MRC National Institute for Medical Research

The Study of Blood Transcriptional Signatures to Improve Medical Management and Understanding of Active Pulmonary Tuberculosis and Similar Respiratory Diseases Including Sarcoidosis

Chloe Isabel Bloom

August 2012

Division of Immunoregulation
MRC National Institute for Medical Research
The Ridgeway
Mill Hill, London
NW7 1AA

Submitted to the University College London for the degree of doctor of philosophy

I, Chloe Bloom confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Table of Contents

Table of Figures

Table of Tables

Acknowledgments

I would like to thank Anne O’Garra, for supporting me throughout, for her motivation, guidance and extensive knowledge and understanding of the project and background behind it. I am most grateful to Marc Lipman who throughout has not only provided the lion’s share of TB patients but as my clinical supervisor has been persistently supportive and encouraging and a pleasure to work with. I would like to thank all the patients and healthy controls who volunteered to participate in this study, without which there would not be a study. I would like to thank the many people who assisted in the recruitment of the patients, in particular Fotini Rozakeas who was incredibly helpful with recruitment and was a great companion to me. I have also been fortunate enough to have a great deal of help from many physicians in helping to recruit patients including Huw Beynon, Marc Lipman, Ling-pei Ho, Rama Vancheeswaran, Melissa.Wickremasinghe, Yvonne West, Diane Bouvry and the Lyon Collaborative Network. I have also had instrumental discussions about sarcoidosis with Rama Vancheeswaran, Melissa.Wickremasinghe, Huw Beynon and Ling-pei Ho. I would like to particularly thank Christine Graham, without whom we could not have achieved this study, for all her hard work and processing of the samples for microarray. I would also like to thank Matthew Berry for his help in my obtaining the MRC funding for this study and input towards my findings.

I would like to thank many people who I have helped me in a many different ways while at MRC National Institute for Medical Research, including those who have worked in Anne’s lab, Paul Redford, Fin McNab, and Leona Gabrysova, all of whom have taught me many practical aspects, with whom I have held many helpful scientific conversions throughout my PhD and have been good companions throughout. I would like to thank the proof readers of my thesis John Ewbank and Simon Blankley. I would like to thank my thesis committee whose guidance has been invaluable, Douglas Young,

Jonathan Stoye and James Turner. In addition many other people have helped me in many different ways during my PhD including John Wills, Kush Meghji, Simon Caidan, Anna Gibson, Ben Kesel, Donna Brown, Marna Roos, and Graham Preece.

I am also very grateful for a number of people at Baylor Institute for Immunological Research including Esperanza Anguiano, Raylynn Cleveland, Yuanyuan Wang, Benjamin Lemoine, Zhaohui Xu, Jose Rossello-urgell, Virginia Pascual, Damien Chaussabel and Jacques Banchereau.

Finally, I would also like to thank Ben, my mum, my dad and George for their support, tolerance and encouragement.

ABSTRACT

Introduction

Tuberculosis is the leading cause of death from curable infectious diseases. New approaches for prevention, diagnosis, and treatment are urgently needed. Understanding the underlying immunopathogenesis is vital to achieve this. Transcriptional profiling of peripheral blood has been successfully applied to inflammatory and infectious diseases to improve understanding of disease mechanisms. Berry *et al.* 2010, recently revealed distinct transcriptional signatures of pulmonary tuberculosis, leading to new knowledge on tuberculosis pathogenesis. Transcriptional profiling also differentiated active TB from other infections and inflammatory diseases. This present study compared whole blood transcriptional profiles of pulmonary tuberculosis to the similar respiratory diseases sarcoidosis, community acquired pneumonia and primary lung cancer.

Methods

Microarray technology and data mining strategies were used to examine whole blood genome-wide transcriptional profiles from patients and controls, before and after treatment.

Results

Transcriptional profiles of tuberculosis and sarcoidosis were comparable to each other but disparate from pneumonia and lung cancer profiles. The dominant genes in the tuberculosis and sarcoidosis profiles were the over-abundance of interferon-inducible genes, the genes showed a higher expression in the tuberculosis patients. The dominant genes in the pneumonia and cancer profiles were the over-abundance of inflammation genes, and under-abundance of protein translation genes in the pneumonia profiles. 144-transcripts were able to distinguish the tuberculosis patients from all other samples with good sensitivity and specificity. The transcriptional profiles from the tuberculosis, pneumonia and sarcoidosis patients significantly changed after receiving successful treatment. The tuberculosis profiles significantly changed by two weeks after treatment initiation, earlier than any validated biomarker of treatment response.

