Distinguishing fungal from bacterial infection: A Mixed Integer Linear Programming approach

Dissertation

zur Erlangung des akademischen Grades

doctor rerum naturalium (Dr. rer. nat)

vorgelegt dem Rat der Medizinischen Fakultät

der Friedrich-Schiller-Universität Jena

von João Pedro Leonor Fernandes Saraiva MSc,

geboren am 07.07.1980 in Mirandela, Portugal

Gutachter

1. Prof. Dr. Rainer König (Jena)

2. Prof. Dr. Reinhard Guthke (Jena)

3. Prof. Dr. Thomas Dandekar (Würzburg)

Tag der öffentlichen Verteidigung: 05.07.2018

To my beloved family and friends.

For their patience and support.

TABLE OF CONTENTS

ABBREVIATIONS 4

ZUSAMMENFASSUNG 7

SUMMARY 9

1. INTRODUCTION 11

1.1. Sepsis and septic shock 11

1.2. Fungal infections 12

1.3. Host response to infection 14

1.4. Lysosome 16

1.5. Pathogen identification 17

1.6. Classification 20

1.7. Feature Selection 24

1.8. Support Vector Machines 25

1.9. Mixed Integer Linear Programming (MILP) 30

2. Objectives 37

3. Materials and methods 38

3.1. Dataset Assembly 38

3.2. Data preprocessing 40

3.3. Support Vector Machine (SVM) Implementation 41

3.4. Machine learning and statistical analysis 44

3.5. Overall performance 46

3.6. Gene expression analysis and refinement of gene signatures 48

3.7. Experimental validation via quantitative reverse transcription PCR (RT-qPCR) 52

3.7.1. Monocyte isolation 52

3.7.2. Preparation of fungi and bacteria 53

3.7.3. Monocyte stimulation assay 53

4. Results 56

4.1. Discriminating infected from non-infected samples 56

4.2. Discriminating fungal from bacterial infected samples 61

4.3. *In silico* validation of the gene signature discriminating fungal from bacterial infected samples 65

4.4. Monocyte-specific fungal immune response 66

4.5. Real time quantitative reverse transcription PCR analysis of monocytes challenged with fungal and bacterial pathogens and cell wall representatives of each microorganism 71

5. Discussion 80

5.1. Combining classifiers improves consistency of gene signatures 80

5.1.1. Infected *versus* Non-infected 80

5.1.2. Fungal *versus* bacterial independent of cell population 81

5.1.3. Fungal *versus* bacterial dependent on cell population 82

5.2. Gene set enrichment analysis of combined classifiers 83

5.3. Lysosome pathway is enriched during fungal infection 86

5.3.1. Functional relevance of the differentially expressed lysosome-related genes 86

5.4. Functional relevance of differentially expressed non-lysosome-direct related genes 89

6. Conclusions and perspectives 91

7. Bibliography 92

8. Appendix 108

9. Ehrenwörtliche Erklärung 125

10. Acknowledgements 126

Abbreviations

Acc	Accuracy
AMM	*Aspergillus* Minimal Medium
BAG3	BAG family molecular chaperone regulator 3
CCL2	Chemokine (C-C motif) ligand 2
CCL3	Chemokine (C-C motif) ligand 3
CCR1	C-C Chemokine Receptor 1
CCR2	C-C chemokine receptor type 2
CCRs	C-C motif receptors
CD164	CD164 molecule
CD36	Scavenger receptor class B member 3
cDNA	Complementary DNA
CO	Carbon monoxide
CO2	Carbon dioxide
CV	Coefficient of variation
CXCL1	Chemokine (C-X-C motif) ligand 1
CXCL2	Chemokine (C-X-C motif) ligand 2
DA	Diagnostic accuracy
DAMP	Damage associated molecular pattern
DAVID	Database for Annotation, Visualization and Integrated Discovery
DCs	Dendritic cells
DNA	Deoxyribonucleic acid
EMBL	European Molecular Biology Laboratory
EV71	Enterovirus 71
FABP5	Fatty acid binding protein 5
FN	False Negative
FP	False Positive
Gb3	Globotriaosylceramide
GEO	Gene Expression Omnibus
GLA	Galactosidase A
HMOX1	Heme oxygenase 1

IFNγ	Interferon gamma
IL	Interleukin
IL12A	Interleukin 12 subunit alpha
IL12B	Interleukin 12 subunit beta
IL-1β	Interleukin 1 beta
IL6	Interleukin 6
IRAK2	Interleukin-1 receptor-associated kinase-like 2
KEGG	Kyoto Encyclopedia of Genes and Genomes
LP	Linear Programming
LPS	Lipopolysaccharides
MALP	Monocyte activating lipopeptide
MAPK	Mitogen-activated Protein kinase
MILP	Mixed Integer Linear Programming
MIP-2	Macrophage inflammatory protein 2
MOI	Multiplicity of Infection
mRNA	Messenger RNA
MyD88	Myeloid differentiation primary response gene 88
NaCl	Sodium chloride
NCBI	National Center for Biotechnology Information
NF-κB	Nuclear factor kappa-light-chain-enhancer of activated B ells
NF-κBIA	NFKB inhibitor alpha
NK	Natural killer
NLR	NOD-like receptors
NPC1	Niemann-Pick disease, type C1
NPV	Negative predictive value
NRQ	Normalized Relative Quantity
PAMP	pathogen associated molecular pattern
PBMCs	Peripheral Blood Mononuclear Cells
PMN	Polymorphonuclear Leukocytes
POL	Pairwise overlap
PPARG	Peroxisome proliferator-activated receptor gamma
PPIB	Peptidylpropyl Isomerase B

PPV	Positive predictive value
PRR	Pathogen recognition receptor
RNA	Ribonucleic acid
RPKM	Reads Per Kilobase of transcript per Million mapped reads
RPMI	Roswell Park Memorial Institute Medium
RQ	Relative quantity
RT-qPCR	Real time quantitative polymerase chain reaction
SCARB2	Scavenger receptor class B member 2
SCARF1	Scavenger receptor class F member 1
Sens	Sensitivity
Spec	Specificity
Spp	Species
SRA	Sequence Read Archive
SVM	Support Vector Machine
Th1	Type 1 helper T cell
Th2	Type 2 helper T cell
TLR	Toll-like receptors
TN	True Negative
TNFAIP3	Tumor necrosis factor alpha-induced protein 3
TNFα	Tumor necrosis factor alpha
TP	True Positive
USA	United States of America

Zusammenfassung

Das Immunsystem ist grundsätzlich für den Schutz des Wirts vor Infektionen verantwortlich. Bei gesunden Individuen kann das Immunsystem im Allgemeinen invadierte Pathogene bekämpfen und beseitigen. Im Gegensatz dazu haben immunsupprimierte Menschen ein erhöhtes Risiko an durch Mikroorganismen verursachten Infektionen zu erkranken. Dies kann bereits durch normalerweise kommensale Organismen geschehen.

In extremen Fällen dringen die invadierten Pathogene in den Blutkreislauf des Wirts ein und verursachen somit eine systemische Infektion mit schwerwiegenden Folgen. Systemische Infektionen können durch verschiedenste Organismen, wie Viren, Pilze oder Bakterien, ausgelöst werden.

Die adäquate Behandlung dieser Infektionen setzt eine schnelle Identifikation des invadierten Pathogens voraus. Das derzeitige Standardverfahren zur Detektion von Pathogenen sind Blutkulturen, die jedoch eine relativ lange Zeit bis zum Erhalt des Ergebnisses benötigen. Die Anwendung von *in situ*-Methoden führt zwar zu einer Identifizierung der pathogen-spezifischen Immunantwort des Wirts, bedarf jedoch häufig heterogener Biomarker, da die Variabilität der verwendeten Methoden und Materialien sehr groß ist. Die Analyse der Genexpressionsprofile von Immunzellen wird immer häufiger eingesetzt. Die Anwendung von Support Vector Machinen (SVMs) erlaubt die Unterscheidung zwischen zwei Infektionsarten. Der Vergleich von Genlisten unterschiedlicher und unabhängiger Studien zeigt einen hohen Grad an Inkonsistenz. Ursachen dafür können verschieden stimulierte Zellarten, verschiedene Pathogene oder anderer Faktoren sein. In dieser Arbeit wurden SVMs in Verbindung mit Gemischt Ganzzahliger Optimierung (Mixed Integer Linear Programming, MILP) angewendet, um konsistente Gensignaturen für die Differenzierung zwischen Pilz- und Bakterieninfektionen zu erstellen. Im ersten Schritt wurden Klassifikatoren verschiedener Datensätze für die Unterscheidung von gesunden und infizierten Proben mittels der Zwangsbedingung, gemeinsame Merkmale auszuwählen, kombiniert. Nach der Etablierung der Methode und der Verbesserung der Konsistenz der Gensignaturen verbessert

wurdeeine generische Gensignatur, die zur Diskriminierung von bakteriellen und fungalen Infektionen, unabhängig von der Art der untersuchten Leukozyten oder dem experimentellen Ablauf, ist, entwickelt. Die erstellte Liste dieser Biomarker zeigte im Vergleich zu Einzel-Klassifikatoren eine um 42% höhere Konsistenz und sagte die infektionsverursachende Pathogenart für einen ungesehenen Datensatz mit einer durchschnittlichen Genauigkeit von 87% voraus. Zuletzt wurde die jeweilige Fokussierung auf ähnliche Leukozytenkompositionen, die die Gensignatur signifikant verändert, überprüft. Wie erwartet waren immun- und inflammatorisch-relevante Signalwege wie beispielsweise die Signalwege für NOD-like und Toll-like Rezeptoren angereichert. Erstaunlicherweise zeigte die Gensignatur des kombinierten Klassifikators ebenfalls eine Anreicherung des lysosomalen Signalwegs, welcher nicht in den Einzel-Klassifikatoren vorkam. Des Weiteren zeigen die Ergebnisse, dass der Lysosomensignalweg nach einer Pilz-Infektion spezifisch in Monozyten induziert ist. Die Analysen von relevanten Genen des lysosomalen Signalwegs mittels quantitativer PCR bestätigte deren erhöhte Genexpression in Monozyten während einer Pilzinfektion.

Im Endergebnis erhöhte der neuenkombinierte Klassifikator die Konsistenz der Gensignaturen im Vergleich zu den Einzel-Klassifikatoren und zeigte darüber hinaus auch Signalwege von Leukozyten, wie beispielsweise Monozyten, auf, die einen geringen Anteil an der Blutzusammensetzung haben.

Summary

The immune system is responsible for protecting the host from infections. In healthy individuals, this system is generally able to fight and clear any pathogen it encounters. In turn, people with a compromised immune system are at higher risk of acquiring infections from microorganisms which are usually commensal in nature. In extreme cases, the invading pathogen can enter the blood stream leading to a systemic infection and ultimately severe consequences. Blood stream infections can be caused by several pathogens such as viruses, fungi and bacteria. Delivery of appropriate treatment requires rapid identification of the invading pathogen. The current gold standard for pathogen identification relies on blood cultures which require a long time to produce a result. The use of *in situ* experiments attempts to identify pathogen specific immune responses but these often lead to heterogeneous biomarkers due to the high variability in methods and materials used (e.g. stimulated cell-type, pathogen strain, culture conditions of the pathogen and experimental protocols). The analysis of gene expression of immune cells during infection has increased over time. Support Vector Machines (SVMs) allow using gene expression patterns to discriminate between two types of infection. Comparing gene lists from independent studies shows a high degree of inconsistency. To produce consistent gene signatures, capable of discriminating fungal from bacterial infection, SVMs using Mixed Integer Linear Programming (MILP) were employed. Firstly combined classifiers from several datasets by joint optimization with the aim to distinguish infected from healthy samples were used. Having employed this method and demonstrated the improvement in consistency of the produced gene signatures the next aim was to discover a generic gene signature that could distinguish fungal from bacterial infections irrespective of the type of the leukocyte or the experimental setup. The produced biomarker list showed an increase in consistency of 42% when compared to single classifiers, and predicted the infecting pathogen on an unseen dataset with an average accuracy of 87%. Lastly, the focus was to determine whether restricting the analysis to data with similar leukocyte compositions would significantly alter the gene signature. As expected,

pathways related to immunity and inflammatory processes such as NOD-like receptor signaling and Toll-like receptor signaling were enriched. Surprisingly, restricting the analysis to datasets comprised of peripheral blood mononuclear cells (PBMCs) and monocytes, the gene signature obtained from the combined classifier also showed an enrichment of genes from the lysosome pathway that was not shown when using independent classifiers. Moreover, the results suggested that the lysosome pathway is specifically induced in monocytes. Real time qPCRs of the lysosome-related genes confirmed the distinct gene expression increase in monocytes during fungal infections.

In conclusion, the combined classifier approach increased the consistency of the gene signatures, compared to single classifiers. This was shown in both discriminating infected from healthy samples as well as in discriminating fungal from bacterially infected cells. Additionally, the combination of classifiers "unmasked" signaling pathways of less-present immune cell types, such as monocytes, when restricting the analysis to only PBMCs and monocyte stimulated datasets.

1. Introduction

1.1. Sepsis and septic shock

Sepsis is a medical condition in individuals with a compromised immune system. Efforts have been made to clearly define sepsis and septic shock. A task force convened by the Society of Critical Care Medicine and the European Society of Intensive Care Medicine was created to tackle this issue (Singer et al. 2016). They defined sepsis as "life-threatening organ dysfunction by a dysregulated host response to infection". Additionally, they define septic shock as "a subset of sepsis in which particularly profound circulatory, cellular, and metabolic abnormalities are associated with a greater risk of mortality than with sepsis alone" (Singer et al. 2016). Patients with sepsis are characterized by having low blood pressure, fever, rapid breathing and altered mental status among others (Levy et al. 2003) In the clinics, diagnosis is carried out by performing blood tests to identify infecting pathogens, organ function and oxygen availability (Rhodes et al. 2017). Additionally, a rapid form of identifying patients with suspected sepsis consists of measuring the qSOFA score which is based on three criteria: blood pressure, breath rate and mental status (Vincent et al. 2009). A single point is assigned to each criteria if the following values are not met: blood pressure ≤ 100 mmHG, breath rate ≤ 22 breaths/min, and altered mentation < 15 (Glasgow coma scale). A qSOFA score ≥2 indicates that a patient is suspected of having sepsis with organ dysfunction with higher risk of poor outcome (Singer et al. 2016). Common treatment relies on the administration of broad-spectrum antibiotics, intravenous fluids to normalize blood pressure as well as insulin to maintain stable blood sugar levels and other supportive procedures (Vincent et al. 2009) .

The invasion of microorganisms into sterile parts of the human body, such as the blood stream, can in general lead to sepsis and septic shock if not treated promptly (Lever and Mackenzie 2007). In a study by Vincent and colleagues (Vincent et al. 2009) the most common source of infections present in patients in intensive care units (ICUs) where shown to be from gram negative bacteria (62%), gram positive bacteria (47%) and fungal pathogens (19%). Sepsis is

among the top 10 leading causes of death in the United States of America (USA) (Jawad et al. 2012). Not many studies have been performed internationally to determine mortality rates, incidence and prevalence but the few that exist, nationwide, refer mortality rates as high as 30% and 80% for sepsis and septic shock, respectively (Jawad et al. 2012). Additionally, a study carried out in the USA also highlighted the elevated costs (20.3 billion US dollars) associated with sepsis (Torio and Andrews 2013). High mortality rates are correlated with the lack of effective treatment and diagnosis. Therefore, it is important to develop novel methods that can rapidly identify the invading microorganism so the adequate treatment can be employed. The use of biomarkers capable of identifying the underlying source of infection would improve substantially the time required for an accurate diagnosis (Bloos and Reinhart 2014).

1.2. Fungal infections

Over the last few decades, the interest in invasive fungal infections has increased due to the threat and mortality rates they pose to immunocompromised individuals (Shoham and Levitz 2005, Horn et al. 2012). The increase in immunocompromised patients, those undergone invasive medical procedures or those treated with broad-spectrum antibiotics, has greatly increased the risk of acquiring fungal infections (Shoham and Levitz 2005, Romani 2011, Brown et al. 2012, Netea et al. 2015). The increase of fungal induced sepsis shows a considerable increase in morbidity and mortality, with *C. albicans* accounting for 10 to 15% of fungal sepsis in the United States of America (Delaloye and Calandra 2014). Fungal species such as *C. albicans* usually are commensal and colonize the mucous membranes and skin of the host, whilst others such as *A. fumigatus*, are ubiquitous molds usually taken up by the host via inhalation (Shoham and Levitz 2005). Virulence factors such as α-(1,3)-glucan, melanin, glucuronoxylomannan, β-glucans and glycosphingolipids, among others, are highly involved in fungal pathogenicity (Hogan et al. 1996) . The role of the fungal cell wall is of great importance in

pathogenicity because it is the structure that establishes first contact to the host carrying antigenic determinants and establishing cross-talk between the human hosts and invading fungi. In healthy individuals, these interactions usually lead to a mounting of an effective immune response (Ruiz-Herrera et al. 2006). However, in individuals with a compromised immune system or whose tissue barriers are disrupted, these fungal organisms may become pathogenic and, in some cases, cause systemic infection possibly leading to the death of the patient (Netea et al. 2015). *Candida* species (spp)., *Aspergillus* spp. and *Cryptococcu*s spp. are among the most frequent causes of invasive fungal infections with *Candida albicans* being ranked fourth in the United States of America as the main cause of nosocomial bloodstream infections (Brown et al. 2012). *C. albicans* and *Aspergillus fumigatus* have been shown to be the most frequent causes of these types of infections in organ transplant patients (Pappas et al. 2010). Significant increased mortality of septic shock patients was observed if arising from candidemia (Kollef et al. 2012). Candidemia occurs when *Candida* species enter the blood stream causing systemic infection (Garey et al. 2006). Patel and co-workers (Patel et al. 2009) displayed a significant increase in survival if appropriate antifungal therapy was administered at the early stage of *Candida albicans* induced septic shock. In a cohort study of critically ill surgical patients with severe sepsis in China, Xie and co-workers (Xie et al. 2008) showed that more than 28% of the patients were identified as having invasive fungal infections. Moreover, out of the 100 identified fungal strains *C. albicans* was the most prevalent fungal species (58%). The authors also demonstrated that invasive fungal infections were associated with higher mortality rates, hospital costs and prolonged stays in the intensive care unit as well as hospital stay in general. However, the toxic effect of antifungals on the host's cells hampers the development of new antifungal therapies due to protein homology and similar protein synthesis between human and fungal cells (Shoham and Levitz 2005). Although the proportion of fungal induced sepsis is less when compared to bacterial induced sepsis, the incidence of fungal infections in septic patients is on the rise (Delaloye and Calandra 2014). Since time is of the essence in the treatment of sepsis, more rapid and precise

diagnostic methods are required in order to deliver the appropriate therapy (antibiotic *versus* antifungal).

1.3. Host response to infection

The human immune system is highly adaptable and a potent mechanism for the clearance of pathogens. The complexity of this system is closely linked to the interconnection of the multitude of organs, cells and pathways and how they tailor immune responses to infecting agents (Nicholson 2016). The overall immune response towards infection has been reviewed extensively (Mogensen 2009), but it is consensual that innate immunity is the first line of defense against infection after the physical barriers are overcome (Rivera et al. 2016) .

The innate immune system is crucial in the early identification and clearance of the invading pathogen and, in later stages of infection, of promoting additional adaptive immune responses. Innate immunity relies on the recognition of pathogen associated molecular patterns (PAMPs) (Mogensen 2009). The latter are identified by pattern recognition receptors (PRRs), present either on the cell surface of immune cells, such as macrophages and dendritic cells (DCs), or in the cytoplasm and trigger pro-inflammatory responses and subsequent activation of downstream signaling cascades (Mogensen 2009). The most studied types of PRRs are Toll-like receptors (TLRs) and NOD-like receptors (NLRs). TLRs are usually present on the cell membrane and are capable of recognizing distinct PAMPs originated from very different pathogens (e.g. viruses, bacteria, fungi) (Delneste et al. 2007, Mogensen 2009, Arias et al. 2017). TLRs recognize lipids (e.g. TLR1, TLR2 and TLR4), nucleic acids (e.g. TLR3, TLR7, TLR9) and proteins (e.g. TLR5) (Gay et al. 2006, Trinchieri and Sher 2007, Barton and Kagan 2009, Mogensen 2009). In turn, NLRs are usually located in the cytoplasm of the cell and play a key in the regulation of the host immune response (Franchi et al. 2009). The interplay and combination of TLRs and NLRs can induce general immune responses such as inflammation but each of them alone provides limited information on what pathogen is the cause of infection. It has been shown that some TLRs, such as TLR2, can recognize both lipopolysaccharide (LPS) and zymosan (which represent cell wall

components of gram negative bacteria and fungi, respectively) (Fritz et al. 2006, Franchi et al. 2009, Mogensen 2009), which limits its use to help discriminating fungal from bacterial infections. The ability to mount an adequate and effective innate immune response relies on the efficient activation of, but not exclusively, neutrophils and monocytes and each account for approximately 62 and 5.5 % of the total number of leukocytes in the blood, respectively (Bhushan 2002). Both have been identified as important antifungal effector cells (Shoham and Levitz 2005). Neutrophils are the main effector cells in fighting *C. albicans* and *A. fumigatus* infections (Traynor and Huffnagle 2001). Monocytes not only fight infections but can also differentiate into other immune cells such as macrophages and DCs which, in turn, are capable of phagocytic activity and provide the necessary stimulus to cells of the adaptive immune system (Shi and Pamer 2011). Monocytes express most PRRs related to fungal (Netea et al. 2008) and bacterial infections (Hessle et al. 2005) but studies have shown that the type of infection will trigger different signaling cascades. Monocytes take a pivotal role in the early recognition of candidiasis, a non-systemic infection caused by any *Candida* species (Netea et al. 2008, Klassert et al. 2014, Ngo et al. 2014).They have been suggested as the most effective mononuclear leukocyte in the killing of *C. albicans* (Netea et al. 2008).

Immune cells exist in the human body in different abundancies. It is possible that the impact of immune cells that are less represented in the blood such as monocytes (approximately 5%) is not well characterized due to the presence of the more abundant leukocytes such as neutrophils and lymphocytes (approximately 62%). Studies have shown that the expression of several genes is immune cell type-specific (Wong et al. 2011, Allantaz et al. 2012, Gardinassi et al. 2016). Other studies have also shown that genes can activate distinct molecular pathways depending on the cell population (Didonna et al. 2016). Cell-type specific gene expression studies have demonstrated that the relative proportion of each leukocyte type invariably has an impact on the global gene expression profile (Palmer et al. 2006). Whilst it is vital to understand how our immune system responds to infection in general, it is also crucial to understand the pathogen-specific host immune responses both dependent as well as independent of leukocyte type and cell population. The ability to clearly identify

what type of pathogen (e.g. fungal or bacterial) allows the employment of more tailored treatments and administration of specific drugs to eliminate the infecting pathogen and thus, improving patient outcome.

1.4. Lysosome

Lysosomes were first discovered in the 1960's by Christian de Duve (Sabatini and Adesnik 2013). These organelles play an essential role in the degradation of extra and intra-cellular components (Schwake et al. 2013). Among others, lysosomes are highly involved in functions such as antigen presentation, innate immunity, autophagy, cholesterol homeostasis, cell signaling and death (Saftig 2006, Parkinson-Lawrence et al. 2010). In innate immunity, lysosomes play an important role by providing the necessary enzymes for pathogen degradation. In addition, lysosomes are also involved in the regulation of inflammatory responses (He et al. 2011). Malfunction of the lysosome leads to several disorders such as Niemann-Pick disease type C and Fabry´s disease (Vellodi 2005). Briefly, individuals with Niemann-Pick disease type C display enlarged spleen but also progressive neurological disease such as dementia (Vanier and Millat 2003). In the case of Fabry´s disease, individuals present a dysfunctional metabolism of sphingolipids which can lead to kidney and heart complications (Kint 1970). Pathogens such as bacteria (Koo et al. 2008), fungi (Kaposzta et al. 1999, Davis et al. 2015) and viruses (Wei et al. 2005) are usually engulfed by phagocytes via phagocytosis. Once phagocytes fuse with lysosomes – originating the so-called phagolysosomes, the enzymes required for pathogen degradation are released. The indigestible material is later released for disposal into the interstitial fluid and blood for recycling or for promoting additional immune responses such as apoptosis (Colbert et al. 2009). Certain pathogens have however, developed strategies to resist the process of degradation and thus evade lysosomal influence in the immune response (Nicholson 2016). *Cryptococcus neoformans*, an opportunistic fungal pathogen, was shown to be able to avoid degradation even when engulfed by macrophages by damaging the lysosome (Kaposzta et al. 1999, Davis et al. 2015). This study showed a correlation between *C. neoformans* replication rates and lysosome damage

which highlighted the benefits of an adequate functioning of the lysosome. Contrastingly, a rapid recruitment of lysosomal compartments to macrophages infected with *C. albicans* demonstrated to be beneficial for the pathogen (Kaposzta et al. 1999). The authors suggested that the acidic environment promoted by the fusion of lysosome to phagosomes promoted the formation of yeast germ tubes allowing the penetration of macrophages and subsequent survival of the fungi (Kaposzta et al. 1999). Lysosomal enzymes such as β-hexosaminidase have also been shown to play an important role in the control of bacteria such as *Mycobacterium marinum* (Koo et al. 2008). The authors showed that the secretion of this enzyme restricted *M. marinum* intracellular growth even when phagosome-lysosome fusion was prevented. Inhibition of phagosome-lysosome fusion, which prevents excessive acidification of the environment, has been suggested as a resistance mechanism for *M. tuberculosis* to avoid killing by macrophages (Vandal et al. 2009).

Lastly, the inhibition of the lysosome has also been shown to enhance human immunodeficiency virus type 1 (HIV-1) infections (Wei et al. 2005)., which further highlights the importance of the lysosome in the clearance of viruses.

The lysosome plays an important role in the clearance of infection. However, both within and between groups of pathogens the regulation and effect of lysosomal activity can have opposite effects. Understanding how these organelles are activated and how they are expressed by different pathogens would provide useful information on pathogen discrimination.

1.5. Pathogen identification

According to the Biomarkers Definitions Working Group from the National Institutes of Health, biomarkers can be defined as "a characteristic that is objectively measured and evaluated as an indicator of normal processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention" (Atkinson A.J. et al. 2001).

Examples of biomarkers range from medical signs, such as blood pressure and fever, to molecular interactions and gene expression alterations (e.g. in

response to a specific treatment or infection). However, biomarkers must be quantifiable in order to guarantee reproducibility (Strimbu and Tavel 2011).

Methods for detecting and identifying pathogens can either be culture or non-culture based. The first consists of growing the microorganism in culture media under controlled conditions whilst the latter consists of detecting and measuring antigens or microbial products (Bursle and Robson 2016) that reflect the presence of a certain pathogen.

Currently, culture based methods such as blood cultures are the "gold standard" for the identification of pathogens in the blood stream. However, this approach can take several days to identify the infectious agent (Kirn and Weinstein 2013) or even be unable to identify the microorganism if the culture media for the invading pathogen is not the most appropriate (Chan and Gu 2011). Additionally, the required number of pathogens in the blood sample might not be sufficient to provide a positive test (Cunnington 2015).

As stated previously, an alternative form of diagnosis would be based on non-culture based methods measuring the host's immune cells unique response to a specific type of infection or by measuring the levels of antibodies and antigens in the blood. Current non-culture based biomarkers for fungal infections include galactomannan, antimannan and β-D-glucan since these are present in the cell wall of fungal pathogens (Chan and Gu 2011, Patterson 2011). Polysaccharide mannans represent more than 7% of the dry weight of *C. albicans* and are highly immunogenic (Bursle and Robson 2016) which demonstrates the usefulness of measuring such compounds. However, the variability of diagnostic accuracy (DA) across different experimental setups present a challenge to accurately identify the pathogen (Chan and Gu 2011).

Polymerase Chain reaction (PCR) is also used as a diagnostic method for pathogen identification. Briefly, PCR is a method for amplifying DNA (i.e. generating many copies of a section of DNA). An advantage of using this technique is that it does not require a great amount of initial DNA. Rapid identification of a pathogen by this method can take up to one working day (Bloos and Reinhart 2014). The use of PCR-based methods for diagnosis of infectious diseases has increased over the years due to its broad-spectrum detection of pathogens, relatively rapid procedure and cost when compared to

the gold standard methods such as blood cultures (Yang and Rothman 2004, Maurin 2012). Despite overcoming some of the limitations of blood cultures such as decreased specificity and time required for pathogen identification, the detection of fungal pathogens via this method is, however, still challenging. PCR-based methods do not distinguish between alive or dead cells since it only detects the presence of DNA or RNA in the blood (Soejima et al. 2008). Further, the fungal cell wall prevents their efficient lysis impeding the release of DNA (Khot and Fredricks 2009). Fungal spores, due to their ubiquitous nature in the air and environment, can lead to false positives either by contaminating reagents or during any step in the whole procedure (Khot and Fredricks 2009). Inversely, the generation of false negatives also has to be considered due to PCR detection limits (i.e. the minimum number of copies of DNA per PCR required for detection) (Khot and Fredricks 2009). Additionally, the sequences for the genes of interest have to be known beforehand (Lorenz 2012).

