UNIVERZA NA PRIMORSKEM
FAKULTETA ZA MATEMATIKO, NARAVOSLOVJE IN
INFORMACIJSKE TEHNOLOOGIJE

ZAKLJUČNA NALOGA
(FINAL PROJECT PAPER)

PLJUČNA FIBROZA KOT POSLEDICA MOTENEGA MITOHONDRIJSKEGA METABOLIZMA IN BIOGENEZE

(LUNG FIBROSIS AS A CONSEQUENCE OF A PERTURBATION OF MITOCHONDRIAL METABOLISM AND BIOGENESIS)

TEA JANKO

2018

ZAKLJUČNA NALOGA (FINAL PROJECT PAPER)

JANKO

UNIVERZA NA PRIMORSKEM
FAKULTETA ZA MATEMATIKO, NARAVOSLOVJE IN INFORMACIJSKE TEHNOLOOGIJE

Zaključna naloga
(Final project paper)

Pljučna fibroza kot posledica motenega mitohondrijskega metabolizma in biogeneze

(Lung fibrosis as a consequence of a perturbation of mitochondrial metabolism and biogenesis)

Ime in priimek: Tea Janko
Študijski program: Bioinformatika
Mentor: doc. dr. Katja Lakota
Somentor: doc. dr. Peter Juvan
Delovni somentor: dr. Gerhard G. Thallinger

Koper, september 2018

Ključna dokumentacijska informacija

Ime in PRIIMEK: Tea JANKO

Naslov zaključne naloge: Pljučna fibroza kot posledica motenega mitohondrijskega metabolizma in biogeneze

Kraj: Koper

Leto: 2018

Število listov: 105 Število slik: 22 Število tabel: 47

Število prilog: 10 Št. strani prilog: 46

Število referenc: 81

Mentor: doc. dr. Katja Lakota

Somentor: doc. dr. Peter Juvan

Delovni somentor: dr. Gerhard G. Thallinger

Ključne besede: pljučna fibroza, sistemska skleroza, ekspresijsko profiliranje, mitohondrijski metabolizem, biogeneza

Izvleček: Pljučna fibroza je progresivno brazgotinjenje pljučnega tkiva, ki se pojavlja pri sistemski sklerozi (SS) in intersticijski pljučni fibrozi (IPF), z omejenimi možnostmi zdravljenja. Patofiziološko to stanje opišemo kot prekomerni nastanek medcelične, katerega povzročajo vztrajno aktivirani fibroblasti, ki diferencirajo v miofibroblaste. Ker povečana beljakovinska sinteza in proliferacija celic zahtevata zvišano regulacijo metaboličnih poti, povezanih s stimulacijo mitohondrijske biogeneze, je bil cilj te raziskave pregledati metabolične motnje in mitohondrijsko biogenezo v pljučnih fibroblastih in posledični učinek na patogenezo SS in IPF. Z bioinformatično analizo (analiza obogatenosti genskih skupin/poti in analiza diferenčne izraženosti genov) dveh javno dostopnih naborov podatkov DNA-mikromrež, so bili pridobljeni seznami obogatenih poti in diferenčno izraženih genov. Za določitev morebitnih funkcijskih interakcij med proteini, ki jih kodirajo diferenčno izraženi geni, je bila uporabljena podatkovna baza STRING. Rezultati analize SS in IPF so pokazali motnje v metaboličnih poteh, ki so pričakovane v visoko proliferativnih celicah – povišana glikoliza/glukoneogeneza, povišan metabolizem purinov in pirimidinov ter povečana replikacija DNA. Poleg tega so rezultati pokazali motnje encimov, vključenih v vse tri stopnje celičnega dihanja (citosolna glikoliza, mitohondrijski cikel citronske kisline in oksidativna fosforilacija).

Opažena je bila tudi sprememba uravnavanja genov, povezanih z metabolizmom sfingolipidov, arginina in prolina ter arahidonske kisline. Vsi pridobljeni rezultati prikazujejo precejšne presnovne spremembe, kar odraža visoko energijsko zahtevo SS in IPF fibroblastov. Prav tako so rezultati pokazali diferenčno izražene gene v mitohondrijski biogenezi, kar indicira, da je ta proces afektiran tako v SS kot v IPF. Kljub temu, je za dokončne zaključke potrebna bolj podrobna preiskava omenjenega procesa.

Key words documentation

Name and SURNAME: Tea JANKO

Title of the final project paper: Lung fibrosis as a consequence of perturbation of mitochondrial metabolism and biogenesis

Place: Koper

Year: 2018

Number of pages: 105 Number of figures: 22 Number of tables: 47

Number of appendices: 10 Number of appendix pages: 46

Number of references: 81

Mentor: Assist. Prof. Katja Lakota, PhD

Co-Mentors: Assist. Prof. Peter Juvan, PhD

Working Co-Mentor: Gerhard G. Thallinger, PhD

Keywords: lung fibrosis, systemic sclerosis, expression profiling, mitochondrial metabolism, biogenesis

Abstract: Pulmonary fibrosis is progressive scarring of lung tissue occurring in systemic sclerosis (SSc) and interstitial pulmonary fibrosis (IPF), with limited treatment options. Pathophysiologically, excessive extracellular matrix (ECM) build up occurs, caused by persistently activated fibroblasts that differentiate into myofibroblasts. Since increased protein synthesis and cell proliferation require upregulation of metabolic pathways linked to the stimulation of mitochondrial biogenesis, the aim of this research was to examine metabolic perturbations and mitochondrial biogenesis in lung fibroblasts and subsequent effect on SSc and IPF pathogenesis. Bioinformatic analysis (gene set enrichment analysis and differential expression analysis) of two publicly accessible DNA microarray datasets produced lists of differentially expressed (DE) genes and enriched pathways. To determine possible functional interactions between the expressed proteins encoded by DE genes, STRING database was used. The results of SSc and IPF analysis showed perturbations in metabolic pathways expected in highly proliferative cells, such as increased glycolysis/gluconeogenesis, increased metabolism of purines, metabolism of pyrimidines and increased DNA replication. Furthermore, results showed perturbations of enzymes involved in all three stages of cell respiration (cytosolic glycolysis, mitochondrial citric acid cycle and oxidative phosphorylation).

In addition, dysregulation of genes associated with sphingolipid metabolism, with arginine and proline metabolism and with arachidonic acid metabolism was observed. Taken together, our results show profound metabolic changes, reflecting high energy demand of SSc and IPF fibroblasts. Lastly, although results show few DE genes in mitochondrial biogenesis, suggesting that this process is affected in both SSc and IPF, this pathway requires more specific examination for definitive conclusions.

Acknowledgements

I hereby express my sincere appreciation and gratitude to the following who have assisted in the creation of this final project paper:

Associate Professor Snežna Sodin-Šemrl, PhD for her wise and insightful comments, guidance, oversight and support;

Gerhard G. Thallinger, PhD for his willingness to be a part of this study, for responding to numerous questions so promptly and extensively, and for giving extremely detailed advice whenever I needed it;

Assistant Professor Peter Juvan, PhD for his constructive criticism and useful comments;

Assistant Professor Katja Lakota, PhD for her willingness to help at any time, for her patience, continuous motivation, immense knowledge, her aspiring guidance and dedicated involvement in every step throughout the process;

My family for their endless support.

I could not have imagined having better advisors, mentors and support system.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SUPPLEMENTARY DATA

LIST OF ABBREVIATIONS

AA – arachidonic acid

AMPK – AMP-activated protein kinase

ATP – adenosine triphosphate

CoA – coenzyme A

cRNA – complementary RNA

CYP – cytochrome P450

DE – differentially expressed

ECM – extracellular matrix

FC – fold change

FDR – false discovery rate

GO-BP – Gene Ontology Biological Processes

GSEA – gene set enrichment analysis

ILDs – interstitial lung diseases

IPF – idiopathic pulmonary fibrosis

LPA – lysophosphatidate

LPPs – lipid phosphate phosphatases

MAPK – mitogen activated protein kinase

MMP – matrix metalloprotease

mRNA – messenger RNA

mtDNA – mitochondrial DNA

NCBI GEO – National Center for Biotechnology Information Gene Expression Omnibus

nDNA – nuclear DNA

NO – nitric oxide

NRF1 – nuclear transcription factor 1

NRF2 – nuclear transcriptional factor 2

OPN – osteopontin

OXPHOS – oxidative phosphorylation

p38 MAPK – p38 mitogen-activated protein kinase

PGC-1β – peroxisome proliferator-activated receptor-gamma coactivator 1 beta

PH – pulmonary hypertension

PPAR – peroxisome proliferator-activated receptor

PPIs – protein-protein interactions

RMA – robust multiarray average

ROS – reactive oxygen species

S1P – sphingosine-1-phosphate

SGPL1 – sphingosine-1-phosphate lyase 1

SIRT1 – Sirtuin 1

SPHK – sphingosine kinase

SSc – systemic sclerosis

SSc-ILD – scleroderma associated interstitial lung disease

TCA – citric acid cycle

Tfam – mitochondrial transcription factor A

TGF- β1 – transforming growth factor β1

TGF-β – transforming growth factor β

UGCG – UDP-glucose ceramide glucosyltransferase

VEGF – vascular endothelial growth factor

1 INTRODUCTION

1.1 Biological background

Pulmonary fibrosis is progressive scarring (extracellular matrix deposition) of lung tissue caused by several different conditions, with limited treatment options and poor prognosis. The aim of this thesis is to investigate the role of mitochondrial metabolism and mitochondrial biogenesis in lung fibrosis associated with two different diagnoses, systemic sclerosis (SSc) and interstitial pulmonary fibrosis (IPF).

SSc is a chronic autoimmune connective tissue disease with an estimate of an annual incidence (in the United States) of 19.3 new cases per million adults per year (Luckhardt & Thannickal, 2015). According to LeRoy et al. (1988), the two main types of the disease are diffuse SSc and limited SSc. The difference is in the extent of cutaneous changes and internal organ involvement (LeRoy et al., 1988). SSc is characterized by fibroproliferative vasculopathy, immunological abnormalities and progressive fibrosis of multiple organs such as lungs and skin (Mostmans et al., 2017). The peak incidence is between 45 and 64 years of age (Luckhardt & Thannickal, 2015).

Silver and Silver (2015) stated that scleroderma-associated interstitial lung disease (SSc-ILD) has become the leading SSc related cause of death. It is likely that it represents a complex interplay between innate and acquired immunity, inflammation and fibrosis, but the exact sequence of events remains uncertain. Females are at higher overall risk for developing SSc, but males are more likely to develop severe SSc-ILD. The prevalence of SSc-ILD is higher in patients with diffuse cutaneous SSc than in those with limited cutaneous SSc. Patients who have anti-topoisomerase I antibodies are also at a higher risk (Silver & Silver, 2015).

Due to poor understanding of molecular pathways involved in interstitial lung diseases (ILDs), a transcriptomic study was conducted by Cho et al. (2011), to identify perturbed gene networks. It revealed strong perturbances in pathways such as transforming growth factor β (TGF-β) pathway, Wnt signalling, focal adhesion, extracellular matrix (ECM)-receptor interactions and mitogen-activated protein kinase (MAPK) signalling. In addition, the results also implied a decrease in general lung metabolic activities (Cho et al., 2011).

Moore and Herzog (2013) stated that IPF is a chronic, progressive, incurable lung disease of unknown etiology (Moore & Herzog, 2013). According to Ryu et al. (2014), it accounts for 20% to 30% of ILDs and occurs mostly in patients between 50 and 85 years of age, predominantly males. In the United States, the incidence of IPF is estimated to be 7 to 17 per 100,000 person-years and estimated median survival after diagnosis is approximately three years (Ryu et al., 2014).

There are only a few expression profiling studies associated with IPF. Since an accurate diagnosis of the disease is very challenging, Meltzer et al. (2011) conducted a study in which the experiments were designed for developing definitive diagnostic and prognostic gene signatures for IPF. The study showed that a statistical method called Bayesian probit regression is a very powerful tool to increase accuracy in diagnosis and prognosis of IPF (Meltzer et al., 2011). Another study by Pardo et al. (2005), focusing on the molecular mechanisms of the disease, demonstrated that osteopontin (OPN) was highly upregulated in bleomycin induced lung fibrosis in mice. This study also analysed the direct effects of OPN on human lung fibroblasts, alveolar epithelial cell migration and proliferation and matrix metalloprotease (MMP) gene expression *in vitro*. Their analysis demonstrated that OPN is highly expressed in IPF lungs and their results suggest that the interaction between MMP-7 and OPN may be involved in the progressiveness of the disease (Pardo et al., 2005).

Although SSc and IPF have different gender predominance, they typically occur at different ages and have different natural history, Herzog et al. (2014) stated that pathophysiologically, they are both characterized by the same process of excessive ECM build up. This impairs the lung's architecture and function, which is the ability to exchange gas and deliver oxygen into the blood, resulting in hypoxia. Major producers of ECM are fibroblasts that, activated with pro-fibrotic stimuli (e.g. TGF-β) differentiate into myofibroblasts. They express ECM, including fibronectin, proteoglycans and collagen types I, III, V and VII (Herzog et al., 2014).

According to Bernard et al. (2015), differentiation of fibroblasts to myofibroblasts could be accompanied by robust metabolic reprogramming (Bernard et al., 2015). *In vitro* experiments showed that changes in mitochondrial biogenesis affect ECM production (Peng et al., 2013) and that metabolic changes affect development of lung fibrosis (Renzoni et al., 2004). It is well known that in fast proliferating cells (as myofibroblasts in fibrotic tissue) the metabolism is reorganised, favouring glycolysis instead of oxidative phosphorylation (OXPHOS). The metabolic reprogramming can be induced by transforming growth factor β1 (TGF-β1) via the p38 mitogen-activated protein kinase (p38 MAPK) dependent pathway and is linked to the stimulation of both mitochondrial biogenesis and glycolysis (Bernard et al., 2015).

Mitochondrial biogenesis is defined as the process through which cells increase their individual mitochondrial mass with growth and division (Ventura-Clapier et al., 2008). Cooper (2000) defined mitochondria as endomembrane systems in eukaryotic cells responsible for cellular energy production via the oxidative breakdown of glucose and fatty acids. These organelles are confined by the outer and the inner membranes which separate them from the cytosol. The inner membrane consists of many folds known as cristae and extends into the matrix of the mitochondrion. Glycolysis, the initial stage of glucose metabolism, occurs in the cytosol. It results in formation of pyruvate which is, together with fatty acids, transported into mitochondrial matrix and converted to acetyl coenzyme A (CoA). Acetyl CoA is then oxidized in the central process of oxidative metabolism called the citric acid cycle (TCA). Most of the energy derived from this metabolism is further produced in the inner mitochondrial membrane as a result of OXPHOS in form of adenosine triphosphate (ATP) molecules (Cooper, 2000). Mitochondria, as the major reactive oxygen species (ROS) producers and antioxidant producers, have a significant role within the cell mediating processes, such as apoptosis (Valero, 2014), which fails to work in fibrotic diseases. This failure subsequently leads to persistence of myofibroblasts and consequently to expansion of the ECM (Bernard et al., 2015).

Ventura-Clapier et al. (2008) stated that mitochondrial self-replication is dependent on nuclear and mitochondrial genome translation for which mitochondria contain a small fraction of a cell’s DNA. It encodes information for 13 protein subunits of the mitochondrial respiratory chain, 22 transfer RNAs and 2 mitochondrial ribosome-coding RNAs. The master regulator of mitochondrial biogenesis is PGC-1α. It activates downstream transcription factors, such as nuclear respiratory factors 1 and 2 (NRF1 and NRF2), leading to transcription of nuclear encoded proteins and of the mitochondrial transcription factor A (Tfam). Tfam then activates transcription and replication of the mitochondrial genome (Ventura-Clapier et al., 2008).

Mitochondrial biogenesis is influenced by different stimuli, including low temperature, caloric restriction without malnutrition which increases AMP-activated protein kinase (AMPK) activity and Sirtuin 1 (SIRT1) activity, hormones, such as thyroid hormone and growth factors such as vascular endothelial growth factor (VEGF) (Guo et al., 2017; Jornayvaz & Shulman, 2010). Dysfunctional mitochondrial biogenesis has been implicated in the pathogenesis of numerous diseases, for instance, hypertrophic cardiomyopathy and heart failure (Pisano et al., 2016), pulmonary arterial hypertension (Yu & Chan, 2017), type 2 diabetes mellitus (Johannsen & Ravussin, 2009) along with Alzheimer’s disease, Parkinson’s disease, Huntington’s disease and amyotrophic lateral sclerosis (Xu et al., 2015). Several pharmacologic substances are available to stimulate the pathways involved in mitochondrial biogenesis. Some of them are on the WADA (World Anti-Doping Agency) list of prohibited substances for athletes, due to their effects on skeletal muscles. It was reported that activation of some of the pathways involved in increased mitochondrial biogenesis such as AMPK signalling pathway (Liang et al., 2017) and the peroxisome proliferator-activated receptor (PPAR) signalling pathway (Dantas et al., 2015; Lakatos et al., 2007), with the addition of activation or upregulation of SIRT1 (Zeng et al., 2017) could be beneficial in fibrotic diseases (for instance SSc and IPF).

Bernard et al. (2015) showed that blockage of mitochondrial biogenesis or glycolysis results in suppression of TGF-β1-induced α-smooth muscle actin and collagen α-2 expression (Bernard et al., 2015) subsequently decreasing ECM production. In addition, Xie et al.

(2015) demonstrated on the TGF-β1-induced pulmonary fibrosis *in vivo* model, that glycolytic suppression diminishes lung fibrosis (Xie et al., 2015). Furthermore, IPF has been consistently associated with the process of aging in which metabolic dysregulation and mitochondrial dysfunction naturally occur (Mora et al., 2017; Zank et al., 2018).

Although several above-mentioned studies were based on cell cultures and *in vivo* models associated with metabolic changes in fibrosis, none of them specifically investigated mitochondrial biogenesis, glycolysis, OXPHOS or other metabolic pathways in SSc and IPF fibroblasts. The additional observation that general symptoms of SSc, such as muscle weakness and fatigue, could be closely associated with disorders of energy metabolism, lead us to investigate mitochondrial dysfunction, as a potential target for new treatments.

1.2 Purpose of the study

The objective of the current study was to determine the scope of metabolic reprogramming that occurs in SSc- and IPF-associated fibroblasts, characterized by lung fibrosis, using bioinformatics databases (two publicly available datasets) and tools.

Specific aims:

1. Determine which pathways are significantly enriched in patients SSc and IPF, using publicly available gene expression data from NCBI GEO (Barrett et al., 2013; Edgar et al., 2002) with accession numbers GSE40839 and GSE44723. These datasets contain cell culture fibroblast samples associated with development of lung fibrosis (SSc and IPF).

2. Determine expression levels of genes associated with mitochondrial metabolic pathways (such as TCA cycle, OXPHOS, β-oxidation, glutaminolysis) and test this set of genes for enrichment using above-mentioned datasets.

3. Determine expression levels of genes associated with mitochondrial biogenesis (such as PPARα, PGC-1α, PGC-1β, NRF1, NRF2, Tfam, AMPK, CaMIKV, eNOS, TORC, calcineurin, p38 MAPK, RIP140, Sin3A, TFB2M, HIF-1) and test this set for enrichment using above mentioned datasets.

2 MATERIALS AND METHODS

2.1 Datasets

2.1.1 Scleroderma associated interstitial lung disease data (GSE40839)

Lindahl et al. (2013) conducted a study, in which primary lung fibroblasts used for analysis of the transcriptome were cultured from control tissue samples and from surgical lung biopsy samples of eight patients with pulmonary fibrosis (SSC-ILD). The control tissue of ten patients undergoing cancer-resection surgery, was histologically normal. The median age was 60 in control patients (range from 52 to 78) and 48 in SSc-ILD patients (range from 38 to 69). Fibroblasts used for the experiments were between passages 2-5: median passage number for the control group was 4.5 (range 3-5) and 4 (range 2-5) for SSc-ILD group. Total RNA was harvested from serum-deprived fibroblasts. Complementary RNA (cRNA) was hybridized to Affymetrix human U133Av2 microarrays (Lindahl et al., 2013).

2.1.2 Idiopathic pulmonary fibrosis data (GSE44723)

Primary cultures of lung fibroblasts used for analysis were isolated from the distal parenchyma of patients with IPF. Fibroblast cell lines were characterized across two phenotypes; stable IPF (six donors) and rapidly progressing IPF (four donors). Primary cells were from passage 11. mRNA from harvested cells was purified and subsequently hybridized to Affymetrix HG-U133 plus 2.0 microarrays (Peng et al., 2013).

2.2 Software used for statistical analysis

BRB-ArrayTools Version 4.5.1 – Stable, an integrated software package implemented as an Excel add-in, was used for the analysis of two different DNA microarray datasets from NCBI GEO (Barrett et al., 2013; Edgar et al., 2002) – GSE40839 (Lindahl et al., 2013) and GSE44723 (Peng et al., 2013). BRB-ArrayTools was developed by the Biometric Research Branch of the Division of Cancer Treatment & Diagnosis of the National Cancer Institute, led by Dr. Richard Simon. The software, among other things, allows for processing and normalization of gene expression data, clustering of genes and samples, visualization of samples using multidimensional scaling and analysis of differential gene expression and enrichment of gene sets (Simon et al., 2007; Simon, 2010).

2.3 Data pre-processing

For both datasets used for analysis (GSE40839 and GSE44723) raw data were obtained from NCBI GEO. Raw data (Affymetrix CEL data file format) and GEO platform (GPL) files were downloaded. The data were imported into BRB-ArrayTools using the data import wizard. Robust multiarray average (RMA) method was used for data normalization.

Annotation of genes was performed using NCBI GEO GPL files (GPL96-57554 for GSE40839 and GPL570-55999 for GSE44723) associated with microarray platforms that were used for the corresponding studies. Based on previous studies by McCarthy and Smyth (2009) and Patterson et al. (2006), a 1.5-fold change threshold (in either direction from the gene's median value across all arrays) was set, below that threshold differential expression is considered unlikely to be of interest for any gene was set (McCarthy & Smyth, 2009; Patterson et al., 2006). The following criteria were used for filtering the data in BRB-ArrayTools: (a) genes of which more than 20% expression data values had at least a 1.5-fold change (in either direction from the gene's median value across all arrays) and (b) genes which had less than 50% of missing values were used for further analysis.

2.4 Analysis design

Pathways (groups of genes) from BioCarta (Nishimura, 2001) and KEGG (Kanehisa & Goto, 2000; Kanehisa et al., 2017; Kanehisa et al., 2016) databases were considered for enrichment analysis. Based on personal preference of KEGG gene set lists and corresponding maps, we chose the KEGG pathway database for further analysis. As mitochondrial biogenesis pathway was not included in that database, it had to be manually added. The list of genes involved in mitochondrial biogenesis (Table A1) was obtained from Reactome (Croft et al., 2014; Fabregat et al., 2018) – Mitochondrial biogenesis (Homo sapiens) pathway.

Analysis for each dataset consisted of three sections:

1. Gene set enrichment analysis (GSEA) on all genes (termed Analysis of KEGG pathways) and on a subset of genes included in any metabolic pathway listed in Table 1 (termed Analysis of Metabolic pathways in the following).

By using a suitable metric, GSEA ranks genes based on the correlation between their expression and the phenotypic class distinction. The gene sets/pathways are defined based on prior biological knowledge (Subramanian et al., 2005) – functional annotation/biological identity of the genes.

Analysis tool uses Efron-Tibshirani's GSA maxmean test and LS/KS permutation test. The first mentioned test uses maxmean statistics to identify DE gene sets. The second mentioned test finds gene sets which have more DE genes among the initial state class and final state class than expected by chance (Simon, 2010).

Table 1*: List of metabolic pathways that were analysed for detection of metabolic changes in SSc or rapid progressing IPF (Metabolic pathways subset)*

Pathway description	Number of genes	Defined gene list
Citrate cycle (TCA cycle)	30	KEGG
D-Glutamine and D-glutamate metabolism	4	KEGG
Fatty acid biosynthesis	6	KEGG
Fatty acid degradation	43	KEGG
Glycolysis/Gluconeogenesis	65	KEGG
Metabolic pathways	1130	KEGG
Mitochondrial biogenesis	35	user
Nitrogen metabolism	23	KEGG
Oxidative phosphorylation	132	KEGG
Pentose phosphate pathway	27	KEGG
Pyruvate metabolism	40	KEGG
Retinol metabolism	64	KEGG
TGF-β signalling pathway	84	KEGG

2. Analysis of genes from Metabolic pathways.

According to Simon (2010), class comparison uses univariate parametric and non-parametric tests, performs random permutations of the class labels and computes the proportion of these random permutations to produce a list of DE genes in one class compared to the other. For each gene in the list, the tool computes the permutation p-value, which is based on before mentioned random permutations (Simon, 2010).

2.1. Additionally, the STRING database was used to determine possible functional interactions between the expressed proteins encoded by DE genes. Simultaneously, lists of DE genes was analysed for functional enrichments – analysis of significantly enriched Gene Ontology Biological Processes GO-BP (The Gene Ontology Consortium et al., 2000; The Gene Ontology Consortium, 2017) and KEGG pathways based on protein-protein interactions (PPIs) was performed.

Interaction predictions are derived from genomic context predictions, high-throughput lab experiments, (conserved) co-expression, automated text mining, and previous knowledge in curated databases (Szklarczyk et al., 2017).

3. Analysis of Mitochondrial biogenesis genes (Table A1).

3.1. In order to subset the data in BRB-ArrayTools, Mitochondrial biogenesis gene list tab-delimited text file (Table A2) had to be custom-assembled and added to the already existing BRB-ArrayTools database. The file had to contain three columns – UniGene cluster IDs, the corresponding gene symbols and GenBank accession numbers of the transcripts, respectively. First two columns were used to search for the UniGene annotation and the last column was used to search GenBank annotation (Simon, 2010).

3.2. To ensure that all genes from Mitochondrial biogenesis gene list were available for the differential expression analysis, after the data import, GPL annotation files for both data sets were revised. Six genes (PRKAG3, PERM1, PPARGC1B, CRTC2, HELZ2, ACSS2) were missing from the GPL96 file associated with GSE40839 dataset, which resulted in a list of 57 genes with 122 corresponding probe sets. In contrast, all 63 genes with 198 corresponding probe sets were present in the GPL570 file associated with GSE44723 dataset. Nevertheless, three genes in both annotation files had to be corrected, in order to be detected when subsetting the data – "LOC100129518 /// SOD2" to "SOD2", "NR1D1 /// THRA" to "NR1D1" and "CALM1 /// CALM2 /// CALM3" to "CALM1". Since both GPL files had the same UniGene cluster and accession numbers for each gene and GPL96 was missing six genes (and corresponding probe set IDs), GPL570 was used to assemble Mitochondrial biogenesis gene list tab-delimited text file used in subsequent analysis.

For both datasets, samples were divided in two classes defined by the disease state; control vs. SSc-ILD in dataset GSE40839 and rapidly progressing IPF vs. stable IPF, in dataset GSE44723. The univariate test used in data analysis sections 2-3 was a two-sample t-test. The significance threshold level α, allows to control a percentage of false positive genes and gene sets. Since gene lists with numerous false positives make interpretation very problematic, the significance threshold level was set at 0.01 (1% of expected false positive DE genes) in all sections except for section 2 of dataset GSE40839, where the level was set at 0.001 (0.1% of expected false positive invalid DE genes).

Results from all analyses are presented in tables with probe set labels, gene symbols, p-values and $\log_2$-fold change (logFC) values. Full name for each gene symbol is available in the online human gene database GeneCards (Stelzer et al., 2016). A p-value below the significance threshold suggests that data provide evidence to reject the null hypothesis and that there is a statistically significant difference in gene expression between the two groups of interest. logFC is a measure describing how much the expression of a gene changes between an initial (i.e. control class in GSE40839 or rapidly progressing IPF in GSE44723) and a final (i.e. SSc-ILD class in GSE40839 or steady IPF in GSE44723) value. In dataset GSE40839 logFC<0 suggests that a certain gene is upregulated and logFC>0 suggests that a certain gene is downregulated in SSc-ILD compared to the control. In contrast, in dataset GSE44723 logFC<0 suggests downregulation of a gene and logFC>0 suggests upregulation of a gene in rapidly progressing IPF compared to steady IPF.

2.5 Visualisation

For visualisation of differentially expressed (DE) genes and corresponding probe sets from the Metabolic pathways subset and the Mitochondrial biogenesis subset, heatmaps were used – each row represents expression of a gene through all samples and each column represents expression levels of genes within a sample. The colour and intensity of the boxes represent gene expression levels using $\log_2$ intensity as a proxy. Heatmaps were generated with the use of the Genesis software (Sturn et al., 2002). Gene expression data associated with each gene in differential expression output gene lists was extracted from BRB-ArrayTools using their plugin for gene expression data. These files had to be modified and saved as a Stanford flat-file in order to import the data into Genesis. First column of the modified files had to be named UNIQID (probe set IDs), second column was optional and was named NAME (symbols of genes associated with probe set IDs) followed by required columns of gene expression data for each sample. Samples were renamed to be more comprehensible (Table A32) – controls, SSc-ILDs, rapidly progressing IPFs and steady IPFs. After data import, samples were divided in two groups for each dataset and hierarchical clustering was performed using “average group linkage (UPGMA)” agglomeration rule and “cluster experiments” calculation parameters. Colour scheme of generated heatmaps was adjusted to a single gradient one. Additionally, the maximum value for saturated colours was set to 15, because the highest gene expression value in all gene lists was 13.84.

