INAUGURAL - DISSERTATION

zur
Erlangung der Doktorwürde
der
Naturwissenschaftlich-Mathematischen Gesamtfakultät
der
Ruprecht - Karls - Universität
Heidelberg

vorgelegt von: Diplom-Biologe Thomas Wolf
aus: Hanau, Deutschland

Molecular Bronchiolitis Obliterans Syndrome Risk Monitoring: A Systems-Based Approach

1. Gutachter: PD. Dr. Rainer König
2. Gutachter: PD. Dr. Nils von Neuhoff

Tag der mündlichen Prüfung:......................

Eidesstattliche Erklärung

Hiermit erkläre ich, Thomas Wolf, an Eides Statt, dass ich die an der Universität Heidelberg vorgelegte Dissertation selbständig und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Die aus fremden Quellen übernommenen Gedanken sind als solche kenntlich gemacht. Die Arbeit wurde bisher in gleicher oder ähnlicher Weise keiner anderen Prüfungsbehörde vorgelegt und auch nicht veröffentlicht.

Heidelberg, den 04. Juni 2012

(Thomas Wolf)

Abstract

The combination of high throughput omics (i.e. genomics or proteomics) and machine learning offers new possibilities for clinical diagnostics and the detection of biomarkers. One disease for which no reliable prognostic marker has been found yet is bronchiolitis obliterans (BO), a clinical manifestation of chronic rejection after lung transplantation. BO is the major limiting factor for long-term survival after lung transplantation, and manifests as a chronic bronchiolar inflammation accompanied by progressive sub-mucosal fibrosis leading to gradual obliteration of the bronchiolar lumen. The resulting reduction in forced expiratory volume per second (FEV_1) is defined as the bronchiolitis obliterans syndrome (BOS). As chronic lung transplant failure occurs more frequently than in other organ transplants, molecular markers for early BO and BOS detection are urgently required to adapt the patients immunosuppressive regimen when airway damage is minimal. To achieve this goal, gene expression in bronchial epithelial cells (microarray anaylsis) and on the proteome level in bronchoalveolar lavage fluid (BALF)(mass spectrometry profiling) were monitored. Analysis of the obtained data sets was performed using novel and established methods from the fields of machine learning and statistics. This thesis also introduces a novel clustering algorithm. In the analysis of gene expression microarrays one problem is the unsupervised discovery of stable and biologically relevant patient subgroups. To this end I developed a novel clustering algorithm. This algorithm focuses on the discovery of a set of patient clusters defined by the consistent up- and down-regulation of a subset of genes. Assessment of cluster stability is done using a bootstrap resampling scheme. This makes it possible to rank the genes in accordance with their clusterwise importance. The algorithm was applied to a publicly available B-cell lymphoma microarray data set and compared to other commonly used clustering algorithms.

Zusammenfassung

Die Kombination aus Hochdurchsatzverfahren und Maschinellem Lernen eröffnet neue Möglichkeiten im Bereich der Biomarker basierten klinischen Diagnostik. Eine der Krankheiten für die noch kein verlässlicher prognostischer Marker gefunden wurde ist Bronchiolitis Obliterans (BO). Hierbei handelt es sich um eine klinische Manifestation der chronischen Transplantatabstoßung nach Lungentransplantation. Bronchiolitis Obliterans, der wichtigste limitierende Faktor für das langfristige Überleben lungentransplantierter Patienten, manifestiert sich als eine chronische Entzündung. Die begleitende progressive bronchioläre submukotische Fibrose führt zu einer Obstruktion des bronchiolären Lumens. Die daraus resultierende Reduktion der forcierten exspiratorischen Einsekundenkapazität (FEV_1) ist als Bronchiolitis Obliterans Syndrom (BOS) definiert. Chronisches Transplantatversagen tritt bei Lungentransplantaten deutlich häufiger auf als in anderen Organen. Deswegen sind molekulare Marker zur Früherkennung dringend erforderlich. Eine solche Früherkennung würde eine Anpassung der immunsuppressiven Therapie ermöglichen, wenn der Schaden der Atemwege noch gering ist. Um dieses Ziel zu erreichen wurde die Genexpression in bronchialen Epithelzellen (Microarray Analyse) und das Proteom der bronchoalveolären Flüssigkeit (BAL) (Massenspektrometrie) untersucht. Die Analyse der dadurch erhaltenen Datensätze erfolgte mittels etablierter und neuartiger Methoden aus den Bereichen Maschinelles Lernen und Statistik. Desweiteren wurde im Rahmen der vorligenden Dissertation ein neuer Clustering-Algorithmus entwickelt. Ein Problem bei der Analyse von Genexpressions Daten ist die Entdeckung stabiler und biologisch relevanter Patienten-Untergruppen. Zu diesem Zweck entwickelte ich einen Clustering-Algorithmus zur Entdeckung von Patienten Untergruppen, die durch die hoch- und herunterregulierung einer Gengruppe definiert sind. Die Bewertung der Stabilität einer Gruppe von Genen erfolgt unter Verwendung eines Bootstrap-Resampling Ansatzes. Dieser Ansatz macht es auch möglich Gene nach ihrer Bedeutung für die jeweiligen Patienten Cluster zu ordnen. Der Algorithmus wurde an einem öffentlich zugänglichen B-Zell-Lymphom Microarray Datensatz getestet, und mit anderen häufig verwendeten Clustering-Algorithmen verglichen.

Danksagung

An dieser Stelle möchte ich mich bei allen Menschen bedanken, die mich auf dem Weg begleitet und zum Gelingen dieser Doktorarbeit beigetragen haben. Als erstes möchte ich Prof. Dr. Roland Eils und Prof. Dr. Brigitte Schlegelberger für das neuartige Thema und die Bereitstellung des Arbeitsplatzes danken. Meinen Betreuern Dr. Marc Zapatka, PD Dr. Nils von Neuhoff und Dr. Benedikt Brors danke ich für die Betreuung während der gesamten Bearbeitungszeit. Die konstruktiven und richtungsweisenden Vorschläge habe ich gerne angenommen und sie stets als hilfreich empfunden. PD Dr. Rainer König, PD Dr. Nils von Neuhoff, PD Dr. Ralf Bischoff und Prof. Dr. Phillip Beckhove danke ich dafür Teil meines Prüfungskomitees zu sein. Desweiteren möchte ich mit bei allen Freunden und Arbeitskollegen bedanken, die durch ihre moralische Unterstützung und stetige Hilfsbereitschaft diese Arbeit mit ermöglicht haben. Mein Dank gilt insbesondere der Forschungsgruppe Computational Oncology am Deutschen Krebsforschungszentrum und der Zell- und Molekularpathologie der Medizinischen Hochschule Hannover. Für die Zusammenarbeit an den Projekten meiner Doktorarbeit möchte ich Tonio Oumeraci, Esteban Czwan, Dr. Britta Skawran und Marlies Eilers danken. Karl-Heinz Gross und Dr. Winfried Hofmann danke ich für die Unterstützung in Computerfragen. Einen recht herzlichen Dank auch an meine Mutter und meine Großmutter, die mich durch die Zeit der Promotion begleitet und unterstützt haben.

Contents

Chapter 1

Introduction

1.1 Motivation

1.1.1 Integrative Analysis of the Bronchiolitis Obliterans Syndrome

The combination of high throughput omics (i.e. genomics or proteomics) and machine learning offers new possibilities for clinical diagnostics and the detection of biomarkers [1,2,3]. One disease for which no reliable prognostic marker has been found yet is bronchiolitis obliterans (BO) [4], a clinical manifestation of chronic rejection after lung transplantation (LTx). BO is the major limiting factor for long-term survival after lung transplantation [5,6,7], and manifests as a chronic bronchiolar inflammation accompanied by progressive sub-mucosal fibrosis leading to gradual obliteration of the bronchiolar lumen. The resulting reduction in forced expiratory volume per second (FEV_1) is defined as the bronchiolitis obliterans syndrome (BOS). As chronic lung transplant failure occurs more frequently than in other organ transplants [4], molecular markers for early BO/BOS detection are urgently required to adapt the patients immunosuppressive regimen when airway damage is minimal. To achieve this goal, gene expression in bronchial epithelial cells (microarray anaylsis) and the proteome level in bronchoalveolar lavage fluid (BALF)(mass spectrometry profiling) were monitored. Analysis of the obtained data sets was performed using novel and established methods from the fields of machine learning and statistics.

1.1.2 Clustering of High Throughput Data (Block Maxx)

In the analysis of gene expression microarrays one problem is the unsupervised discovery of stable and biologically relevant patient subgroups. To this end I developed a novel clustering algorithm. This algorithm focuses on the discovery of a set of sample clusters defined by the consistent up- and down-regulation of a subset of genes. Assessment of cluster stability is done using a bootstrap resampling scheme, which also makes it possible to rank the genes in accordance with their clusterwise importance. The algorithm was applied to a publicly available B-cell lymphoma microarray data set and compared to other commonly used clustering algorithms.

1.2 Publications

The BOS related results obtained from the proteomic data set, as presented in this thesis, were published in *Transplantation* and entitled Proteomic Bronchiolitis Obliterans Syndrome Risk Monitoring in Lung Transplant Recipients [8]. A manuscript covering the microarray part of the BOS study is currently in preparation. The Block Maxx clustering algorithm described in this thesis was partly inspired by another novel biclustering method of mine, which was published at the *Sixth IEEE International Conference on Data Mining - Workshops (ICDMW'06)* in 2006 as Global Biclustering of Microarray Data [9]. I was first author in reference [9] and shared first author in reference [8]. A description of the Block Maxx algorithm will be prepared as a manuscript. I was also involved in the following publications: Proteomic analysis of field cancerization in pharynx and oesophagus: a prospective pilot study published in *The Journal of pathology* [10], Bronchoalveolar lavage fluid of lung cancer patients: Mapping the uncharted waters using proteomics technology published in *Lung cancer* [11], and Multi-parametric analysis and modeling of relationships between mitochondrial morphology and apoptosis published in *PloS one* [12].

1.3 Thesis Outline

In chapter 1 the biological background, existing computational methods and basic topics are introduced. The focus lies on the bronchiolitis obliterans syndrome (BOS), ensemble methods in biomarker detection and the basic principles of cluster analysis. Additionally, basic concepts in transcriptomics and proteomics are presented. In chapter 2, the methods applied in this thesis are presented, specifically the developed approach to obtain suitable biomarkers for the early detection of BOS and the novel clustering approach referred to as Block Maxx. Chapter 3 depicts the results of this thesis and presents expression processes related to BOS as well as the performance of Block Maxx on a clinical microarray data set. The results are discussed and an outlook is given in chapter 4.

1.4 R Programming Language

R is a highly extensible programming language for statistical computing [13]. The software is open source and allows for data input, calculation and graphical display. Functionality can be added by extension via packages. R is freely available under the GNU General Public License and can be obtained via `http://www.r-project.org/`. Specific R package extensions for biological data analysis are offered by the Bioconductor project [14,15]. Bioconductor is open source and open development. More information can be found under `http://www.bioconductor.org/`. If not specified otherwise, all calculations in this thesis were performed using R.

1.5 The Respiratory System

The lung is the primary respiratory organ in mammals and allows the organism to transport oxygen from the atmosphere into the bloodstream, while enabling the release of carbon dioxide into the atmosphere. The respiratory tract of the lung (Figure 1.1) can be divided into the proximal conductive, nonrespiratory airways (nose, pharynx, larynx, trachea, bronchi, and nonalveolar bronchioles) and the distal respiratory region (respiratory bronchioles, alveolar ducts, and alveoli). The respiratory tract is covered by a heterogeneous epithelium. This pseudostratified epithelium [16] lining the proximal conductive airways is progressively replaced by a simple cuboid cell layer [17] in the more distal airways. The alveoli are then covered by a very thin epithelial lining. Following a progressively subdividing system of bronchi and bronchioles these thin walled airsacks provide an extensive surface, where the gas exchange takes place. The alveolar epithelium [18] additionally provides a barrier that protects the host from inhaled foreign agents and contributes to the maintenance of lung fluid balance. Its regenerative abilities allow for normal cell turnover and restoration of alveolar function after lung injury [19].

1.5.1 Study of the Lung Epithelium

In lung pathology the common mean of studying the proteins secreted by the lung epithelium of the alveolar and distal airways is the sampling of the epithelial lining (ELF) by bronchoalveolar lavage (BAL) [21]. BAL is a relatively non-invasive diagnostic procedure for which a flexible fiberoptic bronchoscope is inserted into the lung and small amounts of physiologic solution are injected. The fluid is then aspirated obtaining the bronchoalveolar lavage fluid (BALF), which consists of cells (both resident alveolar cells and recruited inflammatory cells), their secreted products and proteins from leakage across the endothelial-epithelial barrier. An alternative method is to directly analyse the bronchial epithelial cells obtained by bronchial brushing [22] during bronchoscopy. The molecular analysis of BALF (BALFomics [21]) and bronchial epithelial cells [23,22] offers great potential [24,25,26,27,28] for research and diagnostics. So far both approaches have been used to study various lung diseases [24,29,30]. The proteins, genes and other substances gained from such studies, which offer insights into disease pathogenesis, are commonly referred to as biomarkers [31].

1.5.2 Lung Diseases

Biomarkers and the insights they offer into disease pathology are urgently needed for the diagnosis and treatment of various lung diseases. For many of which the only treatment available to end stage patients is the transplantation of a donor lung [7]. Those diseases

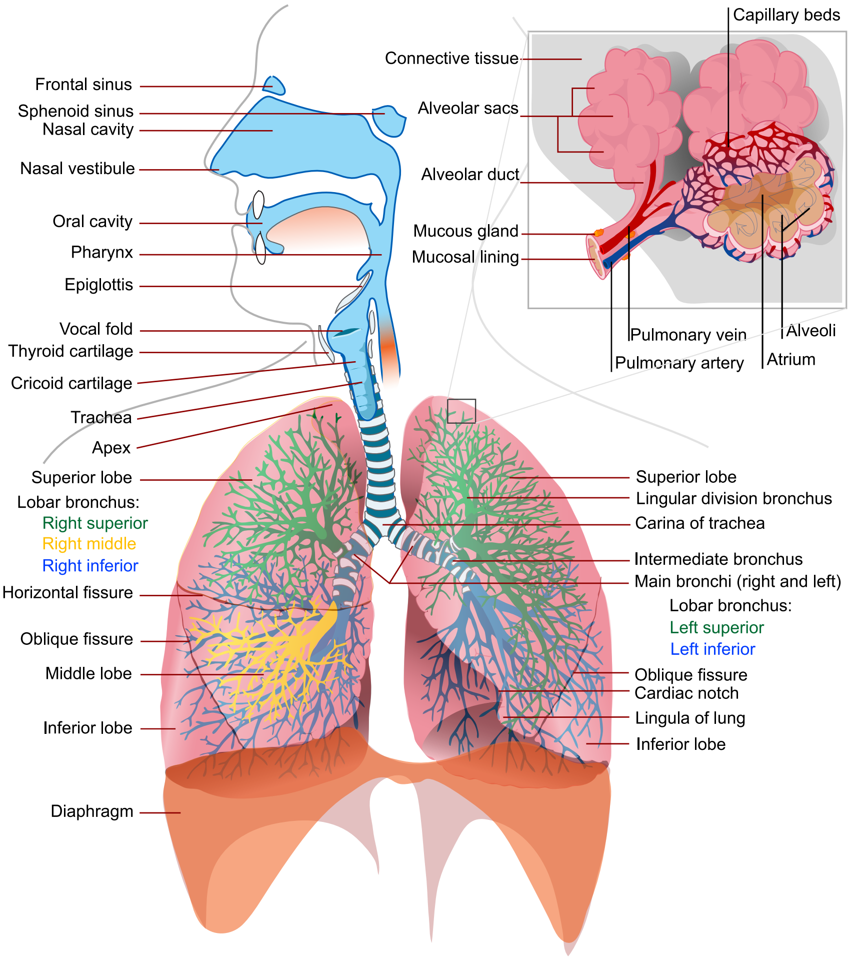


Figure 1.1. An overview of the complete human respiratory system (`http://en.wikipedia.org/wiki/Respiratory_system`)[20]

with lung transplantation as an accepted treatment option can be divided into four groups: septic lung disease (i.e. Cystic Fibrosis (CF) [32]), restrictive lung diseases (i.e. idiopathic pulmonary fibrosis (IPF) [33]), obstructive lung diseases (i.e. chronic obstructive pulmonary disease (COPD) [34]) and pulmonary vascular diseases (i.e. pulmonary hypertension [35]).

1.5.3 Lung Transplantation

Today clinical lung transplantation (LTx) remains the final therapeutic option for patients suffering from a wide range of progressive lung disorders [7,36]. After several trial experiments on dogs in the first half of the 1950s, the first clinical lung transplantation (LTx) was performed in 1963 by Hardy at the University of Mississippi [37,38]. This transplantation was initially successful and showed only a low level organ rejection. The patient died 18 days later from multi organ failure due to a bronchiolar lung tumor. Further lung transplantations were performed between 1963 and 1974. Of the more than 40 patients receiving a transplant during this time, only 2 survived the first month post-transplantation. Main causes of death were among others transplant rejection and respiratory insufficiency. Due to those failures programs for lung transplantations were suspended till the early 1980s. Back then the use of Cyclosporin A as an immunsuppressor led to a new age of lung transplantation [39]. An era which began in 1983, when Dr. Joel Cooper and the Toronto Lung Transplantation Group achieved the first long term success after a single lung transplantation (SLTx)[40]. In 1985 this same group also performed the first long term success in double lung transplantation (DLTx)[41]. After those early successes the number of lung transplantations rose consitently [42], thus establishing lung transplantations as a treatment of end stage lung diseases [7,36]. Even though quite a few medical advances have been made, organ rejection caused by the hosts immune system still represents the final frontier of lung transplantation [4].

1.5.4 Immune System

The immune system prevents infectious agents like bacteria, viruses and parasites from causing harm to the body. The main principles behind the immune system can be divided into the innate and adaptive immune system. Innate immunity provides a first line of defense against many pathogens and starts immediately after a pathogen has entered the body. The immunological recognition allows for the detection of such infectious agents (antigens). This recognition step is achieved by leukocytes of the innate immune system, which are responsible for an immediate immune response, and by the lymphoctes of the adaptive immune system. The adaptive immune system offers a way to overcome the limitations of the innate immune system. This is necessary since the innate immune system provides only limited protection as many pathogens evolve rapidly, altering the molecular patterns of surface molecules. Adaptive immunity offers a solution to these limitations as a great diversity of pathogens can be recognized. B lymphocytes (B-cells) and T lymphocytes (T-cells) are the main effectors of adaptive immunity. Receptors on their surface allow them to recognize foreign substances. A substance that can be recognized by cells of the adaptive immune system is called antigen and the binding region is called epitope. The most important part of the adaptive immune system is the immunological memory. This allows a person, who has once been exposed to a specific infectious agent, to make a more immediate and stronger response against subsequent exposure. Further actions required for the containment and elimination of the infections are provided by the immune effector functions. Members of the immune effector functions are the complement system of blood proteins, antibodies and lymphocytes among other involved white blood cells. To prevent the immune system from causing unwanted conditions such as allergy and autoimmune diseases, the system controls itself by immune regulation [43].

1.5.5 Autoimmune Diseases

To prevent the immune system from attacking the hosts own cells, central self-tolerance safeguards have to be put in place. These mechanisms focus the adaptive immune system on pathogens and steer it away from self [44]. This is essential since the random gene rearrangment, which occurs during lymphocyte development in the primary lymphoid organs (thymus and bone-marrow), leads to the generation of lymphocytes (T-cells and B-cells) with affinity for self antigens. Those potentially dangerous lymphocytes are removed from the pool or controlled by various mechanisms [43]. In the case of immature B-cells this immunological tolerance to self is established in the bone marrow. Thus B-cells recognizing self-molecules present in the bone marrow are removed from the B-cell repertoire by negative selection (clonal deletion). A much more rigorous selection process is performed in the case of T-cells, which undergo both positive and negative selection in the thymus. This makes sure that T-cells selected for survival recognize self-major histocompatibility complex (MHC) molecules but do not recognize self peptides. The breakdown of those self tolerance mechanisms leads to an attack on the normal tissues of the body. This in turn can lead to a variety of immune diseases like systemic lupus erythematosus [45], autoimmune thyroid disease [46] or type I diabetes mellitus [47]. In the case of type I diabetes mellitus, the most extensively studied chronic autoimmune disorder, T-cell mediated destruction of pancreatic islet β-cells occurs. Processes of organ and tissue destruction, as mediated by the hosts immune system, can also occur in the case of allograft rejection. In this case a response to nonself antigens of the transplant occurs [43].

1.5.6 Allograft Rejection

While current advances in immunosuppression proved to be quite effective in the short term, improvements for long-term graft survival are still an ongoing concern [48]. For lung transplants the average 5 year graft survival is around 46.3% [4]. This low graft survival rate is mainly due to the host immune systems recognition of nonself antigens (allorecogntion) and the resulting immune response mediated injury of the transplanted tissue (allograft rejection) [49]. The process of allograft recognition is enabled by the T-cells abillity to recognize genetically different MHC (major histocompatibility complex) molecules. In literature two different pathways (1. direct pathway 2. indirect pathway) are described [49]. For the direct pathway [50] the intact donor MHC molecules on APCs (antigen-presenting cells) in the transplanted tissue are recognized by alloreactive T-cells. In the indirect pathway [51] antigens (Ag) derived from donor MHC molecules are processed by the host APCs and presented to alloreactive T-cells in a self restrictive manner. Those processes signal the activation and transition of the lymphoid tissues naive T-cells into effector cells. In response to the inflammatory signals the allograft is also infiltrated by other cells of the immune system (i.e. neutrophils, macrophages, and natural killer (NK) cells) which promote further injury by proinflammatory mechanisms. The types of rejection occuring after organ transplantation can be grouped into three categories. Hyperacute rejection is a rare and fatal type of rejection which occurs within minutes to hours after transplantation. The cause of this rapidly occurring process are pre-formed antibodies against the donors MHC or ABO blood groups [43]. This process of complement-mediated injury to the graft epithelium activates inflammatory and coagulative cascades. Those cascades are the cause for the extensive thrombosis observed in the graft vessels [43]. Another form of rejection is a cellular mediated immune process referred to as acute rejection [49]. This normally occures within the first months after transplantation. Most patients suffer from at least one episode of acute rejection and even later episodes are still quite frequent [52]. The third rejection type is chronic rejection, which manifests clinically as bronchiolitis obliterans (BO) [4], a submucosal fibrosis of the respiratory bronchioles. Chronic rejection is the main cause for late graft failure and occurs in half of the lung transplant patients within 5 years after transplantation [4]. This rejection type is believed to be caused by chronic cytokine production and smooth vascular muscle proliferation [53]. Another type of graft failure is caused by the graft-versus-host disease (GHVD) [54,43]. The GHVD is essentially characterized as the reverse of graft rejection [55]. This means that immunologically competent cells (i.e. mature T-cells) present in the allograft recognize the transplant recepient as foreign. Thus resulting in a severe inflammatory disease. Acute GHVD is, despite a significant amount of donor-derived lymphoid tissue and cells being transfered, an uncommon and usually fatal occurence after lung transplantation [56,57,58]. GVHD caused by bone marrow transplantation was found to be a risk factor for bronchioltis obliterans (BO) development [59]

1.5.7 Bronchiolitis Obliterans Syndrome

In the mid-1980s bronchiolitis obliterans (BO) was first described in a small cohort of heart lung transplant recipients at Stanford University [60]. Today BO, an accepted pathological manifestation of chronic allograft rejection, is the main reason for transplant dysfunction [4]. Histologically it manifests as a patchy submucosal fibrosis of the respiratory bronchioles, which results in total or subtotal occlusion of the airway lumen [4]. This excessive fibropoliferation of the small airways is due to ineffective epithelial regeneration and tissue repair after injury and inflammation of epithelial cells and subepithelial structures [61]. In parallel with "injury response", a concept proposed to explain the chronic dysfunction of various organ allografts, injury to the airways may be caused by alloimmune-dependent mechanisms, alloimmune-independent mechanisms or a combination of both [62]. The current opinion is that bronchiolitis obliterans represents various causes which lead to similar histologic and clinical results [60]. In addition to the immunological risk factors like repeated acute rejection, human leukocyte antigen (HLA) incompatibility, several non immunological risk factors have been described [63]. The lung allograft is particularly susceptible to fungal and bacterial infection [64,65,66,67]. Lung allograft airway colonization by *fungal aspergillus species* and *bacterial pseudomonas aeruginosa* have already been associated with BO [68]. A viral contributor in the form of *cytomegalovirus* (CMV) pneumonitis has also been correlated with development of BO [69]. Despite all the research describing the development of BO the histologic validation is still only feasible in the end stages. This diagnostic problem is caused by the circumstance that the disease temporarily presents itself only by cellular infiltration of the airways or inactive fibrotic processes. To circumvent those hindrances the term brochiolitis obliterans syndrome (BOS) was introduced, as opposed to the term bronchiolits obliterans (BO), which is used when histologic correlation can be proven [60,4]. BOS is defined as a significant decrease in one second capacity (Forced Expiratory Volume in 1 second, FEV_1) in relation to the best FEV_1-scores after transplantation. FEV_1 scores are usually given as a percentage of the best FEV_1 observed after lung transplantation (baseline). Thus BOS is commonly interpreted as chronic rejection without histological evidence. The BOS classification is divided into 5 stages (Table 1.1) [70,60].

Table 1.1. BOS-stages

BOS-stage	$FEV_1(\%)$
0	> 90
0-p	$81 - 90$
1	$66 - 80$
2	$51 - 65$
3	≤ 50

Of those stages BOS 0-p (potential BOS) was introduced in 2001 to allow for the early detection of BOS [71 ,72,73] . For all diagnostic stages other factors which also lead to a decrease in organ function (i.e. acute rejection or infections) have to be excluded as confounding factors. This diagnostic complexity shows that more accurate biomarkers for BO and BOS detection are urgently needed.

1.6 Biomarkers and High-Throughput Analysis

Recent advances in the field of multi-omics approaches (i.e. genomics, transcriptomics, sequencing, and proteomics) opened up new worlds of opportunities in biomedical research [31,3,74] and biomarker detection. But due to their giant size and complexity, those worlds can only be explored using biostatistics and bioinformatics. One of the treasures to be discovered in the process are new biomarkers for human diseases. Biomarkers can be defined as a dynamic and informational substance (i.e. nucleic acids-based gene expression analysis, peptides, proteins) that can be quantitatively linked to pathogenic progression. Most substance based biomarkers in use today were discovered in blood (plasma, serum and whole blood)[75] but other fluids and sample types are also gaining prominence (i.e. cerebrospinal fluid, tissue biopsies, urine, saliva and bronchoalveolar lavage fluid)[76,24,77,78,10]. While the term biomarker itself has only been established recently, there have always been diagnostic tools monitoring organ functions or general changes in biological structures. For instance body temperature is an indicator for fever [79] or the forced expiratory volume per second (FEV_1) a measure of lung function [7]. Before a molecular biomarker can be used for diagnostic purposes its reliability has to be stringently validated [31,80]. This is far from trivial, especially in the case of a biomarker panel composed of several biomakers [81]. To actually put biomarker research into clinical use, it might be far more feasible to follow an integrative approach between molecular and established clinical diagnostics [82,83]. In the following, common approaches for high throughput analysis of gene expression and monitoring of the proteome are introduced.

1.6.1 Analysis of Gene Expression

The majority of current research is focused on the gene transcription level [84]. This gene expression analysis, which is often referred to as transcriptomics, is usually based on the level of mRNA under varied conditions. Of the techniques used for high throughput transcriptomics microarrays are the most commonly used approach [85,86].

An Introduction to Microarray Technology, Analysis and Validation

Microarrays allow for the measurement of the expression level of a large number (many thousands) of genes simultaneously, and are based on the concept of hybridization. A DNA microarray is a chip of microscopic single strand DNA spots (probes) attached to points on a solid surface. Those DNA (cDNA) strands are complimentary to the mRNA of a specific gene, which allows both nucleic acid strands to bind each other by hydrogen bonds between corresponding nucleotides (Watson-Crick base pairing). For the experiment the whole RNA of a cell at a specific time is isolated and multiplied by polymerase chain reaction (PCR) [84,85,86], with the mRNA or corresponding cDNA being labeled by fluorescent dyes. Now the samples are spotted and hybridized on the microarray chip and, after washing of the non hybridized pieces, scanned by a laser. The intensity of the signal represents the level of a specific mRNA, since each spot is made up of many identical probes. The most common microarray types are known as cDNA and oligonucleotide arrays. For cDNA arrays preamplified cDNA is attached to points on a solid surface chip. These chips are based on competitive hybridization for which a reference sample is used. The reference RNA, which is also hybridized to the microarray, is marked by a different dye. Thus the laser scanning results in ratio data which describes the differential expression between both samples. Recently oligonucleotide arrays (i.e. Affymetrix) gained popularity in the field of transcriptomics. Those arrays allow for the analysis of a single sample, with the probes (sequences of about 25 nucleotides length) being synthesized directly on the chip. Most microarray data sets are in the form of a set of experiments which describe the properties of specific biological entities or conditions (i.e. patients or cell lines). Each experiment is defined by a set of genes associated with the entities. Such data can be easily represented as a data matrix A with one row for each gene and one column for each entity. The matrix entries a_{ij} represent condition-specific expression levels of a gene. After data normalization, which is essential to achieve comparability between different experiments, various data analysis approaches (i.e. gene association analyis, network inferration, model building) can be applied to the obtained data.

Normalization

Results from individual experiments need to be normalized with respect to each other to account for experimental variation in RNA amounts, specific activity of probe labels, and standard handling errors [87]. Individual intensities are adjusted to be able to make comparisons both within the array as well as between arrays in the experiment. These adjustments are necessary to remove purely technical differences that do not represent biological variation and to be able to identify true differentially expressed genes [88]. Over many years of microarray experiments various normalization approaches have been developed [89,90,88,91,92]. After normalization is completed, differentially expressed genes or functional groups, classifying the samples into meaningful clusters, can be identified [93, 94,95,96].

Quantitative Real Time PCR

High throughput transcriptomic approaches like microarrays and RNA sequencing [97] offer nearly unrivaled possibilities for whole genome profiling. As the obtained expression profiles may vary depending on the applied methods and platforms a further validation step becomes necessary. For this quantitative-real-time-polymerase-chain-reaction (qRT-PCR) is the common method of choice. The qRT-PCR is a replication method for nucleic acids, based on the common PCR (polymerase-chain-reaction), which additionally offers the possibility for real time quantification of a target gene. After cDNA synthesis [98] this quantification is performed using fluorescent reporter probes, which specifically detect only DNA fragments containing the target genes sequence of interest. Thus the detected fluorescence directly correlates with the number of synthesized DNA target sequences. The main quantification step is performed during the exponential phase of the PCR, for which each cycle shows an exponential increase in the amount of target sequence. The logarithmized number of cycles after which the fluorescence first rises significantly above the background fluorescence is denoted as CT (threshold cycle). In addition to the gene sequence of interest an internal control gene (housekeeping gene) is also measured, which can be used to normalize for experimental variability. This can be done using the following equation:

$$\Delta CT = CT_{target-gene} - CT_{control-gene}$$

The gene expression in a sample can now be described as an n-times expression change when compared to a reference (i.e diseased vs. healthy reference). For this the $\Delta\Delta CT$ value is calculated.

$$\Delta\Delta CT = \Delta CT_{diseased} - \Delta CT_{reference}$$

From this the expression ratio between a sample and the respective reference can be calculated.

$$Ratio = 2^{-\Delta\Delta CT}$$

Thus obtaining a relative quantification of gene expression.

1.6.2 Differential Expression

The identification of genes that are differentially expressed between two distinct experimental conditions has been a major topic in the scientific community [93]. The easiest approach is to use the intuitive fold change criterion [95].

Fold Change

The fold change is a number which decribes how much a quantity changes compared to a reference value. For example a measured value of 200 and a reference value of 100 would be described as a fold change of 2. For microarray studies the fold change (FC) of a feature (gene) f_i is calculated by dividing the mean/median $\bar{f_i}^{class}$ over all samples in the class of interest by the mean/median of the reference class $\bar{f_i}^{reference}$.

$$FC_{f_i} = \frac{\bar{f_i}^{class}}{\bar{f_i}^{reference}} \tag{1.1}$$

If the genes are equally expressed in both classes, the FC equals 1. For logarithmized expression values the FC can be calculated by:

$$log(FC_{f_i}) = log_2(\bar{f_i}^{class}) - log_2(\bar{f_i}^{reference}) \tag{1.2}$$

In this case the FC equals 0 if the genes are equally expressed in both classes of interest. While the FC is an established measure of differential expression in biological research [99,100,101], it does not provide a significance estimate or take into consideration the level of expression variance. While statistical tests (i.e. t-test, Wilcoxon rank sum test) are more statistically justified than the FC [102], they do not take into consideration that even statistically significant changes are unlikely to have a biological effect if the FC is very low [94]. Further it was shown that for validation purposes genes exhibiting small degrees of change show lower correlation between microarray and qRT-PCR results [103]. One solution to the shortcomings of the FC, while still keeping its advantages, is the rank product (RP) test [94,104].

Rank Product Test

The rank product (RP) test [94,104,105] is a non-parametric statistical test which detects genes that are consistently found among the most up-regulated (or down-regulated). For this we assume a data set with two experimental conditions disease (D) and control (C). Given n_D and n_C replicates, the expression ratios of each gene are calculated for all possible comparisons ($k = n_D \times n_C$) between both conditions:

$$\frac{D_1}{C_1}, \frac{D_1}{C_2}, \ldots, \frac{D_{n_T}}{C_{n_C}} \tag{1.3}$$

After ranking the gene expression ratios within each comparison, the rank product for each gene g in k comparisons $i = 1, \ldots, k$ can be calculated by

$$RP_g^{up} = (\prod_{i=1}^{k} r_{g,i}^{up})^{1/k} \tag{1.4}$$

where $r_{i,g}^{up}$ is the position of gene g in the list of genes in the ith replicate sorted by decreasing FC, i.e. $r^{up} = 1$ for the most strongly up-regulated gene. Analogously RP_g^{down} is calculated from the list of genes sorted by increasing FC, i.e $r^{down} = 1$ for the most strongly down-regulated gene. Genes with small $RP(\bar{r})$ values are supposed to have the highest biological relevance for up- and down-regulation respectively. Based on the RP a significance level can be assigned to the differential expression of each gene. For this a permutation scheme is used to estimate how likely a given RP value is in a random experiment. For this we generate p permutations of k rank lists of length n. Then we calculate the rank products of the n genes in the p permutations and count how often the rank products of the genes in the permutations are smaller or equal to the observed rank product. This allows us to determine a p-value. To correct for multiple testing [106,107] a respective Benjamini-Hochberg false discovery rate (FDR) was associated with each gene [105]. The FDR adjusted p-value p^* is calculated for each gene according to the following procedure:

Benjamini-Hochberg False Discovery Rate

Step 1: Rank the p-values in a decreasing order $p_1 > p_2 \ldots > p_n$

Step 2: The highest p-value is not adjusted: $p_1^* = p_1$

Step 3: Corrected p-value p_2 is calculated by $p_2^* = p_2(\frac{n}{n-1})$

Step 4: Corrected p-value p_3 is calculated as in Step 3 by $p_3^* = p_3(\frac{n}{n-2})$

Continue till all p-values are adjusted. After correction for multiple testing a p-value cut-off of $\alpha^* = 0.05$ equals a false positive rate of 5%.

Gene Annotation

Even after multiple testing correction [106], microarray studies often return very large lists of differentially expressed genes. To allow for the functional analysis such genomic studies, the functional annotations [108] of numerous genes have been made available in various online resources [96]. Those annotations are systematically categorized and describe either hierarchical classification systems (i.e. Gene Ontology (GO)[109]) or biochemical pathway representations (Kyoto Encyclopedia of Genes and Genomes (KEGG)[110] and Biocarta [111]). Those annotation categories can be used to assess the biological significance of a gene list. This type of analysis can be performed using specialized bioinformatics tools such as the Database for Annotation, Visualization and Integrated Discovery (DAVID) [112]. DAVID consists of an integrated biological knowledgebase and analytical tools aimed at systematically extracting biological meaning from gene lists [96]. The DAVID analysis (Figure 1.2) is essentially based on gene enrichment analysis [96]. This method assumes that if a biological category or pathway is influenced in a given study, the co-functioning genes have a higher potential to be detected as differentially expressed. The degree of enrichment is assessed by comparison against a background list. The background list should encompass all the genes that can potentially be selected for the annotation process, which in most cases is all genes from a chip. Given that for example 10% of the differentially expressed genes are kinases vs. 1% in the background, the significance of enrichment can than be tested by established statistical methods, like chi-squared, Binomial probability and Hypergeometric distribution [Huang2009a]. In the case of DAVID a Fisher exact test is used with multiple testing correction being performed using the Bonferroni method [107,106] or several other less conservative approaches [107,106]. Thus a list of functional groups significantly enriched in a gene list of interest is returned.

1.6.3 Analysis of Protein Expression

In addition to the transcriptomic methods dealing with gene expression, there is also a wide array of analogous proteomics methods (i.e peptide arrays) measuring the level of protein in a sample [113,78]. Proteomics is defined as the systematic analysis of all proteins in any defined biological compartment [114]. This offers a complimentary ap-

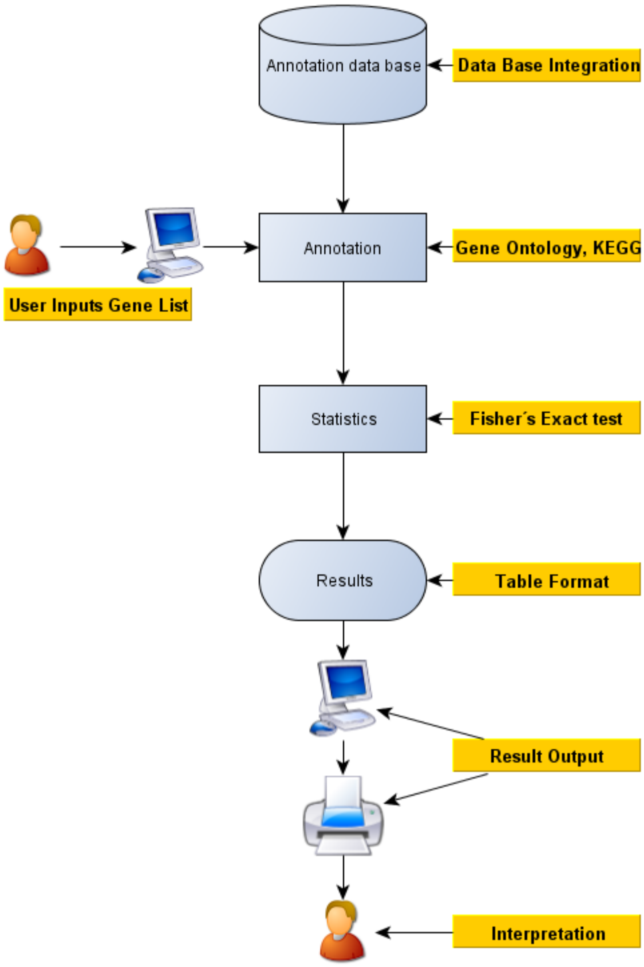


Figure 1.2. The annotation workflow as performed by using DAVID [112]

proach to gene chip microarrays, since there is often a poor correlation between mRNA abundance in a cell or tissue and the quantity of the corresponding functional protein. A relationship further complicated by the complexity of the proteome, due to alternate splicing of the transcripts, sequence polymorphism and posttranslational modifications. Proteomics is believed to offer great potential for the burgeoning field of clinical proteomics [78,81]. This research direction aims to gain insights into the complex processes of pathobiology by the discovery of clinically relevant protein biomarkers.

Mass Spectrometry

One of the workhorses of biomarker detection [115] are mass spectrometry (MS)- based technologies. Modern mass spectromic methods make it possible to generate proteome profiles of biological tissues and fluids. There are different kinds of mass spectrometers [116] but the most popular approach of mass spectrometry for biomarker detection in various body fluids is the MALDI-TOF (Matrix Assisted Laser Desorption and Ionisation - Time of Flight) method [117]. This method offers easy sample preparation and acquisition while also being tolerant towards contaminants such as salts, detergents or common buffers [115]. Even though samples can be analyzed in a highly reproducible [118] fashion by mass spectrometry, reproducibility strongly depends on the sampling process, standardized measurement and a stringently controlled data analysis [119].

Technical Details

In the case of MALDI-TOF [117] the samples of interest are mixed with a matrix consisting of an UV absorbing compound and spotted as a spot on a steel sample slide. The spots are dried and the steel sample slide is inserted into the sample chamber under high vacuum. Laser pulses are fired on the spots and the compound transfers the energy from the excitation laser to the sample molecules, which vaporises and ionises them. After the ions stop vibrating the high voltage is turned on, which attracts and accelerates the ions through the time of flight tube to a detector. The time between the laser pulse and the arrival of the ions at the detector is measured by a high precision timer. Since this time is proportional to the mass of the molecule divided by its charge (m/z), the molecular weight can be calculated. The results from a MALDI-TOF experiment can be represented as a spectrum where the height (intensity) of a peak represents the number of ions present for a specific m/z value.

Mass Spectrometric Data Preprocessing

As shown in [120,121] mass spectrometry can be composed into three components

$$f(i,j) = b(i,j) + s(i,j) + \varepsilon(i,j). \tag{1.5}$$

where $f(i,j)$ is the observed value, $b(i,j)$ is the baseline value, $s(i,j)$ is the true signal and $\varepsilon(i,j)$ is the noise for ith sample at jth m/z ratio. Noise of MS spectra can be caused by many different factors such as system artifacts, thus making preprocessing necessary for a successful data analysis [121,122,123]. The baseline decribes the low frequency component of the observed signal and is mainly caused by chemical noise and ion overload.

Baseline Correction

It is often the case that a raw spectrum shows an elevated baseline. Thus baseline correction was performed by flattening the base profile of each spectrum by removal of baseline slope and offset [**Sauve**]. For this the baseline is estimated by partitioning the m/z range on the logarithmic scale into n equally spaced intervals, finding the local minimum within each interval, and smoothing these local minima by either local regression (loess) or local interpolation. To make sure that the spectrum rests on the zero horizontal line, this baseline is subtracted from each raw spectrum. Resulting negative values are set to zero.

Interpolation

The varying m/z resolution and ranges between spectra prevent a straightforward representation as a data matrix. Thus, after restricting all spectra to the smallest common m/z range, a linear interpolation method [124] can be used to approximate the missing data in spectra positions with low resolution. This makes it possible to represent all spectra in a single data matrix. See Figure 1.3 for details.

Normalization

Various experimental sources can lead to systemic differences between spectra. Thus a normalization method based on the total ion count becomes necessary, which allows for the comparison of the absolute peak intensities of different spectra [125,126]. For this the area under curve (AUC) of each spectra is calculated for all m/z values [127].

$$AUC = \sum_{i=1}^{n} \gamma_i \tag{1.6}$$

where γ_i ,the signal at the ith m/z ratio, is used to measure the protein concentration in mass-spectrometry data. Using this approach all spectra were rescaled by division to the same AUC value.

Peak Recognition

After normalization it is important to detect peaks in the spectrum [128], which are likely to represent proteins. For this the mean spectrum over all available spectra is calculated. This spectrum is then smoothed using the moving average of the k nearest neighbours [129], which helps to remove spurious peaks. Afterwards a local maximum is identified in the smoothed spectrum, which is considerd as a peak when its signal to noise

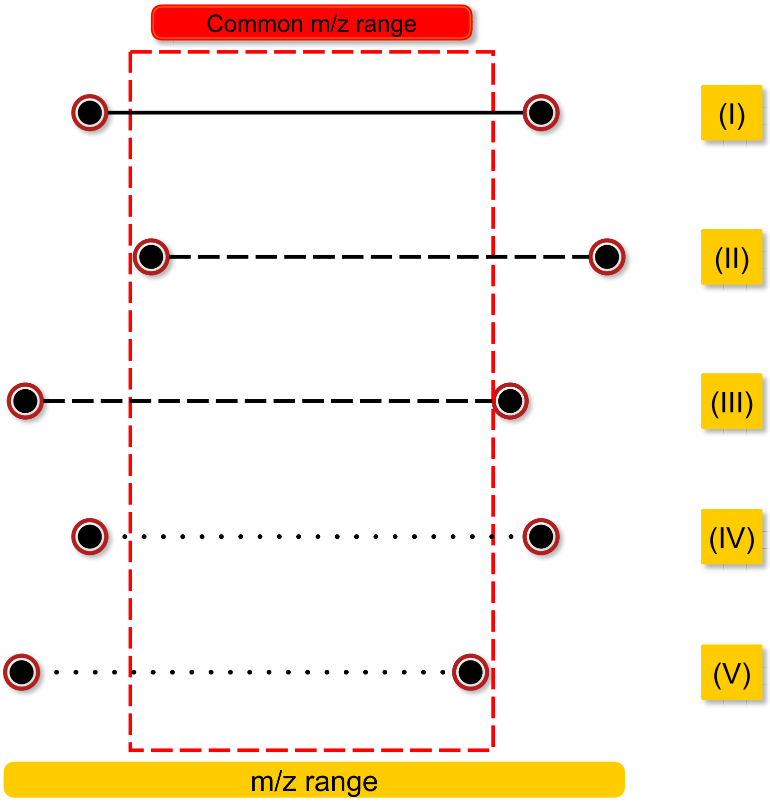


Figure 1.3. Resampling of mass spectra: Five spectra are shown exemplary as lines between the mass range endpoints. The spectra are ordered according to decreasing resolution (I-V). All spectra are restricted to the same mass range between the largest starting point (here of spectrum (II)) and the smallest end point (here of spectrum (V)) of all spectra (I-V). The spectrum with the highest resolution (here spectrum I) is chosen as the master resolution. All other spectra are interpolated to match this master resolution [124].

ratio [130,128] and intensity exceed a user defined threshold. The noise is estimated as the median of the absolute deviation (MAD) of points within a mass window [129,128], while the intensity threshold is set to a percentage of the highest peak intensity. After this step the data can be represented as a data matrix A with one row for each peak and one column for each entity. The matrix entries a_{ij} represent condition-specific intensity levels at the highest point of a peak, for a specific m/z level. The preprocessed data can now be used by various data analysis approaches.

ELISA

To further validate the results obtained from mass spectrometry, an enzyme linked immunosorbent assay (ELISA) is often applied [131,132]. This represents the most reliable and sensitive protein based testing platform [133] currently available. All ELISA tests require an especially well-characterized antibody for the detection of the protein of interest (antigen), which is not always available. Those antibodies, previously linked with an enzyme, bind the protein of interest and the enzyme catalyses a reaction to mark the presence of the antigen. This reaction processes an added enzyme specific substrate and leads to a detectable change in color. This color change signal, as measured by a photometer, allows for an exact detection of antigen concentration.

1.7 Data Mining and Machine Learning

The field of machine learning, which has its roots in artificial intelligence (AI), deals with systems that use previous experience to improve performance on a specific task. In medical research the applications of interest include diagnosis of patients and the grouping of patient cohorts into pathologically relevant subgroups. Those goals are reflected in two main types of data analysis, as defined in [134]:

1.Clustering (Unsupervised Learning) gathers objects into groups (clusters), such that the objects within a cluster are more similar to each other than they are to objects belonging to a different cluster.

2.Prediction (Supervised Learning) assigns an appropriate label (categorization) to new objects given their attributes. The model for the respective decision making uses information extracted from the relationship between attribute values and labels of a set of sample objects.

First we will take a closer look at the concepts behind clustering (unsupervised learning).

1.7.1 Unsupervised Learning

If the task of a research project is to partition a collection of unlabeled data into meaningful clusters, methods from the field of cluster analysis (unsupervised learning) provide an appropriate solution. Cluster analysis as described in [135,134] is the organization of a collection of patterns (usually represented as a vector of measurements, or a point in a multidimensional space) into clusters based on similarity. Thus patterns within a cluster are more similar to each other than they are to a pattern belonging to a different cluster. A pattern p_i consists of a vector of d measurements $p_i = (p_{i1}, \dots, p_{id})$. In the case of gene expression data each gene can be represented as a vector of expression levels under a number of different conditions. Accordingly a condition can be represented as a vector of genewise expression levels. A large number of algorithms to cluster unlabelled data sets has been proposed [136,135] and applied to biological data sets [137,138,139]. Among the most popular methods are hierarchical (agglomerative) clustering [135], k-means [135] and Partitioning Around Medoids (PAM) [140].

Similarity Measures for Clustering

A measure of similarity between two patterns drawn from the same feature space is essential for most clustering procedures. The most common way is to measure the dissimilarity (or distance) between two patterns p_i and p_j, using a measure defined on the feature space. The most popular metric is the Euclidean distance:

$$d_2(p_i, p_j) = \sqrt{\left(\sum_{k=1}^{d} (p_{i,k} - p_{j,k})^2\right)} = ||p_i - p_j||_2 \qquad (1.7)$$

Agglomerative Clustering

The most common approach to microarray data clustering is called agglomerative clustering or hierarchical clustering [135]. The idea behind this is based on a simple principle: Given n data points, the algorithm starts with n clusters by assigning each point to a cluster of its own. Then a bottom-up procedure is performed, which at each stage successively merges the pairs of clusters (i.e. a set of points) closest to each other. Since each step returns a clustering solution with one cluster less, all data points are joined into a single cluster after $n-1$ operations. This approach creates clustering solutions for all possible numbers of clusters $k, 1 \leq k \leq n$, which can be represented as a dendrogram structure by the order of merges.

K-Means Algorithm

K-means [135] is a partitional clustering algorithm which produces a single partition of the data, for which a fixed number of clusters k has to be predefined. To describe the algorithm we first introduce the concept of a cluster centroid. Given a cluster with n points (patterns) $p_1, \dots, p_n$ the cluster centroid is defined as the arithmetic mean μ of the vectors:

$$\mu = \frac{1}{n} \sum_{i=1}^{n} p_i \qquad (1.8)$$

K-means as described in reference [135]

Step 1: Initialization of k Clusters
Each point $p_1, \ldots, p_n$ is assigned to one of the k clusters $C_1, \ldots, C_k$. This is done by randomly choosing k points as cluster centroids $z_1, \ldots, z_k$, and assigning the remaining points to the cluster centroid with the lowest Euclidean distance.

Step 2: Recalculation
For each cluster Cluster C_r the centroid is calculated anew according to equation 1.8.

Step 3: Partitioning
The partitioning is adapted to the new centroids, each point is assigned to the centroid with the lowest Euclidean distance.

Step 4:
Steps 2 and 3 are repeated until the centroids no longer move.

Each new partitioning reduces the sum of the Euclidean distance from the points to their respective cluster centroid but the k-means algorithm is not guaranteed to find the optimal solution (global minimum). Depending on the initialization only local suboptima are found. Because of that it is rerun several times with different starting centroids and the best solution is selected. It is also possible to use another distance measure than the Euclidean distance [141,142,143].

PAM algorithm

The PAM clustering algorithm [140] is similar to k-means [135] but operates directly on a dissimilarity matrix. Before the actual clustering procedure PAM defines k medoids. A medoid is usually defined as an object with minimal average distance to all other objects in the cluster. After the starting medoids are defined, each object in the data set is assigned to the nearest medoid. Thus putting object i into cluster v_i, when medoid m_{vi} is nearer than any other medoid m_w. The algorithm proceeds in two steps:

BUILD-step: This step sequentially selects k centrally located objects to be used as initial medoids.
SWAP-step: If the objective function can be reduced by interchanging (swapping) a selected object with an unselected object, then the swap is carried out.

This is continued until the objective function can no longer be decreased. Thus defining the final data partition.

1.7.2 Supervised Learning (Classification and Prediction)

The goal of prediction [144] is to predict a target attribute for new objects based on the values of the other attributes. The relationship between the target attribute and those other attributes is then learned from a training set of data, for which the target attribute is already known. The training data captures an empirical dependency between the ordinary attributes and the target attribute, for which the data mining technique builds an explicit model of the observed dependenices. In the case of categorial target values the prediction is called classification. As opposed to continous numerical target values, for which the prediction is referred to as regression. There are a number of popular classification methods, which have already been applied to problems from the field of biology and medicine [145,121,146].

Cross-Validation

To get an unbiased assessment of a classifiers predicitve quality, it is required to test its performance on an independent data set [147,148]. Since in a clinical setting the availability of samples is often limited, withholding a substantial proportion of the data for testing purposes might potentially reduce the quality of the predictive model. Thus the approach of k-fold cross-validation can be used to predict how well a classifier will perform in practice. For this the original sample set is divided into k subsamples. While one subsample is retained as the test data for model validation, the remaining $k-1$ subsamples (training set) are used to learn the classifier. This ensures that the classification result for each sample is unbiased by knowledge of the particular sample. The validation procedure is repeated k times (the folds), with each subsample being used exactly once for the validation set. In stratified k-fold cross-validation, the folds are selected such that each fold contains roughly the same proportions of class label types. The classification results obtained for each sample, when the particular sample was not part of the training set, can be used to rate the classifier. Common criteria include accuracy, sensitivity, specifity and the area under curve (AUC) [149]. Overfitting of the predictor can be a major limitation in supervised learning. One of the main reasons for cross-validation is to test if the model was not overfitted. Overfitting means that the model was optimized for the available test data but poorly predicts independent data. This happens when the model picks up random variations that do not present true relationships. Another important thing to keep in mind is that all preprocressing and feature selection steps using class knowledge should be included in the cross-validation. Otherwise a substantial bias is introduced to the cross-validation results [150], which leads to an overestimation of prediction accuracy.

Decision Trees

Classification using decision trees such as CART (classification and regression trees)[146], is quite popular in in the Machine Learning Community. A decision tree classifies a pattern by performing a sequence of simple tests, where the tests performed at subsequent levels depend on the outcome of the previous tests. For the class of binary trees there are two types of tests: (i) equality tests for categorial attributes, (ii) inequality tests on a single real-valued attribute. An example of the first type is *color* = *red*, and an example of the second type is $length \geq 24cm$. Representing the rules as a tree structure T, each tree node t represents a rule used for testing a variable X_i from a set of input variables $X_1, X_2, \ldots, X_D$. Using those rule sets the classifier partitions the input space R^r into cuboid regions for predicting an output variable Y. Hereby the decision node at the top of the tree, containing only outgoing edges, is called root node. To predict the output variable y for a sample x, drop x down T and follow its path till a terminal (leaf) node is reached. Each leaf node contains only incoming edges and represents a specific class which is associated with a certain partitioning based on a sequence of tests. The class associated with a specific leaf node is determined by the majority class of samples from the training set which, according to the rule set, would be assigned to this leaf node. A new sample assigned to this leaf node is then classifed accordingly. See Figure 1.4 for an example.

