UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

FACULDADE DE FARMÁCIA

GIULIANO NETTO FLORES CRUZ

GENOMIC ANALYSIS OF MACROPHAGE GENE SIGNATURES DURING IDIOPATHIC PULMONARY FIBROSIS DEVELOPMENT

PORTO ALEGRE, RS

2018

Giuliano Netto Flores Cruz

Genomic analysis of macrophage gene signatures during idiopathic pulmonary fibrosis development

Undergraduate monograph presented as
partial requirement for the achievement of
the Bachelor degree in Pharmacy.

Academic Advisor: Prof. Paulo J. Saraiva, Msc.
Co-advisor: Prof. Alexandre M. Fuentefria, PhD.
Co-advisor: Prof. Otavio J. Saraiva, Msc.

Porto Alegre, RS
2018

ACKNOWLEDGMENTS

I thank to my Family, Professors, Friends, and God. Hopefully this work will minimally honor the blissful experience of having you in my life.

Amo todos vocês.

ABSTRACT

Idiopathic Pulmonary Fibrosis (IPF) is a chronic, progressive, irreversible lung disease. After diagnosis, the interstitial condition commonly presents 3-5 years of life expectancy if untreated. Despite the limited capacity of recapitulating IPF, animal models have been useful for identifying related pathways relevant for drug discovery and diagnostic tools development. Using these techniques, several immune-related mechanisms have been implicated to IPF. For instance, subpopulations of macrophages and monocytes-derived cells are recognized as centrally active in pulmonary immunological processes. One of the most used technologies is high-throughput gene expression analysis, which has been available for almost two decades now. The “omics” revolution has presented major impacts on macrophage and pulmonary fibrosis research. The present study aims to investigate macrophage dynamics within the context of IPF at the transcriptomic level. Using publicly available gene-expression data, we applied modern data science approaches to (1) understand longitudinal profiles within IPF models; (2) investigate correlation between macrophage genomic dynamics and IPF development; and (3) apply longitudinal profiles uncovered through multivariate data analysis to the development of new sets of predictors able to classify IPF and control samples accordingly. Principal Component Analysis and Hierarchical Clustering showed that our pipeline was able to construct a complex set of biomarker candidates that together outperformed gene expression alone in separating treatment groups in an IPF animal model dataset. We further assessed the predictive performance of our candidates on publicly available gene expression data from IPF patients. Once again, the constructed biomarker candidates were significantly differentiated between IPF and control samples. The data presented in this work strongly suggest that longitudinal data analysis holds major unappreciated potentials for translational medicine research.

TABLE OF CONTENTS

1. INTRODUCTION

Idiopathic Pulmonary Fibrosis (IPF) is a chronic, progressive, irreversible lung disease. After diagnosis, the interstitial condition commonly presents 3-5 years of life expectancy if untreated (WOLTERS et al., 2018). Alveolar damage is characterized by dilatation of the bronchi, tissue remodeling and parenchymal fibrosis which seriously impair gas exchange. The first time the current name was used it referred to radiography tests suggestive of pulmonary fibrosis of unknown etiology, although the disease had been recognized over one hundred years before - named cirrhosis of the lung at the time (ROBBINS, 1948, WOLTERS et al. (2018)). Over three decades after the nomenclature was first published, in 1976 it gained widespread use after a review article (WOLTERS et al., 2018).

Currently, radiographic and histopathological patterns of usual interstitial pneumonia (UIP) with apparently no secondary causes form the basis for IPF definition (RICHELDI; COLLARD; JONES, 2017) and the disease presents sporadic and familial forms. From the edges of the lungs - base and periphery -, it spreads all over the pulmonary tissue causing dry cough, fatigue and dyspnea. The latter is nearly universal in the history of IPF patients and may progress over a period of years, while the first two may be absent in the early stages. Universal findings also include low diffusion capacity of the lung for carbon monoxide (DLCO) and bilateral Velcro-like crackles. The chest radiograph may present nonspecific changes or bilateral basal reticular abnormalities (LEDERER; MARTINEZ, 2018). The difficulty of the diagnosis is illustrated by the fact that clinicians often mistake the presence of dyspnea as indicator of heart failure or chronic obstructive pulmonary disease (COPD), failing to consider interstitial lung disease, which delays IPF detection significantly.

Although its cause is not clear, there is currently a recognized increase in IPF prevalence worldwide (WOLTERS et al., 2018). As a disease of aging, this may be due to population increasing phenomena, but also to recent improvements in disease recognition (LEDERER; MARTINEZ, 2018). Unusual before the 50 years of age, IPF prevalence almost doubles for every decade of life thereafter (WOLTERS et al., 2018). In North America and Europe, the incidence of IPF ranges from 3 to 9 cases per 100000 person-years, while South America and East Asia show less than 4 cases per 100000 person-years (HUTCHINSON et al., 2015). Assuming these numbers as conservative, a recent Brazilian study tried to calculate more precise estimates for the local reality. Using data from the 2010 Brazilian National Census and rates reported by previous studies, the authors suggested annual incidence of IPF between 6.841 and 9.997 cases per 100000 people, whereas the prevalence was estimated at a range of 13.9-18.3 cases per 100000 people (BADDINI-MARTINEZ; PEREIRA, 2015). In terms of mortality, another study estimated 1.2 deaths per 100000 people in 2010, although this is certainly an underestimation as the authors limited their analysis to data from the Information Technology Department of the Brazilian Unified Health Care System (DATASUS), which does not reflect private medicine practices (RUFINO et al., 2013). It is clear, however, that Brazil's data does not oppose the world tendency of IPF prevalence increase.

The number of disease cases has risen along with the number of identified risk factors. Regarding hospitalization, over three quarters of the cases are due to respiratory events, with acute exacerbations as the main cause (BROWN et al., 2015; SONG et al., 2011). In-hospital mortality rate has been reported to reach 50 %, with five-year survival from initial diagnosis lower than 20 % (SONG et al., 2011). In such a complex scenario, multidimensional indexes that include sex, age and physiological abnormalities may be useful to predict mortality (RICHELDI; COLLARD; JONES, 2017). Risk factors also include environmental exposures to dust and air pollution, smoking, chronic viral infections (e.g. Epstein–Barr virus, cytomegalovirus and Kaposi sarcoma-associated herpesvirus) and other comorbidities. Of note, one third of inherent individual risk of IPF have been attributed to genetic variants. Mutations in genes associated with telomere length - such as TERT, TERC, PARN, and RTEL1 - are related to higher higher risk of IPF. The same holds true for genes responsible for cell adhesion, integrity, and mechanotransduction (e.g. DSP, AKAP13, CTNNA, and DPP9) (LEDERER; MARTINEZ, 2018).

Mucin 5B overexpression in small-airway epithelial cells is a universal finding in patients with IPF, which has led to the hypothesis that impaired mucociliary clearance may be linked to changes in microbiome and innate immune responses that promote IPF (LEDERER; MARTINEZ, 2018; MOLYNEAUX et al., 2017). Although the gene expression pattern appears to be genotype-independent, single-nucleotide polymorphisms in the MUC5B (Mucin 5B) gene are notable risk factors for

IPF, even though - paradoxically - they may also predict slower disease progression (PELJTO et al., 2013). Host-microbiome interactions are also highlighted by the central role of macrophages in lung fibrosis development along with the finding that IPF risk is also increased by mutations in the TOLLIP gene, which encodes a protein associated with toll-like receptor family pathways (LEDERER; MARTINEZ, 2018). Lung of patients typically present higher bacterial loads, differences in microbiota composition and diversity, and even increases in potentially pathogenic bacteria (e.g. Staphylococcus spp. and Streptococcus spp.). Finally, epigenetic reprogramming has been associated with pathogenesis of IPF (RICHELDI; COLLARD; JONES, 2017). Upregulation of lung development genes, trans-regulatory methylation marks near transcription factors, and potentially pathogenic modifications in miRNAs have been pointed as complex reprogramming processes often associated with pro-fibrotic pathways.

Overall, it is well accepted that IPF arises from recurrent, subclinical epithelial injury, especially in genetically-predisposed individuals with accelerated epithelial aging (LEDERER; MARTINEZ, 2018). In such patients, the repetitive alveolar damage can induce pro-fibrotic epigenetic reprogramming, persistent cell senescence, production of pro-fibrotic molecules as well as activation of mesenchymal cells (RICHELDI; COLLARD; JONES, 2017). Recent studies, however, have been suggesting that the trigger for lung fibrosis might not be strictly related to exogenous aggression, which is yet to be further investigated (KULKARNI et al., 2016; NAIKAWADI et al., 2016). Still, altered migration, proliferation and mixed activation profiles of epithelial cells, especially alveolar epithelial type 2 cells, are a hallmark of IPF (RICHELDI; COLLARD; JONES, 2017). These and other pathological discoveries will be critical for early diagnosis and treatment development.

Currently, diagnosis of IPF is mainly focused on high-resolution computed tomographic (CT) imaging after identification of history and physical examination that are suggestive of interstitial lung disease (LEDERER; MARTINEZ, 2018). Once the suspicion is confirmed, further physical examination is performed to rule out related disorders such as chronic hypersensitivity pneumonitis and connective-tissue disease, as well as autoimmune conditions. Antinuclear antibodies, rheumatoid factor, antibodies against cyclic citrullinated peptides, Scl-70, SSA-Ro, SSB-La, U1-RNP, Jo-1 and other immunological tests comprise the serological examination routine, even though the presence of these biomarkers may not be enough to confirm

IPF absence (LEDERER; MARTINEZ, 2018). Often, lung biopsy is not performed in face of usual interstitial pneumonia (UIP) in high-resolution CT associated with suitable history, these being enough for IPF diagnosis. However, obvious CT patterns are not always present and invasive methodologies are needed. If histologic UIP is present, IPF diagnosis is confirmed. Otherwise, multidisciplinary discussions are encouraged as means of increasing diagnostics and prognosis assessment (LEDERER; MARTINEZ, 2018).

Finally, in terms of pharmacotherapy, IPF management has been through recent international standardization. In 2015 the American Thoracic Society, European Respiratory Society, Japanese Respiratory Society and Latin American Thoracic Association (ATS/ERS/JRS/ALAT) released international IPF therapy guidelines, updating a previous version from 2011 (RAGHU et al., 2015b). The guidelines strongly favored the recommendation for use of Pirfenidone and Nintedanib (RICHELDI; COLLARD; JONES, 2017). While the latter is a tyrosine kinase inhibitor taken twice daily, the former acts on various pathways, inhibiting TGF-β production and downstream signaling, among other effects (LEDERER; MARTINEZ, 2018). In addition to the need of three daily doses, Pirfenidone may cause anorexia, nausea and photosensivity, and may have its blood levels increased by CYP 1A2 inhibitors. Nintedanib causes diarrhea, risk of bleeding and arterial thrombosis, and a low risk of gastrointestinal perforation. Both treatments require liver-function monitoring and have similar efficacy profiles - reducing forced vital capacity decline rate by nearly half, clearly insufficient for stopping disease progression.

1.1. IPF IMMUNOLOGICAL BACKGROUND

Despite the limited capacity of recapitulating IPF, animal models have been useful for identifying related pathways relevant for drug discovery and diagnostic tools development. Using these techniques, several immune-related molecules have been implicated to IPF, including transforming growth factor beta (TGF-β), connective-tissue growth factor (CTGF), tumor necrosis factor alpha (TNF-α), fibroblast growth factor 2 (FGF2), platelet-derived growth factor (PDGF), several matrix metalloproteinases and chemokines (LEDERER; MARTINEZ, 2018; RICHELDI; COLLARD; JONES, 2017). The fibrotic process itself has its own singularities as well. A wide range of interleukins and other immunologically active mediators have been proposed as critical for fibrotic processes. Among others, these include IL-33, IL-17A, IL-25, TSLP, and IL-13 (Li2014; CAMELO et al., 2017; GURCZYNSKI; MOORE, 2017; HAMS; BERMINGHAM; FALLON, 2015). Regarding immune cell populations, macrophages have been shown to play central roles in pulmonary fibrosis and the elucidation of their biological dynamics is an active area of current research effort (KUROWSKA-STOLARSKA et al., 2009; LEE et al., 2018; MISHARIN et al., 2013, 2017; VENOSA et al., 2016; WYNN; BARRON, 2010).

Subpopulations of macrophages and monocytes-derived cells are recognized as centrally active in pulmonary immunological processes (BRAGA; AGUDELO; CAMARA, 2015; HUSSELL; BELL, 2014; MARTINEZ; GORDON, 2014; MISHARIN et al., 2013; SYRBU; THRALL; SMILOWITZ, 1996). Indeed, these cells can be source of pathophysiological information throughout disease progression for two main reasons. Firstly, they act as sentinels for foreign aggression, which assures detectable variance in animal models from the very beginning of the induced lesion and inflammation (MARTINEZ; GORDON, 2014). Secondly, as anti-inflammatory and pro-fibrotic mechanisms progress, the above-mentioned cells are also fundamentally involved in both the regulatory functions and the fibrosis promotion (CAMELO et al., 2017; LUZINA et al., 2015). Most recruited monocytes differentiate into macrophages upon tissue arrival. These cells respond differently across distinct phases of pulmonary immunological activity so that understanding their dynamics throughout the course of lung aggression is critical (WYNN; BARRON, 2010). Although not limited to a dichotomous model, M1- and M2-like cells typically promote and modulate these mechanisms, and their deep phenotypic profiling is crucial for the understanding of IPF and other fibrosis-related conditions.

1.2. MACROPHAGES AND FIBROSIS DEVELOPMENT

The definition of common macrophage subpopulations is currently under scientific scrutiny and revision (MARTINEZ; GORDON, 2014). Traditionally, macrophages were thought to be divisible into two groups, depending on the activation stimuli to which they are exposed (ABBAS; LICHTMAN; PILLAI, 2017). Classical activation is triggered by exposure to microbial toll-like receptor ligands such as lipopolysaccharide (LPS) and cytokines commonly released by T helper 1 cells (TH1), especially Interferon-gamma (IFN-γ). These signals enhance the antimicrobial and tumoricidal properties of what is then named M1 macrophage. M1 cells promote potentially harmful inflammation, but are crucial to fight, for instance, viral infections and cancer (ALFANO et al., 2013; GINHOUX et al., 2016; MALE et al., 2013). Alternative activation occurs when macrophages are exposed to cytokines characteristic of T helper 2 cells (TH2), notably IL-4 and IL-13, and the then called M2 macrophages inhibit inflammation and promote tissue repair and fibrosis (DELVES et al., 2017). The M1 vs M2 paradigm has been used worldwide and certainly contributed to the advancement of modern immunology. The M1 or M2 responses were referred to as analogous to the TH1 and TH2 responses, and the preferential differentiation of a group of cells in a given environment into one of the phenotypes is termed as "macrophage polarization" (GINHOUX et al., 2016).

However, as these cellular subtypes were first defined with controlled *in vitro* experiments, more complex phenotypic experimental characterizations started to give birth to the model questioning. Subdivisions of M2 macrophages arose from the observation of subtly distinct phenotypes, depending on the anti-inflammatory stimulus used. M2a phenotype is traditionally triggered by IL-4 and IL-13; M2b phenotype is induced by IL-10 exposure; and M2c cells are generated with a combination of immune complexes and LPS. Flow cytometry analyses now include several markers for each phenotype and their subdivisions, but these are often conflicting between studies (MISHARIN et al., 2013; TARIQUE et al., 2015; VENOSA et al., 2016). It is now clear that many homeostatic and pathological situations do not support M1 or M2 phenotypes dichotomy, and that in many cases these cells present high phenotypic plasticity (GINHOUX et al., 2016).

Macrophage activation is influenced by their ontogeny (i.e. if they derive from yolk sac, as the microglia, fetal liver monocytes, as lung macrophages and Kupffer cells, from both, as Langerhans cells, or if they are replaced with adult bone marrow monocytes, as in the gut, heart an dermal macrophages). Tissue-specific signals also drive macrophage activation both in homeostasis and disease and, importantly, the same stress signals with the same kinetics result in differentially programmed macrophages if these have been exposed to different microenvironments or previous stimuli. Taken together, these recent advances lead the understanding of macrophage biology to a multidimensional model of activation, with major microenvironment particularities and transcriptional programs significantly variable across human and mouse normal tissues and pathological conditions.

Faced with the need for standardization, a group of macrophage biology researchers suggested a reviewed nomenclature and experimental guidelines for subpopulation studies (MURRAY et al., 2014). The proposal was based on three principles: "source of macrophages, definition of the activators, and a consensus collection of markers to describe macrophage activation". As it remains a challenge to define macrophage phenotypes that describe accurately the cellular function across time and environmental conditions, the authors proposed nomenclature designation based on the stimuli to which the cells are exposed, for *in vitro* studies. For instance, traditional M1 macrophages would now be named M(LPS) or M(IFN-γ) cells, while M2 would be further divided into M(IL-4), M(IL-4+IL-13), M(IL-10), and others. For *in vivo* studies, the cell names would explicitly declare their multiple markers rather then forcing a fit into M1 or M2 spectra. Many confusions are avoided by this approach. A simple example, the expression of Arginase-1 has been used to describe M2 - or M(IL-4) - macrophages, while it is well known that the enzyme is also expressed by M1 spectrum cells as well as resident macrophages (MURRAY et al., 2014). Additionally, the authors suggest terms to be avoided as these may further confuse classification: "regulatory macrophages" has been used to refer to M2-like cells, despite the fact that all macrophages show regulatory functionalities at some point and even within the M2 spectrum the regulatory functions are considerably heterogeneous. Although their nomenclature standards have been extensively encouraged, macrophages are still presented as M1-like or M2-like in many occasions, especially when accurate classification is not achieved (BECKER et al., 2015; ITALIANI; BORASCHI, 2014; MALAVIYA et al., 2016; MURRAY, 2017; VENOSA et al., 2016; WERMUTH; JIMENEZ, 2015). In such a controversial scenario, novel approaches are currently needed to further improve macrophage classification.

1.3. NOVEL APPROACHES IN MACROPHAGE BIOLOGY RESEARCH

One of the promising techniques applied to macrophage study is high-throughput gene expression analysis. High-throughput technologies have been available for almost two decades now, and these have deeply challenged our current understanding of macrophage biology (KIDD et al., 2014; PEVSNER, 2015; STABLES et al., 2011). A remarkable work from Xue and colleagues (2014) is an outstanding example of these advances. Using diverse *in vitro* stimuli, the authors performed single-cell RNA sequencing on almost 300 macrophages (XUE et al., 2014). Weighted Correlation Network Analysis identified 49 stimulus-specific gene modules that could be used as gene sets for enrichment assessment on data from patients and animal models. This approach deeply extended M1- versus M2-paradigm towards a “spectrum model of human macrophage activation” and was able to identify a refined, activation-independent core signature for human and murine macrophages.

In fact, module analysis has been widely used to investigate transcriptional profiles of immunological processes. First designed using k-nearest neighbors clustering algorithm for microarray data, it identifies groups of genes that are coordinately expressed in a given dataset while supporting systems-scale analysis for translational research (CHAUSSABEL et al., 2008). Currently, this type of procedure is performed by widely used software which implement diverse set of machine learning algorithms (LANGFELDER; HORVATH, 2008). Additionally, it has been adapted to RNA sequencing and single-cell RNA sequencing data. The wide range of algorithms currently used has been comprehensively reviewed elsewhere (SAELENS; CANNOODT; SAEYS, 2018).

Apart from transcriptome studies, proteomic data analysis has yielded promising results as well (COURT et al., 2017). Despite the difficulty of validating direct relationships between gene and protein expression patterns, the one agreement is that M1 versus M2 model is not enough to explain macrophage polarization repertoire (KAMAL et al., 2018; MARTINEZ; GORDON, 2014; TARASOVA et al., 2016). Nevertheless, as Next Generation Sequencing became broadly available at relatively low costs, proteome research is still to match high-throughput transcriptomic techniques in terms of scaling and data generation capacity. Overall, multi-omics approaches start to gradually emerge as the amount of data currently being generated by far exceeds the data analysis resources available for the scientific community.

1.4. BIOINFORMATICS AND FUNCTIONAL GENOMICS

According to Jonathan Pevsner (2015), “functional genomics is the genome-wide study of the function of the DNA (including genes and nongenic elements) as well as the nucleic acid and protein products encoded by DNA” (PEVSNER, 2015). The author also states that functional genomics relies primarily on the use of high-throughput technologies, such as Next Generation Sequencing (NGS) and microarray. More traditional techniques such as real-time polymerase chain reaction are used as means of validation. Finally, Pevsner emphasizes that functional genomics plays a fundamental role in solving one of the ultimate problems in modern biology: understanding the relationship between genotype and phenotype (PEVSNER, 2015).

It is not surprising that these approaches have been broadly applied to macrophage biology studies (FONSECA; SEIDMAN; GLASS, 2017). The massive amount of data generated, though, represents a challenge for researchers. This scenario has led to the advance of the bioinformatics field, which stands at the interface between molecular biology and computer science. Briefly, bioinformatics seeks the analysis of molecular sequences - which can derive from DNA, RNA, or proteins - to answer a broad range of biological questions. Genomics is dedicated to the analysis of DNA sequences of organisms - the genomes -, while transcriptomics analyzes the transcriptome and proteomics, the proteome - and so forth. Going further, Functional Genomics takes advantage of genome-wide assays to understand gene, transcript, and protein functions. Although there are obvious overlaps among the terms, a first perspective of the big picture of bioinformatics suggested by Pevsner is the cell itself: the central dogma of molecular biology states that the relationships between DNA, RNA and proteins ultimately generate cellular phenotype. In genomics, the central dogma is translated into the relationships between the genome, transcriptome and proteome. This approach greatly enhances the complexity of cellular phenotype modeling, and hence computational methods are needed (PEVSNER, 2015). Still, strict definition of these and related terms is often controversial and beyond the scope of this work.

1.4.1. Bioinformatics Development

The development of bioinformatics software is probably one of the most exciting areas of recent scientific advances. A popular approach is web-development for scientific computing. Open, web-available tools such as Basic Local Alignment Search Tool (BLAST) at National Center for Biotechnology Information (NCBI) make it possible for researchers with no programming experience to perform complex bioinformatics analyses (QUEREDA et al., 2016). The paper that introduced BLAST in 1990 counts over 74000 citations as of November, 2018 - similar to the citation numbers from Cesar Victora, one of the Brazilian scientists with the highest citation counts at Google Scholar currently (VICTORA, 2018). However, programming abilities are a restrictive must for those seeking to further customize their analyses or to develop their own algorithms. Although several academic majors take advantage of bioinformatics development, from biology to health sciences, programming skills are often neglected. In Brazil, few are the undergraduate programs that include bioinformatics as a discipline and even fewer are the specialized and well developed bioinformatics graduate programs. This reality is now trending to change as computational biology applications gain major highlights in media and academic routine.