Functional interactions between DE genes, using STRING analysis, are presented as networks. According to Szklarczyk et al. (2017), each network node represents all the proteins produced by a single gene locus. Small nodes represent proteins of unknown 3D structure, while large nodes represent proteins of which 3D structure is somewhat known or predicted. Edges represent protein-protein associations – known, predicted or other interactions are marked with distinct colours (Szklarczyk et al., 2017).

To visualise comparison of the experiments in one phenotype class versus the experiments in another phenotype class, scatterplots were used. They show the average log-ratio within one class on the x-axis versus the average log-ratio within the other class on the y-axis. These averages are taken on a gene-by-gene basis, and each gene (or probe sets representing the same gene) is represented by a single point in the resulting scatterplot (Simon, 2010). This visualisation method was used for second section (Metabolic pathways) and third section (Mitochondrial biogenesis) analysis.

With the use of PathVisio 3.3.0 software (Kutmon et al., 2015; van Iersel et al., 2008), pathway data models (schemes of biological pathways), with coloured DE genes, were produced (Figures A7, A8, A9, A10, A11, A12). These schemes were used for visual representation when discussing the most informative pathways, associated with mitochondrial metabolism.

3 RESULTS

3.1 Pathway enrichment analysis

GSEA was used to identify significantly enriched or depleted groups of genes that may have an association with the pathogenesis of SSc and IPF.

KEGG pathways of interest - based on our hypothesis of changed metabolism in lung fibroblasts in SSc - are Glycolysis/gluconeogenesis (hsa00010), Fat digestion and absorption (hsa04975), Ether lipid metabolism (hsa00565), Fructose and mannose metabolism (hsa00051) and DNA replication (hsa03030). Pathways of our interest in idiopathic pulmonary fibrosis research are Pyrimidine metabolism (hsa00240), DNA replication (hsa03030), One carbon pool by folate (hsa00670), Purine metabolism (hsa00230) and Osteoclast differentiation (hsa04380).

Pathways that are not individually mentioned or discussed in this section represent different medical conditions that have no anatomical or histological meaning for SSc-ILD or IPF when comparing lung fibroblasts, for example, pathways of Pancreas cancer and Infectious trypanosomiasis. Nevertheless, they may include some of the same genes as the discussed pathways. The number of known pathways is lower than the number of known diseases and one pathway can have different effects on different cell types, which implies that one pathway or gene can be involved in any number of disease states. Pleiotopy is the term describing one gene affecting several seemingly unrelated phenotypic traits. Due to this overlap, our main goal is to identify pathways in fibroblasts which have the greatest impact on SSc-ILD and IPF based on pathophysiology of the two diseases.

3.1.1 Scleroderma associated interstitial lung disease pathways (GSE40839)

After the data import and normalization, 22,283 probe sets were available for the analysis; 3,621 of them passed the filtering criteria and remained for the subsequent analysis.

3.1.1.1 Analysis of KEGG pathways

Thirty-eight out of 190 investigated KEGG gene sets were marked as enriched (Table A3). The position of each pathway (1st to 38th) is based on LS permutation p-value. Cytokine-cytokine receptor interaction and cell adhesion molecules (CAMs), Chemokine signalling pathway, Antigen processing and presentation, Toll-like receptor signalling pathway, NOD-like receptor signalling pathway, Cytosolic DNA-sensing pathway, Natural killer cell mediated cytotoxicity, Autoimmune thyroid disease, Allograft rejection and Graft-versus-host disease are all Immune system pathways and are involved in environmental information processing. Phagosome pathway is identified as a transport and catabolism process. There are also pathways belonging to four types of diseases; endocrine/metabolic disease (type I diabetes mellitus), immune diseases (autoimmune thyroid disease, allograft rejection and graft-versus-host disease), cardiovascular (viral myocarditis) and infectious diseases (hepatitis C and Leishmaniasis). Most of these pathways/diseases are already well studied and therefore have their own KEGG pathway, but are not implicated in our researched pathology (analysis of cultures of lung fibroblast cells). One example is Osteoclast differentiation (17th place on the list of enriched pathways, p=0.00005). There are two genes in this pathway (TGFB1 (Table A4) – upregulated and STAT1 (Table A5) – downregulated in SSc-ILD compared to controls) which are also major stimuli for profibrotic fibroblast activation which stimulates their differentiation into myofibroblast. In each of the two mentioned processes, the genes play a completely different role in different cell types. Thus, such pathways are not further investigated or analysed in this study. Nevertheless, they are an additional source of information regarding possible relations among other clinical symptoms or complications of the disease.

Enriched pathways of interest are Glycolysis/gluconeogenesis (22nd, p=0.00071), Fat digestion and absorption (30th, p=0.00893), Ether lipid metabolism (31st, p=0.00898), Fructose and mannose metabolism (32nd, p=0.01912) and DNA replication (38th, p=0.53578).

3.1.1.2 Analysis of Metabolic pathways

After application of the subsetting criteria defined in analysis design (methods section), 351 probe sets remained available for GSEA. Eight out of 93 total investigated gene sets were marked as enriched (Table 2).

***Table 2**: Enriched Metabolic pathways by GSEA (α=0.01) in SSc-ILD compared to controls, sorted by LS permutation p-value*

Pathway description	Number of probe sets	LS permutation p-value
Fc gamma R-mediated phagocytosis	6	0.00043
Ether lipid metabolism	9	0.00211
Glycolysis/Gluconeogenesis	29	0.00292
Fructose and mannose metabolism	8	0.00698
Metabolism of xenobiotics by cytochrome P450	6	0.01958
Glycerophospholipid metabolism	10	0.04252
Fat digestion and absorption	10	0.04468
Lysosome	9	0.09520

Fc gamma R-mediated phagocytosis pathway was not chosen as a pathway of interest for the Metabolic pathways subset. However, it includes three genes, two of which (PPAP2B and PPAP2A) were found as DE in further differential expression analysis (Table A28). Ether lipid metabolism and Glycerophospholipid metabolism are both processes within Lipid metabolism. Fructose and mannose metabolism is part of Carbohydrate metabolism, Fat digestion and absorption is included in Digestive system and Lysosome takes part in Transport and catabolism. None of these pathways were previously specifically characterized to be changed in SSc-ILD, but they include certain genes, which are known to be associated with this disease. Those genes are PPAP2B, PPAP2A, AGPS, PAFAH1B1, TPI1, TSTA3, PFKP, AKR1B1, GMPPA, ATP6V0B, SGSH and GNS. They were found as DE in further differential expression analysis (Table A28).

The two further investigated pathways in our research are Glycolysis/gluconeogenesis (Tables A7 and A8) and Metabolism of xenobiotics by cytochrome P450 (Table A9). Glycolysis/gluconeogenesis pathway has the same DE genes as in previous analysis based on all genes – 18 upregulated (Table A6) and 11 downregulated probe sets (Table A7). All three genes of Metabolism of xenobiotics by cytochrome P450 pathway (ADH5, ALDH1A3 and ADH1B) are also included in Glycolysis/gluconeogenesis pathway. They are all downregulated in SSc-ILD compared to controls with the same parametric p-values and logFC values. ADH5 and ADH1B are also included in fatty acid degradation and retinol metabolism.

3.1.2 Idiopathic pulmonary fibrosis pathways (GSE44723)

After the data import and normalization, 54,675 probe sets were available for the analysis; 7,792 of them passed the filtering criteria and remained for the subsequent analysis.

3.1.2.1 Analysis of KEGG pathways

Twenty-nine out of 203 investigated gene sets were marked as enriched (Table A11). Pyrimidine metabolism pathway, DNA replication, One carbon pool by folate and Cell cycle pathway are all expected to be changed in cells with rapid proliferation. Additionally, Base excision repair, Nucleotide excision repair and Mismatch repair pathways are shown to be enriched. We observe a few pathways, such as Progesterone-mediated oocyte maturation pathway, Oocyte meiosis and Type II diabetes mellitus, which are not linked to pathogenesis of IPF, but include certain common genes.

Enriched pathways of interest in our research are Pyrimidine metabolism (1st, p=0.00001) (Tables A12 and A13), DNA replication (2nd, p=0.00001) (Table A14), One carbon pool by folate (9th, p=0.00025) (Tables A15 and A16), Purine metabolism (21st, p=0.00841) (Tables A17 and A18), and Osteoclast differentiation (27th, p=0.06275) (Tables A19 and A20). These pathways include some genes which were found as DE in further analysis (Sections

3.2 and 3.3). Those genes are RRM1, POLE2, PRIM2, CMPK2, TYMS, DTYMK, POLE3, PRIM1, POLA1, ATIC, DHFR, IMPDH2, PAICS, PFAS and PGM2 which is also included in Glycolysis/gluconeogenesis pathway.

3.1.2.2 Analysis of Metabolic pathways

After application of the subsetting criteria defined in analysis design (methods section), 508 probe sets remained available for GSEA. Nine out of 112 investigated gene sets were marked as enriched (Table 3).

***Table 3**: Enriched Metabolic pathways by GSEA (α=0.01) in stable IPF compared to rapidly progressing IPF, sorted by LS permutation p-value*

Pathway description	Number of probe sets	LS permutation p-value
Pyrimidine metabolism	40	0.00001
DNA replication	10	0.00003
One carbon pool by folate	14	0.00027
Purine metabolism	53	0.00037
Nucleotide excision repair	5	0.00121
Folate biosynthesis	5	0.00547
Base excision repair	6	0.00886
Mucin type O-Glycan biosynthesis	26	0.03494
Bladder cancer	8	0.30927

Eight pathways, not including Folate biosynthesis, are also marked as enriched in previous analysis of all genes. The difference is in total number of genes in each pathway, and the level of their expressions (Tables A21, A22, A23, A24, A25, A26 and A27). These alterations do not provide any additional information regarding inclusion of genes found as DE in further analysis (Sections 3.2 and 3.3).

3.1.3 Comparison of pathway enrichment analysis of both datasets

We observe enriched metabolic pathways – Carbohydrate and Lipid metabolism in SSc-ILD and Nucleotide metabolism with the addition of Metabolism of cofactors and vitamins in IPF. Another commonality is affected genetic information processing (enriched DNA replication pathway). When focusing on the differences, we observe changes in Digestive system and Xenobiotics metabolism in SSc-ILD which are not apparent in IPF and changes in organismal development, which are evident in IPF and not in SSc-ILD.

3.2 Analysis of genes from Metabolic pathways

Our aim was to identify genes which contribute to the pathologic activation of lung fibroblasts in patients with SSc-ILD and IPF. Thus, we included genes associated with eleven metabolic pathways (Table 1) and a major profibrotic pathway (TGF-β signalling pathway) as a control.

3.2.1 Genes from Metabolic pathways associated with Scleroderma associated interstitial lung disease (GSE40839)

Comparison of gene expression levels among fibroblasts with profibrotic phenotype (SSc-ILD class) and unaffected fibroblasts (control class) was performed. It resulted in a list of 101 probe sets representing 75 DE genes (Table A28). Their logFC values range from -5.80 to 3.69. Thirty-seven probe sets (36.6%), representing 26 DE genes, have logFC values greater than 0, which means they are downregulated in SSc-ILD class compared to control class. Sixty-four probe sets (63.4%), representing 49 genes, have logFC values less than 0, which means they are upregulated in SSc-ILD class compared to control class. There are 20 genes represented by multiple probe sets (PPAP2B, TPI1, ADH5, PTGES, GLS, ENO1, PRPS1, PAICS, PGK1, DCN, ACLY, MAN1A1, GALNT10, BCAT1, PPAP2A, SMAD3, PTGS1, GFPT1, ID4 and GNS). All probe sets representing 75 genes are visualised in a heatmap based on their expression values (Figure A1).

Of the 84 genes included in KEGG's TGF-β signalling pathway, which is recognized as the main profibrotic pathway, nine are represented by 14 probe sets. Genes ID1, ID3, SMAD7, INHBA, ID4 and TGFB1 are upregulated, while genes TGFBR2, DCN and SMAD3 are downregulated. Altered expression of these genes leads to overall activation of TGF-β signalling pathway. Other 87 probe sets representing 66 genes are all involved in metabolic pathways. Out of 66 genes, four are included in Fatty acid degradation (ADH5, ACOX3, ADH1B and ALDH1B1). three of them (ADH5, ADH1B and ALDH1B1) are also included in Glycolysis/gluconeogenesis, which has total of ten DE genes – additional seven genes being TPI1, ENO1, ALDH1A3, PFKP, PGK1, GPI and LDHA. Genes PFKP and GPI are furthermore associated with the Pentose phosphate pathway which includes one additional DE gene (PRPS1). Genes GLS (also included in D-glutamine and D-glutamate metabolism) and GLUL are associated with Nitrogen metabolism. Genes AKR1B1, LDHA and ALDH1B1 are included in Pyruvate metabolism. Genes ACLY and IDH2 are involved in TCA cycle. Genes ATP6V0B, ATP5G1 and NDUFS1 are included in Oxidative phosphorylation pathway.

After dividing 66 DE genes into two groups (downregulated and upregulated genes), STRING was used to detect possible connections among genes within each group. These connections, which are based on known and predicted PPIs are visually presented in two separate schemes (Figures A3 and A4). Analysis of enriched GO-BP and KEGG pathways, also based on PPIs, was performed. It showed major upregulation in TGF-β signalling pathway and processes associated with Nucleotide, Pyrimidine, Arginine and proline metabolism (Tables 4 and 5), while downregulation was observed in Tyrosine and Cytochrome P450 xenobiotic metabolism and Carboxylic acid anabolism (Tables 6 and 7).

Table 4: *Functionally enriched GO-BP in the network of proteins encoded by upregulated group of DE genes*

Pathway description	**Count in gene set**	**False discovery rate (FDR)**
Nucleotide metabolic process	21	8.46e-19
Carbohydrate derivative metabolic process	26	8.59e-19
Nucleobase-containing small molecule metabolic process	21	2.3e-18
Nucleoside metabolic process	17	3.16e-16
Nucleoside triphosphate metabolic process	15	9e-16

Table 5: *Functionally enriched KEGG pathways in the network of proteins encoded by upregulated group of DE genes*

Pathway description	**Count in gene set**	**FDR**
Metabolic pathways	39	1.88e-37
Pyrimidine metabolism	7	5.62e-07
TGF-β signalling pathway	6	2.74e-06
Amino sugar and nucleotide sugar metabolism	5	5.84e-06
Arginine and proline metabolism	5	1.4e-05
Glycolysis/gluconeogenesis	5	1.51e-05
Biosynthesis of amino acids	5	2.82e-05
Fructose and mannose metabolism	4	3.57e-05
Purine metabolism	6	7.84e-05

Table 6: *Functionally enriched GO-BP in the network of proteins encoded by downregulated group of DE genes*

Pathway description	**Count in gene set**	**FDR**
Single organism biosynthetic process	15	1.91e-08
Small molecule biosynthetic process	9	1.58e-06
Carboxylic acid biosynthetic process	8	1.58e-06
Lipid biosynthesis process	9	1.15e-05
Small molecule metabolic process	14	3.55e-05
Monocarboxylic acid biosynthesis process	6	9.26e-05

***Table 7**: Functionally enriched KEGG pathways in the network of proteins encoded by downregulated group of DE genes*

Pathway description	**Count in gene set**	**FDR**
Metabolic pathways	23	1.13e-23
Tyrosine metabolism	4	1.89e-05
Metabolism of xenobiotics by cytochrome P450	4	0.000107
Drug metabolism – cytochrome P450	4	0.000107

The following scatterplots show average log-ratio between SSc-ILD class and the Control class, with an emphasis on extremely upregulated genes (Figure 1a) and extremely downregulated genes (Figure 1b).

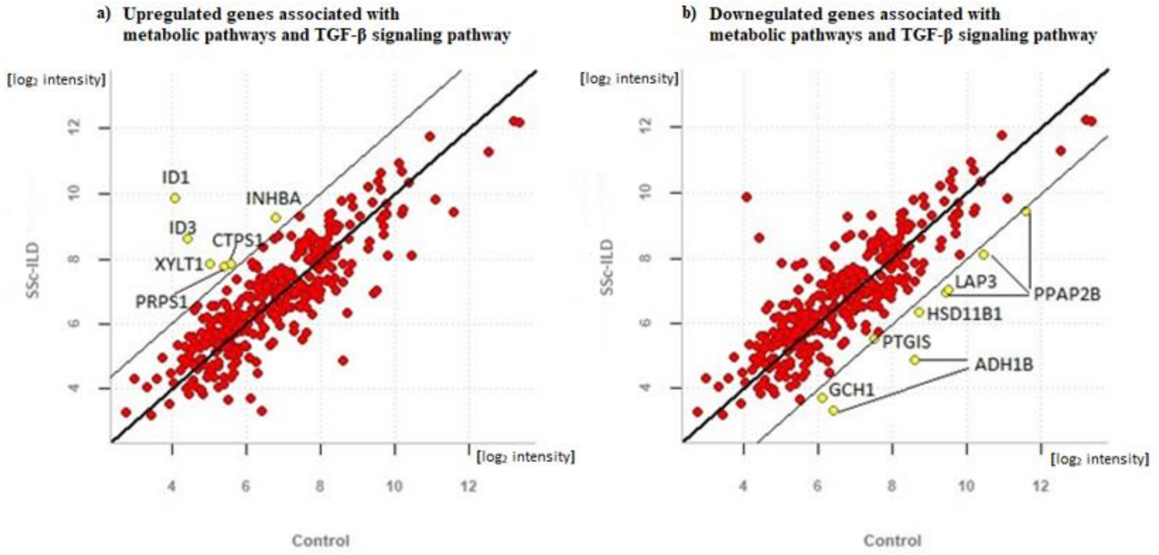

Figure 1: Scatterplots for upregulated and downregulated genes (in SSc-ILD compared to controls) associated with Metabolic pathways and TGF-β pathway
Definition of up/downregulation is based on a logFC>4 and logFC<4 respectively, which is visualised as a line parallel to the identity line. The farther away the point is from identity line, the larger the difference is between its expression in SSc-ILD class and control class. Points above the identity line represent genes with higher expression values in SSc-ILD. Points below the identity line represent genes with higher expression values in controls).

The six upregulated genes shown in the scatterplot (Figure 1a) are ID1, ID3, XYLT1, CTPS1, INHBA and PRPS1. Their interactions are shown in the following scheme produced by STRING analysis (Figure 2).

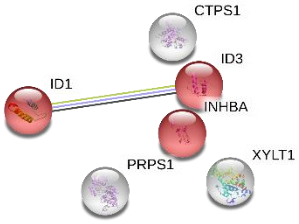

Figure 2: Protein-protein interactions between six upregulated genes (in SSc-ILD compared to controls)
In this network, there are six proteins and one predicted protein-protein association. Interaction between ID1 and ID3 is marked with three distinct colours. Black line represents co-expression, light purple line indicates protein homology, and yellow line represents connection based on textmining. Red coloured nodes represent proteins included in TGF-β pathway.

Genes ID1 (inhibitor of DNA binding 1, HLH protein) and ID3 (inhibitor of DNA binding 1, HLH protein 3) are farthest from the identity line (Figure 1a), compared to other upregulated genes and are both included in TGF-β signalling pathway. They are transcriptional regulators (repressors) associated with cell growth, apoptosis, senescence and differentiation. Among upregulated genes is also INHBA (Inhibin Beta A subunit) which encodes a member of the TGF-β superfamily of proteins. All the above-mentioned genes (ID1, ID3 and INHBA) are red coloured in Figure 2. Other genes are all included in different pathways. Another upregulated gene is XYLT1 (Xylosyltransferase 1) which encodes a protein that catalyses a transfer reaction necessary for biosynthesis of glycosaminoglycan chains in fibroblasts. The last two mentioned upregulated genes which are very closely positioned in the scatterplot are PRPS1 (Ribose-phosphate pyrophosphokinase 1) and CTPS1 (CTP synthase 1). They both encode enzymes which are involved in nucleotide biosynthesis (Stelzer et al., 2016).

The six downregulated genes shown in the scatterplot (Figure 1b) are GCH1, PTGIS, ADH1B (two probe sets), HSD11B1, LAP3 and PPAP2B (three probe sets). Their interactions are shown in the following scheme (Figure 3).

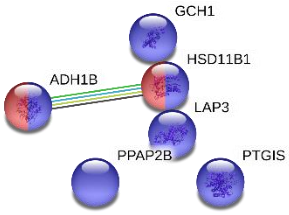

Figure 3: Protein-protein interactions between six downregulated genes (in SSc-ILD compared to controls)
In this network, there are six proteins and one predicted protein-protein association. Interaction between ADH1B and HSD11B1 is marked with four distinct colours. Black line represents co-expression, yellow line represents connection based on textmining, light blue line represents known interaction from curated databases and green line shows predicted interaction based on gene neighbourhood. All purple marked nodes represent proteins included in metabolic pathway and two red marked nodes represent proteins involved in metabolism of xenobiotics by cytochrome P450.

ADH1B (alcohol dehydrogenase 1B (Class I), Beta polypeptide) gene is the farthest from the identity line (Figure 1b), which indicates the greatest difference in its expression in control group compared with SSc-ILD group. Another two downregulated genes are HSD11B1 (hydroxysteroid 11-Beta dehydrogenase 1) and PTGIS (Prostaglandin I2 Synthase). They are associated with Metabolism of xenobiotics by cytochrome P450. PPAP2B (phosphatidic acid phosphatase type 2B) has a role in Metabolic pathways controlling the synthesis of glycerophospholipids (component of membranes – important during rapid growth) and triacylglycerols (storage of metabolic energy). Although among the six most downregulated genes, GCH1 (GTP cyclohydrolase 1), associated with eNOS activation and regulation and LAP3 (leucine aminopeptidase 3), presumably involved in the processing and regular turnover of intracellular proteins, are not involved in any of enriched pathways described in previous section (Stelzer et al., 2016).

3.2.2 Genes from Metabolic pathways associated with Idiopathic pulmonary fibrosis (GSE44723)

Similar to analysis of the SSc dataset, comparison of gene expression levels among fibroblasts with profibrotic phenotype (steady IPF class and rapidly progressing IPF class) was performed. It resulted in a list of 59 probe sets representing 42 genes (Table A29). Their logFC values range from -2.06 to 3.64. Forty-one probe sets (70%) representing 30 genes have logFC values greater than 0, which means they are upregulated in rapidly progressing IPF class compared to steady IPF class. Eighteen probe sets (30%) representing 12 genes have logFC values less than 0, which means they are downregulated in rapidly progressing IPF class compared to steady IPF class. There are 13 genes with multiple probe sets (GALNT7, ME2, GALNT6, BMP2, MEF2C, RDH10, PTGS1, DHFR, PAICS, TYMS, DTYMK, HADH and GK). As in analysis of previous dataset, all probe sets representing 42 genes are visualised in a heatmap (Figure A2). Notably, sample IPF 4 - annotated as rapidly progressing - clusters with the steady state samples with IPF 6 exhibiting the most similar expression profile.

Out of 42 DE genes, two are included in TGF-β signalling pathway (BMP2 and THBS1). They are both downregulated in rapidly progressing IPF class. We observe that upregulated gene MEF2C and downregulated gene TBL1X are directly implicated in Mitochondrial biogenesis pathway. Only gene ME2, which is shown to be upregulated, is involved in Pyruvate metabolism. Furthermore, four genes are included in Oxidative phosphorylation. Three of them (PPA1, COX15 and ATP6V0E2) are downregulated and one of them (TC1RG1) is upregulated. Out of these four genes, only PPA1 is not additionally implicated in KEGG’s gene list of Metabolic pathways, which includes 36 remaining DE genes. One of those genes (HADH), which is upregulated, is involved in Fatty acid degradation. Amongst remaining 36 DE genes are also downregulated gene ALDH1A3 and upregulated gene PGM2, which are additionally included in Glycolysis/gluconeogenesis pathway.

As in analysis of SSc dataset, STRING was used to identify possible connections among genes within two groups of genes (downregulated and upregulated). PRIM2 was excluded from upregulated group because STRING database does not include protein by this identifier in organism Homo sapiens. Connections among remaining proteins are visually presented in two separate schemes (Figures A5 and A6). Additionally, analysis of significantly enriched GO-BP and KEGG pathways showed upregulation in Oxidative phosphorylation and processes associated with Nucleotide biosynthesis, Purine and Pyrimidine metabolism (Tables 8 and 9). Downregulation was observed in Carbohydrate metabolic processes, Glycoprotein metabolism, Retinoic acid biosynthetic processes and in Lipid biosynthesis (Tables 10 and 11).

Table 8*: Functionally enriched GO-BP in the network of proteins encoded by upregulated group of DE genes*

Pathway description	Count in gene set	FDR
Nucleotide biosynthetic process	10	1.1e-09
Single-organism metabolic process	23	1.12e-08
Nucleoside phosphate biosynthetic process	9	2.01e-08
Nucleotide metabolic process	11	3.07e-08

Table 9*: Functionally enriched KEGG pathways in the network of proteins encoded by upregulated group of DE genes*

Pathway description	Count in gene set	FDR
Metabolic pathways	26	2.74e-27
Purine metabolism	10	1.97e-12
Pyrimidine metabolism	8	9.25e-11
One carbon pool by folate	4	8.92e-07
DNA replication	4	9.49e-06
Oxidative phosphorylation	3	0.0358
Mucin type O-Glycan biosynthesis	2	0.0358

Table 10*: Functionally enriched GO-BP in the network of proteins encoded by downregulated group of DE genes*

Pathway description	Count in gene set	FDR
Carbohydrate derivative metabolic process	7	0.00348
Glycoprotein metabolic process	5	0.009
Single organism metabolic process	10	0.009
Retinoic acid biosynthetic process	2	0.0114
Lipid biosynthetic process	5	0.0114

***Table 11**: Functionally enriched KEGG pathways in the network of proteins encoded by upregulated group of DE genes*

Pathway description	**Count in gene set**	**FDR**
Metabolic pathways	9	3.2e-07

The following scatterplots show average log-ratio between rapidly progressive IPF class and steady IPF class, with an emphasis on extremely upregulated genes (Figure 4a) and extremely downregulated genes (Figure 4b).

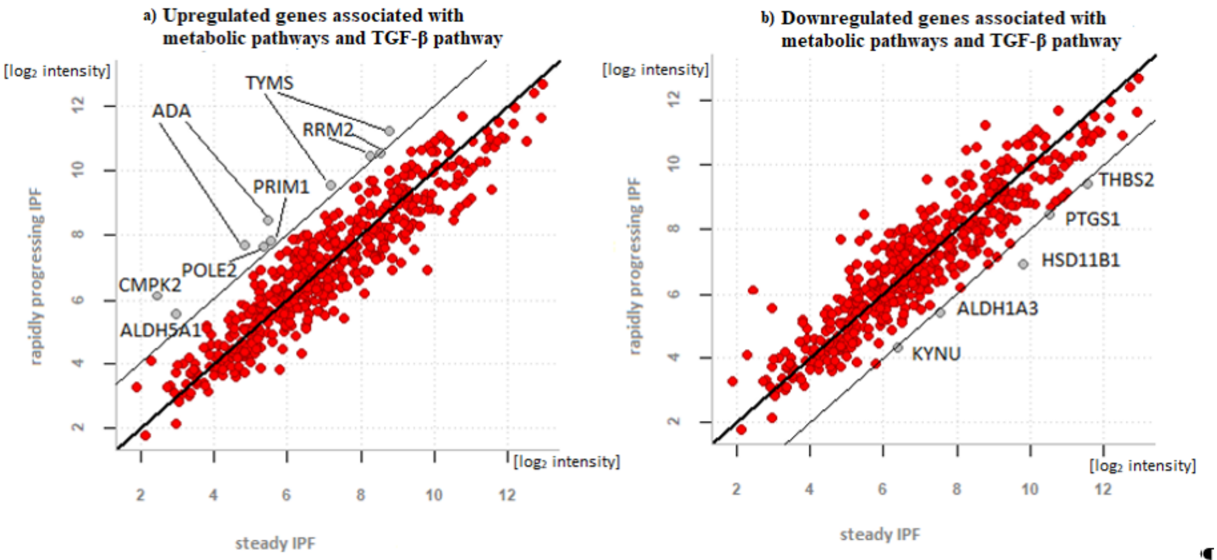

Figure 4: Scatterplots for upregulated and downregulated genes (in rapidly progressing IPF class compared to steady IPF class) associated with Metabolic pathways and TGF-β pathway
Definition of up/downregulation is based on a logFC>4 and logFC<4 respectively, which is visualised as a line parallel to the identity line. The farther away the point is from identity line, the larger the difference is between its expression in rapidly progressing IPF class and steady IPF class. Points above the identity line represent genes with higher expression values in rapidly progressing IPF. Points below the identity line represent genes with higher expression values in steady IPF fibroblasts.

The seven upregulated genes shown in the scatterplot (Figure 4a) are ADA (two probe sets), TYMS (two probe sets), RRM2 (two probe sets), PRIM1, POLE2, CMPK2, ALDH5A1. Based on the STRING analysis, these seven genes encode proteins which are at least partially biologically connected, as a group. Their interactions are shown in the following scheme (Figure 5).