The use of transcriptomic data (i.e. data generated from measuring the abundance of mRNA transcripts in samples from the host with or without any stimulation) has increasingly been used to identify novel biomarkers (Saraiva et al. 2017, 2016, Dix et al. 2015, Linde et al. 2016). Generating transcriptional profiles are mainly achieved through DNA microarrays (Quackenbush 2006) or RNA sequencing (Wang et al. 2009). DNA Microarrays are the most common method in gene expression profiling but as RNA sequencing technology (RNA-Seq) becomes increasingly available so could the method. Contrastingly to microarrays, high-throughput DNA sequencing methods such as RNA-Seq can directly determine the sequence of cDNA, present very low noise, have a high range for detection of gene expression level, require a low amount of RNA and have a relatively low cost for mapping transcriptomes of large genomes (Wang et al. 2009). It has been shown that RNA-Seq outperforms microarrays in the detection of low abundant transcripts, identification of genetic variants as well as avoiding the issues related to probe cross-hybridization and limited detection range of individual probes that exist in microarrays (Bursle and Robson 2016).

Irrespective of the method used for measuring gene expression, changes in the host's cells phenotype during infection is often correlated to changes in gene expression (Jenner and Young 2005). This change can either be

pathogen and cell-type independent (general response) or pathogen and/or cell-type specific (specialized response). Dix and co-workers used a machine learning based approach and identified genes with which bacterial from fungal infections could be distinguished as well as infected from non-infected samples in whole-blood cell cultures (Dix et al. 2015). A transcript for the S100 calcium-binding protein (*S100B*) was identified as a biomarker gene for identifying invasive aspergillosis in hematological patients (Linde et al. 2016). Several *in situ* expression profiling studies have been undertaken (Zaas et al. 2010, Smeekens et al. 2013, Dix et al. 2015) to gain insight into the distinct gene regulation of the host response of immune cells after fungal and bacterial infection. However, the gene lists that were generated by these high throughput methods lacked consistency when comparing the results across studies from different labs. In this context, consistency is defined as, for the same infection similar biomarkers or gene signatures are identified in data, even if generated in different labs or at different conditions. Hence, even such controlled cell culture studies show high heterogeneity. This may be due to the different laboratory settings like different multiplicity of infection (MOI, ratio of number of pathogen cells to the number of immune cells of the host), different pathogen strains and species, different treatments (heat killed, living pathogens, surface molecule extracts such as lipopolysaccharides or glucans), or different time points of sample extraction after infection. Still, the major aim of all these approaches is to find a gene signature, with which the infection can be identified, independent of the specific settings in the laboratories, to improve diagnosis in patients.

1.6. Classification

In the field of machine learning and statistics, classification problems are considered as instances of supervised learning. The general goal is to identify to which class a sample belongs to. Microbiologically, the data used for classification would comprise instances (samples) and features (e.g. transcript, protein or metabolite measurements). It is considered a supervised machine learning approach because the learning algorithm trains on data whose labels

for the samples are known. Used data can be of binary, categorical or continuous nature (Kotsiantis et al. 2006, Maglogiannis 2007). The classifier will, based on the variables, predict to which class (label) new "unseen" data belongs to. In simple terms, the training of the classifier "studies" the expression pattern of the data, usually whose labels are known, after which it will use the learned information and predict the labels of samples of an unknown dataset based solely on the features. The classifiers' evaluation is of critical importance and usually based on the accuracy of prediction (number of correct predictions divided by the total number of predictions) (Kotsiantis et al. 2006). Ideally, the classifier uses one dataset for training and an independent, unused dataset for testing to avoid overfitting. Overfitting usually occurs when the model is too complex due to the excess of parameters compared to the number of observations. In other words, the model is excessively tailored to the training data which leads to poor generalization. Measuring the complexity of the model can be determined by the Vapnik-Chervonenkis (VC) dimension. A large VC dimension represents a more complex model. In turn, the more complex a model, the better it can separate the data points in the training set. To get a better understanding of VC-dimension the concept of shattering must first be elucidated. A set of classifiers C (e.g. set of linear classifiers) shatters n instances if for each of the possible class labels (class 1 and class 2) there exists at least one classifier from our set of classifiers ($c \in C$) that can separate the instances into their classes. For n instances, the number of possible class combinations is 2^n. The maximum number of instances n which can be used to separate the classes in a data set is considered the VC dimension. Consider a dataset X composed of three instances in a two dimension space. The VC-dimension is equal to 3 since we can find at least one set of 3 instances all of whose classes can be separated by a line (Figure 1).

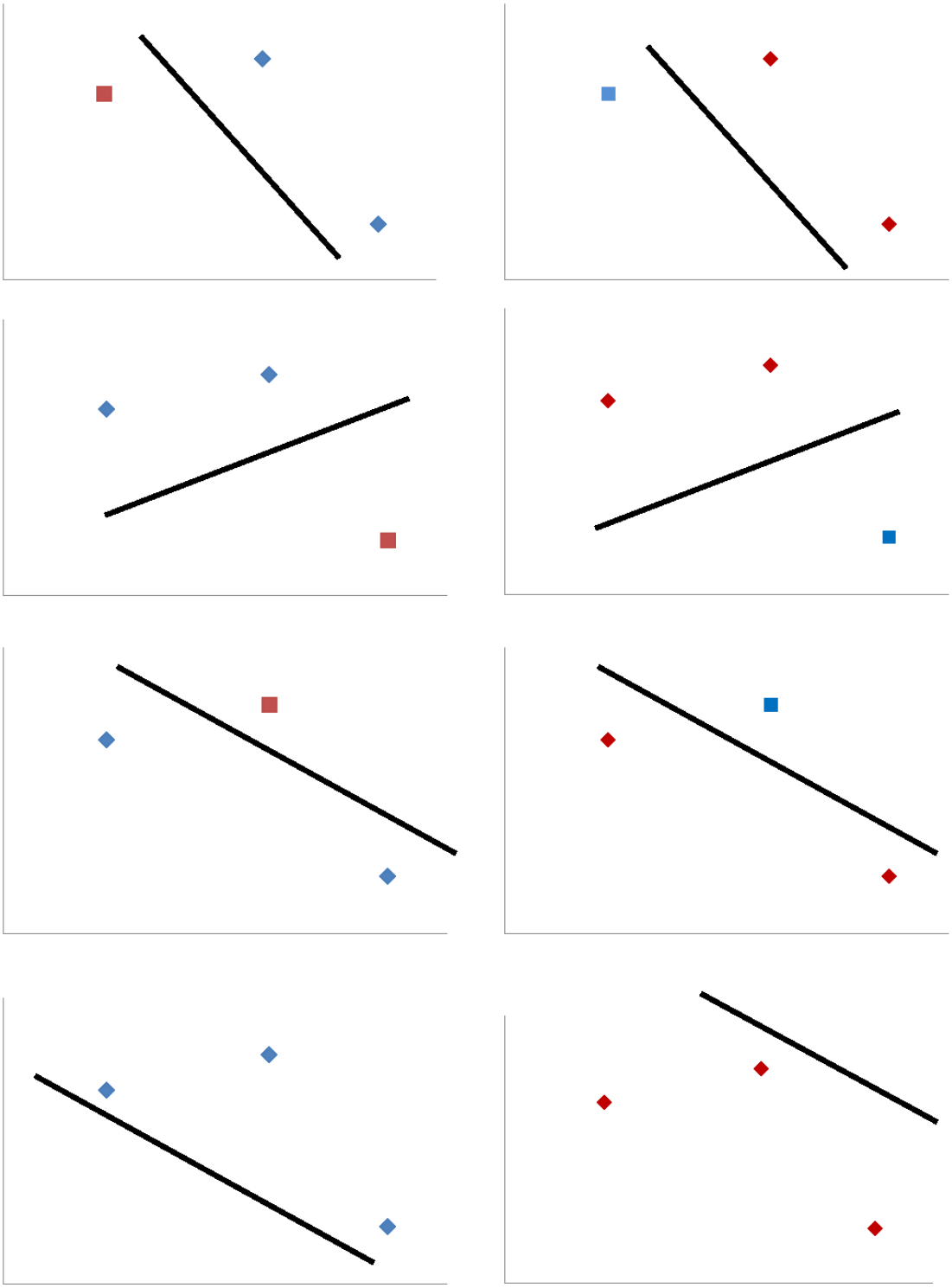

Figure 1 - The line that separates the n instances (number of samples) in the dataset X with all possible class assignments (red – class 1, blue – class 2).

However, when using a margin to separate classes the VC-dimension is calculated by,

$$VC < \frac{D^2}{w^2} + 1$$

where D is the diameter of the sphere in which the instances exist and w is the margin width (Vapnik 1995)(Figure 2).

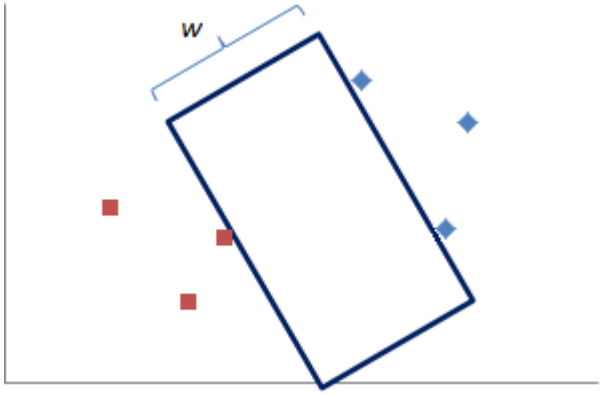

Figure 2 – Linear classifier with margin width w that separates both classes (red – class 1, blue – class 2) in the data

To note that as the margin width increases, the VC-dimension decreases and, consequently, the model´s complexity. Generalization refers to the capability of the trained model to be applied to data not used during the learning process. One should be aware that performance of the model is determined by its predictive capabilities on unseen data. Therefore, having a high performance on the training data as the result of overfitting, can lead to a poor performance on the test data. The higher the generalization the better it will perform when making predictions on new data.

Nevertheless, it is also possible to build and test classifiers using a single dataset. Methods for dealing with this scenario exist such as *k*-fold cross-validation, leave-one-out cross-validation and Monte Carlo cross-validation. In *k*-fold cross-validation the original data is split into *k* parts of equal size. During each iteration one part is used for testing whilst all others are used for training of the classifier. Using a leave-one-out cross-validation, the number of parts into

which the data is divided equals the total number of observations. As in *k*-fold cross-validation, the number of iterations will be equal to the number of observations and each part is used once as a testing set and the remaining parts for training. The Monte Carlo method differs from the above mentioned due to its independence of the number of iterations. This method relies on a simple random sampling. Simple random sampling consists of selecting every individual randomly by chance, with each one of these having the same probability of being selected. However, this might result in the overlap of sample subsets during the training procedure since in each iteration the samples have the same probability to be chosen irrespective of their selection in the previous iteration. Independently of the method chosen for cross-validation, the general aim is to decrease the problem of overfitting and increase generalizability (Kotsiantis et al. 2006, Maglogiannis 2007).

Many methods of classification have been developed over the years such as decision trees, neural networks, k-nearest neighbor, random forest and Support Vector Machines (SVMs) (Fernández-Delgado et al. 2014). In the present work SVMs were used and are explained more in detail in the subsection 1.8

1.7. Feature Selection

Omic data generated from high throughput technologies such as microarrays or RNA-Seq is highly dimensional due to the measurement of the expression levels of thousands of genes. Feature selection is a method that aims to identify the most relevant features in the data and exclude the irrelevant ones. To note that feature selection does not change the variable representation but basically selects a subset of them. Thus, by identifying the most relevant features, the model performance and construction speed are improved (Saeys et al. 2007).

In classification problems, feature selection methods mainly exist in three forms filter, wrapper and embedded, and have been nicely reviewed by Sayes and colleagues (Saeys et al. 2007).

Briefly, filter methods select feature subsets by calculating relevance scores (e.g. based on variance of the features) rather than the error rates and exclude the lowest ranking ones. These features are then used as the input for the

classifier. What one can note from this description is the independence of the feature selection process from the classification step. In other words, the feature subset is not constrained to any specific prediction model. This also allows the use of the feature subset in different classifiers. However, the generalizability of the feature subset from filter methods usually results in lower prediction performances when compared to other feature selection methods (Saeys et al. 2007).

Wrapper methods extract subsets of features from the available search space and then test how well they perform in the classification step. It is considered a wrapper method exactly because the search for the subsets of features is dependent on the classification model. Genetic algorithms are an example of a wrapper method. As an example consider the following: A dataset consists of the gene expression of 10000 genes in 20 samples (divided into two classes). The objective function is to select the genes whose expression best discriminates between these two classes. First, a random amount of *k* groups of *n* genes are randomly assigned. Next, the fitness (i.e. the capacity of the genes in the group to discriminate the two classes) of each *k* group is calculated. Elements *n* of different groups are then exchanged and the groups are again evaluated. At the end of this iterative process (which can be decided by the allowed number of cross-over of genes between groups) the fittest group has the highest probability of being selected.

Lastly, embedded methods are similar to wrapper methods since the search for subsets of features is dependent on the performance of the classifier. However, in this case, feature selection is performed intrinsically as a step during the training of the classifier (e.g. adding a penalty if the number of features is too high in order to obtain a certain performance value) (Saeys et al. 2007). Examples of embedded methods include decision trees and weight vector usage of SVMs (Chow et al. 2001, Guyon et al. 2002, Saeys et al. 2007).

1.8. Support Vector Machines

The method of Support Vector Machines (SVMs) is a supervised machine learning method broadly used in biological context. Besides the possibility to

build not only linear classifiers but also nonlinear ones through use of the kernel trick (explained further below), one of its main advantages is its generalizability by implementation of a margin. One common application of SVMs is the classification based on gene expression profiles. In simple terms and in the context of data comprised of infected and healthy samples, an SVM will "study" the gene expression pattern and determine how well the expression of certain genes can separate the samples according to their infection status. The better these features can be used to predict the status of an unknown sample the higher the generalization of the classifier. Noble (Noble 2006) stated that only four basic concepts were required to understand SVMs: (i) separating hyperplane, (ii) maximum-margin hyperplane, (iii) soft margin and (iv) kernel function.

As an example, gene expression data of samples from two conditions (infected and healthy) are used to "train" the SVM to identify the expression pattern that best differentiates the two classes. If the expression pattern of certain features (genes) is discriminative for the two classes then it should be able to correctly classify new samples whose status (infected or healthy) is unknown based on their expression patterns.

The higher the number of features the higher the probability that the SVM might find a feasible solution that is capable of separating the data points into two classes.

(i) Considering that we have linearly separable data composed of two conditions (Figure 3), the expression values of the identified features during the classification problem can then be used to predict the status of an unknown sample (blue point highlighted by the blue arrow in Figure 3). For this, one just needs to see in which side of the line the selected features expression values of the unknown sample falls (in this case in "green" group).

However, data obtained from high-throughput technologies (e.g. microarrays, RNA-Seq) generates gene expression values for large amounts of features (genes). This increase in the number of features results in a higher dimension space and a plane is required to separate the features (separating hyperplane).

The optimal hyperplane is defined by Vapnik (Vapnik 1982) as the "linear decision function with maximal margin between the vectors of the two classes".

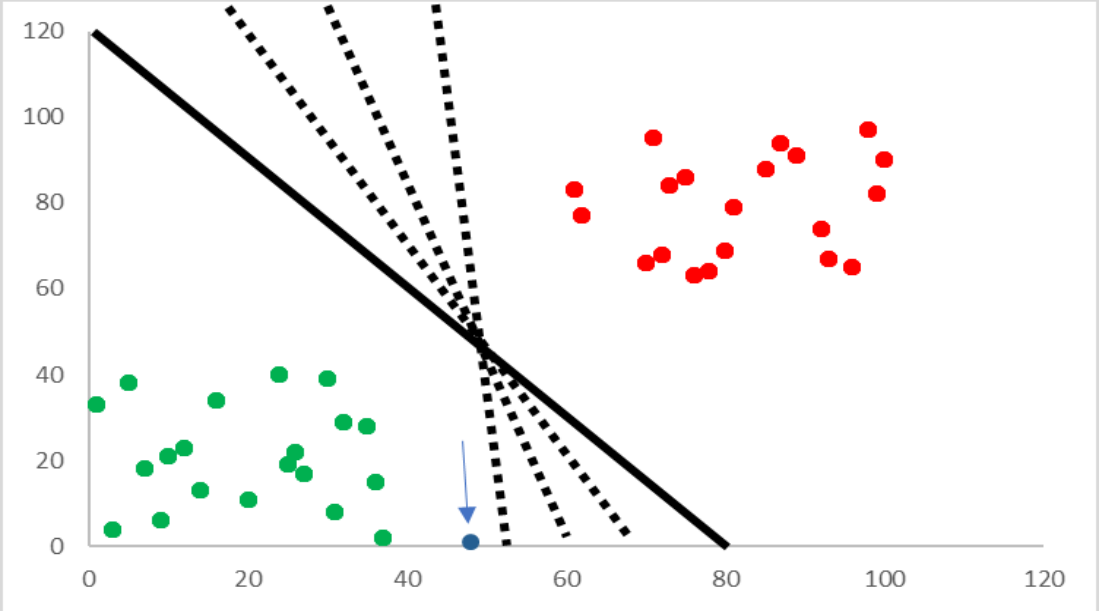

Figure 3 – Two-dimensional representation of expression profile of 2 classes (A: red and B: green) where each dimension is the expression value of a given gene. Blue dot represents the new data.

In this study, this is between vectors of infected and healthy samples, and of fungal and bacterial samples. Such hyperplanes can easily be constructed by considering very few samples from the training data.

(ii) In two dimensions, the classifier will identify the separating line that distinguishes the samples based on their expression profiles (black solid line in Figure 3). However, many lines may exist that achieve that goal (black dotted lines in Figure 3). In the case of SVMs, the selected line will be the one that maximizes the distance *w* from any of the expression profiles (Figure 4).

(iii) Ideally, all data could be divided into two groups just by a straight line. Unfortunately, this is not possible in some cases since no line (i.e. hyperplane) might exist that separates the two classes (Figure 5).

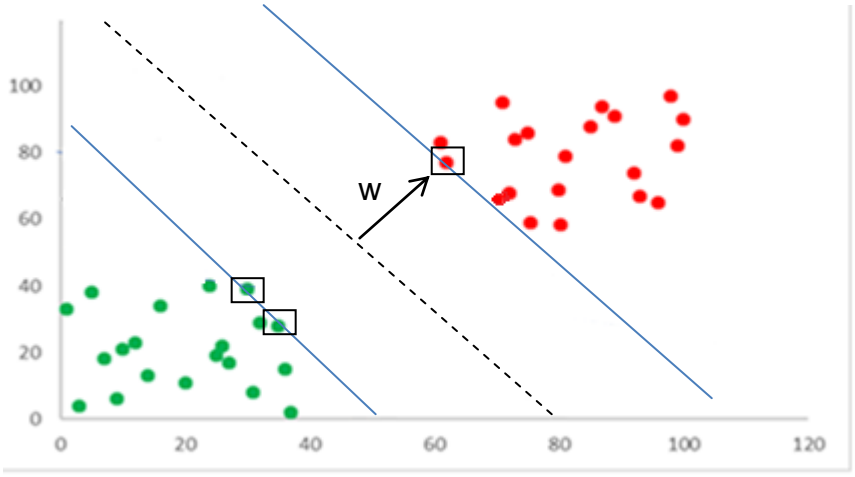

Figure 4 - The maximum-margin hyperplane is defined by the space that adopts the maximal distance *w* from any of the points (in this case marked in black boxes) to the separating line (dotted line).

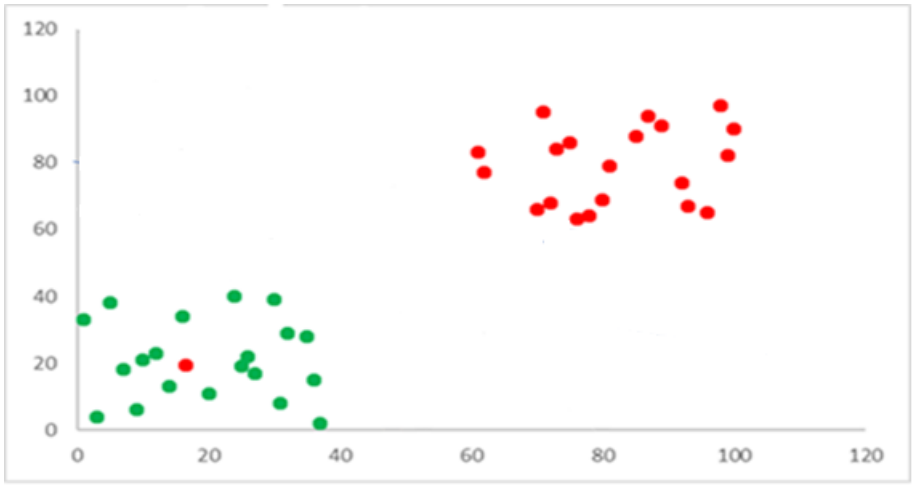

Figure 5 – Linearly inseparable data. The presence of one red point in the green cluster no longer allows the data to be linearly separated.

SVMs circumvent this by allowing the introduction of misclassifications. The space in which the SVM allows samples to wrongfully be placed is called a soft margin. The larger the margin, the more stable it will be when adding new data. In the case of Figure 5, the allowance of one misclassification would result in the same maximum-margin hyperplane shown in Figure 4. One should be aware that there exists a trade-off between the number of allowed misclassifications (and the size of the margin) and the degree of confidence that the classifier will identify new samples accurately. This cost function

controls the relative weight between maximizing the margin and degree of confidence that new samples will be correctly classified. Increasing the number of misclassifications might result in a feasible hyperplane and even on an increase of the optimal margin but this could also lead to worse performance and generalization and hence pay a cost. Inversely, decreasing the number of allowed misclassifications might result on a smaller margin hyperplane but improve the classification of the training samples.

Finally, in cases where the separation of the data points is not possible by a straight line (Figure 6), (iv) kernel functions are employed. The kernel functions, in simple terms, projects the data in a space with higher dimensions in order to find one in which the separation between classes is optimal and linearly separable. The kernel function does calculations only with the kernel products not requiring the calculation of vectors in higher dimensions. As an example, the kernel functions can project 1-dimensional data on higher dimensions simply by squaring the original expression values (Figure 7).

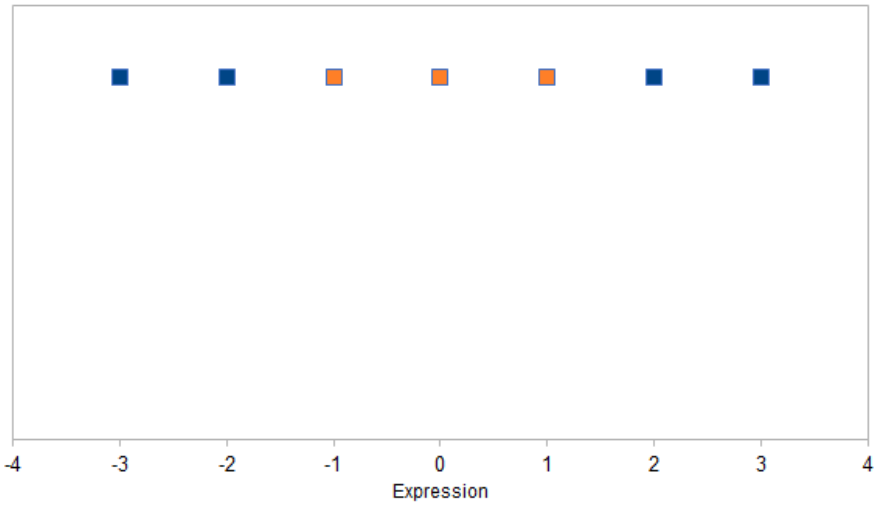

Figure 6 - A non-separable one-dimension data (group 1: orange; group 2: blue)

By doing so, the SVM has now identified a separating line that distinguishes between the two classes of data points (orange and blue) (Figure 7).

In summary, SVMs scale well to larger datasets due to their sparseness of solutions, allow the use of kernels to operate in higher dimension spaces and

take advantage of prior knowledge (i.e. by training on data with known class labels) (Pavlidis et al. 2004). The reduction of the VC-dimension of these classifiers by margin optimization also leads to a decrease in model complexity and consequent increase in generalizability (as explained in 1.6 Classification).

SVMs using high-dimensional kernels also have been shown to outperform other classification methods (Brown et al. 2000).

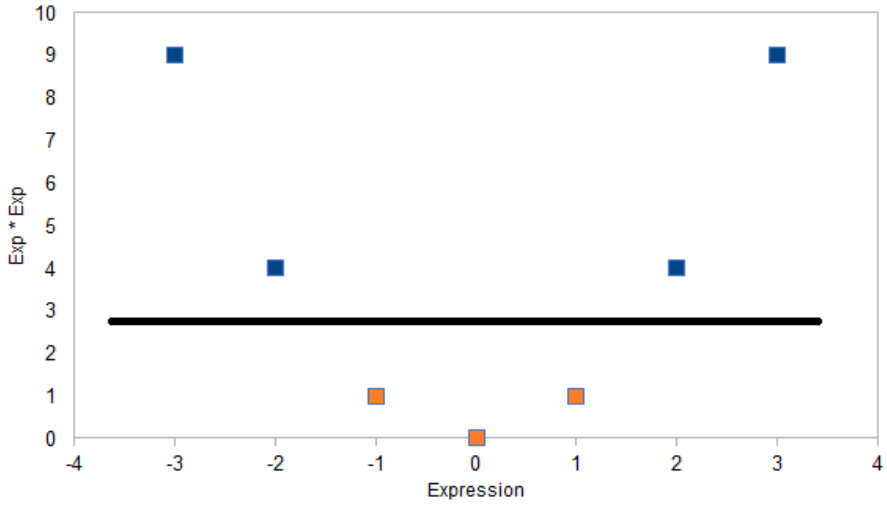

Figure 7 – Separation of the non-separable data in Figure 6 by squaring the data values (group 1: orange; group 2: blue).

However, SVMs are not without limitations: they handle only binary classification problems (Noble 2006); and the running times increase exponentially when the amount of data doubles.

1.9. Mixed Integer Linear Programming (MILP)

Mixed integer linear programming allows formulating linear optimization problems where a subset of variables is restricted to be integer. MILPs have gained increasing interest in the field of machine learning (Gordon et al. 2005, Schacht et al. 2014, Poos et al. 2016). All MILPs can be written in the form,

Objective function: $minimize\ c^t\ x$

Linear constraints: $Ax = b$

Boundaries: $l \leq x \leq u$

Integrality constraints: Some or all x_j must be integer values.

where *c*, *b* are vectors and *A* is a matrix. The solution is also limited by the upper (*u*) and lower (*l*) boundaries and the integrality constraints x_j allow the models to ascertain the discrete nature of some decisions (e.g. binary variables) (Gurobi Optimization 2016).

A simple example in which the usefulness of MILP is evident is the 0-1 Knapsack problem and is formulated as follows:

$$maximize \sum_{i=1}^{n} v_i\ x_i \tag{1}$$

$$\text{subject to } \sum_{i=1}^{n} w_i\ x_i \leq W \text{ and } x_i \in \{0,1\} \tag{2}$$

In this case, a bag exists with a maximum weight capacity *W*. The objective is to maximize the total value $\sum v_i x_i$ of the items (which can only be selected once) to place in the knapsack without exceeding the maximum allowed weight *W*. This is a special kind of MILP because all variables are binary and only one constraint exists. Despite its apparent simplicity, it is still an NP-hard problem which requires efficient solvers (Garey and Johnson 1979). The decision problem form of the knapsack problem is NP-complete. NP-complete (nondeterministic polynomial time problem) is a decision problem whose solutions can be verified rapidly (polynomial time) although without an efficient form of obtaining said solution. In other words, the time required to solve the decision problem increases rapidly with the size of the problem itself. NP-hard problems are optimization problems whose solutions are at least as hard as the decision problem to obtain.

Another example of the usefulness of employing MILPs is in the mapping of pathway networks onto 2-dimension lattice grids. Pathway analysis commonly only lists the genes that comprise the pathway without considering their interactions. By mapping pathways onto 2-dimension lattice grids, pathway analysis can be performed whilst considering their topological structure and how elements of the network interact (Piro et al. 2014). Usually gene expression profiling by high throughput technologies such as microarray or RNA-Seq identifies gene expression patterns that distinguish two conditions. Gene set enrichment analysis is then performed on the identified gene lists but does not consider the topology of the networks. In the study by Piro and colleagues (Piro et al. 2014) the authors aimed to identify enriched pathways that show differential regulation on a global scale but also specifically affected by the redirection of metabolic fluxes taking into consideration the topological information of the data. Consider Figure 8 as an illustrative example of how a metabolic network is embedded into a 2-dimensional grid and how the following MILP problem is formulated.