Regarding programming languages, one is particularly outstanding in the field of Functional Genomics. The R programming language is both a language and an environment for statistical computing and graphics, similar to the S language previously developed by John Chambers and colleagues at Bell Laboratories (TEAM, 2018). Like S, in R one can program their own functions and extend base functionality through the use of packages - which are just R code with certain functionalities validated and encapsulated. Also, it is possible to link C and C++ code to these packages so that computationally-intensive tasks can be performed. The idea of programming “environment” comes from the production a coherent and well-planned framework in which statistical computing can be run and developed. This can contrast with other programming languages such as python, which can be applied to statistics and data analysis in spite of its broader range of applications. Python is another very popular data science language with extensive machine learning algorithm development, although its reach in the field of functional genomics may not be as extensive as in the case of R.

The success of R within bioinformatics field has a particular reason. In 2001, a group of researchers, bioinformaticians, statisticians, and data scientists released Bioconductor, an open-source, open development software project dedicated to the analysis of high-throughput genomic data (GENTLEMAN et al., 2004). To date, the package repository contains over 1600 software packages, which undergo continuous automated testing in addition to formal initial review (HUBER et al., 2015). Bioconductor also supports the rapid development of standard workflows combining highly complex data structures and statistical inference tools, regression, network analysis, machine learning and data visualization, which is especially important for reproducible research. It is deeply documented at three levels: whole workflows combining multiple tools, packages vignettes providing the narrative for the package usage with code and data analysis examples, and manual pages that serve as reference for detailed descriptions of all inputs and outputs for the packages functions. With enough experience, users can become developers and share their work with others through the repository. The choice of R language is justified by its high-level statistical and graphical utilities, which yields rapid prototyping creativity, flexibility and reproducibility unmatched by web-based tools software and general-purpose languages (HUBER et al., 2015). The whole Bioconductor structure and development culture are focused on reproducible research and data analysis, which translate into good practices for documentation and software development that well enforced by the users and developers community.

1.4.2. Microarray data analysis using Bioconductor

DNA Microarray is a genome-wide gene expression measurement technique that emerged by 2000, although it was first developed in the previous decade at Stanford University and National Institute of Health (NIH) (PEVSNER, 2015). It has been one of the most widely used tool for genomic studies worldwide (SINHA, 2014). On the surface of a solid support, several nanograms of DNA are immobilized in a grid-like array. The RNA extracted from biological samples is usually converted to complementary DNA (cDNA) - or cRNA, depending on the platform -, labeled with fluorescence, and selectively hybridized to the array. Each transcript should have a corresponding nucleic acid molecule to which hybridize, although often more than one probe maps to the same gene.

Either for technical reasons, or because a particular gene may have more undergo alternative splicing and thus different expression values are expected, probesets are particularly common on the Affymetrix platform (LIU et al., 2003). Once the microarray is washed, image analysis quantifies the fluorescence signals, and the spot intensities are assumed to correlate with the initial quantity of sample mRNA. The amount of starting material varies across technologies, but for many cases about 1-3 µg (micrograms) of total RNA is needed and the yielded hybridization material usually consists of 5 ng of cDNA.

Data analysis seeks the identification of differentially expressed genes and broad patterns of gene expression (PEVSNER, 2015). The Minimum Information About a Microarray Experiment (MIAME) provides good practices for experiment description, including experimental and microarray design, sample preparation, hybridization procedures, image analysis, and normalization controls. Microarray data is public available mainly through ArrayExpress and Gene Expression Omnibus from the European Bioinformatics Institute (EBI) and NCBI, respectively. Following MIAME is a requirement for using these databases to share your own data. Once microarray data is acquired, it must be properly normalized, undergo inferential statistics (*e.g.* t-tests, analysis of variance), exploratory analysis (e.g. unsupervised learning as clustering, dimensionality reduction), and classification (e.g. supervised analyses, support vector machines). These procedures ultimately lead to biological confirmation, which may be performed by non-high-throughput technologies such a RT-PCR. As all steps require complex calculations and extensive computing, many software options are available at Bioconductor, including platform-specific workflows and high-dimensional statistics tools.

Microarray data distribution is often non-parametric and thus data normalization is essential for sample and experiment comparisons (QUACKENBUSH, 2002). This is because of differences in the labeling efficiency, the amount of starting material, cDNA quality, signal detection, and so forth. Many techniques have been developed to solve this issue. Variance Stabilization and Normalization (VSN) assumes the variance for a specific probe mainly depends on its mean expression level and uses a linear transformation procedure to keep variance approximately constant (HUBER et al., 2002). Such a technique is broadly applied and is implemented in R through the VSN package - available from Bioconductor.

In 2003, Rafael Irizarry introduced the Robust Multiarray Analysis as a method of background correction, quantile normalization, and probeset summarization of probe intensities from Affymetrix platform raw data (IRIZARRY et al., 2003). As a non-parametric approach, quantile normalization makes no assumptions on the expression distributions. For each array, each probe expression measurement is assigned to a quantile. Normalization results from converting original probeset values to their corresponding quantile values. Using a convolution model, RMA is able to distinguish true probeset signal from noise. A improved version, GCRMA, increases RMA's accuracy by adjusting nonspecific hybridization using sequence information (PEVSNER, 2015). After comparing over 30 algorithms for Affymetrix microarray data, Irizarry and colleagues demonstrated the leading capacity of RMA and GCRMA procedures (IRIZARRY; WU; JAFFEE, 2006). Both methods can be easily applied through the affy package (GAUTIER et al., 2004).

Another issue with high-throughput data analysis is the amount of statistical tests. When measuring differential gene expression, one may perform over 20000 t-tests or Mann-Whitney and Wilcoxon tests, for instance - one for each probeset or gene. In this scenario, with a p-value threshold of 0.05, one can expect around 1000 false-positive rejections of the null hypothesis - an unacceptably high rate of type I error. In order to control for these, one must consider correction of p-values.

The Bonferroni procedure is used to control the Family Wide Error Rate, which can be defined as the probability of making at least one type I error among all tests (IRIZARRY; LOVE, 2015). It is considered a rather conservative correction as it sets a new significance cutoff by dividing our previous one - 0.05 - by the number of statistical tests performed. Thus, one must not expect high statistical power with a resulting . A more common approach is to control the False Discovery Rate (FDR), which is the proportion of false calls among one's positive results - amount of errors over the number of rejections of the null. In gene expression experiments, means that 5% of transcripts that were called significant are actually not differentially expressed (PEVSNER, 2015). For a dataset with 20000 genes and 100 significantly induced or repressed genes, such an FDR value would yield only 5 type I errors. The Benjamini-Hochberg procedure, easily applied in R, ranks p-values and assures an FDR below a given value of the analyst choice - typically 0.05 (IRIZARRY; LOVE, 2015).

A final topic on microarray data analysis deserves consideration - regarding the detection of differentially expressed genes. In R, this task is commonly performed using linear models. The most common Bioconductor package for this purpose is limma (RITCHIE et al., 2015). It computes ordinary t-statistics for linear model fits to all genes and then uses Bayesian modeling to moderate residual variance. As limma has been available for almost two decades now, novel packages became publicly available trying to extend and improve limma capabilities. Alternative approaches emerged for time-course gene expression analysis as it is particularly complex. MaSigPro is another R package which uses two regression steps for this specific scenario. First, it fits a global regression model - typically polynomial - and, secondly, it performs step-wise regression to observe group differences and statistically significant longitudinal profiles of gene expression (CONESA et al., 2006). Finally, longitudinal gene set analysis software has been developed by Hejblum and colleagues (HEJBLUM; SKINNER; THIÉBAUT, 2015). The Time-Course Get Set Analysis (TcGSA) package, available from the Comprehensive R Archive Network (CRAN), extends limma and MaSigPro techniques through random effects modeling with maximum likelihood estimates. It is capable of handling unbalanced repeated measures of gene expression and takes into account potential heterogeneity of expression profile within gene sets. The identification of differences in longitudinal expression patterns across factors of interest is thereby made possible.

1.4.3. Microarray data quality assessment

Quality assessment for gene expression data must be computed for all arrays after preprocessing. The preprocessing is platform-specific, but generally includes background subtraction, between array intensity adjustment (normalization), probeset summarisation and log2 transformation (PEVSNER, 2015). Next, exploratory data analysis is performed in order to find any outliers that were not successfully handled by preprocessing – see Figure 3. Outliers are tipically detected using the R package arrayQualityMetrics, which computes several different measures that reflect biases in the data, normalization failures and noise in particular arrays (KAUFFMANN; GENTLEMAN; HUBER, 2009). The package generates an interactive HTML report with deeper descriptions for each measurement applied to all arrays, although dataset-specific reasoning is advisable as each experiment has its own design particularities (SINHA, 2014). An example of “good-quality” microarray data is presented in Figure 1. Density plots (left panel) for all arrays are overlapped, and the consistent data distribution is also made clear by box plots (right panel). Simulation from (SINHA, 2014).

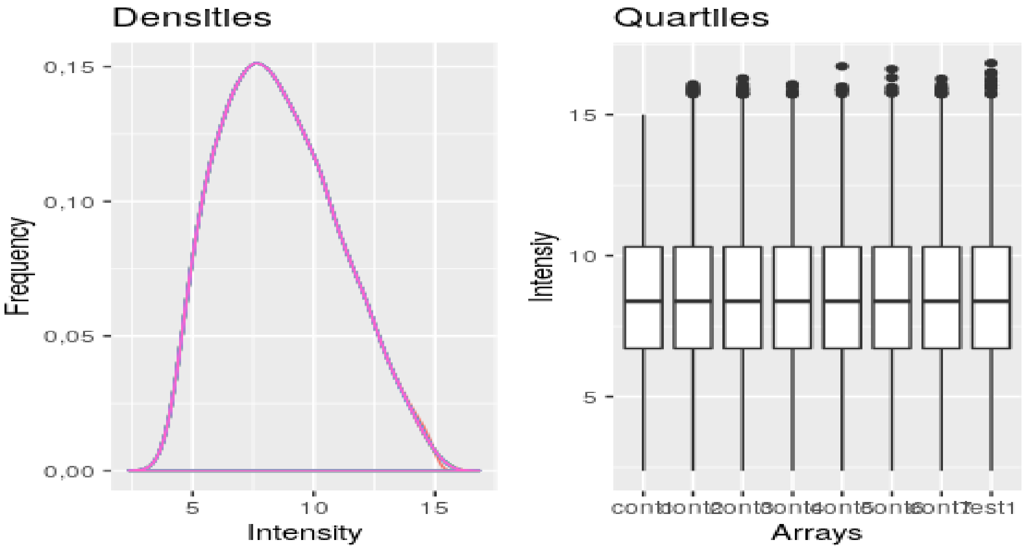

Figure 1. Example of successfully normalized microarray data.

1.4.4. RNA sequencing data analysis

RNA sequencing (RNA-seq) is the implementation of Next Generation Sequencing (NGS) applied to RNA expression analysis (PEVSNER, 2015). Initially, isolation of fragmented messenger RNA and acquisition of cDNA are performed. The experimental procedures vary depending on the platform and on the experiment objectives (i.e. qualitative or quantitative), but they often involve target enrichment, which consists of removal of ribosomal RNA and selecting 3'-end transcripts with long poly-A tails. Once the double-stranded cDNA library is prepared along with platform-specific adaptor sequences, the DNA can be amplified for subsequent sequencing. Many platforms are available, and they vary largely in terms of sequencing chemistry, base-call quality, read length, and many strengths and weakness that need to assessed based on the experiment objectives (METZKER, 2010). Of note, read lengths for RNA-seq experiments are often around 50 base pairs (single-end), although novel transcriptome assembly and annotation projects may benefit from paired-end sequencing with larger reads (CHHANGAWALA et al., 2015).

Bioconductor offers many pipelines for RNA-seq data, which differs from microarray mainly because the common linear models - based on Gaussian distribution - fail to accurately describe data structure. This is due the fact that NGS technologies yield data in the form of read counts (i.e positive integers, as opposed to fully continuous variables) and these tend to better fit into Poisson or negative binomial distributions (ANDERS; HUBER, 2010). Recently, Costa-Silva and colleagues assessed several available software tools for RNA-seq analysis (COSTA-SILVA; DOMINGUES; LOPES, 2017). Using quantitative RT-PCR as reference, the authors concluded that read mapping, a crucial step in data preprocessing, is satisfactorily performed by all software evaluated. Regarding modeling for statistical detection of differentially expressed genes, the best-performing options were limma+voom (LAW et al., 2014), NOIseq (TARAZONA et al., 2015), and DESeq2 (LOVE; HUBER; ANDERS, 2014).

Limma, which stands for Linear Models for Microarray Data, was not originally developed for RNA-seq data. However, the so-called voom method is able to generate precision weights for each observation by estimating mean-variance relationships of log-counts. When entering these into Bayesian modeling pipeline, microarray-derived analytical tools become as accurate as count-based analysis methods - most common strategy for RNA-seq (LAW et al., 2014). For this reason, one may perform both microarray and RNA-seq data analysis using the same Bioconductor-availabe limma package.

Another Bioconductor package, NOISeq, combines non-parametric methodology with empirical Bayes modeling to build its NOISeqBIO pipeline (TARAZONA et al., 2015). It estimates a statistic Z whose distribution is a mixture of those from (1) invariant genes, and (2) genes whose expression changes between conditions. Given a Z-score for a particular gene, the probability of differential expression can then be calculated using Bayes Rule.

Finally, DESeq2 extends its previous version (DESeq) by first fitting a generalized linear model for each gene in an expression matrix - assuming negative binomial distribution (or gamma-Poisson distribution) (LOVE; HUBER; ANDERS, 2014). The mean parameter is correcter by a scaled normalization factor, which accounts for sources of technical biases, including GC content differences, gene length, and sequencing depth between samples. The model uses a logarithmic link so that, in the simplest instance of control versus treatment experiment, it returns coefficients that indicate the gene’s overall expression strength plus its fold change between the conditions as binary logarithm (log2 Fold Change). Next, DESeq2 uses shared variance information across genes while assuming similar dispersion for genes with similar average expression strength. It shrinks gene-wise dispersions towards a predicted value based on expression strength similarities - a process that is weighted using Bayes approach. A similar shrinkage method is used to correct log2 fold changes thereby removing overestimation of expression changes for low read counts. According to the work from Costa-Silva and colleagues, DESeq2 pipeline yields 93% of specificity and 84% of true positive rate, thus being the top-performing method among those included in the study (COSTA-SILVA; DOMINGUES; LOPES, 2017).

1.5. GENOMIC CHARACTERIZATION OF MACROPHAGES

1.5.1. Macrophage Genomic Integrative Analysis

Genomic integrative analysis is a computationally-expensive and rather complex task. First, one must integrate different technologies (e.g. microarray and RNA sequencing) from several distinct platforms (e.g. Affymetrix, Illumina, Agilent) (WALSH et al., 2015). Second, although greater number of samples yields higher statistical power, potential confounding factors must be taken into account. When it comes to lung injury, a wide range of animal models and human conditions have already been tested, and their respective datasets should be treated with care. For instance, the widely used bleomycin-induced IPF model shows enrichment of traditionally M1-associated genes at very early stages (BAUER et al., 2015). Fungal infection models, on the other hand, show divergent genomic markers with potentially protective roles associated with genes from the M2 spectrum (BHATIA et al., 2011; MARGALIT; KAVANAGH, 2015). Both cases, though, may lead to pulmonary fibrosis through macrophage activity (GIESECK; WILSON; WYNN, 2017; IWASAKI; FOXMAN; MOLONY, 2016; WYNN; VANNELLA, 2016).

Other challenges include the adequacy of sample sizes, pre-processing techniques, statistical analysis, modeling validation, experimental design as well as the lack of a comprehensive framework for the execution of genomic meta-analyses (RAMASAMY et al., 2008). Of note, the term “meta-analysis” refers to cases when the researcher analyzes each dataset separately and draws conclusions based on the combination of final statistical results, whereas “cross-platform normalization” is used to describe the integration of raw data (“merging”) from multiple sources for combined downstream analysis (WALSH et al., 2015). Here, we use “integrative analysis” to denote both terms interchangeably as not all datasets analyzed are suitable for merging.

The applicability of integrative analysis for elucidating reproducible macrophage dynamics and even predicting clinical outcome based on the enrichment of their gene signatures has been previously tested (BECKER et al., 2015). Using data from human-derived macrophages challenged with two sets of in-vitro activation stimuli, namely “classical” (IFN-γ + LPS; TNF-α) and “alternative” (IL-4; IL-13), the authors were able to establish prognostic values in diverse clinical settings such as viral infections and asthma. Noteworthy, however, is the fact that gene signatures were still relatively limited by the M1 versus M2 paradigm, which hinders interpretation at the cellular and molecular levels. After all, how to understand the heterogeneity within such microenvironments and, furthermore, how to address similar macrophage subsets that are constantly overlooked (or that are yet to be described)? How comprehensive should an integrated analysis be to assure robustness of detected gene expression patterns? The answers to these questions may eventually lead to better pharmacology development and health care regarding many life-threatening, macrophage-related diseases.

Recently, an elegant work integrated several datasets from human biopsies as well as data from wide range of mouse strains within the context of LPS exposure (BUSCHER et al., 2017). Surprising was not the high level of gene expression variability across strains, but the ability to nevertheless infer the degree of polarization of macrophages in transcriptome samples. To do so, the authors looked at the expression levels of IL-12b and arginase-1, known as M1- and M2-markers, respectively. After correction by population expression mean, the quotient between the two molecules’ RMA (robust multi-array average: quantile normalized, background-corrected, log2 transformed intensities) represented what was named Polarization Factor Ratio (PFR). Next, the authors identified gene sets that were highly correlated with the PFR measurements. Those gene sets could then be used to describe the activation state of tumor-associated macrophages in cancer biopsy samples and even predict patient survival.

Buscher’s paper (2017) is an example of successful integrative analysis applied to the molecular study of macrophage biology. When it comes to IPF, their findings are further supported by protein-level assessment approaches as the behavior of immune cells in such conditions has been extensively studied

(MISHARIN et al., 2013; Mittar2011; LANDI et al., 2014; YU et al., 2016). Venosa and colleagues (2016) characterized the macrophage subpopulations in BAL fluid from Wistar rats exposed to nitrogen mustard (NM) – Figure 2 (VENOSA et al., 2016). In their study, infiltrating M1-like macrophages rapidly increased until three days after the treatment, which correlated with the upregulation of proinflammatory M1 genes and tissue injury. The infiltrating M1 cells started being replaced after the third day, and an accumulation of M2-like macrophages was seen by the 7th day after NM exposure. A persistent increase of M2-like cells was observed until the 28th day after NM, and that response was correlated with M2 genes upregulation and fibrosis development. Figure 2 shows the time-course profile of such cells.

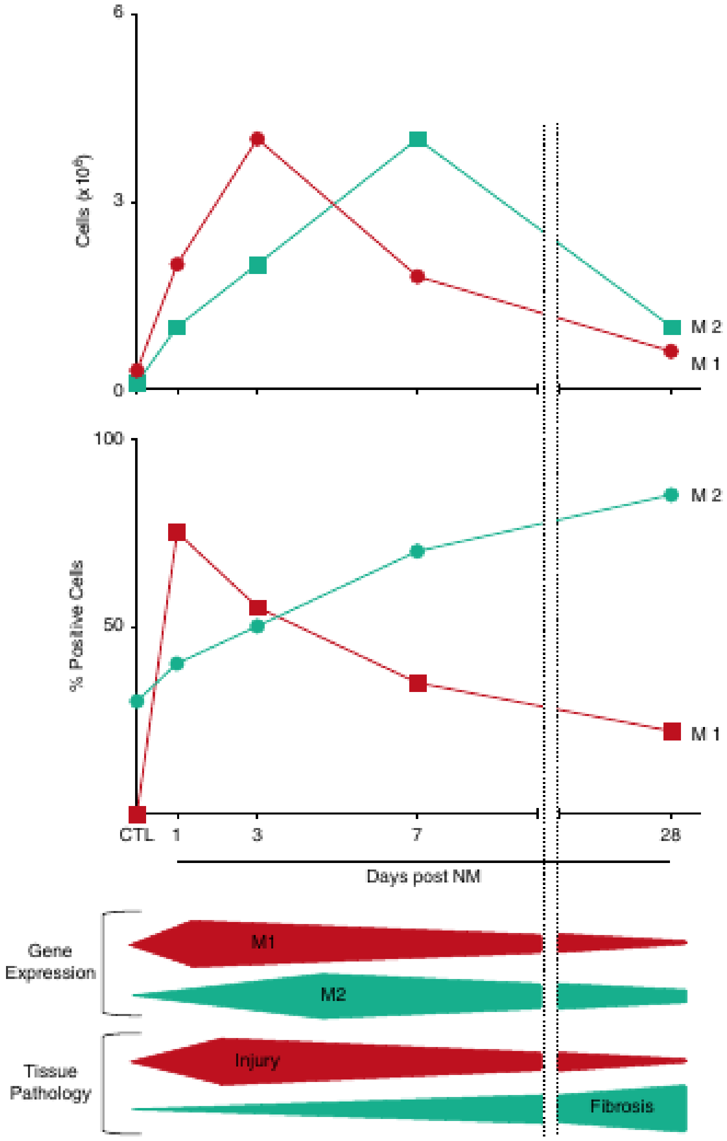

Figure 2. Macrophage dynamics in IPF model. Adapted from: (VENOSA et al., 2016).

Many other studies confirm Venosa's data (BAUER et al., 2015; BRAGA; AGUDELO; CAMARA, 2015; GIESECK; WILSON; WYNN, 2017; HAMS; BERMINGHAM; FALLON, 2015; KOLAHIAN et al., 2016; LEE et al., 2018; LUZINA et al., 2015; MALAVIYA et al., 2016; NIE et al., 2017; PENG et al., 2013; WILLIAMSON; SADOFSKY; HART, 2014; WYNN; VANNELLA, 2016). However, one may note the repetitive assumptions over the M1 versus M2 paradigm for most of the past works. High-throughput technologies combined with single-cell techniques are constantly being applied to yield more precise and informative data. Such a promising methodology has already revealed molecular heterogeneity much greater than previously predicted - for both innate and adaptive immune responses (CHEVRIER et al., 2017; LU et al., 2015; NEU et al., 2017). As an example, mature T helper 17 (Th17) cells have been demonstrated to develop a wide range of transcription programs, which opposed previous conceptions of high gene expression similarity among antigen-specific T cells (HAN et al., 2014).