Construction of a tree classifier requires a labelled training set

$$L = (x_1, y_2), \ldots, (x_n, y_n) \tag{1.9}$$

where x_n is an object measured in the input variables $X_1, X_2, \ldots, X_D$ with $n = 1, \ldots, N$. The class label of each object is defined by y_n which can take a value $k \in 1, \ldots, l$. In the case of a binary classification problem an object can belong to either of two classes i.e. $k \in 1, 2$. The tree model is then constructed by partitioning the training set of measurement vectors into "purer" subsets. Herefore every possible value of each variable is considered for each split. The goodness of a split is evaluated by the achieved decrease in impurity. The most established choice of impurity within a tree node t [146] is given by the Gini index

$$I_G(t) = \sum_{j=1}^{l} p_j(1 - p_j(t)) \tag{1.10}$$

where we assume that there are l classes and $p_1, p_2, \ldots, p_l$ are the proportions of samples in the l classes. Then the Gini index, as used by CART, is a measure of how often a randomly chosen element from the set would be incorrectly labelled if it were randomly labelled according to the distribution of labels in the subset. Thus it is computed by summing the probability of each item being chosen times the probability of a mistake in categorizing that item. It reaches its minimum (zero) when all cases in the node fall into a single target category. Based on the respective impurity function a feature and split are rated according to the decrease in impurity [146]

$$\triangle(s, t) = I(t) - h(t_L)I(t_L) - h(t_R)I(t_R) \tag{1.11}$$

where s is a split of node t, $h(t_L)$ and $h(t_R)$are the propotions of the samples in the left and right daughter nodes of node t, respectively. The split that leads to the highest decrease in impurity is chosen for the tree. By recursively using the node-splitting procedure, the tree is usually overgrown (too many descendant nodes), which is likely to overfit the training data. Thus a pruning steps is necessary which removes some nodes to achieve an optimal bias-variance trade-off. The decision which nodes to remove is usually based on an independent test set or cross-validation.

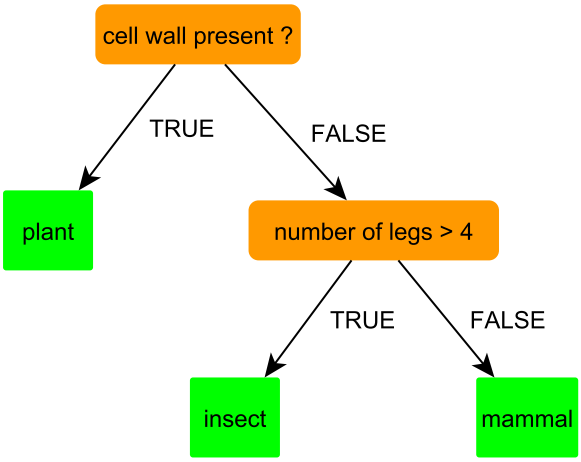


Figure 1.4. A simplified decision tree model for the discrimination between plants, insects and mammals.

1.8 Ensemble Learning

Decision trees can easily be converted into a set of logical formulae. A property that is very desirebale since it allows the user to understand how examples are classified, which is nearly impossible for many other classifiers [146]. Their explanatory power and the fact that they derive rules that can be expressed as logical formulae is the main reason why decision trees are used. They are often not competitive with other classification approaches in terms of their classifcation accuracy. This is especially true for high dimensional data such as microarray or mass spectrometry measurements where a single tree is not able to reflect the inherent complexity. Another disadvantage is the unstable topology of decision trees. A minor change to the training set might result in a substantially different tree model [151]. To overcome those limitations ensemble learning methods generate many classifiers and aggregate their results [152]. Two well-known methods are Boosting [153] and Bagging [154] of classification trees. In Boosting, successive trees give extra weight to points incorrectly predicted by earlier predictors. In the end, a weighted vote is taken for prediction. In Bagging successive trees do not depend on earlier trees. Each tree is independently constructed using a bootstrap sample of the data set. A classification tree T_k is grown for each perturbation Λ_k of the learning set, $k = 1, 2, \ldots, K$; a test observation x is dropped down each tree; and the classifier predicts the class of that observation by that class that enjoys the largest number of total votes over all of the trees. Meaning that in the end a simple majority vote is taken for prediction. An improved variant of the Bagging approach, called random forest [155], was introduced by Leo Breiman [147,156,157].

1.8.1 Random Forest

Like bagging the random forest[155] tree ensemble constructs each tree using B bootstrap samples Λ_b of the learning set Λ. In addtion to this well known randomization scheme it also introduces a randomization component into tree construction itself, such that for the tree T_b each node is split using the best split among a randomly chosen subset of m variables. These are the only two tuning parameters for a random forest: the number m of variables randomly chosen as a subset at each node and the number B of bootstrap samples. The procedure is relatively insensitive to a wide range of of values of m and B. If the class of interest contains only a small proportion of the observations, as is often the case in a clinical setting, each bootstrap sample will contain only very few of the minority class observations. To prevent the resulting poor class prediction for the minority class, the majority class is undersampled (balanced random forest (BRF)) for each bootstrap sample [158,159]. After classifier construction, a random forest predictor score (RF-score) based on the percentage of trees voting for a specific class is calculated. Most of the time a simple major vote (RF-score $\geq 50\%$) is used for class assignment. While the random forest algorithm seems to apply a counterintuitive strategy, it turns out to perform very well compared to other classifiers [160,161] and is robust against overfitting [155]. The random forest methodology framework can even be applied to methods other than decision trees [162]. Another advantage is an internal unbiased estimation of classifier performance similar to cross-validation. At each bootstrap iteration the data not in the bootstrap sample (out of bag (OOB) data) is used to estimate the error rate. The mean error estimation over all bootstrap iterations is referred to as the OOB error.

1.8.2 Random Survival Forest

An extension of the random forest method to right censored data is given by random survival forests [163]. It follows the same principles of constructing an ensemble of base learners but uses survival trees [163] instead of standard decision trees [146]. Those survival trees are a special type of decision trees, which are constructed using an observation data set of time, status, event [164] and associated predictor covariates [165]. For this we denote survival time and censorship information for each of the n individuals within h by

$$(t_1, \delta_1), \ldots, (t_n, \delta_n) \tag{1.12}$$

An individual with no event at time t_l is considered censored ($\delta_l = 0$), while an event at time t_l is denoted by $\delta_l = 1$. The splits at a node h are of the form $x \leq c$ or $x > c$. Given the value x_l of x for an individual $l = 1, \ldots, n$, the number of observations assigned to each of the two daughter nodes created after a split is given by:

$$n_1 = \#(l : x_1 \leq c) \tag{1.13}$$

$$n_2 = \#(l : x_1 > c) \tag{1.14}$$

The splits at each node are chosen such that the resulting daughter nodes show a maximum survival difference. One survival based measure of node seperation, which can be used to choose the best split, is given by the log-rank score test. The best split can now be chosen according to the log-rank score test [166], which serves as a measure of node separation. For this we assume the predictor x has been ordered such a way that $x_1 \leq x_2 \leq \ldots, \leq x_n$. We can now compute a ranking for each survival time t_l,

$$a_l = \delta_l - \sum_{k=1}^{\delta_l} \frac{\delta_k}{n - \delta_k + 1} \tag{1.15}$$

where $\Gamma_k = \#\{t : T_t \leq T_k\}$. The log rank score is defined as

$$S(x, c) = \frac{\sum_{k<=c} a_l - n_1 \bar{a}}{\sqrt{n_1 (1 - \frac{n_1}{n}) s_a^2}} \tag{1.16}$$

with the sample mean and sample variance of $\{a_l : l = 1, \ldots, n\}$ are given by $\bar{a}$ and a_s^2 respectively. The resulting measure of node seperation $|S(x, c)|$ is maximized over x and c, yielding the best split. The tree reaches a saturation point when a terminal node contains at least 3 events with an unique event free time.

Ensemble Estimation

Similar to the random forest method, where the predicted class is defined by a majority vote over all trees, the ensemble cumultative hazard function is produced by the average over all trees. For each tree the cumultative hazard estimates are calculated terminal nodewise. For a node h with distinct event times let $d_{l,h}$ and $Y_{l,h}$ be equal to the number of events and individuals at risk at time $t_{l,h}$. The cumultative hazard for the node h can now be defined as:

$$\hat{H}_h(t) = \sum_{t_{l,h<t}} \frac{d_{l,h}}{Y_{l,h}} \tag{1.17}$$

The survival rate at a time (t) can then be calculated by the product limit (PL) method [167]:

$$\hat{S}_h(t) = \prod_{t_{l,h<t}} (1 - \frac{d_{l,h}}{Y_{l,h}}) \tag{1.18}$$

The rules making up the tree assign each individual i with the predictor x_i to a specific terminal node, from which the individuals cumultative hazard $(\hat{H}(t|x_i))$ is determined. The estimate described so far is based on one tree only, while the final ensemble cumultative hazard is averaged over all trees $(b = 1, ..., ntree)$ in the ensemble. This ensemble hazard $(\hat{H}_e(t_x, i)$ is calculated by:

$$\hat{H}_e(t|x_i) = \frac{1}{ntree} \sum_{b=1}^{ntree} H_b(t, x_i) \tag{1.19}$$

An OOB estimate of the ensemble cumultative hazard $(\hat{H}_e^*)$ can be obtained for each individual by averaging only over those trees for which the individual in question was part of the OOB set. Ensemble survival rates can be calculated analogously to ensemble cumultative hazard.

Concordance error rate

The quality of a predictive model is generally reflected by its prediction error on a test set or by cross-validation. In the case of random forests or random survival forests this can be done using the OOB principle. An important aspect of this procedure is the measure used to rate the prediction error. In the case of a predictive survival model this is provided by the concordance index (C-index). The concordance index reflects the probability that, given a randomly selected pair of individuals, the individual that suffers from an event first also had the worse predicted outcome. The concordance error rate is calculated by 1 minus the C-index and takes values between 0 and 1, with 0 reflecting perfect concordance. A concordance error of 0.5 represents a prediction not better than random guessing. To compute the concordance index we first have to define what constitutes a worse prediction outcome. This can be done using the following approach. Let $t_1^*, \ldots, t_N^*$ denote all unique event times in the data. Individual i is said to have a worse outcome than j if

$$\sum_{k=1}^{N} \hat{H}_e^*(t_k^*|x_i) > \sum_{k=1}^{N} \hat{H}_e^*(t_k^*|x_j). \tag{1.20}$$

The concordance error is calculated as follows [163]:

1: Create all pairs of observed responses over all of the data.

2: Omit those pairs where the shorter event time is censored. Also omit pairs i and j if $T_i = T_j$ unless $\gamma_i = 1$ and $\gamma_j = 0$ or $\gamma_i = 0$ and $\gamma_j = 1$. The total number of permissible pairs is denoted by Permissible-Pairs

3: For each permissible pair in which the shorter event time had the worse predicted outcome, add 1 to the running sum p. If the predictions are tied count add 0.5 to the sum. Concordance $= p$ now denotes the total sum over all permissible comparissons.

4: Define the concordance index (C-index) as

$$C = \frac{Concordance}{Permissible - Pairs} \tag{1.21}$$

and the concordance error by 1 - C

The concordance index can also be applied to uncensored data where it estimates the Mann-Whitney parameter [168]. In this case it also follos the same idea as the Kendall rank correlation coefficient [169]. A possible C-index application on uncensored data is the comparisson of microarray and qRT-PCR results.

1.9 Biclustering

Unsupervised clustering methods are among the most used algorithms in the analysis of high dimensional data sets. These methods allow for the discovery of unknown patterns hidden in the data. In the case of microarray data the most common approaches used include: (1) inferring gene function from clusters of genes similarly expressed across samples [170] and (2) searching sample groups that show similar expression patterns across the genes [171]. While the clustering of genes [140] and the clustering of samples [139] have been been the topic of various research efforts [172], the complexity and high dimensionality of the data [173] still make it a very challenging problem. This complexity prevents common clustering algorithms like k-means or hierarchical clustering from discovering the structures of interest [135]. In recent years new approaches i.e unsupervised random forest clustering [174] or trimmed-k-means [175] introduced concepts like feature weighting or outlier removal. Another concept, that gained prominence during the last decade, is biclustering [138]. The term biclustering describes the simultaneos clustering of the rows and the columns of a data matrix. The first biclustering algorithm was published by Hartigan [176]. Since then numerous new methods were introduced [138,9,177,178,179]. Most clustering techniques cluster a group of genes according to their expression levels under multiple conditions or a group of conditions based on a number of genes. But the clustering of genes and conditions runs into a difficulty. The problem arises from the general principles of cellular processes. Groups of genes that are co-regulated in a certain subset of experimental conditions can be completely independent with respect to other conditions. Such local expression patterns might enable us to find new genetic pathways normally not found by ordinary clustering algorithms. Such local patterns in microarray data can be found by an approach called biclustering , where each bicluster is defined by a subset of genes and a subset of conditions. Classic clustering algorithms cluster the rows and columns independently. The genes are clustered by their expression levels under all conditions. Similarily, each condition is defined by the activity of all the genes. Biclustering algorithms cluster the rows and columns simultaneously. Each gene in a bicluster is selected only a subset of the genes. This enables biclustering algorithms to identify groups of genes that show similar activity patterns under a specific subset of the experimental conditions. This is an improvement over conventional clustering, since only a small set of the genes participates in a cellular process. Additionally a cellular process might be active only in a subset of the conditions. One major disadvantage of biclustering is that they use random seeds, which causes inconsistency and results vary considerably even when obtained from the same dataset [177]. While the new developments in clustering substantially improved on existing methods, the complex assumptions and models used often make the results hard to interpret and validate by researchers [180]. An argument known thus far from discussing the advantages and disadvantages of the fold change compared to more advanced statistical methods [95,94]. In this thesis the novel biclustering algorithms Block Maxx is presented. This algorithm combines the advantages of biclustering with the easier interpretability and usability of conventional clustering approaches.

Chapter 2

Methods

To identify possible prognostic markers for the early detection of the bronchiolitis obliterans syndrome, bioinformatics approaches were applied to high throughtput data obtained from a collective of lung transplanted patients. The clinical and experimental parts were performed at the Hannover Medical School (MHH). All patients enrolled gave their informed consent to participate in this study. Approval from the Ethics Committee of the Hannover Medical School (MHH) has been obtained. BAL and bronchial brushing was performed by the Department of Respiratory Medicine, MHH [24] following European Respiratory Society Task Force guidelines [181]. The data sets include microarray expression data from bronchial epithelial cells and mass spectrometry screening of bronchoalveolar lavage fluid (BALF). Extensive clinical data was provided for all analysed samples. Due to the complexity of the clinical and high throughput data, machine learning methods were adapted specifically to the respective research topics.

2.0.1 Patient Demographics

A total of 77 bronchial epithelial cell samples from 53 patients (26 female, 37 male) were collected from 372 to 1042 days (611 ± 140) after LTx (4 bilateral, 49 single) during routine clinical surveillance. Of these patients, 18 were affected by cystic fibrosis, 23 had lung emphysema, 12 pulmonary fibrosis, 9 alpha-1-PI deficiency and 15 suffered from various other conditions. See Table 2.1 for details. For the BALF profiling 146 samples from 82 patients (42 female, 40 male) were collected 169 to 1000 days (mean 471 days 177) after LTx (73 bilateral, 9 single) during routine clinical surveillance. Of these patients, 19 were affected by cystic fibrosis, 24 had lung emphysema, 15 pulmonary fibrosis, 9 alpha-1-PI deficiency and 15 suffered from various other conditions. See Table 2.3 for details. The majority of patients taking part in this study provided both alveolar epithelial cell and BALF samples.

2.0.2 Study Design

LTx patients were grouped according to their BOS status. An FEV_1 below the International Society for Heart and Lung Transplantation (ISHLT) criteria for BOS ($FEV_1 \leq$

80% of baseline) was considered persistent, if all consecutive FEV_1 measurements over a period of at least one year also showed an $FEV_1 \leq 80\%$. Thereby, and through comprehensive follow-up, other factors (e.g. infection, bronchial stenosis, acute rejection) were excluded as the main underlying reason for FEV_1 decline. Alveolar epithelial cell samples taken after a persistent FEV_1 measurement ($\leq 80\%$) were considered BOS-positive (19 samples from 13 patients). Alveolar epithelial cell samples were considered BOS-negative if they were taken at least two years before an irreversible drop to $FEV_1 \leq 80\%$ (35 samples from 24 patients). Samples not complying with these strict criteria were labeled as BOS-unclassified (23 samples from 16 patients). BALF samples for which a FEV_1 measurement ($\leq 80\%$) became persistent within one year after bronchoscopy, were considered BOS-positive (16 samples from 10 patients). Samples were considered BOS-negative if they were taken at least three years before an irreversible drop to FEV_1 $\leq 80\%$ (38 samples from 26 patients). Samples not complying with these strict criteria were labeled as BOS-unclassified (92 samples from 46 patients). See Table 2.4 for deatils.

Table 2.1. Clinical Data Microarrays (Patients)

Patient demographics		**Hazard ratio (95% Cl)**	**p-value**
Subjects	53		
Sex M/F	27/26	1.54 (0.78 - 3.04)	0.22
Age (yr) at Ltx	45 ± 12	1 (0.97 - 1.03)	0.97
Diagnosis pretransplantation			
Alpha-1-PI deficiency	6	1.48 (0.56-3.88)	0.43
Cystic fibrosis	13	1.35 (0.62-2.94)	0.46
Lung emphysema	16	0.9(0.4-2.02)	0.8
Pulmonary fibrosis	8	0.73 (0.25- 2.11)	0.56
Various conditions	10	0.71(0.25 - 2.04)	0.52
Transplant type			
Single	7	0.36 (0.05-2.67)	0.32
Bilateral	70	2.76 (0.37-20.27)	0.32
Known BOS development			
No	41		
Yes	36		

CI, confidence interval; M, male; F, female; BOS, bronchiolitis obliterans syndrome; LTx, lung transplantation; yr, years.

Table 2.2. Classwise Clinical Data Microarrays (Samples)

Class distribution				
				p-value
Classes	**BOS-positive**	**BOS-negative**	**Unclassified**	(BOS+/BOS-)
				(Wilcoxon)
	13	24	16	
Transplant type				
Single	0	3	1	
Bilateral	13	21	15	
Diagnosis pretransplantation				
Alpha-1-PI deficiency	1	3	2	
Cystic fibrosis	3	5	5	
Lung emphysema	5	7	4	
Pulmonary fibrosis	2	3	3	
Various conditions	2	6	2	
Sex M/F	7/6	11/13	9/7	
Age (yr) at Ltx	48 ± 13	45 ± 12	42 ± 13	0.27
Samples	19	35	23	
days after lung transplant	576 ± 149	615 ± 119	633 ± 163	0.09

M, male; F, female; BOS, bronchiolitis obliterans syndrome; LTx, lung transplantation; yr, years.

Table 2.3. Clinical Data Mass Spectrometry (Patients)

Patient demographics		**Hazard ratio (95% Cl)**	**p-value**
Subjects	82		
Sex M/F	40/42	1.54 (0.78 - 3.04)	0.22
Age (yr) at Ltx	44 ± 13	0.98 (0.96-1.01)	0.24
Diagnosis pretransplantation			
Alpha-1-PI deficiency	9	1.66 (0.68-4.03)	0.26
Cystic fibrosis	19	1.86 (0.9-3.84)	0.09
Lung emphysema	24	0.58(0.25-1.34)	0.2
Pulmonary fibrosis	15	0.82 (0.34-1.99)	0.66
Various conditions	15	0.67(0.23-1.91)	0.45
Transplant type			
Single	9	1.25 (0.43-3.55)	0.68
Bilateral	73	0.8(0.28-2.29)	0.68
Known BOS development			
No	48		
Yes	34		

CI, confidence interval; M, male; F, female; BOS, bronchiolitis obliterans syndrome; LTx, lung transplantation; yr, years.

Table 2.4. Classwise Clinical Data Mass Spectrometry (Samples)

Class distribution				
				p-value
Classes	**BOS-positive**	**BOS-negative**	**Unclassified**	(BOS+/BOS-)
				(Wilcoxon)
	10	27	67	
Transplant type				
Single	2	2	9	
Bilateral	8	25	58	
Diagnosis pretransplantation				
Alpha-1-PI deficiency	1	5	7	
Cystic fibrosis	3	6	16	
Lung emphysema	3	8	18	
Pulmonary fibrosis	2	3	13	
Various conditions	1	5	13	
Sex M/F	4/6	11/16	34/33	
Age (yr) at Ltx	43 ± 15	44 ± 12	43 ± 14	0.92
Samples				
days after lung transplant	579 ± 193	412 ± 143	477 ± 179	0.03

M, male; F, female; BOS, bronchiolitis obliterans syndrome; LTx, lung transplantation; yr, years.

2.1 Gene Expression Profiling

First I will introduce the steps performed for the microarray analysis (gene expression profiling).

2.1.1 Bronchial brush specimens

During bronchoscopy epithelial cells were obtained from the airway mucosa using a sheated bronchial specimen brush. This bronchial brush was advanced through the operating channel of the bronchoscope, and pushed forward and back after being positioned in a bronchial segment. The brush was retracted into the tip before removal from the bronchoscope. This procedure was repeated five times, each time in a different bronchial region. All obtained epithelial cells were removed from the brush by shaking in saline solution and stored at −80 °C [22].

2.1.2 Microarray Profiling

RNA was extracted from the epithelial cells and replicated by polymerase chain reaction (PCR) to obtain a sufficient amount of RNA for a high quality gene expression analysis [182] The expression was measured using two-colour, whole-genome, cDNA microarrays, for which RNA (labelled using Cy3) from an epithelial cell sample is co-hybridized against reference RNA (labelled using Cy5) of the first cell sample obtained after LTx from the respective patient. Thus all expression values are ratios relative to those first samples. These reference samples were taken during the first year post lung transplantation. The experiments were performed as described in [183,184,22], using the human cDNA chips of the Stanford Functional Genomics Faculty [185]. After microarray experiments, the fluorescence intensities of Cy5 and Cy3 were measured on a GenePix 4000 dual-laser scanner (Axon instruments, Foster City, Ca, USA) and the GenePix Pro 4.1 imaging software (Axon instruments). This software allows for the extraction of intensities for each printed cDNA location from the microarray scan. All 43000 spots, on each of the chips built using the Resgen Clone set (Stanford Functional Genomics Facility, Stanford, CA, USA), were quality controlled by the software. The software was able to flag a spot as having bad quality because of shape, size and several other parameters based on spot intensity [186]. Only probes flagged as good in more than 50 percent of all cases were used for further analysis. Normalization was performed using the vsn (variance stabilization and normalization) method [89]. The non-redundant gene-oriented UniGene clusters [187] contain GenBank sequences that represent a unique gene [188]. The expression of multiple probes with the same Unigene cluster ID was averaged (median) to represent genewise expression [189,190].

2.1.3 BOS-Specific Transcriptome Alterations

To filter directly for BOS-specific genome alterations between BOS-positive and BOS-negative samples differential expression was assessed using the rank product test [94,104].

The p-values were adjusted using Benjamini-Hochberg false discovery rate (FDR) [105] and adjusted p-values ≤ 0.05 were considered signifcant. In addition to the p-value, genes were only considered as up-regulated for a mean and median FC of $log_2(FC) > log_2(1)$ (down-regulated for a mean and median FC of $log_2(FC) < log_2(1)$).

2.1.4 BOS-Specific Pathways

After selecting all genes significantly regulated between BOS-positive and BOS-negative samples, a functional analysis was performed using the DAVID online resource [112]. Here KEGG pathway groups were of special interest, since those represent the interaction of a gene set in relation to a specific biological process or disease. The obtained results reflect the significant overrepresentation of KEGG groups [110] among the list of significantly deregulated genes. This was was assessed by comparisson to the entire list of analysed and annotated genes on the microarray chips.

BOS-Pathway Extension

While the processes causing BOS development might be similar to the detected pathways, there are still differences in the exact genes/proteins involved in disease progression. I took this idea into consideration by searching for genes functionally related to the members of the enriched KEGG pathways. This was done using the DAVID functional classification tool. In addition to the basic enrichment analysis described so far, DAVID also implements the functional classification algorithm for functional similarity clustering of genes [191,192]. It uses over 75,000 terms from 14 functional annotation databases to generate a gene-to-gene similarity matrix of shared functional annotation. This similarity matrix is used by a heuristic procedure to partition the genes into the optimal number of k functionally related genes clusters. Genes showing a weak relationship to the remaining genes are removed by the algorithm. The heuristic follows a fuzzy [193] approach, allowing a gene to pariticipate in more than one cluster. To give some guidance in the interpretation of the obtained results an enrichment score is calculated for each cluster. Herefore the geometric mean of the involved functional terms enrichment p-values is calculated. The minus log transformation of this geometric mean gives the enrichment score. Even if a high score represents biological importance, it should not be used for a hard decision concerning a clusters relevance. The final decsion should also take into consideration a more global picture of the expected biology. Using the predifined quality settings, I selected the clusters which showed a Gene Enrichment score (GE) above a user defined threshold and contained at least one member of the enriched KEGG pathways. Genes from these clusters were used to extend the significantly enriched KEGG pathways ("pathway extension").

2.1.5 BOS Risk Monitoring

BOS-free time was defined as the time difference between the date of bronchoscopic sampling and a persistent $FEV_1 \leq 80\%$. For samples taken before an irreversible drop to $FEV_1 \leq 80\%$, the mean follow-up time was 573 ± 493 days, and the follow-up for statistical analysis was limited to 3 years after bronchoscopy. For multivariate analysis of the relation between BOS progression and the respective molecular processes, an integrative random survial forest (RSF) was constructed [163,194]. This model incorporated all KEGG pathway related (direct KEGG members and "pathway extension") genes and the FEV_1 score. The RSF method is an ensemble classifier that consists of many survival trees. Each tree is independently constructed using a bootstrap sample of the data. An ensemble cumulative hazard estimate is obtained by combining information from all trees. At each bootstrap iteration, the data not in the bootstrap sample is used to estimate the error rate by concordance error (1 - C-index) [194]. As the KEGG data base describes interactions of genes in regard to a specific disease/pathway, the RSF methodology was perfectly suited to our needs. This is due to the idea that it automatically models the nonlinear interactions between the variables of interest and returns a variable importance measure (VIMP) for each feature [194,195,163,196].

Variable Ranking

Analogous to the mean decrease in accuracy(MDA) measure introduced for random forest (RF) [155], a variable importance (VIMP) [194,195,163,196] can also be calculated for random survival forests (RSF). In the original definition variable importance is calculated by permuting a variable and then calculating the change in prediction error. For RSF random node assignment is used instead [195,163]. For a variable x, this works as follows. When a case is dropped down a tree it is randomly assigned to a daughter node whenever the tree splits on x. The predictions from the randomized trees are obtained and averaged over the entire forest. Now the VIMP is defined as the prediction error from the noised up forest minus the error of the original forest. As for standard RF a positive VIMP indicates an informative variable and is calculated using the OOB error. The error itself is calculated according to the C-index, which estimates the probability of correctly classifying two cases. Cases with a higher predicted risk should thus also show a shorter BOS free time. Variables with a positive VIMP > 0 were considered potentially informative in regard to BOS progression, while variables with a VIMP ≤ 0 were considered uninformative. A variable with an importance value higher than the absolute value of the lowest negative-scoring variable was considered significantly informative and important. This is based on the rational that the importance of irrelevant variables varies randomly around zero [197]. Further for each of the potentially informative features univariate analysis was performed by univariate Cox proportional hazard models.

Interactions as Identified by RSF

It is a recent idea that for discovering biological interactions [198] the interaction parameters in statistical modeling should be jointly interpreted with the main effects [199]. In traditional parametric modeling these are usually taken to mean the product of two variables in a model [200]. Random forest and other tree based methods offer the advantage that the structure of decision trees gives a broader notion of interaction. For such trees interaction means the ability to model the outcome differently over subgroups defined by a trees partitioning of the data space [200]. As a tree based method random survival forests can thus be used to rate the interactions between the potentially informative features (VIMP > 0) of interest [194,195,163,196]. The interaction for a pair of variables (u and w) is hereby analysed using the concept of a joint VIMP. This joint VIMP (υ_f) is defined as the difference between the OOB prediction error when both predictor variables are noised up and the prediction error without noising up. The sum of the two single variables VIMPs ($\upsilon_u + \upsilon_w$) is called the additive importance or additive VIMP. A large absolute difference between the joint VIMP and the additive VIMP ($|(\upsilon_u + \upsilon_w) - \upsilon_f|$) indicates a potential interaction between the two variables. In the case of two highly correlated and influential variables u and w, the selection of one variable early in the tree growing process likely precludes the other variable from being selected. In the example case that 50 percent of the trees were grown using u, with the remaining trees being grown using w. Now noising up u and w individually would influence only half of all trees, while noising up both simultaneously would influence the entire set of trees. In this case the joint noising up of u and w would lead to a subtantially higher increase in prediction error, as compared to the additive decrease in prediction error. Thus the difference between the joint VIMP and the additive importance would be positive ($(\upsilon_u + \upsilon_w) > \upsilon_f$). A negative difference, on the other hand, can be observed if there is a high overlap between the subtrees of u and w. In the case of u being noised up, the additional noising up of w would not lead to a substantial decrease in prediction error and vice versa. This is due to the idea that the noising up of a variable u (or w) already influences all splits made further down the tree, including splits made on w (or u). Consequentially in the case of high overlap a negative difference between the joint VIMP and the additive importance ($(\upsilon_u + \upsilon_w) < \upsilon_f$) would be detected. Thus highly positive differences reflect the idea that both variables are correlated, while highly negative values hint to a complementary interaction with respect to BOS progression. This is due to the idea that after splitting on one variable, a split based on the specific complementary variable is most likely to further improve prediction. A concept with is reflected in the overlapping subtrees. The differences of interest, between the joint VIMP and the additive importance, were calculated for all pairs of potentially informative variables (VIMP > 0).

Interpretation of Interactions as Identified by RSF

A new approach was developed in this thesis to make the concept of variable interaction, as identified by a random survival forest (RSF) model, more applicable to biological research. To obtain an interpretable interaction score, the differences between joint VIMP and the additive importance were normalised by Z score transformation [201,202,87,203]. Using these interaction scores two graphs [204,205] were constructed. The nodes represent variables and the absolute values of the interactions scores (|interaction score|) serve as edge adjacency weights [206]. This was done for positive and negative interaction scores separately. For these graphs only positive (negative) interaction scores were considered as edges. In the case of negative (positive) interaction scores no edge was introduced. The first graph (positive interaction scores) helps to represent correlation between variables with regard to BOS development. The second graph (negative interaction scores) allows for the characterization of the complementary information between predictive variables. The graph backbone [207] composed of the most important edges of the network can be identified in the form of a minimum spanning tree [208]. This allows for the identification of the network substructure that best reflects the principles behind BOS development and progression. A spanning tree for each graph, which connects all the edges together, was constructed using Prim's algorithm for finding minimum spanning trees in weighted graphs [209,206]. The resulting minimum spanning tree showed a sum over all distance weights ($\frac{1}{adjacency}$) of the edges in that spanning tree less than or equal to the sum of distance weights for every other spanning tree [209]. To identify the most important variables (nodes) in each minimum spanning tree the connectivity of each node was assessed by its node strength. This is defined as the sum of the adjacency weights of edges connected to the node [210,211]. The nodes with a relative node strength of ≥ 0.8 , as compared to the node with the highest node strength, were considered hub nodes of their respective minimum spanning tree. Hub nodes in the first tree graph (positive interaction scores) are considered to be the predictors which best represent the set of predictive variables. Hub nodes from the second tree graph (negative interaction scores) offer supplementary information to other predictive variables. The graph was visualized using the yEd graph editor software (yWorks).

2.1.6 Risk Groups

Both random forest and random survival forest offer the possibility to calculate pairwise proximities between observations (samples). Two observations are considered similar if they often fall in the same terminal nodes. Thus a proximity measure can be calculated by the fraction of two observations being assigned to the same terminal node. After calculating the distance (1-proximity) between the observations, the partitioning around medoids (PAM) [174] algorithm was used to cluster the patients (BOS-negative and BOS-unclassified) into two risk groups (high risk and low risk)[174,10]. The difference in BOS free progression between both groups was visualized using Kaplan Meier curves [164], with significance of differences in BOS free progression being tested by a log rank test [164]. For each of the genes selected as potentially informative according to the RSF model (VIMP > 0), significance of differential expression between the two risk groups was assessed by a Wilcoxon rank sum test.

2.1.7 Quantitative Real Time PCR

Differential expression of selected genes was further validated by quantitative real time PCR (qRT-PCR) (Solaris Assay, Thermo Fisher Scientific). This procedure, which is the established method for the validation of microarray results [98], was performed using a Light Cycler device (version 2.0, Roche, Penzberg, Germany). Detailed information about the procedure can be found in [Skawran2008a]. The relative quantification was determined according to the $\Delta\Delta$ - CT method, using the LightCycler software (version 4.05). The TATA-box binding (TBP) gene was used as an internal control, and the RNA from the first post LTx sample of the respective patient was used as reference. The results from a fold change based rank product (RP) test were considered validated if the qRT-PCR values correlated significantly with the microarray data (spearman rank correlation test, rho > 0, p-value ≤ 0.05) and showed a (median) $log_2(FC) > log_2(2)$ between the groups of interest. Differential expression, as detected by a Wilcoxon rank sum test,was considered validated if the RT-PCR correlated significantly with the microarray data(spearman rank correlation test, rho > 0, p-value ≤ 0.05) and showed a significant differential expression (Wilcoxon rank sum test, p-value ≤ 0.05) between the groups of interest. Concordance between RT-PCR and microarray expression values was further assessed using the concordance index (C-index) [212,194,95]. A C-index around 0.7 was considered as an acceptably good concordance between the results [213].

2.2 Proteome Screening

This section details the steps performed for the MALDI-TOF mass spectrometry profiling (proteome screening).

2.2.1 MALDI-TOF Profiling

After BALF (bronchoalveolar lavage fluid) sample delivery on ice, cells and mucus were removed by centrifugation. Pellet and supernatant were stored at $-80\,^{\circ}$C. All samples went through one freeze-thaw cycle before proteomic analysis. Supernatant proteome constituents were isolated using superparamagnetic particles (MB-WCX,Bruker Daltonics, Bremen, Germany). Samples were processed in duplicate according to modified manufacturers protocols, taking into account the high salt concentration and low protein content (mean 0.1 ± 0.08 μg/μL). A linear MALDI-TOF mass spectrometer (MicroflexLT, Bruker Daltonics) was used for profile spectra acquisition. Spectra were preprocessed, as described in Chapter 1.6.3, using the PROcess R-package [129]. Outlier spectra showing elevated/decreased intensity sum or variance over all m/z values were selected by boxplot analysis [214,215]. If visual inspection of these spectra confirmed the bad quality, these were removed from further analysis [124]. For further anaylsis and data mining the mean (median) spectrum of the remaining technical replicates for each sample was calculated [216].

2.2.2 BOS-Specific Proteome Alterations

BOS normally occurs after the first postoperational year [217,218]. Thus I used the rank product (RP) test [94,104] on samples taken before a persistent $FEV_1 \leq 80$ percent, comparing those obtained during the first postoperational year to samples from later years. When filtering directly for BOS-specific proteome alterations between BOS-positive and BOS-negative samples. To adjust for the significant difference in time after LTx between BOS+/BOS- (Table 2.4), comparisons were restricted to samples collected within a timeframe of one year. The p-values were adjusted using false discovery rate (FDR) [105], and adjusted p-values of p ≤ 0.05 were considered significant.

2.2.3 BOS Predictor Model

For BOS classification, I constructed a random forest (RF) predictor model [155,219,220] based on all significantly regulated peaks (BOS-negative vs BOS-positive and after the first postoperational year). The RF method is an ensemble classifier that consists of many decision trees. Each tree is independently constructed using a subsample of the data. For our analysis the size of the subsample selected from each group was equal to 90% of the smallest class (BOS+). Each subsample was drawn without replacement [221]. A random forest predictor score (RF-score) based on the percentage of trees voting for a specific class, was used for classifying a sample. A simple major vote (RF-score $\geq$ 50%) defined class assignment. At each bootstrap iteration, the data not in the bootstrap sample is used to estimate the error rate. The mean error estimation over all bootstrap iterations is referred to as the out of bag (OOB) error.

Variable Ranking

The random forest method allows for the assesment of a variables importance by looking how much prediction error increases when the OOB data for that variable is permuted while all others are let unchanged. The necessary calculations for this mean decrease in acccuracy (MDA) measure are performed tree by tree. Let T_b be the tree classifier constructed for the bootstrap sample Λ_b. First drop the OOB observations corresponding to Λ_b down the tree T_b, record the resulting classifications, and compute the OOB error rate, $PE_b(OOB)$. Next, randomly permute the OOB values on the jth variable X_j while leaving the data on all other variables unchanged. If X_j is important, permuting its observed values will reduce our ability to successfully classify each of the OOB obseravtions. We then drop the altered OOB observations down the tree T_b, record the resulting classifications, and compute the OOB prediction error rate, $PE_b(OOB_j)$, which should be larger than the error rate of the unaltered data. A raw T_b-score for X_j can be computed by the difference between those two OOB error rates,

$$raw_b(j) = PE_b(OOB_j) - PE_b(OOB), b = 1, 2, \ldots, B \tag{2.1}$$

Finally, we average the raw scores over all the B trees in the forest,thus obtaining the MDA score for the jth variable:

$$MDA(j) = \frac{1}{B}\sum_{b=1}^{B} raw_b(j) \tag{2.2}$$

A high positive MDA value signifies an important variable, while a score ≤ 0 describes uninformative variables. This procedure is repeated for each variable, allowing us to rank the variables according to their importance in a classification setting. A common procedure is to scale the MDA values by Z transformation [87], but in the context of this thesis the unscaled values are used. As shown in [222] unbalanced class size can introduce a bias to the calculated MDA. To make the MDA score unbiased by class size, a separate MDA for each respective class was calculated. The final MDA was then obtained by averaging (mean) over the classwise MDAs.

Variable Selection

Training the random forest on the BOS-positive and BOS-negative samples, the peaks were chosen such that the area under curve (AUC) [222] of the random forests out of bag (OOB) predictions was minimized and peaks considered uninformative according to their RF-based mean decrease inaccuracy score (MDA ≤ 0) were excluded (Listing 2.1). For this an RF model was learned using only the highest ranking (low RP) peaks from each of the four lists of interest (up/down BOS+/BOS- and up/down after the first postoperational year). Subsequent models were learned on a set of peaks increased by the next highest ranked peak from each list. This was repeated as long as it also led to an increase in the AUC of the OOB predictions. Peaks with MDA ≤ 0, according to the RF model learned using the respective peaks, were removed from the list showing the highest AUC. The remaining selected peaks were used to learn the final RF classifier. An unbiased estimation of classifier performance was obtained by a specifically adapted cross-validation scheme, with all feature selection steps using class information being included in the validation.

Listing 2.1. This listing describes the algorithm behind the newly developed feature selection approach.

```

Select the feature lists of interest.

Order the features (peaks) in each separate list according
to their RP score (from lowest to highest).

$p = 1$

Select the $p$ highest ranked features
from each list of interest.
Call this set of features $G$.

   Fit a random forest $F$ to the data using $G$.
   $A_{new}$ = calculate the AUC of $F$ based on the $OOB_F$

```

```
  Call the best current feature set $G^o$
  $G^o = G$

  Call the best current random forest classifier $F^o$
  $F^o = F$

continue = TRUE

while(continue == TRUE)

    $p = p + 1$

    Augment $G$ to include the $p$ highest ranked
    genes from each list of interest.
    Call this new set $G^+$.

    Fit a random forest $F^+$ to the data using $G^+$.
    $A_{new}$ = calculate the AUC
    for $F^+$ based on the $OOB_F^+$

    if  $A_{new} > A_{old}$

            $A_{old} = A_{new}$
            $G^o = G^+$
            $F^o = F^+$

    else

            continue = FALSE

    end if

end while


  Calculate the unscaled $MDA$ of each feature
  in $G^o$ according to $F^o$ .
  Remove the genes with  $MDA \leq 0$ from $G^o$
  Call this set $G^*$.

  Fit the final random forest $F^*$ to the data using $G^*$.
```

Cross-Validation

To get an estimate of classifier performance, I used a cross-validation approach that included all feature selection steps using class information. At each iteration all samples from a respective patient (BOS-positive, BOS-negative and BOS-unclassified) were not part of the training set. Thus the RF predictor score (RF-score) between 0% (= BOS-negative) and 100% (= BOS-positive) assigned to each sample (including the ones labelled as BOS-unclassified) during cross-validation, is not biased by previous knowledge of the patient/sample in question. Discriminative power of BOS-positive and BOS-negative classification was evaluated using acuracy, sensitivity, specifity and a receiver operating characteristic (ROC) curve [149]. Since the assigned RF scores were unbiased, they were used for further validation by risk assessment on the samples taken before an irreversible drop to $FEV_1 \leq 80\%$.

2.2.4 BOS Risk Monitoring

BOS-free time was defined as the time difference between date of bronchoscopic sampling and a persistent $FEV_1 \leq 80\%$. For samples taken before an irreversible drop to $FEV_1 \leq 80\%$, the mean follow-up time was 877 days ± 388, and the follow-up for statistical analysis was limited to 3 years after bronchoscopy. Kaplan-Meier plots were used to visualize the BOS-free time distributions, and the significance of differences in BOS-free time was evaluated by log-rank tests. Univariate and multivariate analysis of factors significantly related to BOS-free time was performed using Cox proportional hazard models.

2.2.5 Identification of BALF Proteome Constituents

Mass identification was enabled by MALDI-TOF/TOF analysis and in-gel tryptic digestion followed by peptide mass fingerprinting. MALDI-TOF/TOF peptide and protein fragment analysis was performed with an Ultraflex I device (Bruker Daltonics). Identification of masses beyond the MALDI-TOF/TOF technical limit of about 3-5 kDa was achieved by horizontal SDS-PAGE of BALF samples (precast 12.5% CleanGels, proprietary neutral SDS buffer system; ETC Electrophorese-Technik, Kirchentellinsfurt, Germany). Gels were stained with colloidal Coomassie (Carl Roth GmbH, Karlsruhe, Germany) and prominent bands up to 170 kDa excised for in-gel tryptic digestion according to standard protocols. Peptide mass mapping using the MASCOT search engine (Matrix Science, London, UK) complemented by MALDI-TOF/TOF data was used to analyze the tryptic digests. Peak identification was further supported [223] with the ExPASy TagIdent tool (SIB, Basle, Switzerland) by database search for unique exact mass matches (pI range 0-14, Mw range 0.5 %). All identification results were further supported by pearson correlation based agglomerative hierarchical clustering.

2.2.6 Enzyme-Linked Immunosorbent Assay

The diagnostic immunoassay for clara cell protein (CCP) was conducted according to the manufacturers instructions (BioVendor, Heidelberg, Germany). Each of the measurements was normalized against the baseline value (first available sample) for the respective patient. BALF samples were divided into two groups. Group 1: ISHLT BOS stage progression in relation to the previous sample (BOS0 -> BOS1 -> BOS2 -> BOS3). Group 2: No ISHLT BOS stage progression in relation to the previous sample. For each sample, the percentage of change in protein concentration in relation to the previous sample was calculated. Significance of detected changes was assessed using the Wilcoxon rank sum test.

2.3 Block Maxx

This thesis presents a novel strategy that clusters the samples based on the consistent up- and down-regulation of a subset of genes (Figure 2.1). This method, using spectral bipartite graph partitioning [224,225,226], allows for the removal of genes/samples that could not be assigned clearly to any of the discovered clusters. My approach is basically a three step procedure: (1) gene filtering (2) clustering (3) assessment of cluster stability/importance. Gene filtering is an integral part of cluster analysis, and strongly affects the clustering results [172]. This is especially important when the number of genes m is many multitudes larger than the number of samples n [172]. Approaches to gene filtering are often based on thresholding the mean or variance [227]. Thus an optimal variance threshold was determined by a combination of hierarchical clustering and cophenetic correlation [228]. The clustering results are judged by their stability according to a clusterwise subsampling strategy [229]. This subsampling strategy is also used to determine each genes importance for a specific cluster. The method offers superior stability and the results are still easy to visualize as heatmaps. Both properties enhance its applicability in day to day biological research. To evaluate the algorithm it was compared to a traditional clustering algorithm (PAM [135,229]), an algorithm that is able to remove outlier samples (trimmed-k-means [175,230,229]), an approach using feature weighting (random forest (RF) clustering)[220,174]. All algorithms were evaluated on a microarray data set of mature aggressive B-cell lymphomas [231].

2.3.1 Data Representation and Gene Filtering

The Block Maxx approach was developed for the analysis of high throughput data sets. These are usually in the form of a set of experiments that describe the properties of specific biological entities (i.e. patients, cell lines). Each experiment is defined by a set of attributes (i.e. gene expression, protein concentration, gene copy number) associated with the entities. Such data can easily be represented as a data matrix (A_{ij}) with one row for each attribute (feature) and one column for each entity. In the following the attributes (features) used are genes while the entities of interest will be referred to as samples. Since for most high througput methods only a subset of the features is

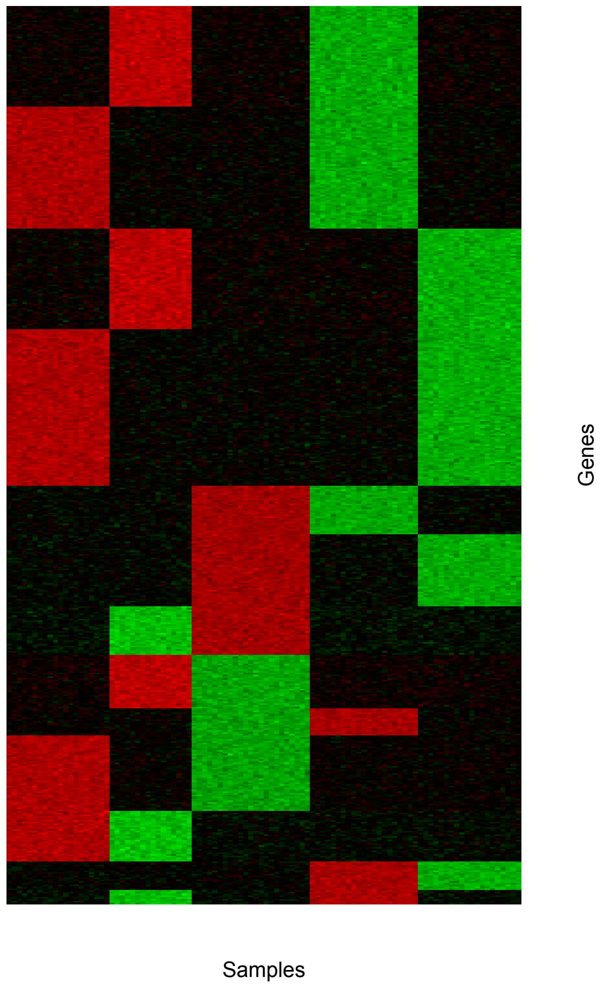


Figure 2.1. A simplified representation of the cluster structure assumed by Block Maxx. For this heatmap red depicts high expression and green shows low expression.

informative, an unsupervised feature selection approach was implemented. The method is based on agglomerative hierarchical clustering [135]. This makes it independent of the number of clusters k. The quality of a selected subset of genes was rated by how well the resulting dendrogram, as generated by hierarchical clustering, represents the patterns hidden in the data. This can be assessed by the cophenetic correlation coefficient[228].

Cophenetic Correlation

Cophenetic correlation rates how faithfully a dendrogram preserves the pairwise distances between the original unmodeled samples (data points). Given an original set of samples X, which has been modeled using a hierarchical clustering method to produce a dendrogram D, the cophenetic correlation can be defined by comparing the euclidean distance $d_2(x_i, x_j)$ between the ith and jth sample with the respective dendrogrammatic distance. The dendrogrammatic distance $t(i,j)$ between the model points of these samples is defined as the height of the dendrogramm node at which the two points are first joined together. Denotig the average of the $d_2(i,j)$ by d and the average of the $t(i,j)$ by t, the cophenetic correlation coefficient c is defined by:

$$c = \frac{\sum_{i<j}(d_2(i,j) - d)(t(i,j) - t)}{\sqrt{[\sum_{i<j}(x(i,j) - x)^2][(t(i,j) - t)^2]}} \tag{2.3}$$

Optimal Gene Set

The variance of gene expression is often used to determine highly informative genes [232]. Thus all genes are ordered in a decreasing manner, according to their expression variance over all samples. The 5 genes with the highest variance are selected to generate the associated dendrogramm of samples, which is rated using the cophenetic correlation coefficient. This is repeated iteratively, sequentially adding 5 more genes , till the cophenetic correlation coefficient c no longer increases. The gene set resulting in the highest cophenetic correlation coefficient is used for all following calculations.

2.3.2 Basic Principles and Data Transformation

The Block Maxx approach assumes that the data matrix contains a structure with k biclusters. Each bicluster is defined by a subset of samples (columns), and a subset of genes (rows). The grouped genes are consistently up- or down-regulated in the respective set of samples. A cluster can simultaneously contain up and- down-regulated genes. Each sample is assigned to exactly one cluster. Genes can be consistently up-regulated in one cluster, while being down-regulated in another cluster. Thus a gene can be a member of up to 2 clusters. Genes that are not differentially expressed in any of the sample clusters are assigned to a background cluster. The assignment to the background cluster is handeled separately for up- and down-regulation. Meaning that while a gene is consistently up-regulated (down-regulated) in one cluster, it is not necessarily consitently down-regulated (up-regulated) in any of the other clusters. Assignment to the background is also performed for samples, if they do not show a consistent over- or under-expression for any of the gene subsets defined by the clustering. As the Block Maxx approach is mainly based on the concept of differential expression, a numerical representation for the up- and down-regulation of a gene in a specfifc sample has to be defined.

Gene Activiation Matrix

Given a data set (A) with m genes (F) and n samples (S) the association of a gene with a specfific sample (a_{ij}) will be represented by a Gene Activation Score (GAS), instead of the raw expression values. This GAS is based on the established Z score [87], which indicates how many standard deviations an observation is above or below the mean for a specific attribute. After calculating the row mean for a gene:

$$\mu_i = \frac{1}{m}\sum_{j=1}^{m} a_{ij} \tag{2.4}$$

as well as the row wise standard deviation

$$\sigma_i = \sqrt{\frac{1}{m}\sum_{j=1}^{m}(a_{ij} - \mu_i)} \tag{2.5}$$

the z-normalized expression value can be assigned by:

$$a'_{ij} = \frac{a_{ij} - \mu_i}{\sigma_i} \tag{2.6}$$

The z transformation normalizes each row to a mean of 0 and a standard deviation of one. Since a gene can be up-regulated in one cluster and down-regulated in another one, we define an up-regulation (F_{up}) and a down regulation (F_{down}) representation for each gene (feature) ($F = (F_{up}, F_{down})$. This leads to an upregualtion $A_{up} = A'$ and a down-regulation matrix $A_{down} = -A'$.

Both matrices (A_{up} and A_{down}) are than combined as:

$$A'' = \begin{bmatrix} A_{up} \\ A_{down} \end{bmatrix} \tag{2.7}$$

For A'' the matrix entries a''_{ij} are range scaled between 0 and 1, thus calculating the final Gene Activation Score (GAS) matrix A^*.

$$a^*_{ij} = \frac{a''_{ij} - a''_{min}}{a''_{max} - a''_{min}} \tag{2.8}$$

For the GAS matrix A^* the global mean (μ^*) over all matrix entries is defined as

$$\mu^* = \frac{0 - A''_{min}}{A''_{max} - A''_{min}} \tag{2.9}$$

with the global standard deviation (σ^*) over all matrix entries being

$$\sigma^* = \frac{1 - A''_{min}}{A''_{max} - A''_{min}} \tag{2.10}$$

Defining a Background

To allow for the removal of genes/samples that can not be assigned clearly to a specific cluster, we introduce the concept of a background. Thus the matrix A^* is extended by $\frac{m}{k}$ background genes (F_{back}) and $\frac{n}{k}$ background samples S_{back}. $F^+ = F, F_{back}$ and $S^+ = S, S_{back}$. The background genes show an activation score of μ^* for the set of S and $\mu^* + \sigma^*$ for S_{back}. For the background genes F_{back} the genes F show an activation score of μ^*.

$$A^+ = \begin{pmatrix} a^*_{11} & a^*_{12} & ... & a^*_{1n} & ... & \mu^* + \sigma^* & ... & \mu^* + \sigma^* \\ a^*_{21} & a^*_{22} & ... & a^*_{2n} & ... & \mu^* + \sigma^* & ... & \mu^* + \sigma^* \\ ... & ... & & ... & & ... & ... & \\ a^*_{m1} & a^*_{m2} & ... & a^*_{mn} & ... & \mu^* & ... & \mu^* \end{pmatrix} \tag{2.11}$$

Genes and samples that are more connected to the background than any of the real clusters will be clustered together with the background samples/genes. Thus they are removed from the detected cluster structure.

2.3.3 Graph Generation and Partitioning

To simultaneously partition genes and samples into k clusters, a spectral clustering approach is used. The spectral clustering algorithm tries to find clusters such that the nodes in a graph are connected by highly weighted edges and the connections between clusters are weak. Weak connections are represented by edges with low weights. The goal is to identify these tightly coupled clusters, and to cut the inter cluster edges. Spectral clustering achieves this goal by clustering the data, using eigenvectors of a similarity/affinity matrix derived from the data set. A biclustering variant of spectral clustering, using the singular vectors of the data matrix, was described and applied by Dhillon [224,225,226]. This approach, based on a bipartite graph representation of a data set, was used to partition A^+ into k clusters.

Graph Model

The matrix A^+ is a gene-by-condition matrix, where a^+_{ij} equals the Gene Expression Scores (GAS). This can be represented as an undirected bipartite graph, with the weights given by the Gene Activation scores. The graph is then defined as $G = F, S, E)$, where $F = \{f_1, f_2, \ldots, f_m\}$ and $S = \{s_1, s_2, \ldots, s_n\}$ are two sets of vertices and E is the set of edges $\{\{f_i, s_j\} : f_i \in F^+, s_j \in S^+\}$. S^+ corresponds to the set of samples and F^+ corresponds to the set of genes. These sets (F^+ and S^+) include the background genes and samples. An edge signifies the association between a condition and a gene. The weight of the edge corresponds to a^+_{ij}. As the graph is supposed to be bipartite, there are no edges between genes and no edges between conditions. To remove edges between samples and genes, with a below average GAS, the entries $a^+_{ij} \leq \mu^*$ are set to 0. The resulting bipartite graph can be represented by the following adjacency matrix:

$$M = \begin{bmatrix} 0 & A^+ \\ A^{+T} & 0 \end{bmatrix} \tag{2.12}$$

Objective of Clustering

For an optimal partitioning the edge weight between members of a cluster should be maximized, while minimizing the edge weight between the clusters. This objective can be defined by the normalized-cut (Ncut) [233]. To define the Ncut we first introduce a specific vertex weight function, where the weight of each vertex ($V = (F, S)$ is equal to the sum of the weights of edges that are incident on it:

$$\text{weight}(i) = \sum_r E_{ir} \tag{2.13}$$

This leads us to the normalized-cut (NCut) objective [233]:

$$N(V_1, V_2) = \frac{\text{cut}(V_1, V_2)}{\sum_{i \in V_1} \sum_r E_{ir}} + \frac{\text{cut}(V_1, V_2)}{\sum_{i \in V_2} \sum_r E_{ir}} \tag{2.14}$$

Graph Partitioning

Presenting the data as a bipartite graph allows for the application of spectral graph theory [234], which can be used to partition the graph into clusters. Thus, since finding the globally minimum solution to the NCut is NP-complete [233], an approximate solution can be found by the singular value decomposstion (SVD) of A^+. The partitioning algorithm, following reference [224], proceeds as follows:

1. **Calculate A^+ and form the matrix**

$$A_n = D_1^{-1/2} A^+ D_2^{-1/2} \tag{2.15}$$

with D_1 and D_2 diagonal matrices such that $D_1(i,i) = \sum_j A^+_{ij}$ and $D_2(j,j) = \sum_i A^+_{ij}$

2. Calculate the singular value decomposition (SVD)

$$\boldsymbol{SVD}(A_n) = U * \alpha * V^T \tag{2.16}$$

and center each column of U and V to a columnwise mean of zero, by subtracting the columnwise mean. This additional step differs from the original approach described in reference [224], as it is necessary to align the backgrounds of the genes and samples with each other.