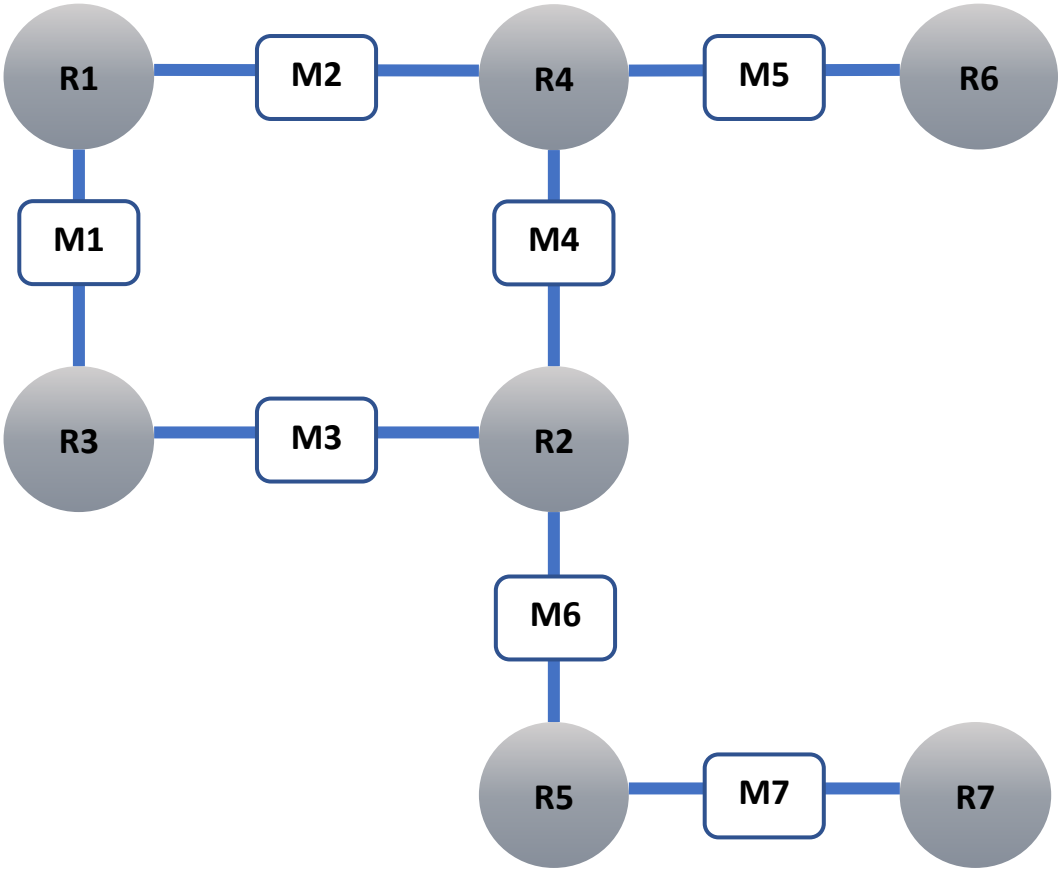

Figure 8 – Example of a metabolic network without considering topology

Integer variables for all edges in the network are introduced that model the Manhattan distance of the two end nodes. Next, binary variables x_{vij} are introduced as an indicator for where on the grid (position (i,j)) the node v should be placed. The objective function will be the minimization of the sum of the Manhattan distances d_{ab} in the grid (equation 3).

$$\min \sum_{(a,b)} d_{ab} \tag{3}$$

$$\sum_{ij} x_{vij} = 1 \tag{4}$$

$$\sum_{v} x_{vij} \leq 1 \tag{5}$$

A grid position must exist for all nodes (equation 4). Each position on the grid can only have, at most, one node (equation 5). Note that equations 6, which compute the Manhattan distances, are not linear. The linearization of the MILP involves converting equations 6 to inequalities.

$$\left|\sum ix_{aij} - \sum ix_{bij}\right| + \left|\sum jx_{aij} - \sum jx_{bij}\right| = d_{ab} \; all \; (a,b) \in E \qquad (6)$$

In the end, the 2-dimensional grid would appear as that illustrated in Figure 9.

Figure 9 – Two-dimensional grid representation of the network exemplified in Figure 8.

Using linear programming for classification was already performed in 1990 by Wolberg and Mangasarian (Wolberg and Mangasarian 1990) for the diagnosis of breast cancer. Multiple criteria linear programming is a classification method commonly used in data mining tasks. Similarly to Support Vector Machines, this method is also based on a set of classified training samples. It uses linear programming for determining the hyperplane which separates two classes. However, this method can only be applied to linearly separable data. Zhang and coworkers (Zhang et al. 2011) modified this classification method to not only deal with nonlinear separable data (by introducing a kernel function) but also to include prior knowledge. Incorporating prior knowledge should, in principle, improve outcomes when classifying nonlinear separable data (Zhang et al.

2011). They used linear constraints both coming from the training problem and from prior knowledge of the underlying classification problem. Considering Figure 4 as an example, prior knowledge, in this case, refers to polyhedral knowledge sets in Figure 10 (green rectangle and black triangle) in the input space of the data which can be expressed as a set of logical rules. Subsequently, the latter is converted into a series of equalities and inequalities in the SVM formulation (Fung et al. 2003, Zhang et al. 2011).

The addition of prior knowledge reduces the search space of the classifier. However, the inclusion of knowledge sets can change the linear classification of the SVM without prior knowledge. The inclusion of knowledge sets decreases the search space of the classifier which can lead to fewer solutions. However, since these knowledge sets are clearly known to identify each class, any new data that falls into these polyhedral sets are most likely to have a high confidence score.

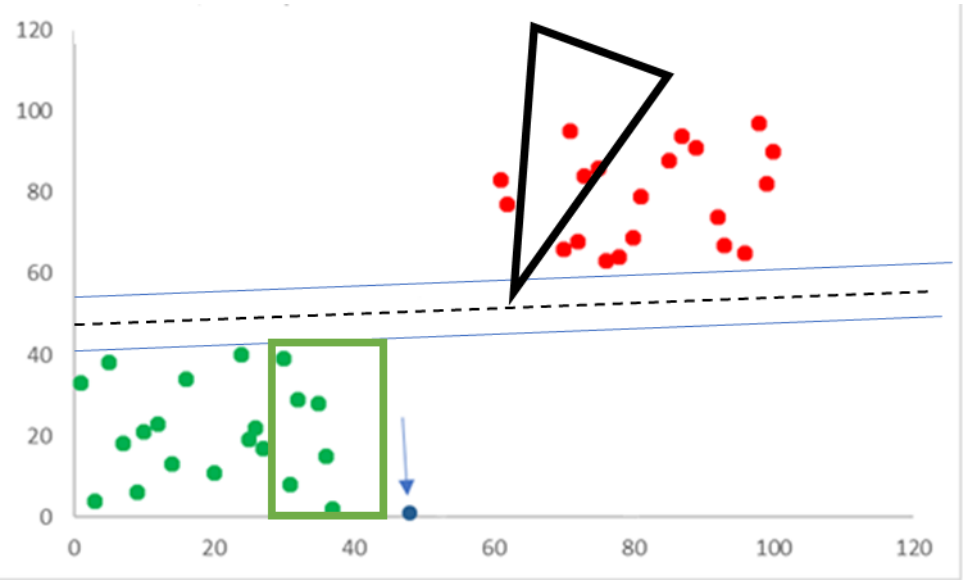

Figure 10 - SVM including prior knowledge (represented as polyhedral sets – green rectangle and black triangle). These knowledge sets are more beneficial if selected although other points can increase the margin hyperplane.

Employment of a constrained based method using Mixed Integer Linear Programming (MILP) has also been used in the inference of gene regulation (Schacht et al. 2014).

2. Objectives

The discovery of new and consistent biomarkers is an essential tool to improve diagnostics in the clinics, especially in the context of sepsis where the rapid identification of the invading pathogen can improve clinical outcome due to a quick and appropriate application of therapy.

However, studies have yet failed to produce a consistent and robust gene signature capable of distinguishing between microorganisms. In this thesis, the aim was to identify novel and robust gene signatures that could be used to distinguish fungal from bacterial infections in the human host. Novel classification methods were applied to produce robust and consistent biomarkers, independently of the cell type.

The goal was also to obtain information on biological functionality of the selected genes in the context of infection and how the host immune system reacted to fungal infections. Here, differential expression analysis coupled with gene set enrichment analysis were performed.

The heterogeneous composition of immune cells may mask pathogen associated molecular patterns (PAMPs) specific to certain cell types. Neutrophils, for instance, account for ~65% of leukocytes. Nevertheless, its action in fighting infection might shadow expression patterns of monocytes, which only make up for ~5% of leukocytes. Thus, determining existing signaling cascades that are specific or enhanced in similar leukocyte type compositions was targeted in addition.

3. Materials and methods

3.1. Dataset Assembly

The normalized gene expression data from three datasets (accession numbers: GSE65088, GSE42606 and GSE69723) was obtained via Gene Expression Omnibus (GEO) from the National Center for Biotechnology Information (NCBI) database. RNA-Seq data was retrieved from NCBI's Sequence Read Archive (SRA). A study performed by Klassert and colleagues (Klassert et al. 2017), and hereon identified as "Klassert", generated RNA-Seq data (accession number SRP076532) which consisted of healthy human blood-derived monocytes stimulated with heat-killed *Aspergillus fumigatus* AF293, *Candida albicans* SC5314 yeast (both at a MOI of 1), *Escherichia coli* serotype O18:K1:H7 (MOI of 10) or left untreated (control). Cells were stimulated for 3 and 6 hours after which their RNA was extracted. On the raw reads a sequence quality analysis was performed using FastQC version 0.10.1 and a read trimming to 150 bp was performed using FASTX Toolkit 0.0.14 and adapter trimming using cutadapt version 1.3. The reads had then been mapped to the reference genome GRCh38/hg38 from the UCSC server and counted for each gene across all samples using HTSeq-count. The read number per gene, total read number per sample and gene length was then used to calculate the Reads Per Kilobase of transcript per Million mapped reads (RPKM) values across all genes and samples. Genes with RPKM values of 0 across all samples were removed. A second dataset (accession number GSE65088) was generated by Dix and co-workers (Dix et al. 2015), hereby identified as "Dix", and consisted of anticoagulated blood from healthy human donors challenged with *C. albicans* SC5314 (1×10^6/mL), *A. fumigatus* ATCC46645 (1×10^6/mL), *E. coli* ATCC25922 (4×10^3/mL) and *S. aureus* (1×10^6/mL). Mock-infected blood samples were used as controls. Samples were taken at 4 and 8 hours post-infection.

Smeekens and colleagues (Smeekens et al. 2013) performed a study in which Peripheral Blood Mononuclear Cells (PBMCs), isolated from blood of healthy human donors, were stimulated with heat-killed *C. albicans* UC820 (1 x

10^6/mL), *Mycobacterium tuberculosis* (10ng/mL) and LPS derived from *E. coli* (10ng/mL). Cells grown in Roswell Park Memorial Institute Medium (RPMI) culture medium were used as controls (accession number GSE42606). Samples were taken at 4 and 24 hours after infection. In this dataset, only the 4-hour time point was considered for our studies since the main focus was the innate immune response. For future reference this dataset will be identified as "Smeekens".

The dataset (accession number GSE69723) generated from the study by Czakai and co-workers (Czakai et al. 2016), and hereby identified as "Czakai", consisted of healthy human blood derived dendritic cells challenged with thimerosal treated *C. albicans* SC5314 (MOI of 1), *A. fumigatus* ATCC46645 (MOI of 1) and *E. coli*-derived LPS (1µg/mL). Samples were collected 6 hours post-challenge. Transcriptomic data generated by us, i.e. Saraiva and colleagues (Saraiva et al. 2016), and hereby identified as "Saraiva", was generated by challenging healthy human blood-derived PBMCs with either heat-killed *C. albicans* MYA-3573 yeast (MOI of 2) or LPS derived from *E. coli* 0111:B4 (10 ηg/mL) (InvivoGen). Four samples were extracted 4 hours post-infection. RNA was extracted using RNAEasy Kit Qiagen and quantity and quality of the total RNA was analyzed using a Nanodrop ND-1000 spectrophotometer (Thermo Fischer Scientific, USA) and a Tape Station 2200 (Agilent Technologies, USA). Lastly, transcriptional data of human blood isolated monocytes challenged with *A. fumigatus* conidia (MOI of 2) and LPS (10 ηg/mL) was downloaded from the European Molecular Biology Laboratory (EMBL) ArrayExpress database (E-MEXP-1103) (http://www.ebi.ac.uk/arrayexpress/experiments/E-MEXP-1103/) and is hereby identified as "Mattingsdal". A total of 5 and 6 samples were extracted 6 hours post-challenge (*A. fumigatus* and LPS, respectively). On the "generic" fungal *versus* bacteria study, this dataset was used for validation of the gene signature whilst in the similar leukocyte study was used for feature selection and training of the classifiers.

3.2. Data preprocessing

Each dataset was controlled if prior normalization had been executed on the expression data. In the absence of normalization, the following was performed: a 1 % quantile was added onto all expression values of the RNA-Seq data and log2 transformed, whilst microarray data was normalized by employing the functions "lumiN" and method "vsn" of the "lumi" R package (Du et al. 2008). Elimination of possible duplicate gene entries was carried out by use of the "avereps" function in the "limma" R package (Ritchie et al. 2015), which calculates the mean expression values for duplicate entries. Genes with an intensity and variance below 40 % were removed. Finally, z-scores were calculated for each gene. The gene list, to be used for feature selection and classification on infected *versus* healthy and "generic" fungal *versus* bacterial studies, consisted of the intersection of the gene lists from the datasets "Smeekens", "Klassert", "Czakai", "Saraiva" and "Dix" and amounted to 1,567 genes. The gene list used for the study of similar leukocyte composition was composed by 1516 genes and was obtained by the intersection of the gene lists from the datasets "Smeekens", "Klassert", "Saraiva" and "Mattingsdal".

In each dataset, the following procedure was employed: In the infected *versus* non-infected sample analysis, cell-infected samples were assigned to group 1 whilst healthy samples were assigned to group 2. In the "Fungal *versus* Bacterial" analysis the samples were grouped into either fungal class (group 1) or bacterial class (group 2). The number of samples in each dataset for each analysis is shown in Table 1 and Table 2. An important aspect of the datasets used in this work is their heterogeneity such as sequencing platforms, type of immune cells in each dataset, number of samples per stimulus and different microorganisms.

Table 1 - Number of samples in each dataset divided into infected and non-infected status

Dataset	Infected class	Non-infected class
Smeekens	73	30
Dix	18	18
Klassert	27	9
Czakai	12	4
Saraiva	8	4

Table 2 – Number of samples in each dataset assigned to fungal or bacterial groups

Dataset	Fungal class	Bacterial class
Smeekens	24	49
Dix	16	20
Klassert	18	9
Czakai	8	4
Saraiva	4	4
Mattingsdal	5	6

3.3. Support Vector Machine (SVM) Implementation

In each analysis, the Mixed Integer Linear Programming (MILP) implementation of the Support Vector Machine (SVM) was realized by the following equations: The objective function was defined as the maximization of the margin of the SVM as seen in Equation 7,

$$obj_{classifier} = ma\,x(t_1 - t_2) \quad (7)$$

with t_1 and t_2 are the margins of class 1 and class 2 to the separating hyperplane, respectively. The objective function was subjected to the following constraints:

$$\sum_{i=1}^{nGenes} n_i g_{ij} \geq t_1 - My_j \ \forall_j \in C_1 \quad (8)$$

$$\sum_{i=1}^{nGenes} n_i g_{ij} \leq t_2 + My_j \ \forall_j \in C_2 \quad (9)$$

Equations 8 and 9, define the constraints applied to the classifier, for both class 1 (C_1) and class 2 (C_2), respectively. The scalar product of the gene expression g_{ij} of sample j with the weight n (for all genes $i \in \{1, \ldots, nGenes\}$) assigned them to a specific side of the margin but only for samples whose variables $y_j \in \{0,1\}$ were equal to 1. If this scalar product was less or equal than t_2 the samples were classified as group 2 and if greater or equal to t_1, classified as group 1. M was a large constant ("big M") that was set to allow exceptions if y_j equaled 1. Equation 10,

$$\sum_{j=1}^{nSamples} y_j \leq k \quad (10)$$

constrained the number of allowed misclassifications k during the training (with *nSamples* training samples) of the classifier. k was set to 10% of the total number of samples $|S|$. To ensure that only genes i whose corresponding variables $x_i \in \{0,1\}$ equaled to 1 were used for classification, constraints of equations 11 and 12 were established,

$$n_i \leq x_i \ \forall \ i \in G \quad (11)$$

$$-n_i \leq x_i \ \forall \ i \in G \quad (12)$$

The number of features (genes) to be determined was constrained by equation 13, in our present study this was set to l=30,

$$\sum_{i=1}^{nSamples} x_i \leq l \tag{13}$$

x and y were defined as binary variables which belong to the set of genes G and samples S by equations 14 and 15, respectively:

$$x_i \in \{0,1\} \forall\, i \in G \tag{14}$$

$$y_j \in \{0,1\} \forall\, j \in S \tag{15}$$

To note, applying these sets of constraints generated a MILP problem and not an ordinary Linear Programming (LP) problem. Selection of consistent genes across all datasets required the combination of two independent MILPs. Each independent classifier was established by applying all previously defined equations. Next, the problems were connected by a combined objective function, equation 16,

$$obj_{combined} = sum\left(obj_{classifier\,1} + obj_{classifier\,2}\right) \tag{16}$$

adding the objective functions of each classifier. Using identical x variables in both classifiers ensured that they use the same set of features, possibly leading to a decrease in performance of the classifiers (Figure 11).

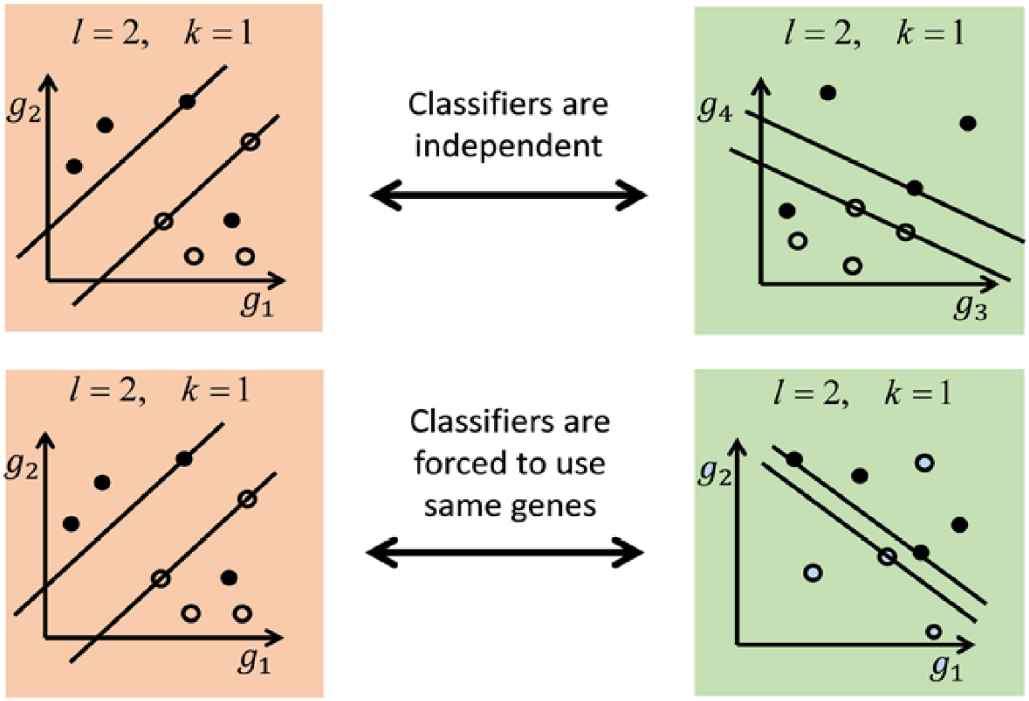

Figure 11 - The upper two SVM classifiers maximize the margin independently from each other (left: g_1 and g_2; right: g_3 and g_4). The lower two maximize the sum of the two margins subject to that both use the same set of genes (g_1 and g_2) for the SVMs.

3.4. Machine learning and statistical analysis

Balancing classes is standard practice when applying Support Vector Machines, because sub-optimal results can be obtained whilst having unbalanced classes in the datasets. The training of SVM classifiers on unbalanced classes may produce models biased towards the class with the highest number of samples (Chawla et al. 2004). To eliminate this problem, a stratification approach was implemented during each classification problem. A *k*-fold cross-validation was employed in which 2/3 of the samples from the minority class (e.g. infected and non-infected or fungal and bacterial) were randomly chosen for training. A 10% sample misclassification was allowed. Random samples of the majority class were selected that amounted to the

number of samples in the minority class. The remaining samples were used for validation and for measuring classifier performance (see 3.5. Overall performance). This procedure was repeated 100 times generating 100 lists of selected genes used as features of the SVMs. For comparing gene lists across single and combined classifiers, the number of selected genes was constrained to l=30. Performance was assessed by the accuracy (percentage of correct predictions on the test set) of the classification on the validation sample sets. Average performance values were calculated for combined classifiers. Comparison of single with combined classifier performances was achieved by their overall average, respectively. Consistency of selected genes was calculated for each pair of lists of selected genes by calculating the pairwise overlap (POL) between the 100 gene lists generated during classification of the two datasets in question. As an example illustrated in Figure 12, every iteration of the "Dix" classifier is intersected with every iteration of the "Smeekens" classifier until all possible combinations are accounted for. The number of intersecting genes in each pairwise calculation is shown above the blue lines. The average POL, in this small example, would be 1.78.

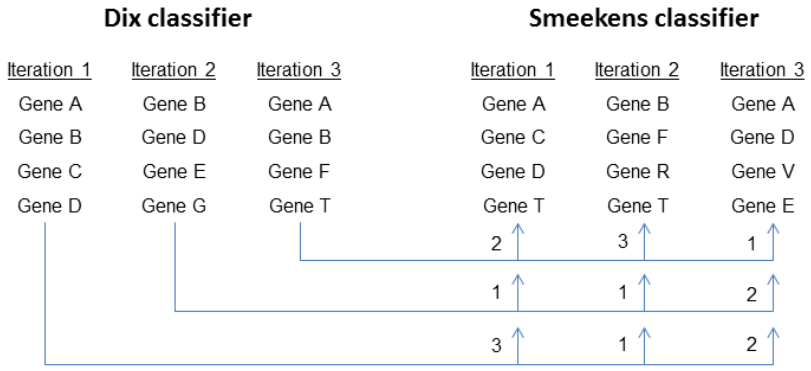

Figure 12 – Illustrative example of the calculation of pairwise overlap between classifiers. Every iteration of the "Dix" classifier was compared to every iteration of the "Smeekens" classifier and the number of genes present in both classifiers is extracted (above blue line and next to the arrow). The total number of pairwise overlap combinations in this example is 9 with an average POL value of 1.78 (16/9 = 1.78).

In the end, the average number of intersecting genes between two independent classifiers is calculated. The mean POL and standard deviations (1σ) were calculated from the list of POL. The final list of intersecting genes was obtained by taking the union of genes from each classifier that were selected in at least 40% of the cross-validation runs.

3.5. Overall performance

Determining the overall performance of the generated models provides an estimation of their generalization error (i.e. correct prediction of samples not included during the training of the classifiers). This would, ideally, be assessed on completely unseen data. An alternative to this, as stated above, is to divide the data into parts before feature selection and classification and one of them used for classification and the other for testing. Several common performance metrics exist such as sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV) and accuracy and are described below. Calculating such measures relies on the classification results exemplified in Table 3 which reports the True positive (TP), True negative (TN), False positive (FP) and False negative (FN) predictions.

Sensitivity reflects the number of samples that were correctly predicted to belong to group 1 (equation 17) divided by the total number of samples of group 1, where specificity reflects the number of samples that were correctly predicted to belong to group 2 (equation 18) divided by the total number of samples of group 2. PPV (equation 19) is the probability that the samples assigned to group 1 indeed belong to that group. Similarly, NPV (equation 20) is the probability that the samples assigned to group 2 truly belong to that group. Finally, accuracy (equation 21) is determined by calculating the proportion of correct sample classifications. Table 3 illustrates where the values are inserted into the confusion matrix.

$$Sensitivity = \frac{TP}{TP + FN} \quad (17)$$

$$Specificity = \frac{TN}{FP + TN} \quad (18)$$

$$PPV = \frac{TP}{TP + FP} \quad (19)$$

$$NPV = \frac{TN}{TN + FN} \quad (20)$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (21)$$

Table 3 – Scheme of a confusion matrix which contains the results of the classifier when comparing two groups. In this case, group 1 is regarded as positive.

		Predicted class	
		Group 1	Group 2
True class	Group 1	True positive (TP)	False Negative (FN)
	Group 2	False positive (FP)	True negative (TN)

As benchmark, the average across all single classifiers, of each of these performance measures was calculated. For the combined classifiers, each pair of classifiers was run within a cross-validation scheme resampling different sets of samples for training and validation and counted the pairwise overlaps of pairs of classifiers which have been run on two other datasets. As an example, the feature lists from a classifier pair of the datasets of Dix and Klassert was compared to the pair of classifiers from Czakai and Smeekens and the number of the same selected features counted. This was performed for all combinations of the different runs of the cross validation and averaged, yielding the average

pairwise overlaps (averaged POL). This was done for all combinations of pairs of datasets and compared to the benchmark. Note that intersections in which the same dataset occurred on both sides were not considered (Figure 13).

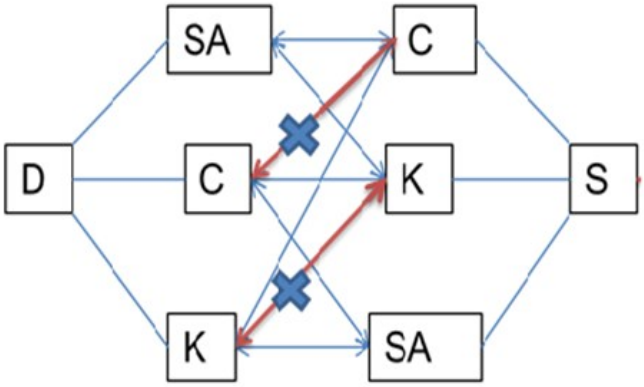

Figure 13 - To compare the benchmark results with the combined approach, the resulting gene list of each combination containing one of the datasets (here exemplarily shown for Dix) was intersected with each combination containing the second dataset (here: Smeekens). Note that intersections in which the same dataset occurred on both sides were not considered (here shown exemplarily for the combinations Dix & Czakai *versus* Smeekens & Czakai; or Dix & Klassert *versus* Smeekens & Klassert). (D: Dix, S: Smeekens, C: Czakai, K: Klassert, Sa: Saraiva).

3.6. Gene expression analysis and refinement of gene signatures

Differential gene expression was calculated using Student`s t-tests and multiple testing correction was performed by the Benjamini-Hochberg method (Benjamini and Hochberg 1995). Genes were considered differentially expressed between two classes if their adjusted p-value was equal or below 0.05.

Refinement of the gene signatures produced during the infected *versus* healthy study was achieved by selecting genes that were present in at least 40% of all combined classifier combinations, followed by excluding genes whose differential expression profile was not consistent in at least 4 out of 5 datasets.

In the generic fungal *versus* bacterial biomarker study (see 4.2), the refinement of the gene signature was assessed by selecting only genes that were consistently differentially expressed (Class1 *versus* Class 2) in all datasets used in feature selection and classification. *In silico* validation of the refined gene signature generated from the generic fungal *versus* bacterial analysis, was determined by employing random forest classifiers (available through the "caret" package, version 6.0-7.1), trained on four datasets ("Dix", "Smeekens", "Klassert" and "Czakai") and tested on "Mattingsdal". This was performed in order to demonstrate that the resulting gene signature could be used to discriminate two classes even when using other common classification methods. The dataset "Saraiva" was not included during this process due to its small sample size (4 fungal samples and 4 bacterial samples).

The study of datasets with similar leukocyte compositions followed the workflow depicted in Figure 14 and is described in the following.

Genes with an intensity and variance below 40 %, in all datasets ("Smeekens", "Saraiva", "Klassert", and "Mattingsdal"), were removed and the others were z-normalized. The datasets were next used in (blue) classification and feature selection and in (green) the determination of differentially expressed genes. From each classifier (single or combined approach), genes that were not selected in at least 20% of the total number of runs were excluded. The resulting gene lists from each group were united and gene set enrichment analysis was performed. Differentially expressed genes were determined in each dataset. Next, three lists of genes were produced representing genes that were differentially expressed in i) all used datasets (S,Sa,K,M), ii) datasets with challenged PBMCs (S,Sa) and iii) datasets with challenged monocytes (K,M). Gene set enrichment analysis was performed for each separate list of differentially expressed genes. The enrichment results derived from the classifiers were then compared to those derived from the differential expression analysis. Ideally, the enriched pathways from both approaches should be highly similar since both approaches identify genes that distinguish samples belonging to two classes. Validation of genes of interest identified during this last study

was determined via real-time qPCR by collaboration partners of the Host Septomics group led by Dr. Hortense Slevogt.

All statistical analyses, except for those concerning the RT-qPCR (see 0), were performed using R software (http://www.r-project.org/) and packages from Bioconductor(Huber et al. 2015).

The MILP approach was implemented in R using the Gurobi interface library and solved with the Gurobi solver (version 6.5.1, www.gurobi.com).

Ascertaining the functional overview of the refined gene signature was achieved using functional annotation tools of the Database for Annotation, Visualization and Integrated Discovery (DAVID, version 6.7, https://david.ncifcrf.gov/home.jsp) (Huang et al. 2009a) with *homo sapiens* genes as background. Briefly, this web-accessible program allows the user to upload gene lists for rapid annotation and analysis. It has the benefit of integrating several sources of annotation data such as KEGG, UniGene, and Gene Ontology among others (Dennis Jr et al. 2003). Besides the association of the uploaded genes to a biological process (i.e. gene ontology terms, KEGG pathways) thus grouping them into a functional category, DAVID also calculates the most enriched pathways by means of a modified Fisher's exact test which is more stringent than the normal Fisher test (Huang et al. 2009).