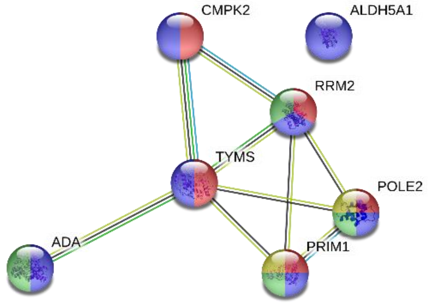

Figure 5: Protein-protein interactions between seven upregulated genes (in rapidly progressing IPF compared to steady IPF)
In this network, there are seven proteins and nine edges (predicted protein-protein associations). The number of interaction indicates that the proteins are at least partially biologically connected, as a group. Interactions between proteins are marked with four distinct colours. Yellow line represents interactions based on text mining, black line indicates interactions based on co-expression, green line shows predicted interaction based on gene neighbourhoods, and light blue line represents known interactions from curated databases.

All proteins are purple marked which represents their inclusion in Metabolic pathways. ALDH5A1 is the only gene, of which proteins have no predicted interactions with the others. Five genes (CMPK2, RRM2, TYMS, POLE2 and PRIM1) are all included in Pyrimidine metabolism pathway (red coloured nodes). Three of them (RRM2, POLE2 and PRIM1) are also involved in DNA replication (yellow coloured nodes) and with the addition of ADA play a role in Purine metabolism (green coloured nodes).

Downregulated genes shown in the scatterplot (Figure 4) are KYNU, ALDH1A3, HSD11B1, PTGS1 and THBS2. As in visualisation of upregulated genes and their interactions, the STRING analysis produced a network of five downregulated genes which is shown in the following scheme (Figure 6).

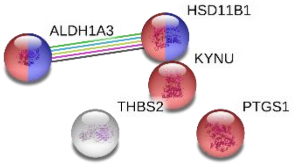

Figure 6: Protein-protein interactions between five downregulated genes (in rapidly progressing IPF compared to steady IPF)
In this network, there are five proteins and one predicted protein-protein association. The lack of associations does not necessarily mean that this group of genes has no important biological connection – their interactions might not be known yet. Interaction between ALDH1A3 and HSD11B1 is marked with five distinct colours. Black line indicates interactions based on co-expression, yellow line represents connection based on textminig, green line shows predicted interaction based on gene neighbourhoods, and light blue line represents known interactions from curated databases which are also experimentally determined (pink line).

Four out of five downregulated genes of which nodes are red coloured are included in Metabolic pathways (ALDH1A3, HSD11B1, KYNU and PTGS1). Two of them (ALDH1A3 and HSD11B1), which are the only ones with protein-protein associations, are also included in Metabolism of xenobiotics by cytochrome P450 pathway (purple coloured nodes). In addition, ALDH1A3 is involved in Glycolysis/gluconeogenesis. Lastly, gene THBS2 is associated with TGF-β signalling pathway.

3.2.3. Comparison of analysis of gene expression levels in both datasets

We observe a similar percentage of downregulated genes (approximately 30%) and upregulated genes (approximately 60%). In addition, we detect four common pathways which include different DE genes; Fatty acid degradation, Glycolysis/gluconeogenesis, Pyruvate metabolism and Oxidative phosphorylation. Out of those four pathways, only Glycolysis/gluconeogenesis has one common DE gene (ALDH1A3), which is downregulated in both SSc-ILD group and rapidly progressing IPF group. There are no other common DE genes in remaining three pathways. STRING predicted protein-protein association analysis shows a higher number of connections between upregulated genes rather than between downregulated genes in both datasets.

3.3 Analysis of Mitochondrial biogenesis genes

Our aim was to identify which of the genes involved in mitochondrial biogenesis contribute the most to the pathologic activation of fibroblasts in patients with SSc-ILD and IPF. Thus, we included 63 genes (Table A1) to detect any meaningful change in their expression.

3.3.1 Mitochondrial biogenesis genes associated with Scleroderma associated interstitial lung disease (GSE40839)

After the data import and normalization, 22,283 probe sets were available for the analysis. After removing all probe sets that are not associated with genes included in our Mitochondrial biogenesis gene list, 121 probe sets (Table A30) remained available for the differential expression analysis. Comparison of gene expression resulted in a list of 18 probe sets, representing 14 genes – 11 upregulated genes (Table 12) and three downregulated genes (Table 13).

***Table 12**: List of probe sets, representing upregulated DE genes (α=0.01) of Mitochondrial biogenesis genes in SSc-ILD compared to controls (sorted by parametric p-value)*

Probe set	Gene symbol	Parametric p-value	logFC
208905_at	CYCS	7.32E-05	-0.97
210046_s_at	IDH2	0.000116	-1.15
201322_at	ATP5B	0.000228	-0.45
218590_at	C10orf2 (TWNK)	0.000315	-0.40
219169_s_at	TFB1M	0.002148	-0.42
216326_s_at	HDAC3	0.005818	-0.27
203737_s_at	PPRC1	0.007069	-0.40
202474_s_at	HCFC1	0.007228	-0.25
202591_s_at	SSBP1	0.007233	-0.58

Probe set	**Gene symbol**	**Parametric p-value**	**logFC**
218605_at	TFB2M	0.008826	-0.47
211984_at	CALM1	0.008960	-0.38

Table 13*: List of probe sets, representing downregulated DE genes (α=0.01) of Mitochondrial biogenesis genes in SSc-ILD compared to controls (sorted by parametric p-value)*

Probe set	**Gene symbol**	**Parametric p-value**	**logFC**
215223_s_at	SOD2	< 1e-07	3.30
216841_s_at	SOD2	2.00E-07	2.96
221477_s_at	SOD2	6.00E-07	2.94
209107_x_at	NCOA1	0.001445	0.42
209105_at	NCOA1	0.002558	0.28
212867_at	NCOA2	0.004628	0.68
209106_at	NCOA1	0.008021	0.42

Among DE genes, we do not observe PPARGC1A, PPARGC1B, NRF1, NRF2 or TFAM – genes which play major roles in mitochondrial biogenesis. However, we do observe upregulation of PPRC1, which encodes a protein belonging to the same family as PPARGC1A, which can activate mitochondrial biogenesis (Stelzer et al., 2016) in response to proliferative signals. We also detect upregulation of TFB1M and TFB2M (nuclear-encoded, mitochondria-targeted transcription factors), genes which are necessary for mitochondrial gene expression – similar to TFAM (Litonin et al., 2010). Additionally, we observe upregulated C10orf2 (TWNK) and a housekeeping gene SSBP1, both involved in mitochondrial DNA replication, along with HDAC3, which plays a critical role in transcriptional regulation. Furthermore, we observe upregulation of mitochondrial proteins ATP5B, CYCS and IDH2 which are involved in OXPHOS, and TCA cycle, Moreover, we detect upregulation of HCFC1, involved in metabolism of proteins. In addition, we observe upregulation of CALM1, which encodes one of the four subunits of phosphorylase kinase (Stelzer et al., 2016). This upregulation is viewed as important, because Ca2+/calmodulim-based signalling is one of the triggers for mitochondrial biogenesis (Michel et al., 2007) – PGC1a activation. Lastly, we detect downregulation of two transcriptional activators NCOA1 and NCOA2, together with SOD2, a gene encoding a major antioxidant protein, which detoxifies superoxide anion radicals generated by mitochondrial respiration (Weisiger & Fridovich, 1973). Of note, HDAC3, HCFC1 and NCOA1 are all involved in chromatin modifying functions – histone acetylation is catalysed by histone acetyl transferases, whereas the reverse reaction is performed by histone deacetylases (Legube & Trouche, 2003; Wysocka et al., 2003)

The following scatterplots show average log-ratio between SSc-ILD class and Control class, with an emphasis on extremely upregulated genes (Figure 7a) and extremely downregulated genes (Figure 7b).

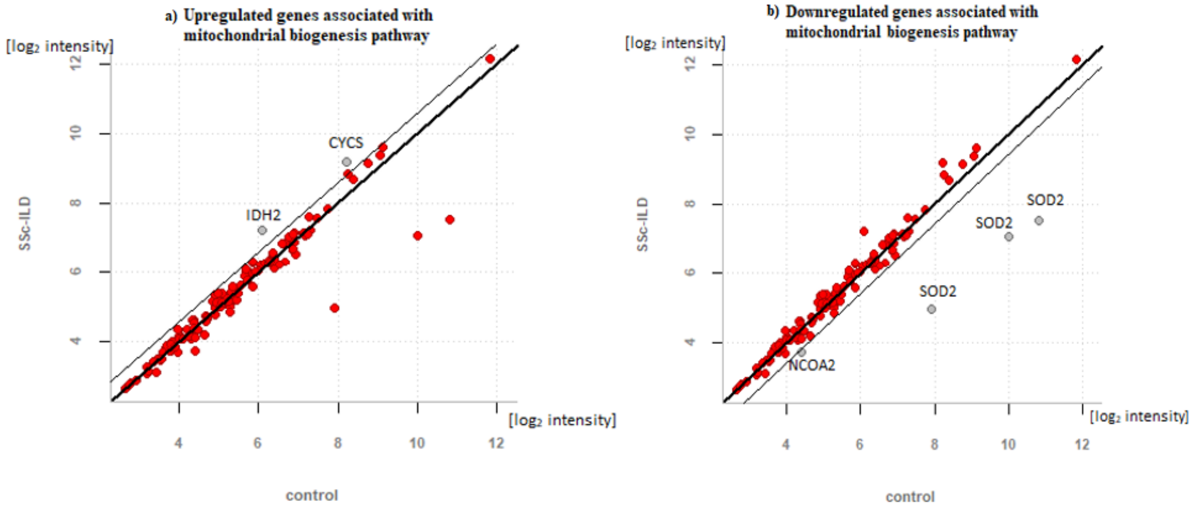

Figure 7: Scatterplots for upregulated and downregulated genes (in SSc-ILD vs. controls associated with Metabolic pathways and TGF-β pathway
Definition of up/downregulation is based on a logFC>1.5 and logFC<1.5 respectively, which is visualised as a line parallel to the identity line. The farther away the point is from identity line, the larger the difference is between its expression in SSc-ILD and controls. Points above the identity line represent genes with higher expression values in SSc-ILD. Points below the identity line represent genes with higher expression values in controls).

With the use of Genesis, the following heatmap was produced (Figure 8).

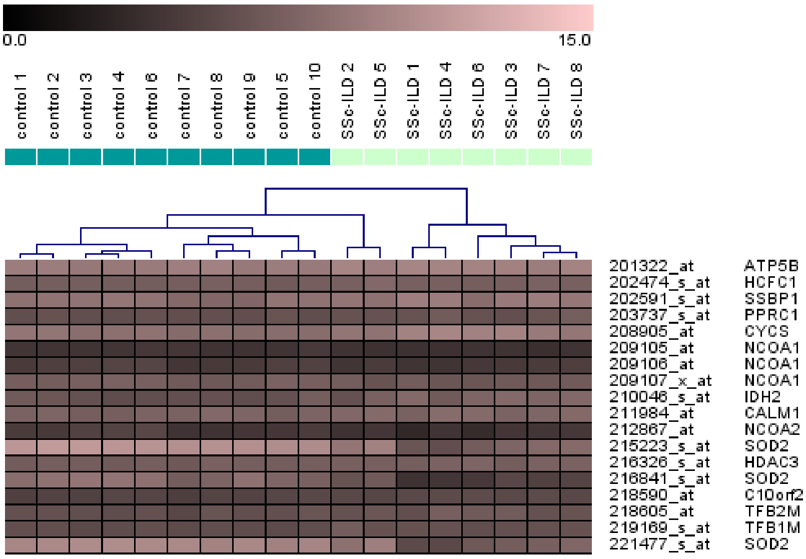

Figure 8: Heatmap of expression values for DE genes (α=0.01) in SSc-ILD and controls
Expression values are represented by black to pink colour gradient, ranging from 2.95 to 11.40 (lowest values in black and highest values in light pink).

3.3.2 Mitochondrial biogenesis genes associated with Idiopathic pulmonary fibrosis (GSE44723)

After the data import and normalization, 54,675 probe sets were available for the analysis. After removing all probe sets that are not associated with genes included in our Mitochondrial biogenesis gene list, 197 probe sets (Table A31) remained available for differential expression analysis. Comparison of gene expression resulted in a list of eight probe sets representing six genes – three upregulated genes (Table 14) and three downregulated genes (Table 15).

Table 14*: List of probe sets, representing upregulated DE genes by differential expression analysis (α=0.01) of Mitochondrial biogenesis genes in rapidly progressing IPF class compared to steady IPF (sorted by parametric p-value)*

Probe set	Gene symbol	Parametric p-value	logFC
200854_at	NCOR1	0.000744	0.51
209199_s_at	MEF2C	0.001744	1.55
209200_at	MEF2C	0.002325	1.30
205811_at	POLG2	0.003354	0.73

Table 15*: List of probe sets, representing upregulated DE genes by differential expression analysis (α=0.01) of Mitochondrial biogenesis genes in rapidly progressing IPF class compared to steady IPF (sorted by parametric p-value)*

Probe set	Gene symbol	Parametric p-value	logFC
204760_s_at	NR1D1	0.0032302	-0.42
201868_s_at	TBL1X	0.0071714	-0.38
1566932_x_at	TFB2M	0.0080112	-0.32
213400_s_at	TBL1X	0.0086882	-0.62

As in previous analysis in this section (dataset GSE40389), we do not observe any statistically significant change in expression of genes, which play major roles in mitochondrial biogenesis. However, we do observe upregulation of TFAM on the scatterplot (Figure 9a), of which upregulation is based on a 1.5-fold change. It is not included in Table 14, because its p-value (0.037379) exceeds the selected significance level.

The two upregulated genes NCOR1 and POLG2 are associated with organelle biogenesis and maintenance, with POLG2 (mitochondrial DNA polymerase-gamma) being additionally implicated in mitochondrial gene expression (Stelzer et al., 2016). As in previous dataset, we observe DE gene TFB2M, which is in contrast downregulated in this dataset. Lastly we notice upregulation of MEF2C, an important transcription factor upregulating transcription of PGC1a in response to various stimuli (Fernandez-Marcos & Auwerx, 2011) and downregulation of TBL1X and NR1D1. None of them are additionally implicated in any of the metabolic pathways discussed so far. NCOR1 and TBL1X are both involved in repression of transcription following retinoid and thyroid receptor signalling and NR1D1 acts as a receptor for heme which stimulates its interaction with the NCOR1/HDAC3 corepressor complex (Stelzer et al., 2016). It also represses expression of PPARGC1A, a potent inducer of heme synthesis (Singh et al., 2016).

The following scatterplots show average log-ratio between rapidly progressive IPF class and steady IPF class, with an emphasis on extremely upregulated genes (Figure 9a) and extremely downregulated genes (Figure 9b).

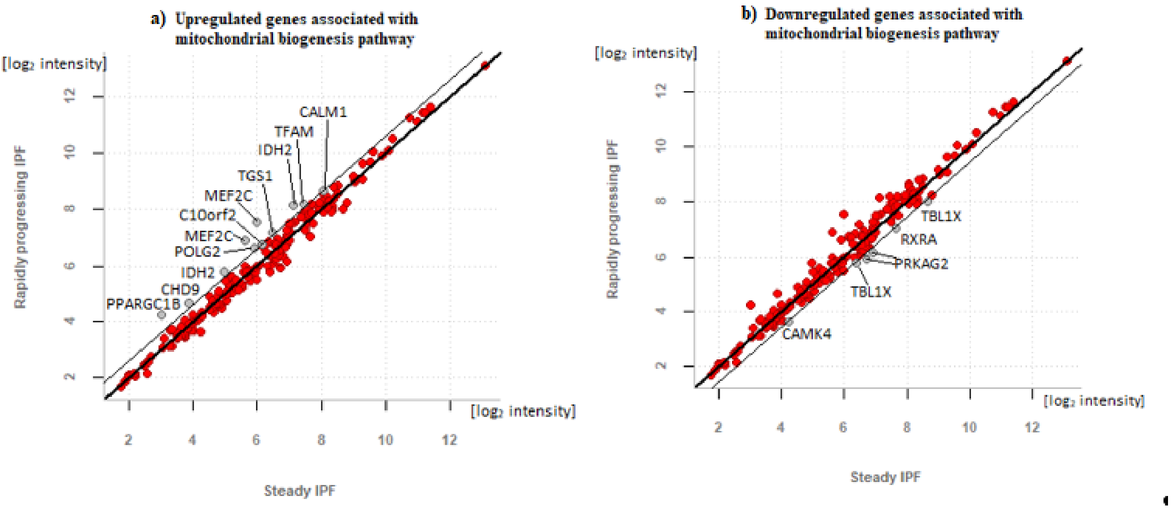

Figure 9: Scatterplots for upregulated and downregulated genes (in rapidly progressing IPF vs. steady IPF) associated with Metabolic pathways and TGF-β pathway
Definition of up/downregulation is based on a logFC>1.5 and logFC<1.5 respectively, which is visualised as a line parallel to the identity line. The farther away the point is from identity line, the larger the difference is between its expression in rapidly progressing IPF and steady IPF. Points above the identity line represent genes with higher expression values in rapidly progressing IPF. Points below the identity line represent genes with higher expression values in steady IPF.

With the use of Genesis, the following heatmap was produced (Figure 10).

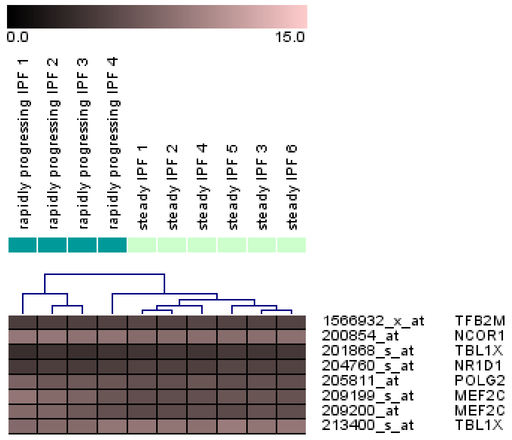

Figure 10: Heatmap of expression values for DE genes (α=0.01) in rapidly progressing IPF and steady IPF
Expression values are represented by black to pink colour gradient, ranging from 3.40 to 9.07 (lowest values in black and highest values in light pink).

3.3.3 Comparison of analysis of gene expression levels in both datasets

We observe greater percentage of probe sets representing DE genes in SSc-ILD controls (approximately 15%) as in rapidly progressing IPF vs. steady IPF (approximately 4%). In addition, we detect one common DE gene (from differential expression analysis; α=0.01) – TFB2M, which is upregulated in SSc-ILD group and downregulated in rapidly progressing IPF group. In addition, we observe one other common gene (in scatterplots Figures 7a and 9a; differential expression based on FC=1.5) – IDH2.

4 DISCUSSION

Fibrosis, a hallmark of SSc and IPF, which is defined by the accumulation of ECM, is seen as the central pathological process in their disease development. With the analysis of publicly available gene expression profiles of lung fibroblasts from patients with both diseases, we examined underlying mechanisms which may lead to their progression and represent potential therapeutic targets. In this study we set a specific question whether Mitochondrial biogenesis and Metabolic pathways play a role in progression of SSc and IPF.

We believe that our study is the first to focus on metabolic genes and their expression profiles of SSc fibroblast cells, while other recent studies have been focusing on metabolism in SSc monocyte-derived macrophages (Moreno-Moral et al., 2018) and IPF lung tissue (Zhao et al., 2017). Our GSEA of all genes in SSc-ILD compared to controls revealed 39 enriched pathways of which four are associated with Carbohydrate and Lipid metabolism. The same analysis of all genes in stable IPF compared to rapidly progressing IPF, revealed six out of 29 enriched pathways which are associated with Nucleotide and Amino acid metabolism. Comparing results of both datasets we have confirmed typical changes expected in highly proliferative cells, such as increased glycolysis, increased metabolism of purines and pyrimidines, along with increased DNA replication.

Keeping in mind that very small changes in expression of enzymatic genes can have greater consequences than huge changes in expression of cytoskeletal genes, we reduced the number of investigated genes. With the analysis of this subset, containing genes associated with metabolism and TGF-β pathway as a positive control, we showed increased expression of enzymes involved in all three stages of cell respiration (glycolysis, TCA and OXPHOS) with predominant increase in glycolysis with 29 of all DE genes (Figures A7, A8 and A9). This may indicate that fibroblasts in SSc have high energy demand. They utilize a metabolic switch occurring in highly proliferative cells, known as the Warburg effect, to produce sufficient energy to function, although glycolysis does not provide the majority of energy – only 2 ATP molecules in contrast with OXPHOS which produces 36 ATP molecules per glucose molecule. According to Jiang (2017), the Warburg effect is observed in many cancer cells where glycolysis is utilised as a primary energy source even in the presence of sufficient amounts of oxygen. This process is called aerobic glycolysis. It is then followed by pyruvate conversion to lactic acid, instead of entering TCA cycle and represents the imbalance between maximum rate of glycolysis and pyruvate oxidation (Jiang, 2017). There are a few substances that actuate OXPHOS rather than glycolysis, such as the polyphenol Resveratrol which decreases the activity of the pentose phosphate pathway and lipogenesis in cells with high proliferation, such as cancer cells (Saunier et al., 2017). We consider this actuation of OXPHOS as a potential therapeutic target for both SSc and IPF.

Furthermore, our results showed dysregulation of genes associated with Sphingolipid metabolism (Figure A10) in both diseases, which implies disruption in sphingosine-1-phosphate (S1P) production. Pyne et al. (2013) stated that S1P is an endogenous bioactive lipid which mediates a variety of biological cell responses, such as cell proliferation, cell migration, cell differentiation and apoptosis. It is generated from sphingosine through sphingosine kinase (SPHK)-activated phosphorylation and may be dephosphorylated by cell surface proteins lipid phosphate phosphatases (LPPs) PPAP2A and PPAP2B (Pyne et al., 2013). Our results showed downregulation of these two genes (in SSc). They hydrolyse lysophosphatidate (LPA), a potent signalling molecule that accelerates lung fibrosis in IPF (Benesch et al., 2016) and acts as critical contributor to scleroderma skin fibrosis (Castelino et al., 2016). Clinical trials regarding antagonists of the LPA1 receptor, have been reported – antagonist SAR100842 (Allanore et al., 2015) as a potential treatment for SSc and antagonist BMS-986020 (Rosen et al., 2017) for treatment of IPF. Additionally, our analysis showed downregulation of UDP-glucose ceramide glucosyltransferase (UGCG) gene and upregulation of sphingosine-1-phosphate lyase 1 (SGPL1) gene in SSc. The results indicate disrupted Fat digestion and absorption. In comparison, we noticed downregulation of gene SPHK1 in IPF, also detected in another study (Zhao et al., 2017), which, as already mentioned, implies disruption of S1P production. Notably, all dysregulated genes mentioned in this paragraph are in close proximity to S1P – they are all involved in the sphingomyelin cycle (Meshcheryakova et al., 2016), suggesting their direct relation to S1P levels. The S1P signalling pathway has already been proposed as a potential therapeutic target in SSc (Pattanaik & Postlethwaite, 2010), IPF (Huang & Natarajan, 2015) and other fibrotic diseases (Gonzalez-Fernandez et al., 2017) with S1P receptor antagonists and SPHK inhibitors as developing drugs.

Our study found increased expression of genes ODC1, AMD1, SRM and ASL in SSc, which are all associated with Arginine metabolism (Figure A11). Arginine metabolites are known to be involved in different sections of fibrotic process. Arginine is converted to ornithine and further to polyamines spermidine and putrescine required for cell proliferation. This conversion process is catalysed by enzymes among which are those encoded by genes ODC1, AMD1, SRM and ASL. In mitochondria, arginine can also be converted to proline and its metabolite hydroxyproline, both essential in collagen synthesis. We showed decreased levels of gene LAP3, encoding enzyme, involved in Proline metabolism (Figure A11). Additional findings in our analysis of SSc fibroblasts are decreased levels of GLUL which catalyses the synthesis of glutamine and increased levels of GLS which catalyses the hydrolysis of glutamine in mitochondria (Figure A11). It has been recently reported that glutaminolysis is required for TGF-β1-induced myofibroblast differentiation and activation (Bernard et al., 2018). We suggest that further exploration of the glutaminase inhibitor CB 839 (Bromley-Dulfano et al., 2013), an agent with potential antineoplastic activity, as an antifibrotic drug, could be beneficial. Comparing pathophysiological mechanisms of SSc and IPF, we could say that our findings are in accord with other studies (Zhao et al., 2017). They also showed increases in arginine metabolites, decreased aspartate levels and elevated glutamate levels in IPF – we did not detect changes in genes associated with these metabolites and enzymes when comparing rapid and slow progressing IPF. This indicates that aforementioned underlying mechanisms are not the main factors that promote faster progression of IPF.

Arachidonic acid (AA), a fatty acid present in cell membranes, is the precursor of a family of biologically and clinically important molecules, known as eicosanoids (including prostaglandins among others). AA is metabolized by the subsequent activities of cyclooxygenase (COX). We believe that our study is the first one to show upregulation of the gene PTGS1/COX1 in SSc, which is a common target of nonsteroidal anti-inflammatory drugs, such as Aspirin. In accordance with Ricciotti and FitzGerald (2011), we interpret that upregulation of PTGS1 (COX1), encoding the enzyme that converts AA to prostaglandin H2 (PGH2), taken together with downregulation of genes PTGES and PTGIS, causes reduced conversion of PGH2 to PGE and PGI. This results in reduced vasodilation and platelet activation. Consequently, there is more than the usual amount of PGH2 available for conversion to PGD2- causing bronchoconstriction, PGF2 and TXA – causing vasoconstriction and platelet activation (Ricciotti & FitzGerald, 2011). We also detected downregulation of COX1 gene in IPF. Altogether, our findings indicate dysregulation in AA metabolism (Tables A12 and A13) and insinuate that this pathway could represent a potential therapeutic target.

Krug et al. (2009) suggested that perturbations in AA metabolic pathways could lead to development of pulmonary hypertension (PH), which is in most cases caused by pulmonary fibrosis. Furthermore, altered production of vasodilator and vasoconstrictor AA metabolites (eicosanoids), such as PGI2, PGD2, PGE2 and PGF2α, plays an important role in pathophysiology of PH, one example being the lack of vasodilator PGI2 and its analogues. An analogue of PGI2 called iloprost, with antithrombotic, antiproliferative and anti-inflammatory characteristics which contribute to pathogenesis of PH, is available for treatment of this disease (Krug et al., 2009). Based on our results and taking into consideration findings of previous studies (C. Foti et al., 2004; R. Foti et al., 2017; Krug et al., 2009; Lasota et al., 2013), we consider treatment of SSc and even IPF with iloprost as a viable option.

Lastly, we sought to identify dysregulated genes in Mitochondrial biogenesis. The results regarding expression of gene IDH2 in SSc were consistent with the results of previous analysis (Metabolic pathways) on much greater number of genes. In both subsets, upregulation of IDH2, an enzyme that catalyses the oxidative decarboxylation of isocitrate to α-ketoglutarate, was detected. Since previous studies suggest an association between ILDs and lung cancer development based on similar characteristics (Archontogeorgis et al., 2012) upregulation of IDH2 is in agreement with findings from Li et al. (2018) who found increased expression of IDH2 in blood lymphocytes from patients with lung cancer compared to controls (Li et al., 2018). Additionally, evidence from previous studies suggests that lung scarring caused by IPF represents a risk factor for lung carcinogenesis (Karampitsakos et al., 2017). IDH converts isocitrate to α-ketoglutarate after which glutamate from glutaminolysis (already identified in our research as dysregulated pathway) enters TCA cycle (Li et al., 2018). There exists one substance called enasidenib, which acts as an inhibitor of mutant IDH2 enzyme and is currently used for treatment of acute myeloid leukaemia in patients with specific mutations in the IDH2 gene (Stein, 2018). Taken together, we propose exploration of treatment with enasidenib for patients with SSc and IPF.

Although our study was carefully prepared and has reached its aims, there were some limitations. First, when determining which genes are included in Mitochondrial biogenesis pathway, we selected genes that are encoded only by nuclear DNA (nDNA) and not by mitochondrial DNA (mtDNA), therefore analysing solely mtDNA encoded genes would be promising for a further and more specific study. Second, regarding the SSc-ILD dataset, the clinical data lacked severity classification, thus leaving us without the option to determine which genes, if any, contribute the most to progression of the disease. In contrast, when studying IPF, we could only compare stable and rapidly progressing phenotypes without a control group. Despite the lack of a control group, we are confident that comparing two different progressions of IPF has great benefits. For example, it gives us the ability to determine the genes with the most significant impact on disease development and progression. It can also contribute to further development of molecular diagnostic testing of the disease. Third, IPF fibroblasts were in culture for several passages (up to the 11th passage) which means that they grew under the same conditions. These cultured fibroblasts lack a wide variety of cytokines emitted by the immune/blood cells that are no longer present after the extraction from a patient. Therefore, with every passage, the fibroblasts may become less activated, which may be the reason why the difference between rapidly progressing and steady IPF is not as significant as expected – the samples do not cluster according to disease state (Figure A2). Another explanation for this unexpected clustering could be that some other cell type, which is not investigated in this study, contributes more to the severity of IPF than fibroblasts.

For future work, we suggest the analysis of IPF versus control and also analysis of different levels of severity in SSc, with the aim to determine if there exists an overlap with the DE genes discovered in this study.

5 CONCLUSION

The purpose of this study was to investigate if there are any changes in mitochondrial metabolism and mitochondrial biogenesis that have a significant role in development and progression of SSc and IPF. Our bioinformatic analysis incorporated publicly available gene expression data from ten patients with histologically normal lung tissue, eight patients with SSc and ten patients with IPF.