The ability to understand the development of these transcription diversities within populations once thought as homogeneous is one of the current challenges in biology and health research. Furthermore, we are still to meet comprehensive and reproducible proteomic characterization of macrophage subpopulations, although relevant advances have been made (BECKER et al., 2012; CHEVRIER et al., 2017; COURT et al., 2017; TARASOVA, 2016). Multi-omic integrative characterizations will therefore build a stronger and more robust body of knowledge to drive macrophage-based therapy and diagnostics (BAKKER et al., 2018).

Here, we seek to determine a vast set of protein-coding genomic signatures that allows the investigation of possibly unacknowledged macrophage activation patterns across multiple datasets of pulmonary fibrosis models and IPF patients. In order to do so, we take advantage of single-cell sequencing data from human macrophages that were artificially stimulated with common and uncommon sets of signaling molecules so that we reach greater depth in our cross-platform characterization of macrophage dynamics (XUE et al., 2014).

1.5.2. Macrophage Gene Signatures

Macrophages have been demonstrated to develop highly complex activation profiles in a diverse set of microenvironments (GINHOUX et al., 2016). As previously discussed, these cells are key players in conditions as IPF (BAUER et al., 2015;

VENOSA et al., 2016; WYNN, 2011). Although many genetic markers are known to play important roles within the macrophage biology context, defining a robust set of gene signatures for the currently known phenotypic subsets remains a challenging task (MARTINEZ; GORDON, 2014). Recent cytometric and genomic approaches have revealed major limitations regarding the classic M1- versus M2-polarization model, which is no longer suitable to explain the biological dynamics of macrophage response (MARTINEZ; GORDON, 2014).

In the pursuit of standardization towards reproducible research, back in 2014 a group of specialists suggested nomenclatures and experimental guidelines for the macrophage activation profiles well-established by then (MURRAY et al., 2014). However, the recent abundance of genomic data has challenged the classical protein-level techniques used to sort macrophage subsets and therefore novel classifications emerged. In the same year that Murray's paper was published, a multi-center work attributed a much higher heterogeneity to macrophages through machine learning algorithms applied to single-cell transcriptome analysis (XUE et al., 2014).

Assessing the transcriptomes from almost 300 *in vitro* stimulated human macrophages, Xue and colleagues used weighted gene co-expression network analysis (WGCNA) to identify 49 co-expression modules, each of which ranging from less than 30 to over 800 distinct genes of size (XUE et al., 2014). Based on Pearson correlation, WGCNA defines gene clusters, known as transcriptional modules, which present specific co-expression patterns across each treatment condition (LANGFELDER; HORVATH, 2008). As an example, these modules can then be used to visualize the comprehensiveness of the M1 versus M2 model. As noted by the authors, stimuli not M1- or M2-associated showed prominent patterns consistent with a rather dynamic spectrum model of cell activation.

In order to achieve greater depth of macrophage phenotypes characterization, in this study the 49 transcriptional modules produced by Xue and colleagues were used as relevant gene sets for further analyses. The Figure 3 shows the distribution of number of genes across the different modules. As reproducibility of gene signatures discovery is particularly challenging, here we also employ assessment of animal models data so that between-species reproducible genomic patterns - presumably more robust - can enrich integrative analysis.

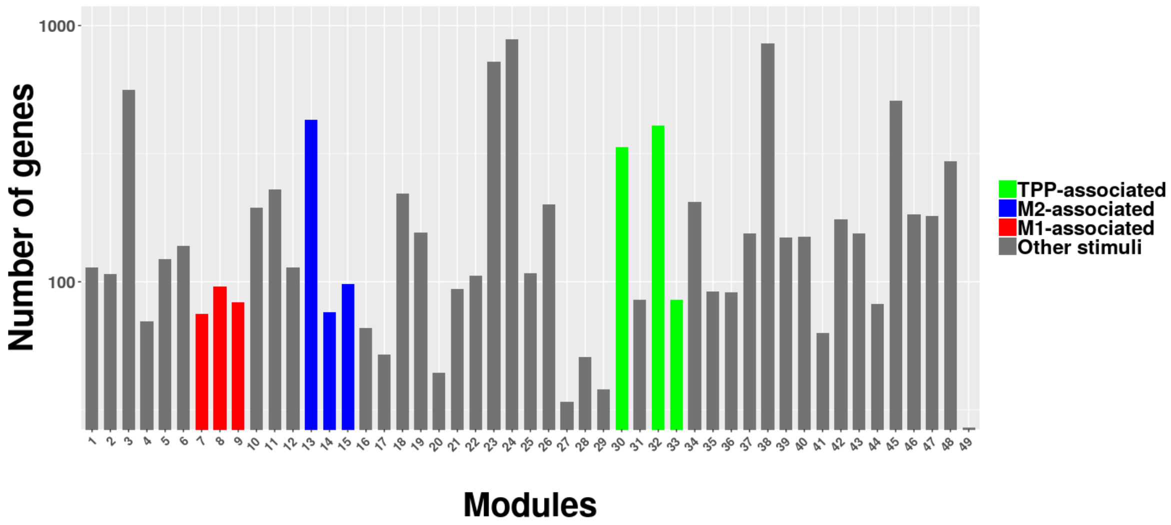

Figure 3. Gene modules distribution - data from Xue and colleagues (2014).

2. OBJECTIVES

Integrative genomic analysis is an interdisciplinary approach that arises from health sciences, engineering, biostatistics, computer sciences, and molecular biology advances. This study is mainly focused on the characterization of macrophage gene expression patterns within the context of Idiopathic Pulmonary Fibrosis, as well as the understanding of cellular subpopulations behavior at the transcriptomics level. As a general objective, we pursue the identification of genomic markers correlated with histopathological kinetics of IPF. Specific objectives are listed below.

- Characterize the temporal profile of gene signatures derived from macrophage subpopulations in animal models of IPF;
- Build numerical factor kinetically correlated with the profiles identified in the previous item;
- Assess the previously built numerical factor in gene expression datasets from IPF patients.

3. MATERIALS AND METHODS

All analyses were performed using R software (3.4.3) and CRAN or Bioconductor packages - Table 1. Complex file parsers were built with python (3.6 or later). Parametric differences were assessed using linear and generalized linear models, T tests, and Tukey's Honest Significant Difference test. False Discovery Rate was controlled for multiple comparisons using Benjamini-Hochberg procedure at 5% level. Non-parametric differences were assessed using Wilcoxon or Mann-Whitney tests. Other statistical procedures were performed according to R packages implementations. All code and figures are publicly available at github.com/giulianonetto/tcc.

Table 1. R packages used in this work.

Package	Utility
arrayQualityMetrics	Microarray quality control
Biobase	Microarray analysis
biomaRt	Data base query
car	General statistics
coin	General statistics
convert	Microarray analysis
dplyr	Data wrangling
GEOquery	Data base query
ggfortify	Plot with statistics
ggloop	Plot iteratively
ggplot2	Plot
ggpubr	Plot
ggrepel	Plot
ggsignif	Plot with statistics
limma	Microarray analysis
maSigPro	Microarray analysis
multcomp	General statistics
nlme	General statistics
rafalib	General statistics
RColorBrewer	Plot
reshape2	Data wrangling
stringr	Data wrangling
tidyr	Data wrangling

4. RESULTS AND DISCUSSION

4.1. IPF ANIMAL MODEL AT THE GENOMIC LEVEL

The bleomycin-induced IPF animal model is widely used to understand lung fibrosis pathology, regardless of its limited capability of mimicking the actual human disease (MOELLER et al., 2008). Bauer and colleagues studied this question when comparing microarray data from one hundred lung samples from IPF patients with rat lungs sampled several at time points after bleomycin and phosphate buffer (PBS) exposure (BAUER et al., 2015). Although they were able to identify disease-relevant translational gene markers, the point of highest rat-human gene expression commonality was at day 7 after rat lung aggression. The authors suggest that these gene signatures can be used to identify IPF patients and to stratify these according to disease severity. Here, we reanalyze their data in order to further understand time course patterns in gene expression and their relation with cellular pathological activity.

Using the arrayQualityMetrics R package, we were able to identify five outliers based on overall expression data and these were removed from further analysis - although the original paper indicated 17 outliers (BAUER et al., 2015). As morphological and cytometric analyses indicate that bleomycin model shows time-related pathological events (IZBICKI et al., 2002; VENOSA et al., 2016), we cut the original data into 5 supposedly divergent phases: namely, “Healthy” for untreated samples, “Injury” for rats killed at early exposure time points (3 and 7 days), “Early Fibrosis” (day 14), “Late Fibrosis” (days 21 through 28), and “Healing” (days 42 through 56). This generally arbitrary classification successfully showed descriptive gene expression patterns in principal component analysis - Figure 4.

The first three principal components separate control and bleomycin samples, explaining over 86% of the total variance. Most importantly, the samples at early time points - i.e. when the recent aggression induces major inflammatory responses - fall well separated from later times as well as from control samples. Notably, the injury-labeled samples fall further beyond others, followed by early fibrosis, late fibrosis, and finally healing-labeled and control samples - almost mimicking the actual time course experimental design and suggesting the impact of measurement times on IPF animal model assessment. Even though the authors indicate that day 7 (injury phase) is the point with maximum similarity among animal model and the actual human lung disease, the variance observed here may not be neglected, especially regarding assessment of IPF candidates and early-diagnosis procedures.

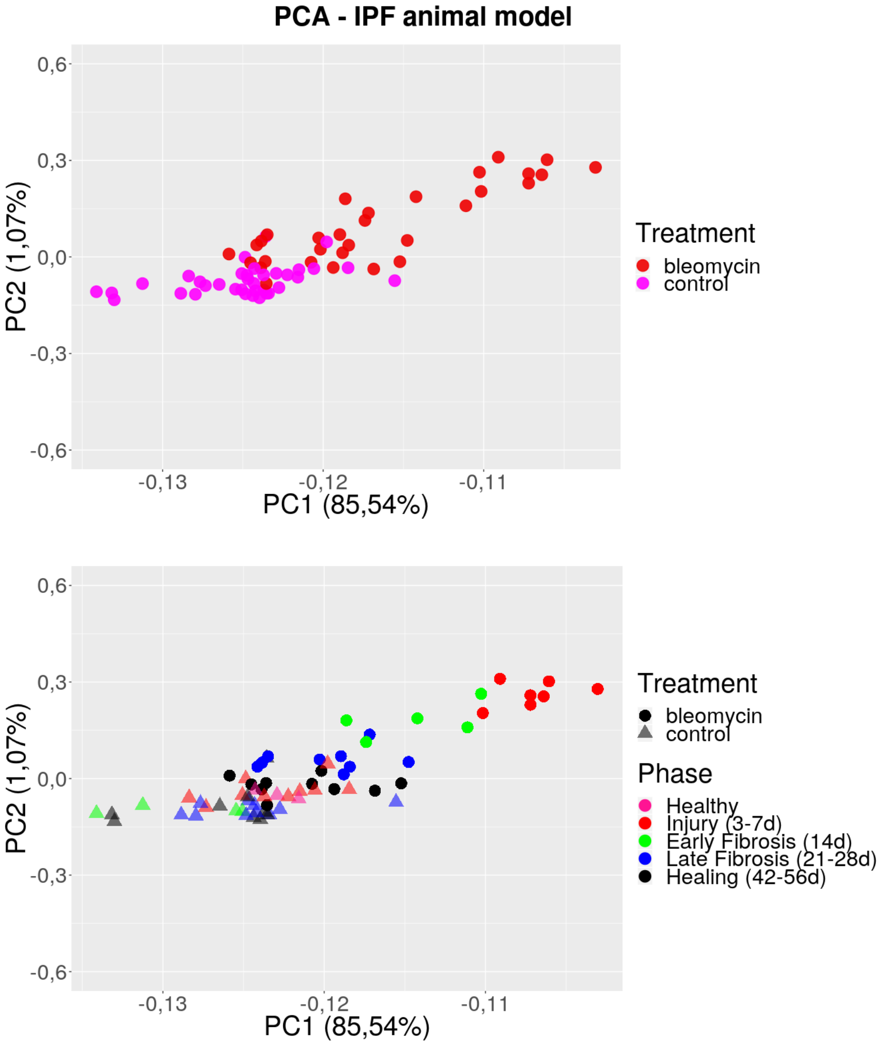

Figure 4. Principal Component Analysis of gene expression from Bauer and colleagues data (2016).

Once disease is installed, one may expect reproducible gene expression patterns, even though this understanding is hindered by the idiopathic characteristic of the condition. However, those patients with developing histopathological characteristics that are yet to be diagnosed as typical IPF may not reflect such genomic patterns. Furthermore, it has been reported that gene signatures differ significantly across IPF patients with progressive and stable conditions (BOON et al., 2009). Thus, it is important to note the importance of longitudinal studies regarding genomic signatures as these may prove themselves helpful when predicting disease onset, progression, and stabilization.

Regarding macrophage biology, several approaches are possible to assess their dynamics in animal models. As previously noted, Venosa and colleagues were able to describe macrophage activity in an animal model of IPF induced by nitrogen mustard (VENOSA et al., 2016). Using data from cytometric, qRT-PCR, and other non-molecular assays, the authors demonstrated the inflammatory profile of infiltrating cells at early time points, while anti-inflammatory and healing profiles where dominant at later times. The proposed kinetics related well with gene expression patterns, although high-throughput technologies were not used.

4.1.1. Macrophage polarization in IPF animal model

Here, the first macrophage characterization addresses the M1 versus M2 paradigm. As proposed by Buscher and colleagues, the Polarization Factor Ratio (PFR) is intended to describe the degree of macrophage polarization towards M1 or M2 spectra (BUSCHER et al., 2017). As a simple model, it derives from the expression levels of M1- and M2-markers. Using Bauer's data, two-way mixed design ANOVA with multilevel modeling did not reveal any time-dependent patterns ($p > 0.05$). However, it did reveal significant differences across treatment groups ($p < 0.005$), which was further confirmed by t-student ($p < 0.01$) and Exact Wilcoxon-Mann-Whitney tests ($p < 0.01$). Figure 5 shows that time courses for both groups fail to trend any direction significantly - perhaps due to noisy data points. However, the curves do not overlap completely, and the boxplots illustrate the distribution differences. In the original paper, Buscher and colleagues demonstrated higher prediction power for the PFR built over IL12b and Arginase 1 expression levels - when comparing to the same score constructed with inducible nitric oxide synthase (iNOS - NOS2 gene, also M1-related) instead of the cited interleukin. Here, the effect size comparison seemed to be shifted, and the PFR (iNOS/Arg1) showed better separation between bleomycin- and PBS-treated animals. Taken together, these data initially indicate that PFR is capable to reflect a slight overall macrophage polarization towards an M1 spectrum in the given pulmonary fibrosis animal model. As a two-gene model, however, such a conclusion is clearly an oversimplification of macrophage and IPF biology.

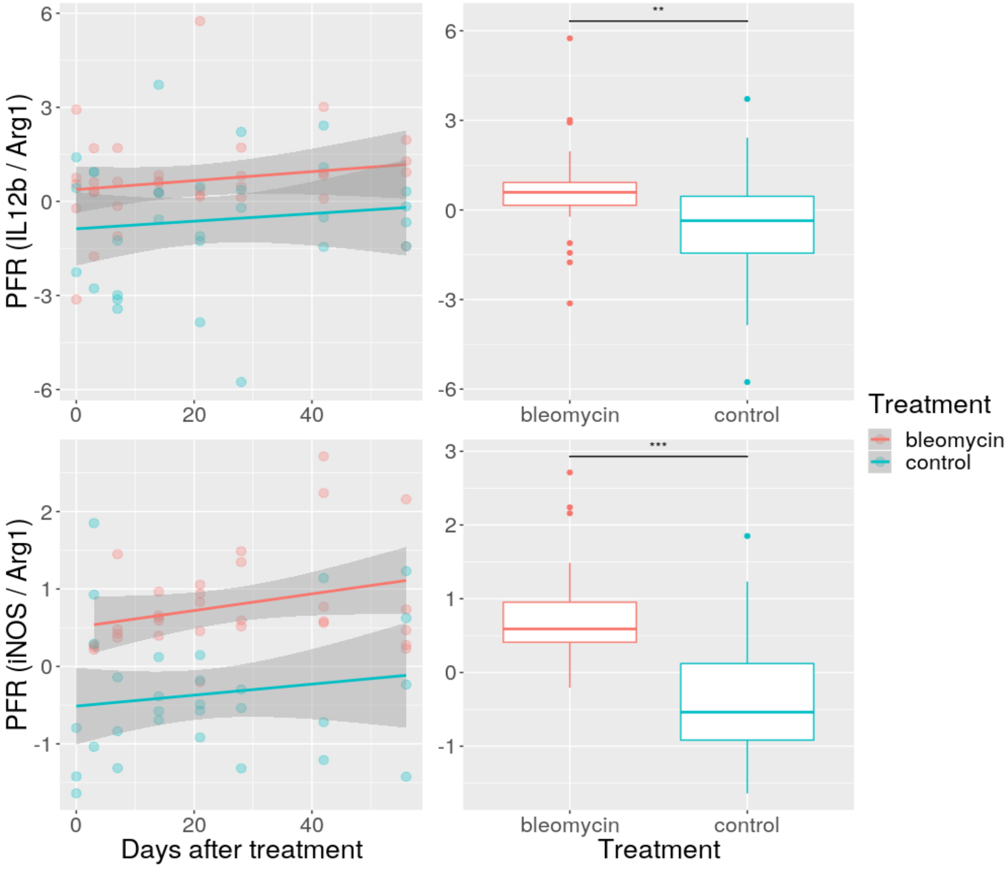

Figure 5. PFR with IPF animal model data from Bauer and colleagues.

** p = 0.007983; *** p = 4.335e-07 (Exact Wilcoxon-Mann-Whitney Test).

In order to further characterize the time-course differences in overall gene expression of bleomycin- and PBS-treated rats, we performed differential expression analysis on Bauer’s data using a two-step statistical method which is especially designed for time course data and is implemented in the Bioconductor package, maSigPro (CONESA et al., 2006). First, the procedure fits a global model for all genes in a given dataset. Then, it applies step-wise regression as a means of variable selection so that it can detect significant differences across study groups and consistent expression profiles across time points. Although the dataset tested contained eight time points for each group, here we relied on a cubic regression model, Higher polynomial degrees have yielded high noisy fitting and possibly high rates of type I error (data not shown), which is somehow expected when working with overly complex polynomials (CONESA et al., 2006). One could argue the use of splines, but these are not available in the maSigPro package. To avoid underfitting, the time points 42 and 56 were excluded from this analysis. A 5% FDR cutoff was used to identify genes with significant differential expression between groups.

MaSigPro also conducts cluster analysis with several strategies to identify similar expression profiles across time. Using the algorithm from mclust R package (Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation - available on CRAN), it can group the time courses into an optimal k number of clusters based on finite normal mixture modeling (SCRUCCA et al., 2016). However, hierarchical clustering showed similar results (with k = 9) and these were taken for further analysis. Figure 6 shows the nine clusters produced by maSigPro and their time course profiles. The dashed lines represent the fitted models, while solid lines show the true median expression values (higher resolution file available in https://github.com/giulianonetto/tcc/tree/master/rmd-files/Development_files/figure-docx/hclust_bauer2015_better.png).

Notably, there are fairly similar clusters (e.g. clusters 6 and 7). Here, however, we are particularly interested in those which show overexpression either at early or later time points, as these may be representative of eventual macrophage polarization patterns. For instance, one may speculate cluster 1 to be filled with genes related to the M1 spectrum, while cluster 4 seems to follow a transitory course and, finally, cluster 9 may represent an M2-polarized environment. Clearly, these are limited speculations once overall expression patterns greatly overlook macrophage dynamics. Therefore, deeper characterization required assessment of gene profiles on a case-by-case basis.

4.1.2. Overall gene expression and immune-related pathways

Based on recent literature, we sought to find genes that have been previously reported as related to immune cells activity and IPF. Figure 7 shows the expression for chemokine (C-C motif) ligand 2 (CCL2), also known as Monocyte chemoattractant protein 1 (MCP1), whose time course profile is representative of a selected set of other significantly differentiated chemokines that were also gathered into cluster 2.

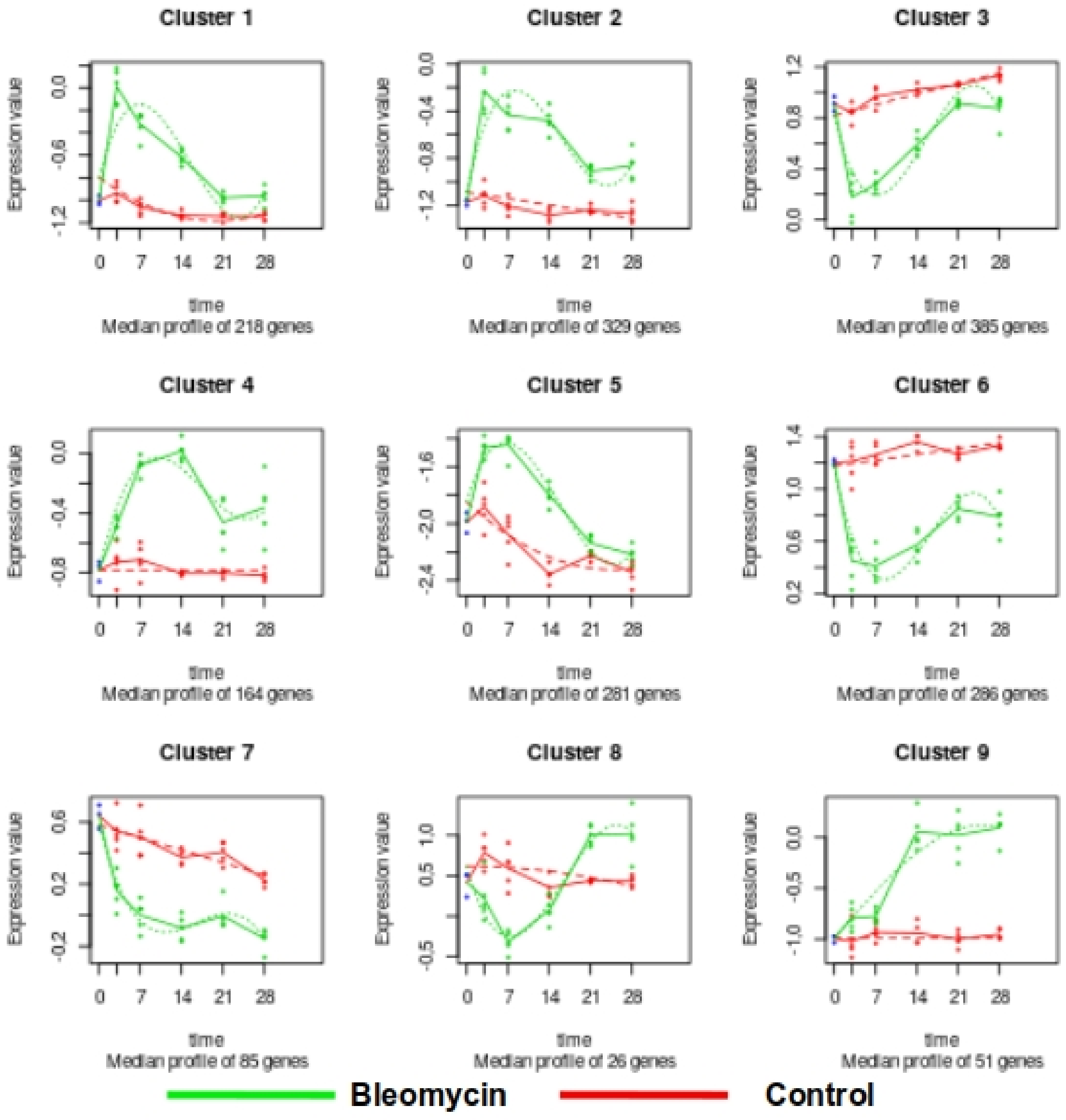

Figure 6. Hierarchical clustering reveals time course expression profiles of differentially expressed genes in IPF animal model.