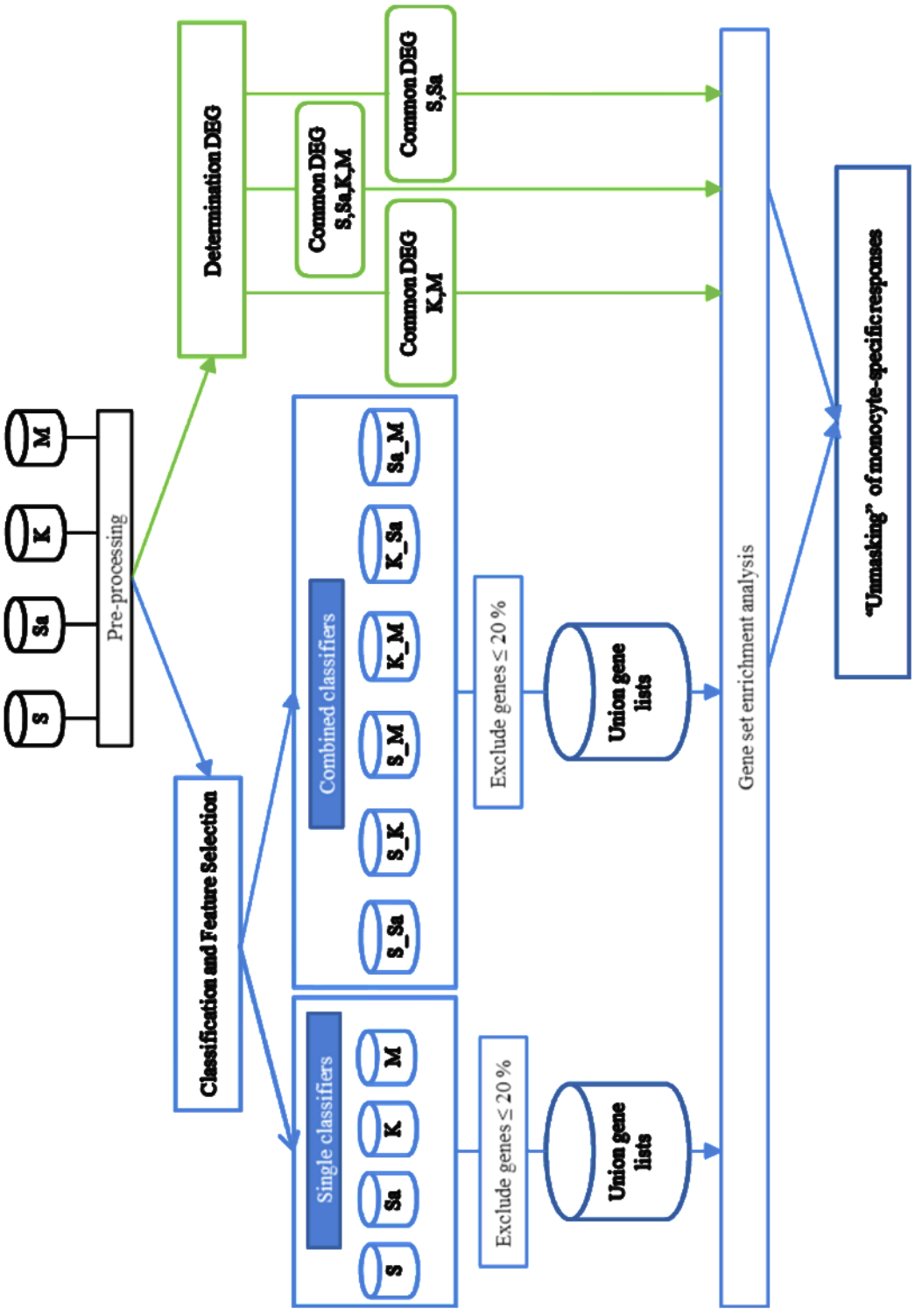

Figure 14 - Workflow for analyzing datasets with similar leukocyte compositions (explanation in page 48) (S: Smeekens, K: Klassert, Sa: Saraiva; M: Mattingsdal)

3.7. Experimental validation via quantitative reverse transcription PCR (RT-qPCR)

Experimental validation of genes of interest was entirely done by the collaboration partners (Tilman Klassert and Cristina Zubíria-Barrera) of the Host Septomics group led by Dr. Hortense Slevogt. The full procedure is described as follows.

3.7.1. Monocyte isolation

Buffy coats of healthy male donors for cell isolation were kindly provided by Dagmar Barz in anonymized form (Institute of Transfusional Medicine of the Jena University Hospital). Human monocytes were isolated from 50 ml buffy coats of four healthy male donors as previously described (Müller et al. 2017). Briefly, ficoll-density gradient centrifugation was used to isolate first peripheral blood mononuclear cells (PBMCs). After restoring the osmolarity of the cells with 0.45% NaCl, remaining erythrocytes were lysed using a hypotonic buffer. Where needed, 5×10^6 PBMCs were seeded in 6-well plates (VWR International, Germany) and allowed to equilibrate for 1h at 37°C 5% CO_2. From the remaining PBMCs, monocytes were then isolated using quadro-MACS (Miltenyi Biotec, UK) by labeling the non-monocytic cells with a cocktail of Biotin-conjugated antibodies and Anti-Biotin Microbeads (Monocyte Isolation Kit II, Miltenyi Biotec, UK). Cell viability of > 98% was assayed by Trypan blue staining. Monocyte concentration was adjusted to 2.5×10^6 cells/ml in RPMI 1640 GlutaMAX medium (Gibco, UK) supplemented with 10% fetal bovine serum (FBS, Biochrom, Germany) and 1% Penicillin/Streptomycin (Thermo Fisher Scientific, USA), 5×10^6 cells were seeded in 6-well plates (VWR International, Germany) and allowed to equilibrate for 1h at 37°C 5% CO_2.

3.7.2. Preparation of fungi and bacteria

Overnight culture from *Escherichia coli* (isolate 018:K1:H7) in LB medium was washed twice in PBS and resuspended in 1 ml RPMI 1640 GlutaMAX medium (Gibco, UK) supplemented with 10% FBS (Biochrom, Germany) at a concentration of 5×10^8 cfu/ml. *Aspergillus fumigatus (*AF293) was grown in Aspergillus Minimal Medium (AMM) Agar-plates for 6 days at 30°C. Conidiospores were harvested by rinsing the plates with sterile 0.05% Tween-20 (Sigma-Aldrich, Germany) and filtered through 70-μm and 30-μm pre-separation filters (Miltenyi Biotec, UK) to get rid of mycelium traces. Spores were washed twice in PBS and cell-concentration was adjusted to 10^7 conidia/ml in RPMI 1640 GlutaMAX medium supplemented with 10% FBS. Conidia were then incubated at 37 °C under shaking for 7 h until cells turned to germ tubes. Germlings were centrifuged and resuspended at 1×10^8 cells/ml in RPMI 1640 GlutaMAX medium supplemented with 10% FBS. Overnight culture of *Candida albicans* (SC5314) in YPD medium was washed twice in PBS and cell concentration was adjusted to 5×10^7 cfu/ml in RPMI 1640 GlutaMAX medium supplemented with 10% FBS.

3.7.3. Monocyte stimulation assay

Pathogens were all heat-killed by incubation at 65°C for 30 min before infection. Monocytes were stimulated with heat-killed pathogens at a pathogen:host ratio of 10:1 for bacteria, 1:1 for *A. fumigatus* germ tubes and *C. albicans* yeasts. In addition, cells were stimulated with pathogen-derived cell wall components: LPS (50 ng/ml) and zymosan (1 µg/ml). After 3 h incubation at 37°C and 5% CO_2, monocytes were lysed for RNA isolation. To analyse the expression level of the genes of interest, total RNA was extracted from 5×10^6 Monocytes using the Qiagen RNeasy mini kit (Qiagen, Germany). Residual genomic DNA was removed by on-column incubation with DNaseI (Qiagen, Germany). A NanoDrop D-1000 Spectrophotometer (Thermo-Fisher Scientific, USA) was then used to assess the amount and quality of the isolated RNA samples. Complementary DNA (cDNA) was synthesized from 1.5 μg of RNA

using the High Capacity cDNA Reverse Transcription Kit (Applied Biosystems, UK) following manufacturer's instructions. To detect the expression of the genes by PCR, specific primers for each target were designed using the online Primer-BLAST tool of the NCBI (http://www.ncbi.nlm.nih.gov/tools/primer-blast/). Possible secondary structures at the primer binding sites were taken into account by characterizing the nucleotide sequence of the regions of interest using the Mfold algorithm (Zuker 2003). The sequences of all primers used for amplification are listed in Table S1 in the supplementary material. For quantification of the relative expression of each gene, we used a CAS-1200 pipetting robot (Qiagen) to set up the qPCR-reactions and a Corbett Rotor-Gene 6000 (Qiagen) as Real-Time qPCR apparatus. Each sample was analysed in a total reaction volume of 20 μl containing 10 μl of 2× SensiMix SYBR Master Mix (Bioline, UK) and 0.2 μM of each primer. The cycling conditions included an initial step of 95°C for 10 min followed by 40 cycles of 95°C for 15 s, 60°C for 20 s and 72°C for 20 s. For each experiment, an RT-negative sample was included as a control. Melting curve analysis and primer efficiency was used to confirm the specificity of the qPCR reactions. The relative expression of the target genes was analysed using the Pfaffl method (Pfaffl et al. 2004, Rieu and Powers 2009). To determine significant differences in the mRNA expression between different experimental conditions, the relative quantity (RQ) for each sample was calculated using the formula $1/E^{Ct}$, where E is the efficiency and Ct the threshold cycle. The RQ was then normalized to the housekeeping gene peptidylprolyl isomerase B (*PPIB*). The stability of the housekeeping gene was assessed using the BestKeeper algorithm (Pfaffl et al. 2004). The normalized RQ (NRQ) values were $\log_2$-transformed for further statistical analysis with GraphPad PRISM v7.02. Statistical analysis was performed using repeated measures one way ANOVA and Bonferroni correction.

Table 4 - Primer sequences for candidate genes employed in qRT-PCR

GENE NAME	SYMBOL	FORWARD PRIMER (5'->3')	REVERSE PRIMER (5'->3')	SIZE (BP)
PEPTIDYL-PROLYL CIS-TRANS ISOMERASE B	*PPIB*	ATGTAGGCCGGG TGATCTTT	TGAAGTTCTCAT CGGGGAAG	219
GALACTOSIDASE ALPHA	*GLA*	AGGAAGAGCCAG ATTCCTGC	GCGAATCCCATG AGGAAAGC	185
SCAVENGER RECEPTOR CLASS B MEMBER 2	*SCARB2*	GCATGCACCCAA ATCAGGAA	GTCGACTCGCCG TCTCTTTA	210
BCL2 ASSOCIATED ATHANOGENE 3	*BAG3*	CCAGAAACCACT CAGCCAGA	CGGAATGGAGAT GTACCCCC	204
PEROXISOME PROLIFERATOR ACTIVATED RECEPTOR GAMMA	*PPARG*	ACAGATCCAGTG GTTGCAGA	AGATGCAGGCTC CACTTTGA	81
FATTY ACID BINDING PROTEIN 5	*FABP5*	AAACCACAGCTG ATGGCAGA	GCTTTCCTTCCC ATCCCACT	92
CLUSTER OF DIFFERENTIATION 164	*CD164*	CCGAACGTGACG ACTTTAGC	GAAGTCTGTCGT GTTCCCCA	234
NPC1 - NPC INTRACELLULAR CHOLESTEROL TRANSPORTER 1	*NPC1*	AGCCACATAACC AGAGCGTT	GAGTGGCTCCCA GTAAGACC	221
HEME OXYGENASE 1	*HMOX1*	AACTTTCAGAAG GGCCAGGT	AGACTGGGCTCT CCTTGTTG	115
C-C MOTIF CHEMOKINE RECEPTOR 1	*CCR1*	TCTTTGGGCTGG TATTGCCT	ACAGCCAGGTCC AAATGTCT	235

4. Results

The focus of this thesis was to identify a consistent gene signature for systemic infection as well as to discriminate between fungal and bacterial infection, either dependent or independent of leukocyte type. For this purpose, a novel constrained based machine learning approach was used which connected two independent classification problems by constraining them to use the same set of genes during feature selection. SVMs were used to implement each classification problem. Each classification problem was set up to maximize the margin of its respective SVM. To note, classifiers were based on different datasets/experiments. Dependency of classifiers on each other limited the search space from which features could be selected and forced them to select genes which were discriminative for both classification problems (Figure 11). In this manner, a "collaborative" selection of genes should enable improving consistency of the biomarker gene list across datasets.

4.1. Discriminating infected from non-infected samples

The classifier was run on different combinations of sample sets (n=100 training sets) obtained from each dataset described in Table 1. For each run, a list of 30 genes was selected by the classifiers which best maximized the separating margin between the classes. The pairwise overlap (POL) of the 100 gene lists from each single classifier was calculated and returned an average of 1.50 (1σ =0.83). For improving the consistency of the gene lists, two single classifiers were combined, each respective to different datasets. At this point, the POL of the gene lists from each of the 100 runs from the single classifier and the combined classifiers were calculated yielding an average POL of 1.99 (1σ =0.65). Additionally, the POL between two combined classifiers was calculated in the same manner as previously stated and returned an average value of 2.14 (1σ =0.61). A considerable increase in consistency of 42% was obtained using the combined approach when comparing the averaged POL of single *vs*. single 1.50 (1σ =0.83) to combined *vs*. combined 2.14 (1σ =0.61). To note this

difference was significant (P=2.95E-10), using a two-sided Kolmogorov-Smirnov test.

Comparing the actual lists of selected genes from the combined to those of the single classifiers also revealed an improved consistency. The composition of the gene lists was assembled from all cross-validation runs of each dataset.

The novel approach also yielded an improved consistency when comparing the actual lists of selected genes. For each single classifier (D: Dix, S: Smeekens, C: Czakai, K: Klassert, Sa: Saraiva), the genes selected in all cross-validation runs were united. Next, the resulting union gene lists for each pair of single classifiers (e.g. Dix *versus* Smeekens) was intersected. The percentage of intersecting genes, genes selected only in the first classifier, and genes selected only in the second classifier were calculated, respectively. The average number of intersecting genes was 8%. This was established as the benchmark and the results are shown in Table 5. Similarly to the pairwise comparison of single classifiers, an average overall intersection of 12 percent was obtained for genes which were commonly selected in the investigated infection datasets using the combined approach. For each combined classifier, the genes from all cross-validations were merged resulting in 6 gene lists. Next, the resulting gene lists from all combined classifiers were intersected, resulting in 377 unique genes, out of which, 33 genes were selected in at least 40% of all runs of all combined classifier combinations. As in the case of combined classifiers, the gene lists from all cross-validations from each single classifier were merged together which resulted in five gene lists. Next, the latter were intersected yielding 149 unique genes. In this case, only 8 genes were selected in 40% of all single classifier combinations. The 33 genes identified through the combined classifier approach were proposed as potential biomarkers for infection.

This was done for each possible combination of single classifiers (9 combinations). This novel approach required the calculation of such intersections between pairs of combined classifiers (see explanation in subsection 3.5). The averaged results are given in Table 5 (below "Combined Approach").

To note, these results were yielded by selecting the number of genes to use for classification of l=30. This parameter seemed not to be crucial as similar improved pairwise overlaps were obtained when selecting l=20 (single classifiers: 0.5 (1σ=0.37), combined classifiers: 0.73 (1σ =0.48)).

Table 5 - Percentages of the intersections of the benchmark and the combined classifier approach

	Benchmark			Combined Approach		
Single	**Dataset 1**	**Dataset 2**	**Intersection**	**Dataset 1**	**Dataset 2**	**Intersection**
D *vs* S	40	54	6	42	47	11
D *vs* C	39	57	4	38	52	10
D *vs* K	38	56	6	37	50	13
D *vs* Sa	23	67	10	33	55	12
C *vs* Sa	33	58	9	39	50	12
C *vs* S	49	44	7	38	51	11
C *vs* K	45	45	9	35	52	13
Sa *vs* S	58	25	17	37	50	13
Sa *vs* K	58	34	8	39	50	11
S *vs* K	44	49	7	34	53	12
Average			**8**			**12**

D: Dix, S: Smeekens, C: Czakai, K: Klassert, Sa: Saraiva

Table 6 shows the 33 selected genes, their differential gene expression and their significance values for each of the investigated datasets. The average accuracy of classification using single classifiers was 90%. When using only the above mentioned proposed gene signature for classification on each dataset, the average accuracy was of 92%.

To improve the robustness of the gene signature, the number of discriminating genes was further decreased by considering only genes whose differential expression profile was consistent in at least 4 out of 5 datasets. The resulting list was composed of 23 genes and yielded an average accuracy of 90% when used for classification on single datasets.

Testing was also performed to determine if the prediction performances of the classifiers suffered greatly by their combination. Single classifier accuracy was,

on average, 90%. Interestingly, a slight increase, to 93%, in performance was obtained when combining classifiers (Table 7).

Table 6 - Selected genes from all combined classifications and their differential expression and regulation.

Gene symbol	Dix		Smeekens		Saraiva		Klassert		Czakai	
	Pval	Reg	Pval	Reg	Pval	Reg	Pval	Reg	Pval	Reg
ADM	2.1E-06	1	6E-11	1	0.00031	1	3.2E-14	1	0.0000035	1
CD83	4.1E-12	1	2E-14	1	0.00019	1	1.7E-15	1	0.0000044	1
MSC	4.3E-07	1	5E-20	1	0.0079	1	3.9E-13	1	0.0000058	1
BTG3	1.8E-06	1	2E-17	1	0.0029	1	4.1E-12	1	3.1E-07	1
ZC3H12C	2.3E-10	1	4E-18	1	0.0008	1	1.4E-13	1	0.00002	1
IRAK2	3.4E-06	1	2E-22	1	0.00012	1	3.6E-12	1	4.8E-07	1
PIM1	5.9E-13	1	4E-17	1	4.8E-06	1	1.6E-11	1	0.00014	1
TRAF1	0.00006	1	1E-26	1	0.00054	1	3.1E-10	1	1.7E-07	1
TXN	2.9E-07	1	4E-18	1	4.4E-05	1	2.9E-12	1	0.000044	1
USP12	0.0043	1	5E-09	1	0.00087	1	7.6E-12	1	0.0000002	1
CXCL1	2.7E-06	1	3E-35	1	2.2E-05	1	1.5E-10	1	0.000014	1
DFNA5	1.9E-05	1	2E-14	1	0.0031	1	4.1E-08	1	0.0000036	1
GJB2	2.7E-11	1	2E-18	1	0.0002	1	3.8E-14	1	0.000097	1
IL1B	0.0021	1	8E-40	1	3.3E-05	1	9.3E-09	1	7E-08	1
IL6	1.6E-13	1	2E-41	1	1.1E-10	1	1.5E-07	1	0.0013	1
MESDC1	0.0035	1	5E-11	1	6.7E-05	1	3.5E-12	1	0.000002	1
PPP1R15A	8.8E-08	1	1E-11	1	0.00016	1	4.9E-12	1	2E-08	1
RGS1	6.6E-19	1	1E-05	1	0.0051	1	7.4E-09	1	0.000016	1
TXNRD1	0.0017	1	2E-10	1	2.2E-05	1	4.3E-06	1	0.0069	1
CD300LF	0.0045	-1	7E-16	-1	0.0005	-1	3.2E-08	-1	0.00042	-1
TMEM170B	1.4E-06	-1	1E-29	-1	0.00015	-1	1.3E-08	-1	0.000014	-1
TRIM8	3.1E-05	-1	5E-28	-1	0.00034	-1	1.3E-13	-1	0.0000081	-1
LTA4H	0.00017	-1	3E-26	-1	3.9E-05	-1	9.5E-09	-1	0.000043	-1
LRRC32	5.9E-07	1	1E-07	1	0.028	0	2.4E-09	1	0.0000018	1
SDC4	1.2E-05	1	6E-07	1	0.017	0	6.7E-11	1	0.0000057	1
YPEL2	0.064	0	8E-15	-1	0.00073	-1	2E-16	-1	1.2E-07	-1
TLR6	1.8E-05	-1	2E-15	-1	0.00013	-1	6.2E-10	-1	0.18	0
YPEL3	0.086	0	4E-17	-1	0.0021	-1	5.5E-13	-1	6.7E-09	-1
FAM117B	0.24	0	2E-12	-1	0.017	0	5.7E-13	-1	1.4E-12	-1
KCTD12	0.04	0	1E-09	-1	0.18	0	1.8E-05	-1	0.14	0
RGS2	0.17	0	8E-24	-1	0.99	0	8E-07	-1	0.085	0
CLEC5A	1.3E-07	-1	0.24	0	0.4	0	0.063	0	3.8E-07	1
JDP2	7.5E-07	-1	0.61	0	0.005	-1	1E-08	-1	0.0000032	1

Reg: regulation; 0: not differentially expressed; 1: up-regulated; -1: down-regulated, in fungal *versus* bacterial infected immune cells. Adjusted P-values (P-val) below 0.05 are considered significant.

Table 7 – Classifier performances

Single		**Combined**	
Dataset	**Accuracy**	**Dataset**	**Accuracy**
Smeekens	0.99	D & S	1
Dix	1	D & K	0.99
Czakai	0.92	D & C	0.99
Klassert	0.97	D & Sa	0.87
Saraiva	0.63	S & K	0.99
		S & C	0.99
		S & Sa	0.81
		K & C	0.98
		K & Sa	0.87
		C & Sa	0.847
Average	**0.90**		**0.93**

D: Dix; S: Smeekens; C: Czakai; K: Klassert; Sa: Saraiva

Gene set enrichment analysis was performed to determine if pathways were significantly represented in the gene signature. Two pathways were significantly enriched: NOD-like receptor signaling and Toll-like receptor signaling with corrected p-values of 0.005 and 0.013, respectively. Both pathways shared the genes *IL1β* and *IL6*, differing only in the presence of *CXCL1* in the NOD-like receptor signaling and TLR6 in Toll-like receptor signaling. Only using genes that were differentially expressed and consistently upregulated, in at least 4 datasets (131 genes), also showed an increased enrichment of the identified pathways (P=1.5E-5 and P=4.4E-5, NOD-like and Toll-like signaling pathways, respectively) as well as the JAK-STAT signaling pathway (P=1.6E-4). The gene lists of both consistently up and downregulated differentially expressed genes can be found in Table A1 in the Appendix.

In summary, combining classifiers improved the consistency of the gene signatures generated for discriminating infected from non-infected samples.

Additionally, genes involved in the regulation of inflammatory signaling pathways were significantly enriched.

4.2. Discriminating fungal from bacterial infected samples

The main goal of this thesis was to identify host genes and pathways that discriminated fungal from bacterial infections, irrespective of the immune cell populations. Feature selection and classification was performed on each individual dataset in Table 2 (except for the "Mattingsdal" dataset which was used for validation) using 100 randomly assigned training sets. As before, a list of 30 genes was generated in each run which best discriminated samples infected with fungal from bacterial pathogens. The pairwise overlap of the 100 generated gene lists from each single classifier returned an average value of 1.09 (1σ=0.35). In this case, a consistent gene signature capable of distinguishing fungal from bacterial infections, independent of leukocyte type was the main objective. To this purpose, the combined classifier approach was employed and returned an average POL of 1.57 (1σ=0.46). The combination of classifiers significantly (P=2.2E-16 using a two-sided Kolmogorov-Smirnov test) improved the POL by 43% when compared to single classifiers. As performed during the infected versus healthy sample analysis, genes that were present in all single classifier gene lists and selected in at least 40% of the runs were extracted. The same procedure was applied to the gene lists from all combined classifiers gene lists. The single and combined classifier final gene lists were comprised of 72 and 88 genes, respectively (Table A2 in the Appendix). Comparison of the two reduced gene lists revealed that out of the 72 single classifier genes, 46 also were selected by the combined approach corresponding to 64%. Gene set enrichment analysis of the resulting lists (intersection, and both classifier-specific approaches) was performed to obtain a functional overview of each of them. Gene Ontology terms such as immune response (P=4.2E-3), purine nucleotide metabolic process (P=2.7E-3) and cell death (P=1.2E-3) were enriched using the combined classifier specific gene list. Single classifier specific genes were enriched in negative regulation of catalytic activity (P=5.5E-3). Special interest was given to genes which were consistently

selected across all classification runs. No KEGG pathways were significantly enriched in either gene list.

For each dataset, differentially expressed genes in fungal *versus* bacterial stimulated samples were determined. In order to improve consistency of the gene signature, genes not differentially expressed in at least 4 datasets were discarded from further analysis. In the gene signature, only 19 genes met this criterion and, out of these, only 12 were consistently up regulated in the datasets. Table 8 shows the list of genes and their regulation across all datasets and how many times they were selected during feature selection. The respective adjusted p-values are shown in Table A3 in the Appendix. Measuring increased gene expression rather than inhibition is easier. Therefore, consistently up regulated genes were ranked as the highest followed by those consistently down regulated. Inconsistently regulated genes across datasets were ranked first if, in most datasets, they were up regulated followed by those down-regulated. Within each of these groups (up and down regulated) they were further ranked according to the average number of runs in which they were chosen during each combined classification problem. The higher the frequency of a gene being used to discriminate between infections is, the higher is the consistency and robustness of the final gene signature.

Table 8 – Refined gene signature and regulation across datasets in fungal *versus* bacteria.

Gene Symbol	Dix	Smeekens	Saraiva	Klassert	Czakai	Average Nº runs
HMOX1	1	1	1	1	1	71
CCR1	1	1	1	1	1	61
GLA	1	1	1	1	1	48
TNFSF14	1	1	1	1	1	60
TBC1D7	1	1	1	1	1	65
SPRY2	1	1	1	1	1	63
EGR2	1	1	1	1	1	60
BCAR3	1	1	1	1	1	59
PAPSS1	1	1	1	1	1	58
RRAGD	1	1	1	1	1	55
DHRS9	1	1	1	1	1	54
SDSL	1	1	1	1	1	53
RNF144B	-1	-1	-1	-1	-1	67
ADA	-1	-1	-1	-1	-1	56
SCARB2	1	1	1	1	-1	64
SOWAHC	1	1	1	1	-1	55
BLVRA	-1	1	-1	-1	-1	64
EDN1	1	1	1	1	-1	97
TNFSF15	1	1	1	1	-1	53

(0: not differentially expressed; 1: up-regulated; -1: down-regulated, in fungal *versus* bacterial infected immune cells)

The combined classifiers had a mean value of 0.96 for sensitivity (Sens), 0.97 for specificity (Spec), 0.97 for positive predictive value (PPV), 0.96 for negative predictive value (NPV) and 0.96 for accuracy (Acc). Comparing performances values between single and combined classifiers showed, at most, a difference of one percent (e.g. accuracy of single classifier for "Smeekens" dataset = 96 %; accuracy of combined classifiers using the "Smeekens" dataset = 97 %). Full performance values for single and combined classifiers are shown in Table 9 and Table 10, respectively.

Summarizing, the combination of classifiers increased the consistency and robustness of the gene signature for discriminating fungal from bacterial infections in human immune cells. GO terms related to immune response and cell death were significantly enriched in the gene signature.

Table 9 - Single Classifier Performances

	Smeekens	Saraiva	Klassert	Dix	Czakai	Average
Sensitivity	0.94	0.94	0.90	0.95	0.97	0.94
Specificity	0.98	0.97	0.99	0.96	0.99	0.98
PPV	0.98	0.97	0.99	0.95	0.99	0.98
NPV	0.94	0.94	0.92	0.95	0.97	0.94
Accuracy	0.96	0.95	0.95	0.95	0.98	0.96

(PPV: Positive Predictive Value; NPV, Negative Predictive Value)

Table 10 - Combined Classifier Performances

	S_Sa	S_K	S_C	Sa_K	Sa_C	K_C	D_S	D_Sa	D_K	D_C	Mean
Sens	0.97	0.95	0.95	0.95	0.98	0.97	0.95	0.97	0.94	0.97	0.96
Spec	0.98	1.00	0.99	0.93	0.93	1.00	0.97	0.94	0.97	0.97	0.97
PPV	0.98	1.00	0.99	0.93	0.94	1.00	0.97	0.93	0.97	0.97	0.97
NPV	0.97	0.95	0.95	0.95	0.98	0.98	0.95	0.97	0.94	0.97	0.96
Acc	0.97	0.97	0.97	0.94	0.96	0.99	0.96	0.95	0.96	0.97	0.96

(*S: Smeekens; Sa: Saraiva; K: Klassert; C: Czakai; D: Dix; Sens: sensitivity; Spec: specificity;PPV: Positive Predictive Value; NPV, Negative Predictive Value)

4.3. *In silico* validation of the gene signature discriminating fungal from bacterial infected samples

To determine the generalizability of the gene signature in predicting the source of infection irrespective of leukocyte type the biomarker list was applied to a new, "unseen" dataset ("Mattingsdal"). To this purpose, microarray data of human monocytes challenged with LPS or *A. fumigatus* was used (www.ebi.ac.uk/arrayexpress, E-MEXP-1103). Samples were extracted 6 h post-infection. Only the consistently up regulated and differentially expressed genes from the biomarker list (Table 8) in all datasets were considered in this process. Performance results are shown in Table 11. All models predicted the fungal infected samples with more than 73% accuracy, yielding an average of 87%. Mean sensitivity and specificity values were 79% and 100%, respectively. Misclassified samples belonged to bacteria-stimulated monocytes. The results clearly show that the used gene signature was capable of discriminating the infecting pathogens with a high level of accuracy.