GSEA of SSc dataset shows 38 enriched gene sets/pathways (Table A3) when analysing all genes and eight enriched pathways (Table 2) when analysing a subset of genes associated with metabolism. The same analysis of IPF dataset shows 29 enriched pathways (Table A11) when analysing all genes and nine enriched pathways (Table 3) when analysing a subset of genes associated with metabolism.

As expected, the results reveal increased glycolysis, increased metabolism of purines and pyrimidines, as well as increased DNA replication in both diseases. Furthermore, the results show perturbed expression of enzymes involved in TCA and OXPHOS. These profound metabolic changes may reflect increased energy demand of highly proliferative cells and corresponding pathways should be further elucidated with the aim to find effective treatment options.

In addition, results confirm changes in pathways that are already therapeutic targets for potential treatments of SSc and IPF, such as Sphingolipid metabolism, AA metabolism and Arginine metabolism.

Lastly, gene expression analysis on genes associated with mitochondrial biogenesis (which was possible only after we created Mitochondrial biogenesis gene list) shows 18 DE genes in SSc dataset and eight DE genes in IPF dataset. Although these results suggest that this process is crucially affected in fibroblasts associated with both diseases, it should be further addressed with more specific experiments, for definitive conclusions.

6 POVZETEK NALOGE V SLOVENSKEM JEZIKU

Pljučna fibroza je progresivno brazgotinjenje pljučnega tkiva, ki se pojavlja pri sistemski sklerozi (SS) in intersticijski pljučni fibrozi (IPF), z omejenimi možnostmi zdravljenja. Patofiziološko to stanje opišemo kot prekomerni nastanek medceličnine, katerega povzročajo vztrajno aktivirani fibroblasti, ki diferencirajo v miofibroblaste. Mitohondrijska biogeneza je opredeljena kot proces, preko katerega celice povečujejo svojo posamezno mitohondrijsko maso z rastjo in delitvijo. Ker povečana beljakovinska sinteza in proliferacija celic zahtevata zvišano regulacijo metaboličnih poti, povezanih s stimulacijo mitohondrijske biogeneze, je bil cilj te raziskave pregledati metabolične motnje in mitohondrijsko biogenezo v pljučnih fibroblastih in posledični učinek na patogenezo SS in IPF.

Za bioinformatično analizo je bil uporabljen program BRB-ArrayTools. Analizirana sta bila dva javno dostopna nabora podatkov DNA-mikromrež (GSE40839 – SS fibroblasti in GSE44723 – IPF fibroblasti).

Analiza je bila razdeljena na tri segmente:

1. Analiza obogatenosti genskih skupin/poti na vseh genih.
2. Analiza diferenčne izraženosti genov vključenih v metabolične poti.
3. Analiza diferenčne izraženosti genov vključenih v mitohondrijsko biogenezo.

Surovi podatki naborov SS in IPF so bili ob uvozu v BRB-ArrayTools logaritemsko transformirani ($\log_2$), normalizirani z metodo RMA in anotirani s pripadajočima GPL datotekama. Za nadaljnjo analizo so bile uporabljene KEGG poti, v katere mitohondrijske biogeneza ni bila vključena. Za analizo omenjene poti, je bilo potrebno narediti seznam genov in ga vključiti v že obstoječo bazo BRB-ArrayTools. Oba nabora podatkov sta bila razdeljena v dve skupini (SS v primerjavi s kontrolno skupino ter hitro napredujoča IPF v primerjavi s počasi napredujočo IPF).

Z analizo vseh treh segmentov so bili pridobljeni seznami obogatenih poti (prvi segment) in diferenčno izraženih genov za vsako podmnožico genov (drugi in tretji segment). Za določitev morebitnih funkcijskih interakcij med proteini, ki jih kodirajo diferenčno izraženi geni, je bila uporabljena podatkovna baza STRING.

Rezultati analize prvega segmenta, za SS v primerjavi s kontrolami, so pokazali 39 obogatenih poti, od katerih so štiri povezane s presnovo ogljikovih hidratov in lipidov. Enaka analiza za hitro napredujočo IPF v primerjavi s počasi napredujočo IPF je pokazala 29 obogatenih poti, od katerih je šest povezanih z metabolizmom nukleotidov in aminokislin. Medsebojna primerjava rezultatov je pokazala motnje v metaboličnih poteh, ki so pričakovane v visoko proliferativnih celicah – povišana glikoliza/glukoneogeneza, povišan metabolizem purinov in pirimidinov ter povečana replikacija DNA.

Rezultati analize drugega segmenta (za SS in IPF) so pokazali motnje encimov, vključenih v vse tri stopnje celičnega dihanja – citosolna glikoliza, mitohondrijski cikel citronske kisline in oksidativna fosforilacija. To nakazuje na visoko energetsko zahtevo fibroblastov, kateri z metaboličnim preklopom na aerobno glikolizo proizvedejo dovolj energije za delovanje. Na podlagi rezultatov, obravnavamo aktivacijo oksidativne fosforilacije kot možno terapevtsko tarčo. Opažena je bila tudi sprememba uravnavanja genov, povezanih z metabolizmom sfingolipidov, arginina in prolina ter arahidonske kisline.

Rezultati analize tretjega segmenta so pokazali diferenčno izražene gene v mitohondrijski biogenezi, kar indicira, da je ta proces afektiran tako v SS kot v IPF. Kljub temu, je za dokončne zaključke potrebna bolj podrobna preiskava omenjenega procesa.

Čeprav je naša raziskava dosegla zadane cilje, ni bila brez omejitev. Prvič, ker smo pri določanju genov vključenih v mitohondrijsko biogenezo izbirali gene, ki jih kodira le jedrna DNA, predlagamo dodatno analizo genov kodiranih z mitohondrijsko DNA. Drugič, podatkovni nabor SS ni imel kliničnih podatkov o resnosti bolezenskega stanja, zato nismo imeli možnosti določiti kateri geni največ prispevajo k napredovanju bolezni. Nasprotno, smo pri podatkovnem naboru IPF primerjali le hitro napredujočo v primerjavi s počasi napredujočo IPF brez kontrolne skupine. Kljub pomanjkanju le te, smo prepričani, da ima primerjanje dveh različnih napredovanj bolezenskega stanja veliko korist. Omogoča nam, da določimo gene, ki imajo na razvoj in napredovanje bolezni največji vpliv.

Za nadaljnje raziskave predlagamo analizo s primerjavo IPF in kontrolne skupine ter analizo s primerjavo različnih stopenj resnosti SS, da bi lahko ugotovili, če obstaja prekrivanje genov, z diferenčno izraženimi geni naše raziskave.

7 REFERENCES

[1] Y. Allanore, A. Jagerschmidt, M. Jasson, O. Distler, C. Denton and D. Khanna, OP0266 Lysophophatidic Acid Receptor 1 Antagonist SAR100842 as a Potential Treatment for Patients with Systemic Sclerosis: Results from a Phase 2A Study, *Annals of the Rheumatic Diseases* 74 (2015), 172-173.

[2] K. Archontogeorgis, P. Steiropoulos, A. Tzouvelekis, E. Nena and D. Bouros, Lung Cancer and Interstitial Lung Diseases: A Systematic Review, *Pulmonary Medicine* 2012 (2012), 315918.

[3] T. Barrett, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, et al., NCBI GEO: archive for functional genomics data sets--update, *Nucleic Acids Research* 41 (2013), D991-D995.

[4] M. G. K. Benesch, X. Tang, G. Venkatraman, R. T. Bekele and D. N. Brindley, Recent advances in targeting the autotaxin-lysophosphatidate-lipid phosphate phosphatase axis in vivo, *Journal of Biomedical Research* 30 (2016), 272-284.

[5] K. Bernard, N. J. Logsdon, S. Ravi, N. Xie, B. P. Persons, S. Rangarajan, et al., Metabolic Reprogramming Is Required for Myofibroblast Contractility and Differentiation, *The Journal of Biological Chemistry* 290 (2015), 25427-25438.

[6] K. Bernard, N. J. Logsdon, G. A. Benavides, Y. Sanders, J. Zhang, V. M. Darley-Usmar, et al., Glutaminolysis is required for transforming growth factor-beta1-induced myofibroblast differentiation and activation, *The Journal of Biological Chemistry* 293 (2018), 1218-1228.

[7] S. Bromley-Dulfano, S. Demo, J. Janes, M. Gross, E. Lewis, A. MacKinnon, et al., Antitumor Activity Of The Glutaminase Inhibitor CB-839 In Hematological Malignances, *Blood* 122 (2013), 4226-4226.

[8] F. V. Castelino, G. Bain, V. A. Pace, K. E. Black, L. George, C. K. Probst, et al., An Autotaxin/Lysophosphatidic Acid/Interleukin-6 Amplification Loop Drives Scleroderma Fibrosis, *Arthritis & Rheumatology* 68 (2016), 2964-2974.

[9] J. H. Cho, R. Gelinas, K. Wang, A. Etheridge, M. G. Piper, K. Batte, et al., Systems biology of interstitial lung diseases: integration of mRNA and microRNA expression changes, *BMC Medical Genomics* 4 (2011), 1-8.

[10] G. M. Cooper, *The Cell: A Molecular Approach* Second edition. Sinauer Associates Sunderland (MA) 2000.

[11] D. Croft, A. F. Mundo, R. Haw, M. Milacic, J. Weiser, G. Wu, et al., The Reactome pathway knowledgebase, *Nucleic Acids Research* 42 (2014), D472-D477.

[12] A. T. Dantas, M. C. Pereira, M. J. de Melo Rego, L. F. da Rocha, Jr., R. Pitta Ida, C. D. Marques, et al., The Role of PPAR Gamma in Systemic Sclerosis, *PPAR Research* 2015 (2015), 124624.

[13] R. Edgar, M. Domrachev and A. E. Lash, Gene Expression Omnibus: NCBI gene expression and hybridization array data repository, *Nucleic Acids Research* 30 (2002), 207-210.

[14] A. Fabregat, S. Jupe, L. Matthews, K. Sidiropoulos, M. Gillespie, P. Garapati, et al., The Reactome Pathway Knowledgebase, *Nucleic Acids Research* 46 (2018), D649-D655.

[15] P. J. Fernandez-Marcos and J. Auwerx, Regulation of PGC-1alpha, a nodal regulator of mitochondrial biogenesis, *The American Journal of Clinical Nutrition* 93 (2011), 884-890.

[16] C. Foti, N. Cassano, A. Conserva, C. Coviello, M. De Meo and G. A. Vena, Diffuse cutaneous systemic sclerosis treated with intravenous iloprost, *Clinical and Experimental Dermatology* 29 (2004), 321-323.

[17] R. Foti, E. Visalli, G. Amato, A. Benenati, G. Converso, A. Farina, et al., Long-term clinical stabilization of scleroderma patients treated with a chronic and intensive IV iloprost regimen, *Rheumatology International* 37 (2017), 245-249.

[18] B. Gonzalez-Fernandez, D. I. Sanchez, J. Gonzalez-Gallego and M. J. Tunon, Sphingosine 1-Phosphate Signaling as a Target in Hepatic Fibrosis Therapy, *Frontiers in Pharmacology* 8 (2017), 579.

[19] D. Guo, Q. Wang, C. Li, Y. Wang and X. Chen, VEGF stimulated the angiogenesis by promoting the mitochondrial functions, *Oncotarget* 8 (2017), 77020-77027.

[20] E. L. Herzog, A. Mathur, A. M. Tager, C. Feghali-Bostwick, F. Schneider and J. Varga, Review: interstitial lung disease associated with systemic sclerosis and idiopathic pulmonary fibrosis: how similar and distinct?, *Arthritis & Rheumatology* 66 (2014), 1967-1978.

[21] L. S. Huang and V. Natarajan, Sphingolipids in pulmonary fibrosis, *Advances in biological Regulation* 57 (2015), 55-63.

[22] B. Jiang, Aerobic glycolysis and high level of lactate in cancer metabolism and microenvironment, 4 (2017), 25-27.

[23] D. L. Johannsen and E. Ravussin, The role of mitochondria in health and disease, *Current Opinion in Pharmacology* 9 (2009), 780-786.

[24] F. R. Jornayvaz and G. I. Shulman, Regulation of mitochondrial biogenesis, *Essays in Biochemistry* 47 (2010), 69-84.

[25] M. Kanehisa and S. Goto, KEGG: kyoto encyclopedia of genes and genomes, *Nucleic Acids Research* 28 (2000), 27-30.

[26] M. Kanehisa, M. Furumichi, M. Tanabe, Y. Sato and K. Morishima, KEGG: new perspectives on genomes, pathways, diseases and drugs, *Nucleic Acids Research* 45 (2017), D353-D361.

[27] M. Kanehisa, Y. Sato, M. Kawashima, M. Furumichi and M. Tanabe, KEGG as a reference resource for gene and protein annotation, *Nucleic Acids Research* 44 (2016), D457-D462.

[28] T. Karampitsakos, V. Tzilas, R. Tringidou, P. Steiropoulos, V. Aidinis, S. A. Papiris, et al., Lung cancer in patients with idiopathic pulmonary fibrosis, *Pulmonary Pharmacology & Therapeutics* 45 (2017), 1-10.

[29] S. Krug, A. Sablotzki, S. Hammerschmidt, H. Wirtz and H. J. Seyfarth, Inhaled iloprost for the control of pulmonary hypertension, *Vascular Health and Risk Management* 5 (2009), 465-474.

[30] M. Kutmon, M. P. van Iersel, A. Bohler, T. Kelder, N. Nunes, A. R. Pico, et al., PathVisio 3: an extendable pathway analysis toolbox, *PLoS Computational Biology* 11 (2015), e1004085.

[31] H. F. Lakatos, T. H. Thatcher, R. M. Kottmann, T. M. Garcia, R. P. Phipps and P. J. Sime, The Role of PPARs in Lung Fibrosis, *PPAR Research* 2007 (2007), 71323.

[32] B. Lasota, S. Skoczynski, K. Mizia-Stec and W. Pierzchala, The use of iloprost in the treatment of 'out of proportion' pulmonary hypertension in chronic obstructive pulmonary disease, *International Journal of Clinical Pharmacy* 35 (2013), 313-315.

[33] G. Legube and D. Trouche, Regulating histone acetyltransferases and deacetylases, *EMBO Reports* 4 (2003), 944-947.

[34] E. C. LeRoy, C. Black, R. Fleischmajer, S. Jablonska, T. Krieg, T. A. Medsger, Jr., et al., Scleroderma (systemic sclerosis): classification, subsets and pathogenesis, *The Journal of Rheumatology* 15 (1988), 202-205.

[35] J. J. Li, R. Li, W. Wang, B. Zhang, X. Song, C. Zhang, et al., IDH2 is a novel diagnostic and prognostic serum biomarker for non-small-cell lung cancer, *Molecular Oncology* 12 (2018), 602-610.

[36] Z. Liang, T. Li, S. Jiang, J. Xu, W. Di, Z. Yang, et al., AMPK: a novel target for treating hepatic fibrosis, *Oncotarget* 8 (2017), 62780-62792.

[37] G. E. Lindahl, C. J. Stock, X. Shi-Wen, P. Leoni, P. Sestini, S. L. Howat, et al., Microarray profiling reveals suppressed interferon stimulated gene program in fibroblasts from scleroderma-associated interstitial lung disease, *Respiratory Research* 14 (2013), 80.

[38] D. Litonin, M. Sologub, Y. Shi, M. Savkina, M. Anikin, M. Falkenberg, et al., Human mitochondrial transcription revisited: only TFAM and TFB2M are required for transcription of the mitochondrial genes in vitro, *The Journal of Biological Chemistry* 285 (2010), 18129-18133.

[39] T. R. Luckhardt and V. J. Thannickal, Systemic sclerosis-associated fibrosis: an accelerated aging phenotype?, *Current Opinion in Rheumatology* 27 (2015), 571-576.

[40] D. J. McCarthy and G. K. Smyth, Testing significance relative to a fold-change threshold is a TREAT, *Bioinformatics* 25 (2009), 765-771.

[41] E. B. Meltzer, W. T. Barry, T. A. D'Amico, R. D. Davis, S. S. Lin, M. W. Onaitis, et al., Bayesian probit regression model for the diagnosis of pulmonary fibrosis: proof-of-principle, *BMC Medical Genomics* 4 (2011), 70.

[42] A. Meshcheryakova, M. Svoboda, A. Tahir, H. C. Kofeler, A. Triebl, F. Mungenast, et al., Exploring the role of sphingolipid machinery during the epithelial to mesenchymal transition program using an integrative approach, *Oncotarget* 7 (2016), 22295-22323.

[43] R. N. Michel, E. R. Chin, J. V. Chakkalakal, J. K. Eibl and B. J. Jasmin, Ca2+/calmodulin-based signalling in the regulation of the muscle fibre phenotype and its therapeutic potential via modulation of utrophin A and myostatin expression, *Applied Physiology, Nutrition, and Metabolism* 32 (2007), 921-929.

[44] M. W. Moore and E. L. Herzog, Regulation and Relevance of Myofibroblast Responses in Idiopathic Pulmonary Fibrosis, *Current Pathobiology Reports* 1 (2013), 199-208.

[45] A. L. Mora, M. Bueno and M. Rojas, Mitochondria in the spotlight of aging and idiopathic pulmonary fibrosis, *The Journal of Clinical Investigation* 127 (2017), 405-414.

[46] A. Moreno-Moral, M. Bagnati, S. Koturan, J. H. Ko, C. Fonseca, N. Harmston, et al., Changes in macrophage transcriptome associate with systemic sclerosis and mediate GSDMA contribution to disease risk, *Annals of the Rheumatic Diseases* 77 (2018), 596-601.

[47] Y. Mostmans, M. Cutolo, C. Giddelo, S. Decuman, K. Melsens, H. Declercq, et al., The role of endothelial cells in the vasculopathy of systemic sclerosis: A systematic review, *Autoimmunity Reviews* 16 (2017), 774-786.

[48] D. Nishimura, BioCarta, *Biotech Software & Internet Report* 2 (2001), 117-120.

[49] A. Pardo, K. Gibson, J. Cisneros, T. J. Richards, Y. Yang, C. Becerril, et al., Up-regulation and profibrotic role of osteopontin in human idiopathic pulmonary fibrosis, *PLoS Medicine* 2 (2005), e251.

[50] D. Pattanaik and A. E. Postlethwaite, A role for lysophosphatidic acid and sphingosine 1-phosphate in the pathogenesis of systemic sclerosis, *Discovery Medicine* 10 (2010), 161-167.

[51] T. A. Patterson, E. K. Lobenhofer, S. B. Fulmer-Smentek, P. J. Collins, T. M. Chu, W. Bao, et al., Performance comparison of one-color and two-color platforms within the MicroArray Quality Control (MAQC) project, *Nature Biotechnology* 24 (2006), 1140-1150.

[52] R. Peng, S. Sridhar, G. Tyagi, J. E. Phillips, R. Garrido, P. Harris, et al., Bleomycin induces molecular changes directly relevant to idiopathic pulmonary fibrosis: a model for "active" disease, *PloS One* 8 (2013), e59348.

[53] A. Pisano, B. Cerbelli, E. Perli, M. Pelullo, V. Bargelli, C. Preziuso, et al., Impaired mitochondrial biogenesis is a common feature to myocardial hypertrophy and end-stage ischemic heart failure, *Cardiovascular Pathology* 25 (2016), 103-112.

[54] N. J. Pyne, G. Dubois and S. Pyne, Role of sphingosine 1-phosphate and lysophosphatidic acid in fibrosis, *Biochimica et Biophysica Acta* 1831 (2013), 228-238.

[55] E. A. Renzoni, D. J. Abraham, S. Howat, X. Shi-Wen, P. Sestini, G. Bou-Gharios, et al., Gene expression profiling reveals novel TGFbeta targets in adult lung fibroblasts, *Respiratory Research* 5 (2004), 24.

[56] E. Ricciotti and G. A. FitzGerald, Prostaglandins and inflammation, *Arteriosclerosis, Thrombosis, and Vascular Biology* 31 (2011), 986-1000.

[57] G. Rosen, L. Sivaraman, P. Cheng, B. Murphy, K. Chadwick, L. Lehman-McKeeman, et al., LPA1 antagonists BMS-986020 and BMS-986234 for idiopathic pulmonary fibrosis: Preclinical evaluation of hepatobiliary homeostasis, *European Respiratory Journal* 50 (2017), 1038.

[58] J. H. Ryu, T. Moua, C. E. Daniels, T. E. Hartman, E. S. Yi, J. P. Utz, et al., Idiopathic pulmonary fibrosis: evolving concepts, *Mayo Clinic Proceedings* 89 (2014), 1130-1142.

[59] E. Saunier, S. Antonio, A. Regazzetti, N. Auzeil, O. Laprevote, J. W. Shay, et al., Resveratrol reverses the Warburg effect by targeting the pyruvate dehydrogenase complex in colon cancer cells, *Scientific Reports* 7 (2017), 6945.

[60] K. C. Silver and R. M. Silver, Management of Systemic-Sclerosis-Associated Interstitial Lung Disease, *Rheumatic Diseases Clinics of North America* 41 (2015), 439-457.

[61] R. Simon, A. Lam, M. C. Li, M. Ngan, S. Menenzes and Y. Zhao, Analysis of gene expression data using BRB-ArrayTools, *Cancer Informatics* 3 (2007), 11-17.

[62] R. Simon, *BRB-ArrayTools Version 4.2 User's Manual*, Biometrics Research Branch National Cancer Institute, The EMMES Corporation, 2010.

[63] S. P. Singh, J. Schragenheim, J. Cao, J. R. Falck, N. G. Abraham and L. Bellner, PGC-1 alpha regulates HO-1 expression, mitochondrial dynamics and biogenesis: Role of epoxyeicosatrienoic acid, *Prostaglandins & Other Lipid Mediators* 125 (2016), 8-18.

[64] E. M. Stein, Enasidenib, a targeted inhibitor of mutant IDH2 proteins for treatment of relapsed or refractory acute myeloid leukemia, *Future Oncology* 14 (2018), 23-40.

[65] G. Stelzer, N. Rosen, I. Plaschkes, S. Zimmerman, M. Twik, S. Fishilevich, et al., The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses, *Current Protocols in Bioinformatics* 54 (2016), 1.30.31-31.30.33.

[66] A. Sturn, J. Quackenbush and Z. Trajanoski, Genesis: cluster analysis of microarray data, *Bioinformatics* 18 (2002), 207-208.

[67] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, et al., Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles, *Proceedings of the National Academy of Sciences of the United States of America* 102 (2005), 15545-15550.

[68] D. Szklarczyk, J. H. Morris, H. Cook, M. Kuhn, S. Wyder, M. Simonovic, et al., The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible, *Nucleic acids research* 45 (2017), D362-D368.

[69] The Gene Ontology Consortium, M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, et al., Gene Ontology: tool for the unification of biology, *Nature Genetics* 25 (2000), 25-29.

[70] The Gene Ontology Consortium, Expansion of the Gene Ontology knowledgebase and resources, *Nucleic Acids Research* 45 (2017), D331-D338.

[71] T. Valero, Mitochondrial biogenesis: pharmacological approaches, *Current Pharmaceutical Design* 20 (2014), 5507-5509.

[72] M. P. van Iersel, T. Kelder, A. R. Pico, K. Hanspers, S. Coort, B. R. Conklin, et al., Presenting and exploring biological pathways with PathVisio, *BMC Bioinformatics* 9 (2008), 399.

[73] R. Ventura-Clapier, A. Garnier and V. Veksler, Transcriptional control of mitochondrial biogenesis: the central role of PGC-1alpha, *Cardiovascular Research* 79 (2008), 208-217.

[74] R. A. Weisiger and I. Fridovich, Superoxide dismutase. Organelle specificity, *The Journal of Biological Chemistry* 248 (1973), 3582-3592.

[75] J. Wysocka, M. P. Myers, C. D. Laherty, R. N. Eisenman and W. Herr, Human Sin3 deacetylase and trithorax-related Set1/Ash2 histone H3-K4 methyltransferase are tethered together selectively by the cell-proliferation factor HCF-1, *Genes & Development* 17 (2003), 896-911.

[76] N. Xie, Z. Tan, S. Banerjee, H. Cui, J. Ge, R. M. Liu, et al., Glycolytic Reprogramming in Myofibroblast Differentiation and Lung Fibrosis, *American Journal of Respiratory and Critical Care Medicine* 192 (2015), 1462-1474.

[77] R. Xu, Q. Hu and G. Wang, Mitochondrial biogenesis involved in neurodegeneration and aging *Gene and Gene Editing* 1 (2015), 103-110.

[78] Q. Yu and S. Y. Chan, Mitochondrial and Metabolic Drivers of Pulmonary Vascular Endothelial Dysfunction in Pulmonary Hypertension, *Advances in Experimental Medicine and Biology* 967 (2017), 373-383.

[79] D. C. Zank, M. Bueno, A. L. Mora and M. Rojas, Idiopathic Pulmonary Fibrosis: Aging, Mitochondrial Dysfunction, and Cellular Bioenergetics, *Frontiers in Medicine* 5 (2018), 10.

[80] Z. Zeng, S. Cheng, H. Chen, Q. Li, Y. Hu, Q. Wang, et al., Activation and overexpression of Sirt1 attenuates lung fibrosis via P300, *Biochemical and Biophysical Research Communications* 486 (2017), 1021-1026.

[81] Y. D. Zhao, L. Yin, S. Archer, C. Lu, G. Zhao, Y. Yao, et al., Metabolic heterogeneity of idiopathic pulmonary fibrosis: a metabolomic study, *BMJ Open Respiratory Research* 4 (2017), e000183.

APPENDIX A – Mitochondrial biogenesis gene list

There are 63 genes involved in mitochondrial biogenesis Reactome pathway. Only 57 genes with 122 corresponding probe sets are included in annotation file (GPL96) for dataset GSE40839 and all 63 genes with 198 corresponding probe sets are included in annotation file (GPL570) for dataset GSE44723.