Conclusions

This study has provided new insight into the parallels and differences of the molecular signatures of these similar respiratory diseases. The findings may have also revealed prospective pragmatic biomarkers for disease diagnosis and treatment monitoring which are being further investigated.

INTRODUCTION

INTRODUCTION

Over nine million new cases of active tuberculosis (TB), and 1.4 million deaths from TB, are estimated to occur around the world every year (WHO 2010). The 2010 World Health Organisation TB report states that TB is a disease of poverty, however it is also recognised that TB is on the rise in some developed countries. Estimates based predominantly on a dermal delayed type hypersensitivity response, the tuberculin skin test, indicate that global prevalence of *Mycobacterium tuberculosis* (*M. tuberculosis)* infection is about 32% (Dye, Scheele et al. 1999). Paradoxically the majority of those thought to be infected are asymptomatic and have no clinical evidence of disease; these individuals are described as having latent TB. Epidemiological studies carried out in both developing and developed countries describe around 5-10% of latent individuals developing active TB during their lifetime, with the highest risk following infection in early adulthood and the lifetime risk declining each year of infection (Vynnycky and Fine 2000) (Comstock, Livesay et al. 1974). The risk is substantially higher in individuals who are immunosuppressed particularly those with HIV co-infection (Corbett, Watt et al. 2003). In the year 2000 these problems were addressed by the World Health Organisation in the plan entitled Stop TB Partnership: The Global Plan to Stop TB 2006-2015 (Young, Perkins et al. 2008). Progression in the fight against TB is severely restricted by a lack of knowledge of both how the host employs mechanisms to kill the bacilli and how the bacterium evades them.

Even in developed countries where there is access to an array of medical facilities correctly diagnosing active TB can also be challenging, time consuming and distressing for the patient. Misdiagnosis and delays in the diagnosis lead to a worse outcome for the patient and potential spread of this infectious disease (Greenaway, Menzies et al. 2002). The difficulty in diagnosing pulmonary TB is the ability to differentiate it from other similar respiratory diseases such as pulmonary sarcoidosis, pneumonia and lung cancer (Wyngarden 1988). Sarcoidosis is also a multisystem granulomatous disorder, but has unknown aetiology, that affects individuals worldwide and is characterized pathologically by the presence of granulomas in involved organs (Iannuzzi, Rybicki et al. 2007). Both sarcoidosis and TB can affect adults within the same age group, who then present with respiratory symptoms and radiological thoracic abnormalities (Iannuzzi, Rybicki et al. 2007) (Anderson, Maguire et al. 2007). Community acquired pneumonia, like TB, is a respiratory infection, and lung cancer is another lung inflammatory disorder (O'Callaghan, O'Donnell et al. 2010). Due to the complexity of these diseases a systems biology approach offers the ability to help unravel the principal host immune responses by simultaneous comparison of the host's transcriptional response to each of these similar diseases. Furthermore the need to apply a genomics approach to improve clinical management of respiratory diseases and elucidate mechanisms of lung pathophysiology is evident by the recent National Heart, Lung, and Blood Institute workshop 'Genomic Medicine and Lung Diseases' (Center, Schwartz et al. 2012). Their overall goal was for 'omics' research to be leading translational lung disease research within the next 5 years.

Gene expression profiling of the peripheral blood has recently been successfully applied to certain inflammatory and infectious diseases, providing new understanding of disease pathogenesis and improved diagnostic and prognostic biomarkers (Pascual, Chaussabel et al. 2010). An earlier study carried out by O'Garra and collaborators of patients with active TB, latent TB and healthy controls used an unbiased comprehensive whole human genome survey of peripheral blood to demonstrate a distinct transcriptional signature in patients with active TB (Berry, Graham et al. 2010). The transcriptional signature was associated with disease severity by demonstrating a correlation with the extent of the radiographic severity. In addition it was observed that after successful antituberculous therapy the transcriptional signature had diminished. This active TB signature therefore has great potential to be developed for diagnostic and treatment monitoring biomarkers. Furthermore 10-20% of the latent TB individuals also revealed the same transcriptional signature as active TB patients. Longitudinal studies are in progress to determine whether this transcriptional signature can predict which latent individuals will progress towards the development of active TB.