Table 11 – Gene signature performance on unseen data

	Dix	Klassert	Smeekens	Czakai	Average
Sens	0.71	0.83	1	0.63	0.79
Spec	1	1	1	1	1
PPV	1	1	1	1	1
NPV	0.67	0.83	1	0.5	0.75
Accuracy	0.82	0.91	1	0.73	0.87

(Sens: sensitivity; Spec: specificity; PPV: Positive Predictive Value; NPV, Negative Predictive Value)

4.4. Monocyte-specific fungal immune response

Whilst it is vital to identify genes whose expression is consistently differentiating between fungal and bacterial infections irrespective of leukocyte type, it is also important to determine which genes consistently discriminate fungal from bacterial infections in specific immune cell populations. The host immune response consists of many players which exist at different ratios (Bhushan 2002). Thus, it is also important to understand how specific immune cells respond to specific infecting pathogens and identify possible differences. To this purpose, classifiers using only PBMC datasets ("Smeekens" and "Saraiva") and monocyte datasets ("Klassert" and "Mattingsdal") as cell population were combined. The "Dix" dataset due to its high heterogeneity (whole blood) and the "Czakai" dataset due to its high specificity (dendritic cells) were disregarded.

A list of 30 genes was generated in each classification run which best discriminated samples infected with fungal from bacterial pathogens. The averaged POL of the 100 generated gene lists of single *versus* single, single *versus* combined and combined *versus* combined classifiers returned values of 0.78 (1σ=0.41), 1.09 (1σ=0.48) and 1.64 (1σ=0.49), respectively. The mean

POL of combined *versus* single already showed an increase in almost 40% when compared to single *versus* single, increasing to 100% when calculating the POLs between combined classifiers. Comparing these results to the previous in the study of infected *versus* non-infected samples (*see* 4.1), an increase in POL of almost 60% was obtained.

Next, determination of pathways that were significantly enriched in both the single and combined classifier gene lists was performed. To this purpose, the genes that were not selected in at least 20% of the total number of all runs of each classifier (single or combined) were discarded. A total of 175 and 164 genes, for single and combined classifiers, respectively, remained (Table A4 in the appendix). The enriched gene sets of single and combined gene signatures are shown in Table 12 and Table 13, respectively. Only one enriched KEGG pathway of the combined classifier gene list was not present in that of the single classifier – the lysosome pathway. For each dataset, differentially expressed genes in fungal *versus* bacterial infected samples was calculated. Intersection of differentially expressed genes was performed not only for all datasets but also based on immune cell population. This resulted in the generation of 3 gene lists (cell population independent, PBMC-specific and monocyte–specific).

The intersection of the differentially expressed genes across all datasets (cell population independent) resulted in a list of 13 genes (*ST3GAL5*, *HMOX1*, *LGALS9*, *GLA*, *HAVCR2*, *TBC1D9*, *ACADVL*, *BCAR3*, *RHOU*, *MGAT2*, *CCL23*, *RGS1*, *SPRY2*) and did not reveal any enriched pathways.

Table 12 – Significantly enriched gene sets for the list of genes from single classifiers.

Pathway	P-value
Chemokine signaling	2.3E-17
Cytokine-cytokine receptor interaction	8.6E-15
Toll-like receptor signaling	2.7E-5
Jak-STAT signaling	7.2E-4
Chronic myeloid leukemia	0.0011
Leukocyte transendothelial migration	0.011
Natural killer cell mediated cytotoxicity	0.192
B cell receptor signaling	0.031
Fc epsilon RI signaling	0.035
Intestinal immune network for IgA production	0.042

Table 13 - Significantly enriched gene sets for the list of genes from combined classifiers.

Pathway	P-value
Toll-like receptor signaling	2.2E-4
Cytokine-cytokine receptor interaction	3.1E-4
Lysosome	0.014
Chemokine signaling	0.027
Jak-STAT signaling	0.042

As stated before, monocytes are vital players in the control of infection, either by promoting inflammation or by differentiation into other immune cells (Shi and Pamer 2011). The processes which they influence, however, can be distinct to those of other more present immune cells such as lymphocytes and may be "masked" due to the overwhelming expression of said cells. Intersecting differentially expressed genes of the datasets encompassing solely monocytes resulted in a list of 720 genes, whilst the intersection of datasets comprised of PBMCs resulted in a list of 57 genes. The enriched gene sets for PBMC-specific and monocyte-specific differentially expressed genes are shown in Table 14. The enriched gene sets in all groups suggested that genes coding for the lysosome were specifically induced by monocytes during a fungal challenge. To note, the combined classifier-originated gene list also showed an enrichment of genes coding for the lysosome. Additionally, the differentially expressed and up-regulated genes (in fungal *versus* bacterial) from the monocyte datasets (Klassert and Mattingsdal) were intersected and gene set enrichment tests were performed. Only two pathways were significantly enriched – the lysosome and Toll-like receptor signaling (P=3.2E-4 and 0.015, respectively). This strengthens the initial finding that cell type specific gene expression is still captured when combining classifiers, without the requirement of performing a cell type specific analysis beforehand. Performing gene set enrichment tests on differentially expressed genes from cell type specific datasets produced the same results.

The lysosome gene set was comprised of 123 genes out of which 13 were differentially expressed and upregulated in both monocyte datasets.

Gene set enrichment was also performed on the gene list that resulted in the intersection of differentially expressed and up regulated genes considering only the datasets of stimulated PBMCs. Results are shown in Table 15.

Table 14 - Enriched gene sets of PBMC-specific and monocyte-specific differentially expressed genes in fungal *versus* bacterial infection

PBMC-specific		**Monocyte-specific**	
Gene set	P-value	Gene set	P-value
Jak-STAT signaling	0.0011	Toll-like receptor signaling	2.5E-5
Toll-like receptor signaling	0.0035	NOD-like receptor	3.5E-5
Cytokine-cytokine receptor interaction	0.046	Hematopoietic cell lineage	2.4E-4
		Cytokine-cytokine receptor interaction	3.9E-4
		Chemokine signaling	0.0018
		Jak-STAT signaling	0.0035
		Lysosome	**0.0044**
		Cytosolic DNA-sensing	0.0049
		MAPK signaling	0.0054
		Adipocytokine signaling	0.016

Table 15 - Enriched pathways using PBMC-specific differentially expressed and up regulated genes in fungal *versus* bacterial induced immune cell response.

Pathway	P-value
Jak-STAT signaling	4.8E-4
Cytokine-cytokine receptor interaction	0.026
Toll-like receptor signaling	0.026

In addition it was also interesting to determine the differences of immune cell population specific responses in fungal *versus* bacterial infections. The identification of pathways that were not enriched using the independent-cell type gene signature suggested that cell type specific responses were possibly being "masked" by the net effect of the whole immune system as expected. Since low-present monocytes (when compared to neutrophils) play an important role in the identification of pathogens (Lauvau et al. 2015) the monocyte-dependent responses were studied in detail.

4.5. Real time quantitative reverse transcription PCR analysis of monocytes challenged with fungal and bacterial pathogens and cell wall representatives of each microorganism

In order to determine if the gene expression patterns are, in fact, a result of the pathogen´s presence, experimental validation was performed via *in vitro* stimulation of monocytes by fungal and bacterial pathogens as well as cell wall representatives of each pathogen (zymosan or LPS, respectively) following RT-qPCR. Experimental procedures and statistical analysis of the RT-qPCR results were performed by collaboration partners at the Host Septomics group led by Dr. Hortense Slevogt and are described below.

The gene signature obtained from the combined classifiers that discriminated fungal from bacterial infections, independent of cell population,

contained four lysosome-related genes. These were selected for real time RT-qPCR analysis. The genes were Galactosidase A (*GLA*), Scavenger receptor class B member 2 (*SCARB2*), Niemann-Pick disease, type C1 (*NPC1*) and CD164 molecule (*CD164*). The real-time RT-qPCR plots are shown in Figure 15.

Additionally, due to their consistent expression across all datasets, five further genes were also selected for RT-qPCR, however, unrelated to the lysosome. These were the BAG family molecular chaperone regulator 3 (*BAG3*), the fatty acid binding protein 5 (*FABP5*), the Peroxisome proliferator-activated receptor gamma (*PPARG*), the heme oxygenase 1 (*HMOX1*) and the C-C chemokine receptor type 1 (*CCR1*) (Figure 18).

The complete table of the real-time RT-qPCR mean expression values across conditions and corresponding p-values is shown in Table A5 in the appendix.

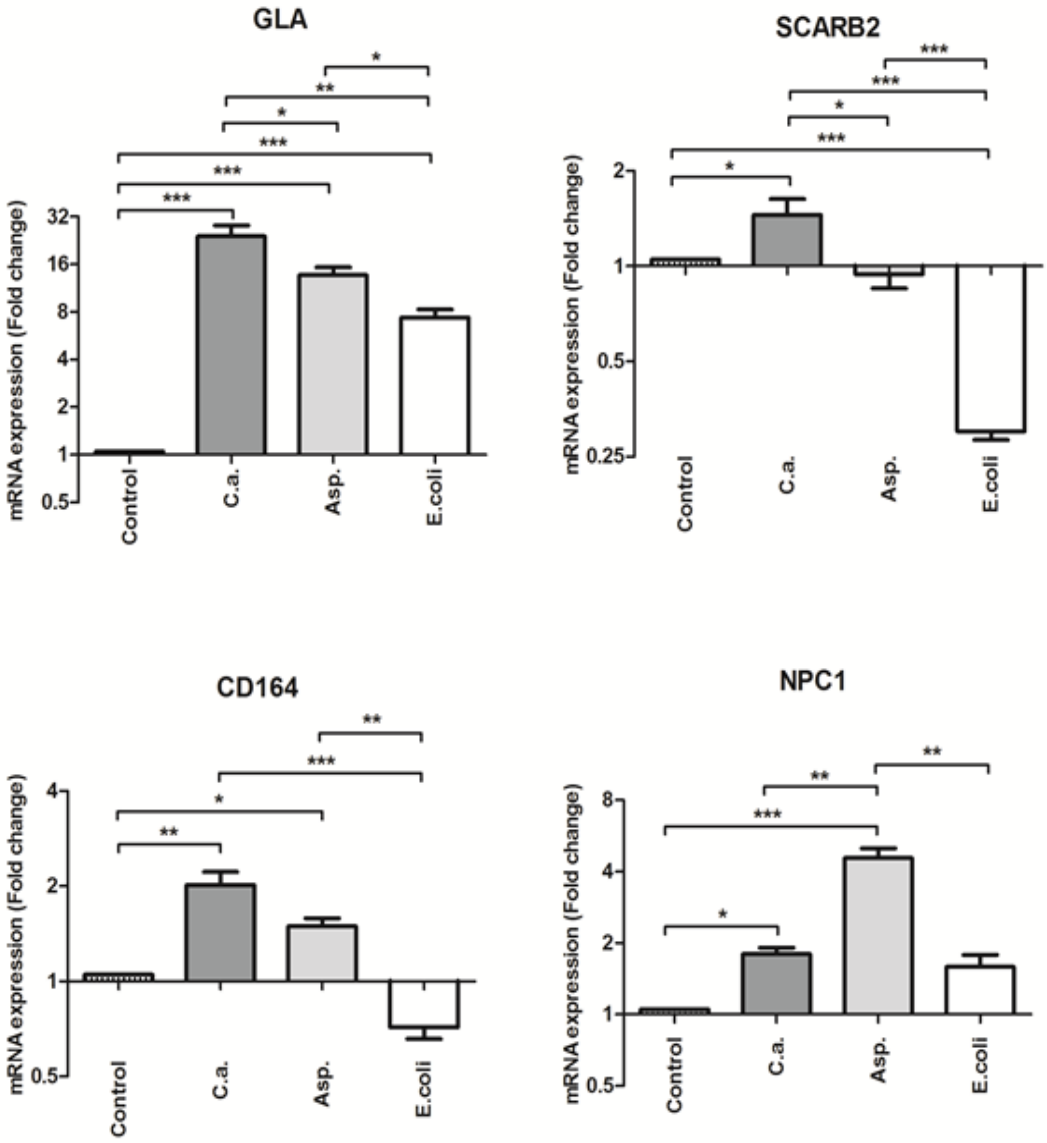

Figure 15 – Lysosomal genes are significantly upregulated in monocytes challenged with fungal pathogens in comparison to bacterial. Relative mRNA expression of *GLA*, *SCARB2*, *CD164* and *NPC1* after monocyte stimulation with *Candida albicans* (C.a.), *Aspergillus fumigatus* (Asp.) and *Escherichia coli* (E. coli). Data were obtained from four independent experiments, each performed with cells from different donors. Results are presented as mean ± SE of the fold change relative to the control (unstimulated cells). Shown is also the statistical significance after repeated measures One-Way ANOVA with Bonferroni correction (***p<0.001; **p<0.01; *p<0.05).

Almost all lysosome related genes showed a significant increase in their expression when the fungi-stimulated group was compared to either the unstimulated controls and/or to the bacteria-challenged samples. *GLA* was significantly up-regulated by both fungal pathogens when compared to either control or *E. coli*-stimulated monocytes. Besides this, the *C. albicans* stimulated monocytes also showed, although not as strongly, a significant increase when compared to *A. fumigatus*. *SCARB2* was highly significantly up-regulated in *C. albicans*-stimulated monocytes when compared *to E. coli*. It also showed a significant increase when compared to controls and to *A. fumigatus*-challenged monocytes. In *E. coli* stimulated monocytes, *SCARB2* was significantly down regulated when compared to controls. *NPC1* showed significant increased expression in *A. fumigatus*-stimulated monocytes when compared to all other challenges. *C. albicans*-stimulated monocytes also showed significant increase of *NPC1* gene expression when compared to controls. Lastly, *CD164* showed significant increased expression in both fungal-challenged monocytes when compared to *E. coli* and controls.

In summary, the expression of the selected genes to be either specifically or significantly more up-regulated in monocytes stimulated by fungal pathogens when compared to monocytes stimulated by bacterial pathogens confirming them as potential biomarkers for fungal *versus* bacterial induced systemic infection could be validated.

To further strengthen the suggestion that the lysosome genes were, in fact, significantly different in monocytes when compared to PBMCs, an additional set of experiments was performed. This consisted of stimulating, in parallel, monocytes and PBMCs from blood of the same donors, with *C. albicans*, *A. fumigatus* and *E. coli*. The fungi-specific pattern observed for lysosome-related genes in monocytes was less evident in PBMCs (Figure 16). The results obtained are in agreement with the readouts from both microarray and RNA-Seq datasets (monocytes *versus* PBMCs). In turn, this could help explain why the lysosome pathway was significantly enriched only using the monocyte datasets.

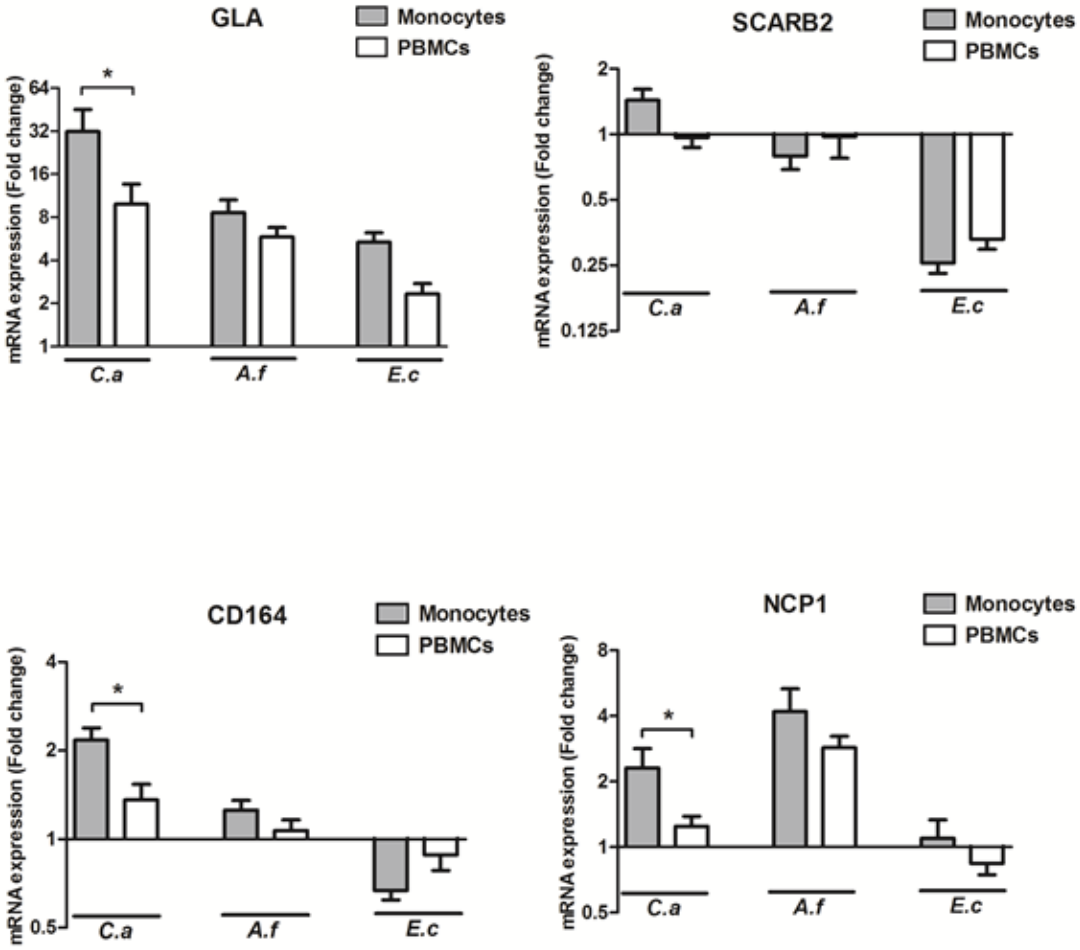

Figure 16 – Lysosomal genes are significantly differentially expressed in monocytes when compared to PBMCs upon challenge with fungal and bacterial pathogens. Comparison of lysosomal-related gene expression levels in pathogen- stimulated PBMCs and monocytes. The validation experiments were repeated with 4 additional donors, from which both monocyte and PBMC fractions were isolated. These were then separately and simultaneously stimulated with *C. albicans* (C.a.), *A. fumigatus* (A.f.) and *E. coli* (E.c.) as detailed in the material and methods section. Results are presented as mean ± SE of the fold change relative to the control (unstimulated cells). Shown is also the statistical significance after repeated measures Two-Way ANOVA with Bonferroni post-hoc test (*$p<0.05$).

Real time RT-qPCR was also performed on monocytes challenged with cell wall components representative for the above-mentioned pathogens (zymosan for fungal pathogens and LPS for gram-negative bacteria). The RT-qPCR barplots

of the lysosome-related genes are shown in Figure 17 and the table with the full results in the appendix (Table A5).

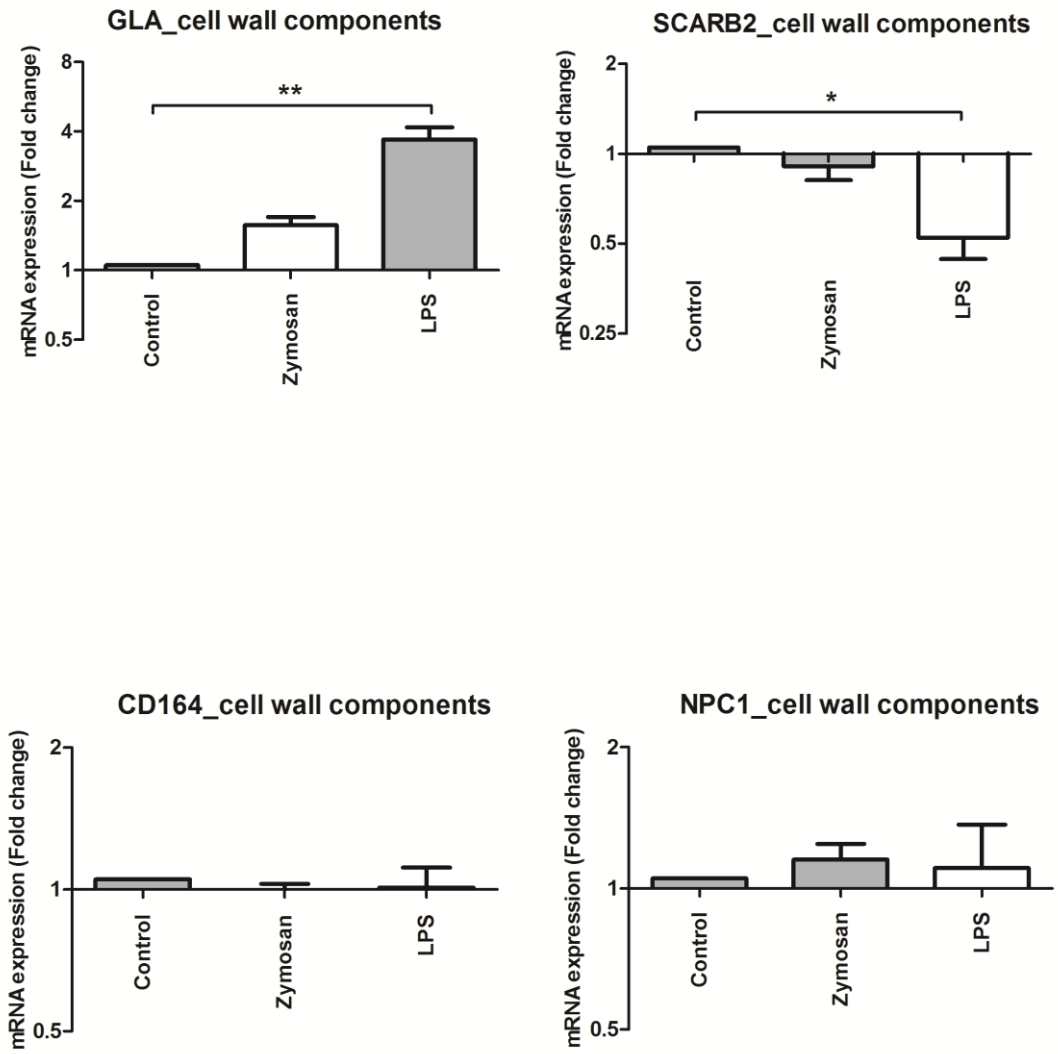

Figure 17 – *GLA* and *SCARB2* genes are significantly differentially expressed (compared to control) in monocytes challenged with bacterial LPS but not with fungal zymosan. Relative mRNA expression of *GLA*, *SCARB2*, *CD164* and *NPC1* after monocyte stimulation with zymosan (1 µg/ml) and LPS (50 ng/ml). Data were obtained from four independent experiments, each performed with cells from different donors. Results are presented as mean ± SE of the fold change relative to the control (unstimulated cells). Shown is also the statistical significance after repeated measures One-Way ANOVA with Dunnett post-hoc test (*$p<0.05$).

GLA gene expression in monocytes was higher expressed in both stimuli (zymosan and LPS) when compared to controls. However, this increase was only significant when monocytes were stimulated with LPS. When comparing the effects of cell wall representatives to inactivated pathogens, only LPS matched that of *E. coli*. Zymosan did not significantly induce *GLA* expression when compared to fungal pathogens. In the case of *SCARB2*, both zymosan and LPS inhibited the expression when compared to unstimulated monocytes. Again, only LPS was capable to inhibit significantly *SCARB2* expression which is in agreement with the results of monocytes challenged with *E. coli*. Monocytes challenged with zymosan exhibited decreased expression of *SCARB2* compared to the controls although not significant and this was opposite of the gene expression profile of monocytes challenged with *C. albicans* (significantly up regulated - Figure 15). *NPC1* expression, although not significant, was increased in both zymosan and LPS challenged monocytes when compared to controls which exhibited the same trend in expression when compared to monocytes challenged with the inactivated pathogens Finally, *CD164* did not show any significantly increased or decreased expression when comparing different pathogenic cell wall components to each other or to controls. This result contrasts to that of monocytes challenged by fungal and bacterial pathogens, which showed a significant increase of *CD164* gene expression during *C. albicans* and *A. fumigatus* and a decrease during *E. coli* stimulations.

In summary, RT-qPCR results show a significant increase in expression of all four lysosome related genes when challenged by both fungal and bacterial pathogens. The exceptions are the gene expressions of *SCARB2* and *CD164*, where a significantly decreased expression was shown during bacterial challenge. Monocytes challenged with cell wall representatives of fungal (zymosan) or bacterial (LPS) species revealed that *SCARB2*, *GLA* and *NPC1* genes (but not *CD164*) displayed a similar expression pattern as the one seen when using inactivated pathogens. However, *GLA* and *SCARB2* expression was only significantly different to controls when challenging monocytes with LPS.

The results for non-lysosome related genes show that all genes were significantly up-regulated in monocytes challenged with either *C. albicans* or *A.*

fumigatus (Figure 18). Inversely, when monocytes were challenged with *E. coli*, these genes, with the exception of *BAG3*, were either significantly (*CCR1* and *HMOX1*) or non-significantly (*FABP5* and *PPARG*) down-regulated (Figure 18).

In summary, monocytes challenged with fungal and bacterial pathogens clearly showed, with the exception of BAG3, different expression patterns of the non-lysosome related genes.

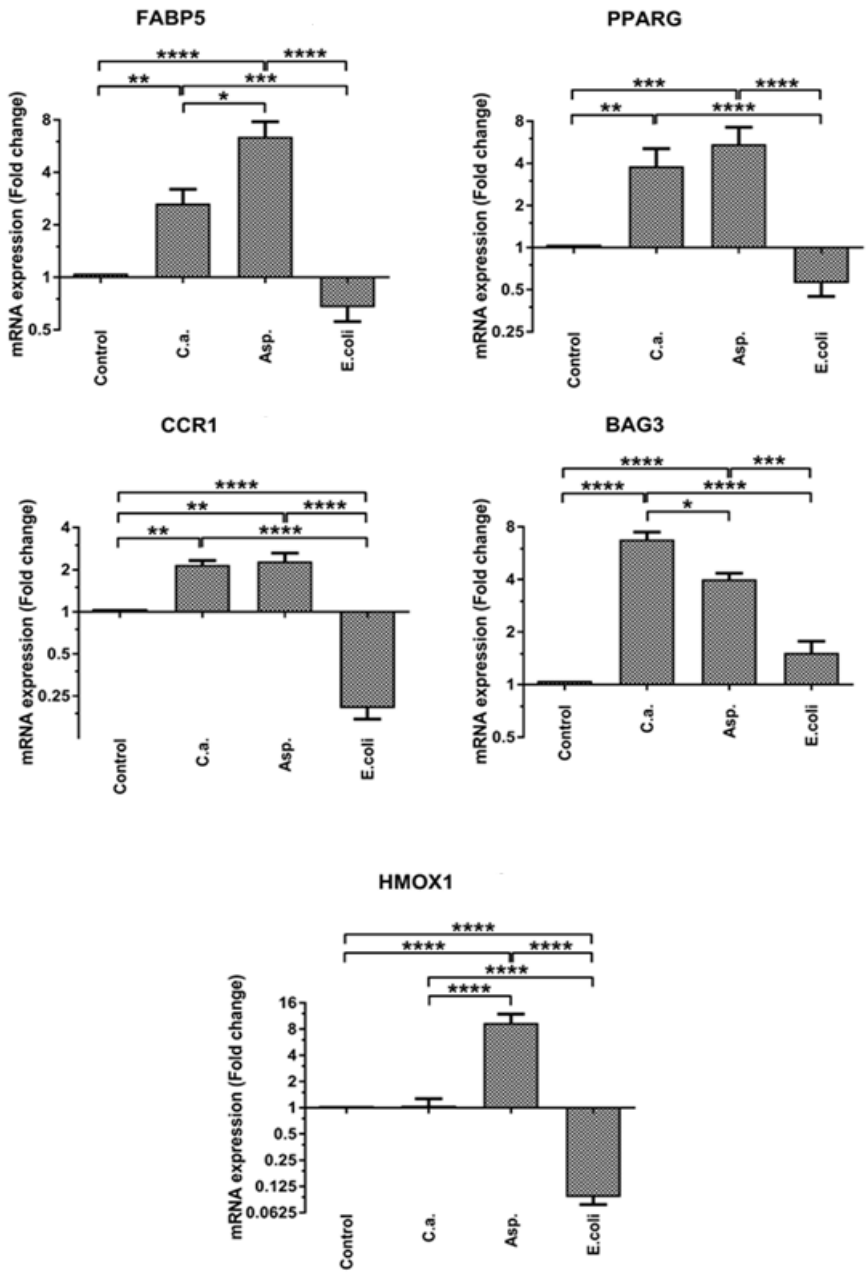

Figure 18 – Fungal pathogens induce increased expression of selected non-lysosome related genes contrastingly to bacterial pathogens. Relative mRNA expression of *FABP5*, *PPARG*, *CCR1*, *BAG3* and *HMOX1* after monocyte stimulation with *Candida albicans* (C.a.), *Aspergillus fumigatus* (Asp.) and *Escherichia coli* (E. coli). Data were obtained from four independent experiments, each performed with cells from different donors. Results are presented as mean ± SE of the fold change relative to the control (unstimulated cells). Shown is also the statistical significance after repeated measures One-Way ANOVA with Bonferroni correction (***p<0.001; **p<0.01; *p<0.05).

