2217/ijr.11.4.

Wells, A. (2015). “Combination therapy in idiopathic pulmonary fibrosis: the way ahead will be hard”. In: *European Respiratory Journal* 45.5, pp. 1208–1210.

Wicky, S. et al. (1994). “CT-guided localizations of pulmonary nodules with methylene blue injections for thoracoscopic resections”. In: *Chest* 106.5, pp. 1326–1328.

Wilkinson, M. D. et al. (2016). “The FAIR guiding principles for scientific data management and stewardship”. In: *Scientific data* 3.

Wu, C. et al. (2013). “Network-based differential gene expression analysis suggests cell cycle related genes regulated by E2F1 underlie the molecular difference between smoker and non-smoker lung adenocarcinoma”. In: *BMC bioinformatics* 14.1, p. 365.

Wu, Tong et al. (2021). “clusterProfiler 4.0: A universal enrichment tool for interpreting omics data”. In: *The Innovation* 2.3, p. 100141. DOI: 10.1016/j.xinn.2021.100141.

Wuyts, W. A. et al. (2013). “The pathogenesis of pulmonary fibrosis: a moving target”. In: *Eur Respir J* 41, pp. 1207–1218.

— (2014). “Combination therapy: The future of management for idiopathic pulmonary fibrosis?” In: *The Lancet Respiratory Medicine* 2.11, pp. 933–942. DOI: 10.1016/S2213-2600(14)70140-2. URL: https://www.thelancet.com/journals/lanres/article/PIIS2213-2600(14)70140-2/fulltext.

Yamaguchi, K. et al. (2017). “Anti-inflammatory actions of gabapentin and pregabalin on the substance P-induced mitogen-activated protein kinase activation in U373 MG human glioblastoma astrocytoma cells”. In: 16.5, pp. 6109–6115.

Yasaei, R. et al. (Jan. 2023). “Gabapentin”. In: *StatPearls [Internet]*. [Updated 2022 Dec 19]. URL: https://www.ncbi.nlm.nih.gov/books/NBK493228/.

Young, T. L. et al. (2021). “Pulmonary delivery of the broad-spectrum matrix metalloproteinase inhibitor marimastat diminishes multiwalled carbon nanotube-induced circulating bioactivity without reducing pulmonary inflammation”. In: *Particle and Fibre Toxicology* 18.1, p. 34. DOI: https://doi.org/10.1186/s12989-021-00427-w.

Yu, Guangchuang et al. (2012). “clusterProfiler: an R package for comparing biological themes among gene clusters”. In: *OMICS: A Journal of Integrative Biology* 16.5, pp. 284–287. DOI: 10.1089/omi.2011.0118.

Zhang, H. et al. (Sept. 2019). “Pharmacokinetic/Pharmacodynamic Integration of Doxycycline Against Mycoplasma hyopneumoniae in an In Vitro Model”. In: *Frontiers in Pharmacology* 10, p. 1088. DOI: 10.3389/fphar.2019.01088.

Zhang, L. et al. (2018). “Macrophages: friend or foe in idiopathic pulmonary fibrosis?” In: *Respiratory research* 19.1, p. 170.

Zhang, Y. et al. (2021). “Rapid and accurate alignment of nucleotide conversion sequencing reads with HISAT-3N”. In: *Genome Research* 31, pp. 1290–1295.

Zhao, H. et al. (2016). “Targeting of discoidin domain receptor 2 (DDR2) prevents myofibroblast activation and neovessel formation during pulmonary fibrosis”. In: *Molecular therapy* 24.11, pp. 1734–1744.

Zhao, S. et al. (2014). “Comparison of RNA-Seq and Microarray in Transcriptome Profiling of Activated T Cells”. In: *PLOS ONE* 9.1, e78644. DOI: 10.1371/journal.pone.0078644.

Zito, P. M. et al. (2023). *Vismodegib*. URL: https://www.ncbi.nlm.nih.gov/books/NBK513360/.

Dataset references

Boesch, M. et al. (2020). “Transcriptomic Profiling Reveals Disease-Specific Characteristics of Epithelial Cells in Idiopathic Pulmonary Fibrosis”. In: *Respiratory research* 21.1, pp. 165–165.

Brereton, C. J. et al. (2022). “Pseudohypoxic HIF Pathway Activation Dysregulates Collagen Structure-Function in Human Lung Fibrosis”. In: *eLife* 11.

Cecchini, M. J. et al. (2018). “Comprehensive Gene Expression Profiling Identifies Distinct and Overlapping Transcriptional Profiles in Non-Specific Interstitial Pneumonia and Idiopathic Pulmonary Fibrosis”. In: *Respiratory research* 19.1, pp. 153–153.

Cho, J. et al. (2011). “Systems Biology of Interstitial Lung Diseases: Integration of mRNA and microRNA Expression Changes”. In: *BMC genomics* 4.1, pp. 8–8.

Christmann, R. B. et al. (2014). “Association of Interferon- and Transforming Growth Factor β–Regulated Genes and Macrophage Activation With Systemic Sclerosis–Related Progressive Lung Fibrosis”. In: *Arthritis & rheumatology (Hoboken, N.J.)* 66.3, pp. 714–725.

DePianto, D. J. et al. (2015). “Heterogeneous Gene Expression Signatures Correspond to Distinct Lung Pathologies and Biomarkers of Disease Severity in Idiopathic Pulmonary Fibrosis”. In: *Thorax* 70.1, pp. 48–56.

— (2021). “Molecular Mapping of Interstitial Lung Disease Reveals a Phenotypically Distinct Senescent Basal Epithelial Cell Population”. In: *JCI Insight* 6.8.

Furusawa, H. et al. (2020). “Chronic Hypersensitivity Pneumonitis, an Interstitial Lung Disease with Distinct Molecular Signatures”. In: *American journal of respiratory and critical care medicine* 202.10, pp. 1430–1444.

Geng, J. et al. (2015). “Down-Regulation of USP13 Mediates Phenotype Transformation of Fibroblasts in Idiopathic Pulmonary Fibrosis”. In: *Respiratory research* 16.1, pp. 124–124.

Konigsberg, I. R. et al. (2021). “Molecular Signatures of Idiopathic Pulmonary Fibrosis”. In: *American journal of respiratory cell and molecular biology* 65.4, pp. 430–441.

Larsson, O. et al. (2008). “Fibrotic Myofibroblasts Manifest Genome-Wide Derangements of Translational Control”. In: *PloS one* 3.9, pp. 3220–.

Lindahl, G. E. et al. (2013). “Microarray Profiling Reveals Suppressed Interferon Stimulated Gene Program in Fibroblasts from Scleroderma-Associated Interstitial Lung Disease”. In: *Respiratory research* 14.1, pp. 80–80.

Luzina, I. G. et al. (2018). “Transcriptomic Evidence of Immune Activation in Macroscopically Normal-Appearing and Scarred Lung Tissues in Idiopathic Pulmonary Fibrosis”. In: *Cellular immunology* 325, pp. 1–13.

Marwick, J. A. et al. (2018). “Neutrophils Induce Macrophage Anti-Inflammatory Reprogramming by Suppressing NF-κB Activation”. In: *Cell death & disease* 9.6, pp. 665–13.

McDonough, J. E. et al. (2019). “Transcriptional Regulatory Model of Fibrosis Progression in the Human Lung”. In: *JCI insight* 4.22.

Megan, E. et al. (2020). “Resident Mesenchymal Vascular Progenitors Modulate Adaptive Angiogenesis and Pulmonary Remodeling via Regulation of Canonical Wnt Signaling”. In: *The FASEB journal* 34.8, pp. 10267–10285.

Meltzer, E. B. et al. (2011). “Bayesian Probit Regression Model for the Diagnosis of Pulmonary Fibrosis: Proof-of-Principle”. In: *BMC medical genomics* 4.1, pp. 70–70.

Parker, M. W. et al. (2014). “Fibrotic Extracellular Matrix Activates a Profibrotic Positive Feedback Loop”. In: *The Journal of clinical investigation* 124.4, pp. 1622–1635.

Peng, R. et al. (2013). “Bleomycin Induces Molecular Changes Directly Relevant to Idiopathic Pulmonary Fibrosis: A Model for ‘Active’ Disease”. In: *PloS one* 8.4, pp. 59348–.

Prasse, A. et al. (2019b). “BAL Cell Gene Expression Is Indicative of Outcome and Airway Basal Cell Involvement in Idiopathic Pulmonary Fibrosis”. In: *American journal of respiratory and critical care medicine* 199.5, pp. 622–630.

Yuanuan, S. et al. (2013). “Syndecan-2 Exerts Antifibrotic Effects by Promoting Caveolin-1—mediated Transforming Growth Factor-β Receptor I Internalization and Inhibiting Transforming Growth Factor-B1 Signaling”. In: *American journal of respiratory and critical care medicine* 188.7, pp. 831–841.

Appendix

Supplementary figures

PCA and MDS plots of the raw data for each non-integrated dataset

The PCA plots of RNA-seq datasets presented here were generated using raw data without normalization using DESeq2. Similarly, the MDS plots for the microarray data were created before batch correction, as an unknown error occurred when attempting to create MDS plots after batch correction using the eUTOPIA-app.

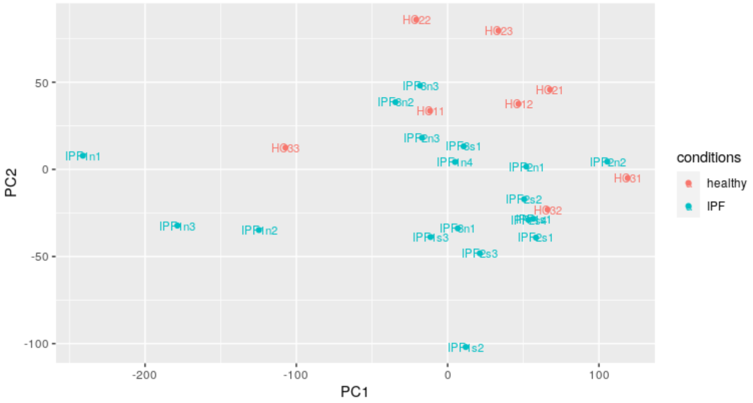

Figure 44: Biopsy dataset GSE99621 PCA plot before normalization.

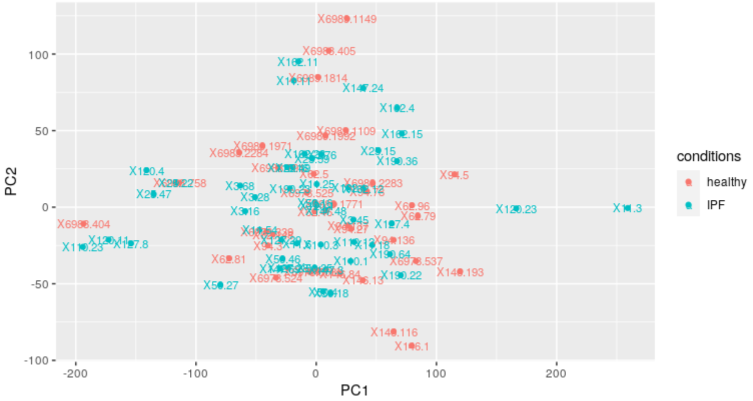

Figure 45: Biopsy dataset GSE124685 PCA plot before normalization.

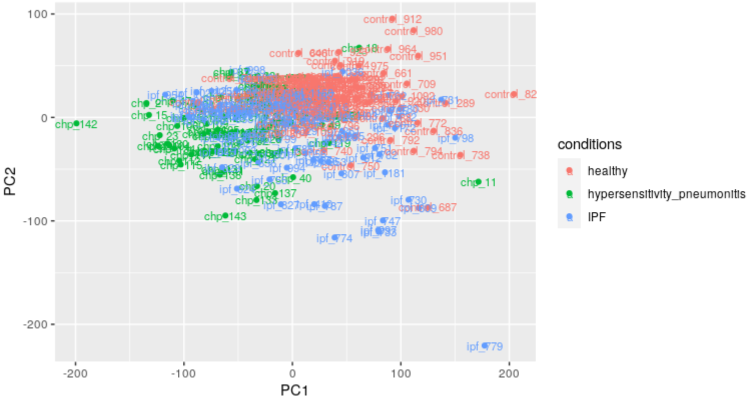

Figure 46: Biopsy dataset GSE150910 PCA plot before normalization.

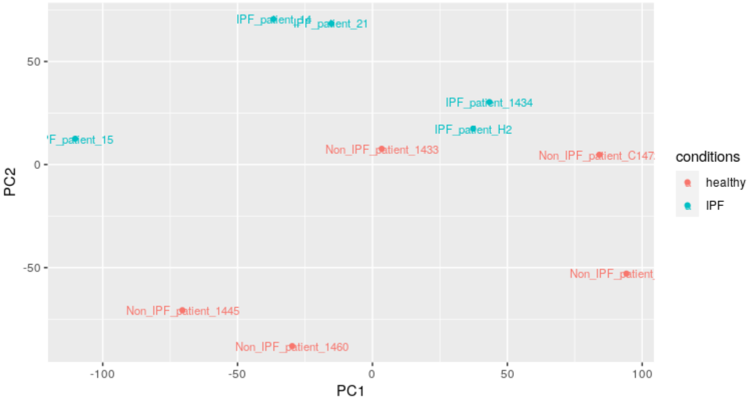

Figure 47: Epithelial dataset GSE151673 PCA plot before normalization.

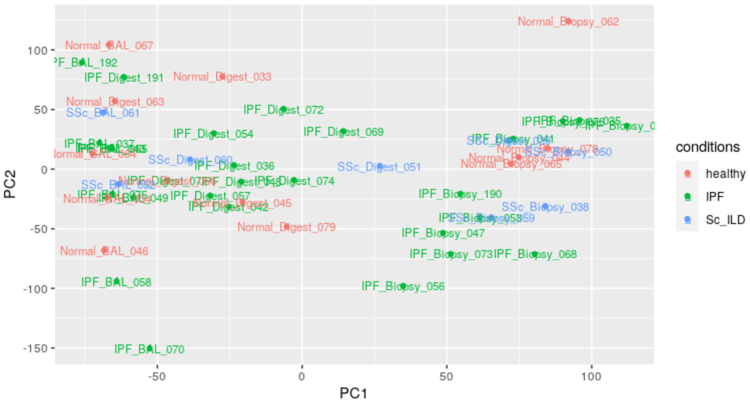

Figure 48: Biopsy and BAL dataset GSE166036 PCA plot before normalization.

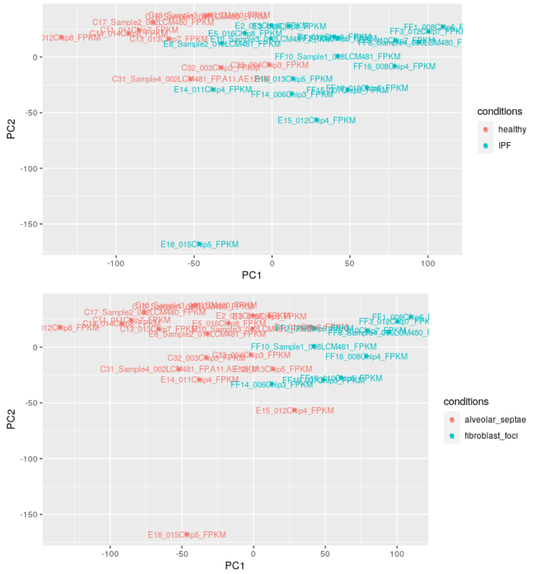

Figure 49: Biopsy dataset GSE169500 PCA plots before normalization. Upper figure is grouped by disease condition and lower figure is grouped bu tissue source: alveolar septae and myofibroblast foci.