Many chemokines have been related to fibrotic processes in general (SAHIN; WASMUTH, 2013). Specifically, CCL2/MCP1 contributes to fibrosis development as chemoattractant to monocytes, macrophages, epithelial cells, and fibroblasts - a role that interacts with other cytokines such ass TGFβ-1, IL4 and IL13 (DELLA LATTA et al., 2015). In fact, CCL2 directly induces fibroblasts to express TGFβ-1, which mediates collagen production. Pirfenidone, a medication currently indicated for IPF treatment (see Introduction), has been shown to inhibit the release of both CCL2 and CCL12, which helps to lower fibrotic process (INOMATA et al., 2014). Not surprisingly, the latter chemokine also showed expression profile compatible with cluster 2 (Figure 7) as its involvement in lung fibrosis has been well established (MOORE et al., 2006). CXCL12, which stands for C-X-C motif chemokine 12, has similar functional properties, although its expression values were somewhat noisier and appeared within cluster 4 - which shows higher increase at slightly later time points (data not shown). Finally, CCL7, CCL22, and CCL24 also clustered into the second group, are all associated with increased fibrosis and are thought as potential targets for immunotherapy development (YOGO et al., 2009; SAHIN; WASMUTH, 2013; AMUBIEYA et al.,2016; GIESECK; WILSON; WYNN, 2017; LEE et al., 2018).

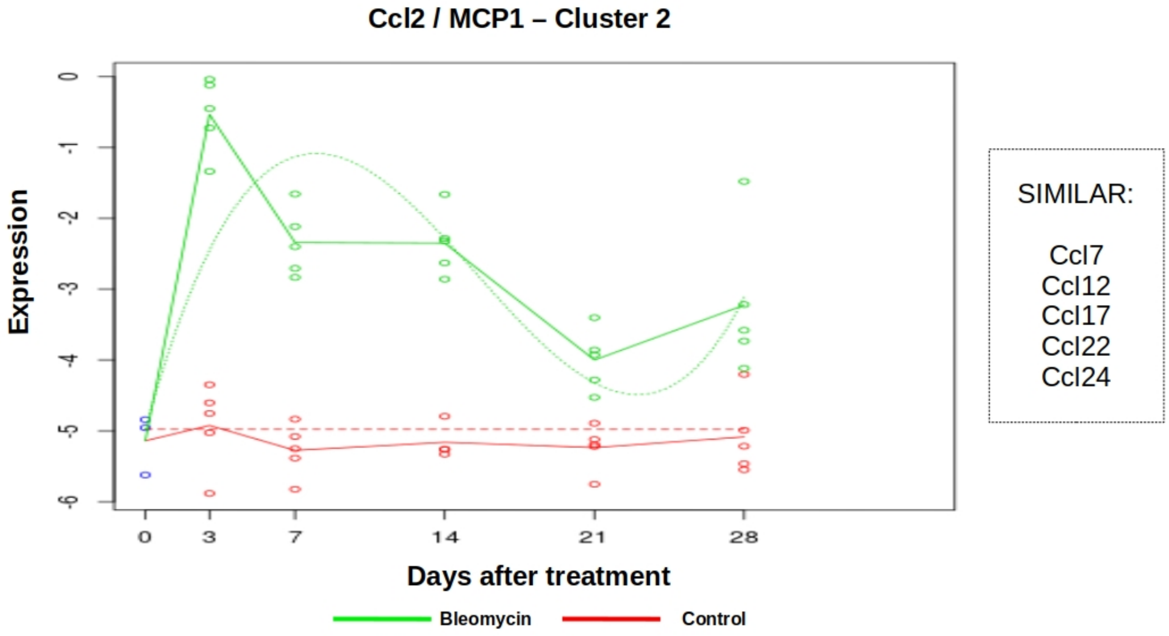

Figure 7. CCL2 (MCP-1) as representative of selected chemokines in Cluster 2. Similarly expressed chemokines listed on the right.

Chemokines ligands act through interaction with chemokine C-C/C-X-C motif receptors, and the expression of these latter molecules is also correlated with IPF (SAHIN; WASMUTH, 2013). Interestingly, the expression of CCR5 was grouped into cluster 1, showing high levels at very early time points - this held true for two probes mapping to the same gene, a common finding in microarray data. This protein has shown reportedly reduction behavior in bleomycin animal models and IPF patients, although has been suggested that its depletion has anti-fibrotic effects (ISHIDA et al., 2007). Another chemokine receptor, CCR2, is associated with CCL2 activity, although it does interact with other chemoattractant agents, including CCL7, CCL8, and CCL13 (SAHIN; WASMUTH, 2013). It is present in monocytes, T helper lymphocytes, and dendritic cells. As CCR2 deficient mice have been reported as protected against lung fibrosis through multiple mechanisms (rather than immune cell trafficking solely), the CCL2-CCR2 axis revealed itself as a potential pharmacological target. Expression profiles of CCR2 and CCR5 are illustrated in Figure 8.

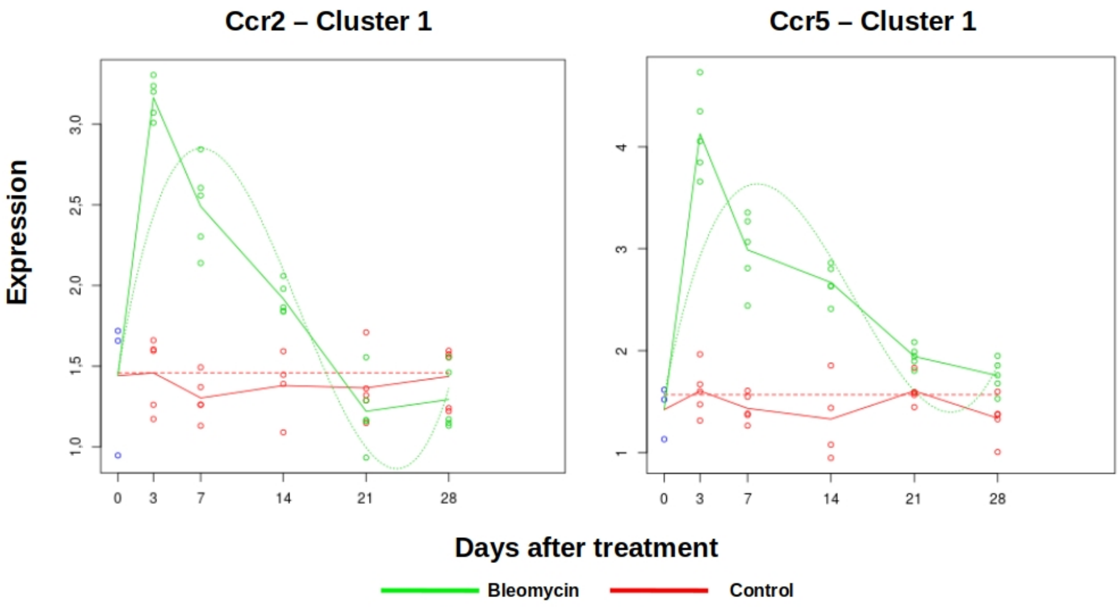

Figure 8. Expression profiles for CCR2 and CCR5.

Not long ago, clinical investigation was carried out to assess the efficacy of Carlumab, an anti-CCL2 antibody (RAGHU et al., 2015a). The evident failure of the treatment suggested that modulating chemokine pathways may turn out much more complex than expected. Recently, Milger and colleagues addressed this unfortunate surprise by measuring CCR2 expression levels on subtypes of immune cells (MILGER et al., 2017). Previous work from their research group revealed that children with interstitial lung disease showed increased CCL2 release as well as CCR2+ CD4+ T-cell frequencies, and these findings were correlated with disease severity and lung function (HARTL et al., 2005). The confounding scenario led the authors to wonder about the “multi-faceted role of CCR2+ cells in lung injury” (MILGER et al., 2017). In the updated work, they found major immunosupressive roles played by CCR2+ CD4+ T cells, which were associated with T regulatory cells. Using adoptive cell assays, they were able to attenuate lung inflammation and fibrotic process development. This is specially surprising as CCR2/CCL2 signals have been extensively associated with pro-inflammatory activity - mainly related to innate immune system. They conclude that depleting such CCR2-depending pathways can no longer be addressed while not considering their heterogeneous functional properties throughout immune cellular system - a claim that might hold true for other related molecules.

Another set of molecules particularly investigated in IPF is the group of metalloproteinases (DANCER; WOOD; THICKETT, 2011). A wide range of these enzymes has been reported as overexpressed in IPF patients and animal models (PARDO et al., 2016). Potential peripheral blood biomarkers for IPF include MMP1 and MMP7 (RICHELDI; COLLARD; JONES, 2017). Their plasma concentrations were able to distinguish IPF from patients with chronic obstructive pulmonary disease, sarcoidosis, and chronic/subacute hypersensitivity pneumonitis - reaching sensivity and specificity values as high as 96.3% and 87.2%, respectively. MMP7 also predicted well subclinical interstitial lung disease, reduced forced vital capacity and carbon monoxide diffusing capacity.

On the other hand, not all studies with bronchoalveolar lavage fluid studies have accused increased MMP1 levels, although microarray data on whole lung tissue does support increase detection (DANCER; WOOD; THICKETT, 2011). Bauer and colleagues, the authors of the study whose data we have been analyzing so far, did find upregulation of MMP7 gene, but not MMP1 (BAUER et al., 2015). Similar results were generated by maSigPro algorithm herein reported, and the gene fell into cluster 1 - Figure 9. Still, while the former is known as a pro-fibrotic agent, the latter represents a paradox that is yet to be solved. Capable of cleaving fibrillar collagens, MMP1 is associated with excessive extracellular matrix degradation - which is opposed to IPF pathogenesis (PARDO et al., 2016). Although partial explanations involve its production location, usually separated from fibroblasts and collagen accumulation, the roles of MMP1 are far from being fully explained. Overexpression of MMP12, another pro-fibrotic agent, is also a common finding in IPF animal models

(PARDO et al., 2016). Here, it was found to be significantly overexpressed across time, being associated to cluster 2 from maSigPro analysis - Figure 9. More time-consistent, however, was the time course profile of MMP14 (cluster 4), whose role in IPF is presently unknown - note its expression increase takes longer than genes in cluster 2. Both enzymes were also identified by Bauer’s analysis with similar time trends.

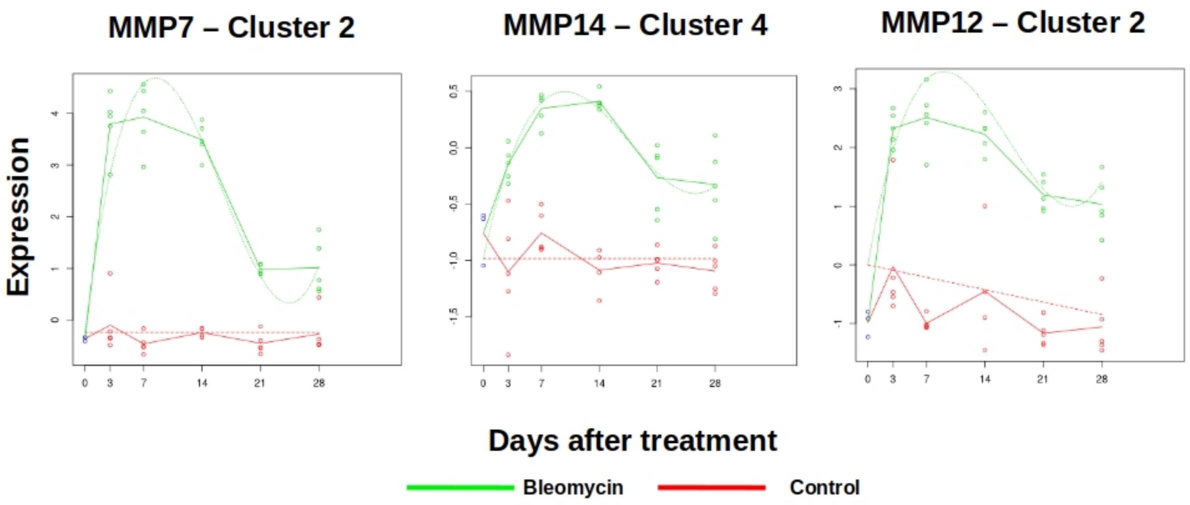

Figure 9. Expression profiles for MMP7, MMP14, and MMP12.

As in the case of chemokines, metalloproteinases have been proven as a complex set of opportunities for pharmaceutical and biomarkers development. Augmenting expression MMP13 and MMP19, two anti-fibrotic enzymes, seems to have therapeutic potential (CRAIG et al., 2015). Inhibiting pro-fibrotic MMPs is also intended. While global inhibition may not be beneficial, monoclonal antibody engineering is an approach under active research (CRAIG et al., 2015; SELA-PASSWELL et al., 2012). The enzymes, though, represent only a fraction of the complexity involved in IPF pathogenesis.

Major cytokines associated with IPF include IL12, IL33, IL1-β1, TGF-β1, IL4, IL13, IL25, and so on (GIESECK; WILSON; WYNN, 2017). The highly complex interaction networks formed by these multiple-origin and multiple-targeted molecules has been extensively addressed, and the current understanding is that wound-healing and pro-fibrotic mechanisms are still to be scrutinized. As illustrated by the CCL2/CCR2 case, these molecular patterns are thought to be highly context dependent, and so the cellular dynamics behind their expression and biological activities builds a major and multifaceted challenge towards pulmonary fibrosis understanding. In this sense, it is important step back and acknowledge that a hallmark of dysregulated fibrogenesis is the excessive deposition of extracellular matrix (ECM). Fibroblasts are prominent players in ECM deposition as they hyperproliferate within damaged tissue, become resistant to apoptosis, and differentiate into pro-fibrotic myofibroblasts - which perpetuates the fibrotic process (KOLAHIAN et al., 2016). While in this inflammation-driven activated status, these cells show hypersensitive responses to a wide range of chemical signals, including many of the above cited molecules, but also to leukotrienes, prostaglandins, and growth factors (KENDALL; FEGHALI-BOSTWICK, 2014). Moreover, myofibroblasts can also produce, for instance, IL1-β1, TGF-β1, IL33 and other allarmins, several chemokines, and even reactive oxygen species. Finally, fibroblast-mediated remodeling of extracellular space greatly contributes to the trafficking of immune cells. In fact, fibroblasts work alongside epithelial/endothelial cells and perivascular macrophages to regulate alveolar repair and fibrosis (KOLAHIAN et al., 2016). It is within this complex scenario that we address cytokine production. Although cell population frequencies can be inferred from gene expression studies, one must keep in mind the multiple sources and targets of immune-related chemicals.

TGF-β1 has been classically associated with fibrotic processes, and many other fibrosis-related cytokines are thought to work through its signaling pathways (FERNANDEZ; EICKELBERG, 2012). It is centrally active in epithelial-mesenchymal transition (EMT), a process through which epithelial cells assume mesenchymal properties, acquire capacity migrate and to differentiate into ECM-producing fibroblasts (KALLURI; WEINBERG, 2009; KOLAHIAN et al., 2016). Inflammatory cytokines, such as IL1-β1, TNF-α, and IFN-γ, have modulatory effects over TGF-β1 production, comprising a network that involves monocytes, macrophages, epithelial cells. Its gene expression profile has been demonstrated both in IPF patients and animal models (LUZINA et al., 2015). Interestingly, using our third-degree polynomial regression model, TGF-β1 overexpression over time could not be detected. As in the case of PFR, however, Exact Wilcoxon-Mann-Whitney test revealed global differences between bleomycin- and PBS-treated rats - Figure 10. Of note, Bauer and colleagues detected differential expression for this gene at day 3 after bleomycin exposure (BAUER et al., 2015).

Note that we have been using maSigPro plotting functions for cases were the generated regression curves produced statistical significance. In this particular case, however, we use the versatile, CRAN-available ggplot2 package, which is an implementation of “The Grammar of Graphs” (WICKHAM, 2015).

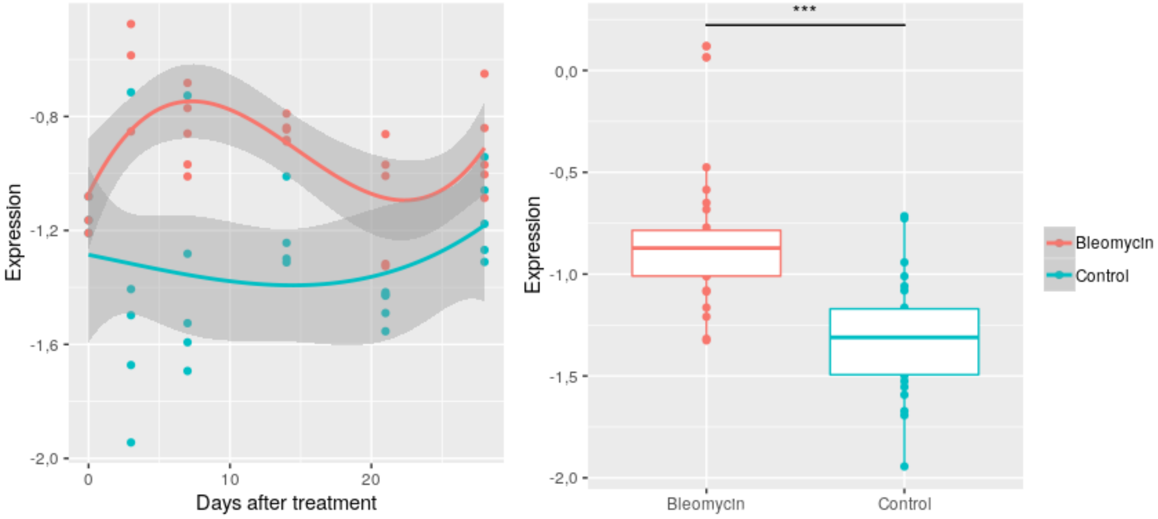

Figure 10. Expression profiles for TGF-α1. *** p < 0.0001

Another important cytokine in the context of IPF is IL1-β1 (LUZINA et al., 2015). As well as TGF-β1, the molecule is produced by and it acts on fibroblasts (KENDALL; FEGHALI-BOSTWICK, 2014). Major amounts are also produced by macrophages (WYNN; BARRON, 2010). In fact, this cytokine is able to reproduce many features of bleomycin-induced pulmonary fibrosis, and the blockade of its signals through monoclonal antibody administration has reduced mice fibrotic development (BYRNE; MAHER; LLOYD, 2016). As a classical inflammation biomarker, IL1-β1 was detected through polynomial regression and placed into cluster 1. Nonetheless, Bauer and colleagues did not reported differential expression for this gene.

As shown in Figure 11, the levels of this cytokine were indeed uncommonly increased in two of samples from day 0, and all the control data looks somewhat noisy. Still, pairwise comparisons using Wilcoxon rank sum test revealed significant differences at days 3 and 7 after treatment (p < 0.01; data not shown). Overall, this is a good example of why microarray experiments often need to undergo validation with qRT-PCR assays, and even of how different statistical and bioinformatics procedures can impact results. Another confusing case includes the detected underexpression of IL33 (cluster 6), which goes against recent literature reports (KOLAHIAN et al., 2016;

LI et al., 2014). Additionally, overexpression at late time points of its receptor gene, Interleukin 1 receptor-like 1 (IL1RL1), was detected in cluster 9 - and also reported by Bauer and colleagues (days 14 and 21).

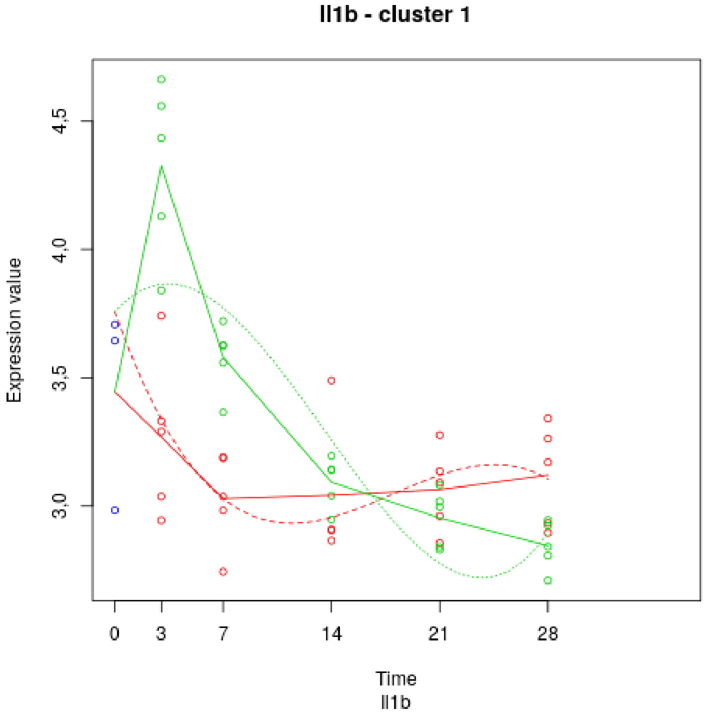

Figure 11. Expression profiles for IL1-β1.