***Table A1**: Important genes associated with Mitochondrial biogenesis from Reactome database*

Symbol	**Name**	**Genes included in GPL96**	**Number of probe sets in GPL96 for each gene**	**Genes include d in GPL570**	**Number of probe sets in GPL570 for each gene**
GABPB1	GA-binding protein transcription factor beta subunit 1	YES	2	YES	2
NRF1	Nuclear respiratory factor 1	YES	4	YES	5
PRKAB1	Protein kinase, AMP-activated, noncatalytic, beta-1	YES	2	YES	2
PRKAG2	Protein kinase, AMP-activated, noncatalytic, gamma-2	YES	2	YES	5
PRKAB2	Protein kinase, AMP-activated, noncatalytic, beta-2	YES	1	YES	3
PRKAG1	Protein kinase, AMP-activated, noncatalytic, gamma-1	YES	1	YES	1
PRKAA2	Protein kinase, AMP-activated, catalytic, alpha-2	YES	1	YES	5
PRKAG3	Protein kinase, AMP-activated, noncatalytic, gamma-3	NO	0	YES	1
CYCS	Cytochrome C, somatic	YES	1	YES	3
PPRC1	Peroxisome proliferator-activated receptor-gamma, coactivator-related protein 1	YES	1	YES	1
HCFC1	Host cell factor C1	YES	2	YES	3
GABPA	GA-binding protein transcription factor, alpha subunit	YES	1	YES	2
PPARGC1A	Peroxisome proliferator-activated receptor-gamma, coactivator 1, alpha	YES	1	YES	2
CREB1	CAMP response element-binding protein 1	YES	4	YES	7
SIRT4	Sirtuin 4	YES	2	YES	2
TFB1M	Transcription factor B1, mitochondrial	YES	1	YES	4

Symbol	Name	Genes included in GPL96	Number of probe sets in GPL96 for each gene	Genes include d in GPL570	Number of probe sets in GPL570 for each gene
PERM1	PPARGC1- and ESRR-induced regulator, muscle, 1	NO	0	YES	1
HDAC3	Histone deacetylase 3	YES	1	YES	1
NCOR1	Nuclear receptor corepressor 1	YES	4	YES	5
NR1D1/THR A	Nuclear receptor subfamily 1, group D, member 1	YES	3	YES	3
POLG2	Polymerase, DNA, gamma-2	YES	1	YES	1
GLUD2	Glutamate dehydrogenase 2	YES	2	YES	2
GLUD1	Glutamate dehydrogenase 1	YES	2	YES	2
SIRT5	Sirtuin 5	YES	2	YES	4
PPARGC1B	Peroxisome proliferator-activated receptor-gamma, coactivator 1, beta	NO	0	YES	4
SIRT3	Sirtuin 3	YES	3	YES	3
MAPK12	Mitogen-activated protein kinase 12	YES	1	YES	3
MAPK11	Mitogen-activated protein kinase 11	YES	3	YES	3
MAPK14	Mitogen-activated protein kinase 14	YES	4	YES	4
CRTC3	CREB-regulated transcription coactivator 3	YES	1	YES	3
CRTC1	CREB-regulated transcription coactivator 1	YES	2	YES	2
CRTC2	CREB-regulated transcription coactivator 2	NO	0	YES	1
ATP5B	ATP synthase, H+ transporting, mitochondrial F1 complex, beta subunit	YES	1	YES	1
PEO1/C10orf 2	Progressive external ophthalmoplegia with mitochondrial DNA deletions, autosomal dominant 3	YES	1	YES	1
IDH2	Isocitrate dehydrogenase 2	YES	2	YES	2
ACSS2	Acetyl-CoA synthetase short chain family, member 2	NO	0	YES	1
SOD2	Superoxide dismutase 2	YES	4	YES	5
POLRMT	Polymerase, RNA, mitochondrial	YES	2	YES	3
TFAM	Transcription factor A, mitochondrial	YES	3	YES	4
ESRRA	Estrogen-related receptor, alpha	YES	2	YES	2

Symbol	Name	Genes included in GPL96	Number of probe sets in GPL96 for each gene	Genes included in GPL570	Number of probe sets in GPL570 for each gene
MEF2D	Myocyte enhancer factor 2, polypeptide D	YES	2	YES	3
MEF2C	Myocyte enhancer factor 2, polypeptide C	YES	3	YES	3
SSBP1	Single-stranded DNA-binding protein 1	YES	2	YES	3
TFB2M	Transcription factor B2, mitochondrial	YES	1	YES	4
ATF2	Activating transcription factor 2	YES	2	YES	3
MED1	Mediator complex subunit 1	YES	2	YES	4
PPARA	Peroxisome proliferator-activated receptor-alpha	YES	2	YES	8
CHD9	Chromodomain helicase DNA-binding protein 9	YES	3	YES	7
TBL1X	Transducin-beta-like 1, x-linked	YES	5	YES	6
TGS1	Trimethylguanosine synthase 1	YES	1	YES	3
HELZ2	Peroxisomal proliferator-activated receptor alpha-interacting cofactor complex, 285-kd subunit	NO	0	YES	4
RXRA	Retinoid x receptor, alpha	YES	2	YES	2
CREBBP	CREB-binding protein	YES	2	YES	4
SMARCD3	SWI/SNF-related, matrix-associated, actin-dependent regulator of chromatin, subfamily D, member 3	YES	1	YES	2
NCOA6	Nuclear receptor coactivator 6	YES	1	YES	2
CARM1	Coactivator-associated arginine methyltransferase 1	YES	1	YES	1
NCOA1	Nuclear receptor coactivator 1	YES	4	YES	4
TBL1XR1	Transducin-beta-like 1 receptor 1	YES	1	YES	6
NCOA2	Nuclear receptor coactivator 2	YES	4	YES	4
CAMK4	Calcium/calmodulin-dependent protein kinase IV	YES	1	YES	3
CALM1	Calmodulin 1	YES	10	YES	11
MTERF1	Transcription termination factor 1, mitochondrial	YES	1	YES	1
ALAS1	Delta-aminolevulinate synthase 1	YES	1	YES	1

Table A2*: Custom made Mitochondrial biogenesis gene list which was added to already existing KEGG gene lists in BRB-ArrayTools database*

<table>
<tr><th>UGCluster</th><th>Symbol</th><th>Accession</th></tr>
<tr><td>Hs.181202</td><td>GABPB1</td><td>NM_002041|NM_005254|NM_016654|NM_016655|NM_181427|XM_005254273|XM_005254274|XM_006720455|XM_006720456|XM_006720457|XM_006720458</td></tr>
<tr><td>Hs.202007</td><td>NRF1</td><td>NM_001040110|NM_001293163|NM_001293164|NM_005011</td></tr>
<tr><td>Hs.6061</td><td>PRKAB1</td><td>NM_006253|XM_005253909</td></tr>
<tr><td>Hs.259842</td><td>PRKAG2</td><td>NM_001040633|NM_016203|NM_024429|XM_005250002|XM_005250003|XM_005250004|XM_005250005|XM_005250006|XM_005250007|XM_005250009|XM_006716021</td></tr>
<tr><td>Hs.50732</td><td>PRKAB2</td><td>NM_005399|NR_103870|NR_103871</td></tr>
<tr><td>Hs.3136</td><td>PRKAG1</td><td>NM_001206709|NM_001206710|NM_002733|NM_212461|XM_005269019|XM_005269020|XM_006719499|XM_006719500</td></tr>
<tr><td>Hs.256067</td><td>PRKAA2</td><td>NM_006252</td></tr>
<tr><td>Hs.434525</td><td>PRKAG3</td><td>NM_017431</td></tr>
<tr><td>Hs.437060</td><td>CYCS</td><td>NM_018947</td></tr>
<tr><td>Hs.146957</td><td>PPRC1</td><td>NM_001288727|NM_001288728|NM_015062|XM_005269656|XM_005269658|XM_006717730|XM_006717731</td></tr>
<tr><td>Hs.83634</td><td>HCFC1</td><td>NM_005334|XM_005274664|XM_006724815|XM_006724816</td></tr>
<tr><td>Hs.78</td><td>GABPA</td><td>NM_001197297|NM_002040|XM_005260938|XM_005260939</td></tr>
<tr><td>Hs.198468</td><td>PPARGC1A</td><td>NM_013261|XM_005248130|XM_005248131|XM_005248132|XM_005248134</td></tr>
<tr><td>Hs.22315</td><td>CREB1</td><td>NM_004379|NM_134442|XR_241289|XR_241290|XR_241292|XR_427071</td></tr>
<tr><td>Hs.50861</td><td>SIRT4</td><td>NM_012240|XM_005253865|XM_006719308|XM_006719309|XM_006719310|XM_006719311|XM_006719312</td></tr>
<tr><td>Hs.279908</td><td>TFB1M</td><td>NM_016020|XM_005267005|XM_005267006</td></tr>
<tr><td>Hs.271462</td><td>PERM1</td><td>NM_001291366|NM_001291367|NM_032722|NR_027693</td></tr>
<tr><td>Hs.388681</td><td>HDAC3</td><td>NM_003883|XM_006714802</td></tr>
<tr><td>Hs.144904</td><td>NCOR1</td><td>NM_001190438|NM_001190440|NM_006311|XM_005256866|XM_005256867|XM_005256868|XM_005256871|XM_005256872|XM_005256873|XM_005256874|XM_005256875|XM_006721601|XM_006721602|XM_006721603|XM_006721604|XM_006721605</td></tr>
<tr><td>Hs.276916</td><td>NR1D1</td><td>NM_001190918|NM_001190919|NM_003250|NM_021724|NM_199334</td></tr>
<tr><td>Hs.437009</td><td>POLG2</td><td>NM_007215|XM_006721651|XR_243630</td></tr>
<tr><td>Hs.525862</td><td>GLUD2</td><td>NM_012084</td></tr>
<tr><td>Hs.355697</td><td>GLUD1</td><td>NM_005271</td></tr>
<tr><td>Hs.282331</td><td>SIRT5</td><td>NM_001193267|NM_001242827|NM_012241|NM_031244|XM_005248967|XM_005248968|XM_005248969</td></tr>
<tr><td>Hs.248652</td><td>PPARGC1B</td><td>NM_001172698|NM_001172699|NM_133263|XM_005268372</td></tr>
<tr><td>Hs.511950</td><td>SIRT3</td><td>NM_001017524|NM_012239|XM_005252835</td></tr>
<tr><td>Hs.432642</td><td>MAPK12</td><td>NM_002969|XM_003846644|XM_005275911</td></tr>
<tr><td>Hs.57732</td><td>MAPK11</td><td>NM_002751|NR_110887</td></tr>
<tr><td>Hs.79107</td><td>MAPK14</td><td>NM_001315|NM_139012|NM_139013|NM_139014|XM_006714998</td></tr>
<tr><td>Hs.567572</td><td>CRTC3</td><td>NM_001042574|NM_022769|XM_005254968</td></tr>
<tr><td>Hs.6051</td><td>CRTC1</td><td>NM_001098482|NM_015321|NM_025021|XM_005259833|XM_005259834|XM_005259835|XM_005259836|XM_006722710</td></tr>
<tr><td>Hs.406392</td><td>CRTC2</td><td>NM_181715|XM_005244946|XM_005244947|XM_005244949|XM_006711199|XM_006711200|XM_006711201|XM_006711202</td></tr>
<tr><td>Hs.406510</td><td>ATP5B</td><td>NM_001686</td></tr>
<tr><td>Hs.22678</td><td>C10orf2</td><td>NM_001163812|NM_001163813|NM_001163814|NM_021830|XM_006717921|XM_006717922|XR_246100</td></tr>
<tr><td>Hs.5337</td><td>IDH2</td><td>NM_001289910|NM_001290114|NM_002168</td></tr>
<tr><td>Hs.14779</td><td>ACSS2</td><td>NM_001076552|NM_001242393|NM_018677|NM_139274|XM_005260455|XM_005260456|XM_006723825|XM_006723826</td></tr>
<tr><td>Hs.384944</td><td>SOD2</td><td>NM_000636|NM_001024465|NM_001024466</td></tr>
<tr><td>Hs.254113</td><td>POLRMT</td><td>NM_005035|XM_005259580</td></tr>
<tr><td>Hs.75133</td><td>TFAM</td><td>NM_001270782|NM_003201|NM_012251|NR_073073</td></tr>
</table>

UGCluster	Symbol	Accession
Hs.110849	ESRRA	NM_001282450|NM_001282451|NM_004451|XM_006718449|XM_006718450
Hs.77955	MEF2D	NM_001271629|NM_005920|XM_005245169|XM_005245170|XM_006711330|XM_006711331|XM_006711332|XM_006711333|XM_006711334
Hs.368950	MEF2C	NM_001131005|NM_001193347|NM_001193348|NM_001193349|NM_001193350|NM_002397|XM_005248511|XM_006714618|XM_006714619|XM_006714620|XM_006714621|XM_006714622|XM_006714623|XM_006714624|XM_006714625
Hs.923	SSBP1	NM_001256510|NM_001256511|NM_001256512|NM_001256513|NM_003143|NR_046269|XM_005250048|XM_005250049|XM_005250050|XM_005250051
Hs.7395	TFB2M	NM_022366
Hs.80285	ATF2	NM_001256090|NM_001256091|NM_001256092|NM_001256093|NM_001256094|NM_001880|NR_045768|NR_045769|NR_045770|NR_045771|NR_045772|NR_045773|NR_045774
Hs.15589	MED1	NM_004774|XM_005257465|XM_006721957
Hs.271640	PPARA	NM_001001928|NM_001001929|NM_001001930|NM_005036|NM_032644|XM_005261653|XM_005261655|XM_005261656|XM_005261657|XM_006724269|XM_006724270|XM_006724271
Hs.59159	CHD9	NM_025134|XM_005256168|XM_005256169|XM_005256170|XM_005256171|XM_005256172|XM_005256174|XM_005256175|XM_005256176|XM_006721280|XM_006721281|XM_006721282|XM_006721283|XR_429731
Hs.76536	TBL1X	NM_001139466|NM_001139467|NM_001139468|NM_005647
Hs.179909	TGS1	NM_024831|XM_005251328|XM_006716485|XM_006716486
Hs.151714	HELZ2	NM_001037335|NM_033405
Hs.20084	RXRA	NM_001291920|NM_001291921|NM_002957|XM_005263409|XM_006717232
Hs.270804	CREBBP	NM_001079846|NM_004380|XM_005255124|XM_005255125|XM_006720848
Hs.444445	SMARCD3	NM_001003801|NM_001003802|NM_003078
Hs.435788	NCOA6	NM_001242539|NM_014071|XM_005260348|XM_006723750|XM_006723751|XM_006723752|XM_006723753|XM_006723754|XM_006723755
Hs.371416	CARM1	NM_199141|XM_005259708
Hs.386092	NCOA1	NM_003743|NM_147223|NM_147233|XM_005264625|XM_005264626|XM_005264627|XM_005264628|XM_006712126
Hs.438970	TBL1XR1	NM_024665|XM_005247771|XM_005247772|XM_005247775|XM_005247776|XM_006713745|XM_006713746
Hs.446678	NCOA2	NM_006540|XM_005251128|XM_005251129|XM_005251130|XM_005251131|XM_005251132|XM_005251133
Hs.440638	CAMK4	NM_001744
Hs.282410	CALM1	NM_001166106|NM_006888|XM_006720258
Hs.97996	MTERF	NM_006980|XM_005250593|XM_005250594|XM_005250595|XM_006716126
Hs.511918	ALAS1	NM_000688|NM_199166|XM_005264944|XM_005264945

APPENDIX B – GSEA of SSc

***Table A3**: Enriched pathways by GSEA (α=0.001) of all genes in SSc-ILD compared to controls, sorted by LS permutation p-value*

Pathway description	Number of gene sets	LS permutation p-value
Cytokine-cytokine receptor interaction	89	0.00001
Chemokine signalling pathway	58	0.00001
Phagosome	73	0.00001
Cell adhesion molecules (CAMs)	57	0.00001
Antigen processing and presentation	43	0.00001
Toll-like receptor signalling pathway	36	0.00001
NOD-like receptor signalling pathway	24	0.00001
Cytosolic DNA-sensing pathway	18	0.00001
Natural killer cell mediated cytotoxicity	48	0.00001
Type I diabetes mellitus	27	0.00001
Hepatitis C	48	0.00001
Autoimmune thyroid disease	24	0.00001
Allograft rejection	24	0.00001
Graft-versus-host disease	26	0.00001
Viral myocarditis	47	0.00001
Leishmaniasis	26	0.00001
Osteoclast differentiation	52	0.00005
RIG-I-like receptor signalling pathway	18	0.00021
Amyotrophic lateral sclerosis (ALS)	18	0.00051
JAK-STAT signalling pathway	51	0.00057
Pancreatic cancer	43	0.00060
Glycolysis/Gluconeogenesis	29	0.00071
Toxoplasmosis	47	0.00073
Endocytosis	75	0.00093
Staphylococcus aureus infection	12	0.00190
Proteasome	14	0.00204
Pathogenic Escherichia coli infection	40	0.00332
Gap junction	51	0.00340
African trypanosomiasis	23	0.00624
Fat digestion and absorption	12	0.00893
Ether lipid metabolism	11	0.00898
Fructose and mannose metabolism	13	0.01912
Chagas disease (American trypanosomiasis)	52	0.02238
Pathways in cancer	155	0.07025
Steroid hormone biosynthesis	14	0.09451
Lysosome	49	0.14518
p53 signalling pathway	59	0.15075
DNA replication	26	0.53578

APPENDIX C – Genes of enriched pathways in SSc

Tables of genes involved in important significantly enriched pathways by GSEA (α=0.001) – datset GSE40839.

Table A4*: Upregulated genes (α=0.001) in Osteoclast differentiation pathway in SSc-ILD compared to controls, sorted by parametric p-value – all genes*

Probe set	Gene symbol	Parametric p-value	logFC
201473_at	JUNB	0.0005496	-0.74
203085_s_at	TGFB1	0.0006889	-1.12
212607_at	AKT3	0.0045982	-0.62
209949_at	NCF2	0.0333401	-0.32
206943_at	TGFBR1	0.0572675	-0.43
204628_s_at	ITGB3	0.0826239	-0.34
203879_at	PIK3CD	0.1209761	-0.43
202429_s_at	PPP3CA	0.1481402	-0.30
204627_s_at	ITGB3	0.3597784	-0.32
211537_x_at	MAP3K7	0.4155386	-0.17
202949_s_at	FHL2	0.4456635	-0.29
220407_s_at	TGFB2	0.6193209	-0.12
204313_s_at	CREB1	0.6726583	-0.10

Table A5*: Downregulated genes (α=0.001) in Osteoclast differentiation pathway in SSc-ILD compared to controls, sorted by parametric p-value – all genes*

Probe set	Gene symbol	Parametric p-value	logFC
AFFX-HUMISGF3A/M97935_3_at	STAT1	< 1e-07	2.06
209969_s_at	STAT1	< 1e-07	2.14
AFFX-HUMISGF3A/M97935_MB_at	STAT1	< 1e-07	2.15
AFFX-HUMISGF3A/M97935_MA_at	STAT1	< 1e-07	2.07
200887_s_at	STAT1	< 1e-07	2.13
AFFX-HUMISGF3A/M97935_5_at	STAT1	2.00E-07	1.90
203882_at	IRF9	3.00E-07	1.64
208944_at	TGFBR2	1.70E-06	0.93
201502_s_at	NFKBIA	8.70E-06	1.83
211676_s_at	IFNGR1	1.05E-05	1.16
210001_s_at	SOCS1	1.16E-05	1.06
209716_at	CSF1	4.71E-05	0.96
202948_at	IL1R1	5.38E-05	1.63
215561_s_at	IL1R1	0.0001191	0.78
204932_at	TNFRSF11B	0.0001532	2.09
209239_at	NFKB1	0.0001625	0.60
201471_s_at	SQSTM1	0.0003801	1.15

Probe set	Gene symbol	Parametric p-value	logFC
210105_s_at	FYN	0.0004628	1.14
201466_s_at	JUN	0.0004828	0.63
204933_s_at	TNFRSF11B	0.0005089	2.23
201465_s_at	JUN	0.0007686	0.57
202743_at	PIK3R3	0.0008968	0.63
201464_x_at	JUN	0.0020500	0.69
202450_s_at	CTSK	0.0024082	1.93
205170_at	STAT2	0.0024738	0.82
207233_s_at	MITF	0.0033579	1.04
213112_s_at	SQSTM1	0.0036446	0.97
207334_s_at	TGFBR2	0.0043350	0.46
216033_s_at	FYN	0.0045883	1.00
209341_s_at	IKBKB	0.0272084	0.53
203752_s_at	JUND	0.0724440	0.42
208510_s_at	PPARG	0.0746871	0.57
203028_s_at	CYBA	0.0774965	0.29
205067_at	IL1B	0.1768759	0.34
221903_s_at	CYLD	0.1811908	0.36
212240_s_at	PIK3R1	0.3174871	0.21
204369_at	PIK3CA	0.4180769	0.20
201648_at	JAK1	0.6447218	0.10
204420_at	FOSL1	0.7430879	0.12

Table A6*: Upregulate genes (α=0.001) in Glycolysis/Gluconeogenesis pathway in SSc-ILD compared to controls, sorted by parametric p-value –all genes*

Probe set	Gene symbol	Parametric p-value	logFC
213011_s_at	TPI1	5.00E-07	-1.03
217294_s_at	ENO1	7.40E-06	-1.03
201037_at	PFKP	9.70E-06	-1.89
201231_s_at	ENO1	1.20E-05	-0.84
200737_at	PGK1	1.32E-05	-0.97
208308_s_at	GPI	2.19E-05	-0.89
200822_x_at	TPI1	2.32E-05	-0.86
217356_s_at	PGK1	3.70E-05	-1.22
200738_s_at	PGK1	5.41E-05	-0.92
200650_s_at	LDHA	9.79E-05	-0.81
209645_s_at	ALDH1B1	0.0008129	-1.29
201313_at	ENO2	0.0015445	-0.79
209646_x_at	ALDH1B1	0.0024024	-0.58
211023_at	PDHB	0.0067421	-0.43
203502_at	BPGM	0.1526975	-0.54
202847_at	PCK2	0.1704186	-0.54
200697_at	HK1	0.2103270	-0.25

Probe set	Gene symbol	Parametric p-value	logFC
202934_at	HK2	0.8117521	-0.07

***Table** A7: Downregulated genes (α=0.01) in Glycolysis/Gluconeogenesis pathway in SSc-ILD compared to controls, sorted by parametric p-value – all genes*

Probe set	Gene symbol	Parametric p-value	logFC
208848_at	ADH5	3.00E-06	1.03
203180_at	ALDH1A3	9.30E-06	1.82
209612_s_at	ADH1B	0.0003134	3.69
208847_s_at	ADH5	0.0005913	0.66
209613_s_at	ADH1B	0.0023055	3.08
209614_at	ADH1B	0.0027065	1.83
202054_s_at	ALDH3A2	0.0210571	0.88
202053_s_at	ALDH3A2	0.0251304	0.51
210544_s_at	ALDH3A2	0.0711939	0.50
201425_at	ALDH2	0.1736907	0.64
201251_at	PKM	0.9665203	0.01

***Table A8**: Upregulated genes (α=0.001) in Glycolysis/Gluconeogenesis pathway in SSc-ILD compared to controls, sorted by parametric p-value –Metabolic pathways genes*

Probe set	Gene symbol	Parametric p-value	logFC
213011_s_at	TPI1	5.00E-07	-1.03
217294_s_at	ENO1	7.40E-06	-1.03
201037_at	PFKP	9.70E-06	-1.89
201231_s_at	ENO1	1.20E-05	-0.84
200737_at	PGK1	1.32E-05	-0.97
208308_s_at	GPI	2.19E-05	-0.89
200822_x_at	TPI1	2.32E-05	-0.86
217356_s_at	PGK1	3.70E-05	-1.22
200738_s_at	PGK1	5.41E-05	-0.92
200650_s_at	LDHA	9.79E-05	-0.81
209645_s_at	ALDH1B1	0.0008129	-1.29
201313_at	ENO2	0.0015445	-0.79
209646_x_at	ALDH1B1	0.0024024	-0.58
211023_at	PDHB	0.0067421	-0.43
203502_at	BPGM	0.1526975	-0.54
202847_at	PCK2	0.1704186	-0.54
200697_at	HK1	0.210327	-0.25
202934_at	HK2	0.8117521	-0.07

***Table A9**: Downregulated genes (α=0.001) in Glycolysis/Gluconeogenesis pathway in SSc-ILD compared to controls, sorted by parametric p-value – Metabolic pathways genes*

Probe set	Gene symbol	Parametric p-value	logFC
208848_at	ADH5	3.00E-06	1.03
203180_at	ALDH1A3	9.30E-06	1.82
209612_s_at	ADH1B	0.000313	3.69
208847_s_at	ADH5	0.000591	0.66
209613_s_at	ADH1B	0.002306	3.08
209614_at	ADH1B	0.002707	1.83
202054_s_at	ALDH3A2	0.021057	0.88
202053_s_at	ALDH3A2	0.02513	0.51
210544_s_at	ALDH3A2	0.071194	0.50
201425_at	ALDH2	0.173691	0.64
201251_at	PKM	0.96652	0.01

***Table A10**: Downregulated genes (α=0.001) in Metabolism of xenobiotics by cytochrome P450 pathway in SSc-ILD compared to controls, sorted by parametric p-value – Metabolic pathways genes*

Probe set	Gene symbol	Parametric p-value	logFC
208848_at	ADH5	3.00E-06	1.03
203180_at	ALDH1A3	9.30E-06	1.82
209612_s_at	ADH1B	0.000313	3.69
208847_s_at	ADH5	0.000591	0.66
209613_s_at	ADH1B	0.002306	3.08
209614_at	ADH1B	0.002707	1.83

APPENDIX D – GSEA of IPF

Table A11: *Enriched pathways by GSEA (α=0.01) of all genes in stable IPF compared to rapidly progressing IPF, sorted by LS permutation p-value*

Pathway description	Number of probe sets	LS permutation p-value
Pyrimidine metabolism	45	0.00001
DNA replication	40	0.00001
Base excision repair	16	0.00001
Nucleotide excision repair	24	0.00001
Mismatch repair	19	0.00001
Cell cycle	103	0.00001
Progesterone-mediated oocyte maturation	54	0.00001
Homologous recombination	17	0.00023
One carbon pool by folate	17	0.00025
Spliceosome	26	0.00049
Non-small cell lung cancer	27	0.00160
Protein processing in endoplasmic reticulum	67	0.00178
Ribosome biogenesis in eukaryotes	11	0.00179
Lysine degradation	24	0.00206
Oocyte meiosis	67	0.00233
Basal transcription factors	6	0.00326
Glioma	49	0.00451
RNA transport	40	0.00454
Bacterial invasion of epithelial cells	27	0.00681
Chronic myeloid leukaemia	39	0.00707
Purine metabolism	89	0.00841
Type II diabetes mellitus	19	0.00929
Ubiquitin mediated proteolysis	35	0.00947
Other glycan degradation	6	0.01252
Protein export	5	0.01688
Mucin type O-Glycan biosynthesis	26	0.04634
Osteoclast differentiation	62	0.06275
Antigen processing and presentation	36	0.06836
Bladder cancer	31	0.08121

APPENDIX E – Genes of enriched pathways in IPF

Tables of genes involved in important significantly enriched pathways by GSEA (α=0.01) – dataset GSE44723.

Table A12: *Downregulated genes (α=0.01) in Pyrimidine metabolism pathway in rapidly progressing IPF compared to stable IPF, sorted by parametric p-value – all genes*

Probe set	Gene symbol	Parametric p-value	logFC
204077_x_at	ENTPD4	0.0217312	-0.67
203234_at	UPP1	0.0247452	-0.94
209474_s_at	ENTPD1	0.3163895	-0.38
223342_at	RRM2B	0.3208927	-0.30
1553994_at	NT5E	0.3890537	-0.45
207691_x_at	ENTPD1	0.4075008	-0.30
205627_at	CDA	0.5090810	-0.30
227556_at	NME7	0.5429829	-0.23
203939_at	NT5E	0.5976266	-0.25
209473_at	ENTPD1	0.6939727	-0.18
201695_s_at	PNP	0.8350958	-0.09
227486_at	NT5E	0.8414048	-0.12

Table A13: *Upregulated genes (α=0.01) in Pyrimidine metabolism pathway in rapidly progressing IPF compared to stable IPF, sorted by parametric p-value – all genes*

Probe set	Gene symbol	Parametric p-value	logFC
225291_at	PNPT1	0.0014094	1.06
201476_s_at	RRM1	0.0014875	0.76
205909_at	POLE2	0.0016366	2.31
205628_at	PRIM2	0.0018440	1.05
226702_at	CMPK2	0.0029050	3.64
1554696_s_at	TYMS	0.0040362	2.36
1553983_at	DTYMK	0.0040484	0.69
202589_at	TYMS	0.0047698	2.43
208828_at	POLE3	0.0052202	0.99
1553984_s_at	DTYMK	0.0056362	1.16
205053_at	PRIM1	0.0058456	2.26
204835_at	POLA1	0.0080029	1.62
203270_at	DTYMK	0.0100524	1.09
201477_s_at	RRM1	0.0111839	1.01
216026_s_at	POLE	0.0112797	1.14
208956_x_at	DUT	0.0143346	0.90
203302_at	DCK	0.0150716	0.82
204441_s_at	POLA2	0.0154821	1.07
206653_at	POLR3G	0.0164935	1.34
212836_at	POLD3	0.0168184	0.95
203422_at	POLD1	0.0168464	1.04

Probe set	Gene symbol	Parametric p-value	logFC
209932_s_at	DUT	0.0169020	0.91
208955_at	DUT	0.0175086	1.77
209773_s_at	RRM2	0.0175463	2.17
201890_at	RRM2	0.0214974	2.02
202338_at	TK1	0.0275547	1.09
1554408_a_at	TK1	0.0291660	1.19
218997_at	POLR1E	0.0397613	0.52
233341_s_at	POLR1B	0.0403370	0.76
217647_at	DHODH	0.1572036	0.65
202613_at	CTPS1	0.2053446	0.40
204646_at	DPYD	0.2191307	0.45
206197_at	NME5	0.5494710	0.21

Table A14*: Upregulated genes (α=0.01) in DNA replication pathway in rapidly progressing IPF compared to stable IPF, sorted by parametric p-value – all genes*

Probe set	Gene symbol	Parametric p-value	logFC
205909_at	POLE2	0.0016366	2.31
203022_at	RNASEH2A	0.0016891	1.79
202726_at	LIG1	0.0017467	1.07
205628_at	PRIM2	0.0018440	1.05
203210_s_at	RFC5	0.0018636	1.59
201202_at	PCNA	0.0019247	1.44
204767_s_at	FEN1	0.0028930	1.60
204768_s_at	FEN1	0.0029900	1.54
202107_s_at	MCM2	0.0030816	1.74
201930_at	MCM6	0.0034859	1.90
204023_at	RFC4	0.0037503	1.56
212141_at	MCM4	0.0044429	0.93
204128_s_at	RFC3	0.0044786	1.69
201528_at	RPA1	0.0046364	1.01
216237_s_at	MCM5	0.0051082	1.94
208828_at	POLE3	0.0052202	0.99
201755_at	MCM5	0.0053360	1.41
203209_at	RFC5	0.0056880	1.60
205053_at	PRIM1	0.0058456	2.26
204127_at	RFC3	0.0059669	1.51
201555_at	MCM3	0.0060818	1.56
222036_s_at	MCM4	0.0063213	1.40
209507_at	RPA3	0.0064839	1.43
219056_at	RNASEH2B	0.0065023	2.17
213647_at	DNA2	0.0066281	1.98
222037_at	MCM4	0.0072396	1.54
204835_at	POLA1	0.0080029	1.62

Probe set	Gene symbol	Parametric p-value	logFC
208795_s_at	MCM7	0.0081063	1.53
210983_s_at	MCM7	0.0096923	1.51
216026_s_at	POLE	0.0112797	1.14
212142_at	MCM4	0.0145609	0.90
204441_s_at	POLA2	0.0154821	1.07
212836_at	POLD3	0.0168184	0.95
203422_at	POLD1	0.0168464	1.04
1053_at	RFC2	0.0181381	1.01
238977_at	MCM6	0.0189354	1.90
203696_s_at	RFC2	0.0459311	0.87
209085_x_at	RFC1	0.0702684	0.50
214060_at	SSBP1	0.2170583	0.24
236675_at	RPA1	0.3757452	0.36

***Table A15**: Downregulated genes (α=0.01) in One carbon pool by folate pathway in rapidly progressing IPF compared to stable IPF, sorted by parametric p-value – all genes*

Probe set	Gene symbol	Parametric p-value	logFC
231202_at	ALDH1L2	0.0314808	-1.47
1556841_a_at	ALDH1L2	0.0326266	-0.74
1559393_at	ALDH1L2	0.0519851	-0.49
220346_at	MTHFD2L	0.5867397	-0.17