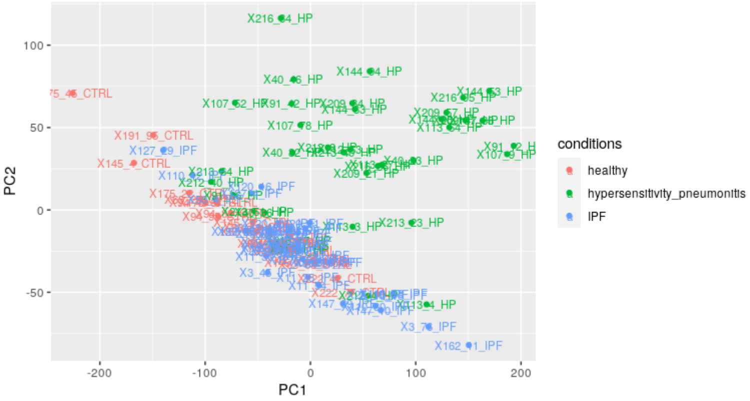

Figure 50: Biopsy dataset GSE184316 PCA plot before normalization.

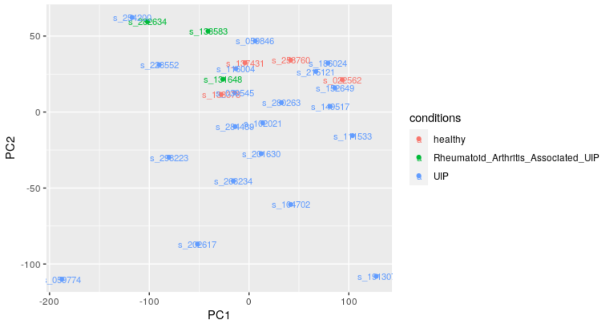

Figure 51: Biopsy dataset GSE199152 PCA plot before normalization.

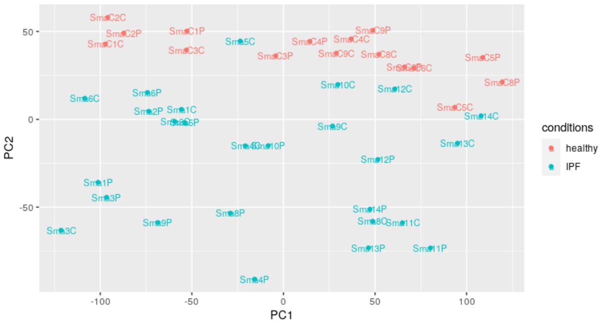

Figure 52: Biopsy dataset GSE199949 PCA plot before normalization.

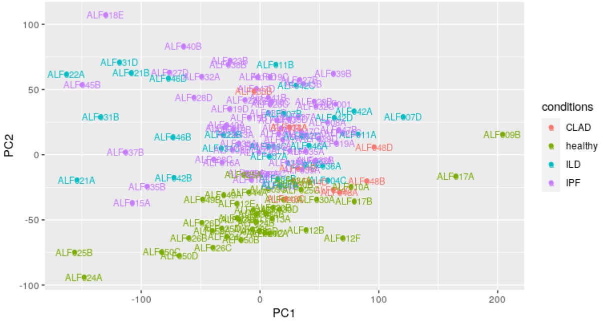

Figure 53: Biopsy dataset GSE213001 PCA plot before normalization.

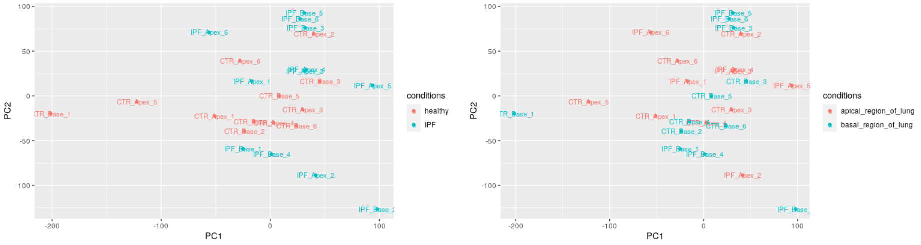

Figure 54: Fibroblast dataset GSE185492 PCA plot before normalization. Left figure is grouped by disease and right figure is grouped by tissue source: apical and basal regions of the lung.

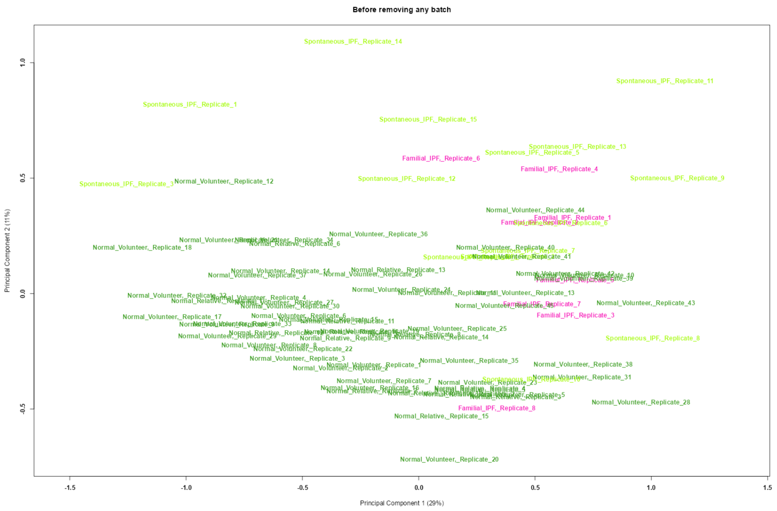

Figure 55: Alveolar macrophage dataset GSE49072 MDS plot before normalization.

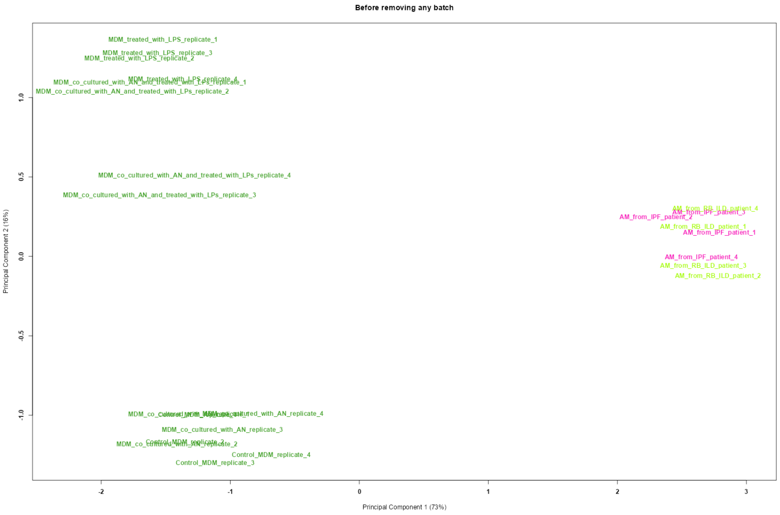

Figure 56: Alverolar macrophage dataset GSE490010 MDS plot before normalization. Experimental set 1: (n=4 per group, 4 match paired groups): Human monocyte-derived macrophages (MDM) co-colture with or without apoptotic neutrophils and with or without LPS (1ng/ml) for 9 hours. Experimental set 2: (n=4 per group, 2 groups):AM from IPF and RB-ILD patients isolated from bronchoalveolar lavage by cell sorting.

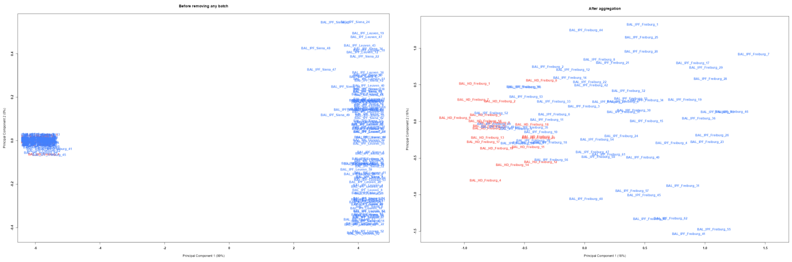

Figure 57: BAL dataset GSE70866 MDS plots before batch correction.The dataset comprises samples from three distinct cities: Freiburg, Sienna, and Leuven, with only the control samples sourced from Freiburg. The left MDS plot represents all the samples, while the right one illustrates only the samples from Freiburg. Despite applying batch correction utilizing city as a covariate, 25000 genes were found to be differentially expressed out of 29000, and hence, only samples from Freiburg were employed due to the batch effect.

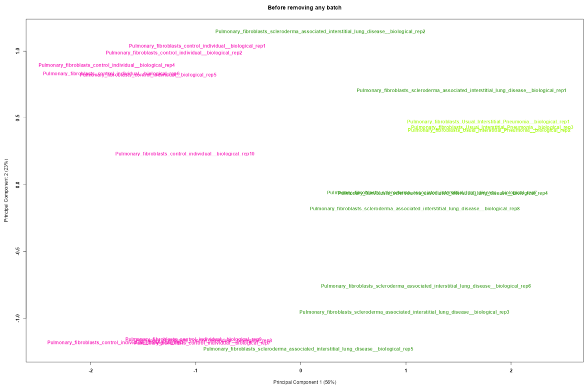

Figure 58: Fibroblast dataset GSE40839 MDS plot before batch correction.

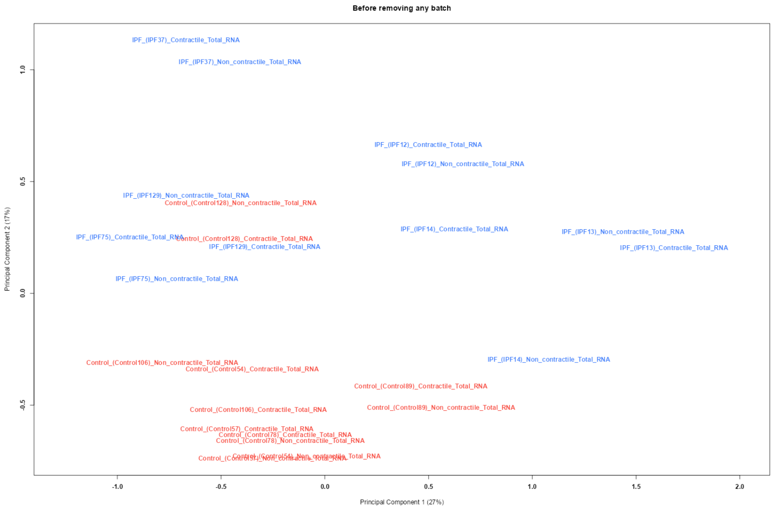

Figure 59: Fibroblast dataset GSE11196 MDS plot before batch correction.

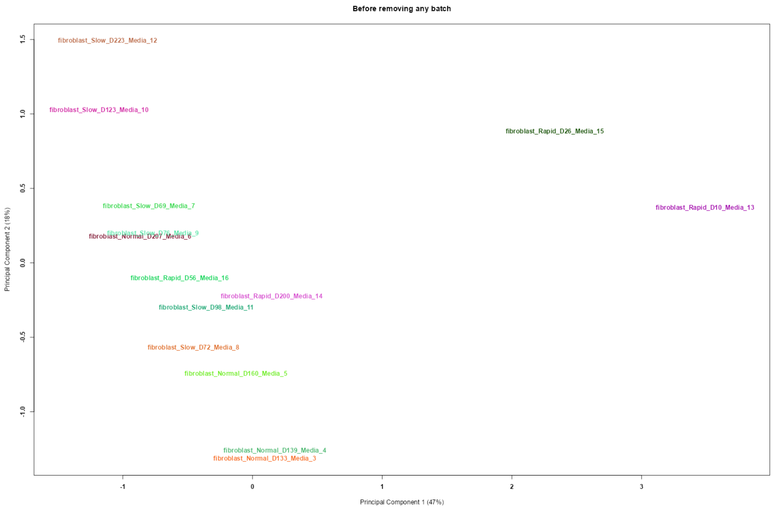

Figure 60: Fibroblast dataset GSE44723 MDS plot before batch correction. Out of the total genes examined, none were found to have an adjusted p-value less than 0.05 across the rapid, slow, and normal groups. However, when comparing the slow and normal groups, only two genes were discovered to have an adjusted p-value less than 0.05.

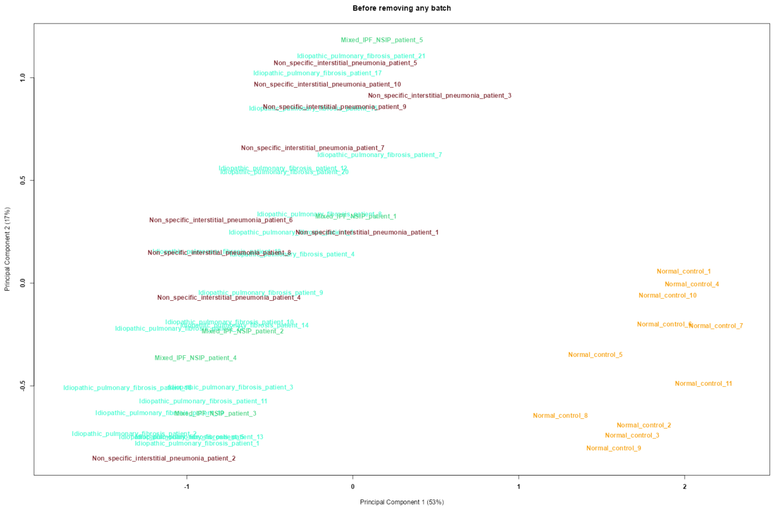

Figure 61: Biopsy dataset GSE110147 MDS plot before batch correction. Three diseases: IPF, NSIP, and Mixed IPF NSIP. For each of these diseases, approximately 20,000 differentially expressed genes were identified out of a total of 30,000 genes.

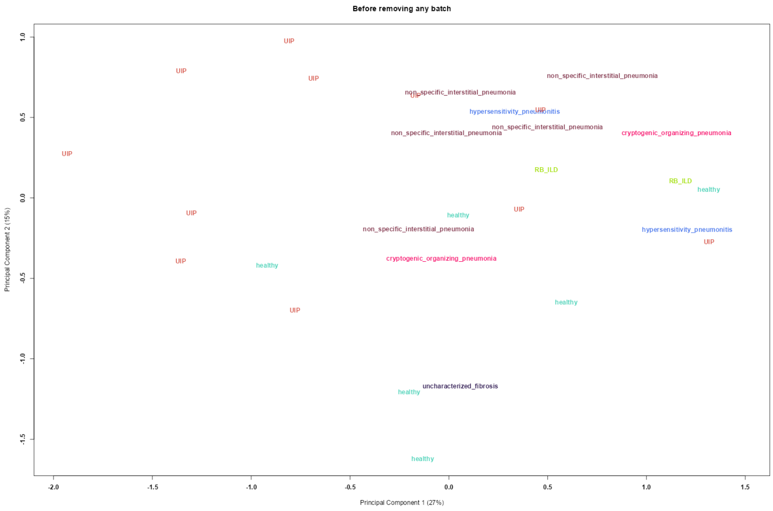

Figure 62: Biopsy dataset GSE21369 MDS plot before batch correction. Differential gene expression analysis was performed by comparing UIP samples to healthy controls.