Finally, other noteworthy immune-related genes that were found to be differentially expressed in Bauer's data include the costimulatory molecules CD80 and CD86, which are prominent in the functionality of antigen-presenting cells (APCs) such as macrophages (COLLINS; LING; CARRENO, 2005); the macrophage surface marker, CD68, a scavenger receptor (VENOSA et al., 2016); the macrophage-modulating, anti-inflammatory apolipoprotein E (ApoE) (BAITSCH et al., 2011; YAO et al., 2016); and the IL-13 receptor subunits alpha-1 and alpha-2 (IL13R-α1/α2). Note that both subunits were overexpressed in Bauer's data, contradicting recent reports for subunit alpha-1 in murine models and IPF patients (KARO-ATAR et al., 2016). Additionally, protective roles have been suggested for both IL13R-α1 and IL13R-α2 in the context of pulmonary fibrosis (KARO-ATAR et al., 2016; LUMSDEN et al., 2015). The expression profiles of the cited molecules are illustrated in Figure 12.

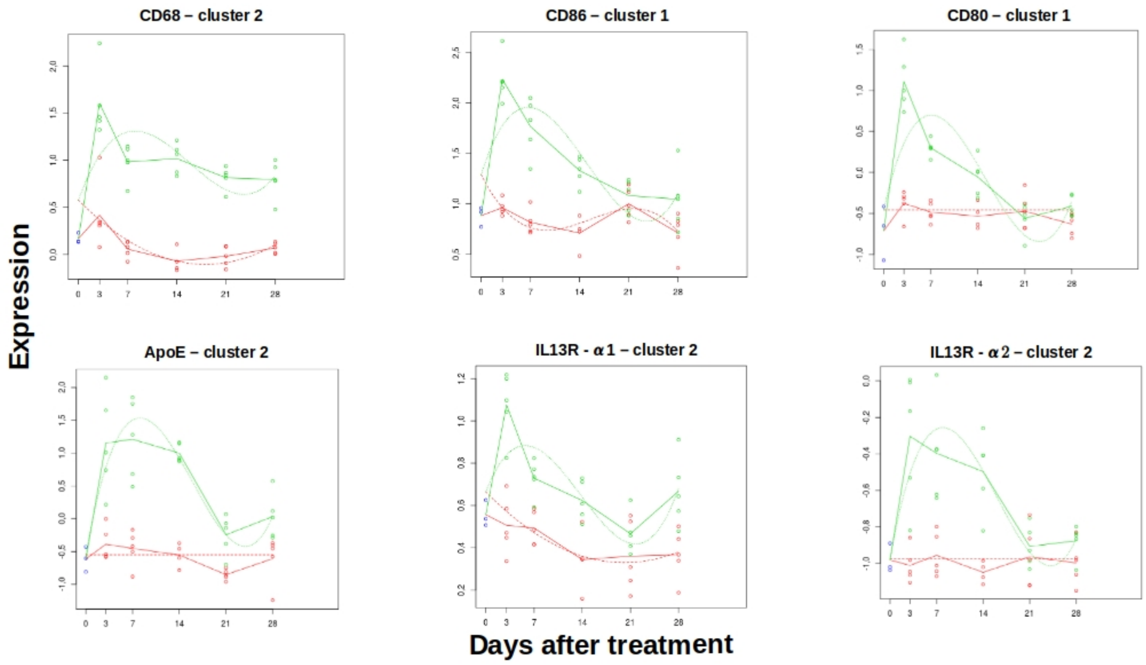

Figure 12. Expression profiles for CD68, CD80, CD86, ApoE, and α1/α2 subunits of the IL13 receptor.

4.2. MACROPHAGE POLARIZATION FACTOR RATIO REVIEWED - A MODULAR PERSPECTIVE

In the last decade, traditional immunology research was faced with innovative systems biology approaches in the work of Chaussabel and colleagues (CHAUSSABEL et al., 2008). As noted by Ena Wang and Francesco M. Marincola (2008) in the Immunity Previews of July 18, 2008:

> *“In summary, Chaussabel et al. (2008) suggest an inductive approach to pathway discovery: Disease-specific gene-expression patterns are identified and condensed into few functional units; these are presumed to represent down-stream effects of biological mechanisms determining the disease status (…). This evidence-based analysis represents a paradigm shift in which system biology (immunology) is approached from the bedside, yielding information most likely to be relevant to human suffering and confronting the basic immunologist and cell biologist with the challenge of aligning experimental observations with the reality of human disease approached in its uncontrollable complexity. Moreover, the modular approach offers practical applications as a global-biomarker-discovery tool that will need to be*

> *aggressively validated in the future.”*
>
> (WANG; MARINCOLA, 2008)

This was the very first introduction to the approach used by Xue and colleagues, over 6 years later, to describe the transcriptome patterns from almost 300 human macrophages through single-cell RNA sequencing technology (XUE et al., 2014). In fact, although the PFR performed well in the recent work form Buscher and colleagues, a 2-gene model is clearly an oversimplification of macrophage biology and hence probably not the best method for describing disease pathological status. In the original paper, the authors actually employed their PFR to select gene signatures from LPS-stimulated murine macrophages and these were ultimately used to assess human disease (BUSCHER et al., 2017). Using Bauer’s IPF animal model data, we identified global differences in “raw” PFR between bleomycin-treated and control rats, even though IL12, arginase 1 and iNOS were not detected as differentially expressed. Now, we employ these new perspectives of immunology to study modular Polarization Factor Ratio (mPFR) candidates, which are based upon Xue’s modules and the clusters derived from longitudinal assessment of Bauer’s data.

4.2.1. Defining modular Macrophage Polarization Factor candidates

Twenty eight candidates to modular Macrophage Polarization Factor (mPFR) were defined according to Equation 1. The difference in sample means for each module (or gene set) was corrected by a baseline factor constructed with a reverse ratio between the global means from the tested modules.

$$\mathbf{mPFR} = \left(\text{Mean }(\text{Set}_1)_{\textbf{sample i}} - \text{Mean }(\text{Set}_2)_{\textbf{sample i}}\right) \times \frac{\text{Mean }(\text{Set}_2)_{\textbf{all samples}}}{\text{Mean }(\text{Set}_1)_{\textbf{all samples}}}$$

Equation 1. Formula for mPFR candidates calculation. Each module is represented by a “gene set”.

Given its complex structure, the complete list of the modules are stored as an R data frame object at github.com/giulianonetto/tcc/rmd-files/data/mPFRcandidates.RDS. It can be downloaded and analyzed using R software - its conversion into text or spreadsheet-like files is not trivial. Basically, there are two columns organized so that the rows assign the two sets used to construct a given mPFR candidate, which is identified by a row name. As a reference, IL12β and Arginase 1 were used for the first candidate, namely Set 1 (S1). The following two candidates, S2 and S3, were constructed with relevant chemokines and cytokines sorted out according to gene expression profiles described in the previous section. The other twenty five candidates were constructed using the nine clusters observed in the longitudinal analysis of Bauer's data. These were identified with a letter C regarding their clustering origin (C1-25). Using this approach, the previously described clusters 1 through 5 were matched with clusters 3, 6, 7, 8, and 9, and the resultant candidates were assigned as C1-C20. Candidates C21 through C23 were formed from cluster 1 against clusters 2, 4, and 5. Finally, candidates C24 and C25 were built with cluster 2 against clusters 4 and 5.

The generation of mPFR candidates respected the reasoning that differences between clusters with overexpression at early time points (e.g. cluster 1) and those with either underexpression (e.g. cluster 7) or late overexpression (e.g. cluster 9) profiles are expected to show higher descriptive power of injured and normal tissues. Although arbitrary, this hypothesis generated twenty eight new variables which were approximately normally distributed. This can be seen in the quantile-quantile (QQ) plot shown in Figure 13. QQ plots are often used to visually compare a given distribution – in this case, Gaussian (IRIZARRY; LOVE, 2015). If the data is normally distributed, the residuals from a fit linear model should also follow such distribution and so the proportion of data that fall into its ranked quantiles can be predicted. In Figure 13, the red lines show the expected values, while dark dots show amount of real data within each quantile. The distribution of candidate S2 was used as representative, but QQ plots for all variables are available at github.com/giulianonetto/tcc/rmd-files/Development_files/figure-docx/LastChapter.

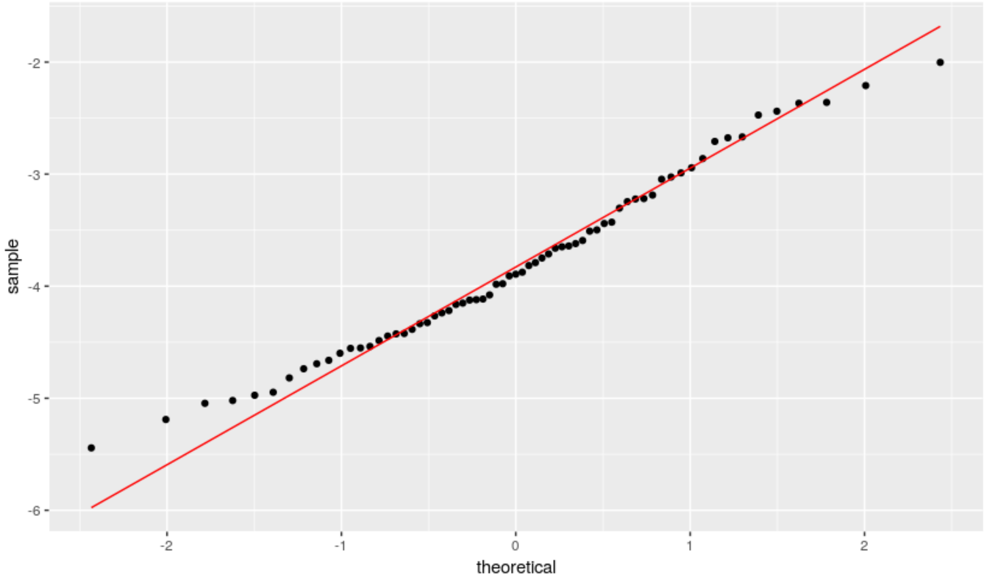

Figure 13. Quantile-quantile plot for C25 candidate showing approximately normal distribution.

4.2.2. mPFR candidates descriptive power

In order to investigate global patterns of newly constructed mPFR candidates, principal component analysis (PCA) was once more applied. Using only the first two principal components, one must notice the radical improvement in Bauer’s data separation as described by the treatment factor – Figure 14. Recall that our first PCA investigated patterns using whole gene expression data. As illustrated in Figure 4, the procedure was able to separate bleomycin- and PBS- treated rats with some group overlap. Also, its pathological time-trend seemed to be revealed when coloring by “disease phase” - e.g. injury, days 3 and 7, and healing, day 56. Here, the separation between control and bleomycin sample is complete. Furthermore, when looking at the time trends, one can recognize days 3 and 7 clustered together, while samples from day 14 form a one-factor group, and the following days seem to distribute accordingly until healing phase (days 42/56) and, finally, control/untreated samples – Figure 15.

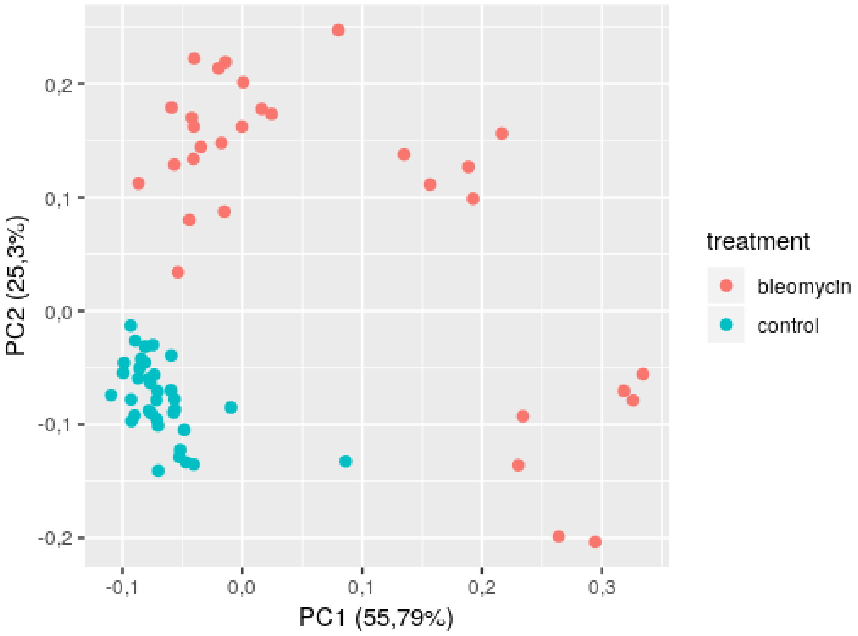

Figure 14. Principal component analysis of bleomycin- and PBS-treated samples using the 28 mPFR candidates. Colors represent treatment conditions.

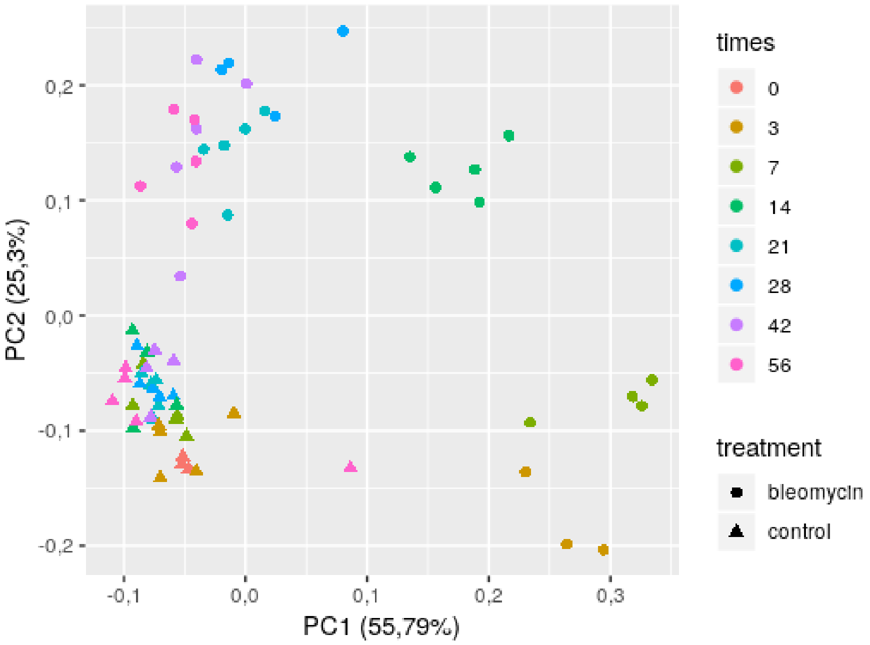

Figure 15. Principal component analysis of bleomycin- and PBS-treated samples using the 28 mPFR candidates. Colors represent time points, while shape now describes the treatment.

Finally, hierarchical clustering was performed in order to further compare descriptive performances between our candidates and the original gene expression matrix. Using the actual genes from Bauer's dataset, the sample-scaled clustering failed to separate well case and control subjects – Figure 16. Surprisingly, when performing the same procedure over the matrix of sample-scaled mPFR candidates, perfect separation between treatment and control samples is revealed – Figure 17. Additionally, the time trend was partially kept consistent with early time points (injury phase) clustered together while clusters from later times (late fibrosis and healing phases) appearing increasingly closer to control-sample clusters. Taken together, these data strongly suggest that the proposed data transformations can potentially build genomic fingerprints for pulmonary fibrosis that are even more powerful than gene expression alone.

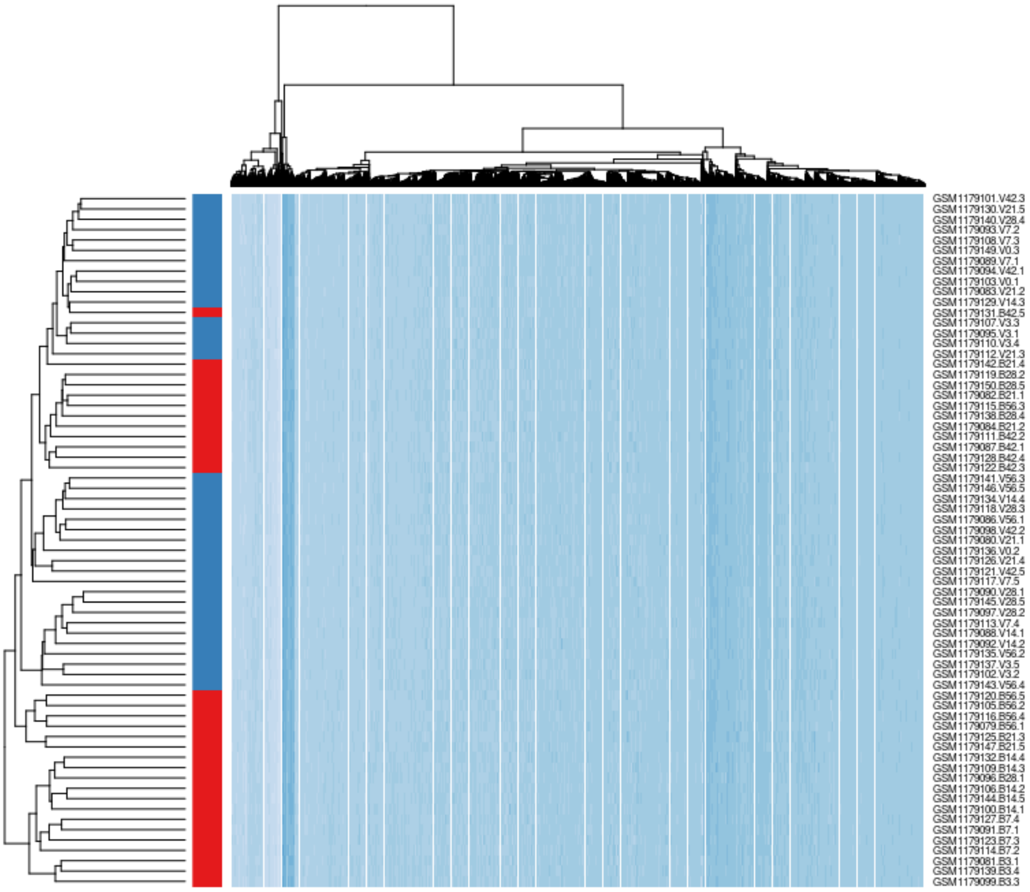

Figure 16. Hierarchical clustering over gene expression matrix shows partial separation between bleomycin (red) and control (blue) samples.

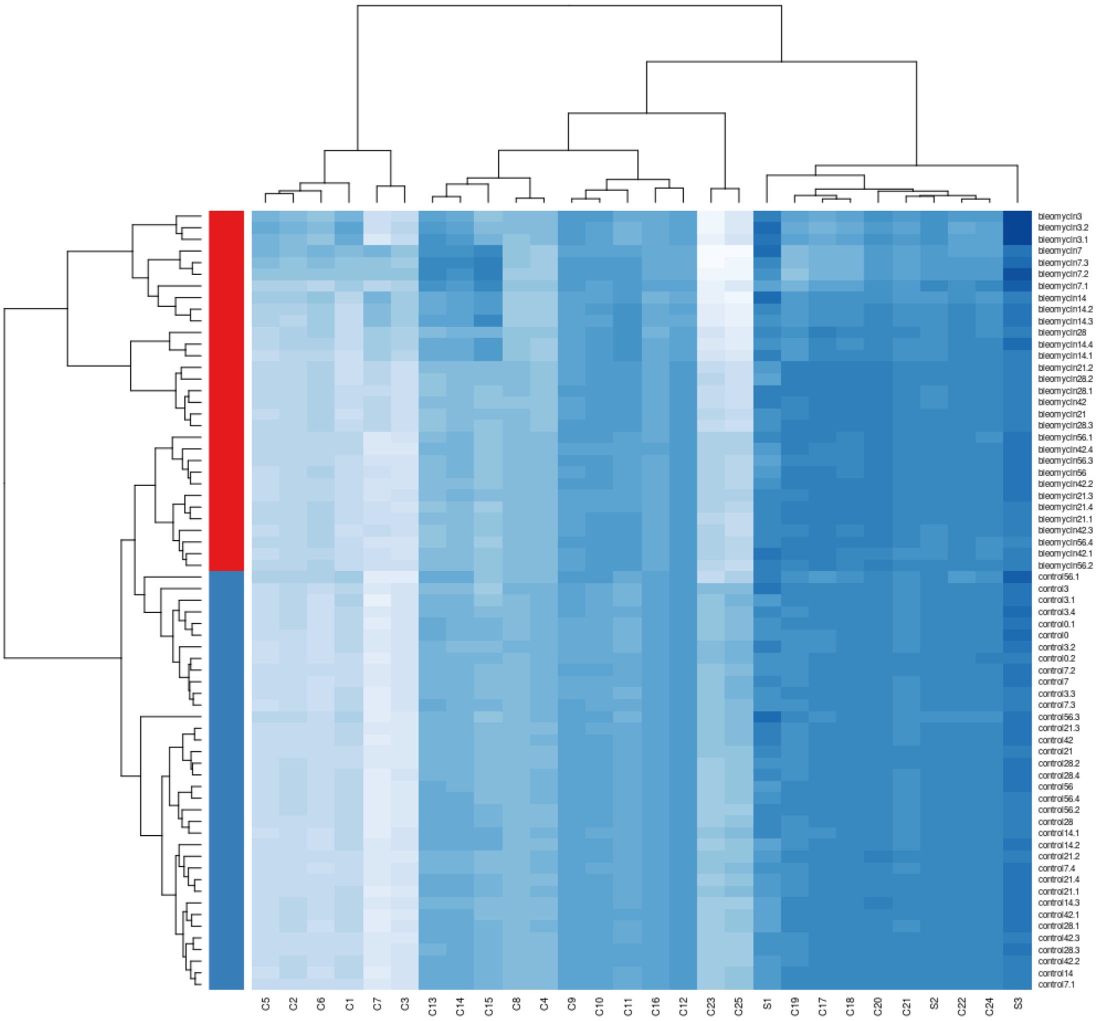

Figure 17. Hierarchical clustering over mPFR candidates matrix shows complete separation between bleomycin (red) and control (blue) samples. Bleomycin-treated samples also show partially consistent time trends.

4.2.3. mPFR candidates as prediction variables

Next, we aimed to assess which of the mPFR candidates were significantly different across treatment groups. As normal distribution was not met for all variables, we used Wilcoxon rank sum test to check for statistical significance with a cutoff of p < 0.05. Still, t-tests were also computed in order to check for results consistency - outliers were removed using linear modeling and Bonferroni correction to detect unusually biased observations. Figure 18 shows the results for the first 4 candidates tested – all boxplots are publicly available at github.com/giulianonetto/tcc/Development_files/figure-docx/LastChapter. Most of the variables showed statistically significance differences between bleomycin- and PBS-treated rats, although the sizes of the differences varied greatly. Of note, bleomycin effect over S1 candidate, used as reference (IL12b- and Arg1-based), was not statistically significant in the robust test performed. The test used Bauer's data from day 7, which is reported as the one with greatest correlation with actual human disease (BAUER et al., 2015).