***Table A16**: Upregulated genes (α=0.01) in One carbon pool by folate pathway in rapidly progressing IPF compared to stable IPF, sorted by parametric p-value – all genes*

Probe set	Gene symbol	Parametric p-value	logFC
208758_at	ATIC	0.0021904	0.88
202533_s_at	DHFR	0.0026556	1.69
1554696_s_at	TYMS	0.0040362	2.36
202589_at	TYMS	0.0047698	2.43
202534_x_at	DHFR	0.0056074	1.62
48808_at	DHFR	0.0066937	1.68
202309_at	MTHFD1	0.0073173	1.24
202532_s_at	DHFR	0.0075526	1.66
238762_at	MTHFD2L	0.0148880	1.08
230097_at	GART	0.0508094	0.76
239562_at	MTHFD2L	0.1382986	0.76
234976_x_at	MTHFD2	0.1676506	0.52
1554841_at	MTHFD2L	0.3724131	0.20

Table A17*: Downregulated genes (α=0.01) in Purine metabolism pathway in rapidly progressing IPF compared to stable IPF, sorted by parametric p-value – all genes*

Probe set	Gene symbol	Parametric p-value	logFC
212522_at	PDE8A	0.0056566	-0.69
236344_at	PDE1C	0.0063397	-1.32
216869_at	PDE1C	0.0094213	-0.86
239218_at	PDE1C	0.0159725	-1.56
204077_x_at	ENTPD4	0.0217312	-0.67
205501_at	PDE10A	0.0244203	-1.15
236300_at	PDE3A	0.0302880	-1.18
222862_s_at	AK5	0.0452287	-1.18
228962_at	PDE4D	0.0572375	-0.58
228507_at	PDE3A	0.0590218	-1.29
243438_at	PDE7B	0.0783267	-0.62
211302_s_at	PDE4B	0.0787966	-0.84
207992_s_at	AMPD3	0.0809413	-0.69
219308_s_at	AK5	0.0898381	-1.18
230109_at	PDE7B	0.0967604	-1.18
203708_at	PDE4B	0.1093486	-1.25
1562227_at	PDE5A	0.2249707	-0.47
205593_s_at	PDE9A	0.2359090	-0.49
209474_s_at	ENTPD1	0.3163895	-0.38
223342_at	RRM2B	0.3208927	-0.30
1553994_at	NT5E	0.3890537	-0.45
233547_x_at	PDE1A	0.3896786	-0.29
207691_x_at	ENTPD1	0.4075008	-0.30
1558680_s_at	PDE1A	0.4407162	-0.25
227088_at	PDE5A	0.4835999	-0.49
206757_at	PDE5A	0.5324234	-0.42
227556_at	NME7	0.5429829	-0.23
208396_s_at	PDE1A	0.5538320	-0.40
204491_at	PDE4D	0.5739310	-0.27
203939_at	NT5E	0.5976266	-0.25
240088_at	PDE5A	0.6057441	-0.17
236234_at	PDE1A	0.6358906	-0.15
1553175_s_at	PDE5A	0.6380457	-0.14
1562228_s_at	PDE5A	0.6654720	-0.15
231213_at	PDE1A	0.6724942	-0.22
209473_at	ENTPD1	0.6939727	-0.18
226325_at	ADSSL1	0.7633995	-0.20
201695_s_at	PNP	0.8350958	-0.09
227486_at	NT5E	0.8414048	-0.12
223272_s_at	NTPCR	0.8567387	-0.04
241994_at	XDH	0.9202875	-0.06

***Table A18**: Upregulated genes (α=0.01) in Purine metabolism pathway in rapidly progressing IPF compared to stable IPF, sorted by parametric p-value – all genes*

Probe set	Gene symbol	Parametric p-value	logFC
201892_s_at	IMPDH2	0.0012528	0.89
225291_at	PNPT1	0.0014094	1.06
201476_s_at	RRM1	0.0014875	0.76
205909_at	POLE2	0.0016366	2.31
205628_at	PRIM2	0.0018440	1.05
208758_at	ATIC	0.0021904	0.88
201013_s_at	PAICS	0.0035761	1.24
201014_s_at	PAICS	0.0039213	1.42
208828_at	POLE3	0.0052202	0.99
213302_at	PFAS	0.0055167	1.53
205053_at	PRIM1	0.0058456	2.26
223358_s_at	PDE7A	0.0072323	1.86
204835_at	POLA1	0.0080029	1.62
225367_at	PGM2	0.0080542	0.68
212175_s_at	AK2	0.0111551	0.70
201477_s_at	RRM1	0.0111839	1.01
216026_s_at	POLE	0.0112797	1.14
202854_at	HPRT1	0.0130152	0.85
203302_at	DCK	0.0150716	0.82
204441_s_at	POLA2	0.0154821	1.07
206653_at	POLR3G	0.0164935	1.34
212836_at	POLD3	0.0168184	0.95
203422_at	POLD1	0.0168464	1.04
209773_s_at	RRM2	0.0175463	2.17
204120_s_at	ADK	0.0190332	0.74
201890_at	RRM2	0.0214974	2.02
225366_at	PGM2	0.0251897	0.66
222317_at	PDE3B	0.0341903	2.23
209433_s_at	PPAT	0.0362300	0.82
224046_s_at	PDE7A	0.0367858	1.20
204639_at	ADA	0.0382945	3.02
218997_at	POLR1E	0.0397613	0.52
233341_s_at	POLR1B	0.0403370	0.76
214582_at	PDE3B	0.0420283	1.76
216705_s_at	ADA	0.0435364	2.86
230097_at	GART	0.0508094	0.76
204119_s_at	ADK	0.0708036	0.53
212174_at	AK2	0.0800793	0.49
209440_at	PRPS1	0.4377859	0.29
230352_at	PRPS2	0.5022550	0.18
228952_at	ENPP1	0.5397365	0.34
206197_at	NME5	0.5494710	0.21

Probe set	Gene symbol	Parametric p-value	logFC
208447_s_at	PRPS1	0.6494109	0.19
229088_at	ENPP1	0.6870518	0.37
224209_s_at	GDA	0.7203912	0.18
205066_s_at	ENPP1	0.9075678	0.11
203741_s_at	ADCY7	0.9435177	0.04
209321_s_at	ADCY3	0.9811355	0.01

Table A19*: Downregulated genes (α=0.01) in Osteoclast differentiation pathway in rapidly progressing IPF compared to stable IPF, sorted by parametric p-value – all genes*

Probe set	Gene symbol	Parametric p-value	logFC
209189_at	FOS	0.003058	-1.40
202450_s_at	CTSK	0.006722	-2.25
204933_s_at	TNFRSF11B	0.007955	-1.89
211676_s_at	IFNGR1	0.008680	-0.64
1552610_a_at	JAK1	0.009024	-0.89
204627_s_at	ITGB3	0.010495	-1.22
1552611_a_at	JAK1	0.010992	-1.00
204932_at	TNFRSF11B	0.011150	-1.51
201648_at	JAK1	0.013787	-0.79
207233_s_at	MITF	0.015150	-0.86
204628_s_at	ITGB3	0.018706	-0.79
201471_s_at	SQSTM1	0.023080	-0.89
227697_at	SOCS3	0.028083	-2.32
202948_at	IL1R1	0.033502	-1.47
206359_at	SOCS3	0.036310	-1.60
39582_at	CYLD	0.048504	-0.51
205205_at	RELB	0.052184	-0.60
213295_at	CYLD	0.059492	-0.58
226066_at	MITF	0.064035	-0.81
225636_at	STAT2	0.064541	-0.79
215561_s_at	IL1R1	0.075050	-0.45
203752_s_at	JUND	0.091650	-0.34
209909_s_at	TGFB2	0.099261	-0.92
213112_s_at	SQSTM1	0.108901	-0.67
228442_at	NFATC2	0.119038	-0.84
229029_at	CAMK4	0.120376	-0.60
224793_s_at	TGFBR1	0.125840	-0.36
39402_at	IL1B	0.143776	-2.25
205067_at	IL1B	0.144697	-2.40
228121_at	TGFB2	0.156321	-0.64
202897_at	SIRPA	0.162888	-0.64
32541_at	PPP3CC	0.163603	-0.40
204813_at	MAPK10	0.178523	-0.58

Probe set	Gene symbol	Parametric p-value	logFC
204638_at	ACP5	0.205251	-0.49
222880_at	AKT3	0.246106	-0.30
210118_s_at	IL1A	0.251436	-1.06
226991_at	NFATC2	0.254158	-0.56
213281_at	JUN	0.322681	-0.43
220407_s_at	TGFB2	0.351414	-0.38
216033_s_at	FYN	0.361407	-0.34
201464_x_at	JUN	0.417200	-0.38
201465_s_at	JUN	0.463437	-0.32
201502_s_at	NFKBIA	0.520934	-0.22
201466_s_at	JUN	0.536944	-0.34
200887_s_at	STAT1	0.542287	-0.22
AFFX-HUMISGF3A/M97935_3_at	STAT1	0.554775	-0.22
210105_s_at	FYN	0.572800	-0.20
212486_s_at	FYN	0.579435	-0.27
208510_s_at	PPARG	0.664412	-0.15
241871_at	CAMK4	0.884169	-0.07

Table A20: *Upregulated genes (α=0.01) in Osteoclast differentiation pathway in rapidly progressing IPF compared to stable IPF, sorted by parametric p-value – all genes*

Gene set	Gene symbol	Parametric p-value	logFC
212239_at	PIK3R1	0.001329	0.83
212240_s_at	PIK3R1	0.002687	0.94
202743_at	PIK3R3	0.003138	1.54
205698_s_at	MAP2K6	0.007669	1.91
210001_s_at	SOCS1	0.009300	0.73
211580_s_at	PIK3R3	0.013292	0.77
230917_at	PLCG2	0.077044	1.40
1552263_at	MAPK1	0.084414	0.43
236561_at	TGFBR1	0.272305	0.41
211105_s_at	NFATC1	0.307455	0.52
209949_at	NCF2	0.402598	0.66
209969_s_at	STAT1	0.947796	0.03

***Table A21**: Downregulated genes in Pyrimidine metabolism (α=0.01) pathway in rapidly progressing IPF compared to stable IPF, sorted by parametric p-value – Metabolic pathways genes*

Probe set	Gene symbol	Parametric p-value	logFC
203234_at	UPP1	0.024745	-0.94
223342_at	RRM2B	0.320893	-0.30
1553994_at	NT5E	0.389054	-0.45
205627_at	CDA	0.509081	-0.30
227556_at	NME7	0.542983	-0.23
203939_at	NT5E	0.597627	-0.25
201695_s_at	PNP	0.835096	-0.09
227486_at	NT5E	0.841405	-0.12

***Table A22**: Upregulated genes (α=0.01) in Pyrimidine metabolism pathway in rapidly progressing IPF compared to stable IPF, sorted by parametric p-value – Metabolic pathways genes*

Probe set	Gene symbol	Parametric p-value	logFC
201476_s_at	RRM1	0.001488	0.76
205909_at	POLE2	0.001637	2.31
205628_at	PRIM2	0.001844	1.05
226702_at	CMPK2	0.002905	3.64
1554696_s_at	TYMS	0.004036	2.36
1553983_at	DTYMK	0.004048	0.69
202589_at	TYMS	0.004770	2.43
208828_at	POLE3	0.005220	0.99
1553984_s_at	DTYMK	0.005636	1.16
205053_at	PRIM1	0.005846	2.26
204835_at	POLA1	0.008003	1.62
203270_at	DTYMK	0.010052	1.09
201477_s_at	RRM1	0.011184	1.01
216026_s_at	POLE	0.011280	1.14
208956_x_at	DUT	0.014335	0.90
203302_at	DCK	0.015072	0.82
204441_s_at	POLA2	0.015482	1.07
206653_at	POLR3G	0.016494	1.34
212836_at	POLD3	0.016818	0.95
203422_at	POLD1	0.016846	1.04
209932_s_at	DUT	0.016902	0.91
208955_at	DUT	0.017509	1.77
209773_s_at	RRM2	0.017546	2.17
201890_at	RRM2	0.021497	2.02
202338_at	TK1	0.027555	1.09
1554408_a_at	TK1	0.029166	1.19
218997_at	POLR1E	0.039761	0.52
233341_s_at	POLR1B	0.040337	0.76
217647_at	DHODH	0.157204	0.65

Probe set	Gene symbol	Parametric p-value	logFC
202613_at	CTPS1	0.205345	0.40
204646_at	DPYD	0.219131	0.45
206197_at	NME5	0.549471	0.21

***Table A23**: Upregulated genes (α=0.01) in DNA replication pathway in rapidly progressing IPF compared to stable IPF, sorted by parametric p-value – Metabolic pathways genes*

Probe set	Gene symbol	Parametric p-value	logFC
205909_at	POLE2	0.001637	2.31
205628_at	PRIM2	0.001844	1.05
208828_at	POLE3	0.005220	0.99
205053_at	PRIM1	0.005846	2.26
204835_at	POLA1	0.008003	1.62
216026_s_at	POLE	0.011280	1.14
204441_s_at	POLA2	0.015482	1.07
212836_at	POLD3	0.016818	0.95
203422_at	POLD1	0.016846	1.04
214060_at	SSBP1	0.217058	0.24

***Table A24**: Downregulated genes (α=0.01) in One carbon pool by folate pathway in rapidly progressing IPF compared to stable IPF, sorted by parametric p-value – Metabolic pathways genes*

Probe set	Gene symbol	Parametric p-value	logFC
220346_at	MTHFD2L	0.58674	-0.17

***Table A25**: Upregulated genes (α=0.01) in One carbon pool by folate pathway in rapidly progressing IPF compared to stable IPF, sorted by parametric p-value – Metabolic pathways genes*

Probe set	Gene symbol	Parametric p-value	logFC
208758_at	ATIC	0.002190	0.88
202533_s_at	DHFR	0.002656	1.69
1554696_s_at	TYMS	0.004036	2.36
202589_at	TYMS	0.004770	2.43
202534_x_at	DHFR	0.005607	1.62
48808_at	DHFR	0.006694	1.68
202309_at	MTHFD1	0.007317	1.24
202532_s_at	DHFR	0.007553	1.66
238762_at	MTHFD2L	0.014888	1.08
230097_at	GART	0.050809	0.76
239562_at	MTHFD2L	0.138299	0.76
234976_x_at	MTHFD2	0.167651	0.52
1554841_at	MTHFD2L	0.372413	0.20

***Table A26**: Downregulated genes (α=0.01) in Purine metabolism pathway in rapidly progressing IPF compared to stable IPF, sorted by parametric p-value – Metabolic pathways genes*

Probe set	Gene symbol	Parametric p-value	logFC
222862_s_at	AK5	0.045229	-1.18
207992_s_at	AMPD3	0.080941	-0.69
219308_s_at	AK5	0.089838	-1.18
223342_at	RRM2B	0.320893	-0.30
1553994_at	NT5E	0.389054	-0.45
227556_at	NME7	0.542983	-0.23
203939_at	NT5E	0.597627	-0.25
226325_at	ADSSL1	0.763400	-0.20
201695_s_at	PNP	0.835096	-0.09
227486_at	NT5E	0.841405	-0.12
223272_s_at	NTPCR	0.856739	-0.04
241994_at	XDH	0.920288	-0.06

***Table A27**: Upregulated genes (α=0.01) in Purine metabolism pathway in rapidly progressing IPF compared to stable IPF, sorted by parametric p-value – Metabolic pathways genes*

Probe set	Gene symbol	Parametric p-value	logFC
201892_s_at	IMPDH2	0.001253	0.89
201476_s_at	RRM1	0.001488	0.76
205909_at	POLE2	0.001637	2.31
205628_at	PRIM2	0.001844	1.05
208758_at	ATIC	0.002190	0.88
201013_s_at	PAICS	0.003576	1.24
201014_s_at	PAICS	0.003921	1.42
208828_at	POLE3	0.005220	0.99
213302_at	PFAS	0.005517	1.53
205053_at	PRIM1	0.005846	2.26
204835_at	POLA1	0.008003	1.62
225367_at	PGM2	0.008054	0.68
212175_s_at	AK2	0.011155	0.70
201477_s_at	RRM1	0.011184	1.01
216026_s_at	POLE	0.011280	1.14
202854_at	HPRT1	0.013015	0.85
203302_at	DCK	0.015072	0.82
204441_s_at	POLA2	0.015482	1.07
206653_at	POLR3G	0.016494	1.34
212836_at	POLD3	0.016818	0.95
203422_at	POLD1	0.016846	1.04
209773_s_at	RRM2	0.017546	2.17
204120_s_at	ADK	0.019033	0.74
201890_at	RRM2	0.021497	2.02
225366_at	PGM2	0.025190	0.66

Probe set	**Gene symbol**	**Parametric p-value**	**logFC**
209433_s_at	PPAT	0.036230	0.82
204639_at	ADA	0.038295	3.02
218997_at	POLR1E	0.039761	0.52
233341_s_at	POLR1B	0.040337	0.76
216705_s_at	ADA	0.043536	2.86
230097_at	GART	0.050809	0.76
204119_s_at	ADK	0.070804	0.53
212174_at	AK2	0.080079	0.49
209440_at	PRPS1	0.437786	0.29
230352_at	PRPS2	0.502255	0.18
228952_at	ENPP1	0.539737	0.34
206197_at	NME5	0.549471	0.21
208447_s_at	PRPS1	0.649411	0.19
229088_at	ENPP1	0.687052	0.37
224209_s_at	GDA	0.720391	0.18
205066_s_at	ENPP1	0.907568	0.11

APPENDIX F – DE genes of Metabolic pathways analysis (SSc and IPF)

Table A28: *DE genes (SSc-ILD vs. controls – upregulated in orange and downregulated in black) by differential expression analysis (α=0.01) of Metabolic pathways genes*

Probe set	Gene symbol	Defined gene list	Parametric p-value	logFC
208937_s_at	ID1	TGF-β signalling pathway	< 1e-07	-5.80
207826_s_at	ID3	TGF-β signalling pathway	< 1e-07	-4.18
212226_s_at	PPAP2B	Metabolic pathways	< 1e-07	2.15
217933_s_at	LAP3	Metabolic pathways	2.00E-07	2.43
204790_at	SMAD7	TGF beta signalling pathway,	3.00E-07	-1.51
201516_at	SRM	Metabolic pathways	5.00E-07	-1.09
204608_at	ASL	Metabolic pathways	5.00E-07	-0.86
213011_s_at	TPI1	Glycolysis/Gluconeogenesis, Metabolic pathways	5.00E-07	-1.03
209355_s_at	PPAP2B	Metabolic pathways	5.00E-07	2.34
212230_at	PPAP2B	Metabolic pathways	6.00E-07	2.44
208941_s_at	SEPHS1	Metabolic pathways	7.00E-07	-0.89
213725_x_at	XYLT1	Metabolic pathways	7.00E-07	-2.84
201577_at	NME1	Metabolic pathways	9.00E-07	-1.15
36936_at	TSTA3	Metabolic pathways	9.00E-07	-0.81
204224_s_at	GCH1	Metabolic pathways	1.10E-06	2.39
208944_at	TGFBR2	TGF beta signalling pathway	1.70E-06	0.93
208848_at	ADH5	Fatty acid degradation, Glycolysis/Gluconeogenesis, Metabolic pathways	3.00E-06	1.03
200790_at	ODC1	Metabolic pathways	3.90E-06	-1.60
210511_s_at	INHBA	TGF-β signalling pathway	6.00E-06	-2.47
207388_s_at	PTGES	Metabolic pathways	6.30E-06	0.97
203157_s_at	GLS	D-Glutamine and D-glutamate metabolism, Metabolic pathways, Nitrogen metabolism	7.20E-06	-1.25
217294_s_at	ENO1	Glycolysis/Gluconeogenesis, Metabolic pathways	7.40E-06	-1.03
204241_at	ACOX3	Fatty acid degradation, Metabolic pathways, PPAR signalling pathway	7.50E-06	-0.94
203180_at	ALDH1A3	Glycolysis/Gluconeogenesis, Metabolic pathways	9.30E-06	1.82
202613_at	CTPS1	Metabolic pathways	9.70E-06	-2.32
201037_at	PFKP	Glycolysis/Gluconeogenesis, Metabolic pathways, Pentose phosphate pathway	9.70E-06	-1.89

Probe set	Gene symbol	Defined gene list	Parametric p-value	logFC
208447_s_at	PRPS1	Metabolic pathways, Pentose phosphate pathway	1.08E-05	-2.40
203159_at	GLS	D-Glutamine and D-glutamate metabolism, Metabolic pathways, Nitrogen metabolism	1.12E-05	-1.56
201231_s_at	ENO1	Glycolysis/Gluconeogenesis, Metabolic pathways	1.20E-05	-0.84
201014_s_at	PAICS	Metabolic pathways	1.24E-05	-0.97
210367_s_at	PTGES	Metabolic pathways	1.28E-05	1.51
200737_at	PGK1	Glycolysis/Gluconeogenesis, Metabolic pathways	1.32E-05	-0.97
218189_s_at	NANS	Metabolic pathways	1.42E-05	-0.94
209440_at	PRPS1	Metabolic pathways, Pentose phosphate pathway	1.89E-05	-1.89
208308_s_at	GPI	Glycolysis/Gluconeogenesis, Metabolic pathways, Pentose phosphate pathway	2.19E-05	-0.89
211813_x_at	DCN	TGF-β signalling pathway	2.29E-05	1.16
200822_x_at	TPI1	Glycolysis/Gluconeogenesis, Metabolic pathways	2.32E-05	-0.86
201893_x_at	DCN	TGF-β signalling pathway	2.50E-05	0.93
210337_s_at	ACLY	Citrate cycle (TCA cycle), Metabolic pathways	2.58E-05	-1.12
200078_s_at	ATP6V0B	Metabolic pathways, Oxidative phosphorylation	2.80E-05	-0.84
210029_at	IDO1	Metabolic pathways	2.89E-05	1.16
208972_s_at	ATP5G1	Metabolic pathways, Oxidative phosphorylation	3.40E-05	-0.79
208116_s_at	MAN1A1	Metabolic pathways	3.44E-05	1.01
207357_s_at	GALNT10	Metabolic pathways	3.65E-05	-1.60
217356_s_at	PGK1	Glycolysis/Gluconeogenesis, Metabolic pathways	3.70E-05	-1.22
212256_at	GALNT10	Metabolic pathways	4.69E-05	-1.64
214390_s_at	BCAT1	Metabolic pathways	5.29E-05	-0.67
200738_s_at	PGK1	Glycolysis/Gluconeogenesis, Metabolic pathways	5.41E-05	-0.92
201272_at	AKR1B1	Metabolic pathways, Pyruvate metabolism	5.78E-05	1.27
207992_s_at	AMPD3	Metabolic pathways	5.89E-05	1.10
209147_s_at	PPAP2A	Metabolic pathways	6.48E-05	1.29
212322_at	SGPL1	Metabolic pathways	6.83E-05	-0.62
208905_at	CYCS	Apoptosis	7.32E-05	-0.97
205396_at	SMAD3	TGF beta signalling pathway	8.30E-05	1.23
204881_s_at	UGCG	Metabolic pathways	8.51E-05	1.14
217993_s_at	MAT2B	Metabolic pathways	8.76E-05	0.72

Probe set	Gene symbol	Defined gene list	Parametric p-value	logFC
201013_s_at	PAICS	Metabolic pathways	8.93E-05	-0.81
201127_s_at	ACLY	Citrate cycle (TCA cycle), Metabolic pathways	9.29E-05	-1.00
200650_s_at	LDHA	Glycolysis/Gluconeogenesis, Metabolic pathways, Pyruvate metabolism	9.79E-05	-0.81
205066_s_at	ENPP1	Metabolic pathways	0.0001024	-1.74
205401_at	AGPS	Metabolic pathways	0.0001060	-1.15
203302_at	DCK	Metabolic pathways	0.0001137	-1.18
210046_s_at	IDH2	Citrate cycle (TCA cycle), Metabolic pathways	0.0001157	-1.15
201128_s_at	ACLY	Citrate cycle (TCA cycle), Metabolic pathways	0.0001273	-1.12
215813_s_at	PTGS1	Metabolic pathways	0.0001358	-1.89
35626_at	SGSH	Metabolic pathways	0.0001362	0.99
202721_s_at	GFPT1	Metabolic pathways	0.000146	-0.81
202722_s_at	GFPT1	Metabolic pathways	0.0001524	-0.84
203270_at	DTYMK	Metabolic pathways	0.000158	-0.81
211896_s_at	DCN	TGF-β signalling pathway	0.0001833	1.20
220751_s_at	FAXDC2	Metabolic pathways	0.0001893	0.90
221760_at	MAN1A1	Metabolic pathways	0.0002171	1.51
217870_s_at	CMPK1	Metabolic pathways	0.0002358	-0.71
205128_x_at	PTGS1	Metabolic pathways	0.0002363	-1.51
208070_s_at	REV3L	Metabolic pathways	3.00E-04	1.30
209335_at	DCN	TGF-β signalling pathway	0.0003125	1.68
209612_s_at	ADH1B	Fatty acid degradation, Glycolysis/Gluconeogenesis, Metabolic pathways	0.0003134	3.69
219374_s_at	ALG9	Metabolic pathways	0.0003382	-0.76
215001_s_at	GLUL	Metabolic pathways, Nitrogen metabolism	0.0003495	1.02
205397_x_at	SMAD3	TGF beta signalling pathway,	0.0003695	1.13
208131_s_at	PTGIS	Metabolic pathways	0.0004129	2.01
209293_x_at	ID4	TGF-β signalling pathway	0.0004158	-0.74
200815_s_at	PAFAH1B1	Metabolic pathways	0.0004176	-0.84
214452_at	BCAT1	Metabolic pathways	0.000559	-0.94
201476_s_at	RRM1	Metabolic pathways	0.0005772	-1.09
208847_s_at	ADH5	Fatty acid degradation, Glycolysis/Gluconeogenesis, Metabolic pathways	0.0005913	0.66
203158_s_at	GLS	D-Glutamine and D-glutamate metabolism, Metabolic pathways, Nitrogen metabolism	0.0005980	-1.06

Probe set	Gene symbol	Defined gene list	Parametric p-value	logFC
201724_s_at	GALNT1	Metabolic pathways	0.0006385	-0.79
218070_s_at	GMPPA	Metabolic pathways	0.0006484	-0.69
212334_at	GNS	Metabolic pathways	0.0006886	0.78
203085_s_at	TGFB1	MAPK Signalling Pathway, p38 MAPK Signalling Pathway, TGF beta signalling pathway	0.0006889	-1.12
201196_s_at	AMD1	Metabolic pathways	0.0006948	-0.76
205404_at	HSD11B1	Metabolic pathways	0.0007209	2.37
205083_at	AOX1	Metabolic pathways	0.0007283	1.01
203039_s_at	NDUFS1	Metabolic pathways, Oxidative phosphorylation	0.0007392	-0.58
209291_at	ID4	TGF-β signalling pathway	0.0007531	-1.36
205571_at	LIPT1	Metabolic pathways	0.0008057	0.79
209645_s_at	ALDH1B1	Fatty acid degradation, Glycolysis/Gluconeogenesis, Pyruvate metabolism	0.0008129	-1.29
210946_at	PPAP2A	Metabolic pathways	0.0008358	1.23
212335_at	GNS	Metabolic pathways	0.0008924	0.83
208828_at	POLE3	Metabolic pathways	0.0009829	-0.92

Table A29*: DE genes (rapidly progressing IPF vs. steady IPF – upregulated in orange and downregulated in black) by differential expression analysis (α=0.01) of Metabolic pathways genes*

Probe set	Gene symbol	Defined gene list	Parametric p-value	logFC
218313_s_at	GALNT7	Metabolic pathways	0.000387	1.37
222587_s_at	GALNT7	Metabolic pathways	0.000829	1.48
209397_at	ME2	Pyruvate metabolism	0.000902	1.13
219956_at	GALNT6	Metabolic pathways	0.000937	1.34
205289_at	BMP2	TGF-β signalling pathway	0.001018	-1.84
210154_at	ME2	Pyruvate metabolism	0.001188	1.12
201892_s_at	IMPDH2	Metabolic pathways	0.001253	0.89
201476_s_at	RRM1	Metabolic pathways	0.001488	0.76
205909_at	POLE2	Metabolic pathways	0.001637	2.31
209199_s_at	MEF2C	Mitochondrial biogenesis	0.001744	1.55
205290_s_at	BMP2	TGF-β signalling pathway	0.001773	-1.89
205628_at	PRIM2	Metabolic pathways	0.001844	1.05
201563_at	SORD	Metabolic pathways	0.001868	1.07
1552378_s_at	RDH10	Metabolic pathways	0.001927	-1.00
208758_at	ATIC	Metabolic pathways	0.002190	0.88
228303_at	GALNT6	Metabolic pathways	0.002225	1.12
238669_at	PTGS1	Metabolic pathways	0.002315	-2.06
209200_at	MEF2C	Mitochondrial biogenesis	0.002325	1.30