3. Construct the l-dimensional data set Z
As the $l = k$ (or $l = k - 1$) singular vectors $u_2, u_3, ..., u_{l+1}$, and $v_2, v_3, ..., v_{l+1}$ often contain k-modal information about the data set. Thus we can form the l-dimensional data set

$$Z = \begin{bmatrix} D_1^{-1/2} U \\ D_2^{-1/2} V \end{bmatrix} \tag{2.17}$$

where $U = [u_2, ..., u_{l+1}$, and $V = [v_2, ..., v_{l+1}]$.

4. Run the k-means algorithm on the l-dimensional data Z.
This obtains the desired k-way multipartitioning. Including the background cluster, k-means is used to search for $k + 1$ clusters.

Genes ($F = (F_{up}, F_{down})$ and samples (S) assigned to the same cluster, as obtained by the k-means step, are considered a bicluster. Genes and samples clustered together with the background genes (F_{back}) and background samples (S_{back}), were not considered part of a bicluster. For the genes up- (F_{up}) and down-regulation (F_{down}) representations were considered separately.

2.3.4 Parameter Optimization

As not every singular vector of the data affinity matrix (A_n) is informative, the choice of the number of singular vectors l is critical to achieve a high quality clustering result. Another important step is to define the optimal number of clusters k^*, a common problem in unsupervised clustering approaches. Before setting these parameters, a measure of cluster quality has to be defined.

Assessing Cluster Stability

To obtain a quality measure of the graph partitioning, which should reflect how well separated and homogenous the clusters are, a subsampling scheme was applied. The cluster stability was only assessed for the samples, a measure for the genes will be introduced later. Each clusters quality was assessed by its reproducibility from a sample subset of the entire data set. This clusterwise approach takes into consideration that in a single clustering only a few clusters are highly stable, while the others prove to be unstable. The data subsets used to test cluster stability are generated by bootstrapping (sampling with replacement), for which duplicates are removed from the randomly selected set. A clustering is calculated from this bootstrap subset, and each cluster is compared to the most similar cluster from the original data set. The metric used to define the similarity between two clusters is the Jaccard index. Given a set of entities (samples) $x_n = x_1, \ldots, x_n$ which are assigned to a set of disjoint cluster subsets, we can define the Jaccard index for two clusters A and B as:

$$\gamma(A,B) = \frac{|A \cap B|}{|A \cup B|}, A, B \subseteq x_n. \tag{2.18}$$

This defines the proportion of entities belonging to both clusters of all entities asigned to at least one of both clusters. Given a set of B bootstrap iterations a clustering is generated for each replicate. For a set of k clusters $C = C_i, \ldots, C_k$ from the original clustering, the maximum similarity to each of the subset clusterings $C_b^* = C_{bj}^*, \ldots, C_{bk}^*$ is calculated for each of the original clusters:

$$\gamma_{C_{ib}} = \boldsymbol{max} \gamma(C_b^*, C_i) \tag{2.19}$$

This generates a sequence $\gamma_{C_{ib}}, b = 1, \ldots, B$. Based on this sequence the mean Jaccard index

$$\bar{\gamma}_{C_i} = \frac{\sum_{i=1}^{B} \gamma_{C,i}}{B} \tag{2.20}$$

serves as a stability measure. A cluster is considered as stable for $\bar{\gamma}_{C_i} \geq 0.75$ The stability for the entire clustering C is defined as:

$$\bar{\gamma}_C = \frac{\sum_{i=1}^{k} \bar{\gamma}_{C_i}}{k} \tag{2.21}$$

A cluster (clustering) is considered as stable for $\bar{\gamma}_{C_i} \geq 0.75$ ($\bar{\gamma}_C \geq 0.75$).

Choosing k and the Number of Singular Vectors

The choice which eigenvectors or singular vectors to choose for data partitioning is still a topic of much discussion in the field of spectral clustering [235]. In this case the number of singular vectors, used for a specific number of clusters k, is decided by the cluster stability score of the samples. Either the first $2:(k)$ or the first $2:(k-1)$ left and right singular vectors are used, depending on which choice shows the higher mean cluster stability over all clusters. After choosing the singular vectors for each number of clusters k, the optimal number of clusters k^* needs to be determined. This was also done based on the clusterwise stability. As reported in [229], a cluster can be considered as stable if it shows a cluster stability score ≥ 0.75. Thus we choose the maximum number of k for which all regular clusters, excluding the background cluster, $(C_{1,\dots,k})$ show a clusterwise stabililty score ≥ 0.75.

2.3.5 Gene Membership Score

As with assessing the cluster stability for the samples, the belonging and importance of a gene to a cluster is quantified by the gene membership score (GMS). The GMS is based on the same principles used to assess the clusterwise stability. During the bootstrap iterations each cluster is assigned the same label as the most similar (samplewise) cluster from the original data set. The genes in the cluster are also labeled accordingly. Thus the GMS gives the percentage of times a gene was assigned a specific label.

$$GMS_{gc} = \sum_{c=1}^{C} B \tag{2.22}$$

The m_{gc} are binary values with $m_{gc} = 1 \; g \in C$

2.3.6 Comparisson and Visualization

To assess the usability of the Block Maxx approach, a previously published microarray study of Burkitt‘s Lymphoma and diffuse large-B-cell lymphoma was used [231]. This data set was generated using Affymetrix U133A gene chips for expression profiling of 220 mature aggressive B-cell lymphomas. On this data set Block Maxx was compared to established algorithms. The PAM (partitioning around medoids) [135] method was included as a standard clustering method. Further methods used for reference employ strategies similar to those used by the Block Maxx approach (trimmed-k-means: outlier removal [175], unsupervised random forest clustering (PAM-RF): feature weighting [174]). The biological/medical validity of the patient clusters was assessed by analysis of survival time [164]. The survival time for the specific cluster was visualized using kaplan meier curves, and statistical significance of differences in survival time was assessesed using the log-rank test. To validate if the gene membership scores actually reflect differential expression, a Wilcoxon rank sum test (one tailed) was used. For each gene and cluster combination the expression values between all samples in the cluster against all samples not in the cluster were compared.

KEGG and Biocarta Association

Validation of the gene clusters was done using the DAVID functional annotation resource [112,96]. Each clusters significant enrichment of pathways (Biocarta and KEGG [110]) was calculated. A cluster with significantly enriched functional groups was rated as biologically relevant.

Chapter 3

Results

3.1 Prediction of the Bronchiolitis Obliterans Syndrome by Transcriptomics Analysis

The bronchial epithelial cell samples of 53 lung transplanted patients were obtained from the airway mucosa using bronchial brushing [22]. The extracted RNA from 77 obtained cell samples, taken at different timepoints after lung transplantation, was replicated by PCR and gene expression was measured using two-colour, whole-genome, cDNA microarrays. For these RNA (labelled using Cy3) from an epithelial cell sample was co-hybridized against reference RNA (labelled using Cy5) of the first cell sample obtained after LTx from the respective patient. The acquired expression profiles were preprocessed and further computational analysis of BOS progression was performed.

3.1.1 Differentially Regulated Genes

After quality control and normalization the rank product test [104,94] was used to compare BOS-positive against BOS-negative samples for all 10387 selected genes. We found 649 genes to be significantly differentially expressed (441 up-regulated and 208 down-regulated) at a p-value ≤ 0.05. The 20 most significant genes from each group (up- and down-regulated) are shown in Table 3.1 a) and 3.1 b) respectively. All differentially regulated genes are shown in Appendix Table 1 and 2. Differential expression, as detected by the rank product test, was validated for 5 genes (1 up-regulated: C15orf48 ; 4 down-regulated: ALDH3A1, SCGB3A2, HBA1 and SFTB) using quantitative real-time PCR (qRT-PCR)(Table 3.2). For all genes the correlation between the microarray measurements and the qRT-PCR results was considered significant (Spearman rank correlation test, p-value ≤ 0.05). All down-regulated genes met the respective FC threshold ($log_2(FC) < log_2(0.5)$). For the up-regulated C15orf48 a tendency for up-regulation was detected (FC > 1.74), but the FC threshold ($log_2(FC) > log_2(2)$) was not fully met. The ordering of the microarray expression values, for each of the validated genes, showed acceptably good concordance [213,212,168] with the qRT-PCR results (C-index $0.76 - 0.92$).

Table 3.1. a) Genes significantly up-regulated in BOS-positive samples

	Gene symbol	UniGene	Gene name	RP	P	FDR	FC
1	SRGN	Hs.1908	Serglycin	322	< 0.001	< 0.001	2.05
2	UPK1B	Hs.271580	Uroplakin 1B	394	< 0.001	< 0.001	2.17
3	C15orf48	Hs.112242	Chromosome 15 open reading frame 48	446	< 0.001	< 0.001	2.1
4	LAPTM5	Hs.371021	Lysosomal protein transmembrane 5	468	< 0.001	< 0.001	1.87
5	CCL3L3	Hs.512304	Chemokine (C-C motif) ligand 3-like 3	478	< 0.001	< 0.001	1.88
6	FPGS	Hs.335084	Folylpolyglutamate synthase	484	< 0.001	< 0.001	1.81
7	CCL3	Hs.514107	Chemokine (C-C motif) ligand 3	622	< 0.001	< 0.001	1.71
8	CCR1	Hs.301921	Chemokine (C-C motif) receptor 1	707	< 0.001	< 0.001	1.59
9	TPSAB1	Hs.405479	Tryptase alpha/beta 1	738	< 0.001	< 0.001	1.65
10	CCL2	Hs.303649	Chemokine (C-C motif) ligand 2	738	< 0.001	< 0.001	1.58
11	IL1B	Hs.126256	Interleukin 1, beta	739	< 0.001	< 0.001	1.64
12	LY6D	Hs.415762	Lymphocyte antigen 6 complex, locus D	752	< 0.001	< 0.001	1.6
13	MXD1	Hs.468908	MAX dimerization protein 1	763	< 0.001	< 0.001	1.66
14	FCER1G	Hs.433300	Fc fragment of IgE, high affinity I, receptor for; gamma polypeptide	775	< 0.001	< 0.001	1.55
15	IFI30	Hs.14623	Interferon, gamma-inducible protein 30	776	< 0.001	< 0.001	1.58
16	SAA1	Hs.632144	Serum amyloid A1	778	< 0.001	< 0.001	1.4
17	CCNA1	Hs.417050	Cyclin A1	791	< 0.001	< 0.001	1.7
18	CENPN	Hs.55028	Centromere protein N	807	< 0.001	< 0.001	1.54
19	HLA-G	Hs.512152	Major histocompatibility complex, class I, G	817	< 0.001	< 0.001	1.41
20	PLEK	Hs.468840	Pleckstrin	818	< 0.001	< 0.001	1.45

Table 3.1. b) Genes significantly down-regulated in BOS-positive samples

	Gene symbol	UniGene	Gene name	RP	P	FDR	$\frac{1}{FC}$
1	ALDH3A1	Hs.531682	Aldehyde dehydrogenase 3 family, member A1	136	< 0.001	< 0.001	2.47
2	SCGB3A2	Hs.483765	Secretoglobin, family 3A, member 2	225	< 0.001	< 0.001	2.15
3	HBA1	Hs.449630	Hemoglobin, alpha 1	286	< 0.001	< 0.001	1.62
4	HBB	Hs.523443	Hemoglobin, beta	294	< 0.001	< 0.001	1.57
5	MID1	Hs.27695	Midline 1 (Opitz/BBB syndrome)	395	< 0.001	< 0.001	1.77
6	SFTPC	Hs.1074	Surfactant protein C	458	< 0.001	< 0.001	1.18
7	NCRNA00086	Hs.374414	Non-protein coding RNA 86	544	< 0.001	< 0.001	1.74
8	ZNF704	Hs.632067	Zinc finger protein 704	639	< 0.001	< 0.001	1.72
9	OLA1	Hs.157351	Obg-like ATPase 1	671	< 0.001	< 0.001	1.57
10	AHSP	Hs.274309	Alpha hemoglobin stabilizing protein	699	< 0.001	< 0.001	1.82
11	VAV2	Hs.369921	Vav 2 guanine nucleotide exchange factor	762	< 0.001	< 0.001	1.58
12	PIPOX	Hs.462585	Pipecolic acid oxidase	767	< 0.001	< 0.001	1.65

13	CES1	Hs.558865	Carboxylesterase 1 (monocyte/-macrophage serine esterase 1)	787	< 0.001	< 0.001	1.6
14	LTF	Hs.529517	Lactotransferrin	795	< 0.001	< 0.001	1.37
15	SCGB1A1	Hs.523732	Secretoglobin, family 1A, member 1 (uteroglobin)	814	< 0.001	< 0.001	1.32
16	SFTPB	Hs.512690	Surfactant protein B	830	< 0.001	< 0.001	1.23
17	TRMT61A	Hs.525610	TRNA methyltransferase 61 homolog A (S. cerevisiae)	864	< 0.001	< 0.001	1.62
18	F2RL2	Hs.42502	Coagulation factor II (thrombin) receptor-like 2	876	< 0.001	< 0.001	1.59
19	APOBEC2	Hs.555915	Apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like 2	895	< 0.001	< 0.001	1.41
20	ZNF700	Hs.528486	Zinc finger protein 700	937	< 0.001	< 0.001	1,41

RP, rank product; P, p-value ;FDR, false discovery rate; FC, mean fold change ; BOS, bronchiolitis obliterans syndrome

Table 3.2. Correlation and Concordance between microarray and qRT-PCR data.

Gene symbol	Rho	P (Spearman)	FC	C-index
C15ORF48	0.68	2.97×10^{-5}	1.74	0.76
ALDH3A1	0.95	$< 2.2 \times 10^{-16}$	$\frac{1}{2.36}$	0.92
SCGB3A2	0.91	$< 2.2 \times 10^{-16}$	$\frac{1}{7.02}$	0.86
HBA1	0.96	$< 2.2 \times 10^{-16}$	$\frac{1}{5.6}$	0.92
SFTPB	0.92	$< 2.2 \times 10^{-16}$	$\frac{1}{3.76}$	0.88

Rho, Spearman rank correlation coefficient; FC, median fold change; P, p-value; C-index, concordance index

3.1.2 Enriched KEGG pathways

Using functional enrichment analysis as implemented in the DAVID online resource [112], I identified several functional groups (pathways as defined in the KEGG database [110]) being significantly highly enriched in the list of differentially expressed genes. A set of 6 KEGG pathways was identified by my analysis (Table 3.3). The list of differentially expressed genes involved in the enriched pathways can be found in Table 3.4. The KEGG Disease pathway map of systemic lupus erythematosus is shown in Figure 3.1. To find other genes that are functionally related to those genes confirmed as KEGG pathway members, but not yet known members of the enriched pathways, I clustered all differentially expressed genes based on functional annotation. Using the funtional classification tool from DAVID [112] and predifined quality settings (High), I selected the clusters which showed a Gene Enrichment score (GE) ≥ 4 and contained at least one member of the enriched KEGG pathways. Thus 6 genes clustered with confirmed members of the significantly enriched KEGG pathways (Table 3.5). The genes from those clusters were selected to extend the list of pathway related genes ("pathway extension", Table 3.6).

Table 3.3. Significantly enriched KEGG pathways

KEGG	Count	FE	P	UniGene
hsa05322:Systemic lupus erythematosus	18	4.03	1.01×10^{-4}	Hs.534322 Hs.467753 Hs.69771 Hs.437275 Hs.372679 Hs.477879 Hs.520048 Hs.008986 Hs.77424 Hs.534847 Hs.351279 Hs.46423 Hs.411472 Hs.485130 Hs.387679 Hs.171182
hsa05332:Graft-versus-host disease	10	5.55	0.004	Hs.534322 Hs.512152 Hs.351279 Hs.77961 Hs.485130 Hs.387679 Hs.171182 Hs.520048 Hs.126256
hsa05320:Autoimmune thyroid disease	9	5.74	0.01	Hs.534322 Hs.512152 Hs.351279 Hs.77961 Hs.485130 Hs.387679 Hs.171182 Hs.520048
hsa05330:Allograft rejection	9	5.47	0.015	Hs.534322 Hs.512152 Hs.351279 Hs.77961 Hs.485130 Hs.387679 Hs.171182 Hs.520048
hsa04940:Type I diabetes mellitus	10	4.4	0.034	Hs.534322 Hs.512152 Hs.351279 Hs.77961 Hs.485130 Hs.387679 Hs.171182 Hs.520048 Hs.126256
hsa04612:Antigen processing and presentation	12	3.56	0.046	Hs.14623 Hs.534322 Hs.434081 Hs.512152 Hs.418123 Hs.351279 Hs.77961 Hs.485130 Hs.387679 Hs.520048 Hs.534255

P, Bonferroni corrected p-value; FE, fold enrichment

Table 3.4. Differentially expressed genes directly involved in enriched KEGG pathways

Gene symbol	UniGene	Gene name	RP	P	FDR	FC	KEGG
IL1B	Hs.126256	Interleukin 1, beta	739	< 0.001	< 0.001	1.64	hsa05332 hsa04940
IFI30	Hs.14623	Interferon, gamma-inducible protein 30	776	< 0.001	< 0.001	1.58	hsa04612
HLA-G	Hs.512152	Major histocompatibility complex, class I, G	817	< 0.001	< 0.001	1.41	hsa04612 hsa05330 hsa05332 hsa04940 hsa05322

FCGR3A	Hs.372679	Fc fragment of IgG, low affinity IIIa, receptor (CD16a)	1001	< 0.001	< 0.001	1.23	hsa05322
C1QB	Hs.8986	Complement component 1, q subcomponent, B chain	1101	< 0.001	< 0.001	1,3	hsa05322
HIST1H2BK	Hs.437275	Histone cluster 1, H2bk	1153	< 0.001	< 0.001	1.57	hsa05322
HLA-DQA1	Hs.387679	Major histocompatibility complex, class II, DQ alpha 1	1161	< 0.001	< 0.001	1.31	hsa04612 hsa05330 hsa05332 hsa04940 hsa05322
HLA-DRA	Hs.520048	Major histocompatibility complex, class II, DR alpha	1277	< 0.001	< 0.001	1.4	hsa04612 hsa05330 hsa05332 hsa04940 hsa05322
H2AFX	Hs.477879	H2A histone family, member X	1346	< 0.001	< 0.001	1.52	hsa05322
FCGR1A	Hs.77424	Fc fragment of IgG, high affinity Ia, receptor (CD64)	1355	< 0.001	< 0.001	1.32	hsa05322
HLA-DMA	Hs.351279	Major histocompatibility complex, class II, DM alpha	1407	< 0.001	0.001	1.22	hsa04612 hsa05330 hsa05332 hsa04940 hsa05322
HLA-DRB1	Hs.534322	Major histocompatibility complex, class II, DR beta 1	1493	< 0.001	0.003	1.26	hsa04612 hsa05330 hsa05332 hsa04940 hsa05322
PSME2	Hs.434081	Proteasome (prosome, macropain) activator subunit 2 (PA28 beta)	1535	< 0.001	0.003	1.36	hsa04612
C4A	Hs.534847	Complement component 4A (Rodgers blood group)	1515	< 0.001	0.011	$\frac{1}{1.26}$	hsa05322
CFB	Hs.69771	Complement factor B	1540	< 0.001	0.003	1.25	hsa05322
CD86	Hs.171182	CD86 molecule	1638	< 0.001	0.006	1.37	hsa05330, hsa05332, hsa04940, hsa05322
HLA-DPB1	Hs.485130	Major histocompatibility complex, class II, DP beta 1	1833	< 0.001	0.02	1.11	hsa04612 hsa05330 hsa05332 hsa04940 hsa05322
HIST1H4C	Hs.46423	Histone cluster 1, H4c	1859	< 0.001	0.022	1.01	hsa05322

C1QC	Hs.467753	Complement component 1, q subcomponent, C chain	1882	< 0.001	0.026	1.17	hsa05322
GRIN2A	Hs.411472	Glutamate receptor, ionotropic, N-methyl D-aspartate 2A	1957	0.001	0.036	1.3	hsa05322
CTSL1	Hs.418123	Cathepsin L1	1973	0.002	0.039	1,2	hsa04612
HLA-B	Hs.77961	Major histocompatibility complex, class I, B	2012	0.002	0.045	1.13	hsa04612, hsa05330 hsa05332 hsa04940 hsa05322
B2M	Hs.534255	Beta-2-microglobulin	2031	0.002	0.049	1.12	hsa04612

RP, rank product; P, p-value ;FDR, false discovery rate; FC, mean fold change

Table 3.5. Six functional gene groups identified by the functional classification tool

Gene Group 1	Enrichment Score: 15.48
UniGene	**Gene name**
Hs.114948	cytokine receptor-like factor 1
Hs.204238	lipocalin 2
Hs.520989	fibrinogen-like 2
Hs.591967	interleukin 18 binding protein
Hs.255462	microseminoprotein, beta-
Hs.21162	retbindin
Hs.439312	phospholipid transfer protein

Gene Group 2	Enrichment Score: 8.77
UniGene	**Gene name**
Hs.512304	chemokine (C-C motif) ligand 3-like 3; chemokine (C-C motif) ligand 3-like 1
Hs.143961	chemokine (C-C motif) ligand 18 (pulmonary and activation-regulated)
Hs.531668	chemokine (C-X3-C motif) ligand 1
Hs.32949	defensin, beta 1
Hs.57907	chemokine (C-C motif) ligand 21
Hs.514107	chemokine (C-C motif) ligand 3
Hs.77367	chemokine (C-X-C motif) ligand 9
Hs.131342	chemokine (C-C motif) ligand 26
Hs.789	chemokine (C-X-C motif) ligand 1 (melanoma growth stimulating activity, alpha)
Hs.943	interleukin 32
Hs.632586	chemokine (C-X-C motif) ligand 10

Gene Group 3	Enrichment Score: 6.78
UniGene	**Gene name**
Hs.418123	**cathepsin L1**
Hs.151254	*kallikrein-related peptidase 7*
Hs.405479	*tryptase alpha/beta 1; tryptase beta 2*
Hs.252549	*cathepsin Z*
Hs.233389	*carboxypeptidase, vitellogenic-like*

Gene Group 4	Enrichment Score: 6.46
UniGene	**Gene name**
Hs.69771	**complement factor B**
Hs.90708	*granzyme A (granzyme 1, cytotoxic T-lymphocyte-associated serine esterase 3)*
Hs.151254	*kallikrein-related peptidase 7*
Hs.69771	**complement component 2**
Hs.405479	*tryptase alpha/beta 1; tryptase beta 2*

Gene Group 5	Enrichment Score: 6.15
UniGene	**Gene name**
Hs.467753	**complement component 1, q subcomponent, C chain**
Hs.384598	*serpin peptidase inhibitor, clade G (C1 inhibitor), member 1*
Hs.534847	**complement component 4A (Rodgers blood group)**
Hs.8986	**complement component 1, q subcomponent, B chain**
Hs.534847	**complement component 4B (Chido blood group)**
Hs.69771	**complement component 2**

Gene Group 6	Enrichment Score: 4.54
UniGene	**Gene name**
Hs.1011	protein Z, vitamin K-dependent plasma glycoprotein
Hs.512587	macrophage stimulating 1 (hepatocyte growth factor-like)
Hs.151254	kallikrein-related peptidase 7
Hs.224698	protein C (inactivator of coagulation factors Va and VIIIa)
Hs.405479	tryptase alpha/beta 1; tryptase beta 2

The members of significantly enriched KEGG pathways are marked in bold. The genes clustered together with members of significantly enriched KEGG pathways are written in italic.

Table 3.6. Pathway extension: differentially expressed genes functionally linked to enriched KEGG pathways

Gene symbol	UniGene	Gene name	RP	P	FDR	FC
TPSAB1	Hs.405479	Tryptase alpha/beta 1	738	< 0.001	< 0.001	1.59
CTSZ	Hs.252549	Cathepsin Z	1387	< 0.001	< 0.001	1.44
SERPING1	Hs.384598	Serpin peptidase inhibitor, clade G (C1 inhibitor), member 1	1543	< 0.001	0.003	1.24
CPVL	Hs.233389	Carboxypeptidase, vitellogenic-like	1662	< 0.001	0.007	1.18
GZMA	Hs.90708	Granzyme A (granzyme 1, cytotoxic T-lymphocyte-associated serine esterase 3)	1681	< 0.001	0.008	1.08
KLK7	Hs.151254	Kallikrein-related peptidase 7	1792	< 0.001	0.015	1.37

RP, rank product; P, p-value ;FDR, false discovery rate; FC, median fold change

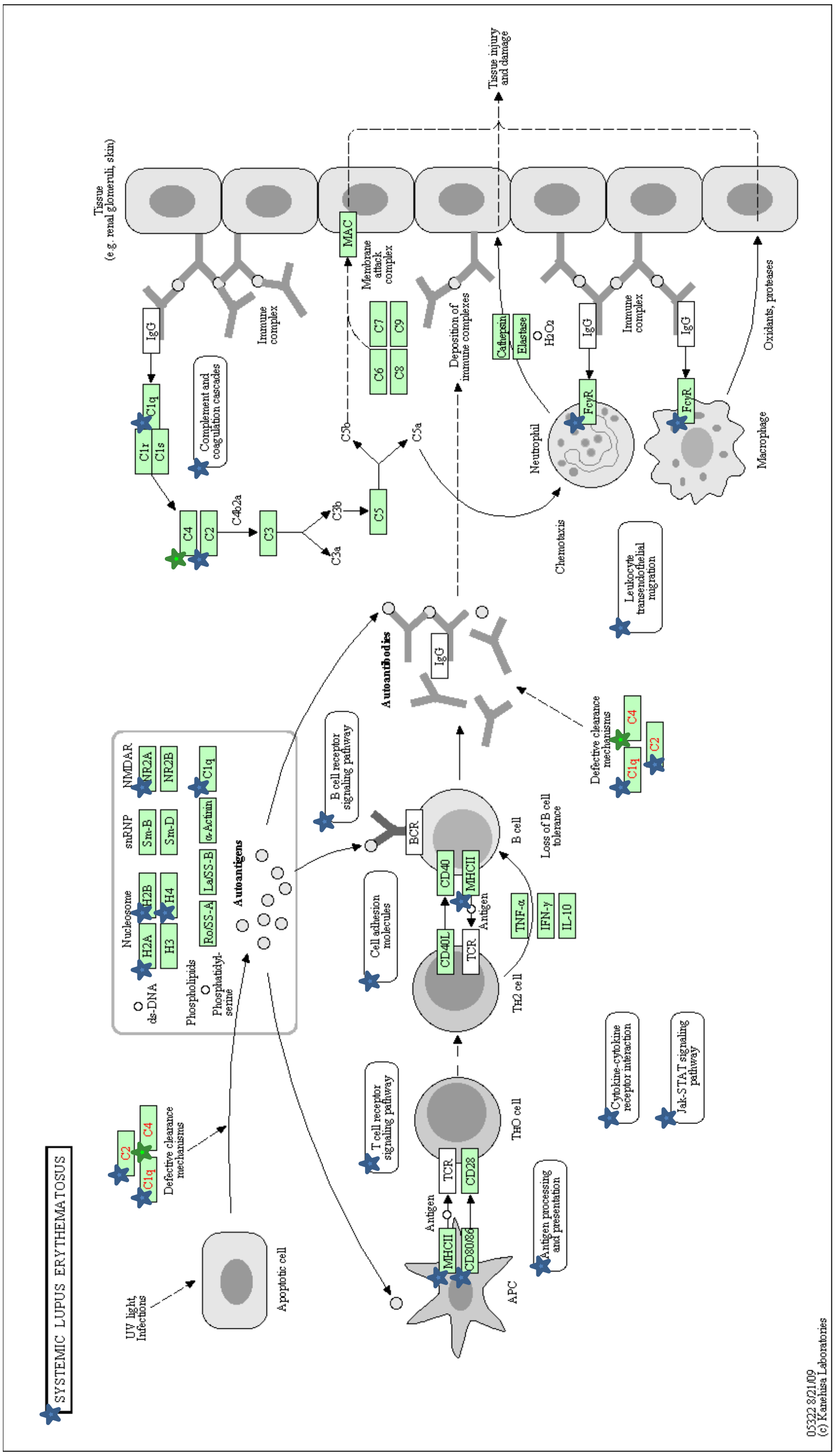


Figure 3.1. The disease pathway map of Systemic Lupus Erythematosus as represented in the KEGG Disease database. In the list of genes significantly deregulated in the BOS+ samples, genes belonging to this particular pathway were found to be significantly enriched. Those genes are marked by stars (blue = up-regualted and green = down-regulated) in the pathway map [110] .

3.1.3 Analysis of BOS-free Progression by Random Survival Forest Modelling

An integrative random survival forest (RSF) model to predict BOS-free progression was constructed and the performance was rated by the out-of-bag (OOB) concordance index (C-index). The model incorporates all significantly deregulated genes which were members of a significantly enriched KEGG pathway, the significantly differentially expressed genes clustered together with members of significantly enriched KEGG pathways by the functional classification tool from DAVID ("pathway extension genes") and the FEV_1 score as the clinical representation of BOS. This integrative approach (genes and FEV_1) showed a concordance error (1 - C-index) of 0.29 (Figure 3.2). To make sure that the observed effects on BOS-free progression are not due to potential confounding factors, Cox proportional hazard models were calculated. The time after lung transplantation (LTx) showed no significant influence on BOS-free time (Hazard Ratio (HR) 1.00 95%, confidence interval (CI) 0.99-1.00, p-value 0.292). Microbial infection (HR 0.84 95%, CI 0.55-1.27, p-value 0.4) and the total number of acute rejections before sampling (HR 0.88 95%, CI 0.29-2.63, p-value 0.816) showed no significant influence on BOS-free time. The patient-specific attributes age at LTx, transplant type and gender had no significant influence on BOS-free time after LTx (Table 2.1).

3.1.4 Model Analysis of the Random Survival Forest for BOS-free Progression

Random forests (RF) and random survival forests (RSF) are sometimes considered to be black-boxes, which is due to them being an ensemble combination of a large number of decision or survival trees. Luckily several methods are available to make the analysis and interpretation of these ensemble methods feasible. The first such approach presented here is the importance of the variables according to the generated random survival forest model.

Variable Importance

The variable importance (VIMP) of each gene/clinical parameter was assessed by the random survival forest (RSF) and is shown in Table 3.7 and Figure 3.2. According to our RSF model 10 features/genes were potentially informative (VIMP $>$ 0, Table 3.7 and Figure 3.2). Of those, 6 were considered significantly informative, with a VIMP higher than the absolute value of the lowest negative-scoring variable [197]: kallikrein 7 (KLK7), HLA class II histocompatibility antigen, DM alpha chain (HLA.DMA),DR alpha chain (HLA.DRA), cathepsin Z (CTSZ) and major histocompatibility complex, class II, DQ alpha 1 (HLA.DQA1). The correlation of each potentially informative gene or clinical feature with BOS free time was additionally tested by univariate Cox proportional hazard models (Table 3.8).

Marginal Plots

The relationship between a predictor variable and an individuals predicted survival, according to to an random survival forest model, can be visualized using marginal plots. For this the predicted BOS-free progression rate, as predicted for each observation at the median follow up time, was plotted against the features respective values (Figure 3.3).

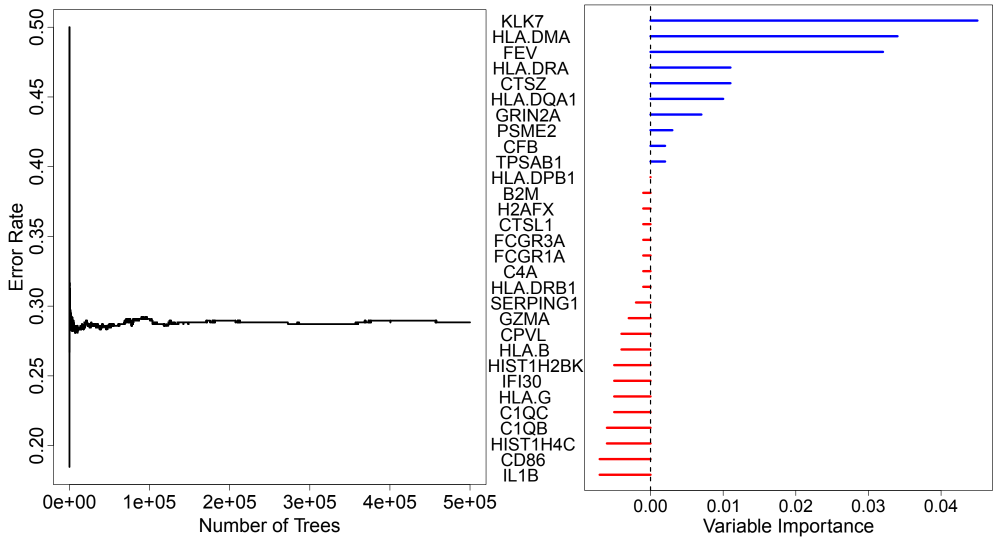


Figure 3.2. The first plot shows the out-of-bag (OOB) concordance error rates of the random survival forest (RSF) for the first b trees. The second plot shows the variable importance (VIMP) for predictors (genes and clinical features). Positive (blue) values represent informative variables, while negative values (red) represent uninformative ones.

Table 3.7. The predictors order according to variable importance (VIMP) as obtained from the random survival forest model

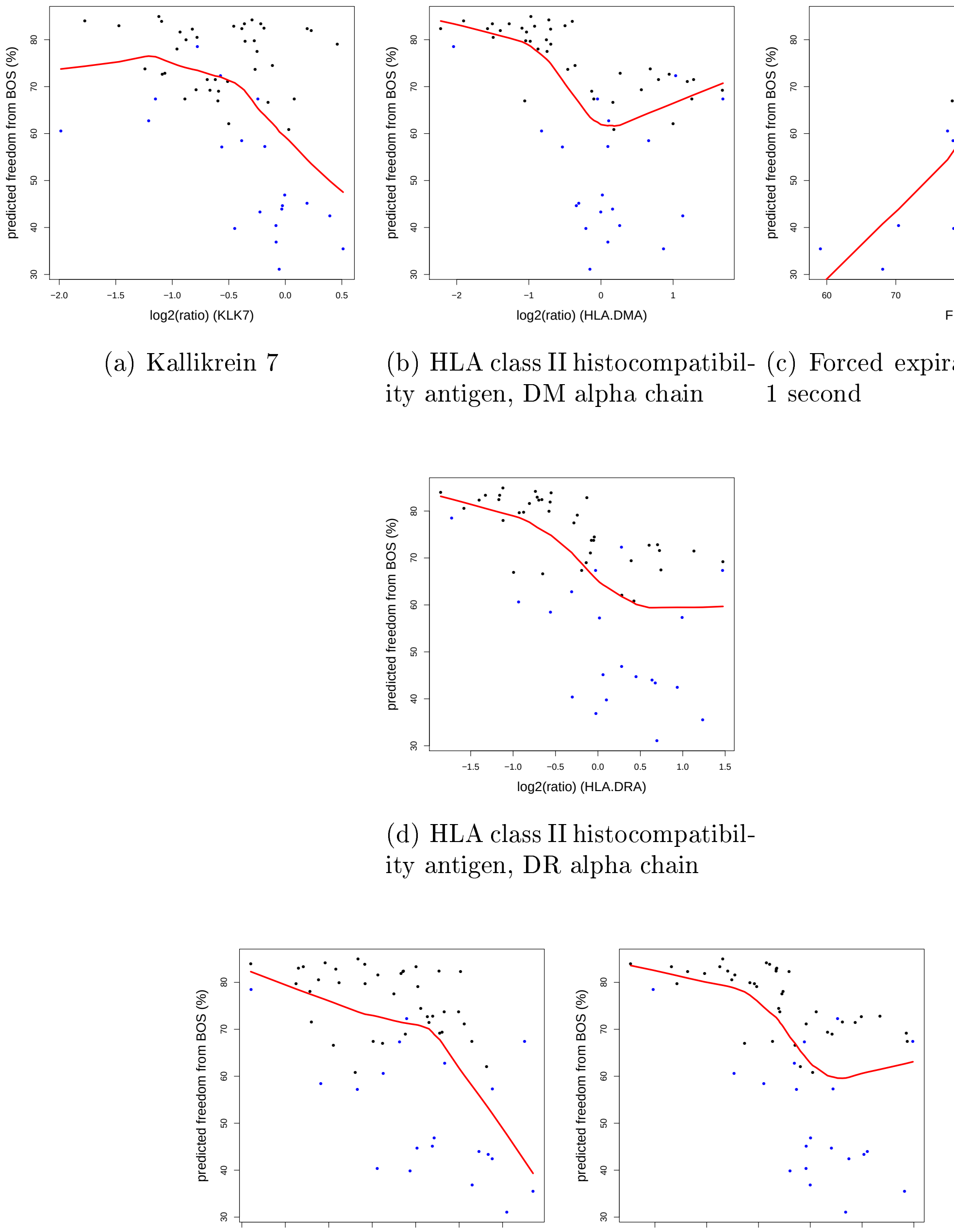
Gene symbol	VIMP	Relative VIMP
KLK7	0.045	1
HLA.DMA	0.034	0.756
FEV_1	0.032	0.711
HLA.DRA	0.011	0.244
CTSZ	0.011	0.244
HLA.DQA1	0.010	0.222
GRIN2A	0.007	0.156
PSME2	0.003	0.067
CFB	0.002	0.044
TPSAB1	0.002	0.044
HLA.DPB1	0.000	0.0000
B2M	-0.001	-0.022
H2AFX	-0.001	-0.022
CTSL1	-0.001	-0.022
FCGR3A	-0.001	-0.022
FCGR1A	-0.001	-0.022
C4A	-0.001	-0.022
HLA.DRB1	-0.001	-0.022
SERPING1	-0.002	-0.044
GZMA	-0.003	-0.067
CPVL	-0.004	-0.089
HLA.B	-0.004	-0.089
HIST1H2BK	-0.005	-0.111
IFI30	-0.005	-0.111
HLA.G	-0.005	-0.111
C1QC	-0.005	-0.111
C1QB	-0.005	-0.111
HIST1H4C	-0.005	-0.111
CD86	-0.007	-0.156
IL1B	-0.007	-0.156

VIMP, variable importance

Table 3.8. Univariate analysis of features selected as informative by random survival forest

FN	**FC**	**P (Wilcoxon)**	**HR (95% CI)**	**P (Cox)**
KLK7 (Kallikrein 7)	$\frac{1}{1.04}$	0.21	2.06 (0.81 - 5.28)	0.21
HLA.DMA (HLA class II histocompatibility antigen, DM alpha chain)	2.27	9.74×10^-12	1.52 (0.99 - 2.33)	0.055
FEV_1 (Clinical Parameter)	$\frac{86.3}{92.8}$	0.159	0.92 (0.88 - 0.96)	4.42×10^{-4}
HLA.DRA (HLA class II histocompatibility antigen, DR alpha chain)	2.24	3.83×10^{-13}	1.9 (1.13 - 3.19)	0.015
CTSZ (Cathepsin Z)	1.63	2.33×10^{-5}	2.39 (1.21 - 4.73)	0.012
HLA.DQA1 (Major histocompatibility complex, class II, DQ alpha 1)	2.82	$< 2.2 \times 10^{-16}$	1.5 (1.05 - 2.14)	0.027
GRIN2A (Glutamate receptor, ionotropic, N-methyl D-aspartate 2A)	1.1	0.4583	2.01 (0.75 - 5.4)	0.166
PSME2 (Proteasome (prosome, macropain) activator subunit 2 (PA28 beta))	1.32	2.23×10^{-6}	1.11 (0.52 - 2.33)	0.793
CFB (Complement factor B)	1.27	0.008	1.34 (0.79 - 2.29)	0.276
TPSAB1 (tryptase alpha/beta 1)	1.29	0.01	1.62 (1.05 - 2.5)	0.03

FN, feature name; P, p-value; FC, median fold change; HR, hazard ratio; CI, confidence interval



(a) Kallikrein 7

(b) HLA class II histocompatibility antigen, DM alpha chain

(c) Forced expiratory volume in 1 second

(d) HLA class II histocompatibility antigen, DR alpha chain

(e) Cathepsin Z

(f) Major histocompatibility complex, class II, DQ alpha 1

Figure 3.3. The marginal plots show the predicted survival, as predicted by the random survival forest (RSF) model, at the median follow up time against the log_2 expression-ratio of the respective gene. Ratio denotes the fold-wise expression change between each sample and the first sample obtained from the respective patient, for both microarray and qRT-PCR results. For FEV_1 the relative percentage of the best FEV_1 observed after lung transplantation (baseline) is given.

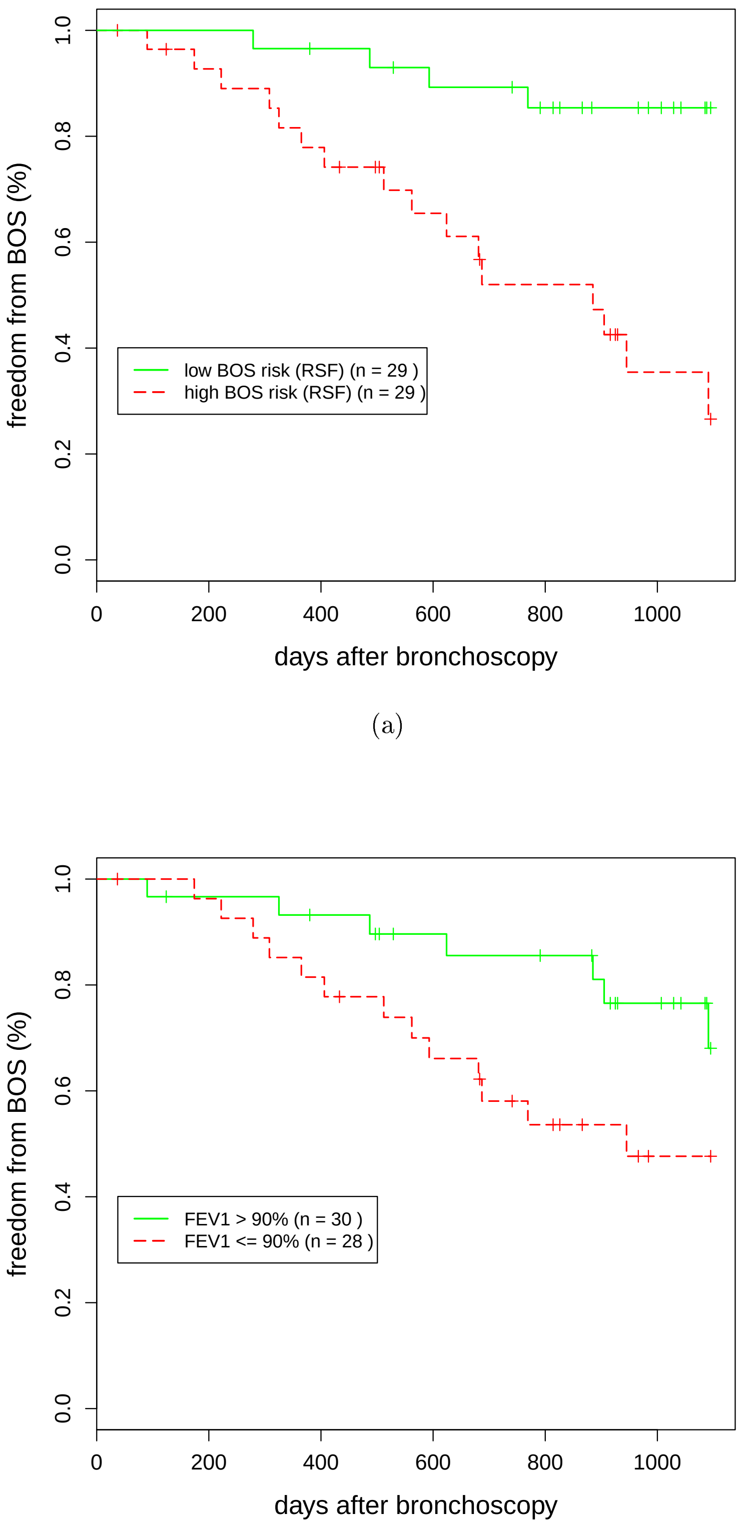
Risk Groups

To get an overview of how the random survival forest (RSF) model partitions the data set, I made use of the proximity information from the RSF model. After transformation of the proximity into a distance ($1 - proximity$), the samples taken before BOS onset (BOS-negative and BOS-unclassified) were clustered into two groups. Both groups showed a significant difference in BOS free time (p-value 2.27×10^{-4}, log-rank test) (Figure 3.4a), and were accordingly labelled as high-risk and low-risk respectively. An FEV_1 score of $\leq 90\%$ (BOS 0-p) was less, but significantly (p-value 0.046, log-rank test), associated with a decrease in BOS free time (Figure 3.4b). Differential expression between both groups (high-risk and low-risk) was assesed for each potentially informative gene or clinical parameter (VIMP > 0), as detected by the RSF model, by means of a Wilcoxon rank sum test (Table 3.8). The expression differences were visualized using a heatmap (Figure 3.5). In the case of cathepsin Z (CTSZ), significant differential expresssion between the high risk and low risk groups (p-value 2.33×10^{-05}, Wilcoxon rank sum test), was confirmed (Table 3.9) on the qRT-PCR level (Spearman rank correlation test, p-value 0.045 and Wilcoxon rank sum test, p-value 1.17×10^{-5}). See Figure 3.6 for the differential expression of cathepsin Z (CTSZ) between random survival forest based clusters, for microarray and qRT-PCR.

Table 3.9. Correlation and concordance between microarray and qRT-PCR data for cathepsin Z.

Gene symbol	Rho	P (Spearman)	FC	P (Wilcoxon)	C-index
CTSZ	0.43	0.045	1.7	1.17×10^{-5}	0.68

Rho, Spearman rank correlation coefficient; P, p-value; C-index, concordance index



(a)

(b)

Figure 3.4. Kaplan-Meier analysis of bronchiolitis obliterans syndrome (BOS)-free time in relation to genomic (random survival forest (RSF)) and diagnostic (FEV_1, BOS 0-p) based risk groups. a) The clustering of the samples according to my random survival forest (RSF) model revealed two groups with a signifcant difference in BOS-free time (p-value 2.27×10^{-4}, log-rank test). b) A forced expiratory volume in 1 sec (FEV_1) less than or equal to 90 percent signature (BOS 0-p) was associated with significantly shorter BOS-free time (p-value 0.045, log-rank test).

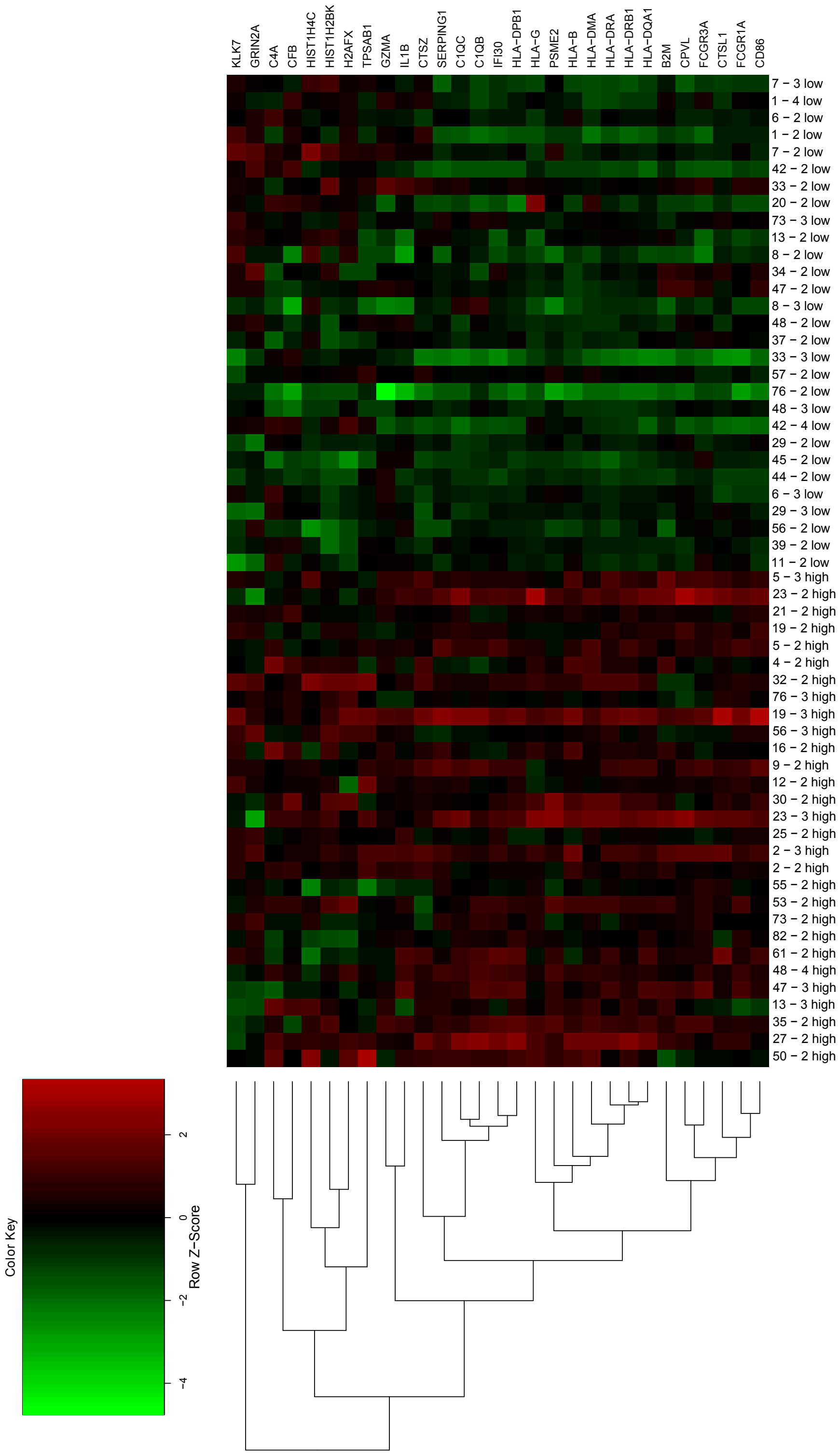
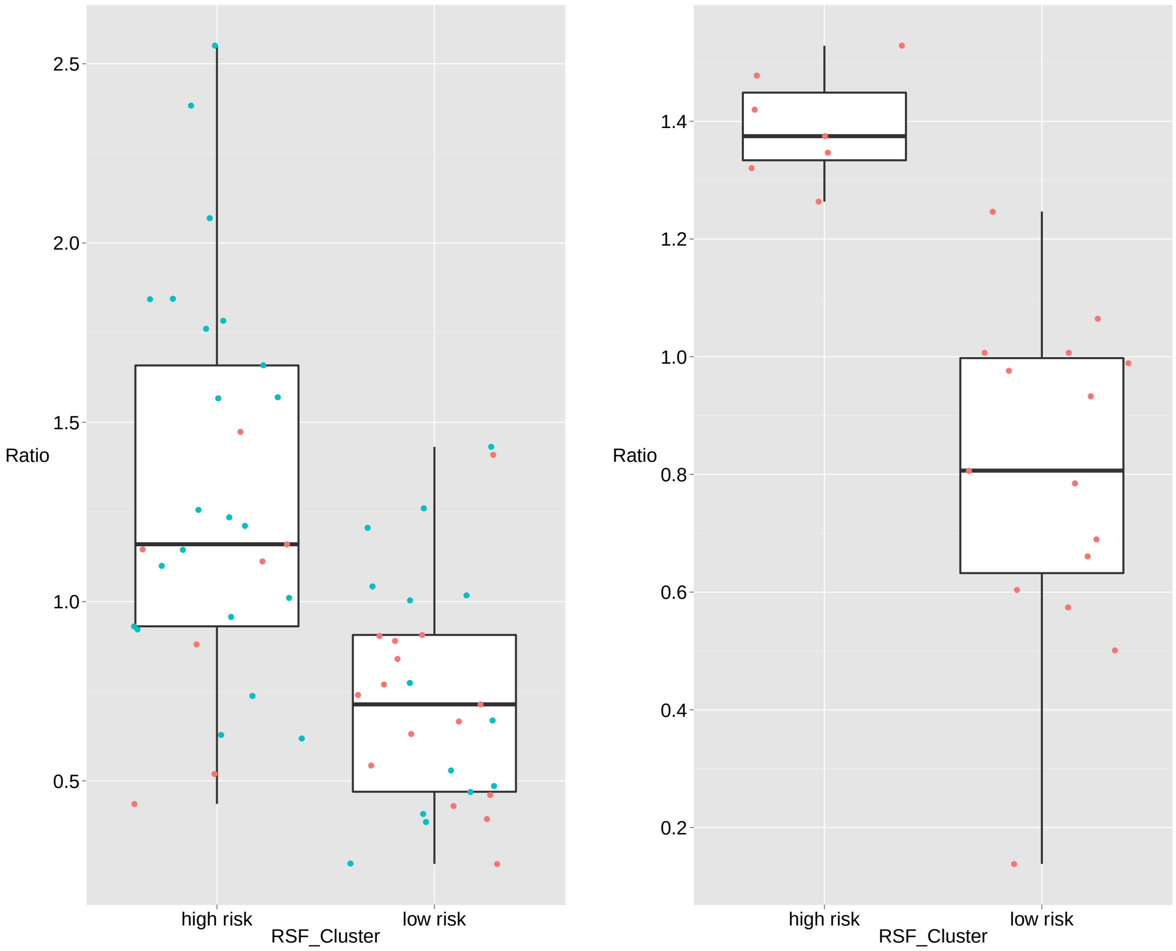


Figure 3.5. Correlation-based heatmap: The heatmap shows the gene expression change between the clusters (high-risk/low-risk) obtained from the random survival forest (RSF) proximity values. The peaks were clustered according to the Pearson correlation coefficient, using the average distance method [236,237].