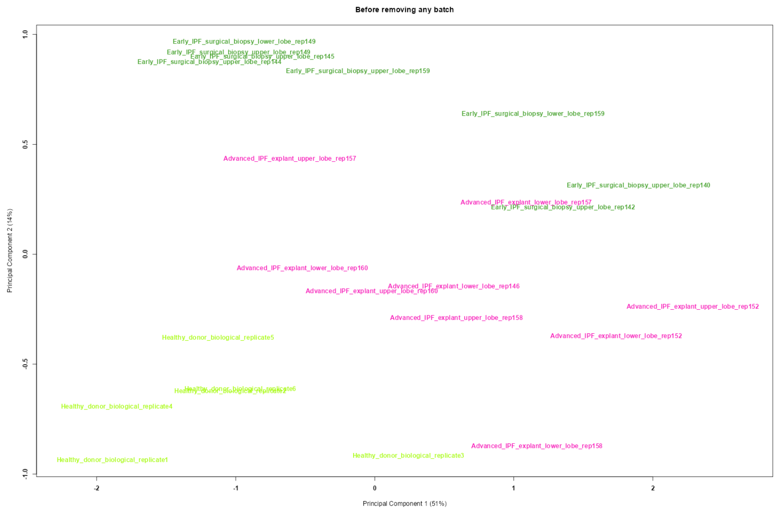

Figure 63: Biopsy dataset GSE24206 MDS plot before batch correction.

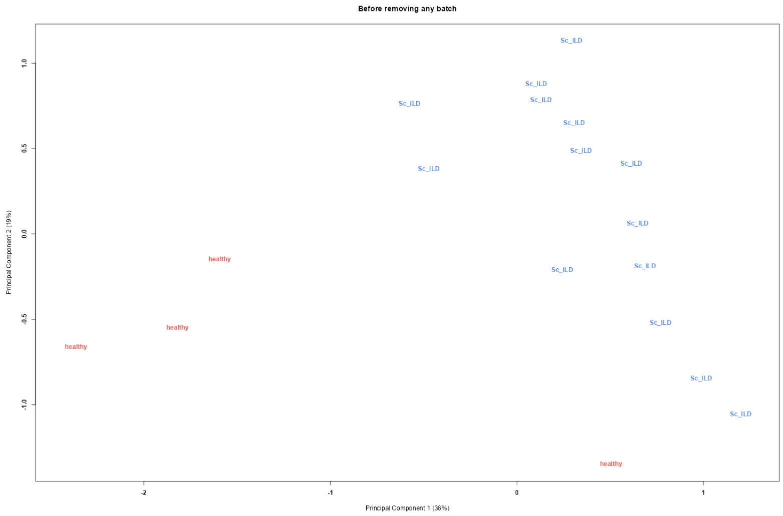

Figure 64: Biopsy dataset GSE76808 MDS plot before batch correction.

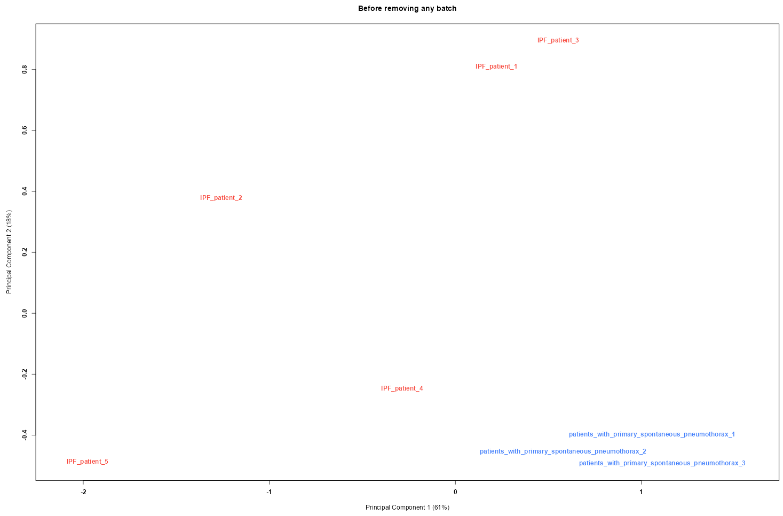

Figure 65: Biopsy dataset GSE72073 MDS plot before batch correction. Primary spontaneous pneumothorax samples as control.

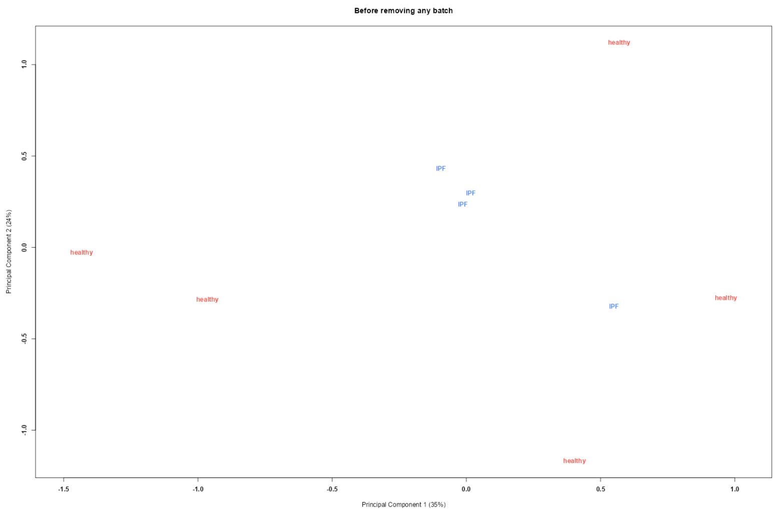

Figure 66: Biopsy dataset GSE94060 MDS plot before batch correction. None of the genes examined in this study exhibited an adjusted p-value less than 0.05.