5. Discussion

As stated in the objectives of this thesis, the main goal was to generate a robust and consistent gene signature capable of discriminating fungal from bacterial infection in the human host. To this purpose a newly developed method capable of increasing consistency in the generated gene signatures across several experimental assays was employed.

MILPs were used for extending the pure classification problem. Allowing integer (in this case binary) requirements on a subset of the variables makes the classification tool much more flexible and powerful. Unlike Linear Programming, MILPs enable the modeling of discrete variables and constraints like e.g. the restriction on a small subset of the features used for classification or allowing a small amount of exceptions in the training set. Using MILPs enabled linking two classifiers, constrained to use the same set of features (genes). Additionally, the combination of classifiers did not require prior preprocessing for minimizing technical differences such as means, ranges and standard deviations, for datasets generated from diverse platforms (Johnson et al. 2007). In contrast to Linear Programming problems, the major disadvantage of MILPs is their complexity, which requires intensive computational power. But in the meanwhile, there exist very efficient solvers which are fast enough to find at least nearly optimal solutions within a given time limit.

5.1. Combining classifiers improves consistency of gene signatures

5.1.1. Infected *versus* Non-infected

To determine the consistency of gene signatures by combining classifiers this novel method was employed to distinguish infected from non-infected samples. The pairwise overlap of the combined classifiers increased by 0.64 when compared to single classifiers. This corresponds to an increase in consistency of 42%. Intersection of the gene lists between single classifiers and

between combined classifiers also showed an average improvement of 50% (i.e. 8 and 12% for single and combined classifiers, respectively).

A decreased performance when combining classifiers was expected, since during feature selection the individual classifiers were "forced" to choose the same genes that discriminated between infected and healthy samples. Interestingly, the combined classifiers displayed a higher accuracy than single classifiers (92% to 90%, respectively).

5.1.2. Fungal *versus* bacterial independent of cell population

The high consistency and accuracy of the combined classifier approach gave confidence to pursue the main goal of the study – generate a consistent gene signature capable of discriminating fungal from bacterial infections in the human host, irrespective of the leukocyte cell population. This becomes greatly important in the context of sepsis where a rapid and accurate identification of the underlying pathogen improves the chance of survival by an appropriate subsequent treatment of the patient. Employing this novel approach, a gene signature was generated capable of distinguishing between fungal and bacterial infected samples with an average accuracy of 96%. The pairwise overlap (number of genes consistently selected across runs) was 43% higher than that of the single classifiers which showed an immediate improvement in feature consistency without introducing prior knowledge into the feature selection and classification problem. This novel method showed no decrease in performance when compared to single classifiers. From the genes consistently selected by single classifiers, 64% were also identified by the combined classifier approach which demonstrated that combining classifiers did not result in a completely different gene list.

The combined classifier approach produced a consistent list composed of 75 genes. Following differential expression analysis and imposing that genes should be differentially expressed in at least 4 datasets, decreased the list to 19 genes of which 12 were consistently up-regulated in all datasets. This refined

gene list was tested on a new "unseen" dataset ("Mattingsdal") and classified the samples with an average accuracy of 87%. Interestingly, misclassification of samples only occurred for bacterial infected samples, shown by a perfect score in terms of sensitivity for fungal infections. This may have an advantage for clinical transfer as the comparably less often occurring fungal systemic infection needs to be precisely identified during sepsis.

In summary, combining classifiers leads to improved consistency of gene signatures with even higher levels of accuracy when compared to single classifiers independently of the immune cell population.

5.1.3. Fungal *versus* bacterial dependent on cell population

Lastly, employment of the novel method on datasets composed of relatively heterogeneous populations of immune cells (PBMCs and monocytes) demonstrated its ability to additionally identify cell-specific signatures. As before in the previous results of this study, a higher pairwise overlap of genes was observed in the gene lists generated from combining classifiers when compared to single classifiers. Combining classifiers for discrimination between fungal and bacterial infections independently of immune cell population, such as PBMCs and monocytes, generated a gene signature enriched for several immune signaling pathways. Among them the lysosome gene set was observed to be specific for monocytes. This was ascertained by the comparison of the enriched signaling pathways of differentially expressed genes in cultures of monocytes against PBMCs, both challenged with fungal or bacterial pathogens. The results were further experimentally validated by employing qPCR and analyzing a set of lysosome-related genes that were either selected by the combined classifier or uniquely differentially expressed in the monocyte challenged datasets. As shown in the results, all the lysosome-related genes (*GLA*, *SCARB2*, *NPC1* and *CD164*) exhibited a significant increase in their expression after fungal challenge when compared to bacterial stimulation, indicating a fungal-specific response by monocytes (Figure 15). Similar results were obtained for other, non-lysosome related genes that were part of the fungal-specific signature and

also these genes could be validated by qPCR (Figure 18). These genes included *BAG3*, *PPARG*, *FABP5*, *HMOX1* and *CCR1*.

5.2. Gene set enrichment analysis of combined classifiers

Inflammatory signaling pathway genes are enriched in infection and help to discriminate fungal from bacterial infections independent of immune cell population.

Gene set enrichment tests revealed that the nucleotide-binding oligomerization domain-like receptor (NLRs) signaling and Toll-like receptor (TLR) signaling pathways were significantly enriched in the gene signature for discriminating infected from non-infected samples. This became even more evident when gene set enrichment tests were performed using the whole list of differentially expressed genes in at least 4 datasets. TLRs and NLRs have long been studied and reviewed in the context of immune response and have been proven to play key roles in pathogen-associated molecular pattern (PAMP) and damage-associated molecular pattern (DAMP) sensing (Netea et al. 2004a, Takeda and Akira 2005, Kanneganti et al. 2007, Rietdijk et al. 2008, Chen et al. 2009). Not only are they part of the pattern recognition receptor family, but they also interact with each other to modulate other cellular processes such as inflammation and cell death. Indeed, the NLR signaling cascade is very similar to that of the TLR since they share several downstream signaling pathways such as Mitogen-activated protein kinase (MAPK) and NF-kB (Chen et al. 2009) and this is corroborated by the results. It has also been suggested that the interplay between these two signaling pathways is important in the clearance of infection due to their synergistic cooperation (Fritz et al. 2006). Netea and colleagues (Netea et al. 2005) showed that NOD2 and TLR2 synergized to promote cytokine production when induced by peptidoglycan (a key component of bacterial cell wall).Toll-like receptors are mainly located on the cell surface or the plasma membrane of intracellular organelles (Trinchieri and Sher 2007) whilst NOD-like receptors are mainly located in the cytoplasm (Shaw et al. 2009). NLRs also have been shown to play a role in inflammasome assembly

and autophagy (Kim et al. 2016). These receptors are found in several immune cells such as lymphocytes and macrophages (Franchi et al. 2009). The importance of these pattern recognition receptors (PRRs) in the mounting of an inflammatory response upon recognition of DAMPs or PAMPs has also been demonstrated by its impact on the NF-Kβ and MAPK signaling pathways (Netea et al. 2004b). This took special interest when the selection of genes (*IL6* and *IL1β*) was identified by this study. The activation of the above mentioned pathways induces a pro-inflammatory and antimicrobial response via induction of interleukin-6 (*IL6*) and interleukin-1 beta (*IL1β*) (Trinchieri and Sher 2007, Carneiro et al. 2008, Chen et al. 2009, Franchi et al. 2009).

In the list of consistently selected genes using the combined classifier approach several cytokines and chemokines were present such as *CXCL1*, *IL1β* and *IL6*, all of which have been shown to participate in the immune response during infection (Cai et al. 2010, Leal et al. 2010, Netea et al. 2010, 2015, Scheller et al. 2011). Cytokines such as IL6 play a key role in the regulation of production of specific immune cells and their respective response towards infection. IL6 activation can lead to neutrophil and macrophage production (Scheller et al. 2011).

Next, focus was given to genes that, despite not being in the gene signature for discriminating infected from healthy samples, were consistently upregulated in at least four datasets and that were present in the enriched pathways. Amongst others, these included chemokines (C-C motif) ligand 2 (*CCL2*) and ligand 3 (*CCL3*), chemokines (C-X-C motif) ligand 2 (*CXCL2*), nuclear factor kappa-light-chain-enhancer of activated B cells (*NF-κB*), nuclear factor of kappa-light-polypeptide gene enhancer in B cells inhibitor, alpha (*NF-κβIA*), tumor necrosis factor, alpha-induced protein 3 (*TNFAIP3*), interleukins 12 subunits alpha and beta (*IL12A* and *IL12B*, respectively), and interleukin-1 receptor-associated kinase-like 2 (*IRAK2*). Both CCL2 and CCL3 are potent chemoattractants of immune cells to sites of infection. This does not necessarily mean that expression of their respective receptors (e.g. CCR2) are always beneficial during pathogen clearance (Szymczak and Deepe 2009), but their downstream effects rely on the combination of signals depending on the type of infection and

ligands. In a study by Szymczak and Deepe (Szymczak and Deepe 2009), CCR2-deficient mice where infected with a dimorphic fungus – *Histoplasma capsulatum* and showed that when one of its ligands (CCL7) was also neutralized, IL4 and fungal burden were increased. NF-κB, a protein complex, plays an important role in the regulation of transcription of DNA and cytokine synthesis. Its role in the regulation of the immune system in response to infection has also been shown, particularly in the control of the transcription of cytokines and antimicrobial effector cells (Gilmore 2006, Hayden et al. 2006, Kanayama et al. 2015, Biswas and Human Bagchi 2016). Interleukins (IL) are a group of cytokines that are secreted by several leukocytes such as macrophages, neutrophils and dendritic cells and are a key player in the modulation of the immune response. IL12 is a heterodimeric cytokine that is composed of two separate genes (*IL12A* and *IL12B*). This interleukin is involved in the differentiation of naive T cells into Th1 cells and in the stimulation of interferon gamma (*IFNγ*) and *TNFα*, both of which can enhance macrophage activity (Sturge and Yarovinsky 2014). Overall, it enhances the cytotoxic activity of both natural killer (NK) cells and CD8+ T lymphocytes (Langrish et al. 2004, Teng et al. 2015). Last but not least, IRAK2 is involved in the activation of NF-κB and MAPK signaling pathways upon infection. The activation of NF-κB is not performed by IRAK2 alone but rather by its association to IL-1R and the MyD88 signaling complex (Muzio 1997). Apart from this role, it is also involved in the activity of IL1 and several TLRs (Meylan and Tschopp 2008). Gene set enrichment analysis also showed that the selected genes were also linked to processes and signaling cascades that are involved in the immune response towards infection.

In summary, the gene signature obtained from combining classifiers to distinguish infected from non-infected samples showed a significant enrichment of the TLR and NLR signaling pathways. Additionally, analysis of differentially expressed genes in all datasets showed that consistently expressed genes were also related to immune responses towards infection.

Gene set enrichment tests of the gene signature generated from the combined classifiers for discriminating fungal from bacterial infections suggested that the MAPK signaling pathway was increased during fungal infections. The MAPK signaling cascade is highly conserved across species, shown by the high sequence similarities of the pathway´s composing genes (Nishida and Gotoh 1993), and has also been shown to be activated during microbial infections leading to pro-inflammatory signals (Ali et al. 2015). However, how these genes interact in a global scale still requires further study.

5.3. Lysosome pathway is enriched during fungal infection

The gene set enrichment tests performed in less heterogeneous immune cell populations (PBMCs and monocytes) showed similar results to single classifiers (e.g. cytokine-cytokine receptor interaction and Toll-like receptor signaling) with the exception of the lysosome pathway. The lysosome was increasingly enriched when only considering monocyte related datasets. Thus, the genes in the gene signature related to the lysosome were further studied.

5.3.1. Functional relevance of the differentially expressed lysosome-related genes

α-Galactosidase A (GLA) is a glycoside hydrolase enzyme encoded by the *GLA* gene. This enzyme hydrolyses the terminal α-galactosyl moieties (especially the α-1,6 linkage) of glycoproteins and glycolipids. Specially, GLA is a lysosomal enzyme that degrades globotriaosylceramide (Gb3) to lactosylceramide, preventing its accumulation in this compartment (Darmoise et al. 2010). Deficiency of this enzyme (GLA) and accumulation of the glycolipid Gb3 in the lysosome of PBMCs has been shown to contribute to diverse physiopathological alterations such as the continuous pro-oxidative and pro-inflammatory state of these cells (De Francesco et al. 2013). Moreover, a pro-inflammatory role of Gb3 could be demonstrated in that study, which was directly mediated by the TLR4-pro-inflammatory signalling pathway (De Francesco et al. 2013). *Candida*

albicans yeast, among other fungi, binds to TLR4 that recognizes short linear O-bound mannan structures present in the fungal cell wall (Netea et al. 2008). Besides this, the GLA product lactosylceramide has been reported to be very abundant on plasma membranes of phagocytes, being involved in the phagocytosis, chemotaxis and superoxide generation during fungal infection (Jimenez-Lucho et al. 1990, Iwabuchi et al. 2015). Our results show that *C. albicans* and *A. fumigatus* induce a significantly higher expression of the GLA gene than *E. coli*, suggesting the importance of this enzyme in monocytes during fungal infection. Thus, GLA may avoid the accumulation of the glycolipid Gb3 in the lysosome as a protective, anti-inflammatory response mechanism of monocytes. Moreover, the conversion of Gb3 to lactosylceramide, as membrane microdomain of immune cells, may increase phagocytosis and clearance of the fungi. Nevertheless, the relevance of this lysosomal enzyme in fungal infection still needs to be clarified.

Scavenger receptor class B member 2 (*SCARB2*) is a gene whose encoding protein the lysosomal integral membrane protein type-2 (LIMP-2/SCARB2) has been shown to be essential for the normal biogenesis and maintenance of lysosomes and endosomes (Gonzalez et al. 2014). As a lysosomal membrane protein, SCARB2 has been reported to act as an entry receptor for Enterovirus 71 (EV71) leading to its internalization to the lysosome (Yamayoshi et al. 2014). Other scavenger receptors, such as CD36 and SCARF1 (human homologs of the murine C03F11.3 and CED-1, respectively), have been shown to bind *C. neoformans* and *C. albicans* via ß-glucan structures, providing protection against these fungal pathogens in a mice model (Croze et al. 1989). Not much is known about the function of SCARB2 during fungal induced immune responses, but the results suggest that this scavenger receptor, like other similar members of this protein family, may play an important role in fungal recognition and internalization to the lysosome.

In this study, other genes encoding lysosomal transmembrane proteins, *CD164* and *NPC1*, were also analysed. Sialomucin core protein 24, also known as endolyn, is encoded by the *CD164* gene. Croze *et al.* reported endolyn to be involved in the maturation of the endosomal-lysosomal compartment (Croze et

al. 1989), while the Niemann-Pick disease type C1 (NPC1) protein encoded by the *NPC1* gene mediated intracellular cholesterol and sphingolipids trafficking into the late endosome and lysosome (Alam et al. 2012). The NPC1 protein is located in the membrane of late endosomes and lysosomes and it might promote the creation and/or movement of these compartments to and from the cell periphery (Ko et al. 2001). In this study, the up-regulation of *CD164* and *NPC1* in human monocytes specifically after fungal challenge was observed, which again suggests the importance of biogenesis and functionality of the lysosome for fungal clearance in monocytes.

Furthermore, analyses were performed to determine whether the most common fungal and bacterial cell wall components (the fungal ß-glucan and the bacterial lipopolysaccharide, respectively) could explain the differential regulation of the genes in question by the different pathogens. In the case of *GLA*, results suggested that LPS was capable of inducing its expression but not ß-glucan which suggests that the higher expression of this gene in fungal versus bacterial challenge is not dependent on the fungal cell wall component. *NPC1* and *CD164* gene expression showed no significant difference to controls. Again, in these cases the presence of fungal ß-glucan and bacterial LPS do not appear to play a role in the expression of these genes during fungal and bacterial infections. Lastly, it was shown that the *E. coli*-derived LPS resembled the downregulation of *SCARB2* already observed after stimulation with *E. coli* cells. In contrast, the fungal ß-glucan component seemed to have no effect on the regulation of this gene. From these results it was concluded that the bacterial liposaccharide seemed to be responsible for the downregulation of *SCARB2*. Zymosan did not induce a significant increased expression of *GLA* and *NPC1* which suggested that other specific fungal epitopes might induce their expression during fungal infection, especially during *C. albicans* infection (with the exception of *NPC1* which was higher expressed after *A. fumigatus* infection).

5.4. Functional relevance of differentially expressed non-lysosome-direct related genes

Most of the genes further analysed in this study are indirectly associated to proper biosynthesis and functionality of the lysosome during fungal infection. In addition, other mechanisms, such as immune cells recruitment, phagocytosis and nutrient metabolism, are also known to be crucial for a successful fungal killing and clearance by phagocytes. Thus, other genes identified in this study to be fungal-challenge specific are involved in those pathways and might play an important role during fungal infection. For instance, *BAG3* encodes the BAG family molecular chaperone regulator 3 (BAG3) protein which regulates macroautophagy for degradation of polyubiquitinated proteins (Gamerdinger et al. 2009). The peroxisome proliferator-activated receptor gamma (*PPARG*) is a gene expressed in macrophages and encodes a protein that plays a central role in regulating fatty acid storage and glucose metabolism (Tyagi et al. 2011). Fatty Acid Binding Protein 5 (FABP5) is a protein encoded by *FABP5* gene and plays a role in the uptake of fatty acids, transport phenomena and fatty acid metabolism (Moore et al. 2015) Additionally, fatty acid binding proteins play a role in inflammation and have been shown to be down regulated in macrophages infected with *Brucella melitensis* (Wang et al. 2011). It has also been shown that loss of *FABP5* promoted a higher anti-inflammatory response in knock out mice (Moore et al. 2015). The increased expression of *FABP5* suggests that fungal infections induce a higher pro-inflammatory response than during bacterial infections. The *HMOX1* gene encodes heme oxygenase-1 (HO-1), which has been shown to be required for immune cell protection against systemic infections (Silva-Gomes et al. 2013). Primarily, HO-1 degrades heme into biliverdin and carbon monoxide (CO). CO has shown different effects during infection. It supports anti-inflammatory cytokine expression (Piantadosi et al. 2011) but may in turn increase the virulence of the infection due to its immunosuppressive effects (Navarathna and Roberts 2010). Additionally, the availability of free iron as a result of high degradation of heme allows its uptake by pathogens proving to be a nutritional benefit (Navarathna and Roberts 2010). Concluding, a higher *HMOX* expression could be an indicator for an

active immune response and defense to infection. The C-C Chemokine Receptor 1 (CCR1), encoded by the *CCR1* gene, has been shown to be widely expressed in immune cells and it was associated with the maintenance of chemokine gradients during infection (Lionakis et al. 2012). Increased expression of CCR1 and its ligands was shown to be significantly induced in Candida-infected organs of mice leading to increased leukocyte accumulation (Lionakis et al. 2012).

In summary, a gene signature was identified being highly relevant in the inflammatory response of human immune cells due to infection. Besides the genes being present in the biomarker list that distinguished infected from healthy samples, a large number of differentially expressed genes also showed a strong interaction and interdependency and link to the immune response. By integrating the combined classifier approach with distinct differential gene expression analysis in studies with specific immune cell populations (PBMCs and monocytes), genes were identified that were up-regulated in monocytes during fungal infection, much more or exclusively in comparison to bacterial infection. Once fungi are phagocytosed, monocytes display transcriptional and translational reprogramming, adapting their physiology, and killing mechanisms to fungal-derived stressors. The up-regulation of fungi-specific genes was observed, which seemed to be important in the fungal-derived reprogramming. Moreover, the application of the combined classifier approach made it possible, for the first time, to identify lysosome-related gene expression as a monocyte-specific footprint of fungal infections and could possibly be used as a diagnostic marker.

6. Conclusions and perspectives

Gene signatures proposed as biomarkers often lack consistency across studies which was demonstrated by the low pairwise overlaps between single classifiers. Throughout this work, this novel proposed approach showed an improvement solely by linking classifiers across datasets. Consistency increased even further when similar cell-type studies were used. This method can also allow combining more than just 2 datasets at a time, with their inherent increase in runtime and complexity. This approach is generic and enables to integrate diverse datasets. This was achieved solely by constraining the classifiers of each of these datasets to use the same sets of features (e.g. genes). This method also allows the integration of additional information such as protein-protein interaction networks. This could provide additional insight on how these genes are connected and increase the extraction of functional relevance during the generation of gene signatures. Another aspect that should be followed up in future studies concerns the optimization of the number of features to use for the classification problem. Additional analysis on the impact of cell wall components and live pathogens might also provide increased insight into the host's response towards each source of infection. The generation of future gene signatures should take into consideration that immune cells might respond differently to certain pathogens. This was the case of lysosomal related genes that were higher induced during fungal infections when compared to bacterial. Additional studies focusing on immune cell response specificity would increase our understanding of the human host response towards different pathogens and possibly lead to newer biomarkers.

As future work, it would be of interest to apply this method to identify co-infected samples.

Finally, the resulting gene signatures make functionally sense and have the potential to be followed up experimentally paving the way to the clinics.

7. Bibliography

Alam MS, Getz M, Safeukui I, Yi S, Tamez P, Shin J, Velázquez P, Haldar K. 2012. Genomic Expression Analyses Reveal Lysosomal, Innate Immunity Proteins, as Disease Correlates in Murine Models of a Lysosomal Storage Disorder. PLoS ONE 7(10).

Ali SR, Karin M, Nizet V. 2015. Signaling cascades and inflammasome activation in microbial infections. Inflammasome 2(1):7–12.

Allantaz F, Cheng DT, Bergauer T, Ravindran P, Rossier MF, Ebeling M, Badi L, Reis B, Bitter H, D'Asaro M, et al. 2012. Expression profiling of human immune cell subsets identifies miRNA-mRNA regulatory relationships correlated with cell type specific expression. PLoS ONE 7(1).

Arias MA, Santiago L, Costas-Ramon S, Jaime-Sánchez P, Freudenberg M, Bagüés JD, P M, Pardo J. 2017. Toll-Like Receptors 2 and 4 Cooperate in the Control of the Emerging Pathogen Brucella microti. Frontiers in Cellular and Infection Microbiology 6.

Atkinson A.J. J, Colburn WA, DeGruttola VG, DeMets DL, Downing GJ, Hoth DF, Oates JA, Peck CC, Schooley RT, Spilker BA, et al. 2001. Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework. Clinical Pharmacology and Therapeutics 69(3):89–95.

Barton GM, Kagan JC. 2009. A cell biological view of Toll-like receptor function: regulation through compartmentalization. Nature reviews. Immunology 9(8):535–42.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate - a practical and powerful approach to multiple testing. J.R.Stat.Soc.Ser.B Stat.Methodol. 57:289–300.

Bhushan V. 2002. First Aid for the USMLE Step 1, 2002: A Student-To-Student Guide McGraw-Hill.

Biswas R, Human Bagchi A. 2016. NFkB pathway and inhibition: an overview. Computational Molecular Biology Computational Molecular Biology Computational Molecular Biology 66(61):1–20.

Bloos F, Reinhart K. 2014. Rapid diagnosis of sepsis. Virulence 5(1):154–160.

Brown GD, Denning DW, Gow N a R, Levitz SM, Netea MG, White TC. 2012. Hidden Killers: Human Fungal Infections. Science Translational Medicine 4(165):165rv13.

Brown MP, Grundy WN, Lin D, Cristianini N, Sugnet CW, Furey TS, Ares M, Haussler D. 2000. Knowledge-based analysis of microarray gene expression data by using support vector machines. Proceedings of the National Academy of Sciences of the United States of America 97(1):262–267.

Bursle E, Robson J. 2016. Non-culture methods for detecting infection. Australian Prescriber 39(5):171–175.

Cai S, Batra S, Lira SA, Kolls JK, Jeyaseelan S. 2010. CXCL1 regulates pulmonary host defense to Klebsiella Infection via CXCL2, CXCL5, NF-kappaB, and MAPKs. Journal of immunology (Baltimore, Md. : 1950) 185(10):6214–25.

Carneiro LAM, Magalhaes JG, Tattoli I, Philpott DJ, Travassos LH. 2008. Nod-like proteins in inflammation and disease. Journal of Pathology 214(2):136–148.

Chan T, Gu F. 2011. Early diagnosis of sepsis using serum biomarkers. Expert review of molecular diagnostics 11(5):487–96.

Chawla N V, Japkowicz N, Drive P. 2004. Editorial : Special Issue on Learning from Imbalanced Data Sets. ACM SIGKDD Explorations Newsletter 6(1):1–6.

Chen G, Shaw MH, Kim Y-G, Nuñez G. 2009. NOD-like receptors: role in innate immunity and inflammatory disease. Annual review of pathology 4:365–398.

Chow ML, Moler EJ, Mian IS. 2001. Identifying marker genes in transcription profiling data using a mixture of feature relevance experts. Physiol Genomics 5(2):99–111.

Colbert JD, Matthews SP, Miller G, Watts C. 2009. Diverse regulatory roles for lysosomal proteases in the immune response. European Journal of Immunology 39(11):2955–2965.

Croze E, Ivanov IE, Kreibich G, Adesnik M, Sabatini DD, Rosenfeld MG. 1989. Endolyn-78, a membrane glycoprotein present in morphologically diverse components of the endosomal and lysosomal compartments: Implications for lysosome biogenesis. Journal of Cell Biology 108(5):1597–1613.

Cunnington AJ. 2015. The Importance of Pathogen Load. PLoS Pathogens 11(1).

Czakai K, Leonhardt I, Dix A, Bonin M, Linde J, Einsele H, Kurzai O, Loeffler J. 2016. Krüppel-like Factor 4 modulates interleukin-6 release in human dendritic cells after in vitro stimulation with Aspergillus fumigatus and Candida albicans. Scientific reports 6(May):27990.

Darmoise A, Teneberg S, Bouzonville L, Brady RO, Beck M, Kaufmann SHE, Winau F. 2010. Lysosomal ??-Galactosidase Controls the Generation of Self Lipid Antigens for Natural Killer T Cells. Immunity 33(2):216–228.

Davis MJ, Eastman AJ, Qiu Y, Gregorka B, Kozel TR, Osterholzer JJ, Curtis JL, Swanson JA, Olszewski MA. 2015. Cryptococcus neoformans-induced macrophage lysosome damage crucially contributes to fungal virulence. Journal of immunology (Baltimore, Md. : 1950) 194(5):2219–2231.

Delaloye J, Calandra T. 2014a. Invasive candidiasis as a cause of sepsis in the critically ill patient. Virulence 5(1):161–169.

Delneste Y, Beauvillain C, Jeannin P. 2007. [Innate immunity: structure and function of TLRs]. Médecine sciences : M/S 23(1):67–73.

Dennis Jr G, Sherman BT, Hosack DA, Yang J, Gao W, Lane CH, Lempicki RA, Dennis G, Sherman BT, Hosack DA, et al. 2003. DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biology 4(9):R60.

Didonna A, Cekanaviciute E, Oksenberg JR, Baranzini SE. 2016. Immune cell-specific transcriptional profiling highlights distinct molecular pathways controlled by Tob1 upon experimental autoimmune encephalomyelitis. Scientific reports 6:31603.

Dix A, Hünniger K, Weber M, Guthke R, Kurzai O, Linde J. 2015a. Biomarker-based classification of bacterial and fungal whole-blood infections in a genome-wide expression study. Frontiers in microbiology 6:171–171.

Dix A, Hünniger K, Weber M, Guthke R, Kurzai O, Linde J. 2015b. Biomarker-based classification of bacterial and fungal whole-blood infections in a genome-wide expression study. Frontiers in microbiology 6:171.

Du P, Kibbe WA, Lin SM. 2008. lumi: a pipeline for processing Illumina microarray. Bioinformatics (Oxford, England) 24(13):1547–8.

Fernández-Delgado M, Cernadas E, Barro S, Amorim D, Amorim Fernández-Delgado D. 2014. Do we Need Hundreds of Classifiers to Solve Real World Classification Problems? Journal of Machine Learning Research 15:3133–3181.

De Francesco PN, Mucci JM, Ceci R, Fossati CA, Rozenfeld PA. 2013. Fabry disease peripheral blood immune cells release inflammatory cytokines: Role of globotriaosylceramide. Molecular Genetics and Metabolism 109(1):93–99.

Franchi L, Warner N, Viani K, Nuñez G. 2009. Function of Nod-like receptors in microbial recognition and host defense. Immunological reviews 227(1):106–28.

Fritz JH, Ferrero RL, Philpott DJ, Girardin SE. 2006. Nod-like proteins in immunity, inflammation and disease. Nature Immunology 7(12):1250–1257.

Fung GM, Mangasarian OL, Shavlik JW. 2003. Knowledge-Based Support Vector Machine Classifiers. Advances in Neural Information Processing Systems 15, NIPS 2002(19):1–9.

Gamerdinger M, Hajieva P, Kaya AM, Wolfrum U, Hartl FU, Behl C. 2009. Protein quality control during aging involves recruitment of the macroautophagy pathway by BAG3. EMBO Journal 28(7):889–901.