The main objective of this current study was to improve our understanding of the immunopathogenesis underlying TB by comparing common and unique transcriptional patterns of active TB to the similar granulomatous disease sarcoidosis, and to the other similar pulmonary diseases pneumonia and lung cancer. This approach may also provide much needed biomarkers to help in the diagnosis and management of both TB and sarcoidosis.

Incidence of TB, sarcoidosis, community acquired pneumonia and lung cancer

The gradual rise in TB over the last 20 years in the UK is thought to be primarily due to migration patterns to the UK (HPA 2010). This has led to over 9,000 new cases of active TB in 2009 (HPA 2010), and an annual incidence rate of around 41.3/100,000 in London in 2003, where nearly three-quarters of cases were born abroad (Anderson, Maguire et al. 2007). Sarcoidosis varies in incidence among geographical regions and can also aggregate in families and specific races, being three to four times more common in blacks (Thomas and Hunninghake 2003). In the United States the annual incidence among whites is 10.9/100,000 and among blacks is 35.5/100,000 (Rybicki, Major et al. 1997). In the UK the annual incidence of sarcoidosis is under half that of

TB, with approximately 3000 new cases of sarcoidosis annually with the highest reported incidence of sarcoidosis occurs in London (Gribbin, Hubbard et al. 2006). However it has been recognised over the last decade that it is not just in the developed countries that both sarcoidosis and TB are present. Sarcoidosis is seen in India, and in some other developing countries, with almost similar frequency as in the West (Jindal, Gupta et al. 2000). As sarcoidosis is a difficult disease to diagnose and often presents little or no symptoms it is likely its prevalence is much higher than reported, particularly in countries with a more limited health care service. Community acquired pneumonia is far more common than TB or sarcoidosis, with an incidence in developed countries of 5-1100/100,000 adults annually (Lim, Baudouin et al. 2009). In the UK the majority of patients are treated in the community, however approximately one third will require hospital admission (Guest and Morris 1997). Lung cancer is both the most common cancer in the world and has the highest mortality (Bray, Ren et al. 2012). In the UK the incidence of lung cancer has steadily fallen since the 1970s, reflecting the fall in the prevalence of male smokers after World War II (82% of men smoked in 1948 compared to 22% today) (CancerStats 2012). However the rates of lung cancer in females have continued to rise. In 2009 the annual incidence rate of lung cancer in the UK was 48/100,000, with a prevalence of just over 41,000 cases of which 44% were women.

Therefore in the UK, TB in comparison to these similar respiratory diseases has a compatible incidence to sarcoidosis and lung cancer, while community acquire pneumonia is the most prevalent.

Challenges in Tuberculosis Diagnosis

Several clinical tests are used to help diagnose TB; each has its own problems.

The tuberculin skin test

It has been shown in experimental mouse models that two weeks after infection with *M. tuberculosis,* memory T cells that specifically recognise *M. tuberculosis* antigens start to appear (Cooper 2009). This immune recall response is detectable and measurable after an intradermal injection of tuberculin, called the tuberculin skin test. The most commonly used tuberculin is the purified protein derivative (PPD). PPD is a precipitate of heat-inactivated concentrate of the culture filtrate of non-species specific *M. tuberculosis* grown in glycerol broth. It contains a mixture of degraded proteins derived from secreted cytosol, cell wall and membrane proteins but there are also polysaccharides and some lipids present (Lee and Holzman 2002). Most of the constituents of PPD are small proteins with molecular masses of approximately 10,000 Daltons, and it is thought due to their small size PPD does not sensitize individuals who have not been exposed to mycobacteria (ATS 2000). The appearance of skin induration at the injection site therefore indicates the presence of a delayed type hypersensitivity response specific for mycobacterial antigens. Cellular infiltration by T cells in combination with other recruited inflammatory cells, such as monocytes and subsequently activated macrophages, results in a maximal cutaneous induration at 48 to 72 hours after inoculation (Vukmanovic-Stejic 2006). This localised immune response shows a predominance of CD4+ T cells with an effector memory cell phenotype (Sarrazin, Wilkinson et al. 2009). The ability to mount such a dermal response to the TST is usually maintained for many years. The TST is one of the oldest diagnostic tests employed in modern medicine and is still widely used throughout the world due to its low cost, however it has many flaws as a diagnostic tool. Even when applying the proper technique and interpretation of findings, cross-reaction due to other mycobacterial antigens commonly gives false-positive results. The dermal immune response cannot be used to discriminate between individuals with latent TB, active TB, those who have been vaccinated with *Mycobacterium bovis* bacillus Calmette-Guerin (BCG) or those exposed to some environmental nontuberculous mycobacteria. This is because PPD contains over 200 antigens also present in the BCG vaccine (an attenuated form of *Mycobacterium bovis*) and many nontuberculous mycobacteria (Richeldi 2006). This is confounded by false negatives where up to 25% of culture confirmed active TB patients have a negative skin test response (Nash and Douglass 1980).