(a)

(b)

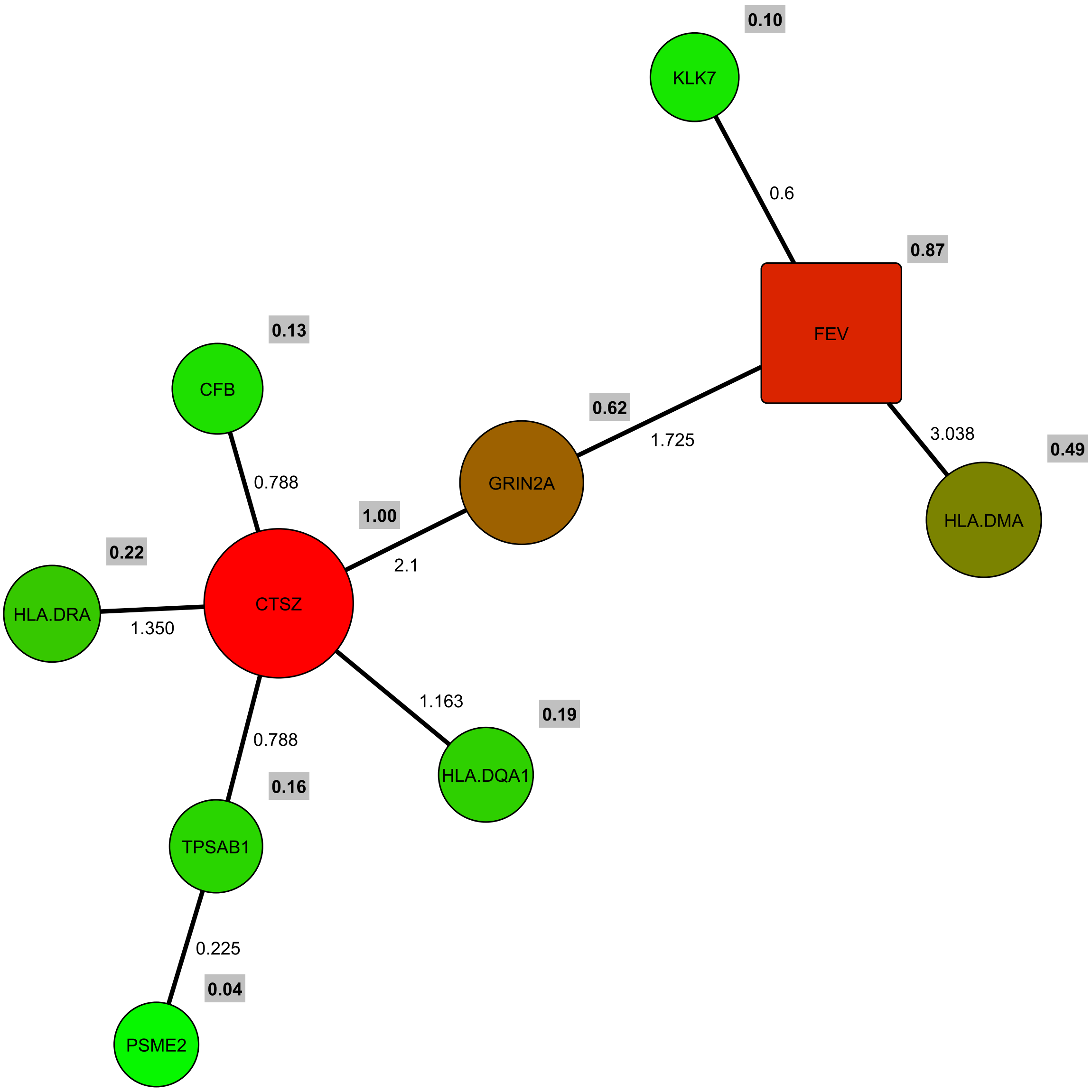
Figure 3.6. Differential expression of cathepsin Z (CTSZ) between random survival forest (RSF) based clusters, for microarrays and qRT-PCR. A significant up-regulation for cathepsin Z (CTSZ) was detected for the cluster, as obtained from the random survival forest (RSF) proximity values, associated with a high BOS risk (p-value 2.27×10^{-4}, log-rank test). This was shown by the microarrays (a) (p-value 2.33×10^{-5}, Wilcoxon rank sum test) and b) the qRT-PCR (p-value 1.17×10^{-5}, Wilcoxon rank sum test) Ratio denotes the fold expression change between each sample and the first sample obtained from the respective patient, for both microarray and qRT-PCR results. a) Expression ratios are from microarray hybridization. Samples where expression has been measured as well by qRT-PCR are shown in red, those where this has not been possible in blue. b) Expression ratios from qRT-PCR, as measured for samples shown in red in a).

Interaction Between Variables

In addition to estimating the variable importance (VIMP) of the features of interest, RSFs can also be used to rate interactions between those features. There are two types of interesting pairwise interactions to be discovered from such an RSF model. Variables which show the same information with regard to BOS prediction (redundancy) and those with complementary interaction. Complementary in the sense that the combined information from both variables leads to an improvement in the prediction of BOS-free progression. In a tree structure complementary means that after splitting on one variable an additional split on a specific complementary variable further down the tree leads to an improvement in prediction performance. The interaction for a pair of variables is analysed using the concept of a joint VIMP [195]. This joint VIMP is defined as the increase in OOB (out of bag) prediction error when both predictor variables are noised up and the prediction error without noising up. Noising up refers to randomly choosing a daughter node when splitting on the variable, instead of using the learned decision rule. The sum of the two single variables VIMPs is called the additive importance. A large absolute difference between the joint VIMP and the additive VIMP (|joint VIMP − additive VIMP|) indicates a potential interaction between two variables [195]. Positive differences (joint Vimp > additive VIMP) reflect the idea that both variables are redundant, while negative values (joint Vimp < additive VIMP) hint to a complementary interaction with respect to BOS progression. To obtain an interpretable interaction score, the differences were normalised by Z score transformation [87] (Table 3.10). Minimum spanning trees, with nodes representing variables and the absolute values of the interactions scores as edge adjacency weights, were calculated for positive and negative interaction scores, separately (Figure 3.7). The first tree graph a) helps to find the variables which best represent the set of predictive variables, while the second graph b) allows for the detection of variables offering complementary information to other predictive variables. The connectivity of each node was assessed by its relative node strength [210] (summed adjacency weight of the connected edges) as compared to the most connected node. Nodes with a relative connectivity > 0.8 were considered hub nodes of their respective graph. For a) the hub nodes were cathepsin Z (CTSZ) and the FEV_1 score (FEV). In tree graph b) kallikrein 7 (KLK 7) was considered a hub node. The protease KLK7 was not a significantly predictive variable when analysed independently of the other selected features. This was shown both for correlation with BOS-free time and for differential expression between the high risk and low risk clusters. In the multivariate RSF model KLK7 was the most informative feature according to the VIMP measure of variable importance. A random survival forest learned without KLK7 being part of the selected feature set showed a concordance error of 33.17, which is a concordance error increase of 4 percent when compared to the feature set including KLK7. The difference of importance for KLK7 between univariate and multivariate assessment of KLK7 importance already hints at showing complementary properties to other variables. This hypothesis was validated by tree graph b). This graph shows KLK7 to be a hub node of complementary information.

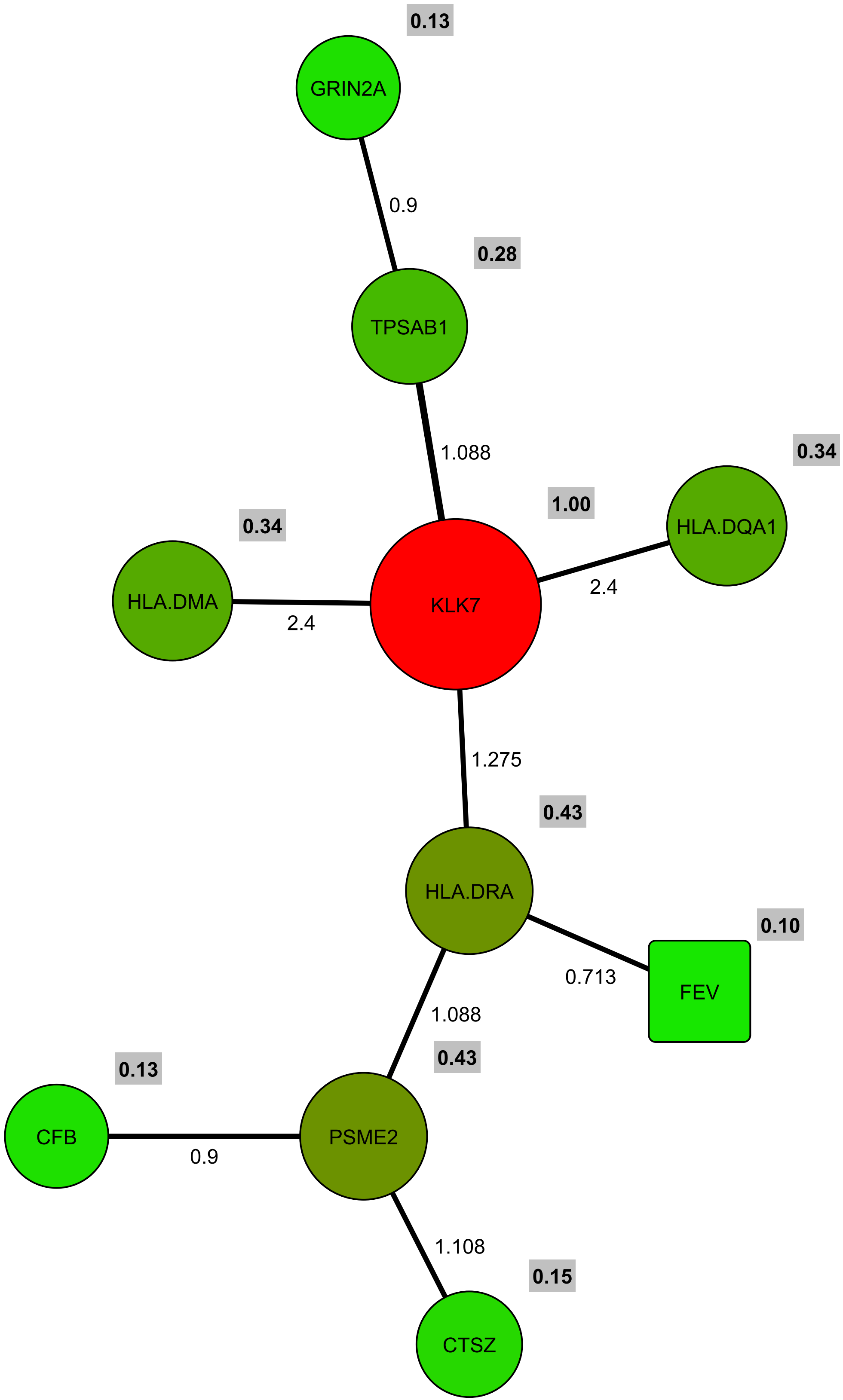
Table 3.10. The pairwise interactions between the important predictors, as detected by the random survival forest model.

Interaction	Paired	Additive	Difference	Interaction Score
HLA.DMA:FEV_1	0.083	0.067	0.016	3.038
CTSZ:GRIN2A	0.030	0.019	0.011	2.100
GRIN2A:FEV_1	0.047	0.038	0.009	1.725
CTSZ:HLA.DRA	0.030	0.022	0.007	1.350
CTSZ:HLA.DQA1	0.026	0.020	0.006	1.163
CTSZ:CFB	0.016	0.012	0.004	0.788
CTSZ:TPSAB1	0.015	0.011	0.004	0.788
CTSZ:HLA.DMA	0.048	0.045	0.004	0.788
CFB:HLA.DQA1	0.014	0.010	0.004	0.788
TPSAB1:HLA.DRA	0.016	0.013	0.003	0.600
KLK7:FEV_1	0.082	0.079	0.003	0.600
GRIN2A:HLA.DRA	0.021	0.019	0.002	0.413
PSME2:HLA.DMA	0.037	0.037	0.001	0.225
CTSZ:FEV_1	0.046	0.045	0.001	0.225
TPSAB1:FEV_1	0.035	0.034	0.001	0.225
TPSAB1:PSME2	0.005	0.004	0.001	0.225
CFB:FEV_1	0.036	0.035	0.001	0.225
HLA.DQA1:FEV_1	0.043	0.042	0.001	0.225
GRIN2A:HLA.DMA	0.042	0.041	0.001	0.225
HLA.DRA:HLA.DMA	0.045	0.045	0.000	0.038
GRIN2A:KLK7	0.053	0.053	0.000	0.038
CFB:HLA.DMA	0.035	0.035	0.000	0.038
HLA.DQA1:HLA.DRA	0.021	0.021	0.000	0.038
CFB:HLA.DRA	0.011	0.013	-0.001	-0.150
PSME2:FEV_1	0.036	0.037	-0.001	-0.150
CFB:KLK7	0.045	0.046	-0.001	-0.150
CFB:GRIN2A	0.007	0.009	-0.001	-0.150
HLA.DQA1:GRIN2A	0.016	0.017	-0.001	-0.150
HLA.DQA1:PSME2	0.012	0.013	-0.001	-0.150
TPSAB1:HLA.DQA1	0.010	0.011	-0.002	-0.338
GRIN2A:PSME2	0.009	0.011	-0.002	-0.338
PSME2:KLK7	0.047	0.049	-0.002	-0.338
CTSZ:KLK7	0.054	0.057	-0.002	-0.338
TPSAB1:CFB	0.001	0.004	-0.003	-0.525
TPSAB1:HLA.DMA	0.032	0.035	-0.003	-0.525
HLA.DRA:FEV_1	0.042	0.046	-0.004	-0.713
HLA.DQA1:HLA.DMA	0.041	0.044	-0.004	-0.713
CFB:PSME2	-0.001	0.004	-0.005	-0.900
TPSAB1:GRIN2A	0.004	0.009	-0.005	-0.900
CTSZ:PSME2	0.009	0.015	-0.006	-1.088
TPSAB1:KLK7	0.041	0.047	-0.006	-1.088
PSME2:HLA.DRA	0.009	0.014	-0.006	-1.088
HLA.DRA:KLK7	0.050	0.057	-0.007	-1.275
HLA.DQA1:KLK7	0.042	0.055	-0.013	-2.400
KLK7:HLA.DMA	0.067	0.080	-0.013	-2.400



a) Minimum Spanning tree of the positive interaction scores

Figure 3.7. Minimum spanning trees of the pair wise interaction scores (positive: redundant and negative: complementary). The edge weights are given by the absolute values of the interaction scores, while the number next to each node shows the relative node weight (summed weight of the connected edges) as compared to the most connected node. This level of connectivity is visualized by node size and node color (from green:lowest to red:highest). Nodes marked red were considered hub nodes of their respective graph.



b) Minimum Spanning tree of the negative interaction scores

Figure 3.7 (continued)

3.2 Prediction of the Bronchiolitis Obliterans Syndrome by Proteome Profiling

The broncholaveolar lavage fluid (BALF) of 82 lung transplanted patients was sampled by bronchoalveolar lavage (BAL) [21]. The proteome constitutents from 146 obtained samples, taken at different time points after lung transplantation, were isolated using superparamagnetic particles (MB-WCX,Bruker Daltonics, Bremen, Germany) and analysed by a linear MALDI-TOF mass spectrometer (MicroflexLT, Bruker Daltonics). The acquired profile spectra were then preprocessed using the PROcess R-package [129], before further computational analysis of BOS pathogenesis was performed.

3.2.1 MALDI-TOF Profiles 1-100 kDa

The peak number in the total mean normalized spectrum was 256 for the 1-10 kDa mass range and 39 for 10-100 kDa mass range with more than 90 percent of all detectable peaks below m/z 20,000. After quality control (Chapter 2.2.1), for each of the mass ranges all spectra of 2 samples (BOS-negative and BOS-unclassified) were excluded from further analysis.

3.2.2 BOS-Specific Proteome Alterations

In the mass range between 1-10 kDa, 23 MALDI-TOF MS peaks (4 peaks, 10-100 kDa) showed significant intensity changes after the first postoperational year as shown in Table 3.11 and 3.13 (10-100 kDa, Table 3.12 and 3.14). Bronchoalveolar lavage fluid (BALF) samples for which an $FEV_1 \leq 80\%$ became persistent within one year after bronchoscopy, were called BOS-positive. An FEV_1 below the International Society for Heart and Lung Transplantation (ISHLT) criteria for BOS ($FEV_1 \leq 80\%$ of baseline) was considered persistent, if all consecutive FEV_1 measurements over a period of at least one year also showed an $FEV_1 \leq 80\%$. Samples were classified as BOS-negative, if they were taken at least three years before an persistent decrease of FEV_1 below 80%. Samples not complying with these strict criteria were labelled as BOS-unclassified. Comparing BOS-positive profiles to negative ones, 24 peaks were found to be differentially regulated in the 1-10 kDa mass range (7 peaks, 10-100 kDa). Those differentially regulated peaks are shown in Table 3.15 and 3.17 (10-100 kDa, Table 3.16 and 3.18). Of all differentially regulated masses (1-10 kDa, Table 3.11, 3.13, 3.15 and 3.17), 13 showed a significant difference in regulation between BOS-positive and BOS-negative at the time of sampling and after the first year post transplantation (Figure 3.9).

Table 3.11. a) Masses with increased intensity after the first year posttransplantation (1-10 kDa)

Peak ranking	Da	RP	FDR	P	Protein identity
	3086.90	49	< 0.01	< 0.001	Serum albumin(ALBU_HUMAN,AA25-51) [a]
7	1309.32 [b]	53	< 0.01	< 0.001	Hemoglobin-beta(HBB_HUMAN,AA33-42) [a]
5	3277.08 [b]	54	< 0.01	< 0.001	Hemoglobin beta(HBB_HUMAN,AA2-32) [a]
	1195.90	58	< 0.01	< 0.001	Hemoglobin-beta(HBB_HUMAN,AA34-42) [a]
	2135.83	59	0.02	< 0.001	Unidentified mass
	3329.27	60	0.02	< 0.001	Unidentified mass
	6798.43	62	0.02	< 0.001	Unidentified mass
	2023.00	63	0.02	< 0.001	Complement C3f fragment(CO3_HUMAN,AA1304-1320) [c]
	2429.70	65	0.03	0.001	Serum albumin (ALBU_HUMAN,AA25-45) [a]

Table 3.12. b) Masses with increased intensity after the first year posttransplantation (10-100 kDa)

Peak ranking	Da	RP	FDR	P	Protein identity
	15154.85	9	< 0.01	< 0.001	Hemoglobin alpha chain [a]

[a] Protein identified in MS/MS experiment.

[b] Peak of random forest biomarker panel.

[c] Exact mass match with the ExPasy TagIdent tool (SIB, Basle, Switzerland).

[d] Exact mass match to previously identified bronchoalveolar lavage fluid component [238]

Da, dalton; RP, rank product; P, p-value; FDR, false discovery rate; Peak ranking (random forest) , classification of samples as BOS+ or BOS- by random forest; BOS, bronchiolitis obliterans syndrome

Table 3.13. a) Masses with decreased intensity after the first year posttransplantation (1-10 kDa)

Peak ranking	Da	RP	FDR	P	Protein identity
1	7921.44 [b]	45	< 0.01	< 0.001	Clara cells 10k Da secretory protein (UTER_HUMAN,AA22-91) [a,c]
	4372.60	46	< 0.01	< 0.001	Unidentified mass
	5593.04	48	< 0.01	< 0.001	Unidentified mass
	1435.54	48	< 0.01	< 0.001	Histatin-3(HIS3_HUMAN,AA33-43) [a]
	1288.20	49	< 0.01	< 0.001	Histatin-3(HIS3_HUMAN,AA34-43) [a,c]
	1336.29	50	< 0.01	< 0.001	Histatin-3(HIS3_HUMAN,AA20-30) [a]
4	3372.41 [b]	55	< 0.01	< 0.001	Neutrophil defensin 2 (DEF2_HUMAN,AA66-94) [c,d]
	5946.94	56	< 0.01	< 0.001	Unidentified mass
	6953.22	58	0,01	< 0.001	Unidentified mass
	7987.27	58	0.01	< 0.001	Unidentified mass
	2666.31	60	0.01	< 0.001	Vimentin(VIME_HUMAN,AA444-466) [a]
	1563.85	60	0.01	< 0.001	Histatin-3(HIS3_HUMAN,AA32-42) [a]
	9755.57	62	0.02	0.001	Unidentified mass
2	3443.58 [b]	63	0.03	0.002	Neutrophil defensin 1 (DEF1_HUMAN,AA65-94) [c,d]

Table 3.14. b) Masses with decreased intensity after the first year posttransplantation (10-100 kDa)

Peak ranking	Da	RP	FDR	P	Protein identity
	13183.97	8	< 0.01	< 0.001	SP-A1 precursor [a]
	12720.94	9	< 0.01	< 0.001	Transthyretin, chain A [a]
	24007.62	9	< 0.01	< 0.001	Unidentified mass

[a] Protein identified in MS/MS experiment.
[b] Peak of random forest biomarker panel.
[c] Exact mass match with the ExPasy TagIdent tool (SIB, Basle, Switzerland).
[d] Exact mass match to previously identified bronchoalveolar lavage fluid component [238]
Da, dalton; RP, rank product; P, p-value; FDR, false discovery rate; Peak ranking (random forest) , classification of samples as BOS+ or BOS- by random forest; BOS, bronchiolitis obliterans syndrome

Table 3.15. Masses with increased intensity in BOS-positive samples (1-10 kDa)

Peak ranking	Da	RP	FDR	P	Protein identity
6	4136.65 [b]	34	< 0.01	< 0.001	Unidentified mass
2	3372.41 [b]	38	0.01	< 0.001	Neutrophil defensin 2 (DEF2_HUMAN,AA66-94) [c,d]
4	3443.58 [b]	42	0.03	< 0.001	Neutrophil defensin 1 (DEF1_HUMAN,AA65-94) [c,d]
3	3202.54 [b]	43	0.03	< 0.001	Unidentified mass

Table 3.16. Masses with increased intensity in BOS-positive samples (10-100 kDa)

Peak ranking	Da	RP	FDR	P	Protein identity
	13183.97	7	0.01	< 0.001	SP-A1 precursor [a]

[a] Protein identified in MS/MS experiment.
[b] Peak of random forest biomarker panel.
[c] Exact mass match with the ExPasy TagIdent tool (SIB, Basle, Switzerland).
[d] Exact mass match to previously identified bronchoalveolar lavage fluid component [238]
Da, dalton; RP, rank product; FDR, false discovery rate; Peak ranking (random forest) , classification of samples as BOS+ or BOS- by random forest; BOS, bronchiolitis obliterans syndrome

Table 3.17. Masses with decreased intensity in BOS-positive samples (1-10 kDa)

Peak ranking	Da	RP	FDR	P	Protein identity
	4372.60	31	< 0.01	< 0.001	Unidentified mass
1	7921.44 [a]	35	< 0.01	< 0.001	Clara cells 10 kDa secretory protein [a] (UTER_HUMAN,AA22-91) [a,c]
	5593.04	35	< 0.01	< 0.001	Unidentified mass
	1435.54	36	< 0.01	< 0.001	Histatin-3(HIS3_HUMAN,AA33-43) [a]
	1288.20	37	< 0.01	< 0.001	Histatin-3(HIS3_HUMAN,AA34-43) [a,c]
	9965.00	46	< 0.01	< 0.001	Unidentified mass
	7612.20	47	< 0.01	< 0.001	Unidentified mass
	2186.62	48	< 0.01	< 0.001	Unidentified mass
	1563.85	48	< 0.01	< 0.001	Histatin-3(HIS3_HUMAN,AA32-42) [a]
5	3277.08 [a]	48	< 0.01	< 0.001	Hemoglobin beta(HBB_HUMAN,AA2-32) [a]
	2470.53	50	< 0.01	< 0.001	Hemoglobin alpha(HBA_HUMAN,AA108-142) [a]
	3476.35	52	0.01	< 0.001	Unidentified mass
7	1309.32 [a]	53	0.01	< 0.001	Hemoglobin-beta(HBB_HUMAN,AA33-42) [a]
	3329.27	53	0.01	0.001	Unidentified mass
	6927.00	55	0.02	0.001	Unidentified mass
	2062.73	55	0.02	0.001	Unidentified mass
	1170.14	59	0.04	0.003	Galanin-like peptide(GALP_HUMAN,AA15-26) [a]
	7987.27	59	0.04	0.003	Unidentified mass
	5946.94	60	0.04	0.003	Unidentified mass
	10021.78	60	0.04	0.003	Unidentified mass

Table 3.18. Masses with decreased intensity in BOS+ samples (10-100 kDa)

Peak ranking	Da	RP	FDR	P	Protein identity
	15154.84	7	< 0.01	< 0.001	Hemoglobin alpha chain [a]
	15860.68	7	< 0.01	< 0.001	Clara cells 10 kDa secretory protein [a,c]
	17474.04	8	< 0.01	< 0.001	Unidentified mass
	14675.63	9	< 0.01	< 0.001	Lysozyme c [a]
	10807.78	9	< 0.01	< 0.001	Calgranulin A [a]
	14299.55	11	0.01	0.001	Unidentified mass

[a] Protein identified in MS/MS experiment.
[b] Peak of random forest biomarker panel.
[c] Exact mass match with the ExPasy TagIdent tool (SIB, Basle, Switzerland).
[d] Exact mass match to previously identified bronchoalveolar lavage fluid component [238]
Da, dalton; RP, rank product; P, p-value; FDR, false discovery rate; Peak ranking (random forest) , classification of samples as BOS+ or BOS- by random forest; BOS, bronchiolitis obliterans syndrome

3.2.3 BOS Predictor Model

From the set of 45 unique peaks showing significant intensity changes (Table 3.11, Table 3.13, Table 3.15 and Table 3.17) between 1-10 kDa, a panel of 7 peaks was selected for the random forest (RF) bronchoalveolar lavage monitoring model (BALmon). These peaks were ranked by variable importance according to their mean decrease in accuracy (MDA) score (Table 3.19). Cross-validation of the predictor ascertained its performance with accuracy: 0.81 (43/53), detection rate (sensitivity): 0.81 (13/16), false positive rate (1 - specificity): 0.19 (7/37) and an area under curve (AUC) of 0.79 (Figure 3.8). Figure 3.10 shows how individual peptide marker intensities change over time for BOS-positive and BOS-negative samples. To make sure that the prediction performance was not influenced by acute rejection or microbial infection, additional statistical tests were performed. The RF score assigned to the BOS-negative and BOS-unclassifed samples during cross-validation did not correlate significantly with the total number of acute rejection episodes until bronchoscopy (rho: 0.07 p-value: 0.459, Spearman correlation test). There was no significant difference in RF score for samples showing acute rejection (p-value: 0.914, Wilcoxon rank test) or microbial infection (p-value: 0.769, Wilcoxon rank test) at the time of bronchoscopy. The FEV_1 value was significantly lower for samples showing acute rejection (p-value: $<$0.001, Wilcoxon rank test) but not microbial infection (p-value: 0.22, Wilcoxon rank test). The FEV_1 correlates significantly with the number of acute rejection episodes until bronchoscopy (rho: -0.28 p-value: 0.002, Spearman rank correlation test).

Table 3.19. Peaks used to classify BOS+ and BOS- samples

Peak ranking	Da	MDA	Protein identity
1	7921.44	0.11	Clara cells 10 kDa secretory protein (UTER_HUMAN,AA22-91) [a,b,c]
2	3443.58	0.10	Neutrophil defensin 1 (DEF1_HUMAN,AA65-94) [a,b]
3	3202.54	0.07	Unidentified mass
4	3372.41	0.03	Neutrophil defensin 2 (DEF2_HUMAN,AA66-94) [a,b]
5	3277.08	0.01	Hemoglobin beta(HBB_HUMAN,AA2-32) [c]
6	4136.65	0.01	Unidentified mass
7	1309.32	0.01	Hemoglobin-beta(HBB_HUMAN,AA33-42) [c]

[a] Exact mass match with the ExPasy TagIdent tool (SIB, Basle, Switzerland).
[b] Exact mass match to previously identified bronchoalveolar lavage fluid component [238]
[c] Protein identified in MS/MS experiment.
Da, dalton; RP, rank product; FDR, false discovery rate; Peak ranking (random forest) , classification of samples as BOS+ or BOS- by random forest; BOS, bronchiolitis obliterans syndrome

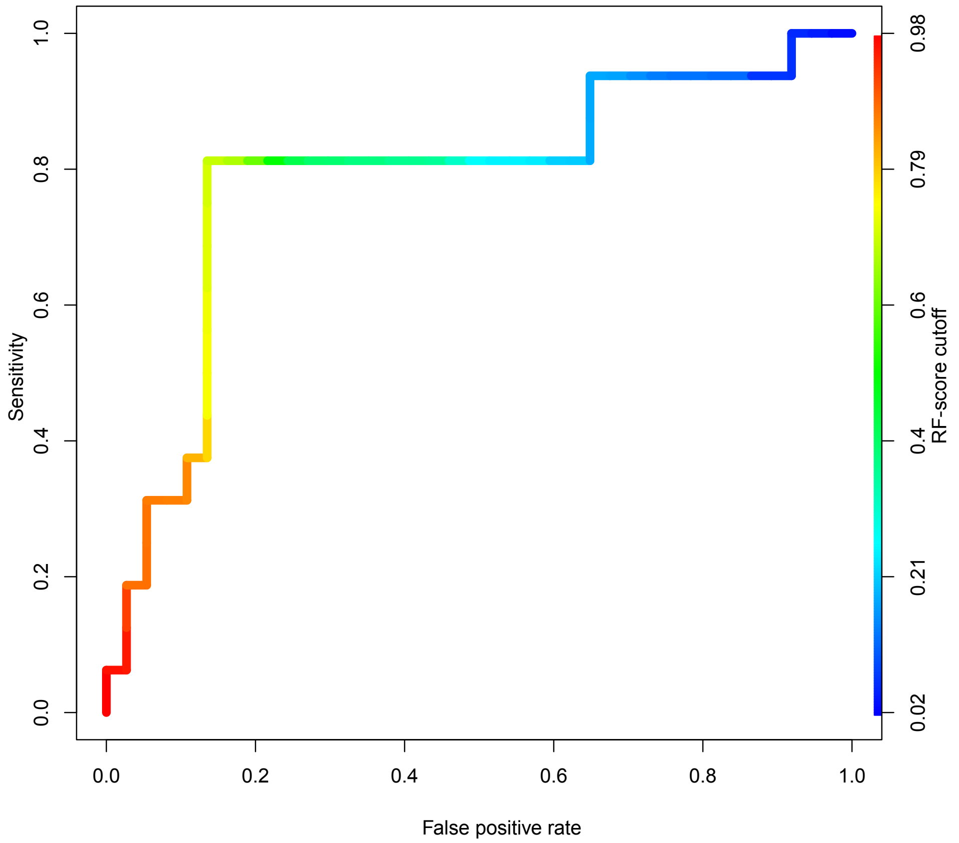


Figure 3.8. Receiver operating characteristics (ROC) curve of bronchoalveolar lavage fluid (BALF) classification into bronchiolitis obliterans syndrome (BOS)-positive and -negative condition by the BALmon biomarker panel. The random forest (RF) predictor, using the BALmon biomarker panel, assigns a score (RF-score) between 0 (0% BOS-negative) and 1 (100% BOS-positive) to all BALF samples. A sample that shows a higher score than a predefined cutoff (RF-score $\geq$ 50%) is classified as BOS-positive. The ROC curve is color-coded according to the level of the cutoff and shows the detection rate (sensitivity) against the false positive rate (1-specifity) for all combinations. An area under curve (AUC) of 0.79 was calculated.

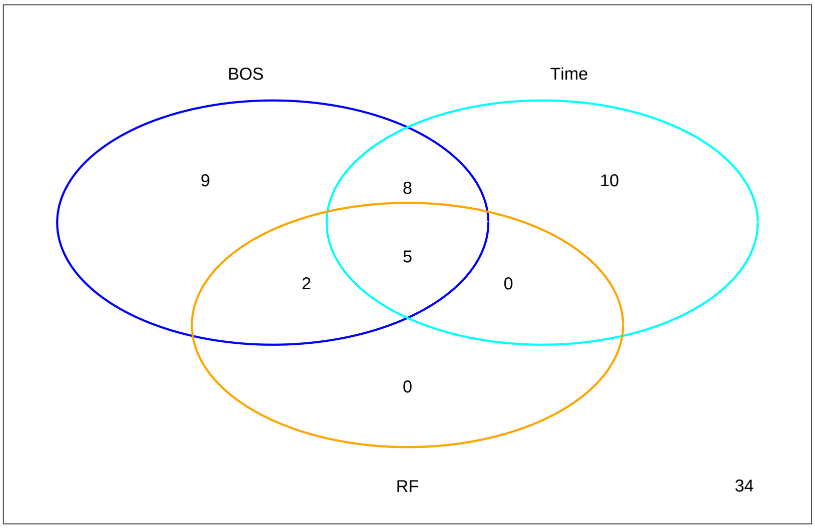
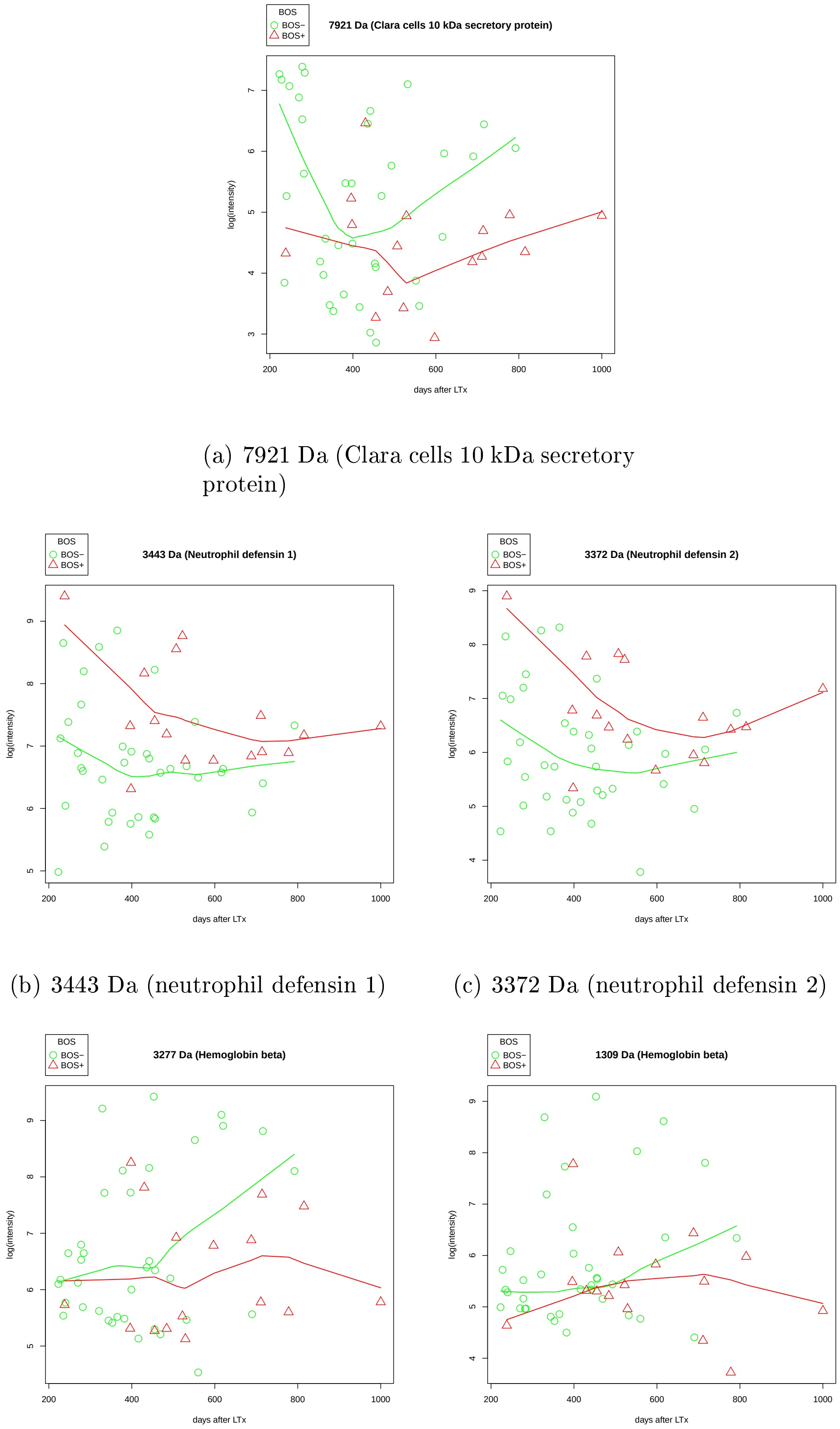


Figure 3.9. The Venn diagramm visualizes the overlap between peaks (1-10 kDa) showing a significant difference in regulation between BOS-positive and BOS-negative at the time of sampling (BOS), after the first year posttransplantation (Time) and the ones selected for the final random forest classifier (RF). [238].



(a) 7921 Da (Clara cells 10 kDa secretory protein)

(b) 3443 Da (neutrophil defensin 1)

(c) 3372 Da (neutrophil defensin 2)

(d) 3277 Da (hemoglobin beta)

(e) 1309 Da (hemoglobin beta)

Figure 3.10. Individual peptide marker intensity patterns over time after lung transplantation.
The scatter plots show how individual peptide marker intensities change over time for BOS-positive (red) and BOS-negative (green) samples.
Solid lines represent regression estimates as obtained by LOESS for local regression [239].

3.2.4 Identification of BALF Proteome Constitutents

Peaks were identified by mass spectrometry, differential signal likely related to differential expression, was obtained from calculation. Thirty-four peaks were identified by direct MALDI-TOF/TOF fragment ion searches, 44 additional mass assignments were based on tryptic in-gel digestion of proteins and 4 additional exact peak masses corresponded to previously identified BALF components. Five out of these 82 polypeptides, could be assigned to predictor masses making up the final BALmon (bronchoalveolar lavage monitoring) random forest model (Table 3.19). The identified and significantly regulated peaks (mass range: 1-10kDa) that belong to the same functional protein group or pathway showed a strong tendency to cluster together, as shown in Figure 3.11.

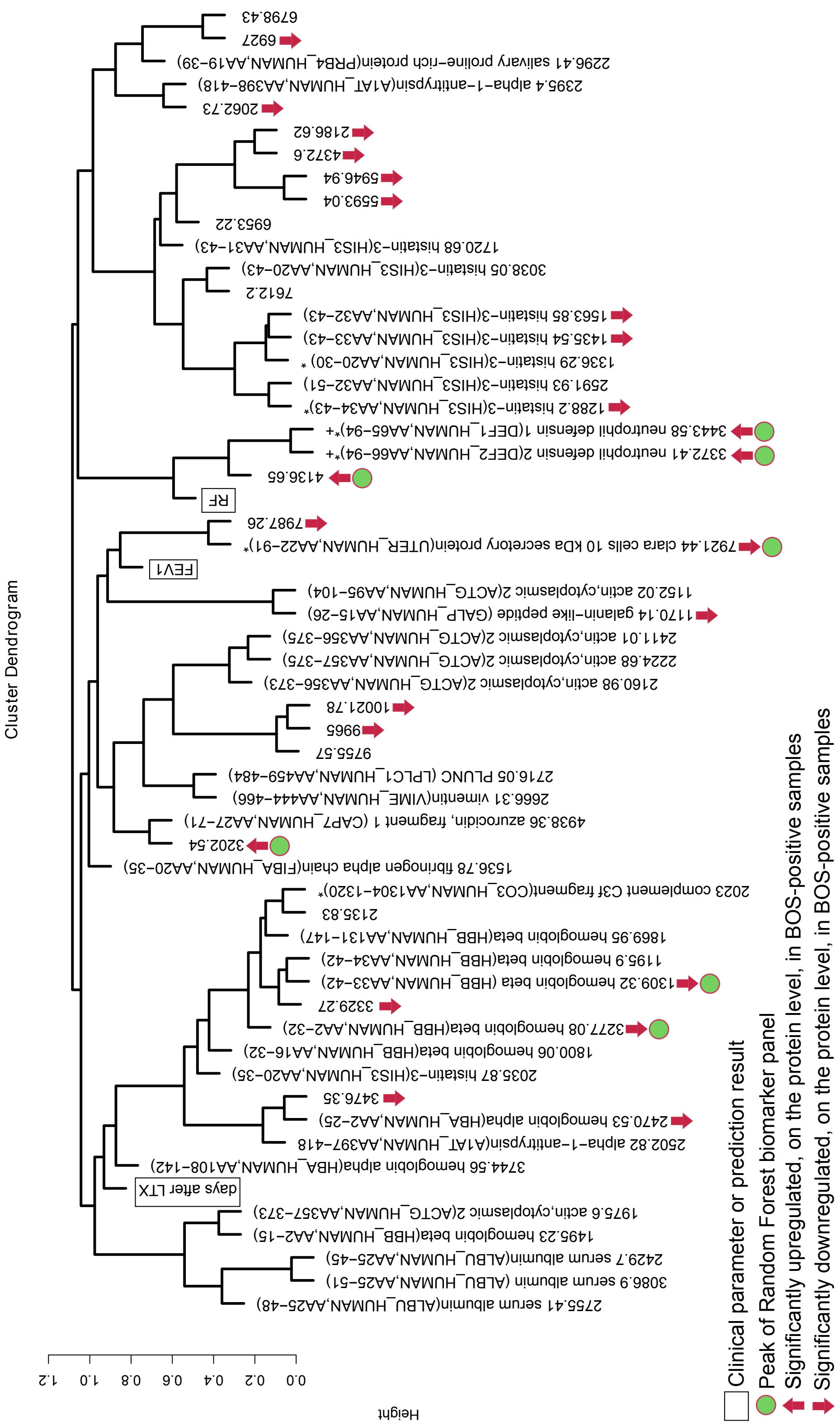


Figure 3.11. Pearson correlation-based average-linkage hierarchical clustering of the BOS biomarker candidates including identified, significant and classifier peaks. The arrows mark significant up- and down-regulation in BOS-positive samples. Peaks not identified by MS/MS identification are marked accordingly: * Exact mass match with the ExPASy TagIdent tool (SIB, Basle, Switzerland).; + Exact mass match to previously identified bronchoalveolar lavage fluid component [238].

3.2.5 BOS Risk Modelling

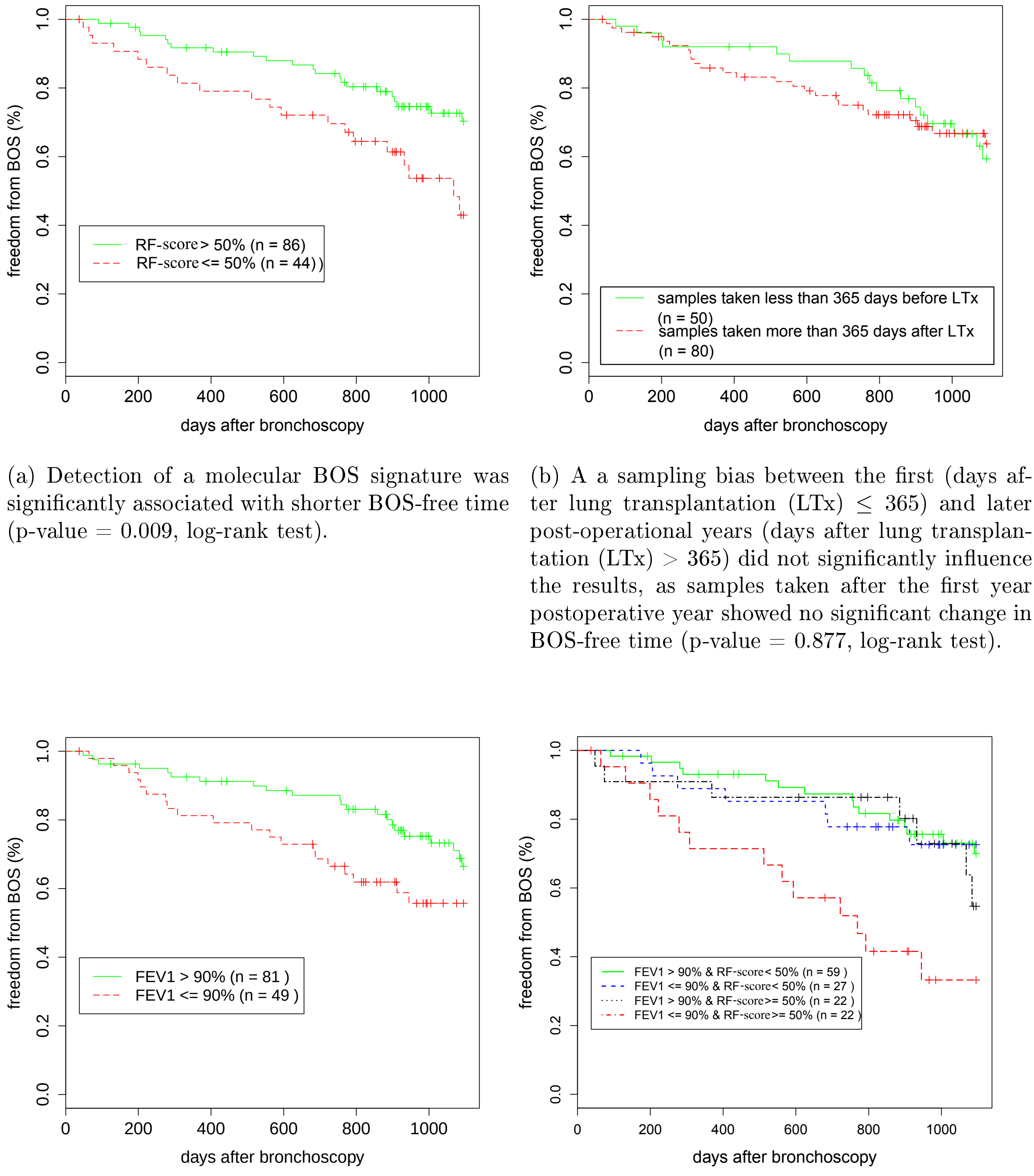
Patient-specific attributes (age at LTx, transplant type, gender and Diagnosis pretransplantation) had no significant influence (p-value $>$ 0.05, univariate Cox proportional hazard model) on BOS-free time after LTx (Table 2.3). A positive classification result (RF $\geq$ 50%) correlated with a significant decrease in BOS-free time (p-value = 0.009, log-rank test) as shown by the Kaplan-Meier plot in Figure 3.12a. This change did not result from a sampling bias between the first and later post-operational years (p-value = 0.877, log-rank test) (Figure 3.12b). An FEV_1 score of $\leq$ 90% (BOS 0-p) was significantly associated with a change in BOS-free time (p-value = 0.046, log-rank test) (Figure 3.12c). Combining both FEV_1 and proteomic signatures (Figure 3.12d), it was possible to stratify patients with BOS 0-p signatures and positive RF-classification results into a group with inferior BOS-free time compared to the BOS 0-p group characterized by negative RF-classification results (p-value = 0.01, log-rank test). The former group shows a short median BOS-free time of 769 days, while for the later group the rate of freedom from BOS never fell below 50 percent. For the univariate Cox models, an RF-score $\geq$ 50% and BOS 0-p showed significant correlation with BOS-free time (Table 3.20). The time after LTx showed no significant influence on BOS-free time (hazard ratio (HR) 1.00,95%-confidence interval (CI) 0.99-1.00, p-value 0.82). Microbial infection (HR 0.82, 95%-CI 0.74-1.38, p-value 0.63), acute rejection (HR 0.91, 95%-CI 0.33-2.56, p-value 0.86) and the total number of acute rejection episodes before sampling (HR 1.01, 95%-CI 0.74-1.38, p-value 0.95) also showed no significant influence on BOS-free time. In the multivariate Cox model using BOS 0-p and an RF-score $\geq$ 50% as factors, only RF-prediction was an independent predictor of BOS progression (RF-score $\geq$ 50%: p -value = 0.02, BOS 0-p: p-value = 0.08).

Table 3.20. Univariate Cox poportional hazard models

	Hazard Ratio	95% Confidence Interval	p-value
BOS-0p	1.84	1.00-3.38	0.05
RF-score $\geq$ 50%	2.19	1.19-4.01	0.01
days after LTx	1	0.99-1.00	0.82

Table 3.21. Multivariate Cox proportional hazard model

	Hazard Ratio	95% Confidence Interval	p-value
BOS-0p	1.71	0.93-3.16	0.08
RF-score $\geq$ 50%	2.07	1.12-3.82	0.02



(a) Detection of a molecular BOS signature was significantly associated with shorter BOS-free time (p-value = 0.009, log-rank test).

(b) A a sampling bias between the first (days after lung transplantation (LTx) $\leq$ 365) and later post-operational years (days after lung transplantation (LTx) $>$ 365) did not significantly influence the results, as samples taken after the first year postoperative year showed no significant change in BOS-free time (p-value = 0.877, log-rank test).

(c) A forced expiratory volume in 1 second (FEV_1) less than or equal to 90 percent signature (BOS-0p) was associated with significantly shorter BOS-free time (p-value = 0.046, log-rank test).

(d) Combining both FEV_1 and biomarker signatures resulted in a more distinct BOS-free time difference than FEV_1 or proteomic criteria alone (p-value = 0.0004, log-rank test).

Figure 3.12. Kaplan-Meier analysis of bronchiolitis obliterans syndrome (BOS)-free time in relation to the absence and presence of predictive factors. The difference between the survival curves (red/green) was assessed by the log-rank test.

3.2.6 ELISA Validation

The MALDI-TOF MS-based results were validated by ELISA for BALF CCP in 13 LTx patients (8 BOS-positive,5 BOS-negative). A strong down-regulation (p-value = 0.009, Wilcoxon rank test) of CCP was associated with BOS progression (Figure 3.13), which is in accordance with the BALmon model. For this BOS progression was defined as an ISHLT BOS stage progression (BOS-0 -> BOS-1 -> BOS2 -> BOS3) in relation to the previous sample.

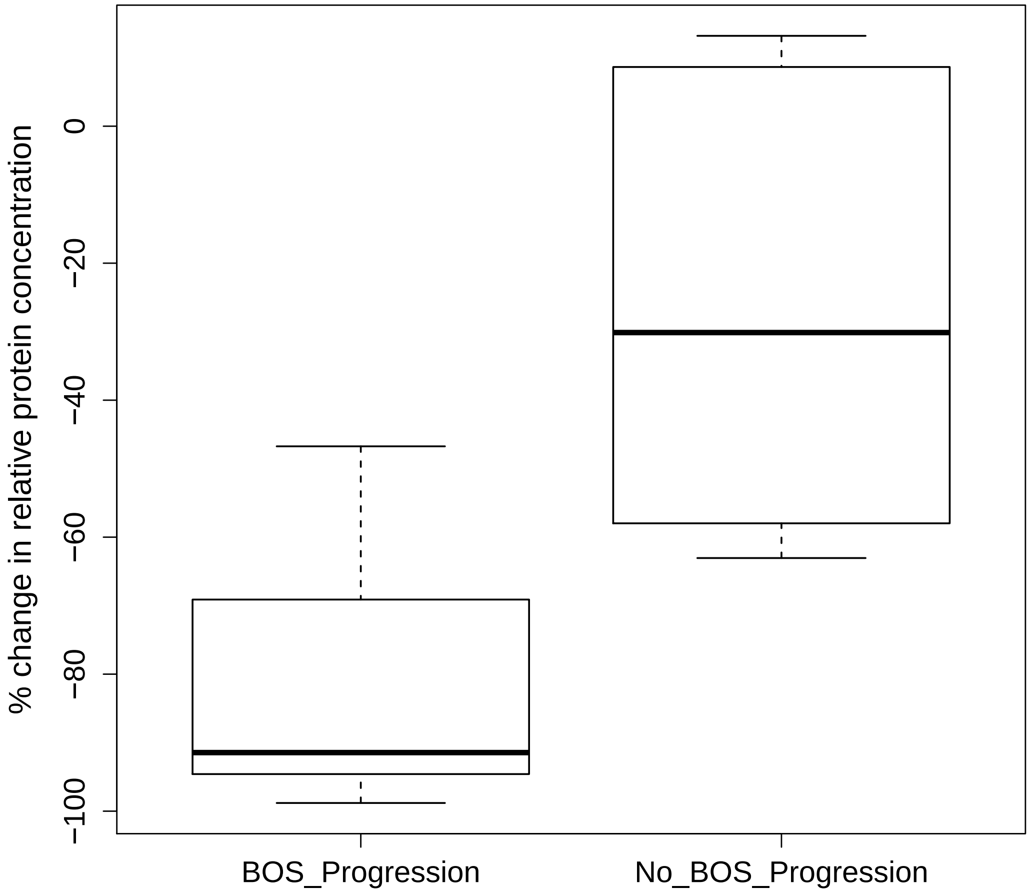


Figure 3.13. ELISA validation: A strong (p-value: 0.009, Wilcoxon rank test) down-regulation of CCP was associated with BOS development which confirms the results from our proteomic data mining.

3.3 Block Maxx

To assess and compare the performance of Block Maxx with several known unsupervised clustering algorithms, a previously published microarray data set was used. This Burkitt lymphoma data set consists of one color Affymetrix gene expression profiles from 220 mature agressive B-cell lymphomas [231].

3.3.1 Gene Selection

To select the most informative genes from the set of genes on the chip, I performed an unsupervised gene selection step. Thus I selected the set of genes which resulted in the dendrogramm, as produced by agglomerative hierarchical clustering, showing the highest concordance correlation (Chapter 2.3.1). This set of 1030 genes, showing the highest variance, was used for all further calculations.

3.3.2 Clustering and Stability Analysis

Using the selected set of genes, clusterings were generated for different unsupervised algorithms (Block Maxx, PAM, PAM-RF (unsupervised random forest clustering) , trimmed-k-means) and numbers of clusters ($k = 2, \ldots, 5$). To evaluate the relative stability of these detected clusters, a bootstrap validation scheme was applied (Table 3.22). The optimal number of clusters for each algorithm (k^*) was chosen as the maximal number of k, for which all (excluding the bakground clusters) clusterwise stabilities were higher or equal to 0.75. The results, according to this criterion, were $k^* = 4$ for Block Maxx and $k^* = 3$ in the case of trimmed-k-means. PAM and PAM-RF (unsupervised random forest clustering) did not meet this threshold for any number of k.

Table 3.22. Clusterwise stabilities

Algorithm	Cluster 1	Cluster 2	Mean	Back
Block Maxx	86.62	87.67	87.15	78.63
Trimmed-k-means	92.28	91.61	91.95	85.04
PAM	74.09	77.51	75.8	
PAM-RF	73.48	79.44	76.46	

Algorithm	Cluster 1	Cluster 2	Cluster 3	Mean	Back
Block Maxx	84.93	83.05	84.32	84.1	77.13
Trimmed-k-means	77.69	78.04	87	80.91	78.68
PAM	76.12	55.22	69.49	66.94	
PAM-RF	78.02	70.99	79.1	76.04	

Algorithm	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Mean	Back
Block Maxx	85.62	82.9	80.27	84.58	83.34	76.4
Trimmed-k-means	71.57	53.34	66.07	75.03	66.5	78.07
PAM	47.29	67.47	73.04	47.99	58.95	
PAM-RF	71.56	65.03	59.7	77.77	68.52	

Algorithm	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Mean	Back
Block Maxx	77.94	77.38	70.56	64.95	84.23	75.01	55.31
Trimmed-k-means	69.12	63.79	43.01	43.78	55.66	55.07	73.73
PAM	48.33	63.47	65.54	47.13	53.56	55.61	
PAM-RF	77.76	68.44	47.48	51.33	74.93	63.99	

Mean, mean clusterwise stability over all clusters; Back, the stability of the background cluster generated by Block Maxx and trimmed-k-means

3.3.3 Survival Analysis

In addition to stability, the ability of the procedure to generate clinically relevant tumour subgroups was used as a performance criterion. Thus I analysed the correlation of clustering patterns, as generated by the compared procedures, with clinical survival data. Significance was assessed using a log rank test and p-values below or equal to 0.05 were considered as significant (Table 3.23). The survival rates were visualized by Kaplan-Meier curves. For each number of k the results obtained from Block Maxx were compared to the reference algorithm with the highest average (mean) clusterwise stability (Figure 3.14). The Block Maxx algorithm showed significant result for $k = 3, \ldots, 5$, trimmed-k-means for $k = 3, \ldots, 5$ and PAM-RF (unsupervised random forest clustering) for $k = 2, \ldots, 5$. In the case of PAM none of the clusterings proved to be significant.