Code

```
####Here is just part of the code I used because
##Some of the parts had to be performed
##For each dataset separately and there
###was also some work done in the console
####I can see looking some parts that I would
##Have bit more sophisticated solutions for some parts now
##Since I gained some experience during this work


#Download the fastq-files from a text file European nucleotide
archive

while read -r line; do echo "$line"; axel -n 10 "$line"; done <
PRJNA513068_tsv.txt

#make the fastqc report

fastqc *.fastq.gz -t 10 -o /home/sinkala/
expressiondata/rnaseq/folder

#Trimming the paired end fastq-files with cutadapt

FASTQ_FILES_R1=$(ls *_1.fastq.gz |cut -d "_" -f1)
FASTQ_FILES_R2=$(ls *_2.fastq.gz |cut -d "_" -f1)

for i in $FASTQ_FILES_R1; do ~/.local/bin/cutadapt -a
```

```
AGATCGGAAGAG -A AGATCGGAAGAG
-q 20 -m 60 -j 2 -o TRIMMED_${i}
_1.fastq.gz -p TRIMMED_${i}_2.fastq.gz ${i}_1.fastq.gz ${i}
_2.fastq.gz; done

#Trimming the single end fastq-files and Ion-Torrent fastq-
files with cutadapt

FASTQ_FILES=$(ls *.fastq.gz |cut -d "." -f1)

for i in $FASTQ_FILES; do ~/.local/bin/cutadapt -q 20 -m 35 -j
2 -o TRIMMED_${i}.fastq.gz ${i}.fastq.gz; done

#Aligning paired end fastq-files with Hisat2

HISAT2_INDEXES=/nasdata/RNA_Seq/RNA_Seq_tools/toninos_pipeline/
indexes/grch38
FASTQ_FILES_R1=$(ls 1.fastq.gz |cut -d "" -f1)
FASTQ_FILES_R2=$(ls2.fastq.gz |cut -d "" -f1)

for i in $FASTQ_FILES_R1; do /nasdata/RNA_Seq/RNA_Seq_tools/
toninos_pipeline/hisat2-2.2.1/hisat2 -q -p 10 -x
$HISAT2_INDEXES/genome -1 ${i}_1.fastq.gz -2 ${i}_2.fastq.gz |
samtools view -Sbh > ${i}.bam; done

#Aligning single end fastq-files and Ion-Torrent with Hisat2

HISAT2_INDEXES=/nasdata/RNA_Seq/RNA_Seq_tools/toninos_pipeline/
indexes/grch38
FASTQ_FILES_R1=$(ls *.fastq.gz |cut -d "." -f1)

for i in $FASTQ_FILES_R1; do /nasdata/RNA_Seq/RNA_Seq_tools/
toninos_pipeline/hisat2-2.2.1/hisat2 -q -p 10 -x
$HISAT2_INDEXES/genome -U ${i}.fastq.gz | samtools view -Sbh >
${i}.bam; done

#Filtering and sorting the uniquely mapped reads

BAM_FILES=$(ls *.bam |cut -d "." -f1)


for i in $BAM_FILES; do samtools view -H ${i}.bam > ${i}
_header.sam; samtools view ${i}.bam |grep -w "NH:i:1" | cat $
{i}_header.sam - |samtools view -@ 40 -Sb - > ${i}_unique.bam;
samtools sort -@ 40 ${i}_unique.bam > ${i}_unique_sorted.bam;
```

```
rm ${i}.sam; rm ${i}_header.sam; rm ${i}.bam; rm ${i}
_unique.bam; done

#Building the count matrix in RStudio.
#Example dataset here GSE166036
#These steps might have been bit different between datasets
#Because sample names differ and the order in metadata
#samples and the actual data samples do not always correspond

setwd("path")
files <- list.files(pattern = "\\.bam")
counts<-Rsubread::featureCounts(files,
isGTFAnnotationFile = TRUE,
annot.ext="/nasdata/RNA_Seq/references/Ensembl_v108_hsapiens/
gtf/Homo_sapiens.GRCh38.108.gtf", GTF.attrType="gene_id",
isPairedEnd=TRUE)

write.table(counts$counts, file="filename.csv", sep="\t")

setwd("/nasdata/sinkala/expressiondata/rnaseq/GSE166036")


raw_matrix <- read.csv("rawcounts_GSE166036.csv", sep="\t")

#change the colnames to SRRxxxxxxxx

column_names<-c()
for (i in 1:length(colnames(raw_matrix))) {
  new_name<-substring(colnames(raw_matrix)[i], 9, 18)
  column_names<-c(column_names, new_name)
}
colnames(raw_matrix)<-column_names

###Read in the metadata to check the conditions

library(xlsx)


metadata <- read.xlsx("/nasdata/sinkala/phenodata/IPF/
updated_datasets/GSE166036_updated_ok.xlsx", sheetIndex = 1)

conditions<-(metadata$disease)


#######ets change the column names of the raw data i
```

```
into sample title########

colnames(raw_matrix) <- metadata$title


#######Filter the low counts########

filter_low_counts <- function(counts.matrix, conditions, method
= "cpm", normalized=FALSE, depth=NULL, cpm=1, p.adj = "fdr"){

  if(is.null(counts.matrix)){stop("Error: please provide a
  numeric count matrix!")}
  if(is.null(conditions)){stop("Error: please provide a factor
  or a vector indicating the conditions!")}
  if(!method %in% c("cpm", "wilcoxon", "proportion"))
  {stop("Error: Please type in one of valid methods!")}


  if (method=="cpm"){
    filtered.counts = NOISeq::filtered.data(counts.matrix,
    factor = conditions, norm = normalized, method = 1,
    cv.cutoff = 100, cpm = cpm, p.adj = p.adj)

  }else if(method=="wilcoxon"){
    filtered.counts = NOISeq::filtered.data(counts.matrix,
    factor = conditions, norm = normalized, method = 2,
    cv.cutoff = 100, p.adj = p.adj)

  }else if(method=="proportion"){

    if(is.null(depth)){stop("Error: indicate a numeric vector
    indicating per sample library depths")}
    if(!class(depth)=="numeric"){stop("Error: please provide
    the depth argument with a numeric vector!")}

    ### Compute librarary depth

    filtered.counts = NOISeq::filtered.data(counts.matrix,
    factor = conditions, norm = normalized, depth = depth,
    method = 3, cv.cutoff = 100, cpm = cpm, p.adj = p.adj)
  }
  return(filtered.counts)
}
filtered_data <- filter_low_counts(counts.matrix =raw_matrix,
conditions = conditions, method = "proportion", normalized =
FALSE, p.adj = "fdr", depth = as.numeric(apply(raw_matrix, 2,
sum)))

write.table(filtered_data, file =
"filtered_data_GSE166036_ensembl.csv", sep = "\t", col.names
=TRUE)

##########Change the ensembl_ids to symbol############

library(org.Hs.eg.db)

ensembl_ids = rownames(filtered_data)


# You can see that the subset data is listed as a integer, we
now need to convert
# this to a vector to pass it into the annotation mapping

ensembl_ids = as.vector(ensembl_ids)

# Using the org.Hs.eg.db we set up mapping info - if you look
at the documentation you
# can also obtain other keytypes

gene_ids <- select(org.Hs.eg.db, keys=ensembl_ids,
                   columns="SYMBOL", keytype="ENSEMBL")


filtered_data$ENSEMBL<-rownames(filtered_data)

matrix_with_gene_id <- merge(filtered_data, gene_ids,
by.x="ENSEMBL", by.y="ENSEMBL")


##Remove rows with NA######

matrix_with_gene_id <- na.omit(matrix_with_gene_id)
matrix_with_gene_id_1<-matrix_with_gene_id[,
2:length(colnames(matrix_with_gene_id))] # REMOVE THE ENSEMBL
COLUMN

aggr_exprmat <- matrix_with_gene_id_1 %>% group_by(SYMBOL) %>%
dplyr::summarise_all(.funs = c(median = "median"))
aggr_exprmat <- as.data.frame(aggr_exprmat)
aggr_exprmat <- na.omit(aggr_exprmat)
```

```
rownames(aggr_exprmat)<-aggr_exprmat$SYMBOL
aggr_exprmat<-aggr_exprmat[,2:length(colnames(aggr_exprmat))]

colnames(aggr_exprmat) <- metadata$title

write.table(aggr_exprmat, file =
"filtered_data_GSE166036_symbol.csv", sep = "\t", col.names
=TRUE)



###PCA before normalisation###
df_pca <- prcomp(aggr_exprmat, center = TRUE, scale. = TRUE)
df_out <- as.data.frame(df_pca$rotation)
df_out$group <- conditions

p<-ggplot(df_out, aes(x=PC1, y=PC2, color=group,
label=rownames(df_out)))
p <- p+geom_point()+geom_text(size=3)
p

### Differential expression analysis ###

setwd("/path")
table<-read.table("normalized_count_matrix.csv", sep = "\t")
metadata <- read.xlsx("/path_to_metadata", sheetIndex = 1)



#######extract ipf and healthy samples###########
####and covariates####

table<-table[,metadata$disease%in%c("IPF","healthy")]

metadata<-metadata[metadata$disease%in%c("IPF","healthy"),]

table<-table[,metadata$disease%in%c("healthy")[metadata
$tissue_source%in%c("alveolar_septae")]+metadata$tissue_source
%in%c("fibroblast_foci")==1]

metadata<-metadata[metadata$disease%in%c("healthy")[metadata
$tissue_source%in%c("alveolar_septae")]+metadata$tissue_source
%in%c("fibroblast_foci")==1,]

conditions<-metadata$disease

tissue<-metadata$tissue_source
```

```
muc5b_genotype<-metadata$muc5b_genotype[metadata$disease%in
%c("IPF","healthy")]

smoking_status<-metadata$smoking_status[metadata$disease%in
%c("IPF","healthy")]

plate<-metadata$plate[metadata$disease%in%c("IPF","healthy")]

sex<-metadata$sex[metadata$disease%in%c("IPF","healthy")]

batch<-metadata$batch[metadata$disease%in%c("IPF","healthy")]

batch<-gsub("-", "_", batch)

race<-metadata$race[metadata$disease%in%c("IPF","healthy")]

immunosuppressant<-metadata$immunosupressant[metadata$disease
%in%c("IPF","healthy")]

##############################################################

table_1<-sapply(table,as.integer)

rownames(table_1)<-rownames(table)

conditions<-(metadata$disease)


colData <- data.frame(treatment=as.vector(conditions),
tissue=as.vector(tissue), sex=as.vector(sex))
rownames(colData) <- colnames(table_1)

table_1<-sapply(table,as.integer)

rownames(table_1)<-rownames(table)



ddsMat <- DESeq2::DESeqDataSetFromMatrix(countData = table_1,
                                         colData = colData,
                                         design = ~sex
                                         +treatment)

dds <- DESeq2::DESeq(ddsMat)
```

```
total_norm_counts <- DESeq2::counts(dds, normalized=TRUE)
write.table(total_norm_counts, file =
"GSE173355_normalized_counts_matrix_symbol_deseq.txt", quote =
FALSE, sep = "\t", row.names = TRUE, col=NA)

res1 <- DESeq2::results(dds, contrast = c("treatment", "IPF",
"healthy"), pAdjustMethod = "fdr", independentFiltering =
FALSE)

write.table(res1, file =
"DEG_results_DESeq2_IPF_vs_healthy_unfiltered_symbol_GSE173355.txt",
row.names = TRUE, sep = "\t", quote = FALSE)

res1_adj <- res1[which(res1$padj<=0.01 &
abs(res1$log2FoldChange)>=0.58),]
print(dim(res1_adj))

write.table(res1_adj, file =
"DEG_results_DESeq2_IPF_vs_healthy_filtered_symbol_GSE173355.txt",
row.names = TRUE, sep = "\t", quote = FALSE)

###########Microarraydatasets_probes_to_geneid################

setwd("/path")

library(xlsx)


dif_table_ipf <-
read.xlsx("ALL_Differential_Expression_Tables_2023-01-24.xlsx",
sheetIndex = 2)

expression_matrix<-
read.table("Expression_Matrix_Normalized_2023-01-24.txt",
sep="\t", header=T, row.names = NULL)


require("biomaRt")
mart <- useMart("ENSEMBL_MART_ENSEMBL")
mart <- useDataset("hsapiens_gene_ensembl", mart)

annotLookup_ipf <- getBM(
  mart=mart,
  attributes=c(
    "affy_hg_u133a",
    "ensembl_gene_id",
    "gene_biotype",
    "external_gene_name"),
  filter = "affy_hg_u133a",
  values = dif_table_ipf$ID,
  uniqueRows = TRUE)


#annotation table
table_annot_ipf <- merge(dif_table_ipf, annotLookup_ipf,
by.x="ID", by.y="affy_hg_u133a")

#Making the annotation table in ensembl id:s
table_annot_ipf_ensembl<-table_annot_ipf[,c(2:9)]
aggr_dif_table_ipf <- table_annot_ipf_ensembl %>%
group_by(ensembl_gene_id) %>% dplyr::summarise_all(.funs =
c(median = "median"))
aggr_dif_table_ipf <- as.data.frame(aggr_dif_table_ipf)
aggr_dif_table_ipf  <- na.omit(aggr_dif_table_ipf )
rownames(aggr_dif_table_ipf)<-aggr_dif_table_ipf
$ensembl_gene_id
aggr_dif_table_ipf<-aggr_dif_table_ipf[,
2:length(colnames(aggr_dif_table_ipf))]

write.table(aggr_dif_table_ipf, file =
"dif_table_ensembl_ipf_GSE11196.csv", sep = "\t", col.names
=TRUE)


#Make the annotation table with gene symbols

table_annot_ipf_symbol<-table_annot_ipf[,c(2:8, 11)]
aggr_dif_table_ipf <- table_annot_ipf_symbol %>%
group_by(external_gene_name) %>% dplyr::summarise_all(.funs =
c(median = "median"))
aggr_dif_table_ipf <- as.data.frame(aggr_dif_table_ipf)
aggr_dif_table_ipf  <- na.omit(aggr_dif_table_ipf)
rownames(aggr_dif_table_ipf)<-aggr_dif_table_ipf
$external_gene_name
aggr_dif_table_ipf<-aggr_dif_table_ipf[,
2:length(colnames(aggr_dif_table_ipf))]

write.table(aggr_dif_table_ipf, file =
"dif_table_symbol_ipf_GSE11196.csv", sep = "\t", col.names
=TRUE)
```

```
#################Expression matrix with
ensembl#################

expression_matrix<-expression_matrix[grep("ENSG",
expression_matrix$row.names),]


rownames_ipf<-c()
for (i in 1:(length(expression_matrix$row.names))) {
  rowname<-strsplit(as.character(expression_matrix
  $row.names[i]), "_")[[1]][1]
  rownames_ipf<-c(rownames_ipf, rowname)
}

rownames(expression_matrix)<-rownames_ipf


expression_matrix<-expression_matrix[,
2:length(colnames(expression_matrix))]


write.table(expression_matrix, file =
"expression_matrix_final_ensembl_GSE11196.csv", sep = "\t",
col.names =TRUE)

########### Expression matrix with GENE-IDS############

library(org.Hs.eg.db)

ensembl_ids = rownames(expression_matrix)


# You can see that the subset data is listed as a integer, we
now need to convert
# this to a vector to pass it into the annotation mapping

ensembl_ids = as.vector(ensembl_ids)

# Using the org.Hs.eg.db we set up mapping info - if you look
at the documentation you
# can also obtain other keytypes

gene_ids <- select(org.Hs.eg.db, keys=ensembl_ids,
                    columns="SYMBOL", keytype="ENSEMBL")


expression_matrix$ENSEMBL<-rownames(expression_matrix)

table_with_gene_id <- merge(expression_matrix, gene_ids,
by.x="ENSEMBL", by.y="ENSEMBL")

table_with_gene_id <- na.omit(table_with_gene_id)



table_with_gene_id_1<-table_with_gene_id[,
2:length(colnames(table_with_gene_id))] # REMOVE THE ENTREZ
COLUMN

aggr_table<- table_with_gene_id_1 %>% group_by(SYMBOL) %>%
dplyr::summarise_all(.funs = c(median = "median"))
aggr_table <- as.data.frame(aggr_table)
aggr_table <- na.omit(aggr_table)
rownames(aggr_table)<-aggr_table$SYMBOL
aggr_table<-aggr_table[,2:length(colnames(aggr_table))]

colnames(aggr_table)<-colnames(expression_matrix)[1:
(length(colnames(expression_matrix))-1)]



write.table(aggr_table, file =
"expression_matrix_final_symbol_GSE11196.csv", sep = "\t",
col.names =TRUE)

#########make the consensus count matrices############
########This was done for RNA-seq and microarray
###Samples separately and for each cell type
###separately

setwd("/path")



GSE185492 <-
read.table("GSE185492_ALL_normalized_counts_matrix_symbol_deseq.txt",
sep="\t", header = T)
rownames(GSE185492)<-GSE185492$X
GSE185492<-GSE185492[,2:length(colnames(GSE185492))]
```

```
GSE44723<-
read.table("expression_matrix_final_symbol_GSE44723.csv",
sep="\t")

GSE40839 <-
read.table("expression_matrix_final_symbol_GSE40839.csv",
sep="\t")

library(purrr)
common_row_names <- Reduce(intersect,
list(rownames(GSE40839),rownames(GSE11196),
rownames(GSE44723)))

GSE11196_subset <- GSE11196[rownames(GSE11196) %in%
common_row_names, ]

GSE44723_subset <- GSE44723[rownames(GSE44723) %in%
common_row_names, ]

GSE40839_subset <- GSE40839[rownames(GSE40839) %in%
common_row_names, ]

write.table(GSE11196_subset,
file="GSE11196_symbol_expression_matrix_subset.txt", sep="\t")

write.table(GSE44723_subset,
file="GSE44723_symbol_expression_matrix_subset.txt", sep="\t")

write.table(GSE40839_subset,
file="GSE40839_symbol_expression_matrix_subset.txt", sep="\t")

####Disease and healthy

disease_GSE185492_subset<-GSE185492[,c(grep(pattern = "IPF",
colnames(GSE185492)))]

healthy_GSE185492_subset<-GSE185492[,c(grep(pattern = "CTR",
colnames(GSE185492)))]

disease_GSE44723_subset<- GSE44723_subset[,c(3,4,5,7)]

disease_GSE44723_subset<-
GSE44723_subset[,c(1,2,6,8,9,10,11,12,13,14)]

healthy_GSE40839_subset<-GSE40839_subset[,c(1:10)]
```

```
disease_GSE40839_subset<-GSE40839_subset[c(19:21)]

write.table(disease_GSE185492_subset,
file="GSE185492_symbol_expression_matrix_disease_subset.txt",
sep="\t")


write.table(healthy_GSE185492_subset,
file="GSE185492_symbol_expression_matrix_healthy_subset.txt",
sep="\t")

######################################################################
#Integrating the differential expression tables for the meta-analysis#
######################################################################

setwd("path")


GSE185492 <-
read.table
("DEG_results_DESeq2_IPF_vs_healthy_unfiltered_symbol_GSE185492_ALL.txt",
header = T)

GSE185492_adj_p<-GSE185492$padj
names(GSE185492_adj_p)<-rownames(GSE185492)

GSE44723<- read.table("dif_table_final_symbol_GSE44723.csv", sep="\t",
header = T)

GSE44723_adj_p<-GSE44723$adj.P.Val
names(GSE44723_adj_p)<-rownames(GSE44723)

GSE11196<- read.table("dif_table_ipf_final_symbol_GSE11196.csv", sep="\t",
header = T)

GSE11196_adj_p<-GSE11196$adj.P.Val
names(GSE11196_adj_p)<-rownames(GSE11196)

GSE40839<- read.table("dif_table_ipf_final_symbol_GSE40839.csv", sep="\t",
header = T)

GSE40839_adj_p<-GSE40839$adj.P.Val
names(GSE40839_adj_p)<-rownames(GSE40839)
```

```
library(purrr)

common_names <- Reduce(intersect,
list(names(GSE185492_adj_p),names(GSE44723_adj_p), names(GSE11196_adj_p),
names(GSE40839_adj_p)))

GSE185492_adj_p_subset <- GSE185492_adj_p[names(GSE185492_adj_p) %in%
common_names]
GSE185492_adj_p_subset<-
GSE185492_adj_p_subset[order(names(GSE185492_adj_p_subset))]

GSE44723_adj_p_subset <-GSE44723_adj_p[names(GSE44723_adj_p) %in%
common_names]
GSE44723_adj_p_subset<-

GSE44723_adj_p_subset[order(names(GSE44723_adj_p_subset))]

GSE11196_adj_p_subset <-GSE11196_adj_p[names(GSE11196_adj_p) %in%
common_names]

GSE11196_adj_p_subset<-
GSE11196_adj_p_subset[order(names(GSE11196_adj_p_subset))]

GSE40839_adj_p_subset <-GSE40839_adj_p[names(GSE40839_adj_p) %in%
common_names]

GSE40839_adj_p_subset<-
GSE40839_adj_p_subset[order(names(GSE40839_adj_p_subset))]

combined_table<-
as.data.frame(cbind(GSE185492_adj_p_subset,GSE44723_adj_p_subset,
GSE11196_adj_p_subset, GSE40839_adj_p_subset))