Gardinassi LG, Garcia GR, Costa CHN, Costa Silva V, de Miranda Santos IKF, Dinis-Oliveira R. 2016. Blood Transcriptional Profiling Reveals Immunological Signatures of Distinct States of Infection of Humans with Leishmania infantum (A Acosta-Serrano, Ed). PLOS Neglected Tropical Diseases 10(11):e0005123.

Garey KW, Rege M, Pai MP, Mingo DE, Suda KJ, Turpin RS, Bearden DT. 2006. Time to Initiation of Fluconazole Therapy Impacts Mortality in Patients with Candidemia: A Multi-Institutional Study. Clinical Infectious Diseases 43(1):25–31.

Garey MR, Johnson DS. 1979. Computers and Intractability: A Guide to the Theory of NP-Completeness (Series of Books in the Mathematical Sciences). Computers and Intractability:340.

Gay NJ, Gangloff M, Weber AN. 2006. Toll-like receptors as molecular switches. Nat Rev Immunol 6(9):693–698.

Gilmore T. 2006. Introduction to NF-jB: players, pathways, perspectives. Oncogene 25:6680–6684.

Gonzalez A, Valeiras M, Sidransky E, Tayebi N. 2014. Lysosomal integral membrane protein-2: A new player in lysosome-related pathology. Molecular Genetics and Metabolism 111(2):84–91.

Gordon GJ, Hong SA, Dud M. 2005. First-Order Mixed Integer Linear Programming. Syntax 26(47):213–222.

Gurobi Optimization I. 2016. Gurobi Optimizer Reference Manual.

Guyon I, Weston J, Barnhill S, Vapnik V. 2002. Gene selection for cancer classification using support vector machines. Machine Learning 46(1–3):389–422.

Hayden MS, West AP, Ghosh S. 2006. NF-κB and the immune response. Oncogene 25(51):6758.

He Y, Xu Y, Zhang C, Gao X, Dykema KJ, Martin KR, Ke J, Hudson EA, Khoo SK, Resau JH, et al. 2011. Identification of a Lysosomal Pathway That Modulates Glucocorticoid Signaling and the Inflammatory Response. Science signaling 4(180):ra44.

Hessle CC, Andersson B, Wold AE. 2005. Gram-positive and Gram-negative bacteria elicit different patterns of pro-inflammatory cytokines in human monocytes. Cytokine 30(6):311–318.

Hogan LH, Klein BS, Levitz SM. 1996. Virulence factors of medically important fungi. Clinical Microbiology Reviews 9(4):469–488.

Horn F, Heinekamp T, Kniemeyer O, Pollm??cher J, Valiante V, Brakhage AA. 2012. Systems biology of fungal infection. Frontiers in Microbiology 3(APR).

Huang DW, Lempicki R a, Sherman BT. 2009a. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature Protocols 4(1):44–57.

Huang DW, Lempicki R a, Sherman BT, Lempicki R a. 2009b. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature protocols 4(1):44–57.

Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, Bravo HC, Davis S, Gatto L, Girke T, et al. 2015. Orchestrating high-throughput genomic analysis with Bioconductor. Nature methods 12(2):115–21.

Iwabuchi K, Nakayama H, Oizumi A, Suga Y, Ogawa H, Takamori K. 2015. Role of ceramide from glycosphingolipids and its metabolites in immunological and inflammatory responses in humans. Mediators of Inflammation 2015.

Jawad I, Lukšić I, Rafnsson SB. 2012. Assessing available information on the burden of sepsis: global estimates of incidence, prevalence and mortality. Journal of global health 2(1):010404–010404.

Jenner RG, Young RA. 2005. Insights into host responses against pathogens from transcriptional profiling. Nature reviews. Microbiology 3(4):281–294.

Jimenez-Lucho V, Ginsburg V, Krivan HC. 1990. Cryptococcus neoformans, Candida albicans, and other fungi bind specifically to the glycosphingolipid lactosylceramide (GAl??1-4Glc??1-1Cer), a possible adhesion receptor for yeasts. Infection and Immunity 58(7):2085–2090.

Johnson WE, Li C, Rabinovic A. 2007. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics (Oxford, England) 8(1):118–127.

Kanayama M, Inoue M, Danzaki K, Hammer G, He Y, Shinohara ML. 2015. Autophagy enhances NFkB activity in specific tissue macrophages by sequestering A20 to boost antifungal immunity. Nature Communications 6:1–14.

Kanneganti TD, Lamkanfi M, Núñez G. 2007. Intracellular NOD-like Receptors in Host Defense and Disease. Immunity 27(4):549–559.

Kaposzta R, Marodi L, Hollinshead M, Gordon S, Silva RP da. 1999. Rapid recruitment of late endosomes and lysosomes in mouse macrophages ingesting Candida albicans. J Cell Sci 112(19):3237–3248.

Khot PD, Fredricks DN. 2009. PCR-based diagnosis of human fungal infections. Expert review of anti-infective therapy 7(10):1201–1221.

Kim YK, Shin JS, Nahm MH. 2016. NOD-Like Receptors in Infection, Immunity, and Diseases. Yonsei medical journal 57(1):5–14.

Kint JA. 1970. Fabry's Disease: Alpha-Galactosidase Deficiency. Science 167(3922):1268.

Kirn TJ, Weinstein MP. 2013. Update on blood cultures: How to obtain, process, report, and interpret. Clinical Microbiology and Infection 19(6):513–520.

Klassert TE, Hanisch A, Bräuer J, Klaile E, Heyl KA, Mansour MK, Mansour MM, Tam JM, Vyas JM, Slevogt H. 2014. Modulatory role of vitamin A on the Candida albicans-induced immune response in human monocytes. Medical microbiology and immunology 203(6):415–24.

Klassert TE, Bräuer J, Hölzer M, Stock M, Riege K, Zubiría-Barrera C, Müller MM, Rummler S, Skerka C, Marz M, et al. 2017. Differential Effects of Vitamins A and D on the Transcriptional Landscape of Human Monocytes during Infection. Scientific Reports 7(January):40599.

Ko DC, Gordon MD, Jin JY, Scott MP. 2001. Dynamic movements of organelles containing Niemann-Pick C1 protein: NPC1 involvement in late endocytic events. Molecular biology of the cell 12(3):601–614.

Kollef M, Micek S, Hampton N, Doherty JA, Kumar A. 2012. Septic shock attributed to Candida infection: Importance of empiric therapy and source control. Clinical Infectious Diseases 54(12):1739–1746.

Koo IC, Ohol YM, Wu P, Morisaki JH, Cox JS, Brown EJ. 2008. Role for lysosomal enzyme β-hexosaminidase in the control of mycobacteria infection. Proceedings of the National Academy of Sciences 105(2):710–715.

Kotsiantis SB, Zaharakis ID, Pintelas PE. 2006. Machine learning: A review of classification and combining techniques. Artificial Intelligence Review 26(3):159–190.

Langrish CL, McKenzie BS, Wilson NJ, De Waal Malefyt R, Kastelein RA, Cua DJ. 2004. IL-12 and IL-23: Master regulators of innate and adaptive immunity. Immunological Reviews 202:96–105.

Lauvau G, Loke P, Hohl TM. 2015. Monocyte-mediated defense against bacteria, fungi, and parasites

Leal SM, Cowden S, Hsia Y-C, Ghannoum MA, Momany M, Pearlman E. 2010. Distinct roles for Dectin-1 and TLR4 in the pathogenesis of Aspergillus fumigatus keratitis. PLoS pathogens 6:e1000976.

Lever A, Mackenzie I. 2007. Sepsis: definition, epidemiology, and diagnosis. Bmj 335(7625):879–883.

Levy MM, Fink MP, Marshall JC, Abraham E, Angus D, Cook D, Cohen J, Opal SM, Vincent J-L, Ramsay G, et al. 2003. 2001 SCCM/ESICM/ACCP/ATS/SIS International Sepsis Definitions Conference. Intensive Care Medicine 29(4):530–538.

Linde JJ, Löffler J, Dix A, Czakai K, Springer J, Fliesser M, Bonin M, Guthke R, Schmitt AL, Einsele H, et al. 2016. Genome-Wide Expression Profiling Reveals S100B as Biomarker for Invasive Aspergillosis. Frontiers in Microbiology 7(March):1–10.

Lionakis MS, Fischer BG, Lim JK, Swamydas M, Wan W, Richard Lee CC, Cohen JI, Scheinberg P, Gao JL, Murphy PM. 2012. Chemokine Receptor Ccr1 Drives Neutrophil-Mediated Kidney Immunopathology and Mortality in Invasive Candidiasis. PLoS Pathogens 8(8).

Lorenz TC. 2012. Polymerase Chain Reaction: Basic Protocol Plus Troubleshooting and Optimization Strategies. Journal of Visualized Experiments : JoVE(63).

Maglogiannis IG. 2007. Emerging artificial intelligence applications in computer engineering : real word AI systems with applications in eHealth, HCI, information retrieval and pervasive technologies

Maurin M. 2012. Real-time PCR as a diagnostic tool for bacterial diseases. Expert Review of Molecular Diagnostics 12(7):731–754.

Meylan E, Tschopp J. 2008. IRAK2 takes its place in TLR signaling. Nature immunology 9(6):581–2.

Mogensen TH. 2009. Pathogen recognition and inflammatory signaling in innate immune defenses. Clinical Microbiology Reviews 22(2):240–273.

Moore SM, Holt V V., Malpass LR, Hines IN, Wheeler MD. 2015. Fatty acid-binding protein 5 limits the anti-inflammatory response in murine macrophages. Molecular Immunology 67(2):265–275.

Müller MM, Lehmann R, Klassert TE, Reifenstein S, Conrad T, Moore C, Kuhn A, Behnert A, Guthke R, Driesch D, et al. 2017. Global analysis of glycoproteins identifies markers of endotoxin tolerant monocytes and GPR84 as a modulator of TNFα expression. Scientific Reports 7(1):838.

Muzio M. 1997. IRAK (Pelle) Family Member IRAK-2 and MyD88 as Proximal Mediators of IL-1 Signaling. Science 278(5343):1612–1615.

Navarathna DHMLP, Roberts DD. 2010. Candida albicans heme oxygenase and its product CO contribute to pathogenesis of candidemia and alter systemic chemokine and cytokine expression. Free Radical Biology and Medicine 49(10):1561–1573.

Netea MG, Van der Graaf C, Van der Meer JWM, Kullberg BJ. 2004a. Recognition of fungal pathogens by Toll-like receptors. European journal of clinical microbiology & infectious diseases : official publication of the European Society of Clinical Microbiology 23(9):672–6.

Netea MG, Kullberg B-J, Van der Meer JWM. 2004b. Proinflammatory cytokines in the treatment of bacterial and fungal infections. BioDrugs : clinical immunotherapeutics, biopharmaceuticals and gene therapy 18(1):9–22.

Netea MG, Ferwerda G, Jong DJ de, Jansen T, Jacobs L, Kramer M, Naber THJ, Drenth JPH, Girardin SE, Kullberg BJ, et al. 2005. Nucleotide-Binding Oligomerization Domain-2 Modulates Specific TLR Pathways for the Induction of Cytokine Release. The Journal of Immunology 174(10):6518–6523.

Netea MG, Brown GD, Kullberg BJ, Gow N a R. 2008. An integrated model of the recognition of Candida albicans by the innate immune system. Nature reviews. Microbiology 6(1):67–78.

Netea MG, Simon A, van de Veerdonk F, Kullberg B-J, Van der Meer JWM, Joosten LAB. 2010. IL-1beta processing in host defense: beyond the inflammasomes. PLoS pathogens 6(2):e1000661.

Netea MG, Joosten LAB, van der Meer JWM, Kullberg B-J, van de Veerdonk FL. 2015. Immune defence against Candida fungal infections. Nature reviews. Immunology 15(10):630–642.

Ngo LY, Kasahara S, Kumasaka DK, Knoblaugh SE, Jhingran A, Hohl TM. 2014. Inflammatory monocytes mediate early and organ-specific innate defense during systemic candidiasis. Journal of Infectious Diseases 209(1):109–119.

Nicholson LB. 2016. The immune system. Essays in Biochemistry 60(3):275–301.

Nishida E, Gotoh Y. 1993. The MAP kinase cascade is essential for diverse signal transduction pathways. Trends in biochemical sciences 18(4):128–131.

Noble WS. 2006. What is a support vector machine? Nature biotechnology 24(12):1565–1567.

Palmer C, Diehn M, Alizadeh AA, Brown PO. 2006. Cell-type specific gene expression profiles of leukocytes in human peripheral blood. BMC Genomics 7(1):115.

Pappas PG, Alexander BD, Andes DR, Hadley S, Kauffman CA, Freifeld A, Anaissie EJ, Brumble LM, Herwaldt L, Ito J, et al. 2010. Invasive fungal infections among organ transplant recipients: results of the Transplant-Associated Infection Surveillance Network (TRANSNET). Clinical infectious diseases : an official publication of the Infectious Diseases Society of America 50(8):1101–1111.

Parkinson-Lawrence EJ, Shandala T, Prodoehl M, Plew R, Borlace GN, Brooks DA. 2010. Lysosomal Storage Disease: Revealing Lysosomal Function and Physiology. Physiology 25(2):102–115.

Patel GP, Simon D, Scheetz M, Crank CW, Lodise T, Patel N. 2009. The effect of time to antifungal therapy on mortality in Candidemia associated septic shock. American journal of therapeutics 16(6):508–511.

Patterson TF. 2011. Clinical utility and development of biomarkers in invasive aspergillosis. Transactions of the American Clinical and Climatological Association 122(210):174–83.

Pavlidis P, Wapinski I, Noble WS. 2004. Support vector machine classification on the web. Bioinformatics (Oxford, England) 20(4):586–7.

Pfaffl MW, Tichopad A, Prgomet C, Neuvians TP. 2004. Determination of stable housekeeping genes, differentially regulated target genes and sample integrity: BestKeeper--Excel-based tool using pair-wise correlations. Biotechnology letters 26(6):509–15.

Piantadosi C a, Withers CM, Bartz RR, MacGarvey NC, Fu P, Sweeney TE, Welty-Wolf KE, Suliman HB. 2011. Heme oxygenase-1 couples activation of mitochondrial biogenesis to anti-inflammatory cytokine expression. The Journal of biological chemistry 286(18):16374–85.

Piro RM, Wiesberg S, Schramm G, Rebel N, Oswald M, Eils R, Reinelt G, König R. 2014. Network topology-based detection of differential gene regulation and regulatory switches in cell metabolism and signaling. BMC Systems Biology 8(1):56.

Poos AM, Maicher A, Dieckmann AK, Oswald M, Eils R, Kupiec M, Luke B, König R. 2016. Mixed Integer Linear Programming based machine learning approach identifies regulators of telomerase in yeast. Nucleic acids research:gkw111-.

Quackenbush J. 2006. Microarray Analysis and Tumor Classification. New England Journal of Medicine 354(23):2463–2472.

Rhodes A, Evans LE, Alhazzani W, Levy MM, Antonelli M, Ferrer R, Kumar A, Sevransky JE, Sprung CL, Nunnally ME, et al. 2017. Surviving Sepsis Campaign: International Guidelines for Management of Sepsis and Septic Shock: 2016 Springer Berlin Heidelberg.

Rietdijk ST, Burwell T, Bertin J, Coyle AJ. 2008. Sensing intracellular pathogens-NOD-like receptors. Current Opinion in Pharmacology 8(3):261–266.

Rieu I, Powers SJ. 2009. Real-Time Quantitative RT-PCR - Design, Calculations, and Statistics. The Plant Cell 21(4):1031–103.

Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. 2015. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research 43(7):e47.

Rivera A, Siracusa MC, Yap GS, Gause WC. 2016. Innate cell communication kick-starts pathogen-specific immunity. 17:356.

Romani L. 2011. Immunity to fungal infections. Nat Rev Immunol 11(4):275–288.

Ruiz-Herrera J, Victoria Elorza M, Valentín E, Sentandreu R. 2006. Molecular organization of the cell wall of Candida albicans and its relation to pathogenicity. FEMS Yeast Research 6(1):14–29.

Sabatini DD, Adesnik M. 2013. Christian de Duve: Explorer of the cell who discovered new organelles by using a centrifuge. Proceedings of the National Academy of Sciences 110(33):13234–13235.

Saeys Y, Inza I, Larrañaga P. 2007a. A review of feature selection techniques in bioinformatics. Bioinformatics 23(19):2507–2517.

Saeys Y, Inza I, Larrañaga P. 2007b. A review of feature selection techniques in bioinformatics. Bioinformatics 23(19):2507–2517.

Saftig P. 2006. Physiology of the lysosome. In Fabry Disease: Perspectives from 5 Years of FOS, Mehta A, , Beck M, , Sunder-Plassmann G (eds). Oxford PharmaGenesis;

Saraiva JP, Biering A, Assmann C, Blaess M, Claus R, Löffler J, Slevogt H, Blaess M, Biering A, Assmann C, et al. 2016b. ScienceDirect across Integrating classifiers classifiers across datasets datasets improves consistency of of biomarker biomarker Integrating classifiers across improves of predictions sepsis Integrating datasets improves of predictions sepsis consistency. IFAC-PapersOnLine 49(26):95–102.

Saraiva JP, Oswald M, Biering A, Röll D, Assmann C, Klassert T, Blaess M, Czakai K, Claus R, Löffler J, et al. 2017. Fungal biomarker discovery by integration of classifiers. BMC Genomics 18.

Schacht T, Oswald M, Eils R, Eichmüller SB, König R. 2014a. Estimating the activity of transcription factors by the effect on their target genes. Bioinformatics (Oxford, England) 30(17):i401-7.

Schacht T, Oswald M, Eils R, Eichmüller SB, König R. 2014b. Estimating the activity of transcription factors by the effect on their target genes. Bioinformatics (Oxford, England) 30(17):i401-7.

Scheller J, Chalaris A, Schmidt-Arras D, Rose-John S. 2011. The pro- and anti-inflammatory properties of the cytokine interleukin-6. Biochimica et biophysica acta 1813(5):878–88.

Schwake M, Schröder B, Saftig P. 2013. Lysosomal Membrane Proteins and Their Central Role in Physiology. Traffic 14(7):739–748.

Shaw MH, Reimer T, Kim Y, Nuñez G. 2009. NIH Public Access. 20(4):377–382.

Shi C, Pamer EG. 2011. Monocyte recruitment during infection and inflammation. Nature reviews. Immunology 11(11):762–74.

Shoham S, Levitz SM. 2005. The immune response to fungal infections. British journal of haematology 129(5):569–82.

Silva-Gomes S, Appelberg R, Larsen R, Soares MP, Gomes MS. 2013. Heme catabolism by heme oxygenase-1 confers host resistance to Mycobacterium infection. Infection and Immunity 81(7):2536–2545.

Singer M, Deutschman CS, Seymour CW, Shankar-Hari M, Annane D, Bauer M, Bellomo R, Bernard GR, Chiche J, Coopersmith CM, et al. 2016. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). JAMA 315(8):801.

Smeekens SP, Ng A, Kumar V, Johnson MD, Plantinga TS, van Diemen C, Arts P, Verwiel ET, Gresnigt MS, Fransen K, et al. 2013. Functional genomics identifies type I interferon pathway as central for host defense against Candida albicans. Nat Commun 4:1342.

Soejima T, Iida K, Qin T, Taniai H, Seki M, Yoshida S. 2008. Method To Detect Only Live Bacteria during PCR Amplification. Journal of Clinical Microbiology 46(7):2305–2313.

Strimbu K, Tavel J a. 2011. What are Biomarkers? Curr Opin HIV AIDS 5(6):463–466.

Sturge CR, Yarovinsky F. 2014. Complex Immune Cell Interplay in the Gamma Interferon Response during Toxoplasma gondii Infection. Infection and Immunity 82(8):3090–3097.

Szymczak WA, Deepe GS. 2009. The CCL7-CCL2-CCR2 axis regulates IL-4 production in lungs and fungal immunity. Journal of immunology (Baltimore, Md. : 1950) 183(3):1964–74.

Takeda K, Akira S. 2005. Toll-like receptors in innate immunity. International immunology 17(1):1–14.

Teng MWL, Bowman EP, McElwee JJ, Smyth MJ, Casanova J-L, Cooper AM, Cua DJ. 2015. IL-12 and IL-23 cytokines: from discovery to targeted therapies for immune-mediated inflammatory diseases. Nature medicine 21(7):719–729.

Torio CM, Andrews RM. 2006. National Inpatient Hospital Costs: The Most Expensive Conditions by Payer, 2011: Statistical Brief #160. In Healthcare Cost and Utilization Project (HCUP) Statistical Briefs Agency for Healthcare Research and Quality (US);

Traynor TR, Huffnagle GB. 2001. Role of chemokines in fungal infections. Medical mycology : official publication of the International Society for Human and Animal Mycology 39(1):41–50.

Trinchieri G, Sher A. 2007. Cooperation of Toll-like receptor signals in innate immune defence. Nature reviews. Immunology 7(3):179–90.

Tyagi S, Gupta P, Saini AS, Kaushal C, Sharma S. 2011. The peroxisome proliferator-activated receptor: A family of nuclear receptors role in various diseases. Journal of advanced pharmaceutical technology & research 2:236–40.

Vandal OH, Nathan CF, Ehrt S. 2009. Acid Resistance in Mycobacterium tuberculosis. Journal of Bacteriology 191(15):4714–4721.

Vanier MT, Millat G. 2003. Niemann–Pick disease type C. Clinical Genetics 64(4):269–281.

Vapnik V. 1982. Estimation of Dependences Based on Empirical Data: Springer Series in Statistics (Springer Series in Statistics), Secaucus, NJ, USA: Springer-Verlag New York, Inc.

Vapnik VN. 1995. The Nature of Statistical Learning Theory, New York, NY, USA: Springer-Verlag New York, Inc.

Vellodi A. 2005. Lysosomal storage disorders. British Journal of Haematology 128(4):413–431.

Vincent J-L, Rello J, Marshall J, Silva E, Anzueto A, Martin CD, Moreno R, Lipman J, Gomersall C, Sakr Y, et al. 2009. International study of the prevalence and outcomes of infection in intensive care units. JAMA 302(21):2323–2329.

Wang F, Hu S, Liu W, Qiao Z, Gao Y, Bu Z. 2011. Deep-Sequencing Analysis of the Mouse Transcriptome Response to Infection with Brucella melitensis Strains of Differing Virulence. PLOS ONE 6(12):e28485.

Wang Z, Gerstein M, Snyder M. 2009. RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics 10(1):57–63.

Wei BL, Denton PW, O'Neill E, Luo T, Foster JL, Garcia JV. 2005. Inhibition of Lysosome and Proteasome Function Enhances Human Immunodeficiency Virus Type 1 Infection. Journal of Virology 79(9):5705–5712.

Wolberg WH, Mangasarian OL. 1990. Multisurface method of pattern separation for medical diagnosis applied to breast cytology. Proceedings of the National Academy of Sciences 87(23):9193–9196.

Wong KL, Tai JJ-Y, Wong W-C, Han H, Sem X, Yeap W-H, Kourilsky P, Wong S-C. 2011. Gene expression profiling reveals the defining features of the classical, intermediate, and nonclassical human monocyte subsets. Blood 118(5):e16-31.

Xie G-H, Fang X-M, Fang Q, Wu X-M, Jin Y-H, Wang J-L, Guo Q-L, Gu M-N, Xu Q-P, Wang D-X, et al. 2008. Impact of invasive fungal infection on outcomes of severe sepsis: a multicenter matched cohort study in critically ill surgical patients. Critical Care 12(1):R5.

Yamayoshi S, Fujii K, Koike S. 2014. Receptors for enterovirus 71. Emerging microbes & infections 3(7):e53.

Yang S, Rothman RE. 2004. PCR-based diagnostics for infectious diseases: uses, limitations, and future applications in acute-care settings. The Lancet Infectious Diseases 4(6):337–348.

Zaas AK, Aziz H, Lucas J, Perfect JR, Ginsburg GS. 2010. Blood gene expression signatures predict invasive candidiasis. Science translational medicine 2(21):21ra17.

Zhang D, Tian Y, Shi Y. 2011. A group of knowledge-incorporated multiple criteria linear programming classifiers. Journal of Computational and Applied Mathematics 235(13):3705–3717.

Zuker M. 2003. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Research 31(13):3406–3415.

8. Appendix

Table A1 – Differentially expressed and consistently up and downregulated genes in, at least, 4 datasets.

Up_regulated			Down_regulated				
ABCA1	*GEM*	*PLEK*	*ABHD14A*	*CRIPT*	*IMP3*	*OSBPL11*	*SLC35A1*
ABTB2	*GJB2*	*PLK3*	*ABHD8*	*CSF1R*	*IMPA2*	*OSBPL2*	*SLC9A3R1*
ANXA5	*GK*	*PNPT1*	*ACAA1*	*CST3*	*ING4*	*P2RY8*	*SMAD5*
ARFGAP3	*GPR137B*	*PPP1R15A*	*ADAP2*	*CTDSP1*	*INPPL1*	*PAFAH2*	*SMUG1*
ARID5A	*HES4*	*PTPN1*	*AGPAT2*	*CYB561D1*	*JTB*	*PARVG*	*SNRNP25*
ARL5B	*ICAM1*	*PTX3*	*AIP*	*CYB561D2*	*KATNB1*	*PCTP*	*SNX30*
ARL8B	*ID2*	*RAPGEF2*	*AKR7A2*	*CYP2S1*	*KCNK6*	*PDHB*	*SPSB2*
ATP13A3	*IDO1*	*RASGEF1B*	*AKTIP*	*DAGLB*	*KCTD15*	*PDK4*	*SS18L2*
BASP1	*IL10*	*RFTN1*	*ANKZF1*	*DBP*	*KDM3B*	*PECAM1*	*STK38*
BATF	*IL12A*	*RGS1*	*AP2S1*	*DCAKD*	*KIAA0141*	*PGLS*	*STMN3*
BAZ1A	*IL12B*	*RIPK2*	*APBB1IP*	*DDX28*	*KIAA0430*	*PHKG2*	*STRA13*
BCL2A1	*IL1A*	*RSAD2*	*APEH*	*DEF6*	*KIAA0513*	*PIK3CG*	*STX10*
CCL2	*IL1B*	*RYBP*	*ARF5*	*DHRS7*	*KLF2*	*PIN1*	*SYK*
CCL20	*IL1RN*	*SDC4*	*ARHGAP1*	*DIRC2*	*KLHDC3*	*PITPNM1*	*TADA3*
CCL22	*IL23A*	*SERPINB9*	*ARHGAP18*	*DIS3L*	*KLHL22*	*PLCB2*	*TAF10*
CCL3	*IL2RA*	*SIAH2*	*ARHGAP19*	*DNASE2*	*LASP1*	*PLOD1*	*TAF4*
CCNA1	*IL6*	*SLAMF1*	*ARHGEF18*	*DNTTIP1*	*LDLRAP1*	*PLSCR3*	*TBC1D10C*
CCR7	*INSIG1*	*SLAMF7*	*ARHGEF6*	*DOCK11*	*LRMP*	*PLXDC2*	*TBCC*
CD274	*IRAK2*	*SLC1A3*	*ARRDC2*	*DOCK2*	*LSM10*	*POLD4*	*TBXAS1*
CD40	*IRF1*	*SLC43A3*	*ASGR1*	*DOK2*	*LSM4*	*POLE3*	*TCF3*
CD80	*ISG20*	*SLCO4A1*	*ATG16L1*	*DPEP2*	*LST1*	*POLR3GL*	*TECPR1*

CD83	ITGB8	SOCS2	ATG16L2	DUSP23	LTA4H	POLR3K	THAP11
CDKN1A	KLF6	SOCS3	ATG9A	ECHS1	LTB4R	PPCS	TMEM14C
CSF2	LAMP3	SPINK1	ATP5O	EIF2B1	LY86	PPIL3	TMEM154
CSTB	LRRC32	STARD8	ATP6V0E2	EPN1	MAST3	PPP1CA	TMEM160
CXCL1	MAFF	STAT4	B9D2	ESYT1	MCEE	PQLC3	TMEM170B
CXCL2	MAMLD1	STX11	BRMS1	EVI2B	MEF2C	PRAM1	TMEM45B
DCUN1D3	MAP3K8	TFRC	CALHM2	FAM102A	METTL7A	PRCP	TNFAIP8L2
DDX60L	MASTL	TMEM140	CALM2	FAM120A	MFNG	PYCARD	TRABD
DFNA5	MCOLN2	TNFAIP3	CAMLG	FAM173A	MFSD1	PYGB	TRAPPC5
DNAJA1	MGLL	TNFRSF4	CARD9	FAM32A	MGST2	RAB34	TRIM8
DRAM1	MMP19	TNFSF9	CASP9	FAM98C	MID1IP1	RABEPK	TRMT12
DUSP5	MSC	TNIP3	CAT	FBXL16	MNT	RASAL3	TSEN54
EBI3	MT2A	TP53BP2	CBL	FGL2	MRFAP1L1	RBM4B	TSPO
EDN1	NFKB1	TRAF1	CCDC69	FOXJ2	MRI1	REEP5	TST
ELOVL7	NFKBIA	TRIM25	CD300LF	FRAT1	MRPL34	REPS2	TXNIP
ETS2	NFKBIZ	TRIM56	CD302	FRAT2	MRPS15	RGS14	TXNRD2
F3	NRIP3	TXN	CD33	FUCA1	MS4A6A	RGS19	TYK2
FJX1	OASL	UBTD2	CDC25B	GAL3ST4	MTIF3	RNASE6	UBAC1
GADD45B	P2RX4	ZBTB43	CDC40	GMFG	MXD4	RNF130	UFC1
GBP1	PFKFB3	ZC3H12A	CDK2AP1	GNAI2	MYO1F	RNF135	VAMP8
GBP2	PIM1	ZC3H12C	CDK2AP2	GPBAR1	NADSYN1	RNF166	WDR81
GCH1	PIM2	ZNFX1	CHST13	GPD1L	NAGA	RNF44	XRCC1
GCLM	PIM3		CLN5	GSTP1	NAGPA	RNPEP	YIPF3
			CLPP	HEBP2	NCKAP1L	RPS6KA4	YPEL2
			CLPTM1L	HHEX	NDUFA2	RRAS	YPEL3
			CLPX	HMHA1	NDUFB10	S100A4	ZBTB48
			CLTB	HSD17B11	NDUFS3	S1PR4	ZFAND2B