Table 3.23. Clusterwise log rank test

Algorithm	P (2 Clusters)	P (3 Clusters)	P (4 Clusters)	P (5 Clusters)
BlockMAXX	0.098	0.039	0.009	0.038
trimmed-k-means	0.068	0.008	0.016	0.02
PAM	0.42	0.657	0.144	0.164
PAM-RF	0.049	0.007	0.009	0.013

P, p-value (log-rank-test)

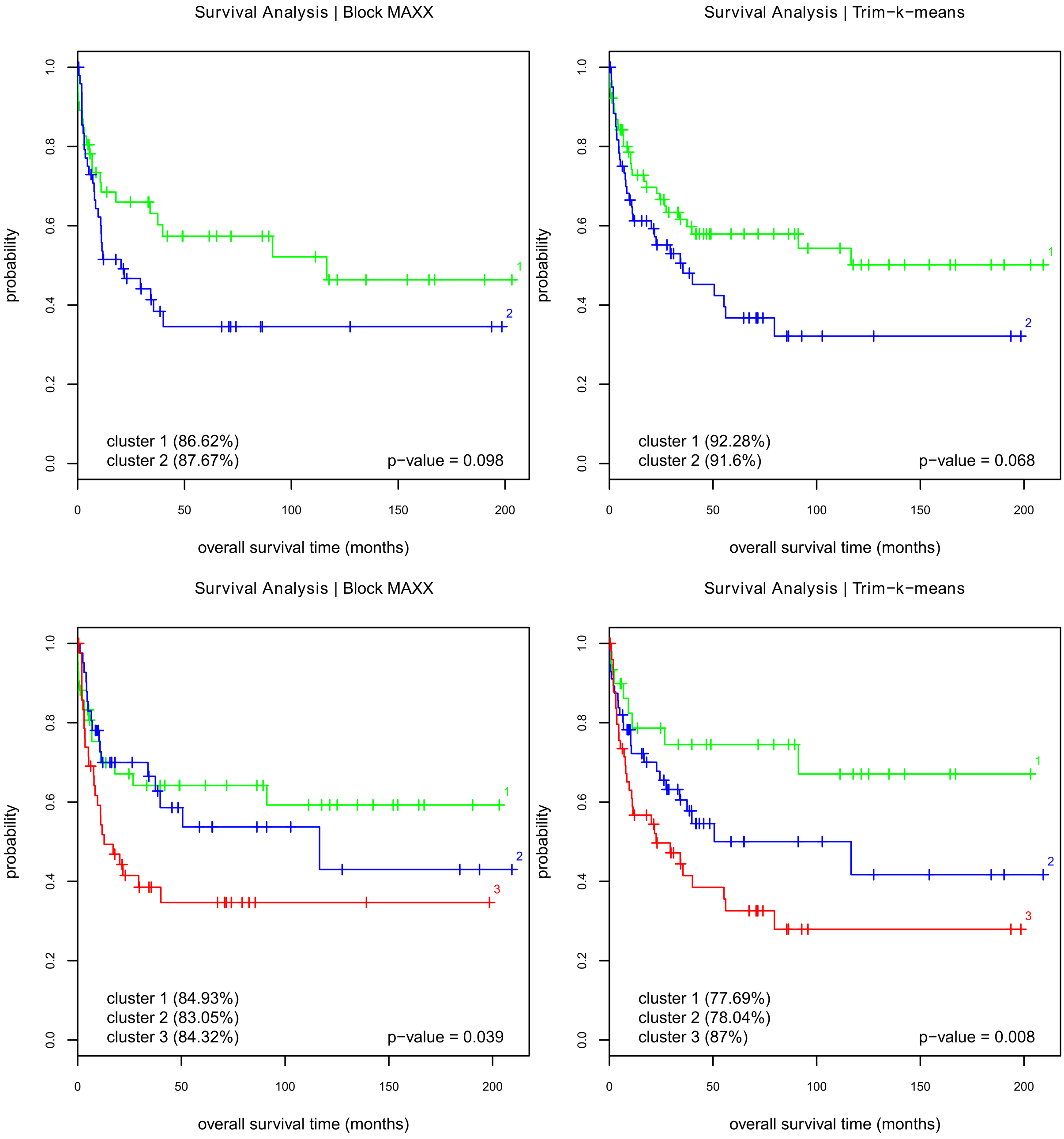


Figure 3.14. Kaplan Meier analysis of survival time in relation to the assignement to a specfifc cluster. The differences between the survival curves was assesed by the log rank test. The p-values for each log rank test are given in each figure and in Table 3.23. (This is a multipage figure.)

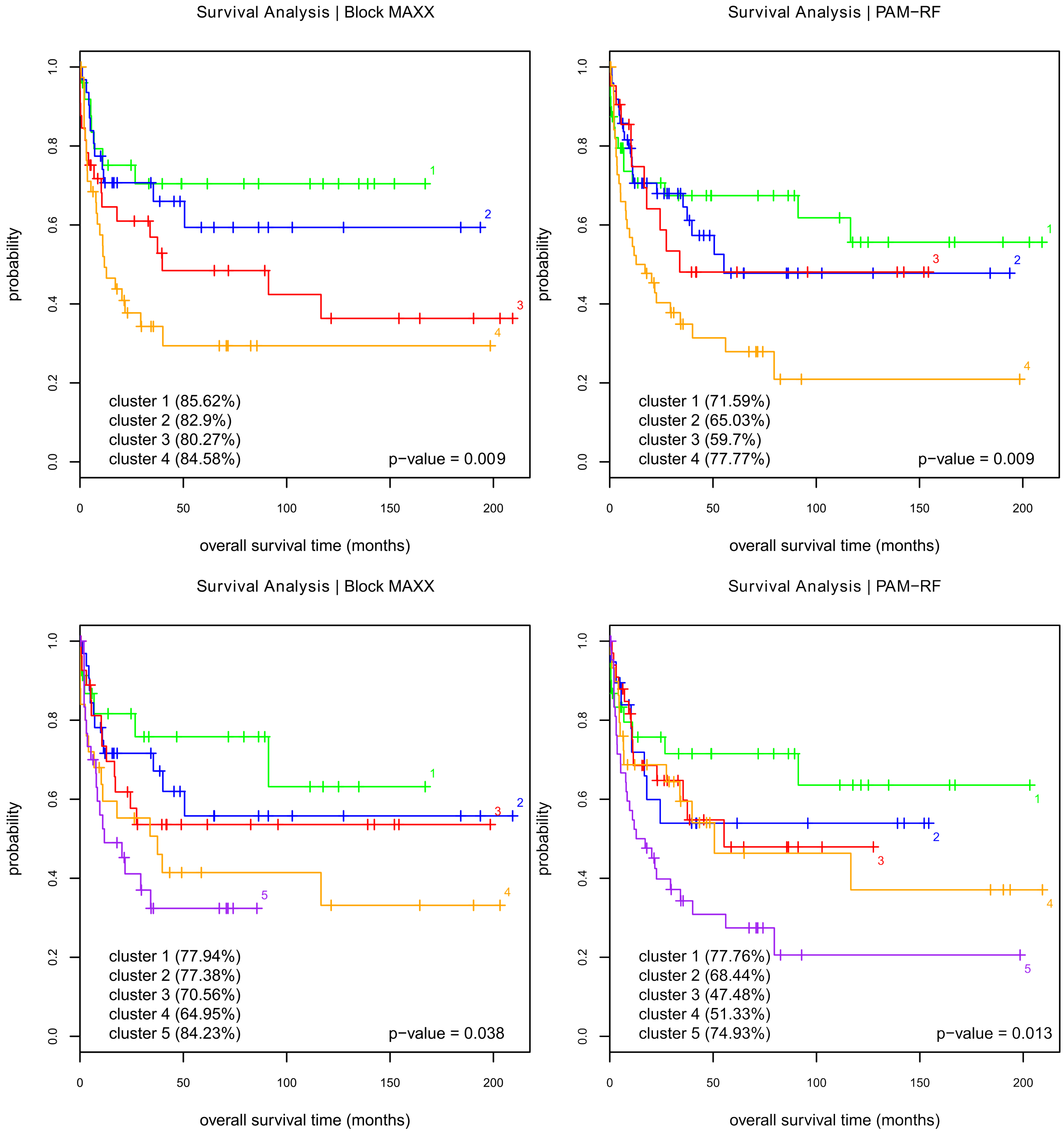


Figure 3.14 (continued)

3.3.4 Visualization

Heatmap

For the number of clusters ($k^* = 4$), detected as optimal for the Block Maxx algorithm, the results were visualized as a heatmap. Since genes can be assigned to up to two clusters, up-regulated in one and down-regulated in another, two respective heatmaps were created (Figure 3.15). In the former genes were ordered according to their up-regulation assignment, while in the later genes were order based on their down-regulation clustering. Both heatmaps use the same sample clustering. The order of the samples is also the same for both heatmaps.

Multidimensional Scaling

The pairwise euclidean distance between all clustered samples was visualized in two dimensional space, using classical multidimensional scaling (MDS) [240,241,242]. The data points were marked by numbers according to their cluster membership (Figure 3.16). For each number of clusters, the results from Block Maxx were compared to the reference algorithm showing the highest average (mean) clusterwise stability.

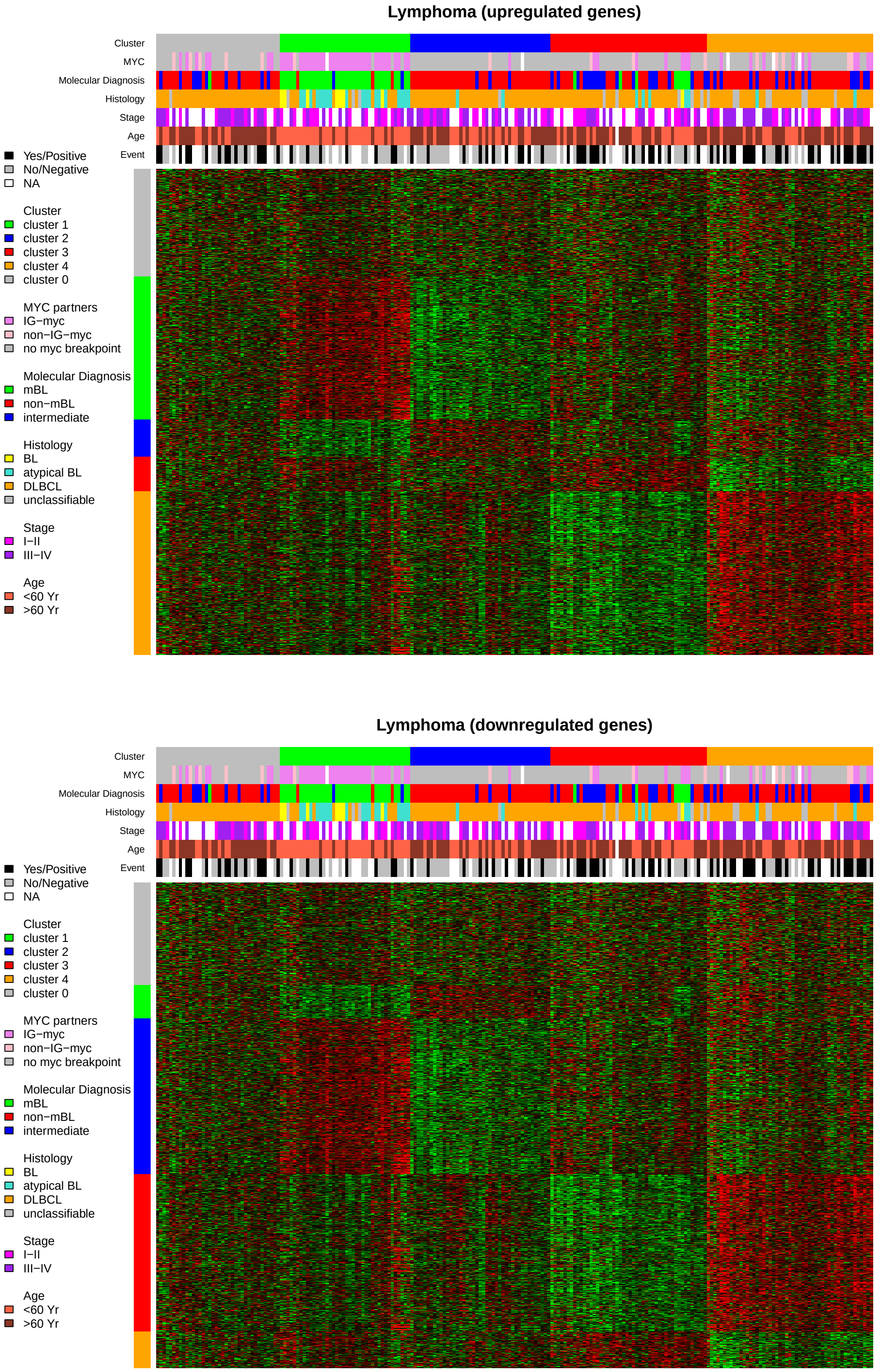
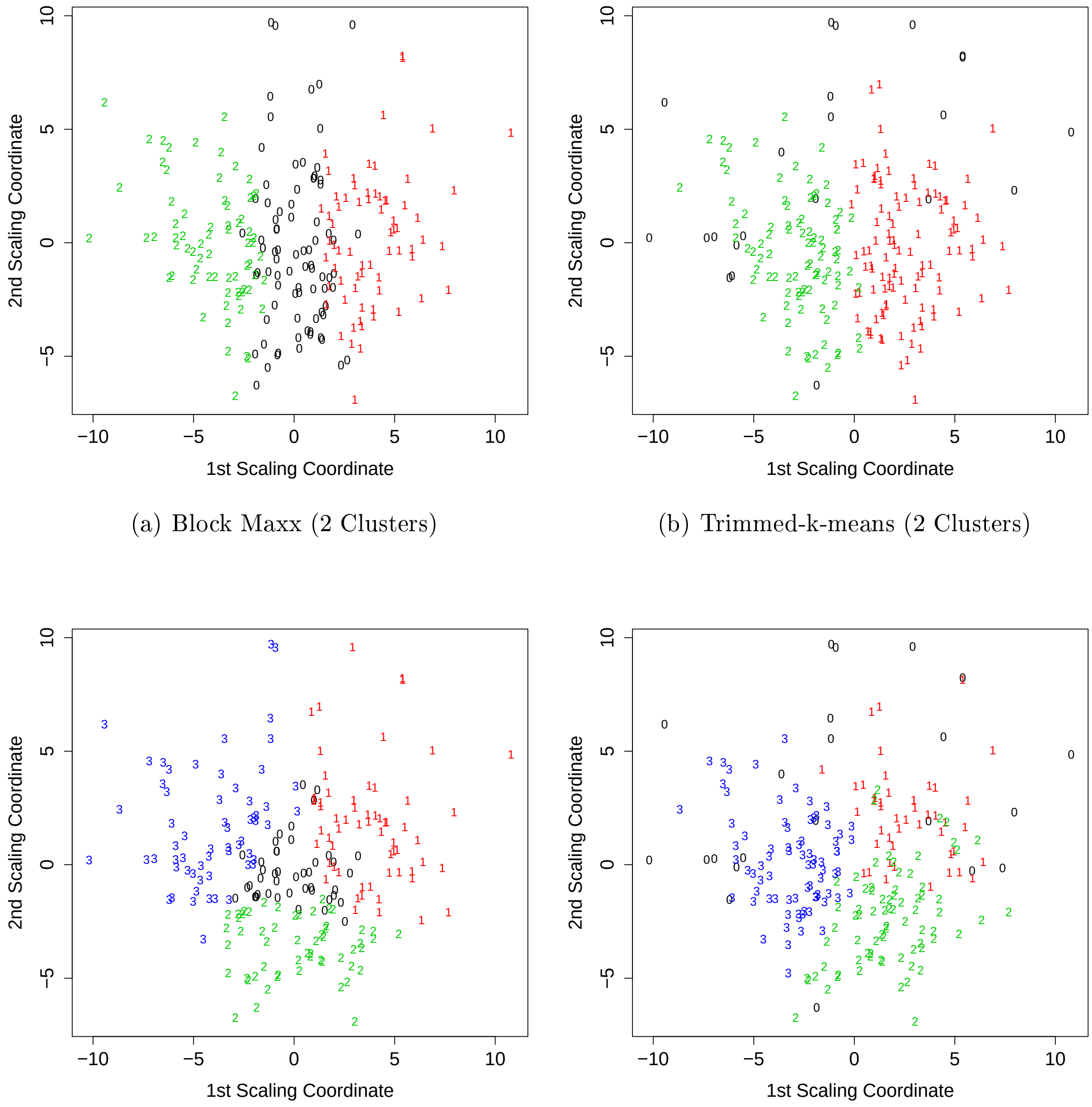


Figure 3.15. The heatmap visualizes the expression of all genes for each cluster. Clusters and samples are ordered according to their cluster membership. Gray denotes the background cluster. Since genes can be members of more than one cluster (up- and down-regulation) two heatmaps were constructed. For these heatmaps red depicts high expression and green shows low expression. (a) Genes were ordered according to up-regulation membership. (b) Genes were clustered according to down-regulation membership.



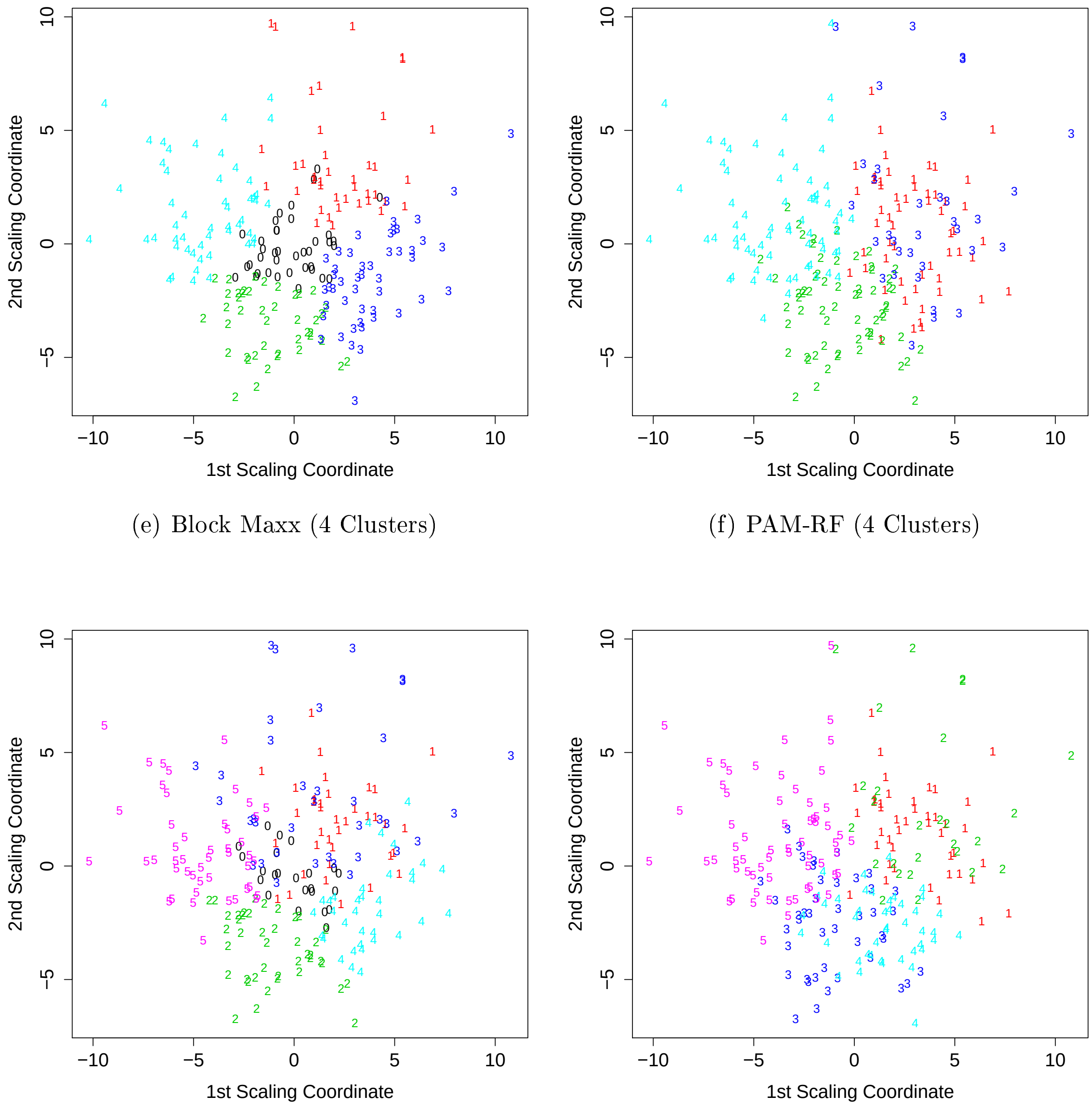
(a) Block Maxx (2 Clusters)

(b) Trimmed-k-means (2 Clusters)

(c) Block Maxx (3 Clusters)

(d) Trimmed-k-means (3 Clusters)

Figure 3.16. The plots visualize the euclidean distances between all samples in a two dimensional space, using multidimensional scaling (0 = background). (This is a multipage figure.)



(e) Block Maxx (4 Clusters)

(f) PAM-RF (4 Clusters)

(g) Block Maxx (5 Clusters)

(h) PAM-RF (5 Clusters)

Figure 3.16 (continued)

3.3.5 Gene Membership

Functional analysis

As with the clustering of the samples, for which clinical relevance was used as a performance criterion, the gene clustering was assessed for functional enrichment of pathways from the KEGG and Biocarta data bases. This was done using the Database for Annotation, Visualization and Integrated Discovery (DAVID) [112]. The significantly functionally enriched pathways for each cluster, up-regulation and down-regulation separately, were reported (Table 3.24).

Table 3.24. Functionally enriched pathways

Cluster 1

Category	Term	Count	FE	P	Regulation
KEGG	hsa05012:Parkinson's disease	21	8.08	2×10^{-11}	up
KEGG	hsa00190:Oxidative phosphorylation	21	7.92	1.48×10^{-11}	up
KEGG	hsa05016:Huntington's disease	22	5.44	4.33×10^{-9}	up
KEGG	hsa03040:Spliceosome	18	6.36	2.23×10^{-8}	up
KEGG	hsa05010:Alzheimer's disease	18	4.86	1.18×10^-6	up
KEGG	hsa03050:Proteasome	8	7.4	8.93×10^-4	up
KEGG	hsa04260:Cardiac muscle contraction	9	5.55	0.002	up
KEGG	hsa03010:Ribosome	9	4.11	0.012	up
BIOCARTA	h_rabPathway:Rab GTPases Mark Targets In The Endocytotic Machinery	4	45.27	0.002	down
BIOCARTA	h_p38mapkPathway:p38 MAPK Signaling Pathway	4	16.46	0.022	down

Cluster 2

Category	Term	Count	FE	P	Regulation
BIOCARTA	h_rabPathway:Rab GTPases Mark Targets In The Endocytotic Machinery	4	41.15	0.003	up
BIOCARTA	h_p38mapkPathway:p38 MAPK Signaling Pathway	4	14.96	0.032	up
KEGG	hsa05012:Parkinson's disease	22	7.88	8.44×10^{-12}	down

KEGG	hsa00190:Oxidative phosphorylation	21	7.37	6.84×10^{-11}	down
KEGG	hsa05016:Huntington's disease	24	5.53	3.95×10^{-10}	down
KEGG	hsa03040:Spliceosome	19	6.25	9.85×10^{-9}	down
KEGG	hsa05010:Alzheimer's disease	18	4.52	3.88×10^{-6}	down
KEGG	hsa03050:Proteasome	8	6.89	0.002	down
KEGG	hsa04260:Cardiac muscle contraction	9	5.17	0.003	down
KEGG	hsa03010:Ribosome	9	3.83	0.02	down

Cluster 3

Category	Term	Count	FE	P	Regulation
KEGG	hsa05012:Parkinson's disease	7	7.86	0.008	up
KEGG	hsa05010:Alzheimer's disease	7	5.51	0.028	up
KEGG	hsa00190:Oxidative phosphorylation	6	6.61	0.025	up

Cluster 4

Category	Term	Count	FE	P	Regulation
KEGG	hsa03010:Ribosome	11	5.48	0.003	up
KEGG	hsa05012:Parkinson's disease	7	7.27	0.014	down
KEGG	hsa05010:Alzheimer's disease	7	5.1	0.045	down
KEGG	hsa00190:Oxidative phosphorylation	6	6.11	0.038	down
KEGG	hsa03040:Spliceosome	6	5.72	0.038	down

FE, fold enrichment ; P, Benjamini-Hochberg False Discovery rate corrected p-value

Stability and Significance

To make sure that the obtained gene membership scores represent differential expression. In addition to stability, a Wilcoxon rank sum test (one-tailed) was used. Expression for all genes were compared between their cluster of membership and all samples not assigned to this cluster. This was done separately for up- and down-regulation. To get an impression how well the gene membership scores represent significantly differential expression, I plotted the gene membership scores against the Benjamini-Hochberg false dicovery rate corrected p-values obtained from the Wilcoxon rank sum test (one-tailed) (Figure 3.17). Genes assigned to the background, when the entire data set was used, passed the 0.75 line in two cases. All of the background genes never passed the 0.95 line. A relation between gene membership and significance was clearly visible in the generated plots. Most of the genes with a gene membership score larger than the 0.75 threshold were significant, while nearly all genes with a gene membership score larger than 0.95 were significant.

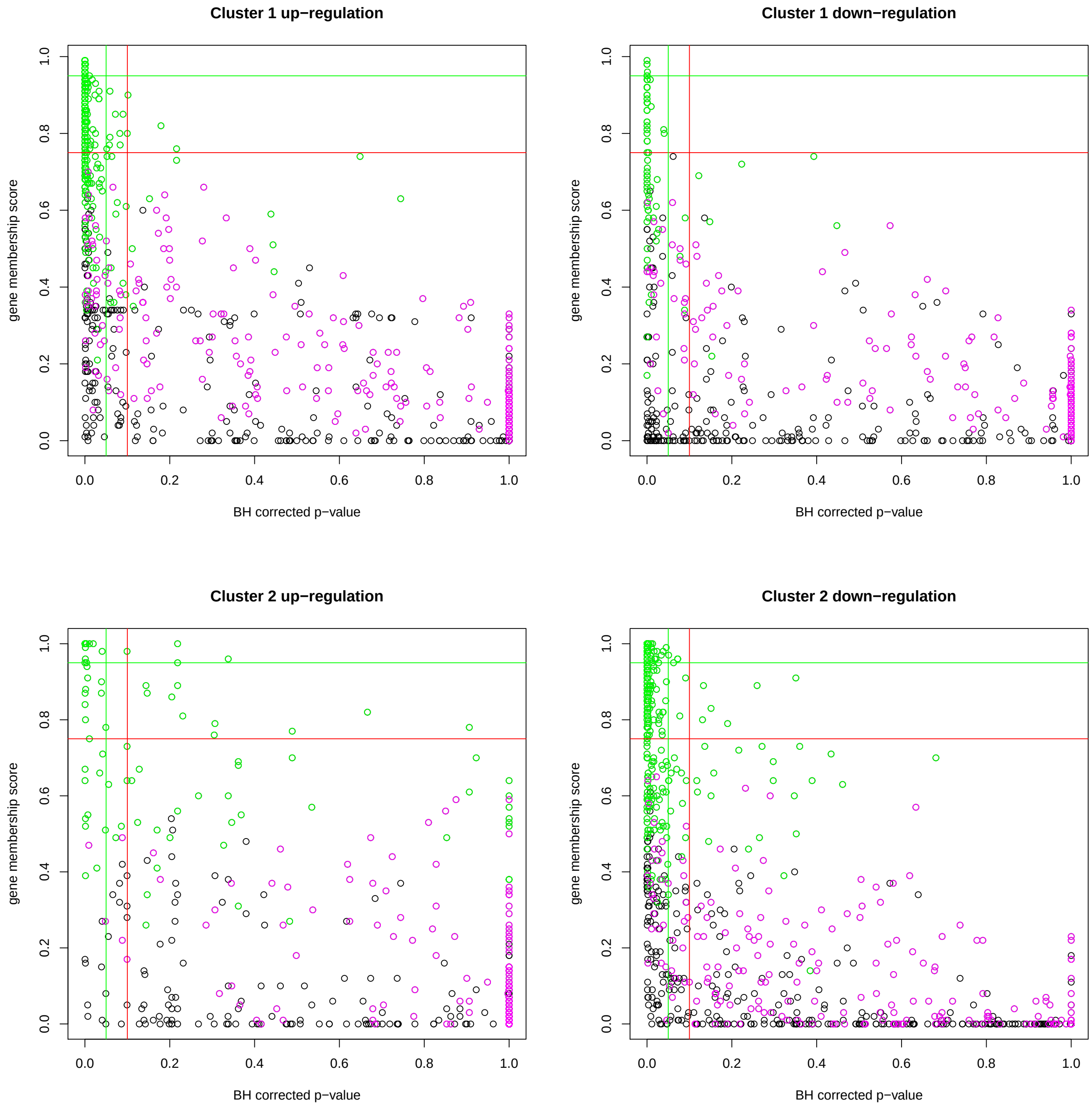


Figure 3.17. The graphs show the relationship between the gene membership score and the significance of the expression difference when compared to other clusters. The p-values were calculated by a Wilcoxon rank sum test (one-tailed) with multiple testing correction according to Benjamini-Hochberg false discovery rate. The green dots represent the genes selected for the respective cluster when the entire data set was used, while red dots represent genes assigned to the background (up- or down-regulation). The black dots are genes that were assigned to another up- or down-regulation cluster. For the p-values the green and red lines represent the 0.05 and 0.1 thresholds respectively. In the case of the gene membership scores a red line signifies a stability of 0.75, while the green line represents a stability of 0.95. (This is a multipage figure.)

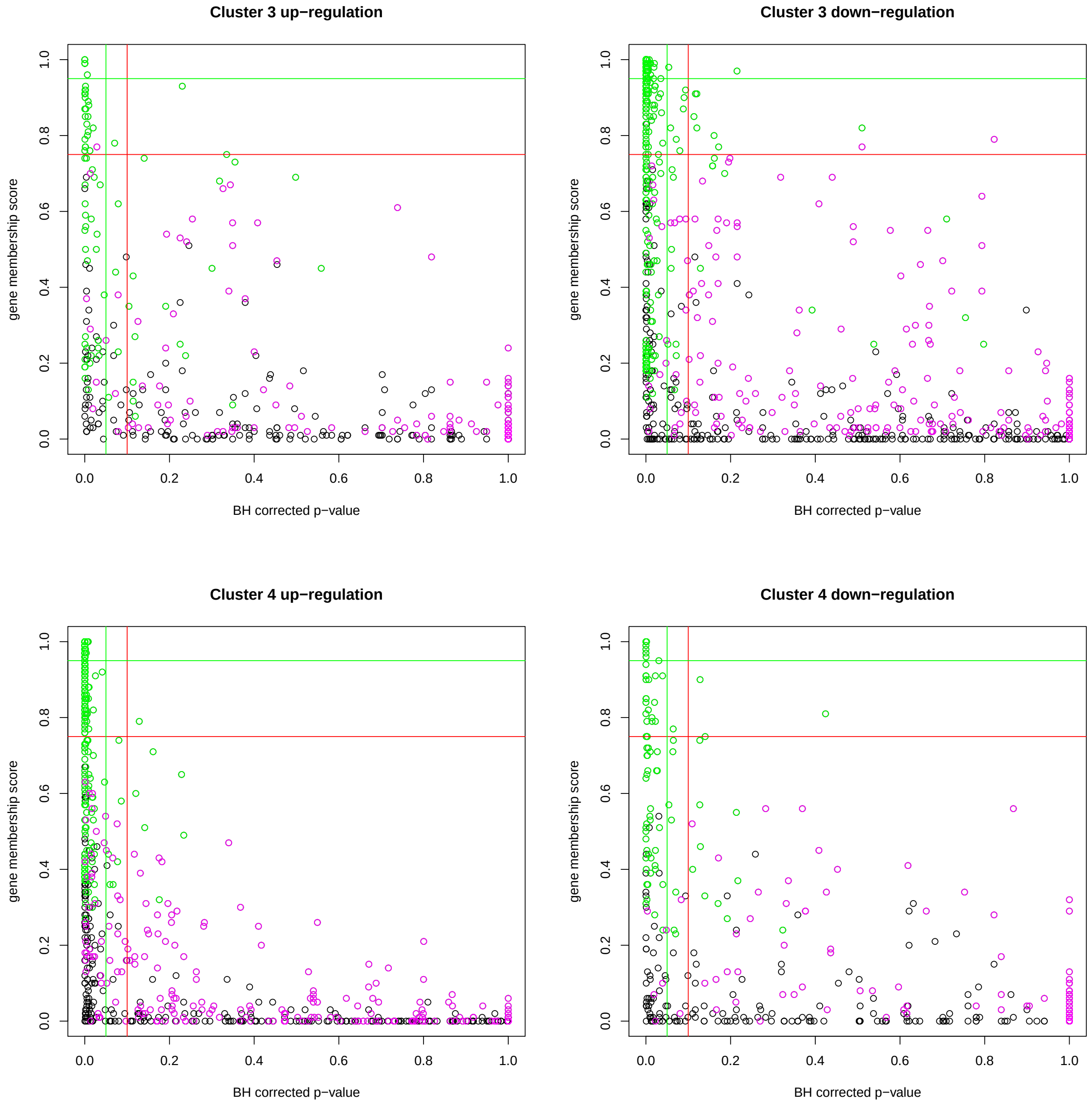


Figure 3.17 (continued)

Chapter 4

Discussion

4.1 Bronchiolitis Obliterans Syndrome

Bronchiolitis obliterans (BO) is a disease for which no reliable prognostic marker has been found yet [4]. It is a clinical manifestation of chronic rejection after lung transplantation (LTx). BO is the major limiting factor for long-term survival after lung transplantation [5,6,7], and manifests as a chronic bronchiolar inflammation accompanied by progressive sub-mucosal fibrosis leading to gradual obliteration of the bronchiolar lumen. The resulting reduction in forced expiratory volume per second (FEV_1) is defined as the bronchiolitis obliterans syndrome (BOS). As chronic lung transplant failure occurs more frequently than in other organ transplants [4], molecular markers for early BO or BOS detection are urgently required to adapt the patients immunosuppressive regimen when airway damage is still minimal. This thesis demonstrates that bronchiolitis obliterans syndrome (BOS) related changes can be discovered for gene expression in bronchial epithelial cells (microarray anaylsis) and from the proteome level in bronchoalveolar lavage fluid (BALF) as determined by mass spectrometric profiling. The detected patterns not only reflect significant differences between BOS-positive and BOS-negative samples, but also show significant correlation with BOS free time. These properties are essential for biomarkers to ultimately supplement established diagnostic procedures [5]. A biomarker panel was obtained by a machine learning approach (random forest), on the basis of a carefully controlled BALF sampling process [243] and mass spectrometric sub-proteome profiles with matrix assisted laser desorption ionization time-of-flight (MALDI-TOF) technology [244,76]. In addition to finding markers that are of putative clinical relevance, I was also interested in a more systems-based approach. This might add to the basic understanding of transplant injury and BOS pathogenesis. This approach was based on gene expression analysis in combination with pathway based methods and model analysis of a random survival forest (RSF) for BOS-free progression. The integrative nature of this thesis lies in the idea that both genomics and proteomics data sets were analysed in combination with clinical data. It has often been reported that mRNA levels do not necessarily correlate with protein abundance [245,246,247199]. As the protein profiles were generated from BALF and gene expression was measured directly from bronchial epithelial cells, a further level of complexity was added to the integration of both data types. This is due to the fact that BALF consisting of a mixture of bronchial epithelial cells and immune cells such as macrophages, neutrophils and lymphocytes, while bronchial brush specimens were shown to contain a relatively pure cell population of more than 90 percent bronchial epithelial cells [22]. This was reported for lung transplants showing no BOS. As infiltration by immune cells is a process that occurs in BOS, a higher percentage of such cells is to be expected.

4.1.1 Clinical Application

Training a random forest classifier on the MALDI-TOF mass spectra of BOS-positive and BOS-negative samples, a panel of biomarker peptides was selected. As shown by cross-validation, this classifier panel was reliably able to differentiate between BOS-positive and BOS-negative samples. Cross-validation of the predictor ascertained its performance with accuracy: 0.81 (43/53), detection rate (sensitivity): 0.81 (13/16), false positive rate (1 - specificity): 0.19 (7/37) and an area under curve (AUC) of 0.79. The selected biomarker panel (BALmon) included hemoglobin (hemoglobin beta (HBB)), secretoglobin (Clara Cell secretory protein (CCP)) and defensin (human neutrophil peptide 1 (HNP1) and human neutrophil peptide 2 (HNP2)) members. Thus the biomarker panel only included proteins and protein families for which the results were shown to be consistent between studies and sample types. This shows the reliability of my newly developed feature selection approach combining rank product and random forest based methodologies. It also supports the idea that, to get reliable results, statistically significant findings should further be judged by their predictive performance [147]. As has already been shown for the included markers CCP, HNP1 and HNP2 [248,238], the classification results were not influenced by acute rejection or microbial infection. The random forest predictor score (RF-score) assigned to the BOS-negative and BOS-unclassified samples during cross-validation did not correlate significantly with the total number of acute rejection episodes until bronchoscopy (rho: 0.07 p-value: 0.459, Spearman correlation test). There was no significant difference in RF-score for samples showing acute rejection (p-value: 0.914, Wilcoxon rank test) or microbial infection (p-value: 0.769, Wilcoxon rank test) at the time of bronchoscopy This is a distinct advantage when compared to FEV_1 based diagnostics. The FEV_1 value was significantly lower for samples showing acute rejection (p-value: $<$0.001, Wilcoxon rank test) but not microbial infection (p-value: 0.22, Wilcoxon rank test). The FEV_1 correlates significantly with the number of acute rejection episodes until bronchoscopy (rho: -0.28 p-value: 0.002, Spearman rank correlation test). The assumption that BOS (bronchiolitis obliterans syndrome) can be predicted before its irreversible clinical symptoms manifest is evident in the well established FEV_1 based BOS 0-p [72,73] criterion, and on the proteome level by the random forest model incorporating the biomarker panel BALmon. For the patient cohort of this study, the BALmon-based classification and the BOS 0-p criterion both showed significant differences in BOS-free time. The combination of both devices seems to suggest an improved accuracy, characterized by a striking decrease in median BOS-free time after positive classification by both criteria. The biomarker panel further showed the ability to stratify the BOS 0-p group into two subgroups differing significantly in BOS-free time (p-value = 0.01, log-rank test). Samples classified as BOS risk positive by only one of both criteria, did not show a significant decrease in BOS-free time. Many studies aim to replace well-known and established clinical tools with emerging biomarker discovery techniques. This study thus implies that it is favorable to follow an integrative approach by combining predictors from clinical parameters with those based on molecular biomarkers, such as BALmon.

4.1.2 Systems-Based Approach

In addition to the search for genes and proteins differentially expressed in BOS-positive samples, a more systems-based approach was performed for the transcriptome data. For this several pathways were identified, which were significantly overrepresented (enriched) among the list of differentially expressed genes. Several of these had already been linked to the bronchiolitis obliterans syndrome, like "systemic lupus erythematosus" [249,250,251,45], "graft vs. host disease" [252] or "allograft rejection" [253]. By annotation based clustering several additional differentially expressed genes could be functionally linked to the enriched pathways ("pathway extension"). Thus further analysis using a random survival forest (RSF) combining FEV_1 diagnostics with known pathway members and these extension genes. The model showed good performance of BOS risk prediction and was able to identify two risk groups defined by distinct molecular patterns. It also allowed for the identification of genes that are predictive of BOS development.

4.1.3 Potential Biomarkers for the Bronchiolitis Obliterans Syndrome

In the following, I will describe several biomarkers that have been found to be associated with BOS, either from gene expression profiling and/or from mass-spectrometric analysis. A summary of the discussed results is shown in Figure 4.1. It should be noted that this analysis only yields mass over-charge (m/z) ratios for these biomarkers, and they need to be identified by complex biochemical studies (Chapter 2.2.5).

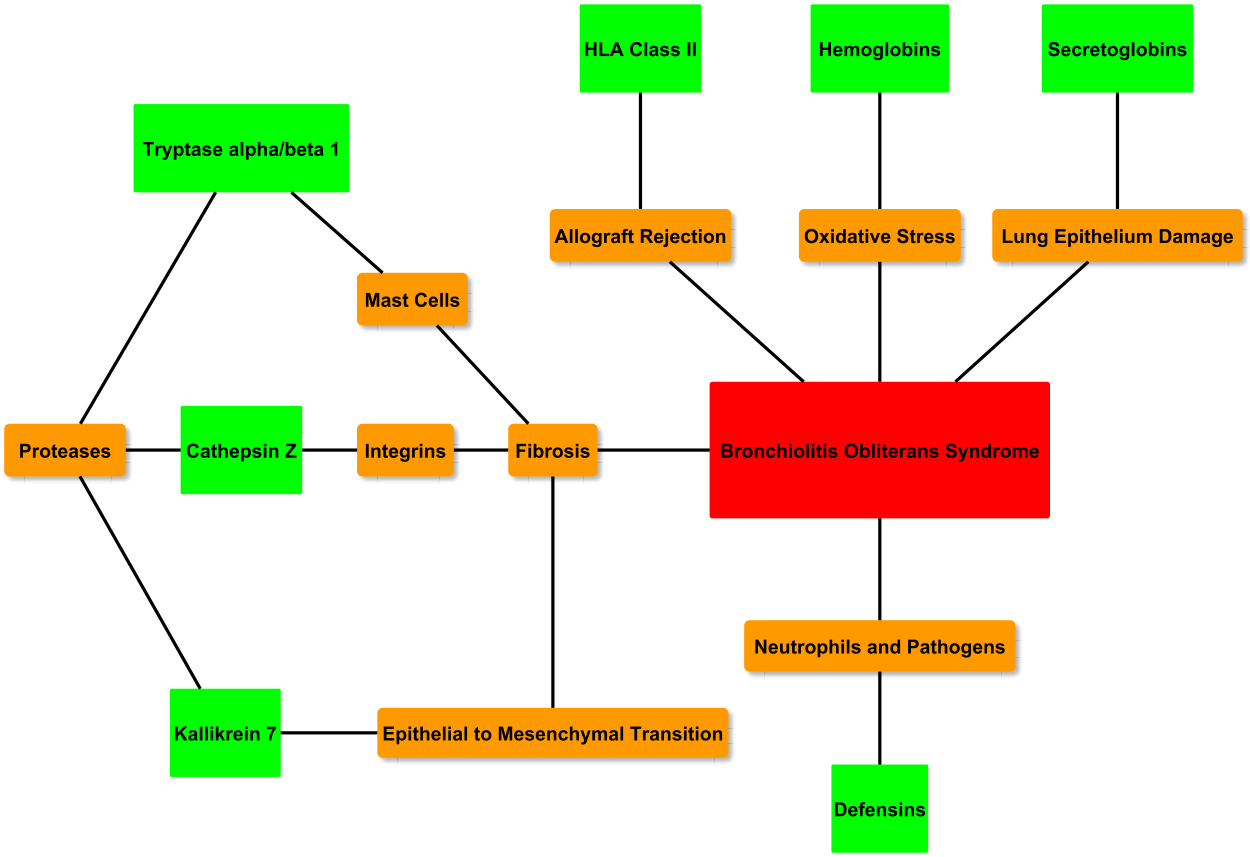


Figure 4.1. The green nodes represent genes and proteins that were shown by random forest and random survival forest model analysis to be involved in BOS development. Orange nodes show terms that link the selected genes and proteins to each other and to BOS (red node). These links were obtained by literature search.

Hemoglobins

I found a markedly reduced expression of hemoglobin (Hb) α (HbA) and β (HbB) chains [254] in BOS-positive samples, as expressed by alveolar type 2 cells and Clara cells [255,256,22,257,258]. The significantly lower hemoglobin gene expression in bronchial epithelial cells was also confirmed at the BALF proteome level by MALDI-TOF profiling. The presence of the HbB protein in BALF was formerly reported in reference [259]. These findings were further supported by the down-regulation of the AHSP (alpha-hemoglobin-stabilizing protein) gene in Bos-positive samples [260,261,262,263]. The alpha-hemoglobin-stabilizing protein (AHSP) serves as a molecular chaperone for free hemoglobin α chains. Current scientific evidence suggests AHSP participation in hemoglobin synthesis and a possible role in neutralization of the toxic effects caused by excess accumulation of free alpha-hemoglobin subunits. Hemoglobins have been shown to play a role in various processes, including gas exchange [264] and protection against oxidative stress [265]. Hb down-regulation was previously correlated with diverse lung diseases [26,259], such as idiopathic pulmonary fibrosis (IPF)[257]. Oxidative stress occurs in pulmonary fibrosis [266,267], where cells are damaged by reactive oxygen species, and has been linked to BOS development and progression [268,269,270]. Up-regulation of HbA and HbB was observed in BALF after the first year post lung transplantation. The up-regulated gene expression of HbA and HbB genes in bronchial epithelial cells was shown to be linked to lung transplantation (LTx) [22], hypoxia in alveolar epithelial cells [258] and in response to oxidative stress in hepatocytes [265]. A down-regulation of hemoglobins in lung transplants that are affected by BOS could thus further facilitate cell damage by oxidative stress.

Secretoglobins

Another protein expressed in Clara cells [271], is the Clara cell secretory protein (uteroglobin or secretoglobin (SCGB) 1A1). Clara cell secretory protein (CCP) is a low-molecular-weight protein secreted into the alveoli in large quantities by Clara cells [272,273]. Clara cells, the primary site of xenobiotic detoxification, form an abundant population of non-ciliated secretory cells within the bronchiolar epithelium of the mammalian lung. They serve as a progenitor pool for bronchial epithelial steady-state maintenance and regeneration after lung injury. Clara cells also provide nonmucinous proteins to the extracellular lining fluid, among them the Clara cell secretory protein (CCP). This protein, one of the most abundant respiratory-tract derived proteins, is suggested to have immunosuppresive and antiprotease properties [274,275,276,277]. Alterations in Clara cell function are often assumed to play a significant role in lung function decrease in airway diseaes, and CCP is known to be deregulated in a number of diseases that damage the lung epithelium [26,25]. This study confirmed the down-regulation of CCP for the bronchiolitis obliterans syndrome [278,238], as the CCP gene was found to be signifcantly down-regulated in BOS-positive samples at both the transcriptomic and proteomic level. Further a significant down-regulation was observed in BALF after the first year post lung transplantation. While on its own CCP is not a reliable predictor for development of BOS

[279], it showed promising performance when combined with other biomarkers as shown in reference [238], and by my BALmon classifer model. The study in [238] predicted BOS using CCP in combination with Lysozyme, for which I found Lysozyme (Lyz) gene expression up-regulated in BOS-postive cell samples and Lysozyme c down-regulated in the BOS-positive BALF samples [280,281]. Secretoglobin (SCGB) 3A2 [282], like CCP a member of the secretoglobin family, also showed a decreased gene expression level in BOS-positive samples. The members of the SCGB gene superfamily, found only in mammals, are all cytokine-like secreted proteins of 10 kDa. The SCGB3A2 (secretoglobin, family 3A, member 2) [282,16,283], which is predominently expressed in the pulmonary airway epithelium, shows anti-inflammatory and growth factor activities. Using pulmonary fibroblasts isolated from adult mice [283], it was shown to inhibit the TGFβ-induced differentiation of fibroblasts to myofibroblasts. This differentiation is a hallmark of the fibrotic process. Thus a down-regulation of SCGB3A2 might result in increased fibrosis in BOS afflicted patients. The SCGB3A1 (secretoglobin, family 3A, member 1), as opposed to the other analysed secretoglobin family members, showed a significantly increased gene expression level in BOS-positive samples. It was shown in [282] that SCGB3A1 also exhibits a different localization pattern than CCP and SCGB3A2. While CCP and SCGB3A1 gene expression serve as a molecular marker for Clara cells distributed throughout the conducting airway epithelium, SCGB3A1 is highly expressed in a subset of airway epithelial cells located primarily in the bronchi [282]. This might serve as possible explanation for the opposing patterns of differential expression in BOS-positive samples.

Pulmonary Surfactant

Pulmonary surfactant [284], a surface-active lipoprotein complex (phospholipoprotein), is formed by type II alveolar cells. This complex was originally known for its role reducing surface tension at the air liquid layer. Recent discoveries revealed surfactant proteins as an important component of the lung immune host defense. In this thesis, surfactant protein A (SPA) [285] was found to be down-regulated on the protein level after the first year post lung transplantation. The down-regulation of SPA has already been linked to BOS in reference [286]. In contrary to reference [286], I found an up-regulation of SPA in BOS-positive BALF samples. Its differential expression was detected only on the protein level, as surfactant protein A was not part of the gene set measured by the microarrays that had been used in our study. Several other genes from the surfactant family were found to be down-regulated in BOS-postive samples: surfactant protein B (SPB) [287], surfactant protein C (SPC) [288] and surfactant protein D (SPD) [289]. As a part of the innate immune system, the Surfactant proteins A and D regulate innate immune cells of the immunologic enviroment of the lung [290,291]. This allows them to modulate an extreme inflammatory reponse, which could otherwise result in a reduced gas exchange by lung damage [284]. The significant deregulation of surfactant protein D could not be confirmed on the protein level by a recent study [279]. Conflicting reports and results of differential surfactant expression in patients afflicted by BOS prevent a clear conclusion concerning their role in BOS development [286,279].

Defensins

Defensins are part of a large family of peptides, which were first known for their broad spectrum of activity against bacteria, fungi and viruses [292]. Recent studies expanded on their role in promotion of innate and adaptive immune reponses, recruitment of inflammatory cells as well as anti-inflammatory effects [292]. Two members of the alpha-defensin category (human neutrophil peptide (HNP) 1-2, or alpha defensin (DEF) 1-2) were found to be up-regulated in BOS-positive BALF samples. Conversely both HNP 1 and 2 were found to be significantly down-regulated in BALF samples after the first year post lung transplantation. HNPs were previously shown to exhibit BOS-related changes but are believed to be involved in other respiratory conditions as well [293,294]. Up-regulation of these two proteins, neutrophil alpha defensin 1 (DEF1 or HNP1) and neutrophil alpha defensin 2, has already been reported in the references [248,238]. Members of the HNP family are constitutively expressed and packaged in azurophil granules of neutrophiles from which they are released in large quantities during neutrophil activation [292]. In both studies [248,238] the HNP levels did not correlate with episodes of acute rejection, *cytomegalovirus* or fungal infection. This is in line with chronic obstructive pulmonary disease (COPD) and idiopathic pulmonary fibrosis (IPF), other diseases with elevated HNP levels, for which infection is not a predominant clinical finding [295,296,248]. Human neutrophil peptides have also been linked to the pathogenesis of acute lung injury [297], which shows acute inflammation with neutrophil inifiltration, fibroproliferation with hyaline membranes and varying degrees of interstitial fibrosis [298]. In a mouse model the effects caused by α-defensins were shown to be mediated by signal transduction through low-density lipoprotein-related receptor (LRP), for which the LRP5 (low-density lipoprotein-related protein 5) gene showed a significantly lower expression level in BOS-positive bronchial epithelial cell samples. Other than for HNP 1-3, the differential expression of human neutrophil peptide 4 (HNP4 or DEF4) was not reported in [248,238]. I found that, HNP4 (DEFA4) gene expression was significantly down-regulated in bronchial epithelial cells of BOS-positive samples. The gene expression of defensin beta 1 (DEFB1) a member of the β-defensins (hBDs), was up-regulated in BOS-positive samples. The hBDs are expressed by mucosal epithelial cells, and are secreted from these cells into the surrounding lining fluid [292]. They support the initial host defense against infection and the maintenance of epithelial integrity [292]. The up-regulated level of defensin beta 2 in the bronchoalveolar lavage of lung transplants has already been linked to the bronchiolitis oblitenans syndrome [299]. The up-regulated gene expression of defensin beta 1(DEFB1) thus further supports the theory that β-defensins play a role in BOS progression.

Histatins

Histatins [300], as secreted by human salivary glands [301], are a group of proteins found in saliva [302] and bronchoalveolar lavage fluid [11] and are a part of human airway surface liquid (ASL) [303]. They show strong antimicrobial activity, especially strong antifungal properties and stimulate wound closure of both oral and non-oral cells [304,305]. The analysed BOS-positive BALF samples showed a significant down-regulation of histatin-3. Despite a specific role of histatins in BOS development and progression being unknown, this finding still adds to the current topics in BOS research. This is due to the circumstance that lung allograft airway colonization by fungal *aspergillus species* and *pseudomonas aeruginosa* have been associated with BOS [68,306]. Histatin-3 also showed a significant down-regulation after the first year post lung transplantation. This, together with the accompanying down-regulation of the human neutrophil peptides (HNP1 and HNP2), is in line with an increased susceptibility of the pulmonary allograft to infection [307]. In a gene expression study histatins were found only in parotid and submandibular glands [308], suggesting a specifity to salivary secretions. Thus, while the findings are interesting in the context of lung transplantation, further research will be necessary to explain the presence and function of histatins in lung fluids.

Calgranulins

For calgranulin A (S100A8), which was suspected as a potential BOS biomarker in reference [238], a significant down-regulation was found in BOS-positive BALF samples. Calgranulin A is a member of the S100 family [309] of calcium binding proteins. These are expressed in multiple cells, including activated macrophages, monocytes and neutrophils. So far altered expression of S100 family members has also been linked to cystic fibrosis [310], COPD [311], idiopathic fibrosis [312] and lung cancer [11,313]. In alveolar epithelial cells gene expression of S100A8 samples was found to be significantly up-regulated, together with S100P [314], S100A10 [315,313] and S100A11 [316]. S100A8, a pro-inflammatory mediator in acute and chronic inflammation [311,317,318], promotes phagocyte migration [319,320,321] and infiltration of granulocytes [322,323] to the sites of tissue damage [324]. It also supports antifungal activities [325], and eight members of the S100 family, including S100A8, show differential gene expression in bronchial epithelial cells after *pseudomonas aeruginosa* exposure [326]. S100A12 (calgranulin C), another pro-inflammatory protein member of the S100 family, showed significantly lower gene expression in BOS-positive epithelial cell samples. In references [311,259] S100A12 was shown to be more associated with acute lung inflammation, while chronic inflammation is mainly mediated by S100A8/A9. Thus the transcriptomic down-regulation of S100A12 would still be in line with BOS being defined by chronic inflammation and rejection processes [327].

Aldehyd Dehydrogenases

Both ALDH3A1 (aldehyde dehydrogenase 3 family , member A1) and ALDH3B2 (aldehyde dehydrogenase 3 family, member B2) were found to show a significantly decreased gene expresion level in BOS-positive epithelial cell samples. The enzymes encoded by these genes belong to the aldehyde dehydrogenase family and are constitutively expressed in the lung [328]. Aldehyde dehydrogenase, which oxidizes mainly aromatic and long chain aliphatic aldehydes, is the second enzyme of the major oxidative pathway of alcohol metabolism. It protects organisms from aldehydes from food and air pollution, and was reported to be a tumor stem cell-associated marker in lung cancer [329,330,331]. Further these enzymes fullfill several noncatalytic duties, including antioxodative functions and structural roles [331]. As cell damage from oxidative stress is linked to BOS development and progression [268,269,270], the down-regulation of ALDH3A1 might partially explain an increased susceptibility to oxidative stress-induced cell damage. While ALDH3A1 and ALDH3B2 gene expression was down-regulated in BOS-positive samples, the expression of ALDH3B2 has already been reported to be up-regulated in transplanted lungs [22]. The same regulation pattern was also found in this thesis for hemoglobins that offer protection against oxidative stress [265]

HLA

Among the genes identified as informative for BOS progression by the random survival forest model (RSF) were several HLA (human leukocyte antigens) genes. The HLA genes were the unifying element between the significantly enriched pathways. Alloreactivity directed toward HLA antigens has already been reported to play a role in the pathogenesis of BOS. The exclusively α-subunit HLA genes (HLA-DMA,HLA-DRA and HLA-DQA1), as selected by the RSF model, were all part of the HLA class II group. Class II HLA antigens are expressed in cilliated bronchial epithelial cells [332,333,334]. During chronic rejection this expression is known to be up-regulated [335,336,337]. A significantly up-regulated gene expression of the HLA Class II genes was confirmed for BOS-positive samples and correlated significantly with an elevated risk of BOS development. While a clear consensus on the the role of HLA mismatching as a risk factor for chronic lung allograft rejection has not yet been reached [338], several studies supplement this idea [60]. In total this thesis further supplements the idea of HLA involvement in BOS risk and progression.