combined_table<-na.omit(combined_table)

write.table(combined_table, file="combined_adj_p_val_fibro.txt", sep="\t")

#Integrating the differential expression tables for the meta-
analysis#
###ALL###

setwd("C:/Users/OWNER/OneDrive - TUNI.fi/Bioteknologia/gradu/data/
meta_analysis/ALL")
```

```
#list_files<-list.files()
#list_files<-list_files[1:(length(list_files)-1)] #CHECK THE LIST_FILES
FIRST

##list_files for all but biopsy

list_files<-c("adj_pval_GSE151673_epithelial.txt",
"adj_pval_integrated_BAL.txt", "adj_pval_integrated_macrophage.txt",
"combined_adj_p_val_fibro.txt")

tables<-list()
for (name in 1:length(list_files)) {

  tables[[name]]<-read.table(list_files[name], sep="\t", header = T)

}



list_names<-list()

for(i in 1:length(tables)){
  list_names[[i]]<-rownames(tables[[i]])
}

common_names <- Reduce(intersect, list_names)


common_p_values<-list()

for(i in 1:length(tables)){

  common_p_values[[i]]<-tables[[i]][rownames(tables[[i]])%in
  %common_names,,drop=FALSE]
  common_p_values[[i]]<-common_p_values[[i]]
  [order(rownames(common_p_values[[i]])),,drop=FALSE]

}


names(common_p_values)<-list_files

combined_table<-as.data.frame(do.call(cbind, common_p_values))
```

```
names_1<-substring(colnames(combined_table), regexpr("GSE",
colnames(combined_table)) -0)

names_1[1]<-"GSE151673_adj_pval"

names_1[2]<-"GSE166036_adj_pval"

names_1[3]<-"GSE70866_adj_pval"

names_1[4]<-"GSE90010_adj_pval"

names_1[5]<-"GSE49072_adj_pval"

colnames(combined_table)<-names_1

setwd("C:/Users/OWNER/OneDrive - TUNI.fi/Bioteknologia/gradu/data/
meta_analysis/ALL_BUT_BIOPSY")

write.table(combined_table,
file="adj_pval_integrated_ALL_BUT_BIOPSY_FINAL.txt", sep="\t")


#########Meta-analysis###############
rm(list=ls())

#### Required libraries ####
suppressMessages(library(esc))
suppressMessages(library(metafor))
suppressMessages(library(metap))
suppressMessages(library(TopKLists))
suppressMessages(library(RankProd))
suppressMessages(library(matrixStats))
suppressMessages(library(igraph))
suppressMessages(library(TopKLists))
suppressMessages(library(minet))
suppressMessages(library(foreach))
suppressMessages(library(parallel))
suppressMessages(library(doParallel))
suppressMessages(library(ggplot2))
suppressMessages(library(plyr))

######## Module 1 - Meta-analysis section ########
##################################################
#This section was mostly offered by Antonio
#Federico, Thank you!
```

```
#' Mean-adjustes transcriptomics data by batch (wrapper of
pamr.batchadjust from the pamr CRAN package)
#'
#' @importFrom pamr pamr.batchadjust
#'
#' @param expr_mat A dataframe with genes on the rows and
samples in the columns.
#' @param samples_label A factor of samples labels in the same
order as the samples in expr_mat columns
#' @param batch_labels A factor of labels indicating the
batches (the studies where the samples are coming from)
#' @return A batch-adjusted expression matrix of the same
dimension of expr_mat
#' @examples
#' \dontrun {
#' calc_effect_size_rank(meta_dataframe)
#' }
#' @export
multi_studies_adjust <-
function(expr_mat, samples_label, batch_labels){

  mylist <- list(x=as.matrix(expr_mat),
  y=as.factor(samples_label),
  batchlabels=as.factor(batch_labels))

  adjusted_mat <- pamr::pamr.batchadjust(data = mylist)

  table_transpose<-as.data.frame(t(expr_mat))

  df_pca <- prcomp(table_transpose, center = TRUE,
  scale. = TRUE)

  p<-ggplot(as.data.frame(df_pca$x),
  aes(x=PC1, y=PC2, color=batch_labels))
  p <- p+geom_point()+ggtitle("Before batch correction")

  table_transpose2<-as.data.frame(t(adjusted_mat$x))
  df_pca2 <- prcomp(table_transpose2, center = TRUE,
  scale. = TRUE)

  p2<-ggplot(as.data.frame(df_pca2$x), aes(x=PC1, y=PC2,
   color=batch_labels))
  p2 <- p2+geom_point()+ggtitle("After batch correction")

  require(gridExtra)
  grid.arrange(p, p2, ncol=2)
  return(adjusted_mat)
}


#' Computes effect sizes for meta-analysis
#'
#' @importFrom esc effect_sizes
#'
#' @param meta_dataframe A dataframe with genes on the rows (as
rownames) and samples in the columns. The columns should
contain p-values deriving from gene-based statistical testing.
#' @return A gene list ranked on the base of the effect size.
#' @examples
#' \dontrun {
#' calc_effect_size_rank(meta_dataframe)
#' }
#' @export
calc_effect_size_rank<- function(meta_dataframe) {
  #index_pval<-which(grepl('pval',colnames(meta_dataframe)))
  #meta_data_pval<-meta_dataframe[,index_pval]
  #colnames(meta_data_pval)<-colnames(meta_dataframe)
  [index_pval]
  #heatmap(as.matrix(meta_data_pval))
  tmp_data<-meta_dataframe
  row.names(tmp_data)<-NULL
  pval<-as.vector(rowMeans(tmp_data))
  tmp <- data.frame(
    pval = pval,
    n =rep(ncol(meta_dataframe),length(pval)),
    studyname = rownames(meta_dataframe)
  )
  effect_sizes_values<-esc::effect_sizes(tmp, p = pval, totaln
  = n, study = studyname, fun = "chisq")
  #check is study label is kept, in case add
  colnames(meta_data_log)
  effect_size_rank<-effect_sizes_values[,c('study','es')]
  rownames(effect_size_rank)<-effect_size_rank$study
  effect_size_rank$study<-NULL
  #ranked final list
  effect_size_rank<-effect_size_rank[order(-effect_size_rank
  $es), , drop = FALSE]
  return(effect_size_rank)
}

#' Computes pvalue-based rank for meta-analysis
#'
#' @importFrom metap sumlog
#'
#' @param meta_dataframe A dataframe with genes on the rows (as
rownames) and samples in the columns. The columns should
contain p-values deriving from gene-based statistical testing.
#' @return A gene list ranked on the base of p-values.
#' @examples
#' \dontrun{
#' calc_pvalue_based_rank(meta_dataframe)
#' }
#' @export
calc_pvalue_based_rank<-function(meta_dataframe){
  #retrieve data
  #index_pval<-which(grepl('pval',colnames(meta_dataframe)))
  #meta_data_pval<-as.data.frame(meta_dataframe[,index_pval])
  #colnames(meta_data_pval)<-colnames(meta_dataframe)
  [index_pval]
  #for each gene combine the p-values by the sum of logs method
  fisher_based_res<-list()
  for(i in 1:length(rownames(meta_dataframe))){
    metap<-metap::sumlog(meta_dataframe[i,])
    fisher_based_res[[i]]<-metap$p
  }
  fisher_based_pvalues<-as.data.frame(fisher_based_res)
  colnames(fisher_based_pvalues)<-rownames(meta_dataframe)
  fisher_based_pvalues<-t(fisher_based_pvalues)
  fisher_based_pvalues<-as.data.frame(fisher_based_pvalues)
  colnames(fisher_based_pvalues)<-"pValue"
  #ranked final list
  fisher_based_pvalues<-
  fisher_based_pvalues[order(fisher_based_pvalues$pValue), ,
  drop = FALSE]
  return(fisher_based_pvalues)
}

#' Computes Rank Product-based rank for meta-analysis
#'
#' @importFrom RankProd RP.advance
#'
#' @param meta_dataframe A dataframe with genes on the rows (as
rownames) and samples in the columns. The columns should
contain adjusted p-values deriving from gene-based statistical
testing.
#' @param class a vector containing the class labels of the
samples. In the two class unpaired case, the label of a sample
is either 0 (e.g., control group) or 1 (e.g., case group). For
one class data, the label for each sample should be 1.
#' @param origin a vector containing the origin labels of the
samples.
#' @return A gene list ranked on the base of adjusted p-values.
#' @examples
#' \dontrun{
#' calc_rank_base_rank(meta_dataframe, class = o, origin = o)
#' }
#' @export
calc_rank_base_rank <- function(meta_dataframe, class, origin){
  index_pval_adj <-
  which(grepl('_adj_pval',colnames(meta_dataframe)))
  meta_data_adpval <-
  as.data.frame(meta_dataframe[,index_pval_adj])
  colnames(meta_data_adpval) <- colnames(meta_dataframe)
  [index_pval_adj]
  cl <- rep.int(1,times = length(colnames(meta_data_adpval)))
  rp.advance.input <- meta_data_adpval
  colnames(rp.advance.input) <- NULL
  rownames(rp.advance.input) <- NULL
  rp.advance.input <- as.matrix(rp.advance.input)
  #origin contains the  labels for different studies
  #origin <- gsub(pattern =
  "_adj_pval",colnames(meta_data_adpval),replacement = "")
  #origin <- gsub("\\_.*","",origin)
  o<- rep(1,dim(meta_data_adpval)[2])
  RP_advance_out <- RankProd::RP.advance(data =
  meta_data_adpval, cl = class, origin = o, calculateProduct
  =T)
  #ranked  final list
  ranks_based_pvalues<-as.data.frame(RP_advance_out$pval)
  rownames(ranks_based_pvalues)<-rownames(meta_data_adpval)
  rank_1<-
  rownames(ranks_based_pvalues[order(abs(ranks_based_pvalues
  $`class1 < class2`)), , drop = FALSE])
  rank_2<-
  rownames(ranks_based_pvalues[order(abs(ranks_based_pvalues
  $`class1 > class2`)), , drop = FALSE])
  borda_list<-list()
  borda_list[[1]]<-rank_1
  borda_list[[2]]<-rank_2
  outputBorda<-Borda(borda_list)
  output_borda<-outputBorda$TopK$mean
  output_borda<-as.data.frame(output_borda)
  rownames(output_borda)<-output_borda$output_borda
  return(output_borda)
}

#' To be compiled
#'
#' @importFrom RankProd RP.advance
#'
#' @param meta_dataframe A dataframe with genes on the rows (as
rownames) and samples in the columns. The columns should
contain adjusted p-values deriving from gene-based statistical
testing.
#' @param class a vector containing the class labels of the
samples. In the two class unpaired case, the label of a sample
is either 0 (e.g., control group) or 1 (e.g., case group). For
one class data, the label for each sample should be 1.
#' @param origin a vector containing the origin labels of the
samples.
#' @return A gene list ranked on the base of
#' @examples
#' \dontrun {
#' calc_rank_base_rank_subsets(meta_dataframe)
#' }
#' @export
calc_rank_base_rank_subsets<-function(meta_dataframe, class,
origin){
  index_pval_adj<-
  which(grepl('_adj_pval',colnames(meta_dataframe)))
  meta_data_adpval<-
  as.data.frame(meta_dataframe[,index_pval_adj])
  colnames(meta_data_adpval)<-colnames(meta_dataframe)
  [index_pval_adj]
  ranks<-list()
  for(i in 1:10){
    data<-meta_data_adpval
    index_col<-sample(colnames(data),size = 5)
    data_partition<-data[,index_col]
    ranks_product<-calc_rank_base_rank(data_partition, class =
    class, origin = origin)
    ranks[[i]]<-ranks_product
  }
  class1<-data.frame()
  class2<-data.frame()
  for(j in 1:length(ranks)){
    if(plyr::empty(class1)){
      class1<-as.data.frame(ranks[[j]][,1])
    } else{
      class1<-qpcR:::cbind.na(class1,ranks[[j]][,1])
    }
    if(plyr::empty(class2)){
      class2<-as.data.frame(ranks[[j]][,2])
    }
    else{
      class2<-qpcR:::cbind.na(class2,ranks[[j]][,2])
    }
  }
  ranks_based_pvalues<-data.frame("class1 < class2"=
  rowMeans(class1),
                                "class1 > class2" =
                                rowMeans(class2))
  rownames(ranks_based_pvalues)<-rownames(meta_data_adpval)
  ranks_based_pvalues<-
  ranks_based_pvalues[order(abs(ranks_based_pvalues
  $class1...class2),abs(ranks_based_pvalues
  $class1...class2.1)), , drop = FALSE]
  return(ranks_based_pvalues)
}


#' Computes a gene rank based on an ensembl of metanalysis methods,
including effect size, p-value and rank product.
#'
#' @importFrom RankProd RP.advance
#' @importFrom metap sumlog
#' @importFrom esc effect_sizes
#'
#' @param meta_dataframe A dataframe with genes on the rows (as rownames)
and samples in the columns. The columns should contain adjusted p-values
deriving from gene-based statistical testing.
#' @param method Statistical method(s) to be included in the ensembl
metanalysis.
#' @param class a vector containing the class labels of the samples. In
the two class unpaired case, the label of a sample is either 0 (e.g.,
control group) or 1 (e.g., case group). For one class data, the label for
each sample should be 1.
```

```
#' @param origin a vector containing the origin labels of the samples.
#' @param metric One statistical metric between "median" and "mean".
#' @return A gene list ranked on the base of the methods chosen for the
metanalysis.
#' @examples
#' \dontrun {
#' run_ensembl_metanalysis(meta_dataframe)
#' }
#' @export
run_ensembl_metanalysis <- function(meta_dataframe,
method=c("effect_size", "pvalue", "rank_product"), class, origin,
metric="median"){
  if (length(method)<2){stop("Error: choose at least two methods among
  effect size, pvalue and rank product!")}
  if (method==c("effect_size", "pvalue", "rank_product")){
    es <- calc_effect_size_rank(meta_dataframe = meta_dataframe)
    pval <- calc_pvalue_based_rank(meta_dataframe = meta_dataframe)
    rankprod <- calc_rank_base_rank_subsets(meta_dataframe = metadf, class
    = class, origin = origin)
    data<-list()
    data[[1]]<-rownames(es)
    data[[2]]<-rownames(pval)
    data[[3]]<-rownames(rankprod)
    names(data)<-c('Effect_size','Fisher_test','Rank_Prod')
    outputBorda<-TopKLists::Borda(data)
    if(metric=="median"){
      final_ranked_genes_median<-as.data.frame(outputBorda$TopK$median)
    }else if(metric=="mean"){
      final_ranked_genes_mean<-as.data.frame(outputBorda$TopK$mean)
    }
    return(final_ranked_genes_median)
  }else if(method==c("effect_size", "pvalue")){
    es <- calc_effect_size_rank(meta_dataframe = meta_dataframe)
    pval <- calc_pvalue_based_rank(meta_dataframe = meta_dataframe)
    data<-list()
    data[[1]]<-rownames(es)
    data[[2]]<-rownames(pval)
    names(data)<-c('Effect_size','Fisher_test')
    outputBorda<-TopKLists::Borda(data)
  }else if(method==c("effect_size","rank_product")){
    es <- calc_effect_size_rank(meta_dataframe = meta_dataframe)
    rankprod <- calc_rank_base_rank_subsets(meta_dataframe = metadf, class
    = class, origin = origin)
    data<-list()
    data[[1]]<-rownames(es)
    data[[2]]<-rownames(rankprod)
    names(data)<-c('Effect_size','Rank_Prod')
    outputBorda<-TopKLists::Borda(data)
  }else if(method==c("pval","rank_product")){
    pval <- calc_pvalue_based_rank(meta_dataframe = meta_dataframe)
    rankprod <- calc_rank_base_rank_subsets(meta_dataframe = metadf, class
    = class, origin = origin)
    data<-list()
    data[[1]]<-rownames(pval)
    data[[2]]<-rownames(rankprod)
    names(data)<-c('Fisher_test','Rank_Prod')
    outputBorda<-TopKLists::Borda(data)
  }
  if(metric=="median"){
    final_ranked_genes_median<-as.data.frame(outputBorda$TopK$median)
  }else if(metric=="mean"){
    final_ranked_genes_mean<-as.data.frame(outputBorda$TopK$mean)
  }
  return(final_ranked_genes_median)
}




######## Module 2 - Feature selection section ########
######################################################

# GO <- "c5.go.v7.2.symbols.gmt"
# REACTOME <- "c2.cp.reactome.v7.2.symbols.gmt"
# KEGG <- "c2.cp.kegg.v7.2.symbols.gmt"
# WIKI <- "c2.cp.wikipathways.v7.2.symbols.gmt"
# MSIGDB <- "msigdb.v7.2.symbols.gmt"
# TRANSCRIPTION_FACTORS<-"c3.tft.v7.2.symbols.gmt"

# myGO <- fgsea::gmtPathways(GO)
# myREACTOME <- fgsea::gmtPathways(REACTOME)
# myKEGG <- fgsea::gmtPathways(KEGG)
# myWIKI <- fgsea::gmtPathways(WIKI)
# myMSIGDB <- fgsea::gmtPathways(MSIGDB)
# myTS <- fgsea::gmtPathways(TRANSCRIPTION_FACTORS)

#' Computes a GSEA of a gene rank against a gene set (provided
through a GMT file).
#'
```

```
#' @importFrom fgsea gmtPathways
#' @importFrom fgsea fgsea
#' @importFrom fgsea plotGseaTable
#'
#' @param gene_list A ranked gene list.
#' @param gmt_file a vector containing the class labels of the
samples. In the two class unpaired case, the label of a sample
is either 0 (e.g., control group) or 1 (e.g., case group). For
one class data, the label for each sample should be 1.
#' @examples
#' \dontrun {
#' compute_gsea(gene_list, gmt_file =
"c2.cp.reactome.v7.2.symbols.gmt")
#' }
#' @export
compute_gsea <- function(gene_list, gmt_file){
  mypath <- fgsea::gmtPathways(gmt_file)
  fgRes <- fgsea::fgsea(pathways = mypath,
                        stats = gene_list,
                        minSize=10,
                        #maxSize=600,
                        nperm=10000)

  print(fgsea::plotGseaTable(mypath[topPathways], gene_list,
  fgRes,  gseaParam=0.05))

  return(fgRes)
  }



#' Computes the threshold of the gene rank based on the
enrichment score of the GSEA.
#'
#' @importFrom fgsea calcGseaStat
#'
#' @param gene_list A ranked gene list.
#' @param fgsea_res fgsea object deriving from fgsea results or
the compute_gsea function.
#' @param background Whole gene set object deriving from the
fgsea::gmtPathways(gmt_file) function.
#' @examples
#' \dontrun {
#' compute_gsea_thresh(gene_list, fgsea_res, background)
#' }
#' @export
```

```
compute_gsea_thresh <- function(geneList, fgsea_res,
background){
  gseaParam=1
  stats <- geneList
  fgsea_res <- fgsea_res[order(fgsea_res$pval, decreasing =
  FALSE),]
  max_vec <- c()
  sigpath <- which(fgsea_res$pval<0.05)
  for(i in 1:length(sigpath)){

    pathway <- background[[fgsea_res$pathway[i]]]
    rnk <- rank(-stats)
    ord <- order(rnk)
    statsAdj <- stats[ord]
    statsAdj <- sign(statsAdj) * (abs(statsAdj)^gseaParam)
    statsAdj <- statsAdj/max(abs(statsAdj))
    pathway <- unname(as.vector(na.omit(match(pathway,
    names(statsAdj)))))
    pathway <- sort(pathway)
    gseaRes <- fgsea::calcGseaStat(statsAdj, selectedStats =
    pathway, returnAllExtremes = TRUE)
    bottoms <- gseaRes$bottoms
    tops <- gseaRes$tops
    n <- length(statsAdj)