Probe set	Gene symbol	Defined gene list	Parametric p-value	logFC
205127_at	PTGS1	Metabolic pathways	0.002477	-1.89
1552306_at	ALG10	Metabolic pathways	0.002594	1.37
202533_s_at	DHFR	Metabolic pathways	0.002656	1.69
226702_at	CMPK2	Metabolic pathways	0.002905	3.64
201013_s_at	PAICS	Metabolic pathways	0.003576	1.24
201014_s_at	PAICS	Metabolic pathways	0.003921	1.42
215813_s_at	PTGS1	Metabolic pathways	0.003942	-1.89
1554696_s_at	TYMS	Metabolic pathways	0.004036	2.36
1553983_at	DTYMK	Metabolic pathways	0.004048	0.69
203228_at	PAFAH1B3	Metabolic pathways	0.004102	1.18
217848_s_at	PPA1	Oxidative phosphorylation	0.004369	0.94
205128_x_at	PTGS1	Metabolic pathways	0.00446	-1.84
201036_s_at	HADH	Fatty acid degradation, Metabolic pathways	0.004616	1.58
202438_x_at	IDS	Metabolic pathways	0.004697	-0.79
202589_at	TYMS	Metabolic pathways	0.00477	2.43
223515_s_at	COQ3	Metabolic pathways	0.004844	1.34
208828_at	POLE3	Metabolic pathways	0.00522	0.99
213302_at	PFAS	Metabolic pathways	0.005517	1.53
202534_x_at	DHFR	Metabolic pathways	0.005607	1.62
1553984_s_at	DTYMK	Metabolic pathways	0.005636	1.16
205053_at	PRIM1	Metabolic pathways	0.005846	2.26
214681_at	GK	Metabolic pathways	0.005848	-1.03
211569_s_at	HADH	Fatty acid degradation, Metabolic pathways	0.006397	1.55
48808_at	DHFR	Metabolic pathways	0.006694	1.68
203180_at	ALDH1A3	Glycolysis/Gluconeogenesis, Metabolic pathways	0.006831	-2.12
202309_at	MTHFD1	Metabolic pathways	0.007317	1.24
202532_s_at	DHFR	Metabolic pathways	0.007553	1.66
221550_at	COX15	Metabolic pathways, Oxidative phosphorylation	0.007553	0.89
201697_s_at	DNMT1	Metabolic pathways	0.007855	1.03
218440_at	MCCC1	Metabolic pathways	0.00795	0.93
204835_at	POLA1	Metabolic pathways	0.008003	1.62
239461_at	GALNT15	Metabolic pathways	0.008041	-1.40
225367_at	PGM2	Glycolysis/Gluconeogenesis, Metabolic pathways	0.008054	0.68
204158_s_at	TCIRG1	Metabolic pathways, Oxidative phosphorylation	0.008148	-0.97
215775_at	THBS1	TGF-β signalling pathway	0.008199	-0.56
213400_s_at	TBL1X	Mitochondrial biogenesis	0.008688	-0.62
235801_at	TUSC3	Metabolic pathways	0.008951	-0.81

Probe set	**Gene symbol**	**Defined gene list**	**Parametric p-value**	**logFC**
213587_s_at	ATP6V0E2	Metabolic pathways, Oxidative phosphorylation	0.009119	1.28
219257_s_at	SPHK1	Metabolic pathways	0.009144	-1.29
226021_at	RDH10	Metabolic pathways	0.009613	-1.03
207387_s_at	GK	Metabolic pathways	0.009827	-1.12

APPENDIX G – List of probe sets available for the analysis of the Mitochondrial biogenesis subset (SSc and IPF)

***Table A30**: List of 121 probe sets, sorted by parametric p-value, available for differential expression analysis (SSc vs. controls) after normalization and subsetting –Mitochondrial biogenesis genes*

Probe set	Symbol	Parametric p-value	logFC
215223_s_at	SOD2	< 1e-07	3.30
216841_s_at	SOD2	2.00E-07	2.96
221477_s_at	SOD2	6.00E-07	2.94
208905_at	CYCS	7.32E-05	-0.97
210046_s_at	IDH2	0.0001157	-1.15
201322_at	ATP5B	0.0002282	-0.45
218590_at	C10orf2	0.0003153	-0.40
209107_x_at	NCOA1	0.0014446	0.42
219169_s_at	TFB1M	0.0021482	-0.42
209105_at	NCOA1	0.0025580	0.28
212867_at	NCOA2	0.0046282	0.68
216326_s_at	HDAC3	0.0058180	-0.27
203737_s_at	PPRC1	0.0070686	-0.40
202474_s_at	HCFC1	0.0072277	-0.25
202591_s_at	SSBP1	0.0072333	-0.58
209106_at	NCOA1	0.0080209	0.42
218605_at	TFB2M	0.0088261	-0.47
211984_at	CALM1	0.0089600	-0.38
219185_at	SIRT5	0.0115379	0.28
203004_s_at	MEF2D	0.0133999	-0.30
200855_at	NCOR1	0.0138047	-0.25
206173_x_at	GABPB1	0.0157815	-0.40
209563_x_at	CALM1	0.0218286	-0.32
215078_at	SOD2	0.0227743	0.31
218292_s_at	PRKAG2	0.0292772	0.30
201834_at	PRKAB1	0.0334026	0.25
208541_x_at	TFAM	0.0334419	-0.25
205731_s_at	NCOA2	0.0335788	0.23
203003_at	MEF2D	0.0340323	-0.34
210449_x_at	MAPK14	0.0384103	-0.22
210188_at	GABPA	0.0407322	-0.27
210045_at	IDH2	0.0416791	-0.18
207243_s_at	CALM1	0.0425208	-0.30
214474_at	PRKAB2	0.0577492	-0.42
203176_s_at	TFAM	0.0614762	-0.30
200655_s_at	CALM1	0.0849745	-0.30
200653_s_at	CALM1	0.0883483	-0.38
221428_s_at	TBL1XR1	0.1009750	0.42

Probe set	Symbol	Parametric p-value	logFC
204618_s_at	GABPB1	0.1072150	-0.32
213400_s_at	TBL1X	0.1074280	0.31
200854_at	NCOR1	0.1103970	-0.27
211280_s_at	NRF1	0.1130610	-0.17
217476_at	NR1D1	0.1184230	-0.15
210249_s_at	NCOA1	0.1248630	0.31
203496_s_at	MED1	0.1415590	-0.29
211985_s_at	CALM1	0.1482650	-0.23
205633_s_at	ALAS1	0.1519780	0.39
204652_s_at	NRF1	0.1591500	-0.15
206106_at	MAPK12	0.1695700	0.14
203783_x_at	POLRMT	0.2005140	-0.14
221010_s_at	SIRT5	0.2075850	0.10
221913_at	SIRT3	0.2169620	0.12
202160_at	CREBBP	0.2179980	0.25
205732_s_at	NCOA2	0.2477040	0.11
200856_x_at	NCOR1	0.2525450	0.15
49327_at	SIRT3	0.2695580	0.18
205446_s_at	ATF2	0.2756580	-0.22
201868_s_at	TBL1X	0.2780400	-0.17
203177_x_at	TFAM	0.2819550	-0.17
202530_at	MAPK14	0.2916050	0.11
213091_at	CRTC1	0.2919330	-0.10
211499_s_at	MAPK11	0.2970560	0.11
209200_at	MEF2C	0.3062120	0.12
221562_s_at	SIRT3	0.3075310	0.10
211279_at	NRF1	0.3192650	-0.07
200622_x_at	CALM1	0.3209210	-0.15
203782_s_at	POLRMT	0.3333940	-0.17
212616_at	CHD9	0.3352560	-0.25
215605_at	NCOA2	0.3605410	0.08
203497_at	MED1	0.3616640	-0.25
202426_s_at	RXRA	0.3645390	-0.14
207968_s_at	MEF2C	0.3769420	0.07
201805_at	PRKAG1	0.4026110	-0.09
213401_s_at	TBL1X	0.4115300	0.08
201867_s_at	TBL1X	0.4275180	0.21
207709_at	PRKAA2	0.4412490	-0.06
201835_s_at	PRKAB1	0.4697280	0.08
204099_at	SMARCD3	0.4712930	-0.09
207159_x_at	CRTC1	0.4872620	-0.07
204651_at	NRF1	0.4899510	-0.06
200623_s_at	CALM1	0.4921500	-0.18
200947_s_at	GLUD1	0.4984920	0.10

Probe set	Symbol	Parametric p-value	logFC
218648_at	CRTC3	0.5117900	-0.10
202449_s_at	RXRA	0.5174090	0.10
203193_at	ESRRA	0.5211590	-0.06
213688_at	CALM1	0.5237480	-0.07
1487_at	ESRRA	0.5399710	-0.06
220047_at	SIRT4	0.5969070	0.04
204314_s_at	CREB1	0.5974750	0.07
200857_s_at	NCOR1	0.5989710	-0.09
215231_at	PRKAG2	0.6020820	-0.03
210447_at	GLUD2	0.6048560	-0.03
219195_at	PPARGC1A	0.6225600	-0.07
204760_s_at	NR1D1	0.6239580	0.07
220586_at	CHD9	0.6527850	0.04
204313_s_at	CREB1	0.6726580	-0.10
214060_at	SSBP1	0.6792820	0.04
202473_x_at	HCFC1	0.6967420	-0.03
204312_x_at	CREB1	0.7274390	-0.04
210349_at	CAMK4	0.7286710	-0.03
219231_at	TGS1	0.7293910	-0.04
214513_s_at	CREB1	0.7298040	-0.04
222248_s_at	SIRT4	0.7325070	0.04
209199_s_at	MEF2C	0.7389180	-0.06
206040_s_at	MAPK11	0.7437450	0.03
206870_at	PPARA	0.7450940	-0.03
212512_s_at	CARM1	0.7479720	-0.04
211561_x_at	MAPK14	0.7731330	-0.04
211500_at	MAPK11	0.8109190	0.03
212615_at	CHD9	0.8152320	0.03
211087_x_at	MAPK14	0.8631680	0.01
200946_x_at	GLUD1	0.9031610	0.03
210771_at	PPARA	0.9095890	0.00
205811_at	POLG2	0.9332510	0.00
31637_s_at	NR1D1	0.9468600	0.03
208979_at	NCOA6	0.9557760	0.01
211808_s_at	CREBBP	0.9568340	0.00
212984_at	ATF2	0.9697980	0.00
215794_x_at	GLUD2	0.9700170	0.00
213710_s_at	CALM1	0.9743620	0.00
201869_s_at	TBL1X	0.9812360	0.00

Table A31*: List of 197 probe sets, sorted by parametric p-value, available for differential expression analysis (SSc vs. controls) after normalization and subsetting –Mitochondrial biogenesis genes*

ProbeSet	Symbol	Parametric p-value	logFC
200854_at	NCOR1	0.000744	0.51
209199_s_at	MEF2C	0.001744	1.55
209200_at	MEF2C	0.002325	1.30
204760_s_at	NR1D1	0.003230	-0.42
205811_at	POLG2	0.003354	0.73
201868_s_at	TBL1X	0.007171	-0.38
1566932_x_at	TFB2M	0.008011	-0.32
213400_s_at	TBL1X	0.008688	-0.62
203782_s_at	POLRMT	0.010707	0.48
203003_at	MEF2D	0.011657	-0.34
226307_at	CRTC2	0.013580	-0.25
202591_s_at	SSBP1	0.014086	0.51
219231_at	TGS1	0.014427	0.70
1566931_at	TFB2M	0.014910	-0.30
200857_s_at	NCOR1	0.023314	0.24
202449_s_at	RXRA	0.023733	-0.62
201867_s_at	TBL1X	0.023998	-0.62
225452_at	MED1	0.024543	0.23
200622_x_at	CALM1	0.026926	0.63
203193_at	ESRRA	0.027585	-0.23
210046_s_at	IDH2	0.027785	1.01
202474_s_at	HCFC1	0.029779	0.42
203004_s_at	MEF2D	0.030244	-0.23
234312_s_at	ACSS2	0.030648	-0.32
200623_s_at	CALM1	0.030960	0.40
233748_x_at	PRKAG2	0.032598	-0.74
225641_at	MEF2D	0.037192	-0.47
203177_x_at	TFAM	0.037379	0.73
244689_at	PPARA	0.038830	-0.30
222634_s_at	TBL1XR1	0.039180	0.51
201869_s_at	TBL1X	0.039400	-0.45
238346_s_at	TGS1	0.040005	0.53
222582_at	PRKAG2	0.041525	-0.79
218292_s_at	PRKAG2	0.041542	-0.81
218590_at	C10orf2	0.042331	0.64
203737_s_at	PPRC1	0.044666	0.40
220047_at	SIRT4	0.045006	-0.29
200947_s_at	GLUD1	0.045294	0.32
206106_at	MAPK12	0.045305	-0.34
200856_x_at	NCOR1	0.046379	0.25

ProbeSet	**Symbol**	**Parametric p-value**	**logFC**
203497_at	MED1	0.047401	0.32
242157_at	CHD9	0.047939	0.78
235890_at	TBL1XR1	0.048837	0.45
213401_s_at	TBL1X	0.051887	-0.36
222633_at	TBL1XR1	0.054116	0.37
203176_s_at	TFAM	0.056028	0.51
243189_at	NRF1	0.060412	0.34
201322_at	ATP5B	0.065566	0.28
223013_at	TBL1XR1	0.072883	0.42
221428_s_at	TBL1XR1	0.073854	0.53
1487_at	ESRRA	0.076617	-0.22
211279_at	NRF1	0.079960	0.36
210045_at	IDH2	0.082099	0.80
216326_s_at	HDAC3	0.082935	0.29
232181_at	PPARGC1B	0.089875	1.21
225456_at	MED1	0.090836	0.31
1558631_at	PPARA	0.099322	-0.18
208905_at	CYCS	0.108071	0.25
229112_at	SIRT5	0.114331	-0.15
204651_at	NRF1	0.116675	0.49
225278_at	PRKAB2	0.118269	-0.36
229029_at	CAMK4	0.120376	-0.60
213710_s_at	CALM1	0.122069	-0.36
207968_s_at	MEF2C	0.130221	0.31
213091_at	CRTC1	0.130297	-0.20
232787_at	HELZ2	0.135262	-0.14
208979_at	NCOA6	0.136726	0.43
1553639_a_at	PPARGC1B	0.138822	0.24
204652_s_at	NRF1	0.142470	0.36
200655_s_at	CALM1	0.143315	0.21
202426_s_at	RXRA	0.146067	-0.23
200855_at	NCOR1	0.155882	0.15
206040_s_at	MAPK11	0.157273	-0.25
202160_at	CREBBP	0.157871	0.26
219185_at	SIRT5	0.164555	0.38
206173_x_at	GABPB1	0.169428	0.33
201835_s_at	PRKAB1	0.175773	0.19
1558027_s_at	PRKAB2	0.180649	-0.30
237289_at	CREB1	0.182522	0.24
212867_at	NCOA2	0.196436	0.40
31637_s_at	NR1D1	0.201438	-0.25
222248_s_at	SIRT4	0.204025	-0.25

ProbeSet	**Symbol**	**Parametric p-value**	**logFC**
215231_at	PRKAG2	0.204459	-0.12
223437_at	PPARA	0.210714	-0.17
226978_at	PPARA	0.212875	-0.18
210447_at	GLUD2	0.215465	-0.42
214060_at	SSBP1	0.217058	0.24
209563_x_at	CALM1	0.219472	0.19
208541_x_at	TFAM	0.244016	0.25
1556340_at	MAPK12	0.244934	-0.22
228177_at	CREBBP	0.257255	-0.14
218605_at	TFB2M	0.261980	0.19
209106_at	NCOA1	0.269216	0.34
236371_s_at	TGS1	0.273943	0.16
204099_at	SMARCD3	0.276655	-0.38
231144_at	SMARCD3	0.285500	-0.12
234301_s_at	TFB1M	0.300686	-0.12
207159_x_at	CRTC1	0.306118	-0.17
217476_at	NR1D1	0.308437	0.10
211087_x_at	MAPK14	0.329513	0.11
1566342_at	SOD2	0.333134	-0.51
210188_at	GABPA	0.352066	0.19
209107_x_at	NCOA1	0.352656	0.18
221477_s_at	SOD2	0.355244	-0.51
238443_at	TFAM	0.359844	0.31
238489_at	PRKAA2	0.364919	-0.10
205732_s_at	NCOA2	0.372989	0.08
1566930_at	TFB2M	0.377298	-0.06
213688_at	CALM1	0.380009	-0.27
1560981_a_at	PPARA	0.382778	0.38
210249_s_at	NCOA1	0.386329	0.19
201805_at	PRKAG1	0.387460	-0.20
201834_at	PRKAB1	0.397346	0.10
49327_at	SIRT3	0.403485	-0.09
225572_at	CREB1	0.406383	0.19
228616_at	POLRMT	0.412571	-0.09
1562442_a_at	SSBP1	0.422454	0.08
204618_s_at	GABPB1	0.426056	0.20
211499_s_at	MAPK11	0.427059	-0.12
221010_s_at	SIRT5	0.428260	0.12
200653_s_at	CALM1	0.435084	0.14
235858_at	CREBBP	0.447006	-0.10
204312_x_at	CREB1	0.449209	0.12
200946_x_at	GLUD1	0.452102	0.15

ProbeSet	Symbol	Parametric p-value	logFC
216841_s_at	SOD2	0.455916	-0.40
232879_at	CRTC3	0.464335	-0.14
219169_s_at	TFB1M	0.479494	-0.15
202530_at	MAPK14	0.493204	0.08
211808_s_at	CREBBP	0.498123	0.07
1569938_at	SIRT5	0.500065	-0.09
205633_s_at	ALAS1	0.510889	0.11
203496_s_at	MED1	0.513221	0.07
1555282_a_at	PPARGC1B	0.524583	0.15
232518_at	HELZ2	0.526083	-0.06
223904_at	PRKAG3	0.526847	-0.09
229415_at	CYCS	0.534438	0.16
205731_s_at	NCOA2	0.538601	-0.09
204313_s_at	CREB1	0.539266	0.11
211280_s_at	NRF1	0.548906	0.14
210449_x_at	MAPK14	0.555744	0.10
210349_at	CAMK4	0.559777	-0.07
232517_s_at	HELZ2	0.561164	-0.07
203783_x_at	POLRMT	0.568684	-0.06
212616_at	CHD9	0.579124	-0.07
223438_s_at	PPARA	0.579326	-0.07
206870_at	PPARA	0.582114	-0.06
233633_at	TBL1XR1	0.589328	-0.06
241619_at	CALM1	0.593738	0.06
207709_at	PRKAA2	0.595948	0.10
209105_at	NCOA1	0.596007	0.12
221913_at	SIRT3	0.596268	-0.04
235388_at	CHD9	0.604945	-0.07
227892_at	PRKAA2	0.606393	-0.25
231224_x_at	PRKAG2	0.614216	-0.07
220586_at	CHD9	0.614506	0.04
1556341_s_at	MAPK12	0.620769	-0.12
1563943_at	PPARGC1B	0.621434	-0.06
228230_at	HELZ2	0.637614	-0.06
219195_at	PPARGC1A	0.649544	-0.25
214474_at	PRKAB2	0.651353	-0.10
229586_at	CHD9	0.655269	-0.06
202473_x_at	HCFC1	0.684349	-0.03
240349_at	PRKAA2	0.685158	-0.09
232022_at	TFB1M	0.688451	-0.07
228075_x_at	TFB1M	0.700160	-0.09
239654_at	CHD9	0.712821	-0.10

ProbeSet	**Symbol**	**Parametric p-value**	**logFC**
212512_s_at	CARM1	0.722940	0.04
207243_s_at	CALM1	0.726569	-0.04
244546_at	CYCS	0.734295	-0.09
215223_s_at	SOD2	0.744542	-0.18
210771_at	PPARA	0.744690	0.08
227428_at	GABPA	0.755201	0.07
211561_x_at	MAPK14	0.767185	0.03
238441_at	PRKAA2	0.772699	-0.12
215794_x_at	GLUD2	0.778838	0.06
225565_at	CREB1	0.783737	-0.04
1568874_at	NCOA6	0.804301	-0.03
1555146_at	ATF2	0.805390	0.03
221562_s_at	SIRT3	0.813687	0.03
212984_at	ATF2	0.815867	0.04
214513_s_at	CREB1	0.823071	0.03
1570293_at	TBL1X	0.830102	0.01
237142_at	PPARA	0.850463	0.01
234313_at	NCOR1	0.853295	-0.01
218648_at	CRTC3	0.861444	-0.01
224501_at	PERM1	0.882620	-0.01
241871_at	CAMK4	0.884169	-0.07
1569141_a_at	PPARGC1A	0.895282	-0.01
215078_at	SOD2	0.912929	-0.03
231177_at	HCFC1	0.939291	0.01
211984_at	CALM1	0.940579	-0.01
204314_s_at	CREB1	0.941928	0.01
211500_at	MAPK11	0.949142	0.00
212615_at	CHD9	0.957975	-0.01
205446_s_at	ATF2	0.970069	-0.01
211985_s_at	CALM1	0.970843	0.01
215605_at	NCOA2	0.975854	0.00

APPENDIX H – Heatmaps (differential expression analysis of Metabolism pathways) and explanatory information regarding the names of samples

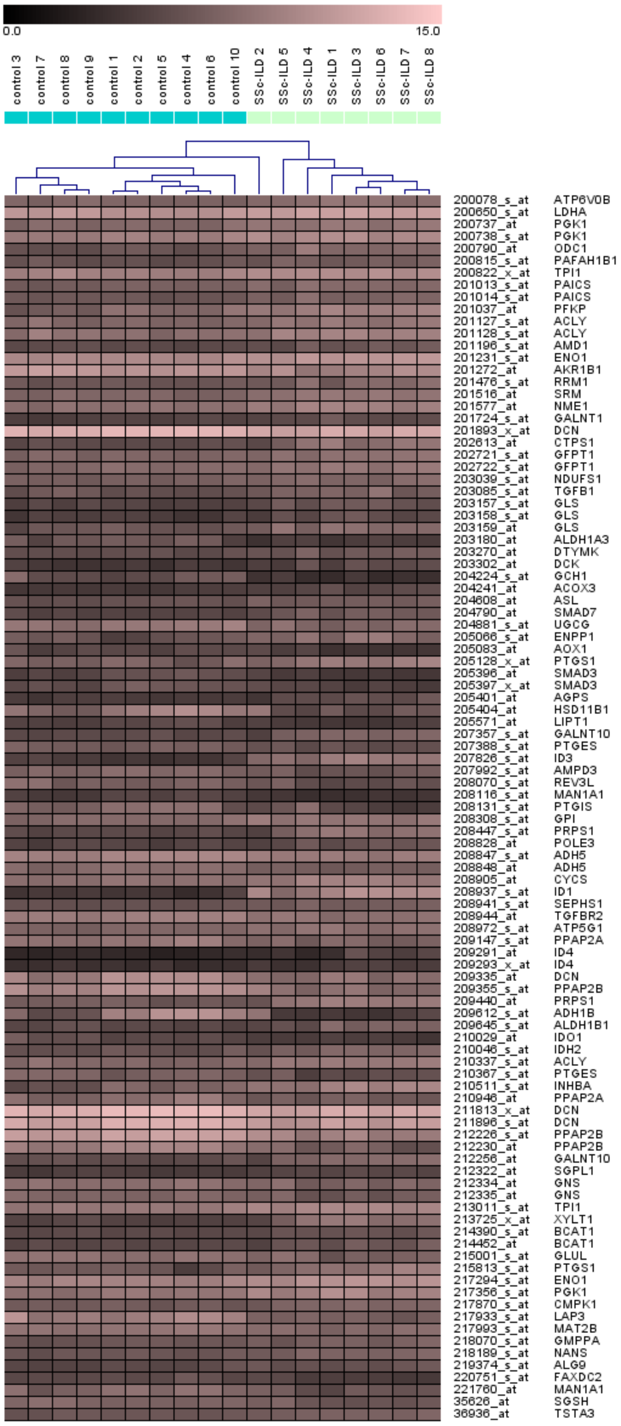

Figure A1: Heatmap of gene expression values for DE genes (α=0.001) in SSc-ILD patients and controls – Metabolic pathways genes

Expression values are represented by black to pink colour gradient, ranging from 2.43 to 13.87 (lowest values in black and highest values in light pink).

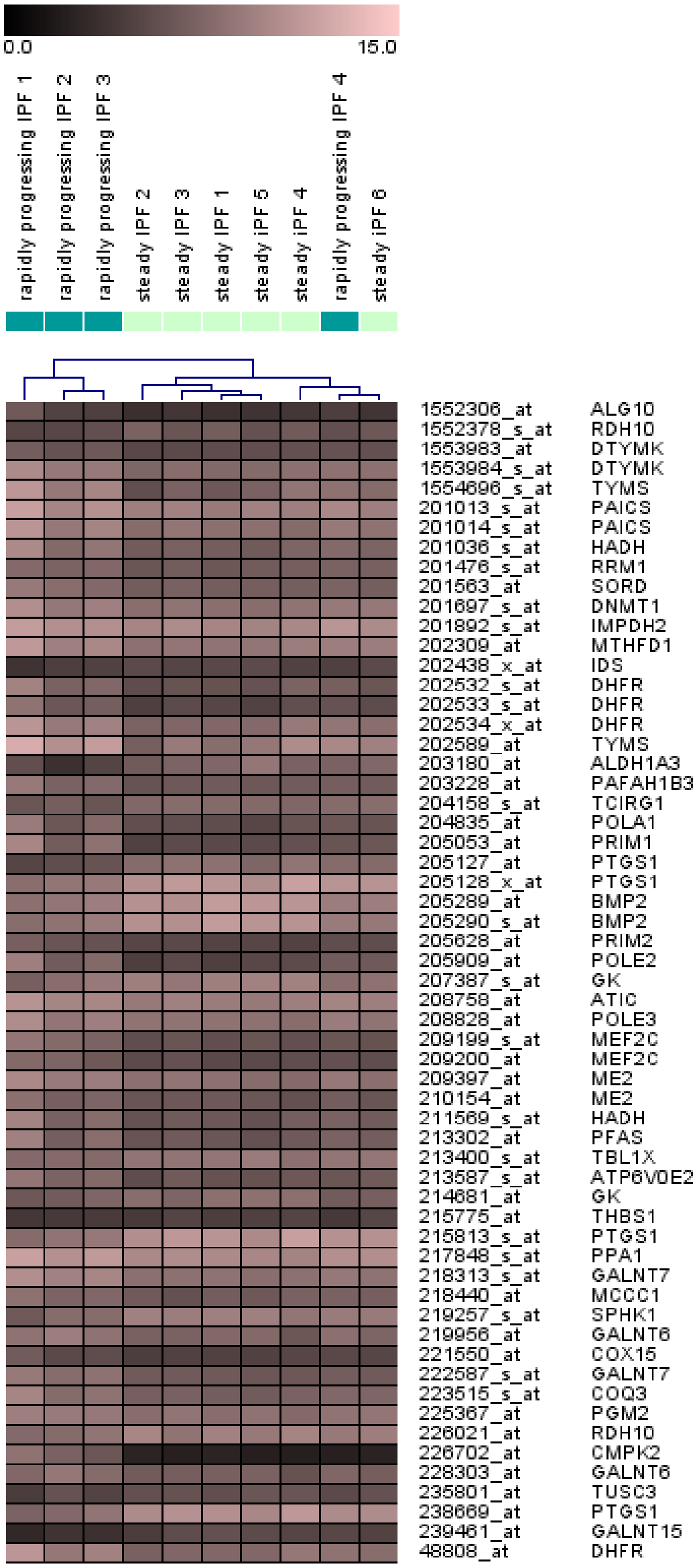

Figure A2: Heatmap of expression values for DE genes (α=0.01) in rapidly progressing IPF and steady IPF – Metabolic pathways genes

Expression values are represented by black to pink colour gradient, ranging from 2.29 to 12.72 (lowest values in black and highest values in light pink).