	CNOT7	*HSPBAP1*	*NDUFS7*	*SASH3*	*ZNF266*
	CNPY3	*HVCN1*	*NOSIP*	*SCRN1*	*ZNF362*
	COMMD3	*ID3*	*NSL1*	*SERP1*	*ZNF467*
	COMMD8	*IFNGR1*	*NT5C*	*SIPA1*	*ZNF792*
	COQ10A	*IL13RA1*	*NUP214*	*SIVA1*	*ZNHIT1*
	CORO1B	*IL16*	*OCEL1*	*SLC2A4RG*	

Table A2 - List of biomarker genes from each type of classifier

Common Genes	**Unique to single classifiers**	**Unique to combined classifiers**
RRAGD	*ANTXR2*	*ATP6V1D*
RGCC	*PPFIBP2*	*BLVRA*
IFNB1	*AGAP3*	*TNFRSF14*
KLHL21	*RIN2*	*RNF144B*
TBC1D7	*GNPDA1*	*ADA*
ADGRE1	*PPIF*	*CXXC5*
TBC1D2	*BATF3*	*CH25H*
HMOX1	*SRXN1*	*SDSL*
HCAR2	*PHACTR1*	*BCAR3*
PELI1	*TAGAP*	*TNFSF15*
DHRS9	*CCL8*	*APOBEC3A*
NCF1	*RGS1*	*TNFSF10*
TNFSF14	*TRIM21*	*DPYSL3*
CCR1	*CD86*	*HK2*
TBC1D9	*SATB1*	*NCF1C*

SPRY2	*GLIPR2*	*RTP4*
SOWAHC	*UBASH3B*	*GLA*
CEBPB	*IVNS1ABP*	*ARHGEF3*
PLCXD1	*ANAPC4*	*NSMAF*
NCOA7	*TTC14*	*KLF4*
CXCL11	*ACADVL*	*TMEM243*
NRIP3	*MGAT2*	*TRMT5*
LGALS9	*CRYGS*	*CMTM7*
TNFSF13B	*WDFY2*	*GPAT3*
IL12B	*CRIPAK*	*TGFBI*
CEP135	*HACD3*	*TMEM106A*
IL27RA		*ANKIB1*
ENC1		*UBA7*
ATP6V0A1		*FAM111A*
PAPSS1		*SLC16A3*
ST3GAL5		*SPP1*
EVL		*TLR7*
SCARB2		*EGR2*
SP140		*CHST12*
NOP16		*CALU*
PCID2		*SLC7A7*
RBCK1		*EMP1*
SLFN12		*PRKAG2*
TRAFD1		*C14orf159*
STK26		*FXYD6*

HAVCR2	*RHOU*
FAM46A	*DDX60L*
EDN1	
CLCF1	
CD40	
PARP4	

Table A3 - List of genes selected from the combined approach and their respective adjusted p-values (≤0.05 was regarded to be significant)

Gene symbol	Dix	Smeekens	Saraiva	Klassert	Czakai
ADA	3.64E-06	1.12E-03	4.92E-01	1.89E-05	1.04E-03
BCAR3	2.02E-06	2.51E-06	2.42E-02	5.47E-05	3.44E-03
BLVRA	3.89E-04	9.37E-03	6.99E-02	4.95E-02	6.14E-06
CCR1	4.01E-03	4.95E-09	2.99E-01	1.37E-08	1.81E-02
DHRS9	1.23E-05	1.79E-07	3.15E-01	3.74E-07	2.94E-03
EDN1	2.57E-02	2.53E-20	1.21E-03	5.08E-01	1.73E-02
EGR2	4.24E-05	2.30E-14	5.26E-01	1.49E-04	2.10E-04
GLA	2.40E-08	1.77E-11	1.80E-02	4.58E-07	1.46E-02
HMOX1	7.92E-07	5.07E-04	3.50E-02	2.05E-06	2.15E-02
PAPSS1	1.06E-02	1.82E-04	5.51E-01	9.31E-09	7.42E-04
RNF144B	5.56E-05	3.66E-03	5.05E-01	3.19E-08	3.14E-03
RRAGD	2.22E-08	1.74E-02	1.57E-01	2.86E-04	2.59E-04

SCARB2	5.71E-01	4.62E-20	4.72E-02	2.82E-04	9.94E-04
SDSL	7.43E-05	2.00E-02	2.22E-01	5.97E-06	1.42E-02
SOWAHC	3.39E-06	2.04E-01	2.66E-02	3.70E-03	3.00E-03
SPRY2	5.64E-08	1.66E-02	9.32E-03	3.21E-05	2.05E-03
TBC1D7	1.88E-11	1.87E-03	1.31E-02	3.55E-03	7.82E-03
TNFSF14	3.79E-11	4.07E-03	4.30E-02	2.18E-05	7.56E-03
TNFSF15	1.16E-08	6.08E-02	9.32E-03	1.70E-05	4.15E-04

Table A4 - Single and combined classifier gene lists

Single classifier			**Combined classifier**		
C5AR1	*GLA*	*SMAD3*	*CCR1*	*UNC93B1*	RPAP2
CCR1	*GLIPR2*	*SMCHD1*	*SCARB2*	*WIPF1*	S100A9
CXCL10	*GNAQ*	*SMCO4*	*STK26*	*ACVR1*	SDSL
EDN1	*GNG2*	*SOWAHC*	*EVL*	*ANAPC4*	SP100
EVL	*GPAT3*	*SP100*	*GLA*	*APOBEC3A*	ST6GALNAC6
FXYD6	*GPR18*	*SP140*	*LGALS9*	*ATP6V0A1*	STAP1
HMOX1	*GRAMD1A*	*SPRY2*	*SERPINA1*	*BATF2*	STAT2
SLC16A3	*HACD3*	*ST3GAL5*	*SPP1*	*BLVRA*	STAT5A
SLFN12	*HAVCR2*	*STAP1*	*VAV1*	*C5AR1*	TGFBI
STK26	*HCAR2*	*STAT5A*	*ANXA1*	*CCL8*	TMEM106A
ACADVL	*HDAC1*	*SUCNR1*	*BLMH*	*CD14*	TMEM243
ACVR1	*HK2*	*SYNJ2BP*	*C1GALT1*	*CD68*	TNFAIP2
ADCY3	*HSPBAP1*	*TAGAP*	*CCL23*	*CDCA4*	TNFSF13B
ADORA2B	*IFNB1*	*TBC1D9*	*CEP135*	*CEBPB*	TPMT

ANAPC4	*IGF2R*	*TGFBI*	*CITED2*	*CENPW*	TRAFD1
ANXA1	*IL12A*	*THOC1*	*EDN1*	*CEP295*	TRANK1
ARHGEF3	*IL27RA*	*TMEM243*	*FAM46A*	*CHMP5*	TRIB2
BCAR3	*IQSEC1*	*TNFAIP2*	*FXYD6*	*CISD1*	TRIM21
BTK	*IRF2*	*TNFRSF1B*	*HAVCR2*	*CISH*	TSC22D1
C12orf10	*IVNS1ABP*	*TNFSF10*	*HMOX1*	*CLEC5A*	TTC14
C14orf159	*JAK2*	*TNFSF13B*	*MOV10*	*CMTM7*	TTYH3
C1GALT1	*KLF4*	*TNFSF15*	*PCID2*	*CRIPAK*	UBA7
CCL20	*KLHL21*	*TP53INP2*	*PELI1*	*CRYGS*	UBASH3B
CCL23	*LGALS9*	*TPMT*	*RHOU*	*CUL4A*	USP11
CCL5	*MAP3K1*	*TRAFD1*	*SATB1*	*CXCL10*	VMO1
CCL8	*MASTL*	*TRANK1*	*SLFN12*	*CXXC5*	WDFY2
CCR4	*MGAT2*	*TRIB2*	*SMCO4*	*DDX60*	XRN1
CCR7	*MOV10*	*TRIM21*	*SOWAHC*	*DHRS9*	ZBTB32
CD247	*MYC*	*TRIM5*	*ST3GAL5*	*DHX58*	ZNF786
CD40	*NAGK*	*TRIP10*	*TAGAP*	*EGR2*	ZRSR2
CD68	*NCF1*	*TTC14*	*TBC1D9*	*EPB41L3*	
CD86	*NCF1C*	*UBA7*	*ACADVL*	*FNDC3A*	
CDCA4	*NCOA7*	*UBASH3B*	*ARHGEF3*	*GLIPR2*	
CDK6	*NDUFAF7*	*UNC93B1*	*BCAR3*	*GPAT3*	
CEBPB	*NDUFV1*	*USP18*	*C12orf10*	*GRHPR*	
CEP135	*NPEPL1*	*VAV1*	*C14orf159*	*HACD3*	
CEP295	*NSMAF*	*VAV3*	*CD40*	*HCAR2*	
CH25H	*NUB1*	*VMO1*	*CD86*	*HK2*	
CISD1	*PAPSS1*	*WDFY2*	*CDK6*	*HSPA6*	

CISH	PARP1	ZNF700
CLCF1	PARP10	ZRSR2
CLDN23	PARP4	
CMTM7	PCID2	
CRIPAK	PCNT	
CRYGS	PELI1	
CXCL1	PIK3CB	
CXCL11	PIK3CG	
CXCL2	PLA2G7	
CXCL3	PLCXD1	
CXCL5	PPBP	
CXCL6	PRKAG2	
CXCL8	RAB3IP	
CXCL9	RABGAP1L	
CXCR4	RBCK1	
CXCR6	RGS1	
DDX60L	RHBDD2	
DHRS9	RNF144B	
EGR2	RPAP2	
EIF2AK3	RPUSD2	
EIF2AK4	S100A9	
ENC1	S1PR4	
ETV3	SATB1	
FAM46A	SCARB2	
FPR2	SDSL	

CLCF1	HSPB1
CLDN23	HSPBAP1
CTSC	IGF2R
DDX60L	IL27RA
EMP1	IQSEC1
ENC1	IRF2
ETV3	MICAL1
GEM	MRPS24
GRAMD1A	MYD88
IFNB1	NCF1
IVNS1ABP	NCF1C
JUN	NDUFAF7
KLF4	NDUFV1
MGAT2	NISCH
NCOA7	NSMAF
PAPSS1	NUB1
PARP10	ORC2
RBCK1	PARP4
RGS1	PCNT
SEPT6	PGD
SLC16A3	PLCXD1
SLC7A7	PPM1M
SP140	PRKAG2
SPRY2	RABGAP1L
TNFRSF1B	RASGRP3

FPR3	*SEPT6*	*TNFRSF9*	*RHBDD2*
GBP3	*SERPINA1*	*TNFSF10*	*RHOH*
GEM	*SLC7A7*	*TRIP10*	*RNF144B*

Table A5 - RT-qPCR mean expression values across conditions and corresponding p-values for all genes of interest.

GENE	BONFERRONI'S MULTIPLE COMPARISONS TEST	MEAN DIFF,	95,00% CI OF DIFF	SIGNIFICANT?	SUMMARY	ADJUSTED P VALUE
GLA	Ctrl vs. LPS	-2,231	-2,751 to -1,711	Yes	****	<0,0001
	Ctrl vs. MALP	-1,886	-2,406 to -1,366	Yes	****	<0,0001
	Ctrl vs. Zym	-3,349	-3,869 to -2,829	Yes	****	<0,0001
	LPS vs. MALP	0,345	-0,175 to 0,865	No	ns	0,315
	LPS vs. Zym	-1,118	-1,637 to -0,5975	Yes	***	0,0003
	MALP vs. Zym	-1,463	-1,982 to -0,9425	Yes	****	<0,0001
	Ctrl vs. C.a.	-4,589	-5,368 to -3,81	Yes	****	<0,0001
	Ctrl vs. Asp.	-3,771	-4,55 to -2,992	Yes	****	<0,0001
	Ctrl vs. E.coli	-2,881	-3,66 to -2,102	Yes	****	<0,0001
	C.a. vs. Asp.	0,8175	0,03874 to 1,596	Yes	*	0,0384
	C.a. vs. E.coli	1,708	0,9287 to 2,486	Yes	***	0,0003
	Asp. vs. E.coli	0,89	0,1112 to 1,669	Yes	*	0,0236
SCARB2	Ctrl vs. LPS	1,283	0,7603 to 1,805	Yes	***	0,0001
	Ctrl vs. MALP	1,046	0,5241 to 1,568	Yes	***	0,0005
	Ctrl vs. Zym	1,533	1,01 to 2,055	Yes	****	<0,0001
	LPS vs. MALP	-0,2363	-0,7584 to 0,2859	No	ns	0,9739
	LPS vs. Zym	0,25	-0,2722 to 0,7722	No	ns	0,8502
	MALP vs. Zym	0,4863	-0,03592 to 1,008	No	ns	0,0724
	Ctrl vs. C.a.	-0,5375	-1,074 to -0,001495	Yes	*	0,0493
	Ctrl vs. Asp.	0,0875	-0,4485 to 0,6235	No	ns	>0,9999
	Ctrl vs. E.coli	1,735	1,199 to 2,271	Yes	****	<0,0001

	C.a. vs. Asp.	0,625	0,089 to 1,161	Yes	*	0,021
	C.a. vs. E.coli	2,273	1,736 to 2,809	Yes	****	<0,0001
	Asp. vs. E.coli	1,648	1,111 to 2,184	Yes	****	<0,0001
PPARG	Ctrl vs. LPS	0,4563	-0,3774 to 1,29	No	ns	0,5924
	Ctrl vs. MALP	-0,05	-0,8837 to 0,7837	No	ns	>0,9999
	Ctrl vs. Zym	0,1163	-0,7174 to 0,9499	No	ns	>0,9999
	LPS vs. MALP	-0,5063	-1,34 to 0,3274	No	ns	0,4286
	LPS vs. Zym	-0,34	-1,174 to 0,4937	No	ns	>0,9999
	MALP vs. Zym	0,1663	-0,6674 to 0,9999	No	ns	>0,9999
	Ctrl vs. C.a.	-1,931	-3,035 to -0,8278	Yes	**	0,0014
	Ctrl vs. Asp.	-2,455	-3,558 to -1,352	Yes	***	0,0002
	Ctrl vs. E.coli	0,8488	-0,2547 to 1,952	No	ns	0,1759
	C.a. vs. Asp.	-0,5238	-1,627 to 0,5797	No	ns	0,8686
	C.a. vs. E.coli	2,78	1,677 to 3,883	Yes	****	<0,0001
	Asp. vs. E.coli	3,304	2,2 to 4,407	Yes	****	<0,0001
CD164	Ctrl vs. LPS	0,2388	-0,1953 to 0,6728	No	ns	0,5837
	Ctrl vs. MALP	0,2613	-0,1728 to 0,6953	No	ns	0,4413
	Ctrl vs. Zym	0,1913	-0,2428 to 0,6253	No	ns	>0,9999
	LPS vs. MALP	0,0225	-0,4116 to 0,4566	No	ns	>0,9999
	LPS vs. Zym	-0,0475	-0,4816 to 0,3866	No	ns	>0,9999
	MALP vs. Zym	-0,07	-0,5041 to 0,3641	No	ns	>0,9999
	Ctrl vs. C.a.	-1,015	-1,58 to -0,4504	Yes	**	0,0011
	Ctrl vs. Asp.	-0,5775	-1,142 to -0,01295	Yes	*	0,0442
	Ctrl vs. E.coli	0,4813	-0,0833 to 1,046	No	ns	0,1113
	C.a. vs. Asp.	0,4375	-0,1271 to 1,002	No	ns	0,1704
	C.a. vs. E.coli	1,496	0,9317 to 2,061	Yes	****	<0,0001
	Asp. vs. E.coli	1,059	0,4942 to 1,623	Yes	***	0,0008
FABP5	Ctrl vs. LPS	-0,5675	-1,595 to 0,4604	No	ns	0,5773
	Ctrl vs. MALP	-0,92	-1,948 to 0,1079	No	ns	0,0881
	Ctrl vs. Zym	-0,08125	-1,109 to 0,9466	No	ns	>0,9999
	LPS vs. MALP	-0,3525	-1,38 to 0,6754	No	ns	>0,9999

	LPS vs. Zym	0,4863	-0,5416 to 1,514	No	ns	0,8758
	MALP vs. Zym	0,8388	-0,1891 to 1,867	No	ns	0,1359
	Ctrl vs. C.a.	-1,405	-2,47 to -0,34	Yes	**	0,0098
	Ctrl vs. Asp.	-2,679	-3,744 to -1,614	Yes	****	<0,0001
	Ctrl vs. E.coli	0,575	-0,49 to 1,64	No	ns	0,6162
	C.a. vs. Asp.	-1,274	-2,339 to -0,2087	Yes	*	0,018
	C.a. vs. E.coli	1,98	0,915 to 3,045	Yes	***	0,0009
	Asp. vs. E.coli	3,254	2,189 to 4,319	Yes	****	<0,0001
BAG3	Ctrl vs. LPS	-0,4325	-0,9285 to 0,06352	No	ns	0,1
	Ctrl vs. MALP	-0,4	-0,896 to 0,09602	No	ns	0,1433
	Ctrl vs. Zym	-1,126	-1,622 to -0,6302	Yes	***	0,0002
	LPS vs. MALP	0,0325	-0,4635 to 0,5285	No	ns	>0,9999
	LPS vs. Zym	-0,6938	-1,19 to -0,1977	Yes	**	0,0067
	MALP vs. Zym	-0,7263	-1,222 to -0,2302	Yes	**	0,0049
	Ctrl vs. C.a.	-2,763	-3,418 to -2,107	Yes	****	<0,0001
	Ctrl vs. Asp.	-2,005	-2,66 to -1,35	Yes	****	<0,0001
	Ctrl vs. E.coli	-0,605	-1,26 to 0,05025	No	ns	0,0756
	C.a. vs. Asp.	0,7575	0,1023 to 1,413	Yes	*	0,0221
	C.a. vs. E.coli	2,158	1,502 to 2,813	Yes	****	<0,0001
	Asp. vs. E.coli	1,4	0,7448 to 2,055	Yes	***	0,0003
NPC1	Ctrl vs. LPS	-0,05375	-0,6742 to 0,5667	No	ns	>0,9999
	Ctrl vs. MALP	-0,365	-0,9854 to 0,2554	No	ns	0,475
	Ctrl vs. Zym	-1,426	-2,047 to -0,8058	Yes	***	0,0002
	LPS vs. MALP	-0,3113	-0,9317 to 0,3092	No	ns	0,7544
	LPS vs. Zym	-1,373	-1,993 to -0,7521	Yes	***	0,0002
	MALP vs. Zym	-1,061	-1,682 to -0,4408	Yes	**	0,0016
	Ctrl vs. C.a.	-0,8438	-1,571 to -0,1168	Yes	*	0,0216
	Ctrl vs. Asp.	-2,191	-2,918 to -1,464	Yes	****	<0,0001
	Ctrl vs. E.coli	-0,67	-1,397 to 0,05697	No	ns	0,0763
	C.a. vs. Asp.	-1,348	-2,074 to -0,6205	Yes	***	0,0009
	C.a. vs. E.coli	0,1738	-0,5532 to 0,9007	No	ns	>0,9999
	Asp. vs. E.coli	1,521	0,7943 to 2,248	Yes	***	0,0004

HMOX1	Ctrl vs. LPS	2,444	1,689 to 3,199	Yes	****	<0,0001
	Ctrl vs. MALP	1,675	0,92 to 2,43	Yes	***	0,0002
	Ctrl vs. Zym	1,021	0,2663 to 1,776	Yes	**	0,0083
	LPS vs. MALP	-0,7688	-1,524 to -0,01375	Yes	*	0,0454
	LPS vs. Zym	-1,423	-2,177 to -0,6675	Yes	***	0,0008
	MALP vs. Zym	-0,6538	-1,409 to 0,1012	No	ns	0,1034
	Ctrl vs. C.a.	-0,07375	-1,063 to 0,9158	No	ns	>0,9999
	Ctrl vs. Asp.	-3,239	-4,228 to -2,249	Yes	****	<0,0001
	Ctrl vs. E.coli	3,406	2,417 to 4,396	Yes	****	<0,0001
	C.a. vs. Asp.	-3,165	-4,155 to -2,175	Yes	****	<0,0001
	C.a. vs. E.coli	3,48	2,49 to 4,47	Yes	****	<0,0001
	Asp. vs. E.coli	6,645	5,655 to 7,635	Yes	****	<0,0001
CCR1	Ctrl vs. LPS	1,26	0,6538 to 1,866	Yes	***	0,0004
	Ctrl vs. MALP	0,6938	0,08753 to 1,3	Yes	*	0,0234
	Ctrl vs. Zym	1,439	0,8325 to 2,045	Yes	***	0,0001
	LPS vs. MALP	-0,5663	-1,172 to 0,03997	No	ns	0,0713
	LPS vs. Zym	0,1788	-0,4275 to 0,785	No	ns	>0,9999
	MALP vs. Zym	0,745	0,1388 to 1,351	Yes	*	0,0153
	Ctrl vs. C.a.	-1,114	-1,903 to -0,3249	Yes	**	0,0063
	Ctrl vs. Asp.	-1,198	-1,986 to -0,4086	Yes	**	0,0038
	Ctrl vs. E.coli	2,29	1,501 to 3,079	Yes	****	<0,0001
	C.a. vs. Asp.	-0,08375	-0,8726 to 0,7051	No	ns	>0,9999
	C.a. vs. E.coli	3,404	2,615 to 4,193	Yes	****	<0,0001
	Asp. vs. E.coli	3,488	2,699 to 4,276	Yes	****	<0,0001

Curriculum vitae

PERSONAL INFORMATION

João Pedro Leonor Fernandes Saraiva

Huygensstr. 22, 04159 Leipzig (Germany)

joao.saraiva@ufz.de

WORK EXPERIENCE

01/02/2013–31/01/2014	Research grant University of Minho, Braga (Portugal) Reconstruction of metabolic network of S. pneumoniae R6.
08/2009–09/2011	Expropriation technician Promapa, Bragança (Portugal) Negotiation of expropriation of terrains needed for construction of highway A4 between government and property owners. Identification of ownership of terrains to be expropriated when absent from public records.
11/2007–04/2009	Assistant manager in the production of table olives Porttable, Freixo de Espada à Cinta (Portugal) Quality control assessment of final product and responsible for facility hygiene. Implementation and supervision of quality control documentation.
06/2007–10/2007	Quality control manager Brigantauto, Bragança (Portugal) Implementation of a quality assessment plan in the company.

EDUCATION AND TRAINING

01/02/2014–Present	PhD in Bioinformatics University of Jena, Jena (Germany)	EQF level 8

Identification of biomarkers for distinguishing fungal from bacterial infections employing Support Vector Machines.

10/2009–12/2012 Master in Bioinformatics EQF level 7

University of Minho, Braga (Portugal)

Master's project: "HpyloriHub - A comprehensive *Helicobacter pylori* Information Resource"

01/2007–04/2007 Quality Control Technician EQF level 7

Superior Insitute of Language and Administration (ISLA), Bragança (Portugal)

Formulate and implement quality control measures and establish documentation pipeline

09/2000–08/2006 Diploma in Biotechnology Engineering EQF level 7

Polytechnic Institute of Bragança, Bragança (Portugal)

Bachelor project: "Germination of *Juniperus oxycedrus* seeds in vitro".

Diploma project: "Biopolimer production".

PERSONAL SKILLS

Mother tongue(s) Portuguese

Other language(s)

	UNDERSTANDING		SPEAKING		WRITING
	Listening	Reading	Spoken interaction	Spoken production	
English	C2	C2	C2	C2	C2
Spanish	C2	C1	B2	B2	A2
German	A2	A2	A2	A2	A1
French	A2	A2	A2	A2	A1

Levels: A1 and A2: Basic user - B1 and B2: Independent user - C1 and C2: Proficient user
Common European Framework of Reference for Languages

Communication Excellent communication skills - Negotiation, regular scientific

skills reports.
Good ability to adapt to new environments (travelling).
Team player - Practicing collective sports, collaboration partners.

Organisational / managerial skills
Good organization and management skills - Assistant production manager at Porttable and Quality control manager at Brigantauto.
Student Representative at Academic Association of the Polytechnic Institute of Bragança (2003-2006)
Organizing team of Scientific conference (Micom2015)
Organizing team of Bioinformatics workshop (Bioinformatics Open Days 2013)

Job-related skills
Good understanding of quality control implementing procedures.
Good mentoring skills - teaching gene expression analysis and R programming language to students.

Digital competence

SELF-ASSESSMENT				
Information processing	Communication	Content creation	Safety	Problem solving
Proficient user	Proficient user	Proficient user	Independent user	Independent user

Digital competences - Self-assessment grid

Good programming skills in R.
Basic understanding of Unix/Linux systems.
Basic programming in Python.

Driving licence B, BE

ADDITIONAL INFORMATION

Courses Scientific Presentation in English

Conferences	Sepsis 2014 (Paris, France) - Poster presentation
	JSMC Symposium (2014) (Jena, Germany)- Poster presentation
	Micom2015 (Jena, Germany) - Poster presentation
	JSMC Symposium (2015) (Bad Sulza, Germany) - Poster presentation
	4th Conference in Constraint-Base Reconstruction and Analysis (2015)(Heidelberg, Germany) - 2 Poster presentations
	Weimar Sepsis Update (2015) (Weimar, Germany) - Poster presentation
	Annual Conference of the Association for General and Applied Microbiology (2016) (Jena, Germany) - Attendance
	6th Conference on Foundations of Systems Biology in Engineering (2016)(Magdeburg, Germany) - Poster presentation
	Micom2017 (Jena, Germany) - Oral presentation

Jena, 30.11.2017

9. Ehrenwörtliche Erklärung

Hiermit erkläre ich, dass mir die Promotionsordnung der Medizinischen Fakultät der Friedrich-Schiller-Universität Jena bekannt ist, ich die Dissertation selbst angefertigt habe und alle von mir benutzten Hilfsmittel, persönlichen Mitteilungen und Quellen in meiner Arbeit angegeben sind, mich folgende Personen bei der Auswahl und Auswertung des Materials sowie bei der Herstellung des Manuskripts unterstützt haben: Prof. Dr. Rainer König und Prof. Dr. Reinhardt Guthke, die Hilfe eines Promotionsberaters nicht in Anspruch genommen wurde und dass Dritte weder unmittelbar noch mittelbar geldwerte Leistungen von mir für die Arbeit erhalten haben, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen, dass ich die Dissertation noch nicht als Prüfungsarbeit für eine staatliche oder andere wissenschaftliche Prüfung eingereicht habe und dass ich die gleiche, in wesentlichen Teilen ähnliche oder eine andere Abhandlung nicht bei einer anderen Hochschule als Dissertation eingereicht habe.

Jena, 30.11.2017 João Saraiva

10. Acknowledgements

The completion of my doctoral thesis represents a hallmark in my life. Therefore, I would like to thank the innumerous people who in some form contributed to my work.

First, I would like to thank Prof. Dr. Rainer König, who accepted me in his Network Modelling group and allowed me to carry out my doctoral thesis. His support, help, advice and overall attention was overwhelming. I would also like to thank Prof. Dr. Reinhardt Guthke for his support and advice whenever I requested it.

I would like to thank all of my current and former colleagues at the Network Modelling group throughout these years. Their companionship, advice and lively discussions helped make this an amazing experience.

Special thank you to Volker Ast for all the fun and interesting discussions at lunch time, not necessarily related to work, which usually lit up the day.

I would like to give special thanks to Dr. Daniela Röll for all the afternoons spent discussing professional matters as well as the social "coffees" so many times helpful for brainstorming. Her attention and time spent on discussing my work was invaluable and will be forever remembered.

I would also like to thank all of my collaboration partners for all their help and cooperation during these years. These are Prof. Dr. Hortense Slevogt as well as members of her group Dr. Tilman Klassert, Cristina Zubíria-Barrera and Maximillian Lautenbach without which this thesis would not be so complete. I would like to give special thanks to Dr. Tilman Klassert for all the time he spent discussing with me the results and pitching new ideas.

An enormous thanks to my family for all the support and love they have given me all of these years and without which my journey would be harder.

I would like to thank Volha Skrahina for always pushing me to do better and providing support, fun times and companionship.

Last but not least I would like to thank all of my friends who, in some way or shape, always were there for me and provided me with plenty of relaxing and fun moments.