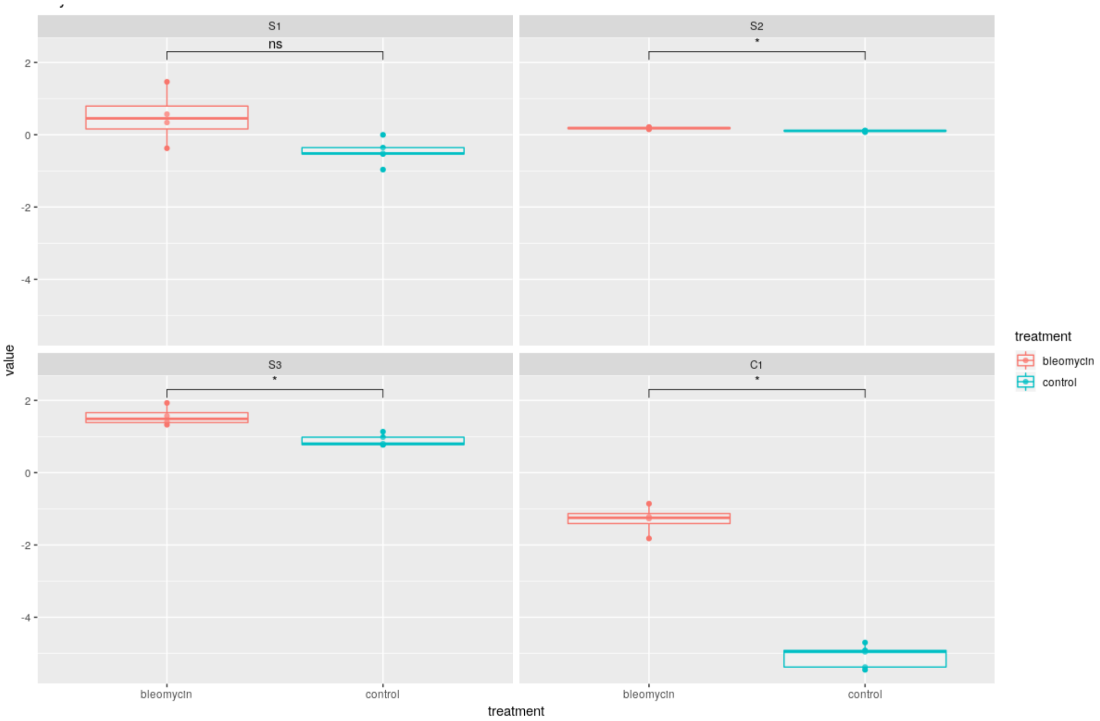

Figure 18. Wilcoxon rank sum test for the first 4 mPFR candidates between bleomycin- and PBS-treated samples. * $p < 0.05$.

As visualization of twenty eight boxplots is not trivial for human interpretation, the relationship between statistical significance and effect size was illustrated as a volcano plot in Figure 19. This graph shows the inverse of the p value in logarithmic scale as a function of effect size. In this case, the p values are computed using regular linear model and therefore are similar to those generated using a t-Student test.

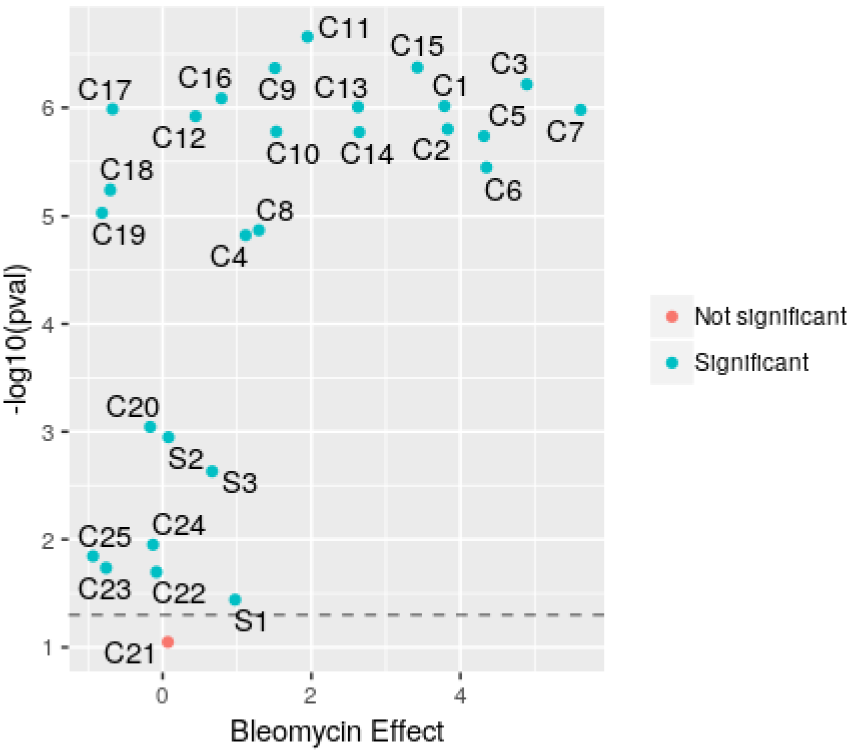

Figure 19. Volcano plot for the 28 mPFR candidates. Dashed line shows p = 0.05, which was used as significance cutoff – also referred in the colors of the points.

Notice that most candidates show positive bleomycin effect, meaning an increase trend following bleomycin exposure. The higher the point in the graph, the lower is the associated p value. Some candidates show negative effect, though. These are mostly based on cluster 9 as “Set2” in equation 1, which showed a late overexpression profile. They also include candidates built from differences between cluster 1/2 (early overexpression) versus cluster 5 (intermediary longitudinal expression profile). As speculated, the greatest effects and lowest p values (upper right corner) were achieved using differences between clusters with early overexpression versus underexpression profiles (e.g. candidates C3 and C7, built from differences between cluster 4 versus cluster 3, and cluster 4 versus cluster 6, respectively – see Figure 6).

As normal distribution was a partially-met assumption for the previous linear model, we wondered whether the fitted values corresponded well to the actual averages. To generally visualize the fitted mPFR performances, we plotted the three-way relationship between statistical significance (inverse of p value in logarithmic scale), effect size (linear model coefficients), and “goodness of fit” (“R squared” from

Pearson's Correlation Coefficient) in Figures 19-20. From the first visualization, one must note that the most significant effects also share the best fits. Also, note the relatively poor performance of the reference candidate, S1. The second plot, known as a "bubble chart", shows that mPFR candidates were satisfactorily fitted for most cases, especially those with the greatest statistical significance and higher effect sizes (greatest -log10(p) values and bleomycin effects - upper right corner). Again, candidates C3 and C7 appear as top performing.

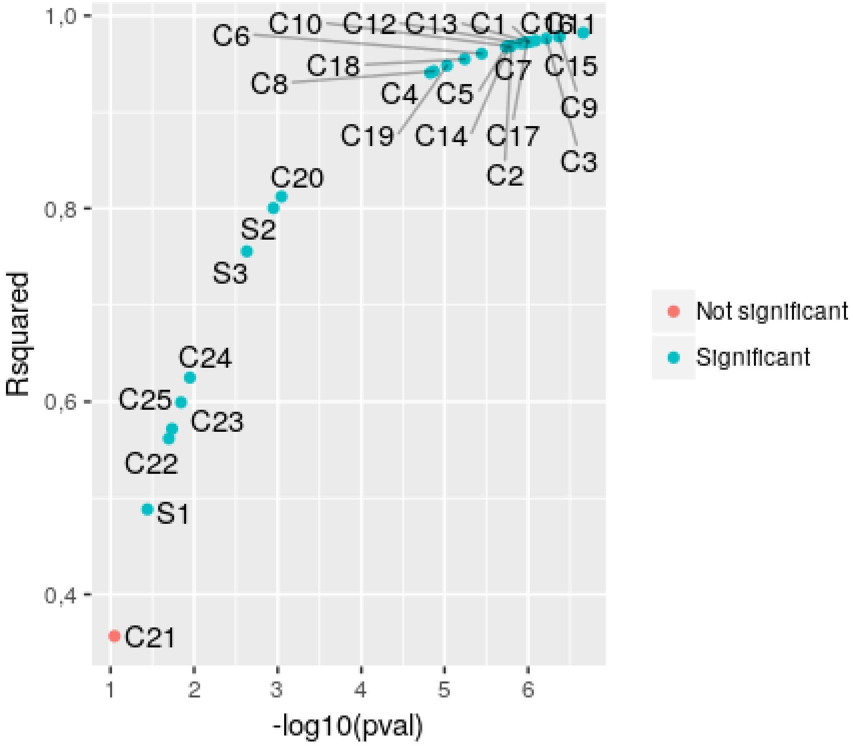

Figure 20. Relationship between statistical significance and goodness of fit, as measured by linear model p value and Pearson's Correlation coefficient, respectively.

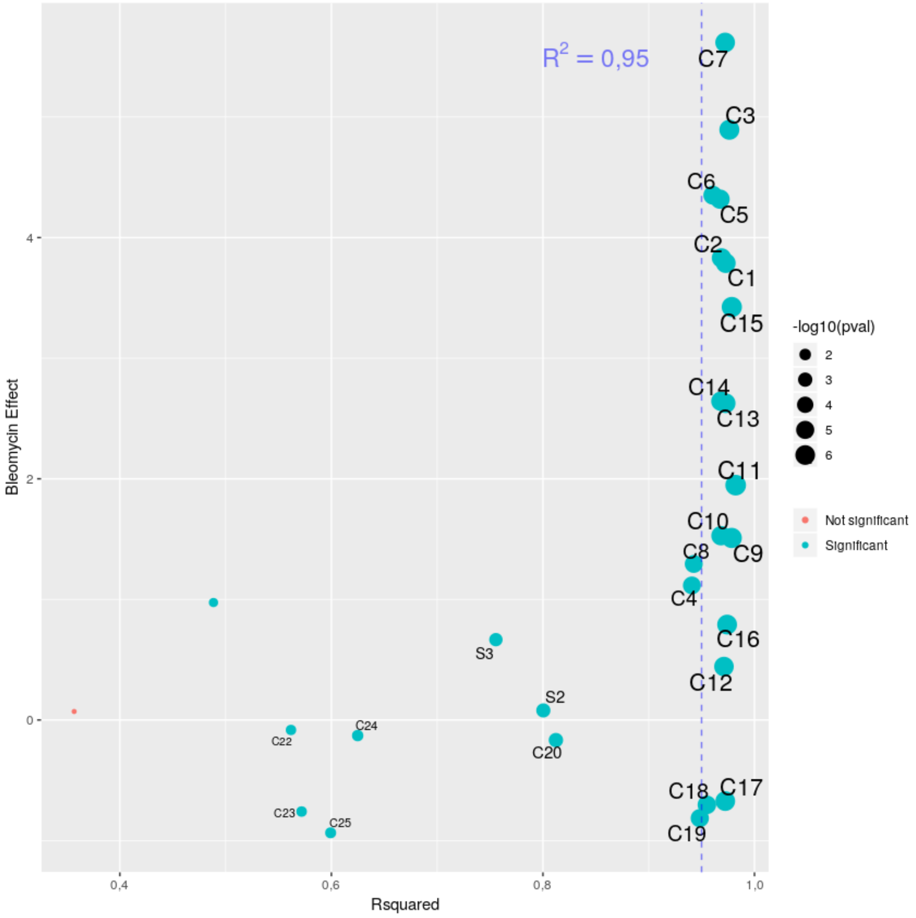

Figure 21. Three-way relationship between effect size, statistical significance, and goodness of fit for all mPFR candidates using Bauer's data. Bubble sizes represent statistical significance, R squared is used as a measure of goodness of fit, and bleomycin effect is the difference in averages between bleomycin- and PBS-treated samples.

4.2.4. Macrophage gene modules as mPFR candidates

Macrophage gene modules produced by Xue and colleagues (2014) using machine learning approach were also used to build mPFR candidates. Consistent with the previous reasoning, we produced candidates with the differences between modules associated with the M1-like activation spectrum versus those related to M2-like spectrum (see Figure 2).

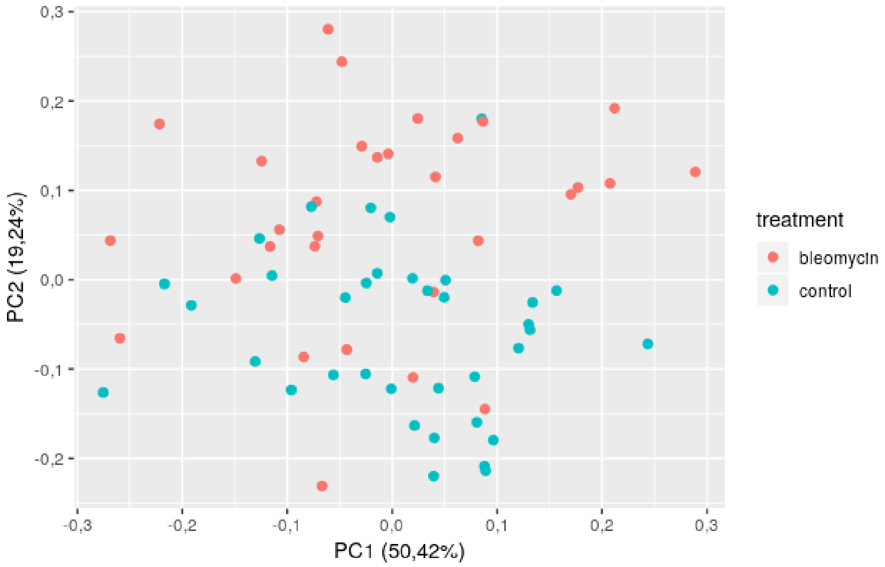

Figure 22. Principal component analysis is not able to separate case and control samples using Xue modules-derived mPFR candidates.

Exploratory data analysis revealed similar distribution as previous mPFR candidates – data not shown. However, PCA failed to separate well bleomycin- and PBS-treated samples based on macrophage modules-derived mPFRs – Figure 21. Still, the mPFR candidate built from module 9 versus module 15 did show statistically significant differences at day 3 and 14 after bleomycin exposure (Wilcoxon rank sum test, p < 0.05). Its very small effect size, though, questions this result's reliability – Figure 22. Fitted linear model showed similar results, with an R2 ~ 0.9 and p value < 0.01 – data not shown. Overall, mPFR candidates derived from macrophage modules were outperformed by those created with maSigPro-derived longitudinal clustering analysis. This is not entirely surprising as the former is basically macrophage-driven, while the latter takes into account gene expression from a wide range of cells. This is also evidence that longitudinal profiles may perform better in generating genomic fingerprints capable of classifying disease-relevant groupings, although such a claim surely needs to be further validated.

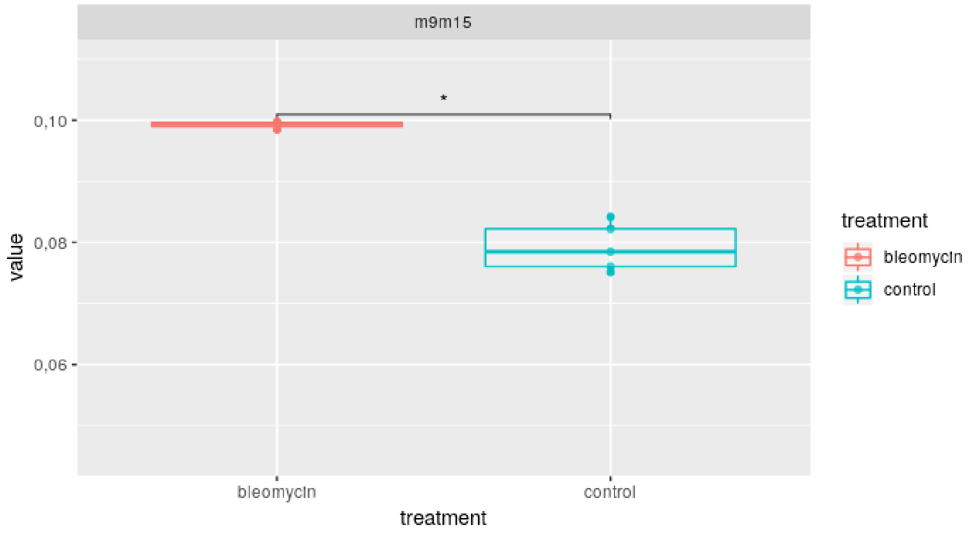

Figure 23. Wilcoxon rank sum test shows significant differences in mPFR candidate, m9m15, between bleomycin and control groups. Data for day 3 after bleomycin exposure. * p < 0.05.

4.3. MODULAR PFR'S AS CLINICAL PREDICTORS – PRELIMINARY RESULTS AND CURRENT CHALLENGES

As mPFR candidates derived from longitudinal cluster analysis showed good performance in segmenting treatment groups in IPF animal model, we further wondered about their performance in human-derived gene expression data. This represented a particularly challenging set of tasks. First, one must recover all orthologs for those genes used to construct mPFR candidates. In a first attempt, out naive exact match algorithm, applied to gene symbols alone, was able to recover 60-90% of the genes comprising each candidate. This is surely not ideal and has major impacts on later results, so a better strategy using official identifiers and fuzzy logic for database querying will probably increase the candidates performance. Mea culpa shall be empathized as one must acknowledge that the programmatic translation of thousands of gene symbols must be anything but trivial. A second challenge includes the interpretation of microarray probes that map to the same gene. In Bauer's dataset, for instance, CCL2 showed over nine probes, all of which with similar gene expression profiles. However, IL13R-a1 also showed more than one correspondent probe, but these were not identical in terms of expression levels. Even though averaging is often advisable – if the data was previously normalized -, this can potentially impact reproducibility of translational assessments. Of course, other major issues with regard to rat-to-human genomic translation also apply. One must not expect to observe identical gene expression profiles for all genes between species, for example. Nevertheless, some level of consistency is expected and this is the basic assumption for genomic studying of animal models.

In order to extract first evidences about the performance of mPFR candidates with humans data, we analyzed microarray gene expression from GSE53845 (DEPIANTO et al., 2015). The dataset is comprised of over 30 samples of IPF patients that underwent biopsy or transplant procedures, plus 8 necropsy-derived control samples. This dataset was chosen given the simplicity of study design. The microarray platform was from Agilent Technologies, the same used in the study from Bauer and colleagues. Raw data was retrieved from GEO database using the GEOquery R/Bioconductor package (HUBER et al., 2015). The same preprocessing and normalization procedures as for Bauer's data were applied.

4.3.1. Exploratory data analysis of DePianto and colleagues data

Principal component analysis was used to assess global trends in original gene expression data. As shown in Figure 23, the first three principal components did show some separation between IPF and control samples, although consistency is not clear. The scaled PCA resulted in most of the variance being explained by the first principal component, and the following components do not differ significantly. This is in accordance with the first PCA performed over Bauer's data, which also showed major between-group overlaps (Figure 4 – please, note color aesthetics are not the same as in the previous plots).

Here, when performing the same PCA with the 28 mPFR candidates previously constructed, the separation between control and IPF samples did not appear to increase significantly, although the spread of data across first and second components did increase sensibly – Figure 24. This is made clear by the scree plot in the figure's lower right panel, which shows the percentage of variance explained by each principal component. Despite the loss of data resultant of poor orthologs translation, this suggests that mPFR candidates might resolve better classification between normal and disease samples in clinical datasets, when comparing with mere normalized gene expression.

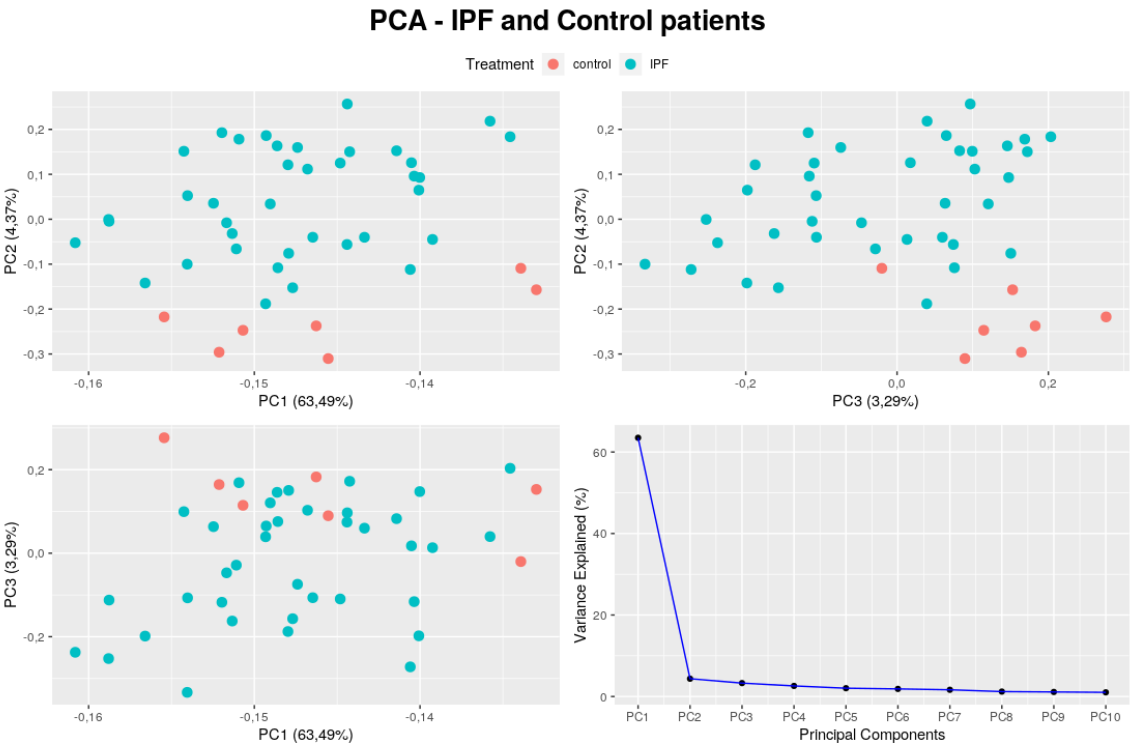

Figure 24. Principal component analysis of DePianto's gene expression data.