Proteases

Among the genes selected by the RSF model were the proteases (peptidases) tryptase alpha/beta 1 (TPSAB1), kallikrein 7 (KLK7) and cathepsin Z (CTSZ). Proteases belong to a group of enzyme families that catalyse protein breakdown. Their activity and expression is governed by several regulative mechanisms that prevent uncontrolled proteolytic action. Proteases play a role in the control of several extracellular and cellular processes. These extracellular processes include extracellular matrix turnover and responses to tissue injury and healing. In the cells they are able regulate gene expression, cell differentiation and proliferation, or induce cell death through limited proteolysis [339].

Tryptases and Mast Cells

A cell type that is especially rich in proteases are mast cells. The mast cell peptidases are stored in and released from secretory granules. Most of the proteolytic enzymes selectively concentrated in the secretory granules of human mast cells show trypsin-

or chemotrypsin like activities [340]. Among these are tryptases alpha and beta, proteases that in humans are both encoded by the same TPSAB1 gene [341]. Tryptases are known to stimulate fibroblast proliferation [342,343]. The expression of the TPSAB1 gene was significantly up-regulated in BOS-positive cell samples and in samples defined as high risk by the RSF model. It was also shown that an increase of TPSAB1 expression correlated significantly with a decrease in BOS free time. In a previous study, elevated tryptase levels in BAL were linked to the development of fibrosis in idiopathic bronchiolitis obliterans organizing pneumonia (BOOP) [344]. The presence of tryptases is a reliable marker for mast cell activation[341]. Mast cells have already been linked to most infiltrative fibrotic lung diseases [345,346,344] and show the ability to produce fibrogenic cytokines [347]. While mast cells are not able to induce fibrosis on their own [348], their close morphologic and functional relationship with fibroblasts further supports their primary involvement in fibrosis [349,350].

Kallikrein 7 and Epithelial-Mesenchymal Transition

Gene expression of the protease kallikrein 7 (KLK7) [351] was significantly up-regulated in BOS-positive bronchial epithelial cell samples. Kallikrein 7 is a chymotrypsin-like secreted serine protease that is known to catalyze the degradation of intercellular adhesive structures in the cornified layer of the skin [352]. KLK7 expression was rated as the most informative feature of the multivariate random survival forest (RSF) model. An up-regulation of KLK7 expression was only predictive of BOS development when considered in combination with other genes. The RSF based model analysis revealed such BOS specific interactions with human leukocyte antigens (HLA-DMA, HLA-DRA and HLA-DQA1) and the tryptase TPSAB1. As already shown for TPSAB1, KLK7 has also been linked to fibrosis. A recent study demonstrated that active kallikrein 7 is involved in extracellular matrix degradation and shows the ability to cleave fibronectin [353]. Another study showed that KLK7 induces epithelial-mesenchymal transition-like changes in prostate cacrinoma cells [352]. Mesenchymal cells are known to migrate into the airspace after acute lung injury, where they deposit connective tissue macromolecules [354]. The occurence of epithelial-mesenchymal transition (EMT) in lung alveolar epithelial cells has been proposed as a mechanism to explain the increased fibroblast numbers, collagen overproduction and fibrosis ocurring in chronic obstructive pulmonary disease (COPD), asthma and pulmonary fibrosis [355]. It was also suggested that for bronchiolitis obliterans EMT could serve as a potential link between injury caused by inflammation and airway remodelling as driven by TGF-β [355]. In this study [355] TGF-β induced remodelling was linked to an reduced expression of epithelial markers as well as an increased expression of vimentin and fibronectin. In the BOS-positive cell samples, as obtained from bronchial epithelial brushing, several differentially expressed genes might support the idea of EMT occurence. These include the up-regulated IL1B (Interleukin 1, beta), FGFBP1 (fibroblast growth factor binding protein 1), LRFN5 (leucine rich repeat and fibronectin type III domain containing 5), ACTG2 (actin, gamma 2, smooth muscle, enteric), PCOLCE2 (procollagen C-endopeptidase enhancer 2) and KRT8 (keratin 8) [356]. A down-regulated gene expression was detected for the cadherin-related family member 1 (CDHR1), which is a member of the cadherin superfamily of calcium-dependent cell-cell adhesion molecules. The down-regulation of epithelial cadherin (E-cadherin), another member of the cadherin superfamily, has already been linked to EMT in several studies [355,357]. During EMT, epithelial cells lose their epithelial properties and gain mesenchymal ones. These mesenchymal properties include the production of metalloproteinases [355] . Two metalloproteinase coding genes (ADAM metallopeptidase domain 9 (ADAM9) and Matrix metallopeptidase 9 (gelatinase B, 92kDa gelatinase, 92kDa type IV collagenase) (MMP9)) were found to be up-regulated in BOS-positive samples. EMT does not seem to occur in lung transplants not suffering from BOS, as on the protein level vimentin was found to be down-regulated in BALF samples taken after the first year post transplantation. Vimentin is a type III intermediate filament (IF) protein and the major cytoskeletal component of mesenchymal cells. Vimentin [358], which plays a role in organ rejection [359] and airway remodeling after lung transplantation [355], serves as a biomarker of mesenchymally derived cells or cells undergoing EMT. Vimentin silencing leads to a change in mesenchymal cell shape [360].

Cathepsin Z and Integrins

Gene expression of the lysosomal cysteine protease cathepsin Z (CTSZ), which is also known as cathespin X (CTSX), was significantly up-regulated in BOS-positive bronchial epithelial cell samples. It was also shown that expression of the CTSZ gene was significantly up-regulated in samples defined as high risk by the random survival forest (RSF) model. An increase of CTSZ expression correlated significantly with a decrease in BOS free time. CTSZ gene expression was rated as one of the most informative features of the multivariate RSF model. The RSF based model analysis revealed that the expression of CTSZ best represents the BOS specific influence of the human leukocyte antigens (HLA-DRA and HLA-DQA1), tryptase alpha/beta 1 (TPSAB1) and complement factor B (CFB). The carboxypeptidase cathepsin Z shows unique properties that set it apart from other cysteine proteases [361]. Cathepsin Z expression is restricted to cells of the immune system, such as dendrictic cells, macrophages and monocytes [362]. Cathepsin Z modulates the activity of the β_2-integrin receptor LFA-1 (Lymphocyte function-associated antigen 1) by sequential cleavage of the C-terminal amino acids of the β_2 intergin subunit [363,364]. Integrins are heterodimers containing two distinct α and β chains. The common integrin β_2 chain (CD18) of all integrins is the protein product of the ITGB2 (integrin beta chain beta 2) gene. The integrin receptor LFA-1 is formed by the combination of β_2 and the alpha L chain (CD11a/CD18, $\alpha_L\beta_2$). The modulation of LFA-1 by CTSZ causes cytoskeletal rearrangement and increased migration of T lymphocytes [363,364]. This is interesting, since T lymphocyte infiltration is commonly described as being linked to BOS development [4]. It was shown in reference [365] that cathepsin Z possesses the ability to activate the β_2-integrin receptor Mac-1 (Macrophage-1 antigen, CD11b/CD18, $\alpha_M\beta_2$)) [366]. This MAC-1 activation promotes the adhesion of monocytes and macrophages to fibrinogen and regulates phagocytosis [363,364]. The participation of fibrinogen in the immune and inflammatory repsonse depends on the specific interaction with leukocyte integrin surface receptors. In macrophages, neutrophils, monocytes and several lmyphocyte subsets the main fibrinogen receptors are MAC-1 [367] and inactivated-C3b (iC3b) receptor 4 (CR4) [368]. The later receptor (CD11c/CD18,$\alpha_x\beta_2$) is formed by the combination of β_2 and the integrin alpha X chain protein. This protein is encoded by the Integrin, alpha X (complement component 3 receptor 4 subunit) (ITGAX,CD11c) gene. Gene expression of ITGAX was found to be significantly up-regulated in the BOS-positive bronchial brushing samples. As BOS is a condition of intraluminal airway fibrosis [4] these findings might prove to be very important for the understanding of BOS pathogenesis.

4.2 Block Maxx

In this thesis, a novel approach for biclustering of gene expression data was presented. This algorithm (Block Maxx) focuses on the discovery of a set of sample clusters defined by the consistent up- and down-regulation of a subset of genes. The method uses spectral bipartite graph partitioning as its framework. This framework was extended by a background model, to allow for the removal of genes and samples that represent noise and can not be clearly assigned to a specific cluster. The algorithm also allows for a cluster to be defined by a *mixture* of up-regulated and down-regulated genes. So far, methods modelling the data as a bipartite graph were only able to find clusters defined by gene expression up-regulation. The algorithm was applied to a publicly available B-cell lymphoma microarray data set [231] and revealed four stable patient clusters. A bootstrap resampling scheme was used to rank the genes in accordance with their clusterwise importance (gene membership score). The ranking was shown to represent significant differential gene expression between the detected clusters. This establishes the cluster membership score as a reliable way to rate gene importance in a bicluster setting. The obtained clustering of the lymphomas also revealed interesting insights into the biology of B-cell lymphomas. The clusters were ordered according to decreasing survival time. Several pathways and gene sets were found significantly enriched in the lists of genes associated with the respective clusters. The first and second cluster, as well as the third and fourth cluster, show opposite patterns of gene expression (Figure 3.15). Genes up-regulated in the first cluster are down-regulated in the second cluster, while genes down-regulated in the first cluster are up-regulated in the second. The same principle applies to the third and fourth cluster. They consequently present opposite patterns of pathway or gene-set enrichment (Table 3.24). The patients in these four clusters also show a significant difference in survival time, where the patients in the first and fourth clusters have the best and worst prognosis of all patients, respectively. The Biocarta "Rab" pathway is overrepresented in the down-regulated genes of the first cluster and in the up-regulated genes of the second cluster. Interestingly, overexpression of Rab genes has been associated with hematological malignancies [369,370], and this correlates with the worse prognosis of the second group, which presents up-regulation of this pathway, as compared to the first cluster. The Biocarta "p38 MAPK" signaling pathway, a known signaling event associated with several tumors, such as breast cancers [371], is also down-regulated and up-regulated in the first and second cluster, respectively, suggesting a protective role of the inhibition of this pathway also in lymphoma. Particularly interesting is the differentially regulation of oxidative phosphorylation in the first and second cluster, as well as in third and fourth cluster. It is known that different cancers, and even cancer subgroups, are selective in the means of producing energy, varying from glycolytic to oxidative types [372]. Interestingly, the KEGG pathway "oxidative phosphorylation" is overrepresented in the up-regulated genes of the first and third clusters and in the down-regulated genes of the second and fourth clusters. This finding suggests two different subgroups of lymphoma based on their ability to produce energy from oxidative phosphorylation.

4.3 Outlook

This study revealed that the clinical implementation of a predictor model such as BALmon, based on proteomic markers, could ultimately supplement established diagnostic procedures. Validation on a separate patient cohort using multiplexed enzyme-linked immunosorbent assay and multiple reaction monitoring mass spectrometry will be performed. This will also allow to confirm BOS-related changes detected on the gene expression level in the proteome of BALF samples. For the Block Maxx algorithm, an R package will be made available. This will allow the community to employ Block Maxx in an easy and open source fashion. Block Maxx will also be extended to allow for semi supervised clustering. Semi-supervised clustering refers to a group of machine learning methods that can make use of a combination of labeled and unlabeled data. Normally a large amount of unlabeled data and a small amount of labeled data. For this the bipartite graph model will be extended to allow edges between samples that are already known to belong to the same clinically defined group.

Bibliography

[1] K. Simmons, J. Kinney, A. Owens, D. A. Kleier, K. Bloch, D. Argentar, A. Walsh, and G. Vaidyanathan, "Practical outcomes of applying ensemble machine learning classifiers to High-Throughput Screening (HTS) data analysis and screening.," *Journal of Chemical Information and Modeling*, vol. 48, no. 11, pp. 2196–2206, 2008.

[2] C. Mballo and V. Makarenkov, "Using machine learning methods to predict experimental high-throughput screening data.," *Combinatorial chemistry high throughput screening*, vol. 13, no. 5, pp. 430–441, 2010.

[3] C. Baumgartner, M. Osl, M. Netzer, and D. Baumgartner, "Bioinformatic-driven search for metabolic biomarkers in disease," *Journal of Clinical Bioinformatics*, vol. 1, no. 1, p. 2, 2011.

[4] J. L. Todd and S. M. Palmer, "Bronchiolitis obliterans syndrome: the final frontier for lung transplantation.," *Chest*, vol. 140, no. 2, pp. 502–8, Aug. 2011.

[5] a. G. N. Robertson, S. M. Griffin, D. M. Murphy, J. P. Pearson, I. a. Forrest, J. H. Dark, P. a. Corris, and C Ward, "Targeting allograft injury and inflammation in the management of post-lung transplant bronchiolitis obliterans syndrome.," *American journal of transplantation : official journal of the American Society of Transplantation and the American Society of Transplant Surgeons*, vol. 9, no. 6, pp. 1272–8, Jun. 2009.

[6] J. Gottlieb, "Update on lung transplantation.," *Therapeutic advances in respiratory disease*, vol. 2, no. 4, pp. 237–47, Aug. 2008.

[7] J Gottlieb, T Welte, M. M. Höper, M Strüber, and J Niedermeyer, "[Lung transplantation. Possibilities and limitations].," *Der Internist*, vol. 45, no. 11, pp. 1246–59, Nov. 2004.

[8] T. Wolf, T. Oumeraci, J. Gottlieb, A. Pich, B. Brors, R. Eils, A. Haverich, B. Schlegelberger, T. Welte, M. Zapatka, and N. von Neuhoff, "Proteomic Bronchiolitis Obliterans Syndrome Risk Monitoring in Lung Transplant Recipients.," *Transplantation*, vol. 92, no. 4, pp. 477–485, Aug. 2011.

[9] T. Wolf, B. Brors, T. Hofmann, and E. Georgii, "Global Biclustering of Microarray Data," *Sixth IEEE International Conference on Data Mining - Workshops (ICDMW'06)*, pp. 125–129, 2006.

[10] M. Roesch-Ely, A. Leipold, M. Nees, D. Holzinger, A. Dietz, C. Flechtenmacher, T. Wolf, M. Zapatka, and F. X. Bosch, “Proteomic analysis of field cancerization in pharynx and oesophagus: a prospective pilot study.,” *The Journal of pathology*, vol. 221, no. 4, pp. 462–470, Aug. 2010.

[11] T. Oumeraci, B. Schmidt, T. Wolf, M. Zapatka, A. Pich, B. Brors, R. Eils, M. Fleischhacker, B. Schlegelberger, and N. von Neuhoff, “Bronchoalveolar lavage fluid of lung cancer patients: Mapping the uncharted waters using proteomics technology.,” *Lung cancer (Amsterdam, Netherlands)*, vol. 72, no. 1, pp. 136–138, Apr. 2011.

[12] Y. Reis, M. Bernardo-Faura, D. Richter, T. Wolf, B. Brors, A. Hamacher-Brady, R. Eils, and N. R. Brady, “Multi-parametric analysis and modeling of relationships between mitochondrial morphology and apoptosis.,” *PloS one*, vol. 7, no. 1, e28694, Jan. 2012.

[13] R. D. C. Team, *R: A Language and Environment for Statistical Computing*, 2010.

[14] J. H. Bullard, “Introduction to Bioconductor R / Bioconductor : A Short Course Bioconductor : Overview ExpressionSet,” *Environments*, pp. 1–33, 2001.

[15] K. S. Pollard, “Introduction to the R language and Bioconductor Acknowledgments,” *Cancer Research*, pp. 1–52, 2005.

[16] R. G. Crystal, S. H. Randell, J. F. Engelhardt, J. Voynow, and M. E. Sunday, “Airway epithelial cells: current concepts and challenges.,” *Proceedings of the American Thoracic Society*, vol. 5, no. 7, pp. 772–7, Sep. 2008.

[17] D. Wang, D. L. Haviland, A. R. Burns, E. Zsigmond, and R. a. Wetsel, “A pure population of lung alveolar epithelial type II cells derived from human embryonic stem cells.,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 11, pp. 4449–54, Mar. 2007.

[18] R. L. Zemans and M. a. Matthay, “Bench-to-bedside review: the role of the alveolar epithelium in the resolution of pulmonary edema in acute lung injury.,” *Critical care (London, England)*, vol. 8, no. 6, pp. 469–77, Dec. 2004.

[19] E. Puchelle, J.-M. Zahm, J.-M. Tournier, and C. Coraux, “Airway epithelial repair, regeneration, and remodeling after injury in chronic obstructive pulmonary disease.,” *Proceedings of the American Thoracic Society*, vol. 3, no. 8, pp. 726–733, 2006.

[20] D. Sinclair, “Atlas der Anatomie des Menschen (Sobotta/Becher),” *Journal of Anatomy*, vol. 112, no. Pt 1, p. 146, 1972.

[21] P. Govender, M. J. Dunn, and S. C. Donnelly, “Proteomics and the lung: Analysis of bronchoalveolar lavage fluid.,” *Proteomics. Clinical applications*, vol. 3, no. 9, pp. 1044–51, Oct. 2009.

[22] B. Skawran, M. Dierich, D. Steinemann, J. Hohlfeld, A. Haverich, B. Schlegelberger, T. Welte, and N. von Neuhoff, “Bronchial epithelial cells as a new source for differential transcriptome analysis after lung transplantation.,” *European journal of cardio-thoracic surgery : official journal of the European Association for Cardio-thoracic Surgery*, vol. 36, no. 4, pp. 715–21, Oct. 2009.

[23] W. R. Fields, R. M. Leonard, P. S. Odom, B. K. Nordskog, M. W. Ogden, and D. J. Doolittle, “Human bronchial epithelial cell transcriptome: gene expression changes following acute exposure to whole cigarette smoke in vitro.,” *American journal of physiology Lung cellular and molecular physiology*, vol. 86, no. 1, pp. 84–91, 2007.

[24] B. Magi, E. Bargagli, L. Bini, and P. Rottoli, “Proteome analysis of bronchoalveolar lavage in lung diseases.,” *Proteomics*, vol. 6, no. 23, pp. 6354–69, Dec. 2006.

[25] E. Kriegova, C. Melle, V. Kolek, B. Hutyrova, F. Mrazek, A. Bleul, R. M. du Bois, F. von Eggeling, and M. Petrek, “Protein profiles of bronchoalveolar lavage fluid from patients with pulmonary sarcoidosis.,” *American journal of respiratory and critical care medicine*, vol. 173, no. 10, pp. 1145–54, May 2006.

[26] D. Merkel, W. Rist, P. Seither, A. Weith, and M. C. Lenter, “Proteomic study of human bronchoalveolar lavage fluids from smokers with chronic obstructive pulmonary disease by combining surface-enhanced laser desorption/ionization-mass spectrometry profiling with mass spectrometric protein identification.,” *Proteomics*, vol. 5, no. 11, pp. 2972–80, Jul. 2005.

[27] D. S. Allen-Gipson, A. A. Floreani, A. J. Heires, S. D. Sanderson, R. G. MacDonald, and T. A. Wyatt, “Cigarette smoke extract increases C5a receptor expression in human bronchial epithelial cells.,” *The Journal of pharmacology and experimental therapeutics*, vol. 314, no. 1, pp. 476–482, 2005.

[28] Y Nakamura, L Tate, R. F. Ertl, M Kawamoto, T Mio, Y Adachi, D. J. Romberger, S Koizumi, G Gossman, and R. A. Robbins, “Bronchial epithelial cells regulate fibroblast proliferation.,” *American Journal of Physiology*, vol. 269, no. 3 Pt 1, pp. L377–L387, 1995.

[29] K. Steiling, A. Y. Kadar, A. Bergerat, J. Flanigon, S. Sridhar, V. Shah, Q. R. Ahmad, J. S. Brody, M. E. Lenburg, M. Steffen, and A. Spira, “Comparison of Proteomic and Transcriptomic Profiles in the Bronchial Airway Epithelium of Current and Never Smokers,” *PLoS ONE*, vol. 4, no. 4, C. Feghali-Bostwick, Ed., p. 10, 2009.

[30] H. Takizawa, “Bronchial epithelial cells in allergic reactions.,” *Current drug targets Inflammation and allergy*, vol. 4, no. 3, pp. 305–311, 2005.

[31] S. E. Ilyin, S. M. Belkowski, and C. R. Plata-Salamán, “Biomarker discovery and validation: technologies and integrative approaches.,” *Trends in biotechnology*, vol. 22, no. 8, pp. 411–6, Aug. 2004.

[32] M. Dennis, D. J. Benos, D. Editor, and F. Abboud, “Review Pathogenesis of Early Lung Disease in Cystic Fibrosis : A Window of IN,” *Annals of Internal Medicine*, pp. 816–822, 2005.

[33] E. B. Meltzer and P. W. Noble, “Idiopathic pulmonary fibrosis.,” *Orphanet journal of rare diseases*, vol. 3, p. 8, Jan. 2008.

[34] D. M. Mannino and a. S. Buist, “Global burden of COPD: risk factors, prevalence, and future trends.,” *Lancet*, vol. 370, no. 9589, pp. 765–73, Sep. 2007.

[35] S. D. Nathan, “Pulmonary Hypertension and Pulmonary Function Testing in Idiopathic Pulmonary Fibrosis,” *Chest*, vol. 131, no. 3, pp. 657–663, Mar. 2007.

[36] B. F. Meyers, J. Lynch, E. P. Trulock, T. J. Guthrie, J. D. Cooper, and G. A. Patterson, “Lung Transplantation : A Decade of Experience,” *Annals of Surgery*, vol. 230, no. 3, pp. 362–371, 1999.

[37] J. D. Hardy, W. R. Webb, M. L. Dalton, and G. R. Walker, “HOMOTRANSPLANTATION OF LUNG IN MAN.,” *Jama The Journal Of The American Medical Association*, vol. 66, pp. 1065–1074, 1964.

[38] D. A. Blumenstock and C Lewis, “The first transplantation of the lung in a human revisited.,” *The Annals of Thoracic Surgery*, vol. 56, no. 6, 1423–1424; discussion 14241425–, 1993.

[39] P. J. Morris, “The impact of Cyclosporin A on transplantation.,” *Advances in Surgery*, vol. 17, pp. 99–127, 1984.

[40] J. D. Cooper, R. J. Ginsberg, M Goldberg, G. A. Patterson, F. G. Pearson, T. R. J. Todd, and Et Al., “Unilateral lung transplantation for pulmonary fibrosis,” *New England Journal of Medicine*, vol. 314, pp. 1140–1145, 1986.

[41] G. A. Patterson, J. D. Cooper, B Goldman, R. D. Weisel, F. G. Pearson, P. F. Waters, T. R. Todd, H Scully, M Goldberg, and R. J. Ginsberg, “Technique of successful clinical double-lung transplantation.,” *The Annals of Thoracic Surgery*, vol. 45, no. 6, pp. 626–633, 1988.

[42] K. R. McCurry, T. H. Shearon, L. B. Edwards, K. M. Chan, S. C. Sweet, M Valapour, R Yusen, and S Murray, “Lung transplantation in the United States, 1998-2007.,” *American journal of transplantation : official journal of the American Society of Transplantation and the American Society of Transplant Surgeons*, vol. 9, no. 4 Pt 2, pp. 942–58, Apr. 2009.

[43] K. Murphy, P. Travers, and M. Walport, *Janeway’s Immunobiology*, C. B. Janeway and M. Ehrenstein, Eds. Garland Science, 2008, vol. 7, p. 928.

[44] K. A. Hogquist, T. A. Baldwin, and S. C. Jameson, “Central tolerance: learning self-control in the thymus.,” *Nature Reviews Immunology*, vol. 5, no. 10, pp. 772–82, 2005.

[45] D. L. Kamen and C. Strange, “Pulmonary manifestations of systemic lupus erythematosus.,” *Clinics in chest medicine*, vol. 31, no. 3, pp. 479–88, Sep. 2010.

[46] B Vaidya, "The Genetics of Autoimmune Thyroid Disease," *Journal of Clinical Endocrinology Metabolism*, vol. 87, no. 12, pp. 5385–5397, 2002.

[47] K Federlin, "Diabetes mellitus and immunology–a manifold interrelation," *Immunitat Und Infektion*, vol. 13, no. 5, pp. 193–199, 1985.

[48] C. Y. Ng, J. C. Madsen, B. R. Rosengard, and J. S. Allan, "Immunosuppression for lung transplantation.," *Proceedings of the American Thoracic Society*, vol. 6, no. 3, pp. 1627–1641, 2009.

[49] D. F. LaRosa, A. H. Rahman, and L. A. Turka, "The innate immune system in allograft rejection and tolerance.," *The Journal of Immunology*, vol. 178, no. 12, pp. 7503–9, 2007.

[50] C. P. Larsen, P. J. Morris, and J. M. Austyn, "Migration of dendritic leukocytes from cardiac allografts into host spleens. A novel pathway for initiation of rejection.," *The Journal of Experimental Medicine*, vol. 171, no. 1, pp. 307–314, 1990.

[51] D. S. Game and R. I. Lechler, "Pathways of allorecognition: implications for transplantation tolerance.," *Transplant Immunology*, vol. 10, no. 2-3, pp. 101–108, 2002.

[52] Y. W. J. Sijpkens, I. I. N. Doxiadis, M. J. K. Mallat, J. W. De Fijter, J. A. Bruijn, F. H. J. Claas, and L. C. Paul, "Early versus late acute rejection episodes in renal transplantation.," *Transplantation*, vol. 75, no. 2, pp. 204–208, 2003.

[53] A. M. Waaga, M Gasser, I Laskowski, and N. L. Tilney, "Mechanisms of chronic rejection.," *Current Opinion in Immunology*, vol. 1, no. 3, pp. 1230–1235, 2000.

[54] J Lew and J. A. Smith, "Mucosal graft-vs-host disease.," *Oral Diseases*, vol. 13, no. 6, pp. 519–529, 2007.

[55] J Prop, C. R. Wildevuur, and P Nieuwenhuis, "Acute graft-versus-host disease after lung transplantation.," *Transplantation Proceedings*, vol. 21, no. 1 Pt 3, pp. 2603–2604, 1989.

[56] H. Luckraz, M. Zagolin, K. McNeil, and J. Wallwork, *Graft-versus-host disease in lung transplantation: 4 case reports and literature review.* 2003.

[57] D. M. Smith, E. D. Agura, K. Ausloos, W. S. Ring, R. Domiati-Saad, and G. B. Klintmalm, *Graft-vs-host disease as a complication of lung transplantation.* 2006.

[58] A. Fossi, L. Voltolini, R. Filippi, L. Luzzi, F. L. Pasini, B. Marchi, G. Gotti, and P. Rottoli, *Severe acute graft versus host disease after lung transplant: report of a case successfully treated with high dose corticosteroids.* 2009.

[59] A. Z. Dudek, H. Mahaseth, T. E. DeFor, and D. J. Weisdorf, "Bronchiolitis obliterans in chronic graft-versus-host disease: analysis of risk factors and treatment outcomes.," *Biology of blood and marrow transplantation journal of the American Society for Blood and Marrow Transplantation*, vol. 9, no. 10, pp. 657–666, 2003.

[60] M Estenne, “Bronchiolitis Obliterans after Human Lung Transplantation,” *American Journal of Respiratory and Critical Care Medicine*, vol. 166, no. 4, pp. 440–444, 2002.

[61] S. A. Yousem, G. J. Berry, P. T. Cagle, D. W. Chamberlain, A. N. Husain, R. H. Hruban, A Marchevsky, N. P. Ohori, J Ritter, and S Stewart, “Revision of the 1990 working formulation for the classification of pulmonary allograft rejection: Lung Rejection Study Group.,” *The Journal of Heart and Lung Transplantation*, vol. 15, pp. 1–15, 1996.

[62] P. F. Halloran, J Homik, N Goes, S. L. Lui, J Urmson, V Ramassar, and S. M. Cockfield, “The "injury response": a concept linking nonspecific injury, acute rejection, and long-term transplant outcomes.,” *Transplantation Proceedings*, vol. 29, no. 1-2, pp. 79–81, 1997.

[63] G. M. Verleden, R. Vos, S. I. De Vleeschauwer, A. Willems-Widyastuti, S. E. Verleden, L. J. Dupont, D. E. M. Van Raemdonck, and B. M. Vanaudenaerde, “Obliterative bronchiolitis following lung transplantation: from old to new concepts?” *Transplant international : official journal of the European Society for Organ Transplantation*, vol. 22, no. 8, pp. 771–9, Aug. 2009.

[64] T. J. Kroshus, V. R. Kshettry, K Savik, R John, M. I. Hertz, and R. M. Bolman, “Risk factors for the development of bronchiolitis obliterans syndrome after lung transplantation.,” *The Journal of Thoracic and Cardiovascular Surgery*, vol. 114, no. 2, pp. 195–202, 1997.

[65] J Gottlieb, F Mattner, H Weissbrodt, M Dierich, T Fuehner, M Strueber, A Simon, and T Welte, “Impact of graft colonization with gram-negative bacteria after lung transplantation on the development of bronchiolitis obliterans syndrome in recipients with cystic fibrosis.,” *Respiratory Medicine*, vol. 103, no. 5, pp. 743–749, 2009.

[66] P Grossi, C Farina, R Fiocchi, and D Dalla Gasperina, “Prevalence and outcome of invasive fungal infections in 1,963 thoracic organ transplant recipients: a multicenter retrospective study. Italian Study Group of Fungal Infections in Thoracic Organ Transplant Recipients.,” *Transplantation*, vol. 70, no. 1, pp. 112–116, 2000.

[67] B. C. Cahill, J. R. Hibbs, K Savik, B. A. Juni, B. M. Dosland, C Edin-Stibbe, and M. I. Hertz, “Aspergillus airway colonization and invasive disease after lung transplantation.,” *Chest*, vol. 112, no. 5, pp. 1160–1164, 1997.

[68] S. S. Weigt, R. M. Elashoff, C Huang, a Ardehali, a. L. Gregson, B Kubak, M. C. Fishbein, R Saggar, M. P. Keane, J. P. Lynch, D. a. Zisman, D. J. Ross, and J. a. Belperio, “Aspergillus colonization of the lung allograft is a risk factor for bronchiolitis obliterans syndrome.,” *American journal of transplantation : official journal of the American Society of Transplantation and the American Society of Transplant Surgeons*, vol. 9, no. 8, pp. 1903–11, Aug. 2009.

[69] L. D. Snyder, C. A. Finlen-Copeland, W. J. Turbyfill, D. Howell, D. A. Willner, and S. M. Palmer, "Cytomegalovirus Pneumonitis Is a Risk for Bronchiolitis Obliterans Syndrome in Lung Transplantation," *American Journal of Respiratory and Critical Care Medicine*, vol. 181, no. 12, pp. 1391–1396, 2010.

[70] M. Estenne, J. R. Maurer, A. Boehler, J. J. Egan, A. Frost, M. Hertz, G. B. Mallory, G. I. Snell, and S. Yousem, "Bronchiolitis obliterans syndrome 2001: an update of the diagnostic criteria.," *The Journal of heart and lung transplantation the official publication of the International Society for Heart Transplantation*, vol. 21, no. 3, pp. 297–310, 2002.

[71] M Estenne, A Van Muylem, C Knoop, and M Antoine, "Detection of obliterative bronchiolitis after lung transplantation by indexes of ventilation distribution.," *American Journal of Respiratory and Critical Care Medicine*, vol. 162, no. 3 Pt 1, pp. 1047–1051, 2000.

[72] R. R. Hachem, M. M. Chakinala, R. D. Yusen, J. P. Lynch, A. A. Aloush, G. A. Patterson, and E. P. Trulock, "The predictive value of bronchiolitis obliterans syndrome stage 0-p.," *American Journal of Respiratory and Critical Care Medicine*, vol. 169, no. 4, pp. 468–472, 2004.

[73] V. N. Lama, S. Murray, J. A. Mumford, K. R. Flaherty, A. Chang, G. B. Toews, M. Peters-Golden, and F. J. Martinez, "Prognostic value of bronchiolitis obliterans syndrome stage 0-p in single-lung transplant recipients.," *American Journal of Respiratory and Critical Care Medicine*, vol. 172, no. 3, pp. 379–383, 2005.

[74] A. Mishra and M. Verma, "Cancer Biomarkers: Are We Ready for the Prime Time?" *Cancers*, vol. 2, no. 1, pp. 190–208, Mar. 2010.

[75] S. M. Hanash, S. J. Pitteri, and V. M. Faca, "Mining the plasma proteome for cancer biomarkers.," *Nature*, vol. 452, no. 7187, pp. 571–9, Apr. 2008.

[76] D. M. Good, P. Zürbig, A. Argilés, H. W. Bauer, G. Behrens, J. J. Coon, M. Dakna, S. Decramer, C. Delles, A. F. Dominiczak, J. H. H. Ehrich, F. Eitner, D. Fliser, M. Frommberger, A. Ganser, M. a. Girolami, I. Golovko, W. Gwinner, M. Haubitz, S. Herget-Rosenthal, J. Jankowski, H. Jahn, G. Jerums, B. a. Julian, M. Kellmann, V. Kliem, W. Kolch, A. S. Krolewski, M. Luppi, Z. Massy, M. Melter, C. Neusüss, J. Novak, K. Peter, K. Rossing, H. Rupprecht, J. P. Schanstra, E. Schiffer, J.-U. Stolzenburg, L. Tarnow, D. Theodorescu, V. Thongboonkerd, R. Vanholder, E. M. Weissinger, H. Mischak, and P. Schmitt-Kopplin, "Naturally occurring human urinary peptides for use in diagnosis of chronic kidney disease.," *Molecular & cellular proteomics : MCP*, vol. 9, no. 11, pp. 2424–37, Nov. 2010.

[77] D. Soo-Quee Koh and G. Choon-Huat Koh, "The use of salivary biomarkers in occupational and environmental medicine.," *Occupational and environmental medicine*, vol. 64, no. 3, pp. 202–10, Mar. 2007.

[78] F. Vitzthum, F. Behrens, N. L. Anderson, and J. H. Shaw, "Proteomics: from basic research to diagnostic application. A review of requirements & needs.," *Journal of proteome research*, vol. 4, no. 4, pp. 1086–97, 2005.

[79] J. E. Barone, “Fever: Fact and fiction.,” *The Journal of Trauma*, vol. 67, no. 2, pp. 406–409, 2009.

[80] H. Mischak, R. Apweiler, R. E. Banks, M. Conaway, J. Coon, A. Dominiczak, J. H. H. Ehrich, D. Fliser, M. Girolami, H. Hermjakob, D. Hochstrasser, J. Jankowski, B. a. Julian, W. Kolch, Z. a. Massy, C. Neusuess, J. Novak, K. Peter, K. Rossing, J. Schanstra, O. J. Semmes, D. Theodorescu, V. Thongboonkerd, E. M. Weissinger, J. E. Van Eyk, and T. Yamamoto, “Clinical proteomics: A need to define the field and to begin to set adequate standards.,” *Proteomics. Clinical applications*, vol. 1, no. 2, pp. 148–56, Feb. 2007.

[81] D. Arnott and M. R. Emmert-Buck, “Proteomic profiling of cancer-opportunities, challenges, and context.,” *The Journal of pathology*, no. July, pp. 16–20, 2010.

[82] X. Chen, L. Wang, and H. Ishwaran, “An Integrative Pathway-based Clinical-genomic Model for Cancer Survival Prediction.,” *Statistics Probability Letters*, vol. 80, no. 17-18, pp. 1313–1319, 2010.

[83] J. M. Nicholls and G. D. Francis, “Anatomical pathology is dead? Long live anatomical pathology.,” *Pathology*, vol. 43, no. 6, pp. 635–641, 2011.

[84] a Schulze and J Downward, “Navigating gene expression using microarrays–a technology review.,” *Nature cell biology*, vol. 3, no. 8, E190–5, Aug. 2001.

[85] A. Introduction, “Editorial Tutorial on Microarray Gene Expression Experiments Gene Expression and Genomic Roots of Microarray Expression Profiling,” *Methods*, pp. 392–399, 2005.

[86] D Repsilber and A Ziegler, “Two-color Microarray Experiments Technology and Sources of Variance Preparing Samples for Array Platforms for the Two- color Array Approach,” *Methods*, pp. 400–404, 2005.

[87] C. Cheadle, M. P. Vawter, W. J. Freed, and K. G. Becker, “Analysis of microarray data using Z score transformation.,” *The Journal of molecular diagnostics : JMD*, vol. 5, no. 2, pp. 73–81, May 2003.

[88] W. Wu, E. P. Xing, C. Myers, I. S. Mian, and M. J. Bissell, “Evaluation of normalization methods for cDNA microarray data by k-NN classification.,” *BMC bioinformatics*, vol. 6, p. 191, Jan. 2005.

[89] A Poustka, “Variance Stabilization and Normalization using the vsn package in R,” *Bioinformatics*, vol. 1, 2002.

[90] M. Mcgee, Z. Chen, J. Cai, T. Ng, and B. Squires, “Improvements to the RMA Algorithm for Gene Expression Microarray Background Correction,” *Seminar*, 2005.

[91] C. Harbron, K.-M. Chang, and M. C. South, “RefPlus: an R package extending the RMA Algorithm.,” *Bioinformatics*, vol. 23, no. 18, pp. 2493–2494, 2007.

[92] Y. Rao, Y. Lee, D. Jarjoura, A. S. Ruppert, C.-G. Liu, J. C. Hsu, and J. P. Hagan, "A comparison of normalization techniques for microRNA microarray data.," *Statistical Applications in Genetics and Molecular Biology*, vol. 7, no. 1, Article22, 2008.

[93] X. Cui and G. a. Churchill, "Statistical tests for differential expression in cDNA microarray experiments.," *Genome biology*, vol. 4, no. 4, p. 210, Jan. 2003.

[94] R. Breitling, P. Armengaud, A. Amtmann, and P. Herzyk, "Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments.," *FEBS letters*, vol. 573, no. 1-3, pp. 83–92, Aug. 2004.

[95] R. Tibshirani, "A comparison of fold-change and the t-statistic for microarray data analysis," *Analysis*, 2007.

[96] D. W. Huang, B. T. Sherman, and R. a. Lempicki, "Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists.," *Nucleic acids research*, vol. 37, no. 1, pp. 1–13, Jan. 2009.

[97] B. J. Haas and M. C. Zody, "Advancing RNA-Seq analysis," *Nature Biotechnology*, VRST '08, vol. 28, no. 5, pp. 421–423, 2010.

[98] M. L. Wong and J. F. Medrano, "Real-time PCR for mRNA quantitation.," *Biotechniques*, vol. 39, no. 1, pp. 75–85, 2005.

[99] S. Zhang and J. Cao, "A close examination of double filtering with fold change and t test in microarray analysis," *BMC Bioinformatics*, vol. 10, no. 1, p. 402, 2009.

[100] D. Landsittel and N. Donohue-Babiak, "Effect of adding fold-change criteria to significance testing of microarray data," *Journal of Statistical Computation and Simulation*, vol. 80, no. 1, pp. 89–97, 2010.

[101] M. R. Dalman, A. Deeter, G. Nimishakavi, and Z.-H. Duan, "Fold change and p-value cutoffs significantly alter microarray interpretations," *BMC Bioinformatics*, vol. 13, no. 2, S11, 2012.

[102] W. H. Hudson, "Statistical Significance Tests," *Social Service Review*, vol. 59, no. 2, pp. 322–322, Jun. 1985.

[103] J. S. Morey, J. C. Ryan, and F. M. Van Dolah, "Microarray validation: factors influencing correlation between oligonucleotide microarrays and real-time PCR," *Biological procedures online*, vol. 8, no. 1, pp. 175–193, 2006.

[104] F. Hong, R. Breitling, C. W. McEntee, B. S. Wittner, J. L. Nemhauser, and J. Chory, "RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis.," *Bioinformatics (Oxford, England)*, vol. 22, no. 22, pp. 2825–7, Nov. 2006.

[105] F. Hong and R. Breitling, "A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments.," *Bioinformatics (Oxford, England)*, vol. 24, no. 3, pp. 374–82, Feb. 2008.

[106] F Bretz, J Landgrebe, and E Brunner, “Multiplicity issues in microarray experiments.,” *Methods of information in medicine*, vol. 44, no. 3, pp. 431–7, Jan. 2005.

[107] X. Gao, “Multiple testing corrections for imputed SNPs.,” *Genetic epidemiology*, vol. 35, no. 3, pp. 154–8, Apr. 2011.

[108] B Brors, “Microarray annotation and biological information on function.,” *Methods of information in medicine*, vol. 44, no. 3, pp. 468–72, Jan. 2005.

[109] S. Y. Rhee, V. Wood, K. Dolinski, and S. Draghici, “Use and misuse of the gene ontology annotations.,” *Nature reviews. Genetics*, vol. 9, no. 7, pp. 509–15, Jul. 2008.

[110] J Wixon and D Kell, “The Kyoto encyclopedia of genes and genomes–KEGG.,” *Yeast (Chichester, England)*, vol. 17, no. 1, pp. 48–55, Apr. 2000.

[111] M Streit, M Kalkusch, K Kashofer, and D Schmalstieg, “Navigation and Exploration of Interconnected Pathways,” *Symposium A Quarterly Journal In Modern Foreign Literatures*, vol. 27, no. 3, A. Vilanova, A. Telea, G. Scheuermann, and T. Möller, Eds., pp. 951–958, 2008.

[112] D. W. Huang, B. T. Sherman, and R. a. Lempicki, “Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.,” *Nature protocols*, vol. 4, no. 1, pp. 44–57, 2009.

[113] P. S. Patel, S. D. Telang, R. M. Rawal, and M. H. Shah, “MINI-REVIEW A Review of Proteomics in Cancer Research,” *Cancer Research*, vol. 6, pp. 113–117, 2005.

[114] K. Kienzl-Wagner, J. Pratschke, and G. Brandacher, “Proteomics-a blessing or a curse? Application of proteomics technology to transplant medicine.,” *Transplantation*, vol. 92, no. 5, pp. 499–509, 2011.

[115] R Bakhtiar and F. L. Tse, “Biological mass spectrometry: a primer.,” *Mutagenesis*, vol. 15, no. 5, pp. 415–30, Sep. 2000.

[116] A. El-Aneed, A. Cohen, and J. Banoub, “Mass Spectrometry, Review of the Basics: Electrospray, MALDI, and Commonly Used Mass Analyzers,” *Applied Spectroscopy Reviews*, vol. 44, no. 3, pp. 210–230, Apr. 2009.

[117] J. Albrethsen, “The first decade of MALDI protein profiling: a lesson in translational biomarker research.,” *Journal of proteomics*, vol. 74, no. 6, pp. 765–73, May 2011.

[118] a Mellmann, F Bimet, C Bizet, a. D. Borovskaya, R. R. Drake, U Eigner, a. M. Fahr, Y He, E. N. Ilina, M Kostrzewa, T Maier, L Mancinelli, W Moussaoui, G Prévost, L Putignani, C. L. Seachord, Y. W. Tang, and D Harmsen, “High interlaboratory reproducibility of matrix-assisted laser desorption ionization-time of flight mass spectrometry-based species identification of nonfermenting bacteria.,” *Journal of clinical microbiology*, vol. 47, no. 11, pp. 3732–4, Nov. 2009.

[119] J. Albrethsen, “Reproducibility in protein profiling by MALDI-TOF mass spectrometry.,” *Clinical chemistry*, vol. 53, no. 5, pp. 852–8, May 2007.

[120] K. a. Baggerly, J. S. Morris, J. Wang, D. Gold, L.-C. Xiao, and K. R. Coombes, "A comprehensive approach to the analysis of matrix-assisted laser desorption/ionization-time of flight proteomics spectra from serum samples.," *Proteomics*, vol. 3, no. 9, pp. 1667–72, Sep. 2003.

[121] G. Ge and G. W. Wong, "Classification of premalignant pancreatic cancer mass-spectrometry data using decision tree ensembles.," *BMC bioinformatics*, vol. 9, p. 275, 2008.

[122] A. Ozcift and A. Gulten, "Assessing effects of pre-processing mass spectrometry data on classification performance.," *European journal of mass spectrometry Chichester England*, vol. 14, no. 5, pp. 267–273, 2008.

[123] K. Coombes, K. Baggerly, and J. Morris, "Pre-processing mass spectrometry data," *Fundamentals of Data Mining in Genomics and Proteomics*, no. January, pp. 79–102, 2007.

[124] P. Bewerunge, "Integrative Data Mining and Meta Analysis of Disease-Specific Large-Scale Genomic,Transcriptomic and Proteomic Data," Dissertation, Universität Heidelberg, 2009.

[125] A. C. Sauve and T. P. Speed, "Normalization, baseline correction and alignment of high-throughput mass spectrometry data," *Genomic Signal Processing and Statistics 2004 GENSiPS 2004 IEEE International Workshop on*, 2004.

[126] A. Cruz-Marcelo, R. Guerra, M. Vannucci, Y. Li, C. C. Lau, and T.-K. Man, "Comparison of algorithms for pre-processing of SELDI-TOF mass spectrometry data.," *Bioinformatics (Oxford, England)*, vol. 24, no. 19, pp. 2129–36, Oct. 2008.

[127] R. Tang, J. P. Sinnwell, J. Li, D. N. Rider, M. D. Andrade, and J. M. Biernacka, "Arthritis using random forests," *Methods*, vol. 5, no. Vi, pp. 1–5, 2009.

[128] C. Yang, Z. He, and W. Yu, "Comparison of public peak detection algorithms for MALDI mass spectrometry data analysis.," *BMC bioinformatics*, vol. 10, p. 4, Jan. 2009.

[129] X Li, R Gentleman, X Lu, Q Shi, J. Iglehart, L Harris, and A Miron, "SELDI-TOF mass spectrometry Protein Data," in *Bioinformatics and Computational Biology Solutions Using R and Bioconductor Springer*, 2005, pp. 91–109.

[130] J. Zhang, E. Gonzalez, T. Hestilow, W. Haskins, and Y. Huang, "Review of Peak Detection Algorithms in Liquid-Chromatography-Mass Spectrometry," *Current Genomics*, vol. 10, no. 6, pp. 388–401, 2009.

[131] M Belghazi, K Bathany, C Hountondji, X Grandier-Vazeille, S Manon, and J. M. Schmitter, "Identification and validation of a potential lung cancer serum biomarker detected by matrix-assisted laser desorption/ionization-time of flight spectra analysis.," *Proteomics*, vol. 3, no. 9, pp. 946–954, 2003.

[132] M. W. Duncan and S. W. Hunsucker, "Proteomics as a tool for clinically relevant biomarker discovery and validation.," *Experimental biology and medicine Maywood NJ*, vol. 230, no. 11, pp. 808–817, 2005.

[133] J. D. Wulfkuhle, L. A. Liotta, and E. F. Petricoin, “Proteomic applications for the early detection of cancer,” *Nature Reviews Cancer*, vol. 3, no. 4, pp. 267–275, 2003.

[134] D Skillicorn, *Understanding Complex Datasets.* Chapman & Hall/CRC, 2007, ch. 3. Singula.

[135] A. K. Jain, M. N. Murty, and P. J. Flynn, “Data Clustering: A Review,” *ACM Computing Surveys*, vol. 31, no. 3, 2000.

[136] J. A. Hartigan and M. A. Wong, “A K-Means Clustering Algorithm,” *Applied Statistics*, C, vol. 28, no. 1, pp. 100–108, 1979.

[137] W. Shannon, R. Culverhouse, and J. Duncan, “Analyzing microarray data using cluster analysis.,” *Pharmacogenomics*, vol. 4, no. 1, pp. 41–52, Jan. 2003.

[138] S. C. Madeira and A. L. Oliveira, “Biclustering algorithms for biological data analysis: a survey.,” *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM*, vol. 1, no. 1, pp. 24–45, 2004.

[139] M. C. P. de Souto, I. G. Costa, D. S. a. de Araujo, T. B. Ludermir, and A. Schliep, “Clustering cancer gene expression data: a comparative study.,” *BMC bioinformatics*, vol. 9, p. 497, Jan. 2008.

[140] A. Thalamuthu, I. Mukhopadhyay, X. Zheng, and G. C. Tseng, “Evaluation and comparison of gene clustering methods in microarray analysis.,” *Bioinformatics (Oxford, England)*, vol. 22, no. 19, pp. 2405–12, Oct. 2006.

[141] A.-C. Doux, J.-P. Laurent, and J.-P. Nadal, “Symbolic Data Analysis With the K-Means Algorithm for User Profiling,” *User Modeling Proceedings of the Sixth International Conference UM97*, A. Jameson, C. Paris, and C. Tasso, Eds., pp. 359–361, 1997.

[142] M. Welling, “Kernel K-means and Spectral Clustering,” *ReCALL*, no. 5, pp. 5–7, 2004.

[143] S. Zhong and B. Raton, “Efficient online spherical k-means clustering,” *Proceedings 2005 IEEE International Joint Conference on Neural Networks 2005*, vol. 5, pp. 3180–3185, 2005.

[144] S. B. Kotsiantis, “Supervised Machine Learning : A Review of Classification Techniques,” *Informatica*, vol. 31, pp. 249–268, 2007.

[145] R. Spang and F. Markowetz, “Microarray Analysis Classification by SVM and PAM,” *Molecular Biology*, pp. 1–9, 2002.

[146] X. Chen, M. Wang, and H. Zhang, “The use of classification trees for bioinformatics,” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 1, no. 1, pp. 55–63, 2011.

[147] L Breiman, “Statistical modeling: the two cultures (with discussion),” *Statistical Science*, vol. 16, pp. 199–231, 2001.

[148] J. S. Borges, A. R. S. Marçal, and J. F. P. Costa, “Comparison of three supervised classification methods for deriving land cover maps from ASTER satellite images,” *Learning*, 2004.

[149] T. Sing, O. Sander, N. Beerenwinkel, and T. Lengauer, “ROCR: visualizing classifier performance in R.,” *Bioinformatics (Oxford, England)*, vol. 21, no. 20, pp. 3940–1, Oct. 2005.

[150] Y. Xiao, J. Hua, and E. R. Dougherty, “Feature selection increases cross-validation imprecision,” *Genomic Signal Processing*, pp. 17–18, 2006.

[151] R.-H. Li and G. G. Belford, “Instability of decision tree classification algorithms,” in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining KDD 02*, ACM Press, 2002, p. 570.

[152] P. Yang, Y. H. Yang, B. B. Zhou, and A. Y. Zomaya, “A Review of Ensemble Methods in Bioinformatics,” *Current Bioinformatics*, pp. 1–18, 2010.

[153] Y. Freund, R. E. Schapire, and P. Avenue, “A Short Introduction to Boosting,” *Society*, vol. 14, no. 5, pp. 771–780, 1999.

[154] L. E. O. Breiman, “Bagging Predictors,” *Machine Learning*, vol. 140, pp. 123–140, 1996.

[155] ——, “Random Forests,” *Machine Learning*, pp. 5–32, 2001.

[156] R. Olshen, “A Conversaton with Leo Breiman,” *Statistical Science*, vol. 16, no. 2, pp. 184–198, 2001.

[157] A. Cutler, “Remembering Leo Breiman,” *The Annals of Applied Statistics*, vol. 4, no. 4, pp. 1621–1633, Dec. 2010.

[158] Y Xie, X Li, E Ngai, and W Ying, “Customer churn prediction using improved balanced random forests,” *Expert Systems with Applications*, vol. 36, no. 3, pp. 5445–5449, 2009.

[159] M. Khalilia, S. Chakraborty, and M. Popescu, “Predicting disease risks from highly imbalanced data using random forest,” *BMC Medical Informatics and Decision Making*, vol. 11, no. 1, p. 51, 2011.

[160] G. Rios and H. Zha, “Exploring Support Vector Machines and Random Forests for Spam Detection,” in *Comparative and General Pharmacology*, D. Heckerman, T. Berson, J. Goodman, and A. Ng, Eds., vol. 129, 2004, p. 2865.

[161] R. Caruana and A. Niculescu-Mizil, “An empirical comparison of supervised learning algorithms,” *Proceedings of the 23rd international conference on Machine learning ICML 06*, ICML ’06, vol. C, no. 1, pp. 161–168, 2006.

[162] S. Li, E. J. Harner, and D. A. Adjeroh, “Random KNN feature selection – a fast and stable alternative to Random Forests,” *BMC Bioinformatics*, vol. 12, no. 1, p. 450, 2011.

[163] H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer, “Random survival forests,” *The Annals of Applied Statistics*, vol. 2, no. 3, pp. 841–860, Sep. 2008.

[164] T. G. Clark, M. J. Bradburn, S. B. Love, and D. G. Altman, "Survival analysis part I: basic concepts and first analyses.," *British journal of cancer*, vol. 89, no. 2, pp. 232–8, Jul. 2003.

[165] M. J. Bradburn, T. G. Clark, S. B. Love, and D. G. Altman, "Survival analysis part II: multivariate data analysis–an introduction to concepts and methods.," *British journal of cancer*, vol. 89, no. 3, pp. 431–6, Aug. 2003.

[166] T Hothorn, "On the exact distribution of maximally selected rank statistics," *Computational Statistics & Data Analysis*, vol. 43, pp. 121 –137, Feb. 2003.

[167] E. L. Kaplan and P Meier, "Nonparametric estimation from incomplete observations," *Journal of the American Statistical Association*, vol. 53, no. 282, pp. 457–481, 1958.

[168] J. A. Koziol and Z. Jia, "The concordance index C and the Mann-Whitney parameter Pr(X>Y) with randomly censored data.," *Biometrical journal Biometrische Zeitschrift*, vol. 51, no. 3, pp. 467–474, 2009.

[169] H. Abdi, "The Kendall Rank Correlation Coefficient," *Cognition*, Encyclopedia of Measurement and Statistics, vol. 11, no. 3, N. Salkind, Ed., pp. 1–7, 1955.

[170] P. D'haeseleer, "How does gene expression clustering work?" *Nature biotechnology*, vol. 23, no. 12, pp. 1499–501, Dec. 2005.

[171] M. Smolkin and D. Ghosh, "Cluster stability scores for microarray data in cancer studies.," *BMC bioinformatics*, vol. 4, p. 36, Sep. 2003.

[172] D. Jiang, C. Tang, and A. Zhang, "Cluster analysis for gene expression data: a survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, no. 11, pp. 1370–1386, Nov. 2004.

[173] G. W. Smith and G. J. M. Rosa, "Interpretation of microarray data: trudging out of the abyss towards elucidation of biological significance.," *Journal of animal science*, vol. 85, no. 13 Suppl, E20–3, Mar. 2007.

[174] T. Shi and S. Horvath, "Unsupervised Learning With Random Forest Predictors," *Journal of Computational and Graphical Statistics*, vol. 15, no. 1, pp. 118–138, Mar. 2006.

[175] J. A. Cuesta-Albertos, A Gordaliza, and C Matran, "Trimmed K-Means: An Attempt to Robustify Quantizers," *Annals of Statistics*, vol. 25, no. 553-576, pp. 553–576, 1997.

[176] J. A. Hartigan, "Direct Clustering of a Data Matrix," *Journal of the American Statistical Association*, vol. 67, no. 337, pp. 123–129, 1972.

[177] M. Sill, S. Kaiser, A. Benner, and A. Kopp-Schneider, "Robust biclustering by sparse singular value decomposition incorporating stability selection.," *Bioinformatics (Oxford, England)*, vol. 27, no. 15, pp. 2089–2097, Jun. 2011.

[178] A. Bhattacharya and R. K. De, "Bi-correlation clustering algorithm for determining a set of co-regulated genes.," *Bioinformatics (Oxford, England)*, vol. 25, no. 21, pp. 2795–801, Nov. 2009.