Table A32*: List of renamed samples used when generating heatmaps in Genesis (SSc and IPF)*

GSE40839	
Original names of samples	**Renamed samples**
GSM1003058	Control 1
GSM1003059	Control 2
GSM1003060	Control 3
GSM1003061	Control 4
GSM1003062	Control 5
GSM1003063	Control 6
GSM1003064	Control 7
GSM1003065	Control 8
GSM1003066	Control 9
GSM1003067	Control 10
GSM1003069	SSc-ILD 1
GSM1003070	SSc-ILD 2
GSM1003071	SSc-ILD 3
GSM1003072	SSc-ILD 4
GSM1003073	SSc-ILD 5
GSM1003074	SSc-ILD 6
GSM1003075	SSc-ILD 7

GSE44723	
Original names of samples	**Renamed samples**
GSM1089614	Rapidly progressing IPF 1
GSM1089615	Steady IPF 1
GSM1089619	Rapidly progressing IPF 2
GSM1089621	Steady IPF 2
GSM1089622	Rapidly progressing IPF 3
GSM1089623	Rapidly progressing IPF 4
GSM1089624	Steady IPF 3
GSM1089625	Steady IPF 4
GSM1089626	Steady IPF 5
GSM1089627	Steady IPF 6

APPENDIX I – STRING schemes (SSc and IPF)

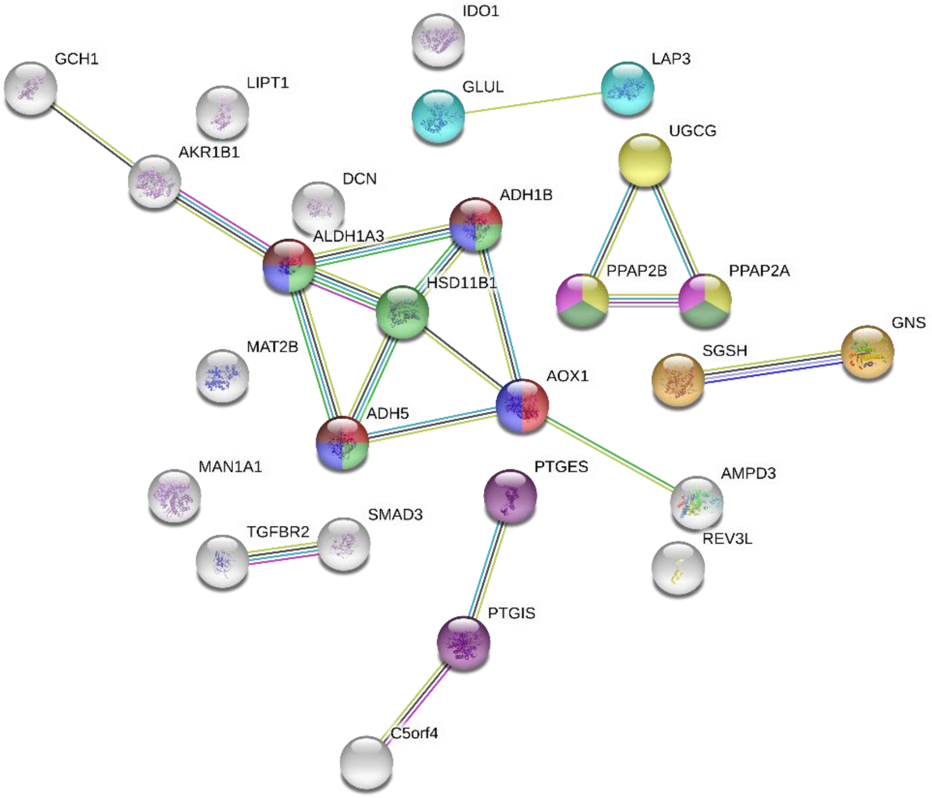

Figure A3: STRING scheme of 26 downregulated genes (SSc-ILD vs. controls) – Metabolic pathways and TGF-β pathway genes

There are 26 nodes (proteins) and 19 edges (protein-protein associations) in this network. Such an enrichment indicates that the proteins are at least partially biologically connected, as a group. Connections between genes associated with specific pathways are nicely visible. 4 genes which are in the centre of the scheme (ALDH1A3, ADH1B, AOX1 and ADH5) all participate in Tyrosine metabolism (red nodes) and Drug metabolism – Cytochrome P450 (light green nodes) pathways. 3 of them (ALDH1A3, ADH1B and ADH5) are included in Glycolysis/gluconeogenesis (brown nodes) and together with the most central gene in the scheme (HSD11B,) form a group of genes involved in Metabolism of xenobiotics by cytochrome P450 (dark blue nodes). A triangle on the right side of the scheme represents Sphingolipid metabolism (yellow nodes) which involves 3 genes (UGCG, PPAP2B and PPAP2A). 2 of them (PPAP2A and PPAP2B) have additional connection due to their association with Fat digestion and absorption (pink nodes) and Ether lipid metabolism (dark green nodes). There are also 2 connected genes above the triangle (GLUL and LAP3) which are associated with Arginine and proline metabolism (light blue nodes) and 2 connected genes below the triangle (SGSH and GNS) which are involved in Glycosaminoglycan degradation (orange nodes). In addition, we observe connection between genes PTGIS and PTGES which are involved in Arachidonic acid metabolism (purple nodes).

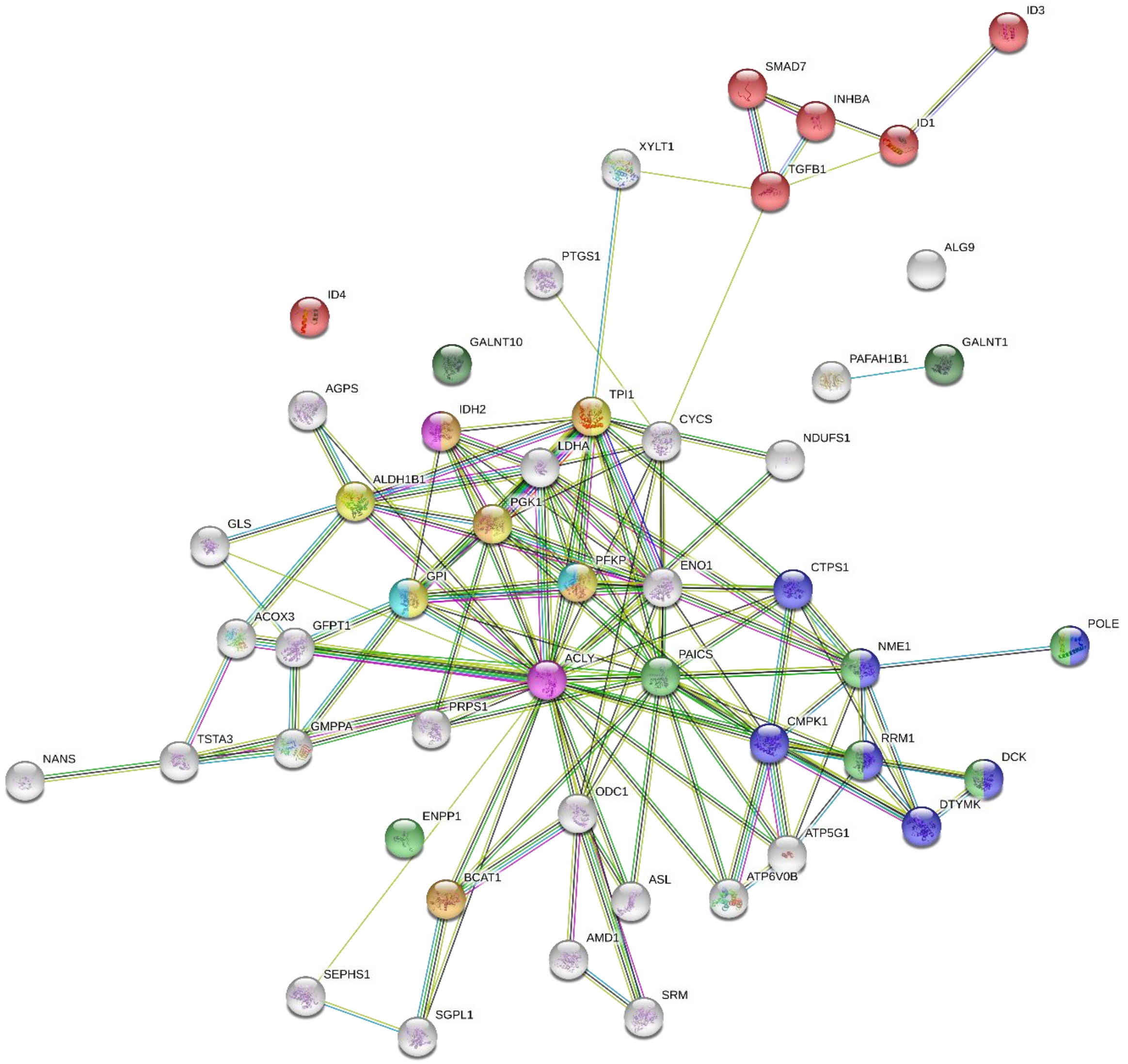

Figure A4: STRING scheme of 49 upregulated genes (SSc-ILD vs. controls) – Metabolic pathways and TGF-β pathway genes

There are 49 nodes (proteins) and 134 edges (protein-protein associations) in this network. Such an enrichment indicates that the proteins are at least partially biologically connected, as a group. There are 5 connected genes in the upper right corner of the scheme (SMAD7, TGFB1, INHBA, ID1 and ID3) which are all involved in TGF-β signalling pathway (red nodes). In the lower right corner, there are 7 connected genes (CTPS1, NME1, POLE, CMPK1, RRM1, DCK, and DTYMK) which are included in Pyrimidine metabolism (dark blue nodes). 4 of them (NME1, POLE, RRM1 and DCK), with the addition of PAICS, form a group of genes involved in Purine metabolism (green nodes). In addition, 5 yellow nodes represent genes involved in Glycolysis/gluconeogenesis and 5 orange nodes represent genes involved in Biosynthesis of amino acids. Genes associated with other pathways (for example TCA cycle – pink nodes, Mucin type O-Glycan biosynthesis – dark green nodes and Pentose phosphate pathway – light blue nodes) mostly form groups of 2 and their connections are not as nicely visible as in pathways described thus far.

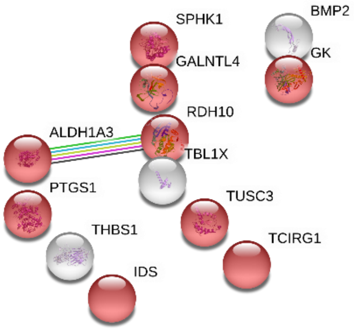

Figure A5: STRING scheme of 12 downregulated genes (rapidly progressing IPF vs. steady IPF) - Metabolic pathways and TGF-β pathway genes

There are 12 nodes (proteins) and 1 edge, which represents predicted association between genes ALDH1A3 and RDH10. Black line indicates interaction based on co-expression, yellow line represents connection based on textminig, green line shows predicted interaction based on gene neighbourhoods and light blue line represents known interaction from curated databases which are also experimentally determined (pink line). All red marked genes are included in Metabolic pathways.

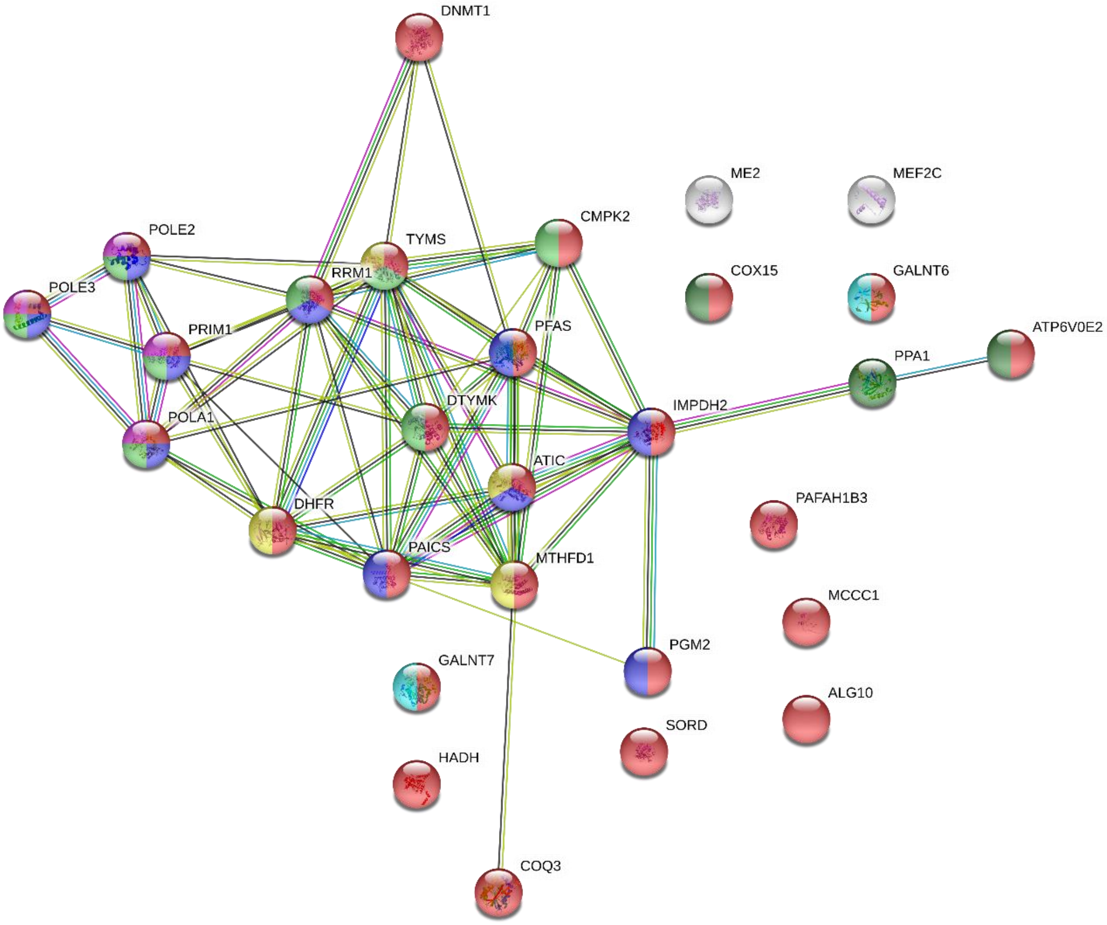

Figure A6: STRING scheme of 29 upregulated genes (rapidly progressing IPF vs. steady IPF) – Metabolic pathways and TGF-β pathway genes

There are 29 nodes (proteins) and 64 edges (protein-protein associations) in this network. Such an enrichment of edges indicates that the proteins are at least partially biologically connected as a group. Nodes are marked with seven distinct colours. Each one represents different KEGG defined pathway. Red colour marks 26 genes included in Metabolic pathways, purple colour marks ten genes included in Purine metabolism, light green colour marks eight proteins included in Pyrimidine metabolism, yellow colour marks four proteins included in One carbon pool by folate pathway, pink colour marks four proteins included in DNA replication, dark green colour marks three proteins included in Oxidative phosphorylation and light blue colour marks two proteins included in Mucin type O-Glycan biosynthesis. There are six distinct colours of edges. Black lines indicate interactions based on co-expression, yellow lines represent connections based on textminig, green lines show predicted interactions based on gene neighbourhoods, light blue lines represent known interactions from curated databases which are also experimentally determined (pink lines) and dark blue lines show predicted interactions based on gene co-occurrence.

APPENDIX J – KEGG schemes with DE genes

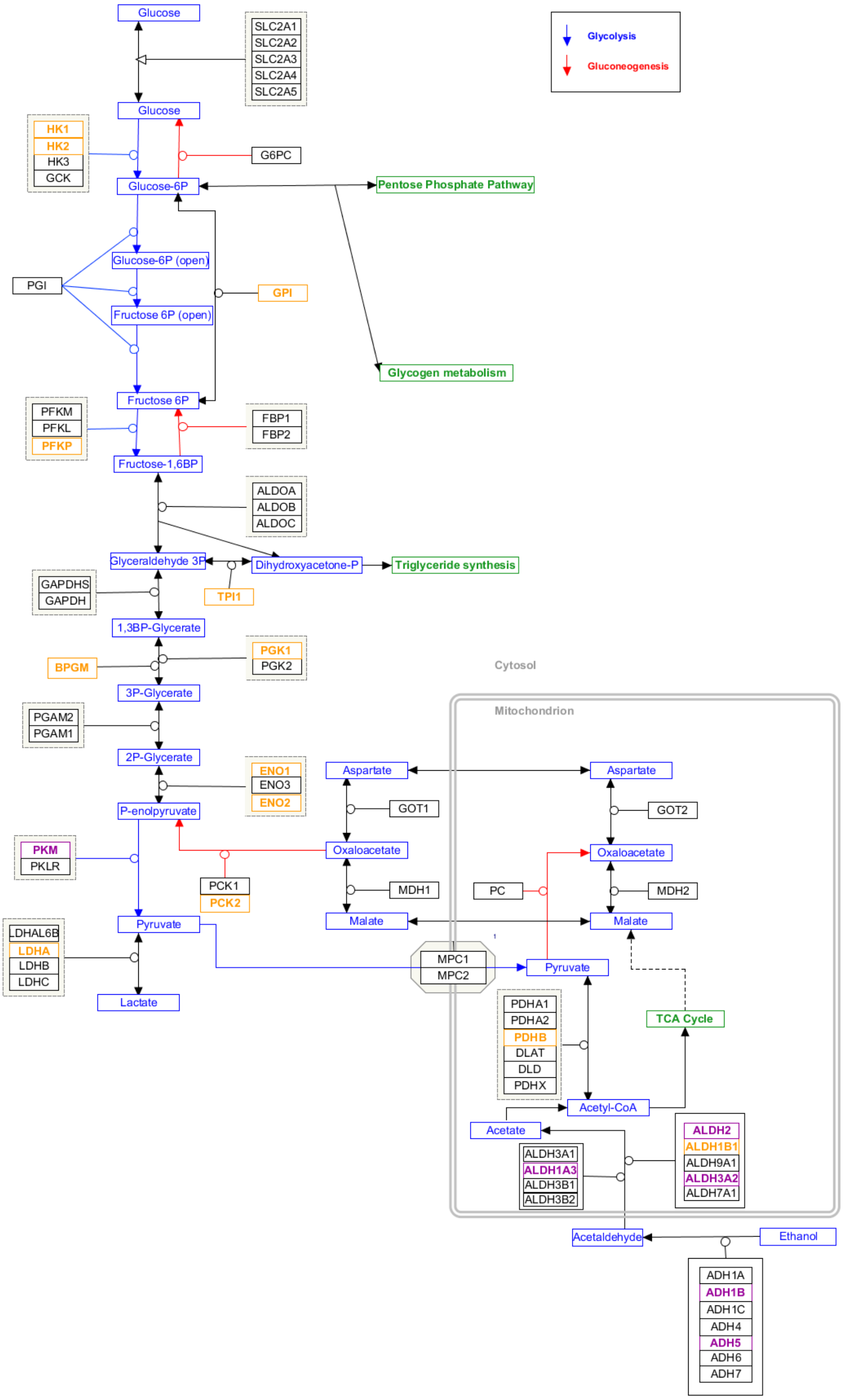

Figure A7: DE genes in Glycolysis/gluconeogenesis pathways
Orange coloured genes are upregulated in SSc-ILD, and purple coloured genes are downregulated in SSc-ILD.

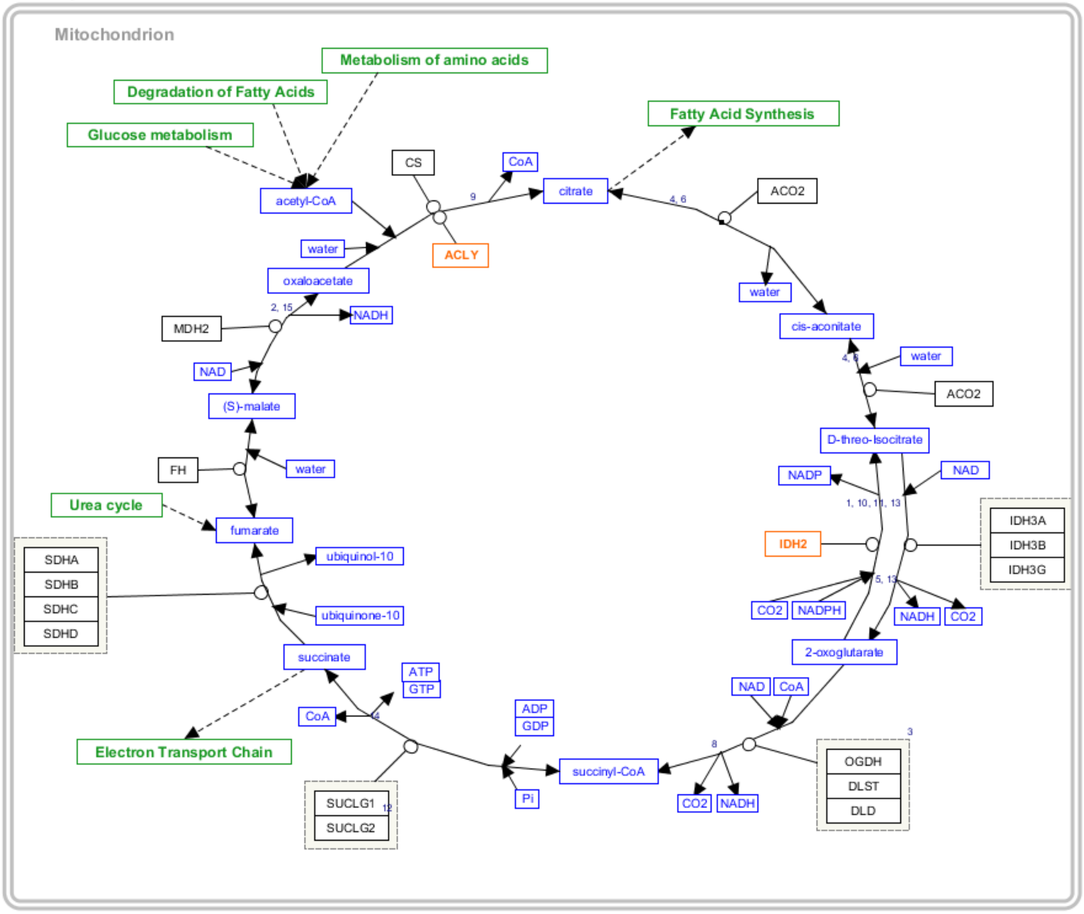

Figure A8: DE genes in TCA cycle pathway
Orange colour represents upregulated genes in SSc-ILD.

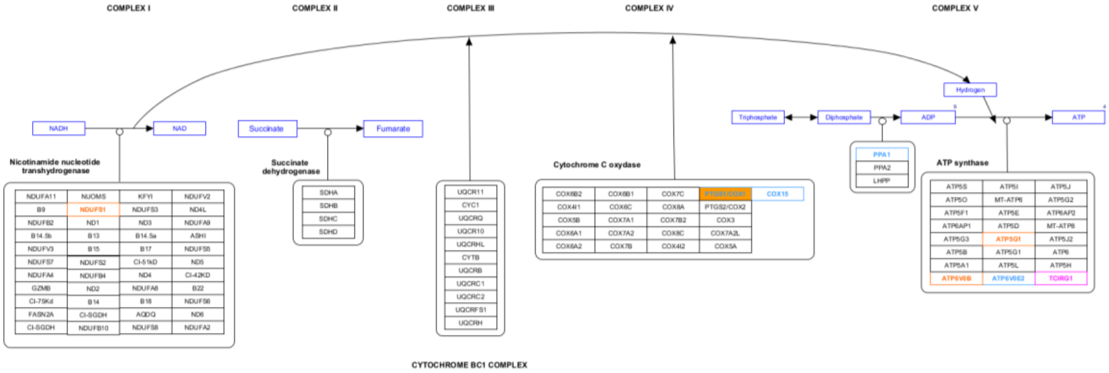

Figure A9: DE genes in OXPHOS pathway
Orange coloured genes are upregulated in SSc-ILD, light blue coloured genes are downregulated in rapidly progressing IPF and pink coloured gene is upregulated in rapidly progressing IPF

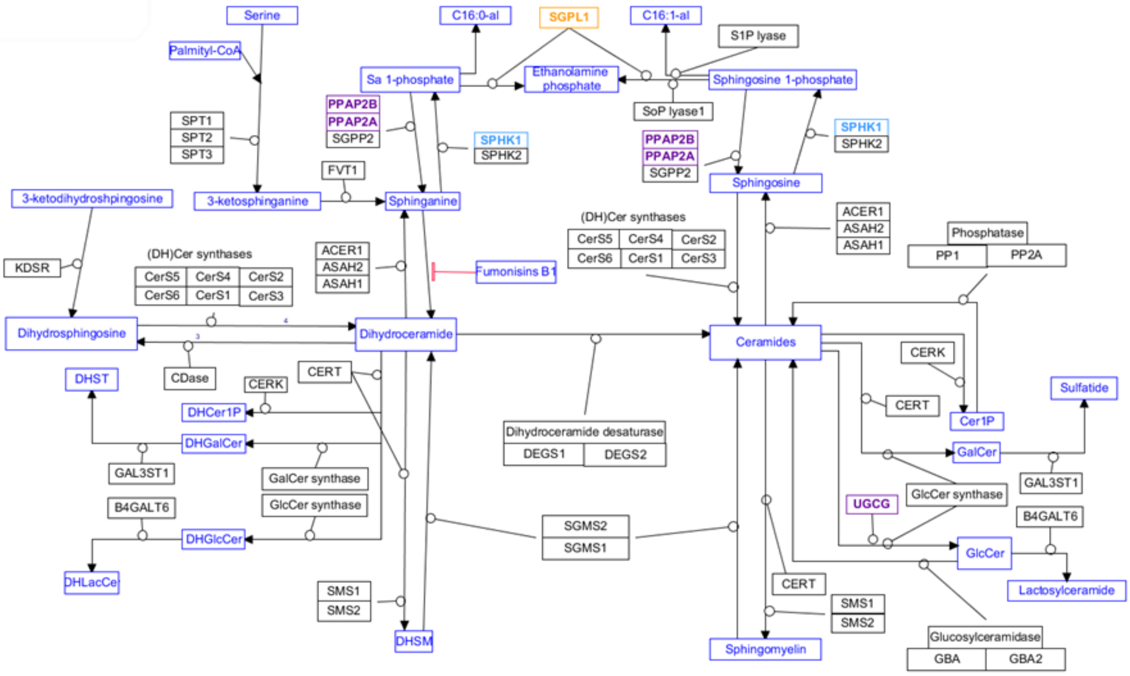

Figure A10: DE genes in Sphingolipid metabolism

Purple coloured genes are downregulated in SSc-ILD, orange coloured gene is upregulated in SSc-ILD and light blue coloured genes are downregulated in rapidly progressing IPF.

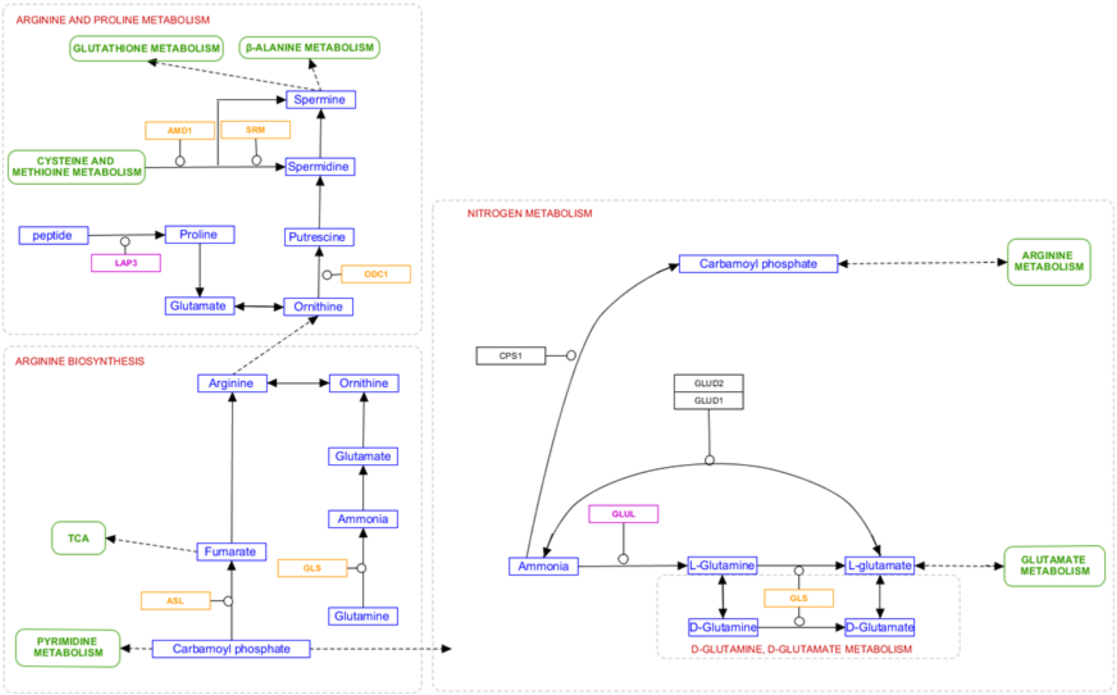

Figure A11: DE genes in Arginine and proline metabolism, Nitrogen metabolism and D-Glutamine, D-Glutamate metabolism

Purple coloured genes are downregulated in SSc-ILD and orange coloured genes are upregulated in SSc-ILD.

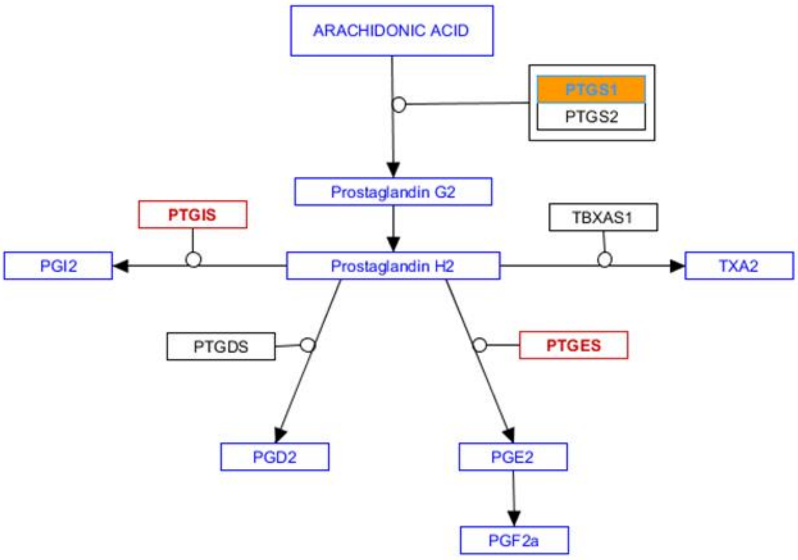

Figure A12: DE genes in Biosynthesis of eicosanoids pathway

Orange colour represents upregulation of a gene in SSc-ILD, light blue colour represents downregulation of a gene in rapidly progressing IPF and red colour represents downregulation of a gene in SSc-ILD.