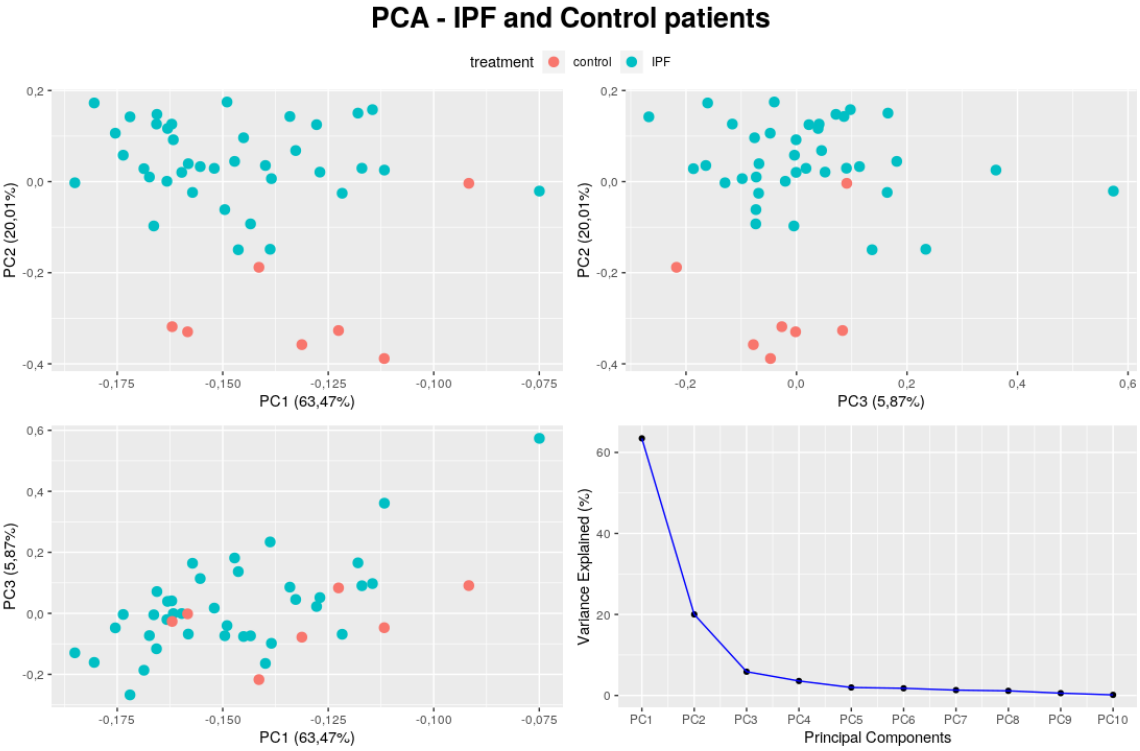

Figure 25. Principal component analysis of mPFR candidates using DePianto's data.

4.3.2. mPFR candidates as prediction variables using humans data

Following the previous workflow, we next assessed the mPFR differences between control and IPF groups. As normal distribution was again partially met (data not shown), Wilcoxon rank sum test was performed. Similar to the results from Bauer's animal model, here we detected statistical significance for several of the mPFR candidates, while the reference (S1) was again non-significant – Figure 25. The same figure as .png file with better resolution and all remaining boxplots are publicly available at github.com/giulianonetto/tcc/rmd-files/Development_files/figure-docx/LastChapter/HUMANS. Notice the major IPF effects over S2 and S3, the ones constructed with semi-arbitrary selection of chemokines, metalloproteinases, interleukins, and other relevant proteins. These were chosen according to their reported relationship with macrophage polarization, but accounting for their clustered profiles in maSigPro longitudinal analyses. This approach seeks to correct for the issues seen in the analysis of macrophage modules from Xue and colleagues (2014). Although macrophages are major players in pulmonary fibrosis (WYNN, 2011), the results from the previous section indicate that overall gene expression profiles need to be taken into account – see Figures 14 and 21.

For instance, the candidate S3 is comprised of CCL2 (MCP1), CCR2, and ILb1 for the Set 1 (see equation 1), while Set 2 included Apolipoprotein E, MMP14, Lipopolysaccharide-biding protein (LBP), and IL1rl1. This suggests that out approach combines recent literature insights with sophisticated multivariate analysis techniques to successfully build a workflow for translational research. Additionally, note that the previously top-performing candidates in animals data, C3 and C7, once again appear as significantly different between IPF patients and controls. Using non-robust linear models, they also show relatively higher Pearson's correlation coefficients and lower p values. However, the goodness of fit was heavily impaired, maybe due to diversity of sample origins – Figure 26. Also, most of Xue's modules showed neither statistical significance or minimally good linear model fits. Wilcoxon tests showed $p < 0.05$ for candidates m8m13, m8m15, m9m15, despite their very small effect sizes (data not shown). As PCA with these macrophage-driven candidates did not show any clear trend, they were dropped from further analyses.

Finally, deeper validation of these results will require bigger datasets and adequacy to software development best practices. Still, the longitudinal assessment of genomic data from animal models seems to hide major potentials for biomarker discovery. Here, it has been developed an analytical workflow for future studies involving innovative longitudinal gene expression analysis, immunopathology, and multivariate statistical analysis successfully applied to translational medicine research. As of these preliminary results, though, it remains to be answered the questions about the actual clinical performance of our 28 modular Macrophage Polarization Factor candidates as IPF biomarkers.

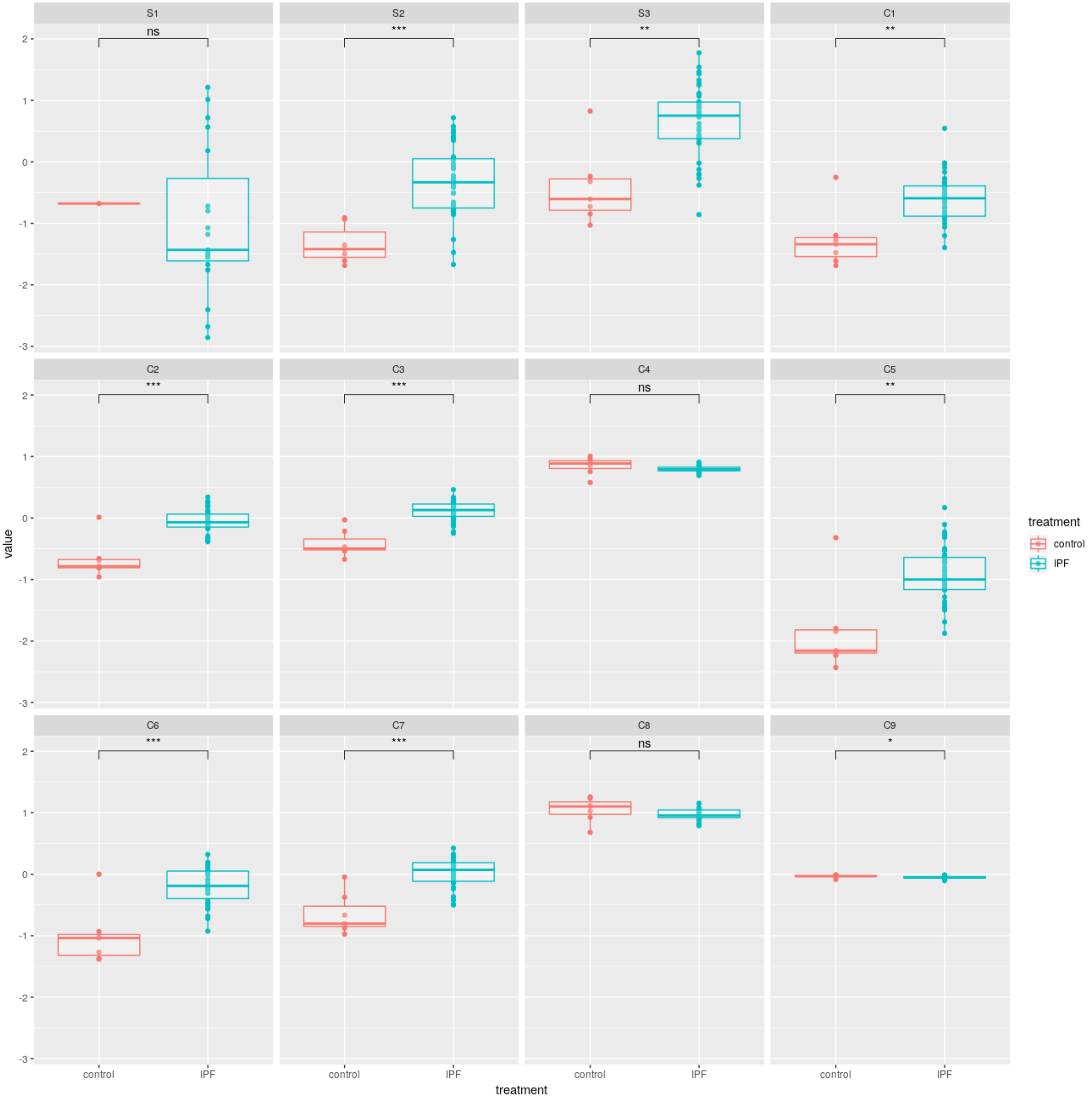

Figure 26. Wilcoxon rank sum test for 12 of the 28 mPFR candidates using clinical data from DePianto and colleagues (2015). * p < 0.05; ** p < 0.01; *** p < 0.001, ns = non-significant.

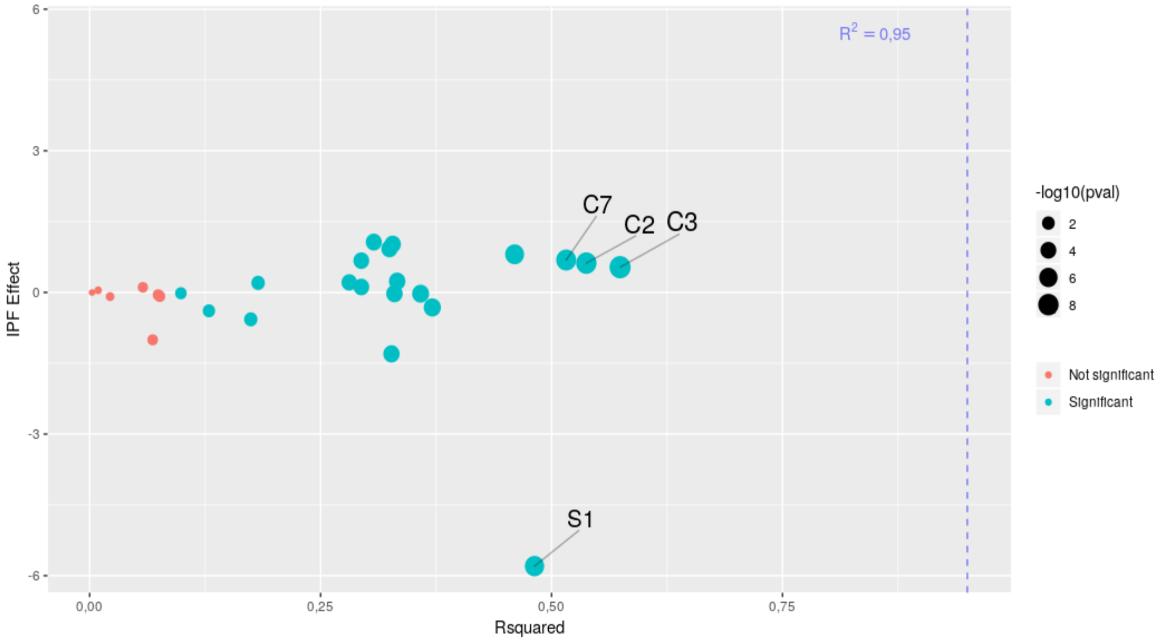

Figure 27. Three-way relationship between effect size, goodness of fit, and statistical significance of mPFR candidates differences between IPF patients and controls. Note the heavily impaired linear model fits.

5. CONCLUSION

Idiopathic Pulmonary Fibrosis is a devastating disease and a current challenge for medical research. Traditional approaches have failed in identifying early diagnosis tools and pharmacological targets that could actually stop disease progression. Functional genomics has appeared in the past decades as a major source of hope, although diversity of analytical procedures, wet lab platforms, and study designs has been added to the intrinsic complexity of genomic data to ultimately hinder research reproducibility. Software development driven by data science approaches has been the major tool with which bioinformaticians are now struggling to extract meaningful and reliable insights from genomic data, Their work along with traditional bench researchers has yielded synergistic results that impact the whole field of biological sciences.

In this context, here we aimed to characterize temporal gene expression profiles in IPF animal models, build numerical factor kinetically correlated with such profiles, and assess the predictive and descriptive performance of such factor. In order to do so, we reasoned that rather than a picture of gene expression at a certain point in time, instead one would record the entire longitudinal trends. This was achieved through assessment of expression data from bleomycin-treated rats in publicly available data (BAUER et al., 2015). As a set of different movies that are classified into subject groupings, the longitudinal expression trends were hierarchically clustered. As opposed to inferences from single time points, these clusters were expected to better describe the *stories* of bleomycin-treated animals.

Baseline-corrected differences between the constructed clusters were computed and shown to successfully classify original treatment groups, an achievement not met by the actual gene expression data. Both robust and traditional hypothesis tests showed significant differences in most of the thereby built 28 modular Macrophage Polarization Factor candidates. In fact, these candidates were constructed using data from whole lung experiments, and assumed to be correlated with macrophage biology as these are major players in pulmonary fibrosis development. When performing the same workflow over machine learning derived macrophage gene modules, however, this assumption was revealed to be flawed (XUE et al., 2014). Although some mPFR candidates constructed from Xue's modules did show statistical significance, their prediction power was limited when compared to the ones built out of longitudinal clustering.

Finally, preliminary results on clinical applicability of the mPFR candidates is presented. Although of limited performance, partially due to poor orthologs recovering, the mPFR candidates showed significant differences between IPF and control patients, especially when using robust statistics techniques. The top-performing candidates were those derived from clusters of opposed profiles - i.e. early overexpression versus late overexpression or underexpression trends. Although further validation is needed, it is clear that the longitudinal analytical procedures herein presented show promising applicability for biomarker and pharmacological target discovery.

Future analysis must include greater number of datasets, more robust statistical methods, and machine learning classification approaches. Here, we successfully described temporal profiles of gene expression in IPF animal model, built numerical candidates kinetically correlated with IPF development, and demonstrated the promising performance of these candidates as predictive and descriptive factors. Larger datasets will be needed in order to further validate the results herein presented. Still, as a movie rather than the picture, it is the study of time courses that will ultimately capture the genomic features of IPF patients, capable of giving these very people more than just hope, but actual time.

REFERENCES

ABBAS, A. K.; LICHTMAN, A. H.; PILLAI, S. **Cellular and Molecular Immunology**. Traducao. 9. ed. [s.l.]: Elsevier Inc., 2017. p. 565

ALFANO, M. et al. Macrophage polarization at the crossroad between hiv-1 infection and cancer development. **Arteriosclerosis, Thrombosis, and Vascular Biology**, [s. l.], v. 33, n. 6, p. 1145–1152, 2013.

AMUBIEYA, O. et al. Eotaxin-2 in lung tissue is associated with disease severity and progression of IPF. **American Thoracic Society**, [s. l.], v. 2, n. Figure 1, p. C37, 2016.

ANDERS, S.; HUBER, W. Differential expression analysis for sequence count data. **Genome Biology**, [s. l.], v. 11, 2010.

BADDINI-MARTINEZ, J.; PEREIRA, C. A. How many patients with idiopathic pulmonary fibrosis are there in Brazil? **Jornal Brasileiro de Pneumologia**, [s. l.], v. 41, n. 6, p. 560–561, 2015. Available at: <http://www.scielo.br/scielo.php?script=sci{_}arttext{\&}pid=S1806-37132015000600560{\&}lng=en{\&}tlng=en>X

BAITSCH, D. et al. Apolipoprotein E (ApoE) induces anti-inflammatory phenotype in macrophages. **Arteriosclerosis, Thrombosis, and Vascular Biology**, [s. l.], v. 31, n. 5, p. 1160–1168, 2011.

BAKKER, O. B. et al. Integration of multi-omics data and deep phenotyping enables prediction of cytokine responses. **Nature Immunology**, [s. l.], v. 19, n. 7, p. 776–786, 2018.

BAUER, Y. et al. A novel genomic signature with translational significance for human idiopathic pulmonary fibrosis. **American Journal of Respiratory Cell and Molecular Biology**, [s. l.], v. 52, n. 2, p. 217–231, 2015.

BECKER, L. et al. Unique proteomic signatures distinguish macrophages and dendritic cells. **PLoS ONE**, [s. l.], v. 7, n. 3, p. 1–12, 2012.

BECKER, M. et al. Integrated Transcriptomics Establish Macrophage Polarization Signatures and have Potential Applications for Clinical Health and Disease. **Scientific Reports**, [s. l.], v. 5, n. July, p. 1–12, 2015. Available at: <http://dx.doi.org/10.1038/srep13351>X

BHATIA, S. et al. Rapid host defense against Aspergillus fumigatus involves alveolar macrophages with a predominance of alternatively activated phenotype. **PLoS ONE**, [s. l.], v. 6, n. 1, 2011.

BOON, K. et al. Molecular phenotypes distinguish patients with relatively stable from progressive idiopathic pulmonary fibrosis (IPF). **PLoS ONE**, [s. l.], v. 4, n. 4, 2009.

BRAGA, T. T.; AGUDELO, J. S. H.; CAMARA, N. O. S. Macrophages during the fibrotic process: M2 as friend and foe. **Frontiers in Immunology**, [s. l.], v. 6, n. NOV, p. 1–8, 2015.

BROWN, A. W. et al. Outcomes after hospitalization in idiopathic pulmonary fibrosis: A cohort study. **Chest**, [s. l.], v. 147, n. 1, p. 173–179, 2015. Available at: <http://dx.doi.org/10.1378/chest.13-2424>X

BUSCHER, K. et al. Natural variation of macrophage activation as disease-relevant phenotype predictive of inflammation and cancer survival. **Nature Communications**, [s. l.], v. 8, n. May, p. 1–10, 2017. Available at: <http://dx.doi.org/10.1038/ncomms16041>X

BYRNE, A. J.; MAHER, T. M.; LLOYD, C. M. Pulmonary Macrophages: A New Therapeutic Pathway in Fibrosing Lung Disease? **Trends in Molecular Medicine**, [s. l.], v. 22, n. 4, p. 303–316, 2016. Available at: <http://dx.doi.org/10.1016/j.molmed.2016.02.004>X

CAMELO, A. et al. IL-33 , IL-25 , and TSLP induce a distinct phenotypic and activation pro fi le in human type 2 innate lymphoid cells. **Blood Advances**, [s. l.], v. 1, n. 10, p. 577–589, 2017. Available at: <http://www.bloodadvances.org/content/bloodoa/1/10/577.full.pdf?sso-checked=true>X

CHAUSSABEL, D. et al. A Modular Analysis Framework for Blood Genomics Studies: Application to Systemic Lupus Erythematosus. **Immunity**, [s. l.], v. 29, p. 150–164, 2008.

CHEVRIER, S. et al. An Immune Atlas of Clear Cell Renal Cell Carcinoma. **Cell**, [s. l.], v. 169, n. 4, p. 736–749.e18, 2017.

CHHANGAWALA, S. et al. The impact of read length on quantification of differentially expressed genes and splice junction detection. **Genome Biology**, [s. l.], v. 16, n. 1, p. 1–10, 2015. Available at: <http://dx.doi.org/10.1186/s13059-015-0697-y>X

COLLINS, M.; LING, V.; CARRENO, B. M. The B7 family of immune-regulatory ligands. **Genome Biology**, [s. l.], v. 6, n. 6, p. 1–7, 2005.

CONESA, A. et al. maSigPro: A method to identify significantly differential expression profiles in time-course microarray experiments. **Bioinformatics**, [s. l.], v. 22, n. 9, p. 1096–1102, 2006.

COSTA-SILVA, J.; DOMINGUES, D.; LOPES, F. M. RNA-Seq differential expression analysis: An extended review and a software tool. **PLoS ONE**, [s. l.], v. 12, n. 12, p. 1–18, 2017.

COURT, M. et al. Proteomic Signature Reveals Modulation of Human Macrophage Polarization and Functions Under Differing Environmental Oxygen Conditions. **Molecular & Cellular Proteomics**, [s. l.], v. 16, n. 12, p. 2153–2168, 2017. Available at: <http://www.mcponline.org/lookup/doi/10.1074/mcp.RA117.000082>X

CRAIG, V. J. et al. Matrix metalloproteinases as therapeutic targets for idiopathic pulmonary fibrosis. **American Journal of Respiratory Cell and Molecular Biology**, [s. l.], v. 53, n. 5, p. 585–600, 2015.

DANCER, R. C. A.; WOOD, A. M.; THICKETT, D. R. Metalloproteinases in idiopathic pulmonary fibrosis. **European Respiratory Journal**, [s. l.], v. 38, n. 6, p. 1461–1467, 2011.

DELLA LATTA, V. et al. Bleomycin in the setting of lung fibrosis induction: From biological mechanisms to counteractions. **Pharmacological Research**, [s. l.], v. 97, p. 122–130, 2015. Available at: <http://dx.doi.org/10.1016/j.phrs.2015.04.012>X

DELVES, P. J. et al. **Roitt’s Essential Immunology**. Traducao. 13. ed. [s.l.]: John Wiley & Sons, 2017. v. 136p. 23–42

DEPIANTO, D. J. et al. Heterogeneous gene expression signatures correspond to distinct lung pathologies and biomarkers of disease severity in idiopathic pulmonary fibrosis. **Thorax**, [s. l.], v. 70, n. 1, 2015.

FERNANDEZ, I. E.; EICKELBERG, O. The Impact of TGF-$^{FDR=0.05}$ on Lung Fibrosis. **Proceedings of the American Thoracic Society**, [s. l.], v. 9, n. 3, p. 111–116, 2012. Available at: <http://www.atsjournals.org/doi/abs/10.1513/pats.201203-023AW>X

FONSECA, G. J.; SEIDMAN, J. S.; GLASS, C. K. Genome wide approaches to defining macrophage indentity and function. **Microbiol Spectr.**, [s. l.], v. 4, n. 5, 2017.

GAUTIER, L. et al. Affy - Analysis of Affymetrix GeneChip data at the probe level. **Bioinformatics**, [s. l.], v. 20, n. 3, p. 307–315, 2004.

GENTLEMAN, R. C. et al. Bioconductor : open software development for computational biology and bioinformatics. **Genome Biology**, [s. l.], v. 5, n. 10, 2004.

GIESECK, R. L.; WILSON, M. S.; WYNN, T. A. Type 2 immunity in tissue repair and fibrosis. **Nature Reviews Immunology**, [s. l.], v. 18, n. 1, p. 62–76, 2017. Available at: <http://dx.doi.org/10.1038/nri.2017.90>X

GINHOUX, F. et al. New insights into the multidimensional concept of macrophage ontogeny, activation and function. **Nature Immunology**, [s. l.], v. 17, n. 1, p. 34–40, 2016.