    xs <- as.vector(rbind(pathway - 1, pathway))
    ys <- as.vector(rbind(bottoms, tops))
    toPlot <- data.frame(x = c(0, xs, n + 1), y = c(0, ys, 0))
    #diff <- (max(tops) - min(bottoms))/8

    max_vec <- c(max_vec, which(names(geneList) %in%
    names(gseaRes$tops)[which(gseaRes$tops==max(gseaRes
    $tops))]))
  }

  return(max_vec)
}




#######Build the networks########

rm(list=ls())

setwd("/nasdata/sinkala/expressiondata/countdata/
adjusted_matrices")
```

```
source("/nasdata/afederico/INfORM_functions.R")

args = commandArgs(trailingOnly=TRUE)

file_path=args[1]


generatematrices=get_ranked_consensus_matrix(gx_table =
read.table(file_path, sep="\t"), iMethods = c("clr"),
                                             iEst = c("pearson"),
                                             iDisc=c("none"),
                                             ncores = 30,
                                             debug_output = TRUE,
                                             updateProgress =
                                             TRUE)



#Jaccard_similarity_index_based_
#intersect_network_between_microarray_and_rnaseq###

########I had the original modules stored in RData object###########

load("list_of_modules_no_intersect_1.RData")

modules_bal_microarray_disease<-list_of_modules[[1]]
modules_bal_rna_seq_disease<-list_of_modules[[3]]


####List of modules 1##############
members=igraph::membership(modules_bal_microarray_disease)
members_list <- list()

for(i in 1:length(modules_bal_microarray_disease)){
  members_list[[i]] <- names(which(members==i))
}
names(members_list) <- paste0("mod",
1:length(modules_bal_microarray_disease))


##########List of modules2##########
members=igraph::membership(modules_bal_rna_seq_disease)
members_list_rnaseq <- list()
for(i in 1:length(modules_bal_rna_seq_disease)){
  members_list_rnaseq[[i]] <- names(which(members==i))
}
names(members_list_rnaseq) <- paste0("mod",
1:length(modules_bal_rna_seq_disease))

###Similarity index matrix between the module lists##

similarity_matrix<-matrix(0,nrow = length(members_list),
ncol=length(members_list_rnaseq),dimnames =
list(names(members_list), names(members_list_rnaseq)))
for(i in 1:length(members_list)){
  for (z in 1:length(members_list_rnaseq)) {
    similarity_matrix[i,z]<-length(intersect(members_list[[i]],
    members_list_rnaseq[[z]]))/length(union(members_list[[i]],
    members_list_rnaseq[[z]]))

  }
}

#####Heatmap and summary of the similarity matrix####
###Just a visualization during the analysis
##Final heatmaps done with ComplexHeatmap

heatmap.2(similarity_matrix, trace = "none")

apply(similarity_matrix, 1, summary)

######Which modules above the threshold#####

which(similarity_matrix>0.1, arr.ind = TRUE)
############################################################

##Intersect between the most similar modules#########

intersect_vector<-c(intersect(members_list[[10]],
members_list_rnaseq[[34]]),intersect(members_list[[5]],
 members_list_rnaseq[[25]]), intersect(members_list[[12]],
 members_list_rnaseq[[17]]),intersect(members_list[[3]],
 members_list_rnaseq[[14]]), intersect(members_list[[2]],
 members_list_rnaseq[[14]]))

#####Load the original networks#####

network_microarray_disease<-readRDS("network_bal_microarray_disease.rds")
network_rnaseq_bal_disease<-readRDS("network_bal_rnaseq_disease.rds")
```

```
##Subgraphs based on the Intersect between the most similar modules##

subgraph_microarray_disease_bal<-
induced.subgraph(network_microarray_disease, vids= intersect_vector)

subgraph_rnaseq_disease_bal<-
induced.subgraph(network_rnaseq_bal_disease, vids= intersect_vector)

#########Intersect between the subgraphs#####

#intersect_bal_dis_jaccard<-
  igraph::intersection(subgraph_microarray_disease_bal,
  subgraph_rnaseq_disease_bal, keep.all.vertices = FALSE)
########################################################

##subgraph union#########

subgraph_union<-
igraph::union(subgraph_microarray_disease_bal,
subgraph_rnaseq_disease_bal)

#####################

########Making the annotations and the
#ranked gene lists of the subgraph union###########

edge_list<-as_edgelist(subgraph_union)

new_graph<-graph_from_edgelist(edge_list)
igraph::vertex_attr(new_graph, name="color") <- "lightgray"
igraph::vertex_attr(new_graph, name="highlightcolor") <- "darkgray"
igraph::edge_attr(new_graph, name="color") <- "lightgray"
igraph::edge_attr(new_graph, name="highlightcolor") <- "darkgray"

annotated_intersect_bal_disease_union<-annotate_iGraph(new_graph)

rank_list_attr=c("betweenness","cc","degree","closeness","eigenvector")

gene_list_union_intersect_bal_disease<-
get_ranked_gene_list(annotated_intersect_bal_disease_union,
rank_list_attr = rank_list_attr)

saveRDS(annotated_intersect_bal_disease_union,
 file="annotated_union_network_intersect_bal_disease.rds")
```

```
write.table(gene_list_union_intersect_bal_disease,
file="ranked_union_gene_list_intersect_bal_disease.txt", sep="\t")



#Parse ranked matrix and get bin_mat and edge_rank
# Get edge rank list and binary inference matrix from edge rank
matrix computed by get_ranked_consensus_matrix().
# parse_edge_rank_matrix parses the edge rank matrix created by
using the internal function get_ranked_consensus_matrix_matrix()
to get a ranked edge list and a binary matrix.

rankMat.parsed=parse_edge_rank_matrix(edge_rank_matrix =
generatematrices, edge_selection_strategy = "default",
                                      mat_weights = "rank", topN =
                                      10, debug_output = TRUE,
                                      updateProgress = TRUE)

conGraph <- get_iGraph(rankMat.parsed$bin_mat)
saveRDS(conGraph, file="network_biopsy_rnaseq_healthy.rds")

get_reactome_from_modules <- function(igraph_modules, geneID="SYMBOL",
pval_cutoff=0.05, outPath, layout="overall") {
  if (file.exists(outPath)){
    setwd(file.path(outPath))
  } else {
    dir.create(file.path(outPath))
    setwd(file.path(outPath))
  }

  cat("The files will be exported in ", getwd())
  members=igraph::membership(igraph_modules)
  if(layout=="overall"){
    sigpath.overall <- data.frame()
    sigpath <- c()
    #pdf(file = paste0(subDir1, "report_annotazione_funzionale.pdf"),
    paper = "a4" , height = 1600, width = 900)
    for(mod in 1:length(igraph_modules)){
      x <- NA
      x <- names(members[members==mod])
      if(length(x)>10) {
        eg = clusterProfiler::bitr(x, fromType=geneID,
         toType="ENTREZID", OrgDb="org.Hs.eg.db")
        print(head(eg))
        sigpath <- ReactomePA::enrichPathway(gene=eg$ENTREZID,
        pvalueCutoff=pval_cutoff, readable=T)
        sigpath <- as.data.frame(sigpath)
        print(head(sigpath))
        if (length(sigpath$ID)>0){
          sigpath$Module <- mod
          sigpath.overall <- rbind(sigpath.overall, sigpath)
          #write.csv(as.data.frame(sigpath),
          file = paste0("Significant_enriched_pathways_module_",
           mod, ".csv"),
          quote = FALSE, row.names = FALSE)
        }

      }

    }
    write.table(sigpath.overall,
    file = "Pathway_results_overall_disease_macro.txt", sep="\t")

  } else if (layout=="single") {
    sigpath.overall <- data.frame()
    sigpath <- c()
    #pdf(file = paste0(subDir1, "report_annotazione_funzionale.pdf"),
    paper = "a4" , height = 1600, width = 900)
    for(mod in 1:length(igraph_modules)){
      x <- NA
      x <- names(members[members==mod])
      if(length(x)>10) {
        eg = clusterProfiler::bitr(x, fromType=geneID,
        toType="ENTREZID", OrgDb="org.Hs.eg.db")
        print(head(eg))
        sigpath <- ReactomePA::enrichPathway(gene=eg$ENTREZID,
        pvalueCutoff=pval_cutoff, readable=T)
        sigpath <- as.data.frame(sigpath)
        print(head(sigpath))
        if (length(sigpath$ID)>0){
          sigpath$Module <- mod
          #sigpath.overall <- rbind(sigpath.overall, sigpath)
          write.csv(as.data.frame(sigpath),
          file = paste0("Significant_enriched_pathways_module_", mod,
           ".csv"),
          quote = FALSE, row.names = FALSE)
        }

      }
    }

  }

  require(ReactomePA)

  return(cat("Analysis completed!!!"))

}




get_bubbleplot_from_pathways <-
function(igraph_modules, geneID="SYMBOL") {

  lst <- list()
  members=igraph::membership(igraph_modules)
  for(mod in 51:length(igraph_modules)){
    x <- NA
    x <- names(members[members==mod])
    if(length(x)>10) {
      convgenes = clusterProfiler::
      bitr(x, fromType="SYMBOL", toType="ENTREZID", OrgDb="org.Hs.eg.db")
      print(head(convgenes))
      lst[[mod]] <- convgenes$ENTREZID
    }
  }

  names(lst) <- seq_along(lst)
  lst[sapply(lst, is.null)] <- NULL

  res <- clusterProfiler::compareCluster(lst, fun="enrichPathway")
  print(clusterProfiler::dotplot(res))
  return(res)

}


rm(list=ls())
setwd("/nasdata/sinkala/expressiondata/countdata/
adjusted_matrices/final_networks")

###########################Read in data###############################

list.files()
```

```
load("list_of_modules_no_intersect_1.RData")
names(list_of_modules)

open_targets<-readRDS("/home/antonio/final_opentargets_parsed.rds")

opecentrality_gene_list<-
read.table("ranked_genelist_macro_micro_disease.txt")
centrality_gene_list<-
centrality_gene_list$ranked_genelist_macro_micro_disease.txt
biopsy_disease<-list_of_modules_biopsy[[1]]
biopsy_healthy<-list_of_modules_biopsy[[2]]

######Get module sizes#####################

module_sizes_disease<-c()
for (i in 1:length(macro_disease)) {
  module_sizes_disease<-c(module_sizes_disease,
  length(macro_disease[[i]]))
}

module_sizes_healthy<-c()
for (i in 1:length(macro_healthy)) {
  module_sizes_healthy<-c(module_sizes_healthy,
  length(macro_healthy[[i]]))
}

####List of modules disease##############

members=igraph::membership(macro_disease)
members_list_disease <- list()

for(i in 1:length(macro_disease)){
  members_list_disease[[i]] <-
  names(which(members==i))
}
names(members_list_disease) <-
paste0("mod", 1:length(macro_disease))


##########List of modules2##########

#members=igraph::
membership(bal_healthy_modules)
members=igraph::membership(macro_healthy)
members_list_healthy<-list()
```

```
#members_list_healthy<- _healthy_modules_filtered
for(i in 1:length(macro_healthy)){
  members_list_healthy[[i]] <- names(which(members==i))
}
names(members_list_healthy) <-
paste0("mod", 1:length(macro_healthy))

#####remove the modules that are <10 genes########

module_indexes_disease<-which(module_sizes_disease>10)

module_indexes_healthy<-which(module_sizes_healthy>10)

members_list_disease<-members_list_disease[module_indexes_disease]

members_list_healthy<-members_list_healthy[module_indexes_healthy]

#####################################

similarity_matrix<-matrix(0,nrow =
length(members_list_disease), ncol=length(members_list_healthy),
dimnames = list(names(members_list_disease), names(members_list_healthy)))
for(i in 1:length(members_list_disease)){
  for (z in 1:length(members_list_healthy)) {
    similarity_matrix[i,z]<-
    length(intersect(members_list_disease[[i]],
     members_list_healthy[[z]]))/
     length(union(members_list_disease[[i]], members_list_healthy[[z]]))

  }
}

gplots::heatmap.2(similarity_matrix, trace = "none")

write.table(similarity_matrix, "similarity_matrix_modules_epi.txt",
 sep="\t")

jaccard_max<-apply(similarity_matrix,1,FUN =max)

dissimilar_modules <- order(jaccard_max)

########################################

##Download the enrichment analysis table and spearate the pathway genes
```

```
pathways<-read.table("Pathway_results_overall_disease_macro.txt",sep="\t")
head(pathways)

pathways_most_dissimilar_module<-pathways[pathways$Module==8,]

most_dissimilar_pathway<-
pathways_most_dissimilar_module[which.min
((as.numeric(as.character(pathways_most_dissimilar_module$p.adjust)))),]

pathway_genes<-most_dissimilar_pathway$geneID

pathway_genes<-
unlist(strsplit(as.character(pathway_genes), "/"))
######################################################################


#SORT THE DRUGS BY MODULE CENTRALITY AND PUT THE
#CENTRALITY RANK IN THE TABLE####

drugs<-open_targets[open_targets$dat.target.gene_info.symbol%in%
pathway_genes,]

drugs<- drugs[complete.cases(drugs$dat.drug.molecule_name),]

drugs_compressed <- distinct(drugs, drugs$dat.drug.molecule_name,
.keep_all = TRUE)





#######Sort the modules based on centrality##############

sorted_modules<-list()
for (i in 1:length(macro_disease)) {
  # Get the module's gene list
  module_genes <- macro_disease[[i]]
  centrality_gene_list<-as.vector(centrality_gene_list)
  # Boolean vector of the genes which module
  # genes are on the gene list and max of those
  module_centrality <- centrality_gene_list[centrality_gene_list
  %in% module_genes]
  sorted_modules[[i]]<-module_centrality
}
```

```
#####Make table with sorted genes of the module and the gene rank#####

most_dissimilar_module<-sorted_modules[[8]]

rank<-as.numeric(seq(1:length(most_dissimilar_module)))

module_rank<-cbind(most_dissimilar_module, rank)

module_rank<-as.data.frame(module_rank)

colnames(module_rank)<-c("gene_name", "rank")

# Extract gene names from most_dissimilar_module column

colnames(drugs_compressed)[1]<-"gene_name"

drugs_compressed_1<-merge(drugs_compressed, module_rank)

###Sort the drugs by the rank######################

gene_names <- module_rank$gene_name

# Find indices of matching genes in dat.target.gene_info.symbol column
gene_indices <- match(drugs_compressed_1$gene_name, gene_names)

# Reorder rows of drugs_compressed based on gene_indices
drugs_compressed_ordered <-
drugs_compressed_1[order(as.numeric(as.vector(drugs_compressed_1$rank))),]

# Filter rows to keep only those that match the gene_names

#ranks<-module_rank[module_rank$gene_name%in%
drugs_compressed_ordered$gene_name,]


drugs_compressed_ordered<-drugs_compressed_ordered[,c(1,2,4,8,9,11,12)]

colnames(drugs_compressed_ordered)<-c("gene", "target_info",
"disease_info", "molecule_type", "drug_phase",
 "drug", "gene_module_rank")

drugs_compressed_ordered<-drugs_compressed_ordered[,c(1,6,7,2,3,4,5)]

######################################################
####Make the networks for plotting#############
```

```
########Parse the networks################
rm(list=ls())
list.files()

centrality_gene_list<-
read.table("ranked_genelist_epithelial_rnaseq_disease.txt")

centrality_gene_list<-centrality_gene_list$x

network<-readRDS("annotated_network_epithelial_rnaseq_disease.rds")
network<-network[[1]]

list.files()
load("list_of_modules_epithelium_datasets.RData")
names(list_of_modules_epithelial)

modules<-list_of_modules_epithelial[[1]]

########Sort the modules

sorted_modules<-list()
for (i in 1:length(modules)) {
  # Get the module's gene list
  module_genes <- modules[[i]]
  centrality_gene_list<-as.vector(centrality_gene_list)
  # Boolean vector of the genes which module genes
   are on the gene list and max of those
  module_centrality <-
  centrality_gene_list[centrality_gene_list %in% module_genes]
  sorted_modules[[i]]<-module_centrality
}

ranked_genes_modules<-c()
for (i in 1:length(sorted_modules)) {
  if (length(sorted_modules[[i]])>50) {
    ranked_genes_modules<-c(ranked_genes_modules,
    sorted_modules[[i]][1:50])
  }
  else{
    ranked_genes_modules<-c(ranked_genes_modules,sorted_modules[[i]])
  }
}
```

```
  # Extract the subgraph containing only the
  #top 100 vertices of the current module
  subgraph <- induced_subgraph(network, ranked_genes_modules)
  plot(subgraph)

################################
netwok_parsed<-subgraph
dirpath<-"subgraph_epi_healthy.gml"
############################Plot the graphs##############################
write.graph(netwok_parsed, file=dirpath, format = "gml")

###########################################

###########Heatmaps_from_similarity_matrices##############
rm(list=ls())
#################
BiocManager::install("ComplexHeatmap")
install.packages("circlize")
install.packages("gridtext")
install.packages("viridis")
###################################3
library(ComplexHeatmap)
library(circlize)
library(gridtext)
library("viridis")

setwd("path")

similarity_matrix_bal<-read.table
("similarity_matrix_modules_epi.txt", sep="\t")

#Extract the modules with more than 10 genes

#similarity_matrix<-similarity_matrix_bal[1:9,]

column_title_map = gt_render(
  paste0("<span style='font-size:25pt;
  color:black'>Jaccard similarity index of disease and
  healthy modules in epithelial samples</span><br>",
         "<br>",
         "<span style='font-size:20pt; color:black'> Healthy modules"))

row_title_map <-gt_render
(paste0("<span style='font-size:20pt; color:black'> Disease modules"))
```

```
ComplexHeatmap::
Heatmap(similarity_matrix_bal,column_title =
column_title_map, row_title = row_title_map,
                        show_row_names = T,
                        show_column_names = T,
                        heatmap_legend_param = list(title =
                        "Jaccard similarity index", labels_gp =
                        gpar(fontsize = 11)),
                        , col = rev(inferno(10)))
```