Interferon gamma release assays & M. tuberculosis specific antigens

The problems of using the TST as a diagnostic tool has led to the development and use of Interferon Gamma Release Assays (IGRAs) that can measure interferon gamma (IFN-γ) production from sensitised T cells in response to stimulation by relatively specific *M. tuberculosis* antigens. The main antigens used are present in the RD1 region of *M. tuberculosis*, a region shown to contribute to the virulence of the bacteria, and considered to be the primary attenuating deletion in BCG (Mahairas, Sabo et al. 1996). The RD-1 locus contains 9 genes including two secreted proteins, early secreted antigen-6 (ESAT-6) and culture filtrate protein-10 (CFP-10), that are shown to be prominent T cell antigens. There are two types of IGRAs that differ from each other mainly with respect to the two techniques used for IFN-γ detection, either an enzyme linked immunospot assay (TSPOT.TB^{TM}, Oxford Immunotec, Oxford, UK) or an enzyme linked immunosorbent assay (QuantiFERON-Gold In Tube, Cellestis, Carnegia, Australia). This study uses the QuantiFERON-TB Gold In-tube ELISA, which involves mixing whole blood with the mycobacterial antigens ESAT-6, CFP-10, and TB 7.7, which are not found in any BCG strains nor in the majority of non-tuberculous mycobacteria. Although the two tests have comparable sensitivity and specificity the ELISA has some advantages over the TSPOT.TB^{TM} as samples can be stored and run in batches, it requires less investment in equipment, it is more cost effective, and the assay is technically easier to perform. The Quantiferon-TB Gold In-tube has a specificity of 99% in non-BCG vaccinated patients and 96% in BCG vaccinated patients (Pai, Zwerling et al. 2008).

Measuring IFN-γ

The amount of IFN-γ secreted in response to *M. tuberculosis* antigens has not been robustly quantified. Some studies demonstrate a lower response in patients with active TB compared to those with latent TB (Hirsch, Toossi et al. 1999; Pathan, Wilkinson et al. 2001; Vekemans, Lienhardt et al. 2001), which reverts after successful TB treatment (Vekemans, Lienhardt et al. 2001). This could be due to a reduced number of T cells present in the blood of patients with active TB (Berry, Graham et al. 2010). The reduction in peripheral T cells may be due to migration of effector T cells to the site of disease or an increased susceptibility to apoptosis (O'Garra 2008). However, some studies show the opposite findings, with latent individuals having a lower IFN-γ response than active TB patients (Cardoso, Antas et al. 2002; Janssens, Roux-Lombard et al. 2007). As the ELISPOT assay measures the number of activated cells not the total amount secreted quantifying this reduces the quantitative capacity of the assay to measure the IFN-γ response. Therefore due to the lack of consistency and difficulty in quantification, measuring IFN-γ alone is not a satisfactory method to distinguish patients with active TB from those with latent TB.