[179] A. Prelić, S. Bleuler, P. Zimmermann, A. Wille, P. Bühlmann, W. Gruissem, L. Hennig, L. Thiele, and E. Zitzler, "A systematic comparison and evaluation of biclustering methods for gene expression data.," *Bioinformatics (Oxford, England)*, vol. 22, no. 9, pp. 1122–9, May 2006.

[180] R. Santamaría, R. Therón, and L. Quintales, "A visual analytics approach for understanding biclustering results from microarray data.," *BMC bioinformatics*, vol. 9, p. 247, Jan. 2008.

[181] E. H. Klech, C Hutter, D. O. Parma, M. P. Warsaw, W Pohl, and G. R. Milano, "Clinical guidelines and indications for bronchoalveolar lavage (BAL): Report of the European Society of Pneumology Task Group on BAL.," *The European respiratory journal official journal of the European Society for Clinical Respiratory Physiology*, vol. 3, no. 8, pp. 937–976, 1990.

[182] F. J. T. Staal, G Cario, G Cazzaniga, T Haferlach, M Heuser, W.-K. Hofmann, K Mills, M Schrappe, M Stanulla, L. U. Wingen, J. J. M. van Dongen, and B Schlegelberger, "Consensus guidelines for microarray gene expression analyses in leukemia from three European leukemia networks.," *Leukemia : official journal of the Leukemia Society of America, Leukemia Research Fund, U.K*, vol. 20, no. 8, pp. 1385–92, Aug. 2006.

[183] B. Skawran, D. Steinemann, A. Weigmann, P. Flemming, T. Becker, J. Flik, H. Kreipe, B. Schlegelberger, and L. Wilkens, "Gene expression profiling in hepatocellular carcinoma: upregulation of genes in amplified chromosome regions.," *Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc*, vol. 21, no. 5, pp. 505–16, May 2008.

[184] B. Skawran, D. Steinemann, T. Becker, R. Buurman, J. Flik, B. Wiese, P. Flemming, H. Kreipe, B. Schlegelberger, and L. Wilkens, "Loss of 13q is associated with genes involved in cell cycle and proliferation in dedifferentiated hepatocellular carcinoma.," *Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc*, vol. 21, no. 12, pp. 1479–89, Dec. 2008.

[185] R. Shyamsundar, Y. H. Kim, J. P. Higgins, K. Montgomery, M. Jorden, A. Sethuraman, M. Van De Rijn, D. Botstein, P. O. Brown, and J. R. Pollack, "A DNA microarray survey of gene expression in normal human tissues," *Genome Biology*, vol. 6, no. 3, R22, 2005.

[186] S. Gaseitsiwe, D. Valentini, S. Mahdavifar, I. Magalhaes, D. F. Hoft, J. Zerweck, M. Schutkowski, J. Andersson, M. Reilly, and M. J. Maeurer, "Pattern Recognition in Pulmonary Tuberculosis Defined by High Content Peptide Microarray Chip Analysis Representing 61 Proteins from M. tuberculosis," *PLoS ONE*, vol. 3, no. 12, D. Unutmaz, Ed., p. 8, 2008.

[187] D. Sohal, A. Yeatts, K. Ye, A. Pellagatti, L. Zhou, P. Pahanish, Y. Mo, T. Bhagat, J. Mariadason, J. Boultwood, A. Melnick, J. Greally, and A. Verma, "Meta-Analysis of Microarray Studies Reveals a Novel Hematopoietic Progenitor

Cell Signature and Demonstrates Feasibility of Inter-Platform Data Integration," *PLoS ONE*, vol. 3, no. 8, M. Rattray, Ed., p. 10, 2008.

[188] P. J. Park, Y. A. Cao, S. Y. Lee, J.-W. Kim, M. S. Chang, R. Hart, and S. Choi, "Current issues for DNA microarrays: platform comparison, double linear amplification, and universal RNA reference.," *Journal of Biotechnology*, vol. 112, no. 3, pp. 225–245, 2004.

[189] A. V. Kapp, S. S. Jeffrey, A. Langerø d, A.-L. Bø rresen Dale, W. Han, D.-Y. Noh, I. R. Bukholm, M. Nicolau, P. O. Brown, and R. Tibshirani, "Discovery and validation of breast cancer subtypes," *BMC Genomics*, vol. 7, no. 2000, p. 231, 2006.

[190] T. Teitz, J. J. Stanke, S. Federico, C. L. Bradley, R. Brennan, J. Zhang, M. D. Johnson, J. Sedlacik, M. Inoue, Z. M. Zhang, S. Frase, J. E. Rehg, C. M. Hillenbrand, D. Finkelstein, C. Calabrese, M. A. Dyer, and J. M. Lahti, "Preclinical Models for Neuroblastoma: Establishing a Baseline for Treatment," *PLoS ONE*, vol. 6, no. 4, P. Dent, Ed., p. 16, 2011.

[191] D. W. Huang, B. T. Sherman, Q. Tan, J. R. Collins, W. G. Alvord, J. Roayaei, R. Stephens, M. W. Baseler, H. C. Lane, and R. A. Lempicki, "The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists," *Genome Biology*, vol. 8, no. 9, R183, 2007.

[192] D. W. Huang, B. T. Sherman, Q. Tan, J. Kir, D. Liu, D. Bryant, Y. Guo, R. Stephens, M. W. Baseler, H. C. Lane, and R. A. Lempicki, "DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists," *Nucleic Acids Research*, vol. 35, no. Web Server issue, W169–W175, 2007.

[193] T. Næ s and B. r.-H. Mevik, "The flexibility of fuzzy clustering illustrated by examples," *Journal of Chemometrics*, vol. 13, no. 3-4, pp. 435–444, 1999.

[194] H. Ishwaran and U. B. Kogalur, "Random Survival Forests for R," *October*, vol. 7, no. October, pp. 25–31, 2007.

[195] H. Ishwaran, "Variable importance in binary regression trees and forests," *Electronic Journal of Statistics*, vol. 1, pp. 519–537, 2007.

[196] H. Ishwaran, U. B. Kogalur, X. Chen, and A. J. Minn, "Random Survival Forests for High-Dimensional Data," *Analysis*, 2011.

[197] C. Strobl, J. Malley, and G. Tutz, "An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests.," *Psychological Methods*, vol. 14, no. 4, pp. 323–348, 2009.

[198] H. J. Cordell, "Detecting gene-gene interactions that underlie human diseases.," *Nature Reviews Genetics*, vol. 10, no. 6, pp. 392–404, 2009.

[199] X. Wang, R. C. Elston, and X. Zhu, “The meaning of interaction.,” *Human Heredity*, vol. 70, no. 4, pp. 269–277, 2010.

[200] X. Chen and H. Ishwaran, “Random forests for genomic data analysis.,” *Genomics*, Apr. 2012.

[201] K. J. Berry and P. W. Mielke, “A Monte Carlo investigation of the Fisher Z transformation for normal and nonnormal distributions.,” *Psychological Reports*, vol. 87, no. 3 Pt 2, pp. 1101–1114, 2000.

[202] C. Cheadle, Y. S. Cho-Chung, K. G. Becker, and M. P. Vawter, “Application of z-score transformation to Affymetrix data.,” *Applied Bioinformatics*, vol. 2, no. 4, pp. 209–217, 2003.

[203] C. F. Bond and K. Richardson, “Seeing the fisher z-transformation,” *Psychometrika*, vol. 69, no. 2, pp. 291–303, 2004.

[204] G. a. Pavlopoulos, M. Secrier, C. N. Moschopoulos, T. G. Soldatos, S. Kossida, J. Aerts, R. Schneider, and P. G. Bagos, “Using graph theory to analyze biological networks.,” *BioData mining*, vol. 4, no. 1, p. 10, Jan. 2011.

[205] W. Huber, V. J. Carey, L. Long, S. Falcon, and R. Gentleman, “Graphs in molecular biology,” *BMC Bioinformatics*, vol. 8, no. Suppl 6, S8, 2007.

[206] G. Csardi and T. Nepusz, “The igraph software package for complex network research,” *InterJournal Complex Systems*, vol. Complex Sy, no. 1695, p. 1695, 2006.

[207] L. K. Gallos, C. Song, S. Havlin, and H. A. Makse, “Scaling theory of transport in complex biological networks.,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 19, pp. 7746–7751, 2007.

[208] M. Mares, “The saga of minimum spanning trees,” *Computer Science Review*, vol. 2, no. 3, pp. 165–221, 2008.

[209] R. C. Prim, “Shortest connection networks and some generalizations,” *Bell System Technical Journal*, vol. 36, no. 6, pp. 1389–1401, 1957.

[210] A Barrat, M Barthélemy, R Pastor-Satorras, and A Vespignani, “The architecture of complex weighted networks,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 11, pp. 3747–3752, 2003.

[211] E. Almaas, “Biological impacts and context of network theory.,” *Journal of Experimental Biology*, vol. 210, no. Pt 9, pp. 1548–1558, 2007.

[212] L. Guo, E. K. Lobenhofer, C. Wang, R. Shippy, S. C. Harris, L. Zhang, N. Mei, T. Chen, D. Herman, F. M. Goodsaid, P. Hurban, K. L. Phillips, J. Xu, X. Deng, Y. A. Sun, W. Tong, Y. P. Dragan, and L. Shi, “Rat toxicogenomic study reveals analytical consistency across microarray platforms.,” *Nature Biotechnology*, vol. 24, no. 9, pp. 1162–1169, 2006.

[213] A. R. Brady, D. Harrison, S. Black, S. Jones, K. Rowan, G. Pearson, J. Ratcliffe, and G. J. Parry, "Assessment and optimization of mortality prediction tools for admissions to pediatric intensive care in the United kingdom.," *Pediatrics*, vol. 117, no. 4, e733–e742, 2006.

[214] M. Frigge, D. C. Hoaglin, and B. Iglewicz, "Some Implementations of the Boxplot," *American Statistician*, vol. 43, no. 1, p. 50, 1989.

[215] H. Wickham and L. Stryjewski, "40 years of boxplots," *Construction*, pp. 1–17, 2011.

[216] M. Hilario, A. Kalousis, C. Pellegrini, and M. Müller, "Processing and classification of protein mass spectra.," *Mass Spectrometry Reviews*, vol. 25, no. 3, pp. 409–449, 2006.

[217] S. D. Nathan, S. D. Barnett, J Wohlrab, and N Burton, "Bronchiolitis obliterans syndrome: utility of the new guidelines in single lung transplant recipients," *J Heart Lung Transplant*, vol. 22, no. 4, pp. 427–432, 2003.

[218] S. Ahmad, O. a. Shlobin, and S. D. Nathan, "Pulmonary complications of lung transplantation.," *Chest*, vol. 139, no. 2, pp. 402–11, Feb. 2011.

[219] D. S. Siroky, "Navigating Random Forests and related advances in algorithmic modeling," *Statistics Surveys*, vol. 3, pp. 147–163, 2009.

[220] A. Liaw and M. Wiener, "Classification and Regression by randomForest," *Glass*, vol. 2, no. December, pp. 18–22, 2002.

[221] C. Strobl, A.-L. Boulesteix, A. Zeileis, and T. Hothorn, "Bias in random forest variable importance measures: Illustrations, sources and a solution," *BMC Bioinformatics*, vol. 8, no. 1, p. 25, 2007.

[222] M. L. Calle, V. Urrea, A.-L. Boulesteix, and N. Malats, "AUC-RF: a new strategy for genomic profiling with random forest.," *Human Heredity*, vol. 72, no. 2, pp. 121–32, 2011.

[223] M.-C. Le Bihan, Y. Hou, N. Harris, E. Tarelli, and G. R. Coulton, "Proteomic analysis of fast and slow muscles from normal and kyphoscoliotic mice using protein arrays, 2-DE and MS.," *Proteomics*, vol. 6, no. 16, pp. 4646–61, Aug. 2006.

[224] I. S. Dhillon, "Co-clustering documents and words using bipartite spectral graph partitioning," *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '01*, no. April 2006, pp. 269–274, 2001.

[225] H Zha, X He, C Ding, M Gu, and H Simon, "Bipartite graph partitioning and data clustering," *Proceedings of the tenth international conference on Information and knowledge management CIKM01*, CIKM '01, vol. pages, no. 2, p. 25, 2001.

[226] M. Wieling and J. Nerbonne, "Bipartite spectral graph partitioning for clustering dialect varieties and detecting their linguistic features," *Computer Speech Language*, vol. 25, no. 3, pp. 700–715, 2011.

[227] A. J. Hackstadt and A. M. Hess, "Filtering for increased power for microarray data analysis," *BMC Bioinformatics*, vol. 10, no. 1, p. 11, 2009.

[228] E. E. Kuramae, V. Robert, C. Echavarri-Erasun, and T. Boekhout, "Cophenetic correlation analysis as a strategy to select phylogenetically informative proteins: an example from the fungal kingdom.," *BMC evolutionary biology*, vol. 7, p. 134, Jan. 2007.

[229] C Hennig, "Cluster-wise assessment of cluster stability," *Computational Statistics & Data Analysis*, vol. 52, no. 1, pp. 258–271, Sep. 2007.

[230] L. A. Garcia-Escudero and A. Gordaliza, "Robustness Properties of k Means and Trimmed k Means," *Journal of the American Statistical Association*, vol. 94, no. 447, p. 956, 1999.

[231] S. S. Dave, K. Fu, G. W. Wright, L. T. Lam, P. Kluin, E.-J. Boerma, T. C. Greiner, D. D. Weisenburger, A. Rosenwald, G. Ott, H.-K. Müller-Hermelink, R. D. Gascoyne, J. Delabie, L. M. Rimsza, R. M. Braziel, T. M. Grogan, E. Campo, E. S. Jaffe, B. J. Dave, W. Sanger, M. Bast, J. M. Vose, J. O. Armitage, J. M. Connors, E. B. Smeland, S. Kvaloy, H. Holte, R. I. Fisher, T. P. Miller, E. Montserrat, W. H. Wilson, M. Bahl, H. Zhao, L. Yang, J. Powell, R. Simon, W. C. Chan, and L. M. Staudt, "Molecular diagnosis of Burkitt's lymphoma.," *The New England Journal of Medicine*, vol. 354, no. 23, pp. 2431–2442, 2006.

[232] H. Redestig, D. Repsilber, F. Sohler, and J. Selbig, "Integrating functional knowledge during sample clustering for microarray data using unsupervised decision trees.," *Biometrical journal Biometrische Zeitschrift*, vol. 49, no. 2, pp. 214–229, 2007.

[233] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888–905, 2000.

[234] U. Luxburg, "A tutorial on spectral clustering," *Statistics and Computing*, vol. 17, no. 4, pp. 395–416, Aug. 2007.

[235] T Xiang and S Gong, "Spectral clustering with eigenvector selection," *Pattern Recognition*, vol. 41, no. 3, pp. 1012–1029, Mar. 2008.

[236] L. Khan and F. Luo, "Hierarchical clustering for complex data," *Artificial Intelligence*, vol. 14, no. 5, pp. 1–19, 2005.

[237] S Seal, S Komarina, and S Aluru, "An optimal hierarchical clustering algorithm for gene expression data," *Information Processing Letters*, vol. 93, no. 3, pp. 143–147, 2005.

[238] Y. Zhang, M. Wroblewski, M. I. Hertz, C. H. Wendt, T. M. Cervenka, and G. L. Nelsestuen, "Analysis of chronic lung transplant rejection by MALDI-TOF profiles of bronchoalveolar lavage fluid.," *Proteomics*, vol. 6, no. 3, pp. 1001–10, Feb. 2006.

[239] R. A. Cohen and N. Carolina, “An Introduction to PROC LOESS for Local Regression,” *North*, vol. 273, 1999.

[240] M. W. Trosset, B. Hendrickson, C.-k. Li, and R. Mathias, “Extensions of classical multidimensional scaling via variable reduction,” *Computational Statistics*, vol. 17, 2002.

[241] I. Borg and P. J. F. Groenen, *Modern Multidimensional Scaling*, Springer, Ed., ser. Springer Series in Statistics 445. Springer, 2005, vol. 94, p. 614.

[242] M Trosset and C Priebe, “The out-of-sample problem for classical multidimensional scaling,” *Computational Statistics & Data Analysis*, vol. 52, no. 10, pp. 4635–4642, 2008.

[243] D. S. Kukafka, G. M. O’Brien, S Furukawa, and G. J. Criner, “Surveillance bronchoscopy in lung transplant recipients.,” *Chest*, vol. 111, no. 2, pp. 377–381, 1997.

[244] C. a. K. Borrebaeck and C. Wingren, “Transferring proteomic discoveries into clinical practice.,” *Expert review of proteomics*, vol. 6, no. 1, pp. 11–3, Feb. 2009.

[245] S. P. Gygi, Y. Rochon, B. R. Franza, and R. Aebersold, “Correlation between protein and mRNA abundance in yeast.,” *Molecular and Cellular Biology*, vol. 19, no. 3, pp. 1720–1730, 1999.

[246] G. Chen, T. G. Gharib, C.-C. Huang, J. M. G. Taylor, D. E. Misek, S. L. R. Kardia, T. J. Giordano, M. D. Iannettoni, M. B. Orringer, S. M. Hanash, and D. G. Beer, “Discordant protein and mRNA expression in lung adenocarcinomas.,” *Molecular cellular proteomics MCP*, vol. 1, no. 4, pp. 304–313, 2002.

[247] Y. Guo, P. Xiao, S. Lei, F. Deng, G. G. Xiao, Y. Liu, X. Chen, L. Li, S. Wu, Y. Chen, H. Jiang, L. Tan, J. Xie, X. Zhu, S. Liang, and H. Deng, “How is mRNA expression predictive for protein expression? A correlation study on human circulating monocytes.,” *Acta Biochimica et Biophysica Sinica*, vol. 40, no. 5, pp. 426–436, 2008.

[248] G. L. Nelsestuen, M. B. Martinez, M. I. Hertz, K. Savik, and C. H. Wendt, “Proteomic identification of human neutrophil alpha-defensins in chronic lung allograft rejection.,” *Proteomics*, vol. 5, no. 6, pp. 1705–13, Apr. 2005.

[249] B Godeau, C Cormier, and C. J. Menkes, “Bronchiolitis obliterans in systemic lupus erythematosus: beneficial effect of intravenous cyclophosphamide.,” *Annals of the rheumatic diseases*, vol. 50, no. 12, pp. 956–8, Dec. 1991.

[250] K El-Reshaid, J. P. Madda, and T. D. Chugh, “Bronchiolitis obliterans organizing pneumonia associated with systemic lupus erythematosus.,” *Annals of Saudi Medicine*, vol. 16, no. 3, pp. 332–334, 1996.

[251] H. Takada, Y. Saito, A. Nomura, S. Ohga, K. Kuwano, N. Nakashima, S. Aishima, N. Tsuru, and T. Hara, “Bronchiolitis obliterans organizing pneumonia as an initial manifestation in systemic lupus erythematosus.,” *Pediatric Pulmonology*, vol. 40, no. 3, pp. 257–260, 2005.

[252] C Chadwick, S. M. Marven, and A. J. Vora, "Autologous blood pleurodesis for pneumothorax complicating graft-versus-host disease-related bronchiolitis obliterans.," *Bone Marrow Transplantation*, vol. 33, no. 4, pp. 451–453, 2004.

[253] R. R. Hachem, "Lung allograft rejection: diagnosis and management.," *Current Opinion in Organ Transplantation*, vol. 14, no. 5, pp. 477–482, 2009.

[254] L. Liu, M. Zeng, and J. S. Stamler, "Hemoglobin induction in mouse macrophages.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 96, no. 12, pp. 6643–6647, 1999.

[255] D. A. Newton, K. M. K. Rao, R. A. Dluhy, and J. E. Baatz, "Hemoglobin is expressed by alveolar epithelial cells.," *The Journal of biological chemistry*, vol. 281, no. 9, pp. 5668–5676, 2006.

[256] M. Bhaskaran, H. Chen, Z. Chen, and L. Liu, "Hemoglobin is expressed in alveolar epithelial type II cells.," *The Journal of Biological Chemistry*, vol. 281, no. 4, pp. 5668–5676, 2006.

[257] N. Ishikawa, S. Ohlmeier, K. Salmenkivi, M. Myllärniemi, I. Rahman, W. Mazur, and V. L. Kinnula, "Hemoglobin α and β are ubiquitous in the human lung, decline in idiopathic pulmonary fibrosis but not in COPD.," *Respiratory research*, vol. 11, p. 123, 2010.

[258] C. L. Grek, D. A. Newton, D. D. Spyropoulos, and J. E. Baatz, "Hypoxia upregulates expression of hemoglobin in alveolar epithelial cells.," *American Journal of Respiratory Cell and Molecular Biology*, vol. 44, no. 4, pp. 439–447, 2011.

[259] D. W. Chang, S. Hayashi, S. A. Gharib, T. Vaisar, S. T. King, M. Tsuchiya, J. T. Ruzinski, D. R. Park, G. Matute-Bello, M. M. Wurfel, R. Bumgarner, J. W. Heinecke, and T. R. Martin, "Proteomic and computational analysis of bronchoalveolar proteins during the course of the acute respiratory distress syndrome.," *American journal of respiratory and critical care medicine*, vol. 178, no. 7, pp. 701–9, Oct. 2008.

[260] V. Baudin-Creuza, C. Vasseur-Godbillon, C. Pato, C. Préhu, H. Wajcman, and M. C. Marden, "Transfer of human alpha- to beta-hemoglobin via its chaperone protein: evidence for a new state.," *The Journal of Biological Chemistry*, vol. 279, no. 35, pp. 36 530–36 533, 2004.

[261] Y. Kong, S. Zhou, A. J. Kihm, A. M. Katein, X. Yu, D. A. Gell, J. P. Mackay, K. Adachi, L. Foster-Brown, C. S. Louden, A. J. Gow, and M. J. Weiss, "Role of alpha-hemoglobin-stabilizing protein in normal erythropoiesis and beta-thalassemia.," *Journal of Clinical Investigation*, vol. 114, no. 10, pp. 1457–1466, 2004.

[262] M. J. Weiss and C. O. Dos Santos, "Chaperoning erythropoiesis," *Blood*, vol. 113, no. 10, pp. 2136–2144, 2009.

[263] M. E. Favero and F. F. Costa, "Alpha-Hemoglobin-Stabilizing Protein: An Erythroid Molecular Chaperone," *Biochemistry research international*, vol. 2011, p. 373 859, 2011.

[264] A. Ben-Tal, “Simplified models for gas exchange in the human lungs.,” *Journal of Theoretical Biology*, vol. 238, no. 2, pp. 474–495, 2006.

[265] W. Liu, S. S. Baker, R. D. Baker, N. J. Nowak, and L. Zhu, “Upregulation of hemoglobin expression by oxidative stress in hepatocytes and its implication in nonalcoholic steatohepatitis.,” *PLoS ONE*, vol. 6, no. 9, e24363, 2011.

[266] C Mastruzzo, N Crimi, and C Vancheri, “Role of oxidative stress in pulmonary fibrosis.,” *Monaldi archives for chest disease Archivio Monaldi per le malattie del torace Fondazione clinica del lavoro IRCCS and Istituto di clinica tisiologica e malattie apparato respiratorio Universita di Napoli Secondo ateneo*, vol. 57, no. 1, pp. 65–76, 1996.

[267] V. L. Kinnula, C. L. Fattman, R. J. Tan, and T. D. Oury, “Oxidative Stress in Pulmonary Fibrosis,” *American Journal of Respiratory and Critical Care Medicine*, vol. 172, no. 4, pp. 417–422, 2005.

[268] J Behr, K Maier, B Braun, M Schwaiblmair, and C Vogelmeier, “Evidence for oxidative stress in bronchiolitis obliterans syndrome after lung and heart-lung transplantation. The Munich Lung Transplant Group.,” *Transplantation*, vol. 69, no. 9, pp. 1856–1860, 2000.

[269] J. Madill, E. Aghdassi, B. Arendt, B. Hartman-Craven, C. Gutierrez, C.-W. Chow, and J. Allard, “Lung transplantation: does oxidative stress contribute to the development of bronchiolitis obliterans syndrome?” *Transplantation reviews Orlando Fla*, vol. 23, no. 2, pp. 103–110, 2009.

[270] J Mallol, V Aguirre, and V Espinosa, “Increased oxidative stress in children with post infectious Bronchiolitis Obliterans.,” *Allergologia Et Immunopathologia*, no. xx, 2011.

[271] B. R. Stripp and S. D. Reynolds, “Clara Cells,” in *Encyclopedia of Respiratory Medicine*, G. J. Laurent and S. D. Shapiro, Eds., Elsevier, 2006, pp. 471–478.

[272] A Bernard, F. X. Marchandise, S Depelchin, R Lauwerys, and Y Sibille, “Clara cell protein in serum and bronchoalveolar lavage.,” *The European respiratory journal official journal of the European Society for Clinical Respiratory Physiology*, vol. 5, no. 10, pp. 1231–1238, 1992.

[273] C Hermans, M Petrek, V Kolek, B Weynand, T Pieters, M Lambert, and A Bernard, “Serum Clara cell protein (CC16), a marker of the integrity of the air-blood barrier in sarcoidosis.,” *The European respiratory journal official journal of the European Society for Clinical Respiratory Physiology*, vol. 18, no. 3, pp. 507–514, 2001.

[274] G Singh and S. L. Katyal, “Clara cells and Clara cell 10 kD protein (CC10).,” *American Journal of Respiratory Cell and Molecular Biology*, vol. 17, no. 2, pp. 141–143, 1997.

[275] M. S. López De Haro, L Alvarez, and A Nieto, "Evidence for the identity of antiproteinase pulmonary protein CCSP and uteroglobin.," *FEBS Letters*, vol. 232, no. 2, pp. 351–353, 1988.

[276] Q. Ye, M. Fujita, H. Ouchi, I. Inoshima, T. Maeyama, K. Kuwano, Y. Horiuchi, N. Hara, and Y. Nakanishi, "Serum CC-10 in inflammatory lung diseases.," *Respiration international review of thoracic diseases*, vol. 71, no. 5, pp. 505–510, 2004.

[277] K. Sarafidis, T. Stathopoulou, E. Diamanti, V. Soubasi, C. Agakidis, A. Balaska, and V. Drossou, "Clara cell secretory protein (CC16) as a peripheral blood biomarker of lung injury in ventilated preterm neonates.," *European Journal of Pediatrics*, vol. 167, no. 11, pp. 1297–1303, 2008.

[278] M. Nord, K. Schubert, T. N. Cassel, O. Andersson, and G. C. Riise, "Decreased serum and bronchoalveolar lavage levels of Clara cell secretory protein (CC16) is associated with bronchiolitis obliterans syndrome and airway neutrophilia in lung transplant recipients.," *Transplantation*, vol. 73, no. 8, pp. 1264–9, Apr. 2002.

[279] A. W. M. Paantjens, H. G. Otten, W. G. J. Van Ginkel, D. A. Van Kessel, J. M. M. Van Den Bosch, J. M. Kwakkel-Van Erp, and E. A. Van De Graaf, "Clara cell secretory protein and surfactant protein-D do not predict bronchiolitis obliterans syndrome after lung transplantation.," *Transplantation*, vol. 90, no. 3, pp. 340–342, 2010.

[280] G Singh, S. L. Katyal, W. E. Brown, D. L. Collins, and R. J. Mason, "Pulmonary lysozyme–a secretory protein of type II pneumocytes in the rat.," *The American review of respiratory disease*, vol. 138, no. 5, pp. 1261–1267, 1988.

[281] E. M. Prager and P Jollès, "Animal lysozymes c and g: an overview.," *Exs*, vol. 75, pp. 9–31, 1996.

[282] S. D. Reynolds, P. R. Reynolds, G. S. Pryhuber, J. D. Finder, and B. R. Stripp, "Secretoglobins SCGB3A1 and SCGB3A2 define secretory cell subsets in mouse and human airways.," *American Journal of Respiratory and Critical Care Medicine*, vol. 166, no. 11, pp. 1498–1509, 2002.

[283] R. Kurotani, S. Okumura, T. Matsubara, U. Yokoyama, J. R. Buckley, T. Tomita, K. Kezuka, T. Nagano, D. Esposito, T. E. Taylor, W. K. Gillette, Y. Ishikawa, H. Abe, J. M. Ward, and S. Kimura, "Secretoglobin 3A2 suppresses bleomycin-induced pulmonary fibrosis by transforming growth factor beta signaling down-regulation.," *The Journal of Biological Chemistry*, vol. 286, no. 22, pp. 19 682–19 692, 2011.

[284] A. M. Pastva, J. R. Wright, and K. L. Williams, "Immunomodulatory Roles of Surfactant Proteins A and D," *Proceedings of the American Thoracic Society*, vol. 4, no. 3, pp. 252–257, 2007.

[285] H. P. Haagsman, "Interactions of surfactant protein A with pathogens.," *Biochimica et Biophysica Acta*, vol. 1408, no. 2-3, pp. 264–77, 1998.

[286] F. Meloni, R. Salvini, A. M. Bardoni, I. Passadore, N. Solari, P. Vitulo, T. Oggionni, M. Viganò, E. Pozzi, and A. M. Fietta, "Bronchoalveolar lavage fluid proteome in bronchiolitis obliterans syndrome: possible role for surfactant protein A in disease onset.," *The Journal of heart and lung transplantation the official publication of the International Society for Heart Transplantation*, vol. 26, no. 11, pp. 1135–1143, 2007.

[287] G. S. Pryhuber, "Regulation and function of pulmonary surfactant protein B.," *Molecular Genetics and Metabolism*, vol. 64, no. 4, pp. 217–228, 1998.

[288] T. Curstedt, "Surfactant protein C: basics to bedside.," *Journal of perinatology official journal of the California Perinatal Association*, vol. 25 Suppl 2, S36–S38; discussion S39, 2005.

[289] E. C. Crouch, "Surfactant protein-D and pulmonary host defense," *Respiratory Research*, vol. 1, no. 2, pp. 93–108, 2000.

[290] F. X. McCormack and J. A. Whitsett, "The pulmonary collectins, SP-A and SP-D, orchestrate innate immunity in the lung," *Journal of Clinical Investigation*, vol. 109, no. 6, pp. 707–712, 2002.

[291] G. L. Sorensen, S. Husby, and U. Holmskov, "Surfactant protein A and surfactant protein D variation in pulmonary disease.," *Immunobiology*, vol. 212, no. 4-5, pp. 381–416, 2007.

[292] T. Tecle, S. Tripathi, and K. L. Hartshorn, "Review: Defensins and cathelicidins in lung immunity.," *Innate immunity*, vol. 16, no. 3, pp. 151–159, 2010.

[293] L. T. Spencer, G. Paone, P. M. Krein, F. N. Rouhani, J. Rivera-Nieves, and M. L. Brantly, "Role of human neutrophil peptides in lung inflammation associated with alpha1-antitrypsin deficiency.," *American journal of physiology. Lung cellular and molecular physiology*, vol. 286, no. 3, pp. L514–20, Mar. 2004.

[294] T Shestakova, E Zhuravel, L Bolgova, O Alekseenko, M Soldatkina, and P Pogrebnoy, "Expression of human beta-defensins-1, 2 and 4 mRNA in human lung tumor tissue: a pilot study.," *Experimental oncology*, vol. 30, no. 2, pp. 153–6, Jun. 2008.

[295] P. S. Hiemstra, S Van Wetering, and J Stolk, "Neutrophil serine proteinases and defensins in chronic obstructive pulmonary disease: effects on pulmonary epithelium.," *The European respiratory journal official journal of the European Society for Clinical Respiratory Physiology*, vol. 12, no. 5, pp. 1200–1208, 1998.

[296] H Mukae, H Iiboshi, M Nakazato, T Hiratsuka, M Tokojima, K Abe, J Ashitani, J Kadota, S Matsukura, and S Kohno, "Raised plasma concentrations of α-defensins in patients with idiopathic pulmonary fibrosis," *Thorax*, vol. 57, no. 7, pp. 623–628, 2002.

[297] K. Bdeir, A. A.-R. Higazi, I. Kulikovskaya, M. Christofidou-Solomidou, S. A. Vinogradov, T. C. Allen, S. Idell, R. Linzmeier, T. Ganz, and D. B. Cines, "Neutrophil alpha-defensins cause lung injury by disrupting the capillary-epithelial barrier.," *American Journal of Respiratory and Critical Care Medicine*, vol. 181, no. 9, pp. 935–946, 2010.

[298] K. Tsushima, L. S. King, N. R. Aggarwal, A. De Gorordo, F. R. D'Alessio, and K. Kubo, "Acute lung injury review.," *Internal medicine Tokyo Japan*, vol. 48, no. 9, pp. 621–630, 2009.

[299] D. J. Ross, A. M. Cole, D. Yoshioka, A. K. Park, J. A. Belperio, H. Laks, R. M. Strieter, J. P. Lynch, B. Kubak, A. Ardehali, and T. Ganz, "Increased bronchoalveolar lavage human beta-defensin type 2 in bronchiolitis obliterans syndrome after lung transplantation.," *Transplantation*, vol. 78, no. 8, pp. 1222–1224, 2004.

[300] M Calderon-Santiago and M. D. De Castro, "The dual trend in histatins research," *TRACTRENDS IN ANALYTICAL CHEMISTRY*, vol. 28, no. 8, pp. 1011–1018, 2009.

[301] M. Ahmad, M. Piludu, F. G. Oppenheim, E. J. Helmerhorst, and A. R. Hand, "Immunocytochemical localization of histatins in human salivary glands.," *The journal of histochemistry and cytochemistry official journal of the Histochemistry Society*, vol. 52, no. 3, pp. 361–370, 2004.

[302] M. Castagnola, R. Inzitari, D. V. Rossetti, C. Olmi, T. Cabras, V. Piras, P. Nicolussi, M. T. Sanna, M. Pellegrini, B. Giardina, and I. Messana, "A cascade of 24 histatins (histatin 3 fragments) in human saliva. Suggestions for a pre-secretory sequential cleavage pathway.," *The Journal of Biological Chemistry*, vol. 279, no. 40, pp. 41 436–41 443, 2004.

[303] a. S. Verkman, Y. Song, and J. R. Thiagarajah, "Role of airway surface liquid and submucosal glands in cystic fibrosis lung disease.," *American journal of physiology. Cell physiology*, vol. 284, no. 1, pp. C2–15, Jan. 2003.

[304] M. J. Oudhoff, J. G. M. Bolscher, K. Nazmi, H. Kalay, W. Van T Hof, A. V. N. Amerongen, and E. C. I. Veerman, "Histatins are the major wound-closure stimulating factors in human saliva as identified in a cell culture assay.," *The FASEB journal official publication of the Federation of American Societies for Experimental Biology*, vol. 22, no. 11, pp. 3805–12, 2008.

[305] M. J. Oudhoff, P. A. M. Van Den Keijbus, K. L. Kroeze, K Nazmi, S Gibbs, J. G. M. Bolscher, and E. C. I. Veerman, "Histatins enhance wound closure with oral and non-oral cells.," *Journal of Dental Research*, vol. 88, no. 9, pp. 846–850, 2009.

[306] P. Botha, L. Archer, R. L. Anderson, J. Lordan, J. H. Dark, P. A. Corris, K. Gould, and A. J. Fisher, "Pseudomonas aeruginosa colonization of the allograft after lung transplantation and the risk of bronchiolitis obliterans syndrome.," *Transplantation*, vol. 85, no. 5, pp. 771–774, 2008.

[307] K. F. Remund, M. Best, and J. J. Egan, “Infections relevant to lung transplantation.,” *Proceedings of the American Thoracic Society*, vol. 6, no. 1, pp. 94–100, 2009.

[308] J. C. VanderSpek, H. E. Wyandt, J. C. Skare, A Milunsky, F. G. Oppenheim, and R. F. Troxler, “Localization of the genes for histatins to human chromosome 4q13 and tissue distribution of the mRNAs.,” *The American Journal of Human Genetics*, vol. 45, no. 3, pp. 381–387, 1989.

[309] I Salama, P. S. Malone, F Mihaimeed, and J. L. Jones, “A review of the S100 proteins in cancer.,” *European journal of surgical oncology the journal of the European Society of Surgical Oncology and the British Association of Surgical Oncology*, vol. 34, no. 4, pp. 357–364, 2008.

[310] M. M. Wilkinson, A Busuttil, C Hayward, D. J. Brock, J. R. Dorin, and V Van Heyningen, “Expression pattern of two related cystic fibrosis-associated calcium-binding proteins in normal and abnormal tissues.,” *Journal of Cell Science*, vol. 91 (Pt 2), pp. 221–230, 1988.

[311] E. Lorenz, M. S. Muhlebach, P. A. Tessier, N. E. Alexis, R Duncan Hite, M. C. Seeds, D. B. Peden, and W. Meredith, “Different expression ratio of S100A8/A9 and S100A12 in acute and chronic lung diseases.,” *Respiratory Medicine*, vol. 102, no. 5, pp. 567–573, 2007.

[312] E. Bargagli, C. Olivieri, M. Cintorino, R. M. Refini, N. Bianchi, A. Prasse, and P. Rottoli, “Calgranulin B (S100A9/MRP14): a key molecule in idiopathic pulmonary fibrosis?” *Inflammation*, vol. 34, no. 2, pp. 85–91, 2011.

[313] X. Yang, N. C. Popescu, and D. B. Zimonjic, “DLC1 interaction with S100A10 mediates inhibition of in vitro cell invasion and tumorigenicity of lung cancer cells through a RhoGAP-independent mechanism.,” *Cancer Research*, vol. 71, no. 8, pp. 2916–2925, 2011.

[314] Y. Tutar, “Dimerization and ion binding properties of S100P protein.,” *Protein and Peptide Letters*, vol. 13, no. 3, pp. 301–306, 2006.

[315] Y. Ito, K. Arai, R. Nozawa, H. Yoshida, T. Higashiyama, Y. Takamura, A. Miya, K. Kobayashi, K. Kuma, and A. Miyauchi, “S100A10 expression in thyroid neoplasms originating from the follicular epithelium: contribution to the aggressive characteristic of anaplastic carcinoma.,” *Anticancer Research*, vol. 27, no. 4C, pp. 2679–2683, 2007.

[316] H Inada, M Naka, T Tanaka, G. E. Davey, and C. W. Heizmann, “Human S100A11 exhibits differential steady-state RNA levels in various tissues and a distinct subcellular localization.,” *Biochemical and Biophysical Research Communications*, vol. 263, no. 1, pp. 135–138, 1999.

[317] J. Zhao, I. Endoh, K. Hsu, N. Tedla, Y. Endoh, and C. L. Geczy, “S100A8 modulates mast cell function and suppresses eosinophil migration in acute asthma.,” *Antioxidants redox signaling*, vol. 14, no. 9, pp. 1589–1600, 2011.

[318] T. Vogl, A. L. Gharibyan, and L. A. Morozova-Roche, "Pro-Inflammatory S100A8 and S100A9 Proteins: Self-Assembly into Multifunctional Native and Amyloid Complexes.," *International Journal of Molecular Sciences*, vol. 13, no. 3, pp. 2893–917, 2012.

[319] J. Roth, T. Vogl, C. Sorg, and C. Sunderkötter, "Phagocyte-specific S100 proteins: a novel group of proinflammatory molecules.," *Trends in Immunology*, vol. 24, no. 4, pp. 155–158, 2003.

[320] D. Foell, M. Frosch, C. Sorg, and J. Roth, "Phagocyte-specific calcium-binding S100 proteins as clinical laboratory markers of inflammation.," *Clinica chimica acta international journal of clinical chemistry*, vol. 344, no. 1-2, pp. 37–51, 2004.

[321] M.-A. Raquil, N. Anceriz, P. Rouleau, and P. A. Tessier, "Blockade of antimicrobial proteins S100A8 and S100A9 inhibits phagocyte migration to the alveoli in streptococcal pneumonia.," *The Journal of Immunology*, vol. 180, no. 5, pp. 3366–3374, 2008.

[322] J Edgeworth, P Freemont, and N Hogg, "Ionomycin-regulated phosphorylation of the myeloid calcium-binding protein p14.," *Nature*, vol. 342, no. 6246, pp. 189–192, 1989.

[323] M. C. Dessing, L. M. Butter, G. J. Teske, N. Claessen, C. M. Van Der Loos, T. Vogl, J. Roth, T. Van Der Poll, S. Florquin, and J. C. Leemans, "S100A8/A9 Is Not Involved in Host Defense against Murine Urinary Tract Infection," *PLoS ONE*, vol. 5, no. 10, J. H. Fritz, Ed., p. 7, 2010.

[324] D. Foell, H. Wittkowski, T. Vogl, and J. Roth, "S100 proteins expressed in phagocytes: a novel group of damage-associated molecular pattern molecules.," *Journal of Leukocyte Biology*, vol. 81, no. 1, pp. 28–37, 2007.

[325] H. Y. Sroussi, G. A. Köhler, N. Agabian, D. Villines, and J. M. Palefsky, "Substitution of methionine 63 or 83 in S100A9 and cysteine 42 in S100A8 abrogate the antifungal activities of S100A8/A9: potential role for oxidative regulation.," *FEMS Immunology and Medical Microbiology*, vol. 55, no. 1, pp. 55–61, 2009.

[326] J. B. Vos, M. A. Van Sterkenburg, K. F. Rabe, J. Schalkwijk, P. S. Hiemstra, and N. A. Datson, "Transcriptional response of bronchial epithelial cells to Pseudomonas aeruginosa: identification of early mediators of host defense.," *Physiological Genomics*, vol. 21, no. 3, pp. 324–336, 2005.

[327] T. Mauad, A. Van Schadewijk, J. Schrumpf, C. E. Hack, S. Fernezlian, A. L. Garippo, B. Ejzenberg, P. S. Hiemstra, K. F. Rabe, and M. Dolhnikoff, "Lymphocytic inflammation in childhood bronchiolitis obliterans.," *Pediatric Pulmonology*, vol. 38, no. 3, pp. 233–239, 2004.

[328] C Chang, L. C. Hsu, V Davé, and A Yoshida, "Expression of human aldehyde dehydrogenase-3 associated with hepatocellular carcinoma: promoter regions and nuclear protein factors related to the expression.," *International Journal of Molecular Medicine*, vol. 2, no. 3, pp. 333–338, 1998.

[329] F. Jiang, Q. Qiu, A. Khanna, N. W. Todd, J. Deepak, L. Xing, H. Wang, Z. Liu, Y. Su, S. A. Stass, and R. L. Katz, "Aldehyde dehydrogenase 1 is a tumor stem cell-associated marker in lung cancer.," *Molecular cancer research MCR*, vol. 7, no. 3, pp. 330–8, 2009.

[330] P. Fuchs, C. Loeseken, J. K. Schubert, and W. Miekisch, "Breath gas aldehydes as biomarkers of lung cancer.," *International journal of cancer Journal international du cancer*, vol. 126, no. 11, pp. 2663–2670, 2010.

[331] G Muzio, M Maggiora, E Paiuzzi, M Oraldi, and R. A. Canuto, "Aldehyde dehydrogenases and cell proliferation.," *Free Radical Biology & Medicine*, vol. 52, no. 4, pp. 735–46, 2011.

[332] G. A. Rossi, O Sacco, B Balbi, S Oddera, T Mattioni, G Corte, C Ravazzoni, and L Allegra, "Human ciliated bronchial epithelial cells: expression of the HLA-DR antigens and of the HLA-DR alpha gene, modulation of the HLA-DR antigens by gamma-interferon and antigen-presenting function in the mixed leukocyte reaction.," *American Journal of Respiratory Cell and Molecular Biology*, vol. 3, no. 5, pp. 431–439, 1990.

[333] O Sacco, S Lantero, L Scarso, V Frangova, V Ottolini, and G. A. Rossi, "The Increased Expression of HLA-DR and ICAM-1 Molecules by Human Bronchial Epithelial Cells, Induced by Activated Mononuclear Cells, is Downregulated by Nedocromil Sodium," *Mediators of Inflammation*, vol. 3, no. 7, S7–S13, 1994.

[334] T Suda, A Sato, W Sugiura, and K Chida, "Induction of MHC class II antigens on rat bronchial epithelial cells by interferon-gamma and its effect on antigen presentation.," *Lung*, vol. 173, no. 2, pp. 127–137, 1995.

[335] P. M. Taylor, M. L. Rose, and M. H. Yacoub, "Expression of MHC antigens in normal human lungs and transplanted lungs with obliterative bronchiolitis.," *Transplantation*, vol. 48, no. 3, pp. 506–510, 1989.

[336] D. S. Milne, A Gascoigne, J Wilkes, L Sviland, T Ashcroft, A. D. Pearson, A. J. Malcolm, and P Corris, *The immunohistopathology of obliterative bronchiolitis following lung transplantation.* 1992.

[337] D. S. Milne, A. D. Gascoigne, J Wilkes, L Sviland, T Ashcroft, A. J. Malcolm, and P. A. Corris, "MHC class II and ICAM-1 expression and lymphocyte subsets in transbronchial biopsies from lung transplant recipients.," *Transplantation*, vol. 57, no. 12, pp. 1762–1766, 1994.

[338] L. D. Sharples, K. McNeil, S. Stewart, and J. Wallwork, "Risk factors for bronchiolitis obliterans: a systematic review of recent publications.," *The Journal of heart and lung transplantation the official publication of the International Society for Heart Transplantation*, vol. 21, no. 2, pp. 271–281, 2002.

[339] T Quillard, K Croce, F. A. Jaffer, R Weissleder, and P Libby, "Molecular imaging of macrophage protease activity in cardiovascular inflammation in vivo.," *Thrombosis and haemostasis*, vol. 105, no. 5, pp. 828–836, 2011.

[340] H. R. P. Miller and A. D. Pemberton, "Tissue-specific expression of mast cell granule serine proteinases and their role in inflammation in the lung and gut," *Immunology*, vol. 105, no. 4, pp. 375–390, 2002.

[341] L. B. Schwartz, D. D. Metcalfe, J. S. Miller, H Earl, and T Sullivan, "Tryptase levels as an indicator of mast-cell activation in systemic anaphylaxis and mastocytosis.," *The New England Journal of Medicine*, vol. 316, no. 26, pp. 1622–1626, 1987.

[342] S Peyrol, J. F. Cordier, and J. A. Grimaud, "Intra-alveolar fibrosis of idiopathic bronchiolitis obliterans-organizing pneumonia. Cell-matrix patterns.," *The American journal of pathology*, vol. 137, no. 1, pp. 155–170, 1990.

[343] S. J. Ruoss, T Hartmann, and G. H. Caughey, "Mast cell tryptase is a mitogen for cultured fibroblasts.," *Journal of Clinical Investigation*, vol. 88, no. 2, pp. 493–499, 1991.

[344] A Pesci, M Majori, M. L. Piccoli, A Casalini, A Curti, D Franchini, and M Gabrielli, "Mast cells in bronchiolitis obliterans organizing pneumonia. Mast cell hyperplasia and evidence for extracellular release of tryptase.," *Chest*, vol. 110, no. 2, pp. 383–391, 1996.

[345] O Kawanami, V. J. Ferrans, J. D. Fulmer, and R. G. Crystal, "Ultrastructure of pulmonary mast cells in patients with fibrotic lung disorders.," *Laboratory investigation a journal of technical methods and pathology*, vol. 40, no. 6, pp. 717–734, 1979.

[346] A Pesci, G Bertorelli, M Gabrielli, and D Olivieri, "Mast cells in fibrotic lung disorders.," *Chest*, vol. 103, no. 4, pp. 989–996, 1993.

[347] S. J. Galli, J. R. Gordon, and B. K. Wershil, "Mast cell cytokines in allergy and inflammation.," *Agents and actions Supplements*, vol. 43, pp. 209–220, 1993.

[348] M. C. Mihm, W. H. Clark, R. J. Reed, and M. G. Caruso, "Mast cell infiltrates of the skin and the mastocytosis syndrome.," *Human Pathology*, vol. 4, no. 2, pp. 231–239, 1973.

[349] F Levi-Schaffer and E Rubinchik, "Mast cell/fibroblast interactions.," *Clinical and experimental allergy journal of the British Society for Allergy and Clinical Immunology*, vol. 24, no. 11, pp. 1016–1021, 1994.

[350] M. Artuc, U. M. Steckelings, and B. M. Henz, "Mast cell-fibroblast interactions: human mast cells as source and inducers of fibroblast and epithelial growth factors.," *The Journal of investigative dermatology*, vol. 118, no. 3, pp. 391–395, 2002.

[351] I. S. Fernández, L. Ständker, W.-G. Forssmann, G. Giménez-Gallego, and A. Romero, "Crystallization and preliminary crystallographic studies of human kallikrein 7, a serine protease of the multigene kallikrein family.," *Acta Crystallographica Section F Structural Biology And Crystallization Communications*, vol. 63, no. Pt 8, pp. 669–672, 2007.

[352] L. Mo, J. Zhang, J. Shi, Q. Xuan, X. Yang, M. Qin, C. Lee, H. Klocker, Q. Q. Li, and Z. Mo, “Human kallikrein 7 induces epithelial-mesenchymal transition-like changes in prostate carcinoma cells: a role in prostate cancer invasion and progression.,” *Anticancer Research*, vol. 30, no. 9, pp. 3413–3420, 2010.

[353] V. C. Ramani and R. S. Haun, “The extracellular matrix protein fibronectin is a substrate for kallikrein 7.,” *Biochemical and Biophysical Research Communications*, vol. 369, no. 4, pp. 1169–1173, 2008.

[354] B Chen, V Polunovsky, J White, B Blazar, R Nakhleh, J Jessurun, M Peterson, and P Bitterman, “Mesenchymal cells isolated after acute lung injury manifest an enhanced proliferative phenotype.,” *Journal of Clinical Investigation*, vol. 90, no. 5, pp. 1778–1785, 1992.

[355] L. A. Borthwick, S. M. Parker, K. A. Brougham, G. E. Johnson, M. R. Gorowiec, C Ward, J. L. Lordan, P. A. Corris, J. A. Kirby, and A. J. Fisher, “Epithelial to mesenchymal transition (EMT) and airway remodelling after human lung transplantation.,” *Thorax*, vol. 64, no. 9, pp. 770–777, 2009.

[356] J. Câmara and G. Jarai, “Epithelial-mesenchymal transition in primary human bronchial epithelial cells is Smad-dependent and enhanced by fibronectin and TNF,” *Fibrogenesis tissue repair*, vol. 3, no. 1, p. 2, 2010.

[357] K Araki, T Shimura, H Suzuki, S Tsutsumi, W Wada, T Yajima, T Kobayahi, N Kubo, and H Kuwano, “E/N-cadherin switch mediates cancer progression via TGF-β-induced epithelial-to-mesenchymal transition in extrahepatic cholangiocarcinoma.,” *British Journal of Cancer*, vol. 105, no. 12, pp. 1885–1893, 2011.

[358] M. J. D. Griffiths, D. Bonnet, and S. M. Janes, “Stem cells of the alveolar epithelium.,” *Lancet*, vol. 366, no. 9481, pp. 249–60, 2005.

[359] B. Mahesh, H.-S. Leong, K. S. Nair, A. McCormack, P. Sarathchandra, and M. L. Rose, “Autoimmunity to vimentin potentiates graft vasculopathy in murine cardiac allografts.,” *Transplantation*, vol. 90, no. 1, pp. 4–13, 2010.

[360] M. G. Mendez, S.-I. Kojima, and R. D. Goldman, “Vimentin induces changes in cell shape, motility, and adhesion during the epithelial to mesenchymal transition.,” *The FASEB journal : official publication of the Federation of American Societies for Experimental Biology*, vol. 24, no. 6, pp. 1838–51, Jun. 2010.

[361] I Santamaría, G Velasco, a. M. Pendás, a Fueyo, and C López-Otín, “Cathepsin Z, a novel human cysteine proteinase with a short propeptide domain and a unique chromosomal location.,” *The Journal of biological chemistry*, vol. 273, no. 27, pp. 16 816–23, Jul. 1998.

[362] J. Kos, A. Sekirnik, A. Premzl, V. Zavasnik Bergant, T. Langerholc, B. Turk, B. Werle, R. Golouh, U. Repnik, M. Jeras, and V. Turk, “Carboxypeptidases cathepsins X and B display distinct protein profile in human cells and tissues.,” *Experimental Cell Research*, vol. 306, no. 1, pp. 103–113, 2005.

[363] Z. Jevnikar, N. Obermajer, M. Bogyo, and J. Kos, "The role of cathepsin X in the migration and invasiveness of T lymphocytes.," *Journal of Cell Science*, vol. 121, no. Pt 16, pp. 2652–2661, 2008.

[364] J. Kos, Z. Jevnikar, and N. Obermajer, "The role of cathepsin X in cell signaling.," *Cell adhesion migration*, vol. 3, no. 2, pp. 164–166, 2009.

[365] N. Obermajer, A. Premzl, T. Zavasnik Bergant, B. Turk, and J. Kos, "Carboxypeptidase cathepsin X mediates beta2-integrin-dependent adhesion of differentiated U-937 cells.," *Experimental Cell Research*, vol. 312, no. 13, pp. 2515–2527, 2006.

[366] T Springer, G Galfré, D. S. Secher, and C Milstein, "Mac-1: a macrophage differentiation antigen identified by monoclonal antibody.," *European Journal of Immunology*, vol. 9, no. 4, pp. 301–306, 1979.

[367] D. C. Altieri, F. R. Agbanyo, J Plescia, M. H. Ginsberg, T. S. Edgington, and E. F. Plow, "A unique recognition site mediates the interaction of fibrinogen with the leukocyte integrin Mac-1 (CD11b/CD18).," *The Journal of Biological Chemistry*, vol. 265, no. 21, pp. 12 119–12 122, 1990.

[368] T. P. Ugarova and V. P. Yakubenko, "Recognition of fibrinogen by leukocyte integrins.," *Annals Of The New York Academy Of Sciences*, vol. 936, pp. 368–385, 2001.

[369] S Culine, N Honoré, A Tavitian, and B Olofsson, "Overexpression of the ras-related rab2 gene product in peripheral blood mononuclear cells from patients with hematological and solid neoplasms.," *Cancer Research*, vol. 52, no. 11, pp. 3083–3088, 1992.

[370] S Culine, N Honoré, V Closson, P Lang, J Bertoglio, A Tavitian, and B Olofsson, "A possible role for the Ras-related Rab2 protein in the immunological events associated with hematological malignancies.," *Nouvelle revue francaise dhematologie*, vol. 35, no. 1, pp. 41–44, 1993.

[371] D. Sukhtankar, A. Okun, A. Chandramouli, M. A. Nelson, T. W. Vanderah, A. E. Cress, F. Porreca, and T. King, "Inhibition of p38-MAPK Signaling Pathway Attenuates Breast Cancer Induced Bone Pain and Disease Progression in a Murine Model of Cancer-induced Bone Pain.," *Molecular Pain*, vol. 7, no. 1, p. 81, 2011.

[372] C. Jose, N. Bellance, and R. Rossignol, "Choosing between glycolysis and oxidative phosphorylation: a tumor's dilemma?" *Biochimica et Biophysica Acta*, vol. 1807, no. 6, pp. 552–561, 2011.