GURCZYNSKI, S. J.; MOORE, B. B. IL-17 in the Lung: The Good, The Bad, and The Ugly 1. **Am J Physiol Lung Cell Mol Physiol**, [s. l.], n. 734, 2017.

HAMS, E.; BERMINGHAM, R.; FALLON, P. G. Macrophage and innate lymphoid cell interplay in the genesis of fibrosis. **Frontiers in Immunology**, [s. l.], v. 6, n. NOV, p. 1–11, 2015.

HAN, A. et al. Linking T-cell receptor sequence to functional phenotype at the single-cell level. **Nature Biotechnology**, [s. l.], v. 32, n. 7, p. 684–692, 2014.

HARTL, D. et al. A role for MCP-1/CCR2 in interstitial lung disease in children. **Respiratory Research**, [s. l.], v. 6, p. 1–12, 2005.

HEJBLUM, B. P.; SKINNER, J.; THIÉBAUT, R. Time-Course Gene Set Analysis for Longitudinal Gene Expression Data. **PLoS Computational Biology**, [s. l.], v. 11, n. 6, p. 1–21, 2015.

HUBER, W. et al. Orchestrating high-throughput genomic analysis with Bioconductor. **Nature Methods**, [s. l.], v. 12, n. 2, p. 115–121, 2015. Available at: <http://dx.doi.org/10.1038/nmeth.3252>X

HUBER, W. et al. Variance Stabilization Applied to Microarray Data Calibration and to the Quantification of Differential Expression. **Bioin**, [s. l.], v. 18, p. 1–14, 2002.

HUSSELL, T.; BELL, T. J. Alveolar macrophages: plasticity in a tissue-specific context. **Nature Reviews Immunology**, [s. l.], v. 14, n. 2, p. 81–93, 2014. Available at: <http://www.nature.com/doifinder/10.1038/nri3600>X

HUTCHINSON, J. et al. Global incidence and mortality of idiopathic pulmonary fibrosis: A systematic review. **European Respiratory Journal**, [s. l.], v. 46, n. 3, p. 795–806, 2015. Available at: <http://dx.doi.org/10.1183/09031936.00185114>X

INOMATA, M. et al. Pirfenidone inhibits fibrocyte accumulation in the lungs in bleomycin-induced murine pulmonary fibrosis. **Respiratory Research**, [s. l.], v. 15, n. 1, p. 1–14, 2014.

IRIZARRY, R. A. et al. Exploration , Normalization , and Summaries of High Density Oligonucleotide Array Probe Level Data. [s. l.], n. June, p. 249–264, 2003.

IRIZARRY, R. A.; LOVE, M. I. **Data Analysis for the Life Sciences**. Traducao. 1. ed. [s.l.]: Leanpub, 2015. p. 1–30 Available at: <http://leanpub.com/dataanalysisforthelifesciences{\%}0Ahttp://ebooks.cambridge.org/ref/id/CBO9781107415324A009>X

IRIZARRY, R. A.; WU, Z.; JAFFEE, H. A. Comparison of Affymetrix GeneChip expression measures. **Bioinformatics**, [s. l.], v. 22, n. 7, p. 789–794, 2006.

ISHIDA, Y. et al. Essential roles of the CC chemokine ligand 3-CC chemokine receptor 5 axis in bleomycin-induced pulmonary fibrosis through regulation of macrophage and fibrocyte infiltration. **American Journal of Pathology**, [s. l.], v. 170, n. 3, p. 843–854, 2007.

ITALIANI, P.; BORASCHI, D. From monocytes to M1/M2 macrophages: Phenotypical vs. functional differentiation. **Frontiers in Immunology**, [s. l.], v. 5, n. OCT, p. 1–22, 2014.

IWASAKI, A.; FOXMAN, E. F.; MOLONY, R. D. Early local immune defences in the respiratory tract. **Nature Reviews Immunology**, [s. l.], v. 17, n. 1, p. 7–20, 2016. Available at: <http://www.nature.com/doifinder/10.1038/nri.2016.117>X

IZBICKI, G. et al. Time course of bleomycin-induced lung fibrosis. **International Journal of Experimental Pathology**, [s. l.], v. 83, n. 3, p. 111–119, 2002.

KABACOFF, R. I. **R in Action: Data analysis and graphics with R**. Traducao. 2. ed. Shelter Island: Manning, 2015.

KALLURI, R.; WEINBERG, R. A. The basics of epithelial-mesenchymal transition.

The Journal of Clinical Investigation, [s. l.], v. 119, n. 6, 2009.

KAMAL, A. H. M. et al. Inflammatory Proteomic Network Analysis of Statin-treated and Lipopolysaccharide-activated Macrophages. **Scientific Reports**, [s. l.], v. 8, n. 1, p. 1–13, 2018. Available at: <http://dx.doi.org/10.1038/s41598-017-18533-1>X

KARO-ATAR, D. et al. A protective role for IL-13 receptor 1 in bleomycin-induced pulmonary injury and repair. **Mucosal Immunology**, [s. l.], v. 9, n. 1, p. 240–253, 2016. Available at: <http://dx.doi.org/10.1038/mi.2015.56>X

KAUFFMANN, A.; GENTLEMAN, R.; HUBER, W. arrayQualityMetrics - A bioconductor package for quality assessment of microarray data. **Bioinformatics**, [s. l.], v. 25, n. 3, p. 415–416, 2009.

KENDALL, R. T.; FEGHALI-BOSTWICK, C. A. Fibroblasts in fibrosis: Novel roles and mediators. **Frontiers in Pharmacology**, [s. l.], v. 5 MAY, n. May, p. 1–13, 2014.

KIDD, B. A. et al. Unifying immunology with informatics and multiscale biology. **Nature Immunology**, [s. l.], v. 5, n. February, p. 118–127, 2014.

KOLAHIAN, S. et al. Immune mechanisms in pulmonary fibrosis. **American Journal of Respiratory Cell and Molecular Biology**, [s. l.], v. 55, n. 3, p. 309–322, 2016.

KULKARNI, T. et al. Alveolar Epithelial Disintegrity in Pulmonary Fibrosis. **American Journal of Physiology - Lung Cellular and Molecular Physiology**, [s. l.], v. 0006, p. ajplung.00115.2016, 2016. Available at: <http://ajplung.physiology.org/lookup/doi/10.1152/ajplung.00115.2016>X

KUROWSKA-STOLARSKA, M. et al. IL-33 Amplifies the Polarization of Alternatively Activated Macrophages That Contribute to Airway Inflammation. **The Journal of Immunology**, [s. l.], v. 183, n. 10, p. 6469–6477, 2009. Available at: <http://www.jimmunol.org/cgi/doi/10.4049/jimmunol.0901575>X

LANDI, C. et al. A system biology study of BALF from patients affected by idiopathic pulmonary fibrosis (IPF) and healthy controls. **Proteomics - Clinical Applications**, [s. l.], v. 8, n. 11-12, p. 932–950, 2014.

LANGFELDER, P.; HORVATH, S. WGCNA: An R package for weighted correlation network analysis. **BMC Bioinformatics**, [s. l.], v. 9, 2008.

LAW, C. W. et al. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. **Genome Biology**, [s. l.], v. 15, 2014.

LEDERER, D. J.; MARTINEZ, F. J. Idiopathic pulmonary fibrosis. **The New England Journal of Medicine**, [s. l.], v. 378, p. 1811–1823, 2018.

LEE, J. et al. Bronchoalveolar lavage (BAL) cells in idiopathic pulmonary fibrosis express a complex pro-inflammatory, pro-repair, angiogenic activation pattern, likely associated with macrophage iron accumulation. **PLoS ONE**, [s. l.], v. 13, n. 4, p. 1–15, 2018.

LI, D. et al. IL-33 promotes ST2-dependent lung fibrosis by the induction of alternatively activated macrophages and innate lymphoid cells in mice. **Journal of Allergy and Clinical Immunology**, [s. l.], v. 134, n. 6, p. 1422–1432, 2014.

LIU, G. et al. NetAffix: Affymetrix probesets and annotations. **Nucleic Acids Research**, [s. l.], v. 31, n. 1, p. 82–86, 2003.

LOVE, M. I.; HUBER, W.; ANDERS, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. **Genome Biology**, [s. l.], v. 15, n. 12, p. 1–21, 2014.

LU, Y. et al. Highly multiplexed profiling of single-cell effector functions reveals deep functional heterogeneity in response to pathogenic ligands. **Proceedings of the National Academy of Sciences**, [s. l.], v. 112, n. 7, p. E607–E615, 2015. Available at: <http://www.pnas.org/lookup/doi/10.1073/pnas.1416756112>X

LUMSDEN, R. V. et al. Modulation of pulmonary fibrosis by IL-13R$^{\beta}$2. **American**

Journal of Physiology-Lung Cellular and Molecular Physiology, [s. l.], v. 308, n. 7, p. L710–L718, 2015. Available at: <http://www.physiology.org/doi/10.1152/ajplung.00120.2014>X

LUZINA, I. G. et al. The cytokines of pulmonary fibrosis: Much learned, much more to learn. **Cytokine**, [s. l.], v. 74, n. 1, p. 88–100, 2015. Available at: <http://dx.doi.org/10.1016/j.cyto.2014.11.008>X

MALAVIYA, R. et al. Macrophages and inflammatory mediators in pulmonary injury induced by mustard vesicants. **Ann N Y Acad Sci**, [s. l.], v. 1374, n. 1, p. 168–175, 2016.

MALE, D. et al. **Immunology**. Traducao. 8. ed. [s.l.]: Elsevier, 2013. v. 39p. 63

MARGALIT, A.; KAVANAGH, K. The innate immune response to Aspergillus fumigatus at the alveolar surface. **FEMS Microbiology Reviews**, [s. l.], v. 39, n. 5, p. 670–687, 2015.

MARTINEZ, F. O.; GORDON, S. The M1 and M2 paradigm of macrophage activation: time for reassessment. **F1000Prime Reports**, [s. l.], v. 6, n. March, p. 1–13, 2014. Available at: <http://www.f1000.com/prime/reports/b/6/13>X

METZKER, M. L. Sequencing technologies the next generation. **Nature Reviews Genetics**, [s. l.], v. 11, n. 1, p. 31–46, 2010. Available at: <http://dx.doi.org/10.1038/nrg2626>X

MILGER, K. et al. Pulmonary CCR2+CD4+T cells are immune regulatory and attenuate lung fibrosis development. **Thorax**, [s. l.], v. 72, n. 11, p. 1007–1020, 2017.

MISHARIN, A. V. et al. Flow cytometric analysis of macrophages and dendritic cell subsets in the mouse lung. **Am J Respir Cell Mol Biol**, [s. l.], v. 49, n. 4, p. 503–510, 2013. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/23672262>X

MISHARIN, A. V. et al. Monocyte-derived alveolar macrophages drive lung fibrosis and persist in the lung over the life span. **The Journal of Experimental Medicine**, [s. l.], v. 214, n. 8, 2017.

MOELLER, A. et al. The bleomycin animal model: a useful tool to investigate treatment options for idiopathic pulmonary fibrosis? Antje. **The International Journal of Biochemistry & Cell Biology**, [s. l.], v. 40, n. 3, p. 362–382, 2008.

MOLYNEAUX, P. L. et al. Host-microbial interactions in idiopathic pulmonary fibrosis. **American Journal of Respiratory and Critical Care Medicine**, [s. l.], v. 195, n. 12, p. 1640–1650, 2017.

MOORE, B. B. et al. The role of CCL12 in the recruitment of fibrocytes and lung fibrosis. **American Journal of Respiratory Cell and Molecular Biology**, [s. l.], v. 35, n. 2, p. 175–181, 2006.

MURRAY, P. J. Macrophage Polarization. **Annual Review of Physiology**, [s. l.], v. 79, n. 1, p. 541–566, 2017. Available at: <http://www.annualreviews.org/doi/10.1146/annurev-physiol-022516-034339>X

MURRAY, P. J. et al. Macrophage Activation and Polarization: Nomenclature and Experimental Guidelines. **Immunity**, [s. l.], v. 41, n. 1, p. 14–20, 2014. Available at: <http://dx.doi.org/10.1016/j.immuni.2014.06.008>X

NAIKAWADI, R. P. et al. Telomere dysfunction in alveolar epithelial cells causes lung remodeling and fibrosis. **JCI Insight**, [s. l.], v. 1, n. 14, p. 2–12, 2016. Available at: <https://insight.jci.org/articles/view/86704>X

NEU, K. E. et al. Single-Cell Genomics : Approaches and Utility in Immunology. **Trends in Immunology**, [s. l.], v. 38, n. 2, p. 140–149, 2017. Available at: <http://dx.doi.org/10.1016/j.it.2016.12.001>X

NIE, Y. et al. AKT2 Regulates Pulmonary Inflammation and Fibrosis via Modulating Macrophage Activation. **The Journal of Immunology**, [s. l.], p. 1601503, 2017.

Available at: <http://www.ncbi.nlm.nih.gov/pubmed/28455433{\%}0Ahttp://www.ncbi.nlm.nih.gov/pubmed/28455433>X

PARDO, A. et al. Role of matrix metalloproteinases in the pathogenesis of idiopathic pulmonary fibrosis. **Respiratory Research**, [s. l.], v. 17, n. 1, p. 23, 2016. Available at: <http://respiratory-research.com/content/17/1/23>X

PELJTO, A. L. et al. Association Between the MUC5B Promoter Polymorphism and Survival in Patients With Idiopathic Pulmonary Fibrosis. **JAMA - Journal of the American Medical Association**, [s. l.], v. 309, n. 21, p. 2232–2239, 2013.

PENG, R. et al. Bleomycin Induces Molecular Changes Directly Relevant to Idiopathic Pulmonary Fibrosis: A Model for “Active” Disease. **PLoS ONE**, [s. l.], v. 8, n. 4, 2013.

PEVSNER, J. **Bioinformatics and Functional Genomics**. Traducao. 3. ed. Chichester: John Wiley & Sons, 2015. Available at: <https://academic.oup.com/bfg/article-lookup/doi/10.1093/bfgp/3.2.187>X

QUACKENBUSH, J. Microarray data normalization and transformation. **Nature Genetics**, [s. l.], v. 32, n. 4S, p. 496–501, 2002.

QUEREDA, J. J. et al. Bacteriocin from epidemic <i>Listeria</i> strains alters the host intestinal microbiota to favor infection. **Proceedings of the National Academy of Sciences**, [s. l.], v. 113, n. 20, p. 5706–5711, 2016. Available at: <http://www.pnas.org/lookup/doi/10.1073/pnas.1523899113>X

RAGHU, G. et al. CC-chemokine ligand 2 inhibition in idiopathic pulmonary fibrosis: A phase 2 trial of carlumab. **European Respiratory Journal**, [s. l.], v. 46, n. 6, p. 1740–1750, 2015a. Available at: <http://dx.doi.org/10.1183/13993003.01558-2014>X

RAGHU, G. et al. An official ATS/ERS/JRS/ALAT clinical practice guideline:

Treatment of idiopathic pulmonary fibrosis: An update of the 2011 clinical practice guideline. **American Journal of Respiratory and Critical Care Medicine**, [s. l.], v. 192, n. 2, p. e3–e19, 2015b.

RAMASAMY, A. et al. Key issues in conducting a meta-analysis of gene expression microarray datasets. **PLoS Medicine**, [s. l.], v. 5, n. 9, p. 1320–1332, 2008.

RICHELDI, L.; COLLARD, H. R.; JONES, M. G. Idiopathic pulmonary fibrosis. **Nature Reviews Disease Primers**, [s. l.], v. 3, p. 1941–1952, 2017. Available at: <http://dx.doi.org/10.1038/nrdp.2017.74>X

RITCHIE, M. E. et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. **Nucleic Acids Research**, [s. l.], v. 43, n. 7, p. e47, 2015.

ROBBINS, L. Idiopathic Pulmonary Fibrosis : Roentgenologic Findinqs ’. **Radiology**, [s. l.], p. 459–467, 1948.

RUFINO, R. et al. Incidence And Mortality Of Interstitial Pulmonary Fibrosis In Brazil. **American Journal of Respiratory and Critical Care Medicine**, [s. l.], v. 33, n. 187, p. A1458, 2013.

SAELENS, W.; CANNOODT, R.; SAEYS, Y. A comprehensive evaluation of module detection methods for gene expression data. **Nature Communications**, [s. l.], v. 9, n. 1, 2018.

SAHIN, H.; WASMUTH, H. E. Chemokines in tissue fibrosis. **Acta Mathematica Hungarica**, [s. l.], p. 1041–1048 Contents, 2013. Available at: <http://dx.doi.org/10.1016/j.bbadis.2012.11.004>X

SCRUCCA, L. et al. mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models. **R Journal**, [s. l.], v. 8, n. 1, p. 289–317, 2016.

SELA-PASSWELL, N. et al. Antibodies targeting the catalytic zinc complex of activated matrix metalloproteinases show therapeutic potential. **Nature Medicine**, [s. l.], v. 18, n. 1, p. 143–147, 2012. Available at: <http://dx.doi.org/10.1038/nm.2582>X

SINHA, P. P. **Bioinformatics with R**. Traducao. Birmingham: Packt Publishing, 2014.

SONG, J. W. et al. **Acute exacerbation of idiopathic pulmonary fibrosis: incidence, risk factors and outcome.** Traducao. [s.l: s.n.]. v. 37p. 356–363 Available at: <http://www.ncbi.nlm.nih.gov/pubmed/20595144>X

STABLES, M. J. et al. Transcriptomic analyses of murine resolution-phase macrophages. **Blood**, [s. l.], v. 118, n. 26, p. 192–209, 2011.

SYRBU, S.; THRALL, R. S.; SMILOWITZ, H. M. Sequential appearance of inflammatory mediators in rat bronchoalveolar lavage fluid after oleic acid-induced lung injury. **Experimental Lung Research**, [s. l.], p. 33–49, 1996.

TARASOVA, N. **Establishing a proteomics-based monocyte assay to assess differential innate immune responses**. 2016. Doctoral Thesis - Karolinska Institutet; Karolinska Institutet, Stockholm, 2016. Available at: <http://repositorio.ufpe.br/handle/123456789/14905>X

TARASOVA, N. K. et al. Establishing a proteomics-based monocyte assay to assess differential innate immune activation responses. **Journal of Proteome Research**, [s. l.], v. 15, n. 7, p. 2337–2345, 2016.

TARAZONA, S. et al. Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. **Nucleic Acids Research**, [s. l.], v. 43, n. 21, 2015.

TARIQUE, A. A. et al. Phenotypic,functional,and plasticity features of classical and alternatively activated human macrophages. **American Journal of Respiratory Cell and Molecular Biology**, [s. l.], v. 53, n. 5, p. 676–688, 2015.

TEAM, R. C. **R: A Language And Environment For Statistical Computing**ViennaR

Foundation for Statistical Computing,, 2018. Available at: <https://www.r-project.org/.>X

VENOSA, A. et al. Characterization of distinct macrophage subpopulations during nitrogen mustard-induced lung injury and fibrosis. **American Journal of Respiratory Cell and Molecular Biology**, [s. l.], v. 54, n. 3, p. 436–446, 2016.

VICTORA, C. **Cesar Victora - Google Scholar Citations**, 2018. Available at: <scholar.google.com.br/citations?user=TdRzT8kAAAAJ{\&}hl=en>. Acesso em: 6 nov. 2018.X

WALSH, C. et al. Microarray Meta-Analysis and Cross-Platform Normalization: Integrative Genomics for Robust Biomarker Discovery. **Microarrays**, [s. l.], v. 4, n. 3, p. 389–406, 2015. Available at: <http://www.mdpi.com/2076-3905/4/3/389/>X

WANG, E.; MARINCOLA, F. M. Bottom Up: A Modular View of Immunology. **Immunity**, [s. l.], v. 29, n. 1, p. 9–11, 2008.

WERMUTH, P. J.; JIMENEZ, S. A. The significance of macrophage polarization subtypes for animal models of tissue fibrosis and human fibrotic diseases. **Clinical and Translational Medicine**, [s. l.], v. 4, n. 1, p. 2, 2015. Available at: <http://www.clintransmed.com/content/4/1/2>X

WICKHAM, H. ggplot2: Elegant Graphics for Data Analysis. [s. l.], 2015.

WILLIAMSON, J. D.; SADOFSKY, L. R.; HART, S. P. The pathogenesis of bleomycin-induced lung injury in animals and its applicability to human idiopathic pulmonary fibrosis. **Experimental Lung Research**, [s. l.], v. 41, n. 2, p. 57–73, 2014.

WOLTERS, P. J. et al. Time for a change: is idiopathic pulmonary fibrosis still idiopathic and only fibrotic? **The Lancet Respiratory Medicine**, [s. l.], v. 6, n. 2, p. 138–153, 2018. Available at: <http://dx.doi.org/10.1016/S2213-2600(18)30007-9>X

WYNN, T. A. Integrating mechanisms of pulmonary fibrosis. **The Journal of**

Experimental Medicine, [s. l.], v. 208, n. 7, p. 1339–1350, 2011. Available at: <http://www.jem.org/lookup/doi/10.1084/jem.20110551>X

WYNN, T. A.; VANNELLA, K. M. Macrophages in Tissue Repair, Regeneration, and Fibrosis. **Immunity**, [s. l.], v. 44, n. 3, p. 450–462, 2016. Available at: <http://dx.doi.org/10.1016/j.immuni.2016.02.015>X

WYNN, T.; BARRON, L. Macrophages: Master Regulators of Inflammation and Fibrosis. **Semin Liver Dis.**, [s. l.], v. 30, n. 3, p. 245–257, 2010.

XUE, J. et al. Transcriptome-Based Network Analysis Reveals a Spectrum Model of Human Macrophage Activation. **Immunity**, [s. l.], v. 40, n. 2, p. 274–288, 2014. Available at: <http://dx.doi.org/10.1016/j.immuni.2014.01.006>X

YAO, X. et al. Emerging roles of apolipoprotein e and apolipoprotein A-I in the pathogenesis and treatment of lung disease. **American Journal of Respiratory Cell and Molecular Biology**, [s. l.], v. 55, n. 2, p. 159–169, 2016.

YOGO, Y. et al. Macrophage derived chemokine (CCL22), thymus and activation-regulated chemokine (CCL17), and CCR4 in idiopathic pulmonary fibrosis. **Respiratory Research**, [s. l.], v. 10, n. February, 2009.

YU, Y.-R. A. et al. Flow Cytometric Analysis of Myeloid Cells in Human Blood, Bronchoalveolar Lavage, and Lung Tissues. **American Journal of Respiratory Cell and Molecular Biology**, [s. l.], v. 54, n. 1, p. 13–24, 2016. Available at: <http://www.atsjournals.org/doi/10.1165/rcmb.2015-0146OC>X