ESPERANTO vocabulary

The ESPERANTO vocabulary was compiled using various test versions of the application, resulting in a less refined vocabulary compared to the final version of the app. As an example, there is significant repetition in the vocabulary. Esperanto vocabulary represented in Table 32.

Table 32: Esperanto vocabulary

label	lab_syn	allowed_features	syn_features
age	age (years)	#PUUTTUU!	#PUUTTUU!
age	birthdate	#PUUTTUU!	#PUUTTUU!
age	birth	#PUUTTUU!	#PUUTTUU!
age	age.ch1	#PUUTTUU!	#PUUTTUU!
antigen_identified	#PUUTTUU!	NO	#PUUTTUU!
antigen_identified	#PUUTTUU!	YES	#PUUTTUU!
batch	batch.ch1	#PUUTTUU!	#PUUTTUU!
cell_jamming	jammed_unjammed.ch1	jammed	jammed
cell_jamming	jammed_unjammed.ch1	unjammed	unjammed
cell_line_treatment	cell_line_treatment	AM_from_IPF	AM from IPF
cell_line_treatment	cell.line.ch1	AM_from_IPF	AM from IPF
cell_line_treatment	source_name_ch1	AM_from_IPF	AM from IPF
cell_line_treatment	cell.line.ch1	AM_from_RB_ILD	AM from RB-ILD
cell_line_treatment	cell_line_treatment	AM_from_RB_ILD	AM from RB-ILD
cell_line_treatment	source_name_ch1	AM_from_RB_ILD	AM from RB-ILD
cell_line_treatment	cell.line.ch1	control_1	#PUUTTUU!
cell_line_treatment	source_name_ch1	control_1	#PUUTTUU!
cell_line_treatment	cell_line_treatment	control_1	#PUUTTUU!
cell_line_treatment	cell_line_treatment	control_lung_fibroblast_control_ECM	#PUUTTUU!
cell_line_treatment	cell.line.ch1	control_lung_fibroblast_control_ECM	#PUUTTUU!
cell_line_treatment	source_name_ch1	control_lung_fibroblast_control_ECM	#PUUTTUU!
cell_line_treatment	source_name_ch1	control_lung_fibroblast_IPF_ECM	#PUUTTUU!
cell_line_treatment	cell.line.ch1	control_lung_fibroblast_IPF_ECM	#PUUTTUU!
cell_line_treatment	cell_line_treatment	control_lung_fibroblast_IPF_ECM	#PUUTTUU!
cell_line_treatment	cell_line_treatment	IPF_lung fibroblast_control_ECM	#PUUTTUU!
cell_line_treatment	cell.line.ch1	IPF_lung fibroblast_control_ECM	#PUUTTUU!
cell_line_treatment	source_name_ch1	IPF_lung fibroblast_control_ECM	#PUUTTUU!
cell_line_treatment	cell_line_treatment	IPF_lung_fibroblast_IPF_ECM	#PUUTTUU!
cell_line_treatment	source_name_ch1	IPF_lung_fibroblast_IPF_ECM	#PUUTTUU!
cell_line_treatment	cell.line.ch1	IPF_lung_fibroblast_IPF_ECM	#PUUTTUU!

Continued on next page

Continued from previous page

label	lab_syn	allowed_features	syn_features
cell_line_treatment	source_name_ch1	MDM_co_cultured_with_apoptotic_neutrophils_9h	MDM co-cultured with apoptotic neutrophils (9h)
cell_line_treatment	cell_line_treatment	MDM_co_cultured_with_apoptotic_neutrophils_9h	MDM co-cultured with apoptotic neutrophils (9h)
cell_line_treatment	cell.line.ch1	MDM_co_cultured_with_apoptotic_neutrophils_9h	MDM co-cultured with apoptotic neutrophils (9h)
cell_line_treatment	cell_line_treatment	MDM_co_cultured_with_apoptotic_neutrophils_and_stimulated_with_LPS_1ng_per_ml_for_9h	MDM co-cultured with apoptotic neutrophils and stimulated with LPS (1ng/ml) for 9h
cell_line_treatment	cell.line.ch1	MDM_co_cultured_with_apoptotic_neutrophils_and_stimulated_with_LPS_1ng_per_ml_for_9h	MDM co-cultured with apoptotic neutrophils and stimulated with LPS (1ng/ml) for 9h
cell_line_treatment	source_name_ch1	MDM_co_cultured_with_apoptotic_neutrophils_and_stimulated_with_LPS_1ng_per_ml_for_9h	MDM co-cultured with apoptotic neutrophils and stimulated with LPS (1ng/ml) for 9h
cell_line_treatment	cell.line.ch1	MDM_no_treatment	MDM no treatment
cell_line_treatment	source_name_ch1	MDM_no_treatment	MDM no treatment
cell_line_treatment	cell_line_treatment	MDM_no_treatment	MDM no treatment
cell_line_treatment	cell_line_treatment	MDM_stimulated_with_LPS_1ng_per_ml_9h	MDM stimulated with LPS (1ng/ml
cell_line_treatment	cell_line_treatment	MDM_stimulated_with_LPS_1ng_per_ml_9h	9h)
cell_line_treatment	cell.line.ch1	MDM_stimulated_with_LPS_1ng_per_ml_9h	MDM stimulated with LPS (1ng/ml
cell_line_treatment	cell.line.ch1	MDM_stimulated_with_LPS_1ng_per_ml_9h	9h)
cell_line_treatment	source_name_ch1	MDM_stimulated_with_LPS_1ng_per_ml_9h	MDM stimulated with LPS (1ng/ml
cell_line_treatment	source_name_ch1	MDM_stimulated_with_LPS_1ng_per_ml_9h	9h)
cell_type	cell type.ch1	alveolar_macrophage	Alveolar Macrophage
cell_type	cell type.ch1	alveolar_macrophage	Alveolar macrophage
cell_type	cell type.ch1	AT_II_cells	AT-II cells
cell_type	cell type.ch1	bronchoalveolar_lavage_cells	bronchoalveolar lavage (BAL) cells
cell_type	cell type.ch1	bulk	#PUUTTUU!
cell_type	cell type.ch1	epithelial_culture	epithelial-culture
cell_type	cell type.ch1	epithelial_culture	cultured epithelial cells
cell_type	cell type.ch1	mesenchymal_progenitor_cell	mesenchymal progenitor cell (MPC)
cell_type	cell type.ch1	monocyte_derived_macrophage	monocyte-derived macrophage
cell_type	cell type.ch1	pulmonary_fibroblasts	fibroblast
cell_type	cell type.ch1	pulmonary_myofibroblasts	Lung myofibroblasts
cell_type	cell type.ch1	single_cell	#PUUTTUU!
cluster	#PUUTTUU!	#PUUTTUU!	#PUUTTUU!
cohort	cohort.ch1	#PUUTTUU!	#PUUTTUU!
collagen_gel_contraction	#PUUTTUU!	contractile	#PUUTTUU!
collagen_gel_contraction	#PUUTTUU!	non_contractile	#PUUTTUU!
contact_fax	#PUUTTUU!	#PUUTTUU!	#PUUTTUU!
contact_zip_postal_code	contact_zip.postal_code	#PUUTTUU!	#PUUTTUU!
data_processing	data_processing	#PUUTTUU!	#PUUTTUU!
disease	diagnosis.ch1	cell_line_NA	#PUUTTUU!
disease	diagnosis.ch1	cryptogenic_organizing_pneumonia	COP
disease	diagnosis.ch1	familial_IPF	Familial Idiopathic Pulmonary Fibrosis (IPF)
disease	diagnosis.ch1	healthy	control donor
disease	diagnosis.ch1	hypersensitivity_pneumonitis	HP
disease	diagnosis.ch1	IPF	idiopathic pulmonary fibrosis (IPF)
disease	diagnosis.ch1	mixed_IPF_NSIP	Mixed IPF-NSIP
disease	diagnosis.ch1	NA_cell_culture	#PUUTTUU!

Continued on next page

Continued from previous page

label	lab_syn	allowed_features	syn_features
disease	diagnosis.ch1	non_specific_interstitial_pneumonia	Non-specific interstitial pneumonia
disease	diagnosis.ch1	primary_spontaneous_pneumothorax	primary spontaneous pneumothorax)
disease	diagnosis.ch1	RB_ILD	RB-ILD
disease	diagnosis.ch1	Rheumatoid_Arthritis_Associated_UIP	RA-UIP
disease	diagnosis.ch1	Sc_ILD	Scleroderma associated interstitial lung disease
disease	diagnosis.ch1	spontaneous_IPF	Spontaneous Idiopathic Pulmonary Fibrosis (IPF)
disease	diagnosis.ch1	UIP	Usual Interstitial Pneumonia
disease	diagnosis.ch1	UIP	UIP/IPF
disease	diagnosis.ch1	uncharacterized_fibrosis	FU
disease	diagnosis.ch1	NDC	NDC
disease	diagnosis.ch1	ILD	ILD
disease	diagnosis.ch1	CLAD	CLAD
disease_state	#PUUTTUU!	early	#PUUTTUU!
disease_state	#PUUTTUU!	familial	#PUUTTUU!
disease_state	#PUUTTUU!	rapid	Rapid progressing fibrosis
disease_state	#PUUTTUU!	slow	Slow progressing fibrosis
disease_state	#PUUTTUU!	spontaneous	#PUUTTUU!
disease_state_discard_maybe	status.ch1	rapid	Rapid
disease_state_discard_maybe	status.ch1	slow	Slow
ethnicity	ethnic group	Hispanic or Latino	Hispanic
ethnicity	ethnic group	Hispanic or Latino	Latino
ethnicity	ethnicity.ch1	Hispanic or Latino	Hispanic
ethnicity	ethnicity.ch1	Hispanic or Latino	Latino
ethnicity	ethnicity.ch1	non_hispanic	non-hispanic
ethnicity	ethnic group	non_hispanic	non-hispanic
ethnicity	ethnic group	Not Hispanic or Latino	not-hispanic
ethnicity	ethnic group	Not Hispanic or Latino	not-latino
ethnicity	ethnic group	Not Hispanic or Latino	other
ethnicity	ethnicity.ch1	Not Hispanic or Latino	not-hispanic
ethnicity	ethnicity.ch1	Not Hispanic or Latino	not-latino
ethnicity	ethnicity.ch1	Not Hispanic or Latino	other
extract_protocol	extract_protocol_ch1	#PUUTTUU!	#PUUTTUU!
extract_protocol_channel2	exctract_protocol_ch2	#PUUTTUU!	#PUUTTUU!
forced_vital_capacity	fvc.group.ch1	#PUUTTUU!	#PUUTTUU!
gender_age_physiology_index	gap.ch1	#PUUTTUU!	#PUUTTUU!
growth_protocol	growth_protocol_ch1	#PUUTTUU!	#PUUTTUU!
hyb_protocol	hyb_protocol	#PUUTTUU!	#PUUTTUU!
immunosupressant	is.ch1	YES	#PUUTTUU!
immunosupressant	is.ch1	NO	#PUUTTUU!
immunosupressant	is.ch1	UNKNOWN	#PUUTTUU!
institution	institution.ch1	#PUUTTUU!	#PUUTTUU!
label	label_ch1	biotin	biotin