Chapter 5

Appendix

Table 1. Genes significantly up-regulated in BOS-positive samples

Gene symbol	Entrez ID	UniGene	RP	P	FDR	Mean FC	Median FC
SRGN	5552	Hs.1908	322	<0.001	<0.001	2.05	1.77
UPK1B	7348	Hs.271580	394	<0.001	<0.001	2.17	1.3
C15orf48	84419	Hs.112242	446	<0.001	<0.001	2.1	1.89
LAPTM5	7805	Hs.371021	468	<0.001	<0.001	1.87	2.84
CCL3L3	414062	Hs.512304	478	<0.001	<0.001	1.88	1.91
FPGS	2356	Hs.335084	484	<0.001	<0.001	1.81	1.66
CCL3	6348	Hs.514107	622	<0.001	<0.001	1.71	1.4
CCR1	1230	Hs.301921	707	<0.001	<0.001	1.59	1.23
TPSAB1	7177	Hs.405479	738	<0.001	<0.001	1.65	1.67
CCL2	6347	Hs.303649	738	<0.001	<0.001	1.58	1.6
IL1B	3553	Hs.126256	739	<0.001	<0.001	1.64	1.32
LY6D	8581	Hs.415762	752	<0.001	<0.001	1.6	1.69
MXD1	4084	Hs.468908	763	<0.001	<0.001	1.66	1.45
FCER1G	2207	Hs.433300	775	<0.001	<0.001	1.55	1.95
IFI30	10437	Hs.14623	776	<0.001	<0.001	1.58	1.62
SAA1	6288	Hs.632144	778	<0.001	<0.001	1.4	1.03
CCNA1	8900	Hs.417050	791	<0.001	<0.001	1.7	1.39
CENPN	55839	Hs.55028	807	<0.001	<0.001	1.54	1.43
HLA-G	3135	Hs.512152	817	<0.001	<0.001	1.41	1.59
PLEK	5341	Hs.468840	818	<0.001	<0.001	1.45	1.31
HIST1H1C	3006	Hs.7644	821	<0.001	<0.001	1.58	1.88
PSAP	5660	Hs.523004	823	<0.001	<0.001	1.49	1.46
TREM1	54210	Hs.283022	834	<0.001	<0.001	1.59	1.44
CD52	1043	Hs.276770	846	<0.001	<0.001	1.51	1.5
SPP1	6696	Hs.313	868	<0.001	<0.001	1.45	1.48
SOCS3	9021	Hs.527973	879	<0.001	<0.001	1.56	1.54
TMPRSS13	84000	Hs.266308	885	<0.001	<0.001	1.54	1.6
TNFAIP6	7130	Hs.437322	905	<0.001	<0.001	1.63	1.35
S100A8	6279	Hs.416073	912	<0.001	<0.001	1.18	1.09
G0S2	50486	Hs.432132	924	<0.001	<0.001	1.48	1.23
CCL18	6362	Hs.143961	927	<0.001	<0.001	1.61	1.44
PLUNC	51297	Hs.211092	928	<0.001	<0.001	1.03	1.02
DDX5	1655	Hs.279806	934	<0.001	<0.001	1.29	1.32
AQP9	366	Hs.104624	937	<0.001	<0.001	1.45	1.2
RNASE1	6035	Hs.78224	959	<0.001	<0.001	1.3	1.17
FCGR3A	2214	Hs.372679	1001	<0.001	<0.001	1.23	1.29
HAUS8	93323	Hs.404088	1007	<0.001	<0.001	1.51	1.59
CD53	963	Hs.443057	1012	<0.001	<0.001	1.44	1.45
LGALS1	3956	Hs.445351	1017	<0.001	<0.001	1.38	1.54
C8orf4	56892	Hs.591849	1022	<0.001	<0.001	1.51	1.57
RAD51L3	5892	Hs.631757	1025	<0.001	<0.001	1.57	1.47
MRC1L1	414308	Hs.461247	1030	<0.001	<0.001	1.15	1.29

CXCR4	7852	Hs.593413	1039	<0.001	<0.001	1.39	1.47
IL1R2	7850	Hs.25333	1059	<0.001	<0.001	1.5	1.2
ADAM9	8754	Hs.591852	1063	<0.001	<0.001	1.55	1.55
CXCL9	4283	Hs.77367	1068	<0.001	<0.001	1.13	1.33
FGFBP1	9982	Hs.1690	1072	<0.001	<0.001	1.46	1.35
MDM4	4194	Hs.497492	1073	<0.001	<0.001	1.59	1.93
HN1	51155	Hs.532803	1080	<0.001	<0.001	1.59	1.57
IDO1	3620	Hs.840	1095	<0.001	<0.001	1.29	1.02
C1QB	713	Hs.8986	1101	<0.001	<0.001	1.3	1.37
HRASLS	57110	Hs.36761	1101	<0.001	<0.001	1.6	1.54
CORO6	84940	Hs.143046	1112	<0.001	<0.001	1.51	1.49
OLR1	4973	Hs.412484	1123	<0.001	<0.001	1.41	1.36
BCL2A1	597	Hs.227817	1127	<0.001	<0.001	1.44	1.34
DAPK2	23604	Hs.237886	1137	<0.001	<0.001	1.43	1.46
SELL	6402	Hs.82848	1142	<0.001	<0.001	1.04	1.04
LCP1	3936	Hs.381099	1143	<0.001	<0.001	1.32	1.42
FHL1	2273	Hs.435369	1152	<0.001	<0.001	1.47	1.54
HIST1H2BK	85236	Hs.437275	1153	<0.001	<0.001	1.57	1.5
LANCL1	10314	Hs.13351	1154	<0.001	<0.001	1.37	1.12
FAM26F	441168	Hs.381220	1155	<0.001	<0.001	1.45	1.34
ADM	133	Hs.441047	1161	<0.001	<0.001	1.48	1.31
HLA-DQA1	3117	Hs.387679	1161	<0.001	<0.001	1.31	1.02
ATP6V1F	9296	Hs.78089	1162	<0.001	<0.001	1.47	1.27
GJA4	2701	Hs.296310	1168	<0.001	<0.001	1.47	1.5
MARCO	8685	Hs.67726	1178	<0.001	<0.001	1.27	1.31
C1orf38	9473	Hs.10649	1187	<0.001	<0.001	1.35	1.54
IL32	9235	Hs.943	1203	<0.001	<0.001	1.45	1.35
YWHAZ	7534	Hs.492407	1210	<0.001	<0.001	1.23	1.11
EPSTI1	94240	Hs.546467	1218	<0.001	<0.001	1.23	1.25
GOLGA5	9950	Hs.104320	1218	<0.001	<0.001	1.55	1.42
CSF2RB	1439	Hs.592192	1222	<0.001	<0.001	1.34	1.13
AIF1	199	Hs.76364	1223	<0.001	<0.001	1.34	1.54
SAG	6295	Hs.32721	1231	<0.001	<0.001	1.57	1.36
FPR1	2357	Hs.753	1236	<0.001	<0.001	1.3	1.39
BCL3	602	Hs.31210	1237	<0.001	<0.001	1.5	1.45
LRFN5	145581	Hs.136893	1249	<0.001	<0.001	1.35	1.63
LYZ	4069	Hs.524579	1252	<0.001	<0.001	1.25	1.21
CST1	1469	Hs.123114	1256	<0.001	<0.001	1.37	1.33
GPR109A	338442	Hs.524812	1266	<0.001	<0.001	1.28	1.27
C3AR1	719	Hs.591148	1270	<0.001	<0.001	1.45	1.31
STXBP2	6813	Hs.515104	1270	<0.001	<0.001	1.4	1.06
A2M	2	Hs.212838	1271	<0.001	<0.001	1.19	1.09
DEFB1	1672	Hs.32949	1273	<0.001	<0.001	1.52	1.53
IFITM1	8519	Hs.458414	1274	<0.001	<0.001	1.22	1.02
HLA-DRA	3122	Hs.520048	1277	<0.001	<0.001	1.4	1.81
AKR1A1	10327	Hs.474584	1292	<0.001	<0.001	1.45	1.51

CYP27B1	1594	Hs.524528	1294	<0.001	<0.001	1.52	1.56
BST2	684	Hs.118110	1294	<0.001	<0.001	1.35	1.36
ASS1	445	Hs.160786	1303	<0.001	<0.001	1.41	1.34
NGFRAP1	27018	Hs.448588	1303	<0.001	<0.001	1.44	1.34
ACTG2	72	Hs.516105	1313	<0.001	<0.001	1.36	1.4
SERPINB3	6317	Hs.227948	1318	<0.001	<0.001	1.24	1.31
SERPINB4	6318	Hs.123035	1318	<0.001	<0.001	1.11	1.25
LITAF	9516	Hs.459940	1320	<0.001	<0.001	1.42	1.41
IL1RN	3557	Hs.81134	1330	<0.001	0.001	1.34	1.24
H2AFX	3014	Hs.477879	1346	<0.001	0.001	1.52	1.41
C12orf44	60673	Hs.9911	1350	<0.001	0.001	1.46	1.39
TFF3	7033	Hs.82961	1354	<0.001	0.001	1.17	1.39
FCGR1A	2209	Hs.77424	1355	<0.001	0.001	1.32	1.29
LCP2	3937	Hs.304475	1357	<0.001	0.001	1.41	1.36
LYZL6	57151	Hs.97477	1358	<0.001	0.001	1.42	1.22
NCF2	4688	Hs.587558	1358	<0.001	0.001	1.25	1.53
PPPDE2	27351	Hs.570455	1362	<0.001	0.001	1.44	1.4
STK40	83931	Hs.471768	1364	<0.001	0.001	1.52	1.6
GPR109B	8843	Hs.458425	1364	<0.001	0.001	1.42	1.57
UCP2	7351	Hs.80658	1364	<0.001	0.001	1.37	1.29
MAF	4094	Hs.134859	1365	<0.001	0.001	1.4	1.45
SPRR1B	6699	Hs.1076	1367	<0.001	0.001	1.36	1.21
CDV3	55573	Hs.518265	1372	<0.001	0.001	1.3	1.6
SLC2A3	6515	Hs.419240	1378	<0.001	0.001	1.38	1.41
CSTA	1475	Hs.518198	1379	<0.001	0.001	1.34	1.48
IGSF1	3547	Hs.22111	1384	<0.001	0.001	1.49	1.52
GMFG	9535	Hs.5210	1386	<0.001	0.001	1.3	1.09
CTSZ	1522	Hs.252549	1387	<0.001	0.001	1.44	1.31
TM4SF5	9032	Hs.184194	1392	<0.001	0.001	1.48	1.4
GUCY1B2	2974	Hs.411573	1393	<0.001	0.001	1.43	1.62
PTPRH	5794	Hs.179770	1393	<0.001	0.001	1.52	1.28
MT1F	4494	Hs.513626	1402	<0.001	0.001	1.38	1.22
LOC100130288	100130288	Hs.120196	1404	<0.001	0.001	1.36	1.75
HLA-DMA	3108	Hs.351279	1407	<0.001	0.001	1.22	1.44
TSPAN13	27075	Hs.364544	1412	<0.001	0.001	1.4	1.68
FAM96A	84191	Hs.439548	1413	<0.001	0.001	1.45	1.61
C11orf9	745	Hs.473109	1416	<0.001	0.001	1.44	1.45
STX10	8677	Hs.43812	1425	<0.001	0.001	1.5	1.35
RPL21	6144	Hs.381123	1435	<0.001	0.001	1.29	1.08
WBP5	51186	Hs.533287	1440	<0.001	0.001	1.45	1.4
WFDC2	10406	Hs.2719	1443	<0.001	0.001	1.33	1.31
GGH	8836	Hs.78619	1449	<0.001	0.002	1.44	1.57
ATOX1	475	Hs.125213	1450	<0.001	0.002	1.35	1.2
POLR2L	5441	Hs.441072	1459	<0.001	0.002	1.31	1.4
MTHFD2	10797	Hs.469030	1461	<0.001	0.002	1.38	1.51
RGS16	6004	Hs.413297	1466	<0.001	0.002	1.46	1.34

APOBEC3B	9582	Hs.226307	1468	<0.001	0.002	1.48	1.47
ALOX5AP	241	Hs.507658	1472	<0.001	0.002	1.23	1.03
TNFRSF1B	7133	Hs.256278	1472	<0.001	0.002	1.28	1.24
GPSM3	63940	Hs.520046	1475	<0.001	0.002	1.33	1.42
MFNG	4242	Hs.517603	1483	<0.001	0.002	1.42	1.6
CCNB2	9133	Hs.194698	1485	<0.001	0.002	1.45	1.42
HLA-DRB1	3123	Hs.534322	1493	<0.001	0.002	1.26	1.44
MMP9	4318	Hs.297413	1499	<0.001	0.002	1.33	1.39
MAZ	4150	Hs.23650	1500	<0.001	0.002	1.38	1.36
MUC12	10071	Hs.489355	1507	<0.001	0.003	1.42	1.37
HCST	10870	Hs.117339	1508	<0.001	0.003	1.39	1.44
RARRES1	5918	Hs.131269	1515	<0.001	0.003	1.27	1.13
GSTM4	2948	Hs.348387	1519	<0.001	0.003	1.41	1.4
EMCN	51705	Hs.152913	1522	<0.001	0.003	1.41	1.25
LOC100289478	100289478	Hs.121915	1522	<0.001	0.003	1.4	1.21
SLA	6503	Hs.75367	1524	<0.001	0.003	1.31	1.21
S100P	6286	Hs.2962	1526	<0.001	0.003	1.23	1.05
CCNL1	57018	Hs.4859	1526	<0.001	0.003	1.14	1.08
TMSB4X	7114	Hs.522584	1531	<0.001	0.003	1.33	1.34
VNN2	8875	Hs.293130	1531	<0.001	0.003	1.14	1.22
PSME2	5721	Hs.434081	1535	<0.001	0.003	1.36	1.55
HSD17B6	8630	Hs.524513	1536	<0.001	0.003	1.4	1.39
MTHFS	10588	Hs.459049	1539	<0.001	0.003	1.43	1.38
CFB	629	Hs.69771	1540	<0.001	0.003	1.25	1.03
CXCL10	3627	Hs.632586	1543	<0.001	0.003	1.15	1.01
SERPING1	710	Hs.384598	1543	<0.001	0.003	1.24	1.27
FCAMR	83953	Hs.145519	1544	<0.001	0.003	1.45	1.45
CAV1	857	Hs.74034	1554	<0.001	0.003	1.11	1.06
CORO1A	11151	Hs.415067	1558	<0.001	0.003	1.17	1.14
KIF5A	3798	Hs.151219	1562	<0.001	0.004	1.41	1.29
CREB3L1	90993	Hs.405961	1562	<0.001	0.004	1.27	1.18
LOC645638	645638	Hs.463652	1565	<0.001	0.004	1.3	1.28
EPR1	8475	Hs.514527	1570	<0.001	0.004	1.4	1.38
SERPINA3	12	Hs.534293	1571	<0.001	0.004	1.39	1.42
L1CAM	3897	Hs.522818	1575	<0.001	0.004	1.2	1.24
UNC93A	54346	Hs.567508	1576	<0.001	0.004	1.42	1.34
CD44	960	Hs.502328	1577	<0.001	0.004	1.4	1.4
ITGAX	3687	Hs.248472	1577	<0.001	0.004	1.25	1.05
GCAT	23464	Hs.54609	1578	<0.001	0.004	1.34	1.39
MRPS35	60488	Hs.311072	1580	<0.001	0.004	1.28	1.61
NDST3	9348	Hs.480596	1582	<0.001	0.004	1.37	1.56
PIN1	5300	Hs.465849	1585	<0.001	0.004	1.28	1.29
ZPBP	11055	Hs.388841	1586	<0.001	0.004	1.48	1.48
CPZ	8532	Hs.78068	1587	<0.001	0.004	1.4	1.37
MRPL14	64928	Hs.311190	1589	<0.001	0.004	1.4	1.47
EHHADH	1962	Hs.429879	1590	<0.001	0.004	1.3	1.18

IGLJ3	28831	Hs.449585	1590	<0.001	0.004	1.31	1.15
CD300A	11314	Hs.9688	1593	<0.001	0.004	1.3	1.32
IFI27	3429	Hs.532634	1594	<0.001	0.005	1.02	1.13
SYCP1	6847	Hs.112743	1594	<0.001	0.005	1.26	1.64
PAFAH1B3	5050	Hs.466831	1597	<0.001	0.005	1.35	1.17
RARRES3	5920	Hs.17466	1599	<0.001	0.005	1.22	1.52
TSTA3	7264	Hs.404119	1600	<0.001	0.005	1.34	1.23
LSP1	4046	Hs.56729	1606	<0.001	0.005	1.27	1.01
MUC2	4583	Hs.315	1606	<0.001	0.005	1.24	1.3
LOC100287697	100287697	Hs.326822	1608	<0.001	0.005	1.39	1.43
SH2D3C	10044	Hs.306412	1609	<0.001	0.005	1.33	1.42
HAS3	3038	Hs.592069	1609	<0.001	0.005	1.34	1.57
KIF23	9493	Hs.270845	1614	<0.001	0.005	1.44	1.38
FCER2	2208	Hs.465778	1616	<0.001	0.005	1.4	1.37
TRNP1	388610	Hs.355747	1619	<0.001	0.005	1.32	1.11
CMPK2	129607	Hs.7155	1620	<0.001	0.006	1.21	1.46
PRDX3	10935	Hs.523302	1622	<0.001	0.006	1.34	1.59
FPR2	2358	Hs.99855	1623	<0.001	0.006	1.27	1.13
DIRAS3	9077	Hs.194695	1625	<0.001	0.006	1.32	1.5
C19orf53	28974	Hs.231616	1628	<0.001	0.006	1.3	1.28
UBE2C	11065	Hs.93002	1630	<0.001	0.006	1.36	1.31
CYTH4	27128	Hs.170944	1632	<0.001	0.006	1.29	1.18
CD86	942	Hs.171182	1638	<0.001	0.006	1.37	1.57
SIRPA	140885	Hs.581021	1644	<0.001	0.006	1.27	1.09
MRPL54	116541	Hs.356578	1645	<0.001	0.006	1.22	1.09
UBE2L6	9246	Hs.425777	1645	<0.001	0.006	1.27	1.15
ROBLD3	28956	Hs.632483	1651	<0.001	0.007	1.4	1.2
RPL13	6137	Hs.410817	1652	<0.001	0.007	1.32	1.34
PLIN2	123	Hs.3416	1655	<0.001	0.007	1.21	1.21
RLBP1	6017	Hs.1933	1656	<0.001	0.007	1.39	1.39
RASGRP4	115727	Hs.130434	1656	<0.001	0.007	1.37	1.3
AGXT	189	Hs.144567	1657	<0.001	0.007	1.29	1.18
SLC16A2	6567	Hs.75317	1660	<0.001	0.007	1.34	1.44
PROC	5624	Hs.224698	1661	<0.001	0.007	1.42	1.36
CPVL	54504	Hs.233389	1662	<0.001	0.007	1.18	1.42
MT1G	4495	Hs.433391	1666	<0.001	0.007	1.27	1.05
CCDC6	8030	Hs.591360	1668	<0.001	0.007	1.39	1.38
SEC13	6396	Hs.166924	1669	<0.001	0.007	1.4	1.33
GLO1	2739	Hs.268849	1669	<0.001	0.007	1.26	1.14
BOLA2	552900	Hs.444600	1671	<0.001	0.007	1.38	1.27
TRMT2B	79979	Hs.496501	1673	<0.001	0.007	1.3	1.23
KRT8	3856	Hs.533782	1675	<0.001	0.008	1.33	1.47
C1QTNF2	114898	Hs.110062	1677	<0.001	0.008	1.42	1.37
RGS1	5996	Hs.75256	1678	<0.001	0.008	1.27	1.37
GZMA	3001	Hs.90708	1681	<0.001	0.008	1.08	1.05
MSN	4478	Hs.87752	1683	<0.001	0.008	1.26	1.27

C19orf70	125988	Hs.356626	1685	<0.001	0.008	1.26	1.39
DHX9	1660	Hs.191518	1687	<0.001	0.008	1.23	1.35
IL18BP	10068	Hs.591967	1689	<0.001	0.008	1.39	1.28
MUC4	4585	Hs.369646	1691	<0.001	0.008	1.26	1.25
PCOLCE2	26577	Hs.8944	1692	<0.001	0.008	1.19	1.35
CHRNB4	1143	Hs.624178	1694	<0.001	0.008	1.38	1.59
LOC100288376	100288376	Hs.519523	1698	<0.001	0.009	1.19	1.24
CCS	9973	Hs.502917	1698	<0.001	0.009	1.42	1.34
TTC13	79573	Hs.424788	1699	<0.001	0.009	1.35	1.53
GK	2710	Hs.1466	1701	<0.001	0.009	1.37	1.43
PPP1R9A	55607	Hs.21816	1704	<0.001	0.009	1.29	1.44
RPL27A	6157	Hs.523463	1705	<0.001	0.009	1.3	1.13
CXCL1	2919	Hs.789	1705	<0.001	0.009	1.21	1.19
SALL1	6299	Hs.135787	1713	<0.001	0.01	1.28	1.27
GLUL	2752	Hs.518525	1716	<0.001	0.01	1.23	1.21
ISG20	3669	Hs.459265	1717	<0.001	0.01	1.29	1.29
IFITM3	10410	Hs.374650	1717	<0.001	0.01	1.19	1.06
FAM96B	51647	Hs.9825	1718	<0.001	0.01	1.31	1.2
PCNT	5116	Hs.474069	1719	<0.001	0.01	1.3	1.27
PDYN	5173	Hs.22584	1721	<0.001	0.01	1.36	1.26
ASAH1	427	Hs.527412	1722	<0.001	0.01	1.07	1.18
SAT1	6303	Hs.28491	1724	<0.001	0.01	1.32	1.21
RAMP2	10266	Hs.514193	1727	<0.001	0.011	1.31	1.26
ZNF555	148254	Hs.47712	1728	<0.001	0.011	1.38	1.33
VSIG4	11326	Hs.8904	1731	<0.001	0.011	1.01	1.13
LOXL1	4016	Hs.65436	1737	<0.001	0.012	1.34	1.32
SPAG11A	653423	Hs.559506	1739	<0.001	0.012	1.39	1.32
ALPI	248	Hs.37009	1742	<0.001	0.012	1.34	1.38
STAT4	6775	Hs.80642	1746	<0.001	0.012	1.13	1.22
OAZ2	4947	Hs.74563	1748	<0.001	0.012	1.33	1.22
CYBA	1535	Hs.513803	1754	<0.001	0.012	1.29	1.17
PRDX4	10549	Hs.83383	1755	<0.001	0.013	1.32	1.48
PTMS	5763	Hs.504613	1762	<0.001	0.013	1.33	1.09
SLC24A6	80024	Hs.286194	1763	<0.001	0.013	1.32	1.62
MS4A4A	51338	Hs.325960	1764	<0.001	0.013	1.22	1.41
RDH10	157506	Hs.244940	1764	<0.001	0.013	1.23	1.31
CLPP	8192	Hs.515092	1769	<0.001	0.013	1.3	1.37
SCGB3A1	92304	Hs.62492	1772	<0.001	0.014	1.16	1.12
MYL6	4637	Hs.632717	1774	<0.001	0.014	1.3	1.55
ARSA	410	Hs.88251	1774	<0.001	0.014	1.35	1.39
BLOC1S1	2647	Hs.94672	1775	<0.001	0.014	1.3	1.3
LOC100288550	100288550	Hs.448226	1776	<0.001	0.014	1.28	1.25
LOC100288974	100288974	Hs.102310	1776	<0.001	0.014	1.36	1.24
KCNJ15	3772	Hs.411299	1776	<0.001	0.014	1.31	1.17
CD37	951	Hs.166556	1777	<0.001	0.014	1.19	1.09
WIPF1	7456	Hs.128067	1780	<0.001	0.014	1.27	1.4

DNTTIP2	30836	Hs.85769	1782	<0.001	0.014	1.13	1.3
TUBB4	10382	Hs.110837	1785	<0.001	0.014	1.38	1.36
RABAC1	10567	Hs.11417	1792	<0.001	0.015	1.28	1.37
GPR182	11318	Hs.483909	1792	<0.001	0.015	1.29	1.29
ACSL1	2180	Hs.406678	1792	<0.001	0.015	1.18	1.36
KLK7	5650	Hs.151254	1792	<0.001	0.015	1.37	1.34
PPEF2	5470	Hs.290873	1792	<0.001	0.015	1.31	1.18
SOD3	6649	Hs.2420	1795	<0.001	0.015	1.34	1.28
IER3	8870	Hs.591785	1795	<0.001	0.015	1.14	1.35
TNNI1	7135	Hs.320890	1796	<0.001	0.016	1.4	1.2
ESPL1	9700	Hs.153479	1802	<0.001	0.016	1.41	1.4
UBB	7314	Hs.356190	1803	<0.001	0.016	1.27	1.39
SLC22A1	6580	Hs.117367	1803	<0.001	0.016	1.39	1.43
NFKBIB	4793	Hs.9731	1805	<0.001	0.017	1.41	1.32
VMO1	284013	Hs.122561	1806	<0.001	0.017	1.2	1.32
C1orf194	127003	Hs.446962	1806	<0.001	0.017	1.26	1.24
DUSP2	1844	Hs.1183	1809	<0.001	0.017	1.29	1.25
RASAL1	8437	Hs.528693	1814	<0.001	0.018	1.31	1.08
LOC100288209	100288209	Hs.601492	1817	<0.001	0.018	1.24	1.46
COX7A2	1347	Hs.70312	1821	<0.001	0.019	1.18	1.54
NCKAP1L	3071	Hs.182014	1822	<0.001	0.019	1.23	1.42
LCN2	3934	Hs.204238	1823	<0.001	0.019	1.24	1.08
S100A11	6282	Hs.417004	1823	<0.001	0.018	1.25	1.3
ATP1A3	478	Hs.515427	1826	0.001	0.019	1.14	1.25
CDCA8	55143	Hs.524571	1828	0.001	0.019	1.36	1.32
HLA-DPB1	3115	Hs.485130	1833	0.001	0.02	1.11	1.68
LGALS9	3965	Hs.81337	1834	0.001	0.02	1.32	1.36
STOML2	30968	Hs.3439	1835	0.001	0.02	1.32	1.36
CCL5	6352	Hs.514821	1835	0.001	0.02	1.22	1.1
CD1A	909	Hs.1309	1836	0.001	0.02	1.29	1.29
NRG2	9542	Hs.408515	1836	0.001	0.02	1.33	1.16
FUT3	2525	Hs.169238	1839	0.001	0.02	1.22	1.25
DMBT1	1755	Hs.279611	1840	0.001	0.02	1.33	1.04
PTTG1	9232	Hs.350966	1843	0.001	0.021	1.34	1.11
CYCS	54205	Hs.437060	1843	0.001	0.021	1.25	1.38
SYNGR2	9144	Hs.464210	1849	0.001	0.022	1.3	1.47
RPL29	6159	Hs.425125	1850	0.001	0.021	1.29	1.31
SURF4	6836	Hs.512465	1850	0.001	0.021	1.29	1.35
CAMK2N1	55450	Hs.197922	1851	0.001	0.021	1.34	1.28
CD3D	915	Hs.504048	1854	0.001	0.022	1.15	1.06
B9D1	27077	Hs.462445	1855	0.001	0.022	1.25	1.13
POLR2I	5438	Hs.47062	1857	0.001	0.022	1.16	1.12
C1orf96	126731	Hs.585011	1859	0.001	0.022	1.38	1.32
HIST1H4C	8364	Hs.46423	1859	0.001	0.022	1.01	1.12
ATP5G2	517	Hs.524464	1861	0.001	0.023	1.26	1.16
CAV2	858	Hs.212332	1866	0.001	0.023	1.16	1.33

TCEB2	6923	Hs.172772	1867	0.001	0.024	1.27	1.26
PTGS1	5742	Hs.201978	1868	0.001	0.024	1.35	1.31
AP1M1	8907	Hs.71040	1869	0.001	0.024	1.31	1.3
PPFIA3	8541	Hs.413748	1871	0.001	0.024	1.33	1.19
CDK5R1	8851	Hs.500015	1873	0.001	0.024	1.08	1.36
ATP8B3	148229	Hs.306212	1873	0.001	0.024	1.32	1.3
TSPO2	222642	Hs.357392	1877	0.001	0.025	1.29	1.43
ADCK1	57143	Hs.413208	1878	0.001	0.025	1.34	1.46
FGR	2268	Hs.1422	1878	0.001	0.025	1.09	1.18
SMPDL3B	27293	Hs.123659	1880	0.001	0.025	1.41	1.32
C1QC	714	Hs.467753	1882	0.001	0.026	1.17	1.25
ZNF131	7690	Hs.535804	1883	0.001	0.026	1.36	1.27
GALNT13	114805	Hs.470277	1883	0.001	0.026	1.26	1.36
FBP1	2203	Hs.494496	1884	0.001	0.026	1.11	1.14
ECHS1	1892	Hs.76394	1887	0.001	0.026	1.31	1.31
CD163	9332	Hs.504641	1889	0.001	0.026	1.14	1.26
TTYH3	80727	Hs.440899	1893	0.001	0.027	1.37	1.33
AGER	177	Hs.534342	1897	0.001	0.028	1.26	1.18
ACP5	54	Hs.1211	1900	0.001	0.028	1.18	1.28
CNPY1	285888	Hs.146751	1901	0.001	0.028	1.28	1.18
FLJ40852	285962	Hs.17589	1901	0.001	0.028	1.31	1.42
LAP3	51056	Hs.570791	1901	0.001	0.028	1.07	1.04
C19orf56	51398	Hs.108969	1903	0.001	0.028	1.19	1.21
PDAP1	11333	Hs.632296	1904	0.001	0.028	1.35	1.48
STOM	2040	Hs.253903	1904	0.001	0.028	1.34	1.17
CAPN14	440854	Hs.468059	1904	0.001	0.028	1.25	1.49
NEFM	4741	Hs.458657	1906	0.001	0.029	1.26	1.39
CEACAM6	4680	Hs.466814	1907	0.001	0.029	1.2	1.21
AMPD1	270	Hs.89570	1911	0.001	0.03	1.26	1.27
MRPL18	29074	Hs.416998	1911	0.001	0.03	1.35	1.39
BTG1	694	Hs.255935	1916	0.001	0.03	1.31	1.34
NIPSNAP1	8508	Hs.173878	1916	0.001	0.03	1.28	1.26
CCL26	10344	Hs.131342	1918	0.001	0.03	1.3	1.14
TUBB3	10381	Hs.511743	1918	0.001	0.03	1.31	1.28
LGALS2	3957	Hs.531776	1921	0.001	0.031	1.28	1.32
FGL2	10875	Hs.520989	1922	0.001	0.031	1.05	1.29
MBNL1	4154	Hs.201858	1922	0.001	0.031	1.29	1.4
RPSA	3921	Hs.449909	1925	0.001	0.032	1.22	1.21
LMX1B	4010	Hs.129133	1926	0.001	0.032	1.36	1.38
KIF9	64147	Hs.373947	1928	0.001	0.032	1.26	1.28
ASF1B	55723	Hs.26516	1930	0.001	0.032	1.33	1.42
CCL21	6366	Hs.57907	1930	0.001	0.032	1.36	1.34
DCTN6	10671	Hs.158427	1931	0.001	0.032	1.3	1.34
NDST1	3340	Hs.222055	1933	0.001	0.033	1.35	1.32
FAM100A	124402	Hs.513313	1934	0.001	0.033	1.32	1.12
NDUFA2	4695	Hs.534333	1934	0.001	0.033	1.25	1.14

LMAN2	10960	Hs.75864	1935	0.001	0.033	1.25	1.2
AIP	9049	Hs.412433	1938	0.001	0.033	1.28	1.44
VIP	7432	Hs.53973	1939	0.001	0.033	1.34	1.17
SRR	63826	Hs.461954	1940	0.001	0.034	1.33	1.4
CHCHD8	51287	Hs.475387	1942	0.001	0.034	1.31	1.3
LOC401397	401397	Hs.117929	1943	0.001	0.034	1.27	1.41
DSE	29940	Hs.458358	1945	0.001	0.034	1.23	1.26
NOP10	55505	Hs.14317	1946	0.001	0.034	1.25	1.23
GCM1	8521	Hs.28346	1948	0.001	0.035	1.33	1.19
AKR1B1	231	Hs.521212	1948	0.001	0.035	1.18	1.22
ETF1	2107	Hs.483494	1949	0.001	0.035	1.22	1.15
NDUFA7	4701	Hs.333427	1952	0.001	0.035	1.17	1.28
NMI	9111	Hs.54483	1953	0.001	0.036	1.24	1.27
LGI3	203190	Hs.33470	1954	0.001	0.036	1.3	1.35
GSS	2937	Hs.82327	1954	0.001	0.036	1.34	1.25
HERC1	8925	Hs.210385	1955	0.001	0.036	1.2	1.4
GRIN2A	2903	Hs.411472	1957	0.001	0.036	1.3	1.28
MEOX1	4222	Hs.438	1962	0.001	0.037	1.24	1.35
CPN1	1369	Hs.2246	1968	0.002	0.038	1.37	1.34
GPNMB	10457	Hs.190495	1968	0.002	0.038	1.15	1.24
MRPL41	64975	Hs.44017	1969	0.002	0.038	1.18	1.34
ZNF701	55762	Hs.235167	1971	0.002	0.038	1.33	1.38
PSMB8	5696	Hs.180062	1972	0.002	0.039	1.18	1.25
CTSL1	1514	Hs.418123	1973	0.002	0.039	1.2	1.32
FAM135B	51059	Hs.126024	1977	0.002	0.04	1.24	1.28
FOXF2	2295	Hs.484423	1980	0.002	0.04	1.21	1.23
GPR87	53836	Hs.591292	1981	0.002	0.04	1.29	1.18
BDNF	627	Hs.502182	1983	0.002	0.041	1.19	1.17
CEP78	84131	Hs.374421	1984	0.002	0.041	1.28	1.17
SLC16A1	6566	Hs.75231	1985	0.002	0.041	1.21	1.14
C3orf14	57415	Hs.47166	1986	0.002	0.041	1.27	1.27
PLAUR	5329	Hs.466871	1986	0.002	0.041	1.28	1.21
S100A10	6281	Hs.143873	1987	0.002	0.041	1.25	1.41
TEK	7010	Hs.89640	1988	0.002	0.041	1.16	1.12
PRAC	84366	Hs.116467	1989	0.002	0.041	1.16	1.04
TIMM8B	26521	Hs.279915	1992	0.002	0.042	1.23	1.42
BIN2	51411	Hs.14770	1992	0.002	0.042	1.15	1.07
BACH1	571	Hs.154276	1994	0.002	0.042	1.13	1.05
TWF2	11344	Hs.436439	1994	0.002	0.042	1.22	1.19
SLC6A6	6533	Hs.529488	1995	0.002	0.042	1.3	1.19
APOL1	8542	Hs.114309	1996	0.002	0.043	1.23	1.08
PLK4	10733	Hs.172052	1997	0.002	0.043	1.3	1.32
HSD3B7	80270	Hs.460618	1997	0.002	0.043	1.29	1.33
PLTP	5360	Hs.439312	1999	0.002	0.043	1.28	1.25
NDUFA13	51079	Hs.534453	1999	0.002	0.043	1.26	1.21
GSTM5	2949	Hs.75652	2001	0.002	0.043	1.34	1.29

MT1E	4493	Hs.534330	2001	0.002	0.043	1.21	1.22
MYBL2	4605	Hs.179718	2005	0.002	0.044	1.34	1.38
OLFM1	10439	Hs.522484	2006	0.002	0.044	1.29	1.19
POLD4	57804	Hs.523829	2006	0.002	0.044	1.24	1.39
FUT10	84750	Hs.458713	2011	0.002	0.045	1.28	1.35
GPER	2852	Hs.20961	2011	0.002	0.045	1.26	1.27
HLA-B	3106	Hs.77961	2012	0.002	0.045	1.13	1.29
TMEM48	55706	Hs.476525	2013	0.002	0.045	1.33	1.41
ESRRA	2101	Hs.110849	2016	0.002	0.046	1.2	1.22
NCF4	4689	Hs.474781	2018	0.002	0.046	1.26	1.26
ATG2A	23130	Hs.370671	2019	0.002	0.047	1.36	1.35
TCP10L	140290	Hs.42034	2020	0.002	0.047	1.28	1.46
BRPF3	27154	Hs.520096	2021	0.002	0.047	1.29	1.25
SALL4	57167	Hs.517113	2021	0.002	0.047	1.35	1.33
MAGEA6	4105	Hs.441113	2022	0.002	0.047	1.33	1.19
EPB41L4B	54566	Hs.591901	2022	0.002	0.047	1.16	1.17
LOC100288413	100288413	Hs.363087	2024	0.002	0.047	1.34	1.22
C12orf65	91574	Hs.319128	2025	0.002	0.048	1.31	1.28
KDM5D	8284	Hs.80358	2025	0.002	0.048	1.24	1.33
SLC6A4	6532	Hs.134662	2026	0.002	0.048	1.21	1.1
NDUFA1	4694	Hs.534168	2029	0.002	0.049	1.19	1.36
DNMT1	1786	Hs.202672	2030	0.002	0.049	1.34	1.31
B2M	567	Hs.534255	2031	0.002	0.049	1.12	1.04
KPNA2	3838	Hs.594238	2031	0.002	0.049	1.26	1.34
TMEM40	55287	Hs.475502	2032	0.002	0.049	1.17	1.23
CDC25B	994	Hs.153752	2033	0.002	0.049	1.27	1.29
LOC644838	644838	Hs.145561	2035	0.002	0.05	1.34	1.23
FABP4	2167	Hs.391561	2035	0.002	0.05	1.09	1.21
SPRYD3	84926	Hs.343334	2036	0.002	0.05	1.25	1.37
SSNA1	8636	Hs.530314	2037	0.002	0.05	1.31	1.27
TBC1D23	55773	Hs.477003	2037	0.002	0.05	1.24	1.11

RP, rank product; P, p-value ;FDR, false discovery rate; FC, fold change

Table 2. Genes significantly down-regulated in BOS-positive samples

Gene symbol	Entrez ID	UniGene	RP	P	FDR	$\frac{1}{MeanFC}$	$\frac{1}{MedianFC}$
ALDH3A1	218	Hs.531682	136	<0.001	<0.001	2.47	2.85
SCGB3A2	117156	Hs.483765	225	<0.001	<0.001	2.15	1.87
HBA1	3039	Hs.449630	286	<0.001	<0.001	1.62	2.37
HBB	3043	Hs.523443	294	<0.001	<0.001	1.57	3.17
MID1	4281	Hs.27695	395	<0.001	<0.001	1.77	1.85
SFTPC	6440	Hs.1074	458	<0.001	<0.001	1.18	1.31
NCRNA00086	399668	Hs.374414	544	<0.001	<0.001	1.74	1.74
ZNF704	619279	Hs.632067	639	<0.001	<0.001	1.72	1.76

OLA1	29789	Hs.157351	671	<0.001	<0.001	1.57	2
AHSP	51327	Hs.274309	699	<0.001	<0.001	1.82	1.26
VAV2	7410	Hs.369921	762	<0.001	<0.001	1.58	1.7
PIPOX	51268	Hs.462585	767	<0.001	<0.001	1.65	1.74
CES1	1066	Hs.558865	787	<0.001	<0.001	1.6	1.83
LTF	4057	Hs.529517	795	<0.001	<0.001	1.37	1.92
SCGB1A1	7356	Hs.523732	814	<0.001	<0.001	1.32	1.4
SFTPB	6439	Hs.512690	830	<0.001	<0.001	1.23	1.84
TRMT61A	115708	Hs.525610	864	<0.001	<0.001	1.62	1.87
F2RL2	2151	Hs.42502	876	<0.001	<0.001	1.59	1.48
APOBEC2	10930	Hs.555915	895	<0.001	<0.001	1.41	1.69
ZNF700	90592	Hs.528486	937	<0.001	<0.001	1.15	1.53
SLC25A36	55186	Hs.144130	943	<0.001	<0.001	1.35	1.43
RTBDN	83546	Hs.21162	950	<0.001	<0.001	1.48	1.6
HFE2	148738	Hs.632436	966	<0.001	<0.001	1.49	1.51
MUC5AC	4586	Hs.534332	969	<0.001	<0.001	1.3	1.09
CXorf57	55086	Hs.274267	978	<0.001	<0.001	1.45	1.71
APBB1	322	Hs.372840	984	<0.001	<0.001	1.5	1.47
WNT6	7475	Hs.29764	992	<0.001	<0.001	1.42	1.85
C14orf132	56967	Hs.6434	993	<0.001	<0.001	1.42	1.42
TBC1D8	11138	Hs.442657	1006	<0.001	<0.001	1.51	1.64
CORO2B	10391	Hs.551213	1007	<0.001	<0.001	1.43	1.56
ZMAT1	84460	Hs.496512	1029	<0.001	<0.001	1.29	1.37
CAMK2G	818	Hs.523045	1069	<0.001	<0.001	1.41	1.49
CDHR1	92211	Hs.137556	1089	<0.001	<0.001	1.47	1.64
SFTPD	6441	Hs.253495	1092	<0.001	<0.001	1.24	1.57
HYDIN	54768	Hs.461229	1100	<0.001	<0.001	1.37	1.25
MST1	4485	Hs.512587	1114	<0.001	<0.001	1.31	1.12
PRKAR2A	5576	Hs.631923	1119	<0.001	<0.001	1.46	1.5
IFIT1	3434	Hs.20315	1124	<0.001	<0.001	1.38	1.33
LOC100130872	100130872	Hs.302963	1129	<0.001	0.001	1.39	1.67
RAB3B	5865	Hs.123072	1132	<0.001	0.001	1.47	1.61
WDR19	57728	Hs.438482	1142	<0.001	0.001	1.37	1.25
ALDH3B2	222	Hs.87539	1144	<0.001	0.001	1.19	1.37
NELL2	4753	Hs.505326	1148	<0.001	0.001	1.39	1.09
CDKN1C	1028	Hs.106070	1153	<0.001	0.001	1.41	1.07
S100A12	6283	Hs.19413	1154	<0.001	0.002	1.24	1.12
MPL	4352	Hs.82906	1166	<0.001	0.002	1.44	1.44
KCNMB4	27345	Hs.525529	1184	<0.001	0.002	1.42	1.7
PTPN14	5784	Hs.193557	1190	<0.001	0.002	1.44	1.51
GSTA3	2940	Hs.102484	1197	<0.001	0.002	1.42	1.33
NUP210	23225	Hs.475525	1202	<0.001	0.002	1.44	1.56
C1orf173	127254	Hs.531182	1202	<0.001	0.002	1.31	1.43
CRLF1	9244	Hs.114948	1212	<0.001	0.002	1.32	1.25
DLGAP2	9228	Hs.113287	1213	<0.001	0.002	1.37	1.52
NLGN4X	57502	Hs.21107	1214	<0.001	0.002	1.39	1.82

SLC35D3	340146	Hs.369703	1227	<0.001	0.002	1.45	1.44
NRG1	3084	Hs.453951	1236	<0.001	0.002	1.39	1.51
RGS22	26166	Hs.120021	1243	<0.001	0.002	1.33	1.55
ZNF233	353355	Hs.466891	1246	<0.001	0.003	1.44	1.47
LRP5	4041	Hs.6347	1246	<0.001	0.003	1.35	1.27
LOC440248	440248	Hs.531509	1254	<0.001	0.003	1.15	1.51
ATOH8	84913	Hs.135569	1258	<0.001	0.003	1.36	1.38
TMEM146	257062	Hs.631842	1275	<0.001	0.003	1.21	1.2
CX3CR1	1524	Hs.78913	1277	<0.001	0.003	1.38	1.21
SNCAIP	9627	Hs.426463	1281	<0.001	0.003	1.46	1.4
C12orf66	144577	Hs.505871	1284	<0.001	0.003	1.44	1.34
USP6	9098	Hs.448851	1299	<0.001	0.003	1.26	1.4
ALAS2	212	Hs.522666	1306	<0.001	0.004	1.43	1.04
CCDC66	285331	Hs.476399	1310	<0.001	0.004	1.26	1.38
GSTA2	2939	Hs.94107	1312	<0.001	0.004	1.34	1.32
COL7A1	1294	Hs.476218	1317	<0.001	0.004	1.4	1.25
DNAH5	1767	Hs.212360	1322	<0.001	0.004	1.25	1.36
CACNA1C	775	Hs.118262	1357	<0.001	0.004	1.37	1.42
NOX4	50507	Hs.371036	1358	<0.001	0.004	1.36	1.47
LOC100287743	100287743	Hs.129095	1359	<0.001	0.004	1.41	1.67
ACSS1	84532	Hs.529353	1362	<0.001	0,004	1.3	1.53
FLRT3	23767	Hs.41296	1371	<0.001	0,004	1.35	1.4
AKR1C2	1646	Hs.567256	1372	<0.001	0,005	1.27	1.06
SNX13	23161	Hs.487648	1390	<0.001	0,005	1.29	1.27
NPIPL3	23117	Hs.611072	1390	<0.001	0,005	1.12	1.66
NAP1L5	266812	Hs.12554	1395	<0.001	0,005	1.4	1.61
ZNF703	80139	Hs.288042	1402	<0.001	0,006	1.3	1.3
TP53INP1	94241	Hs.492261	1406	<0.001	0,006	1.32	1.39
C1orf168	199920	Hs.437655	1410	<0.001	0,006	1.25	1.21
C18orf10	25941	Hs.436636	1413	<0.001	0,006	1.38	1.38
PRR15L	79170	Hs.368260	1428	<0.001	0,007	1.3	1.17
METRN	79006	Hs.533772	1438	<0.001	0,008	1.13	1.32
KRCC1	51315	Hs.469254	1440	<0.001	0,008	1.38	1.33
SNTB1	6641	Hs.46701	1441	<0.001	0,008	1.36	1.32
SLC29A1	2030	Hs.25450	1454	<0.001	0,008	1.42	1.58
DNAH7	56171	Hs.97403	1460	<0.001	0,008	1.24	1.2
PLEKHG1	57480	Hs.189781	1461	<0.001	0,008	1.18	1.7
MBLAC2	153364	Hs.64004	1474	<0.001	0,009	1.39	1.49
MCF2L2	23101	Hs.208267	1474	<0.001	0,009	1.2	1.27
DNAH3	55567	Hs.526500	1478	<0.001	0,009	1.23	1.58
DSC1	1823	Hs.567260	1479	<0.001	0,009	1.25	1.3
NRCAM	4897	Hs.21422	1486	<0.001	0,009	1.37	1.49
C12orf4	57102	Hs.302977	1488	<0.001	0,009	1.32	1.25
RNPC3	55599	Hs.632423	1492	<0.001	0,009	1.19	1.37
NKAIN1	79570	Hs.470259	1499	<0.001	0,010	1.34	1.5
WIF1	11197	Hs.284122	1504	<0.001	0,010	1.33	1.29

SNRPN	6638	Hs.592473	1504	<0.001	0,010	1.16	1.31
CYP1B1	1545	Hs.154654	1512	<0.001	0,011	1.41	1.23
FBXL3	26224	Hs.508284	1514	<0.001	0,011	1.29	1.17
C4A	720	Hs.534847	1515	<0.001	0,011	1.26	1.31
RBM18	92400	Hs.415842	1522	<0.001	0,011	1.25	1.28
C21orf7	56911	Hs.222802	1523	<0.001	0,011	1.29	1.37
FAM120A	23196	Hs.372003	1524	<0.001	0,011	1.4	1.52
NAT10	55226	Hs.577281	1524	<0.001	0,011	1.32	1.4
CDC42SE2	56990	Hs.508829	1525	<0.001	0,011	1.39	1.36
DENND5B	160518	Hs.118166	1535	<0.001	0,012	1.33	1.37
WDSUB1	151525	Hs.20848	1540	<0.001	0,012	1.33	1.28
PPY	5539	Hs.558368	1549	<0.001	0,013	1.26	1.34
ZNF439	90594	Hs.528731	1552	<0.001	0,014	1.19	1.28
SLC6A20	54716	Hs.413095	1554	<0.001	0,014	1.3	1.36
PON3	5446	Hs.440967	1554	<0.001	0,013	1.25	1.32
FAM83A	84985	Hs.379821	1555	<0.001	0,013	1.29	1.4
HSD17B13	345275	Hs.284414	1570	<0.001	0,015	1.22	1.25
GPCPD1	56261	Hs.636359	1571	<0.001	0,015	1.24	1.44
OXTR	5021	Hs.2820	1572	<0.001	0,015	1.29	1.33
ANAPC16	119504	Hs.426296	1573	<0.001	0,015	1.19	1.46
FOLR1	2348	Hs.73769	1579	<0.001	0,015	1.13	1.17
PTCD2	79810	Hs.126906	1579	<0.001	0,015	1.34	1.44
PPP2CA	5515	Hs.105818	1586	<0.001	0,016	1.31	1.29
LOC100288779	100288779	Hs.11729	1587	<0.001	0,016	1.29	1.25
SNRPN	6638	Hs.632166	1588	<0.001	0,016	1.36	1.41
CXXC4	80319	Hs.12248	1595	<0.001	0,017	1.3	1.37
ZNF771	51333	Hs.148584	1597	<0.001	0,017	1.26	1.31
CHRFAM7A	89832	Hs.510853	1598	<0.001	0,017	1.19	1.24
ARGLU1	55082	Hs.508644	1598	<0.001	0,017	1.15	1.31
SESN1	27244	Hs.591336	1607	<0.001	0,018	1.32	1.17
MLLT11	10962	Hs.75823	1615	<0.001	0,019	1.32	1.51
HDAC10	83933	Hs.26593	1615	<0.001	0,019	1.31	1.51
ACY3	91703	Hs.126265	1617	<0.001	0,019	1.28	1.43
PRKCA	5578	Hs.531704	1619	<0.001	0,019	1.3	1.35
VSTM2L	128434	Hs.517029	1641	<0.001	0,021	1.22	1.28
RNLS	55328	Hs.149849	1651	<0.001	0,022	1.25	1.32
SLC25A37	51312	Hs.596025	1653	<0.001	0,022	1.25	1.24
C18orf16	147429	Hs.559280	1654	<0.001	0,022	1.36	1.4
ABCA13	154664	Hs.226568	1655	<0.001	0,022	1.16	1.32
LOC440354	440354	Hs.552700	1657	<0.001	0,022	1.32	1.32
FXYD2	486	Hs.413137	1662	<0.001	0,023	1.29	1.13
PM20D2	135293	Hs.356247	1664	<0.001	0,023	1.31	1.42
JHDM1D	80853	Hs.308710	1666	<0.001	0,023	1.2	1.44
LOC100132004	100132004	Hs.434447	1670	<0.001	0,024	1.31	1.33
SLC22A9	114571	Hs.502772	1673	<0.001	0,024	1.32	1.31
CYP4X1	260293	Hs.439760	1676	<0.001	0,025	1.26	1.28

TNS3	64759	Hs.520814	1678	<0.001	0,025	1.23	1.35
FAM111B	374393	Hs.186579	1680	<0.001	0,025	1.33	1.4
NRP1	8829	Hs.131704	1680	<0.001	0,025	1.27	1.17
SEZ6L2	26470	Hs.6314	1681	<0.001	0,025	1.23	1.45
GOLGA8A	23015	Hs.182982	1687	<0.001	0,026	1.24	1.31
DEFA4	1669	Hs.591391	1687	<0.001	0,026	1.24	1.1
CLEC3B	7123	Hs.476092	1689	<0.001	0,026	1.1	1.13
CX3CL1	6376	Hs.531668	1689	<0.001	0,026	1.19	1.24
MPP6	51678	Hs.533355	1694	<0.001	0,027	1.38	1.25
CXorf42	158801	Hs.442518	1694	<0.001	0,027	1.35	1.37
PDXP	57026	Hs.632762	1700	<0.001	0,028	1.31	1.35
MGC39584	441058	Hs.130535	1702	<0.001	0,028	1.33	1.44
LOC100286909	100286909	Hs.445414	1705	<0.001	0,028	1.16	1.22
ATRN	8455	Hs.276252	1710	<0.001	0.029	1.22	1.37
MSMB	4477	Hs.255462	1718	<0.001	0.031	1.08	1.34
SESTD1	91404	Hs.30977	1719	<0.001	0.031	1.22	1.37
CDK5RAP1	51654	Hs.435952	1724	<0.001	0.031	1.29	1.35
LGALS3	3958	Hs.531081	1724	<0.001	0.031	1.11	1.28
GMPR	2766	Hs.484741	1728	<0.001	0.031	1.23	1.28
KIAA1407	57577	Hs.477159	1742	<0.001	0.034	1.19	1.15
CYP24A1	1591	Hs.89663	1744	<0.001	0.034	1.24	1.31
USP51	158880	Hs.40061	1747	<0.001	0.035	1.33	1.45
RAI2	10742	Hs.446680	1751	<0.001	0.035	1.27	1.32
ALCAM	214	Hs.591293	1752	<0.001	0.036	1.31	1.22
ODF2L	57489	Hs.149360	1754	<0.001	0.036	1.15	1.2
TMX3	54495	Hs.440534	1755	<0.001	0.036	1.2	1.17
PPP3R1	5534	Hs.280604	1755	<0.001	0.036	1.2	1.35
MUC5B	727897	Hs.523395	1757	<0.001	0.036	1.12	1.04
C1orf111	284680	Hs.97784	1766	<0.001	0.038	1.34	1.26
SIPA1L2	57568	Hs.268774	1768	<0.001	0.039	1.29	1.32
RAPGEFL1	51195	Hs.632254	1770	<0.001	0.039	1.29	1.28
SMYD1	150572	Hs.516176	1775	<0.001	0.040	1.12	1.33
VAT1L	57687	Hs.461405	1778	<0.001	0.040	1.21	1.31
LOC100130522	100130522	Hs.352602	1781	<0.001	0,041	1.32	1.37
AGTR1	185	Hs.477887	1783	<0.001	0.041	1.28	1.26
CD3G	917	Hs.2259	1783	<0.001	0.041	1.34	1.22
C1QL2	165257	Hs.433493	1785	<0.001	0.041	1.29	1.23
STK33	65975	Hs.501833	1789	<0.001	0.042	1.19	1.17
GABRE	2564	Hs.22785	1795	<0.001	0.043	1.21	1.31
NHLRC2	374354	Hs.594372	1795	<0.001	0.043	1.23	1.27
HDAC8	55869	Hs.310536	1796	<0.001	0.043	1.32	1.49
CBLN2	147381	Hs.569851	1796	<0.001	0.043	1.22	1.08
FAM110B	90362	Hs.154652	1798	<0.001	0.043	1.31	1.23
KCNJ16	3773	Hs.463985	1803	<0.001	0,044	1.19	1.43
LOC645195	645195	Hs.536063	1807	<0.001	0.045	1.33	1.26
IQGAP1	8826	Hs.430551	1809	<0.001	0.046	1.25	1.17

PROZ	8858	Hs.1011	1812	<0.001	0.046	1.16	1.24
CLK1	1195	Hs.433732	1815	<0.001	0.046	1.17	1.34
PAN2	9924	Hs.273397	1815	<0.001	0.046	1.29	1.27
FLJ31104	441072	Hs.482141	1816	<0.001	0.046	1.28	1.36
C5orf4	10826	Hs.519694	1817	<0.001	0.046	1.28	1.23
SEMA5A	9037	Hs.27621	1819	0.001	0.047	1.13	1.05
ZNF592	9640	Hs.79347	1820	0.001	0.047	1.31	1.23
POLR3B	55703	Hs.62696	1823	0.001	0.048	1.28	1.45
LGR6	59352	Hs.497402	1824	0.001	0.048	1.21	1.25
VNN1	8876	Hs.12114	1825	0.001	0.049	1.01	1.1
AGA	175	Hs.207776	1826	0.001	0.049	1.24	1.27
STIM2	57620	Hs.135763	1828	0.001	0.049	1.29	1.24
LOC388152	388152	Hs.405809	1829	0.001	0.049	1.22	1.26
LOC100287465	100287465	Hs.507676	1830	0.001	0.049	1.34	1.27
P2RY13	53829	Hs.546396	1830	0.001	0.049	1.12	1.23
ABCC3	8714	Hs.463421	1830	0.001	0.049	1.28	1.13

RP, rank product; P, p-value ;FDR, false discovery rate; FC,fold change