Continued on next page

Continued from previous page

label	lab_syn	allowed_features	syn_features
label	label_ch2	biotin	biotin
label	label_ch2	Cy3	Cy3
label	label_ch1	Cy3	Cy3
label	label_ch1	Cy5	Cy5
label	label_ch2	Cy5	Cy5
label	label_ch2	Cy5_Cy3	#PUUTTUU!
label	label_ch1	Cy5_Cy3	#PUUTTUU!
label_protocol	label_protocol_ch1	#PUUTTUU!	#PUUTTUU!
label_protocol_channel2	label_protocol_ch2	#PUUTTUU!	#PUUTTUU!
lane	#PUUTTUU!	#PUUTTUU!	#PUUTTUU!
library_preparation	extract_protocol_ch1.1	#PUUTTUU!	#PUUTTUU!
microscopic_appearance	macroscopic.appearance.ch1	normal	#PUUTTUU!
microscopic_appearance	macroscopic.appearance.ch1	scarred	scarred
molecule	#PUUTTUU!	Heavy_polyribosomal_RNA	Contractile Heavy polyribosomal RNA
molecule	#PUUTTUU!	Heavy_polyribosomal_RNA	Non-contractile Heavy polyribosomal RNA
molecule	#PUUTTUU!	PolyA_RNA	#PUUTTUU!
molecule	#PUUTTUU!	polysome_associated_RNA	#PUUTTUU!
molecule	#PUUTTUU!	total_RNA	total RNA
molecule	#PUUTTUU!	total_RNA;polyA_RNA	#PUUTTUU!
molecule_ch2	#PUUTTUU!	polyA RNA	polyA RNA
molecule_channel2	source_name_ch2	polyA_RNA	polyA RNA
molecule_extract_protocol	extract_protocol_ch1.1	#PUUTTUU!	#PUUTTUU!
molecule_source	#PUUTTUU!	#PUUTTUU!	#PUUTTUU!
muc5b_genotype	#PUUTTUU!	#PUUTTUU!	#PUUTTUU!
organism	organism_ch1	Homo_sapiens	Homo sapiens
organism_channel2	organism_ch2	Homo_sapiens	Homo sapiens
patient_id	subject id	#PUUTTUU!	#PUUTTUU!
patient_id	patient	#PUUTTUU!	#PUUTTUU!
patient_id	patient id	#PUUTTUU!	#PUUTTUU!
patient_id	patient number	#PUUTTUU!	#PUUTTUU!
patient_id	patient_number	#PUUTTUU!	#PUUTTUU!
patient_id	subject_id	#PUUTTUU!	#PUUTTUU!
plate	#PUUTTUU!	#PUUTTUU!	#PUUTTUU!
race	#PUUTTUU!	American Indian/Alaska Native	American Indian or Alaska Native
race	#PUUTTUU!	asian	4
race	#PUUTTUU!	Asian	A
race	#PUUTTUU!	Asian	oriental
race	#PUUTTUU!	black	3
race	#PUUTTUU!	Black/African American	black/aa
race	#PUUTTUU!	Black/African American	black/african american
race	#PUUTTUU!	Black/African American	B
race	#PUUTTUU!	Black/African American	AA

Continued on next page

Continued from previous page

label	lab_syn	allowed_features	syn_features
race	#PUUTTUU!	Black/African American	Black/AA
race	#PUUTTUU!	caucasian	White
race	#PUUTTUU!	hispanic	1
race	#PUUTTUU!	Native Hawaiian/Pacific Islander	Native Hawaiian or Pacific Islander
race	#PUUTTUU!	other	6
race	#PUUTTUU!	Other	#PUUTTUU!
sample_description	description	IPF_apex	IPF.Apex
sample_description	group.ch1	IPF_apex	IPF.Apex
sample_description	description	NDC_base	NDC.Base
sample_description	group.ch1	NDC_base	NDC.Base
sample_description	group.ch1	ILD_apex	ILD.Apex
sample_description	description	ILD_apex	ILD.Apex
sample_description	group.ch1	NDC_apex	NDC.Apex
sample_description	description	NDC_apex	NDC.Apex
sample_description	description	IPF_base	IPF.Base
sample_description	group.ch1	IPF_base	IPF.Base
sample_description	group.ch1	ILD_base	ILD.Base
sample_description	description	ILD_base	ILD.Base
sample_description	description	CLAD_apex	CLAD.Apex
sample_description	group.ch1	CLAD_apex	CLAD.Apex
sample_description	description	CLAD_base	CLAD.Base
sample_description	group.ch1	CLAD_base	CLAD.Base
sample_description	description	IPF_NA	IPF.NA
sample_description	group.ch1	IPF_NA	IPF.NA
sample_description	#PUUTTUU!	healthy_apex	#PUUTTUU!
sample_description	#PUUTTUU!	healthy_base	#PUUTTUU!
sample_description	description	NDC_NA	NDC.NA
sample_description	group.ch1	NDC_NA	NDC.NA
sample_id	sample_id	#PUUTTUU!	#PUUTTUU!
sample_source	sample_description	healthy control 1 area 1	#PUUTTUU!
sample_source	sample_source	healthy control 1 area 1	#PUUTTUU!
sample_type	#PUUTTUU!	#PUUTTUU!	#PUUTTUU!
scan_protocol	scan_protocol	#PUUTTUU!	#PUUTTUU!
sequence_read_archive	relation.1	#PUUTTUU!	#PUUTTUU!
sex	gender	F	female
sex	gender	F	Female
sex	Sex.ch1	F	female
sex	Sex.ch1	F	Female
sex	Sex.ch1	M	male
sex	Sex.ch1	M	Male
sex	gender	M	male
sex	#PUUTTUU!	unknown	#PUUTTUU!

Continued on next page

Continued from previous page

label	lab_syn	allowed_features	syn_features
sex	gender	M	Male
smoking_status	smoking_ever_never.ch1	N	N
smoking_status	smoking_ever_never.ch1	Y	Y
smoking_status	smoking_ever_never.ch1	ex-smoker	Ex-Smoker
source_name_channel2	#PUUTTUU!	Stratagene_Universal_Human_Reference_RNA_(catalog number 740000)	#PUUTTUU!
supplementary_file	supplementary_file	#PUUTTUU!	#PUUTTUU!
surface_density	#PUUTTUU!	#PUUTTUU!	#PUUTTUU!
survival_status	survival.status..0...censored..1...death.ch1	censored	0
survival_status	survival.status..0...censored..1...death.ch1	death	1
taxid_channel2	taxid_ch2	9606	9606
time	time point	#PUUTTUU!	#PUUTTUU!
time_to_death_days	time.to.death..days..ch1	#PUUTTUU!	#PUUTTUU!
timepoint_days	timepoint.ch1	#PUUTTUU!	#PUUTTUU!
tissue	#PUUTTUU!	BAL_cells	BAL cells
tissue	#PUUTTUU!	blood	#PUUTTUU!
tissue	#PUUTTUU!	lung	donor lungs
tissue	#PUUTTUU!	lung	fibrotic (IPF) lungs
tissue	#PUUTTUU!	lung	Alveolar Macrophage
tissue	#PUUTTUU!	lung_cultured_fibroblasts	cultured human fibroblasts
tissue	#PUUTTUU!	lung_cultured_myofibroblasts	#PUUTTUU!
tissue	#PUUTTUU!	lung_epithelial_cell_culture	epithelial-culture
tissue	#PUUTTUU!	lung_epithelial_cells_differentiated_from_mesenchymal_stem_cell_like_cells	Epithelial cells differentiated from mesenchymal stems cell-like cells
tissue	#PUUTTUU!	lung_lower_lobe	lung
tissue	#PUUTTUU!	lung_lower_lobe	lower lobe
tissue	#PUUTTUU!	lung_upper_lobe	lung
tissue	#PUUTTUU!	lung_upper_lobe	upper lobe
tissue_source	tissue_source	alveolar_septae	control alveolar septae
tissue_source	tissue_source	alveolar_septae	IPF alveolar septae
tissue_source	site.ch1	alveolar_septae	control alveolar septae
tissue_source	site.ch1	alveolar_septae	IPF alveolar septae
tissue_source	site.ch1	apical_region_of_lung	apical region of lung
tissue_source	tissue_source	apical_region_of_lung	apical region of lung
tissue_source	site.ch1	basal_region_of_lung	basal region of lung
tissue_source	tissue_source	basal_region_of_lung	basal region of lung
tissue_source	tissue_source	biopsy	#PUUTTUU!
tissue_source	site.ch1	biopsy	#PUUTTUU!
tissue_source	tissue_source	central	central
tissue_source	site.ch1	central	central
tissue_source	tissue_source	control_lung_necropsy	control lung necropsy
tissue_source	site.ch1	control_lung_necropsy	control lung necropsy
tissue_source	tissue_source	digest	#PUUTTUU!
tissue_source	site.ch1	digest	#PUUTTUU!

Continued on next page

Continued from previous page

label	lab_syn	allowed_features	syn_features
tissue_source	site.ch1	explant	#PUUTTUU!
tissue_source	tissue_source	explant	#PUUTTUU!
tissue_source	tissue_source	FFPE_tissue	FFPE tissues
tissue_source	site.ch1	FFPE_tissue	FFPE tissues
tissue_source	site.ch1	fibroblast_foci	IPF fibrobalst foci
tissue_source	tissue_source	fibroblast_foci	IPF fibrobalst foci
tissue_source	tissue_source	healthy_donor	#PUUTTUU!
tissue_source	site.ch1	healthy_donor	#PUUTTUU!
tissue_source	site.ch1	healthy_lung_BAL	Bronchoalveolar lavage
tissue_source	tissue_source	healthy_lung_BAL	Bronchoalveolar lavage
tissue_source	tissue_source	healthy_lung_biopsy	Biopsy
tissue_source	site.ch1	healthy_lung_biopsy	Biopsy
tissue_source	tissue_source	healthy_lung_digest	Digest
tissue_source	site.ch1	healthy_lung_digest	Digest
tissue_source	tissue_source	IPF_lung_BAL	Bronchoalveolar lavage
tissue_source	site.ch1	IPF_lung_BAL	Bronchoalveolar lavage
tissue_source	tissue_source	IPF_lung_biopsy	IPF lung biopsy
tissue_source	tissue_source	IPF_lung_biopsy	Biopsy
tissue_source	site.ch1	IPF_lung_biopsy	IPF lung biopsy
tissue_source	site.ch1	IPF_lung_biopsy	Biopsy
tissue_source	site.ch1	IPF_lung_digest	Digest
tissue_source	tissue_source	IPF_lung_digest	Digest
tissue_source	tissue_source	IPF_lung_transplant	IPF lung transplant
tissue_source	site.ch1	IPF_lung_transplant	IPF lung transplant
tissue_source	site.ch1	Lung_tissue_sample_from_lung_transplant_patient	#PUUTTUU!
tissue_source	tissue_source	Lung_tissue_sample_from_lung_transplant_patient	#PUUTTUU!
tissue_source	site.ch1	Lung_tissue_sample_from_the_patient_with_ILD	Lung tissue sample from the patient with ILD
tissue_source	tissue_source	Lung_tissue_sample_from_the_patient_with_ILD	Lung tissue sample from the patient with ILD
tissue_source	tissue_source	necropsy	#PUUTTUU!
tissue_source	site.ch1	necropsy	#PUUTTUU!
tissue_source	site.ch1	peripheral	peripheral
tissue_source	tissue_source	peripheral	peripheral
tissue_source	site.ch1	routine_lung_volume_reduction	routine lung volume reduction
tissue_source	tissue_source	routine_lung_volume_reduction	routine lung volume reduction
tissue_source	tissue_source	Sc_ILD_lung_BAL	Bronchoalveolar lavage
tissue_source	site.ch1	Sc_ILD_lung_BAL	Bronchoalveolar lavage
tissue_source	site.ch1	Sc_ILD_lung_biopsy	Biopsy
tissue_source	tissue_source	Sc_ILD_lung_biopsy	Biopsy
tissue_source	tissue_source	Sc_ILD_lung_digest	Digest
tissue_source	site.ch1	Sc_ILD_lung_digest	Digest
tissue_source	tissue_source	transplant	#PUUTTUU!
tissue_source	site.ch1	transplant	#PUUTTUU!

Continued on next page

Continued from previous page

label	lab_syn	allowed_features	syn_features
tissue_source	site.ch1	Uninvolved_lung_tissue_sample_from_lung_cancer_patient	Uninvolved lung tissue sample from lung cancer patient
tissue_source	tissue_source	Uninvolved_lung_tissue_sample_from_lung_cancer_patient	Uninvolved lung tissue sample from lung cancer patient
tissue_source	tissue_source	Apex	apex
tissue_source	site.ch1	Apex	apex
tissue_source	site.ch1	Base	Base
tissue_source	tissue_source	Base	Base
treatment	#PUUTTUU!	co_cultured_with_apoptotic_neutrophils	co-cultured with apoptotic neutrophils
treatment	#PUUTTUU!	co_cultured_with_apoptotic_neutrophils_and_stimulated_with_LPS	co-cultured with apoptotic neutrophils and stimulated with LPS
treatment	#PUUTTUU!	control	#PUUTTUU!
treatment	#PUUTTUU!	NA_patient_cells	#PUUTTUU!
treatment	#PUUTTUU!	no_treatment	#PUUTTUU!
treatment	#PUUTTUU!	stimulated_with_LPS	stimulated with LPS
treatment_protocol	treatment_protocol_ch1	#PUUTTUU!	#PUUTTUU!
treatment_protocol_channel2	treatment_protocol_ch2	#PUUTTUU!	#PUUTTUU!
age_group	age group.ch1	60_	Senior Adult (>=60)
age_group	age group.ch1	19_59	Adult [19-59]
disease_subtype	diseasesubtype.ch1	hypersensitivity_pneumonitis	HP
disease_subtype	diseasesubtype.ch1	NSIP	#PUUTTUU!
diseasenormal	#PUUTTUU!	#PUUTTUU!	#PUUTTUU!
dlco_perc	#PUUTTUU!	#PUUTTUU!	#PUUTTUU!
dlco_perc_corrected	#PUUTTUU!	#PUUTTUU!	#PUUTTUU!
fvcprebd_perc	#PUUTTUU!	#PUUTTUU!	#PUUTTUU!
leftright	leftright.ch1	right	Right
leftright	leftright.ch1	unclassified	Unclassified
leftright	leftright.ch1	left	Left
library_size	lib.size.ch1	#PUUTTUU!	#PUUTTUU!
lungweight_mg	lungweight mg.ch1	#PUUTTUU!	#PUUTTUU!
norm_factors	#PUUTTUU!	#PUUTTUU!	#PUUTTUU!
processing_date	#PUUTTUU!	5/1/2018	5/1/2018
rin	#PUUTTUU!	#PUUTTUU!	#PUUTTUU!
rnaconcentration_ngul	#PUUTTUU!	#PUUTTUU!	#PUUTTUU!
severity	severity.ch1	advanced	Advanced
severity	severity.ch1	unknown	Unknown
severity	severity.ch1	severe	Severe
severity	severity.ch1	moderate	Moderate
volum_ul	#PUUTTUU!	14	14
#PUUTTUU!	#PUUTTUU!	culture	#PUUTTUU!
psl	#PUUTTUU!	YES	#PUUTTUU!
psl	#PUUTTUU!	NO	#PUUTTUU!
psl	#PUUTTUU!	UNKNOWN	#PUUTTUU!