The multiplicity of TB

Neither the IGRAs nor the TST can distinguish between those individuals with active TB and those with latent TB. These available diagnostic tests can only tell us which individuals have been exposed such that they have subsequently developed an acquired immune response. Post mortem studies in humans with latent TB have shown a wide range of recovery of viable bacilli (Barry, Boshoff et al. 2009**).** Indeed at the time of immunological testing the bacilli may no longer be viable. Hence the current definition of ‘latent TB’ really includes a spectrum of individuals from those who have possibly completely expelled the infection to those who may contain active replicating bacteria but remain asymptomatic (Barry, Boshoff et al. 2009). One animal model that has been found to have a similar pathology to human TB, and a similar spectrum of lesion types, are the cynomolgus macaques. The observed histopathologic features of the granulomas in *M. tuberculosis* infected macaques also reflected a dynamic and heterogeneous process during both active and latent infection (Lin, Rodgers et al. 2009).

It is recommended that the majority of latently infected individuals be treated with antibiotics for 3-9 months (NICE 2011). This incurs a significant cost to the health care system and subsequent risk of side effects. However this strategy has been proven to be necessary because preventive treatment of individuals with latent TB diminishes the risk of subsequent development of active TB by about 90% (Richeldi 2006). Although currently we are unable to identify which latent TB individuals will benefit from prophylactic treatment, research is currently in progress to determine a suitable stratification.

As discussed above there is evidence of heterogeneity in the clinical and histological presentation of TB (Barry, Boshoff et al. 2009). This concept was recently strengthened by the demonstration of molecular heterogeneity corresponding to the clinical heterogeneity (Berry, Graham et al. 2010). This study, from O’Garra and collaborators, demonstrated for the first time through whole genome peripheral blood gene expression profiling the heterogeneity of the transcriptional host response in patients with TB (Berry, Graham et al. 2010). In fact the heterogeneity of the active TB signature was unambiguously explained by the severity of their disease, as defined by the extent of their radiographic changes. The heterogeneity of the latent signature maybe explained by the spectrum of latent disease and as such those with a signature similar to active TB may have had subclinical disease with a high short-term possibility of reactivation, however this hypothesis requires verification.

Unfortunately current diagnostic tests are a long way from distinguishing between these diverse ranges of latent individuals, limiting the ability of new and available technologies to evolve our clinical management. For example, a recent study using polychromatic flow cytometry was able to show discriminating differences in cytokine profiles of *M. tuberculosis*-specific CD4+ T cells from patients with active TB versus latent TB (Harari, Rozot et al. 2011). Regrettably the cohort of latent patients recruited for this study were a specific subset of the spectrum, identified by screening rather than probable exposure to *M. tuberculosis* therefore it is likely the cohort only had limited exposure and thus little or no infection. Classifying patients along this spectrum would have a huge impact on clinical management by proposing only to treat those who have a high risk of developing disease. In addition a more accurate classification will aid research studies to appropriately interpret findings when comparing between the various subtypes of TB.

Diagnosing active TB

Making a diagnosis of active TB requires a completely different approach from that of latent TB. The diagnosis is usually suspected in those with classical symptoms including a productive cough, drenching night sweats and weight loss, along with the typical radiological abnormalities such as cavities, densities and thoracic lymphadenopathy (Wyngarden 1988). The gold standard for the diagnosis of active TB is *M. tuberculosis* cultured from specimens such as sputum, lung washings, lymph node biopsy, cerebrospinal fluid – although notably culture can take up to 6 weeks due to the slow growing nature of the bacteria (Pfyffer, Cieslak et al. 1997).

Rapid detection of mycobacterial bacilli can be achieved by the 125 year old sputum microscopy smear test, but this is not very specific as it cannot distinguish *M. tuberculosis* from other mycobacteria, nor very sensitive as it is only able to detect 60% of culture positive pulmonary *M. tuberculosis* (Young, Perkins et al. 2008). A major clinical problem when trying to diagnose active pulmonary TB is the lack of an adequate sputum sample, resulting in many patients undergoing an invasive procedure to obtain lung washings where possible (Tamura, Shimada et al. 2010). In the USA public health surveillance identified that only 70% of pulmonary TB is diagnosed by bacterial culture, therefore a significant number of patients receive empirical antituberculous treatment (CCDC 2007). In developing countries such as South Africa, the number of patients with a diagnosis of pulmonary TB confirmed by culture is far lower, at approximately 50% (WHO 2010). A relatively new development in TB diagnosis is the use of PCR to detect *M. tuberculosis* and common drug resistance strains (Taegtmeyer, Beeching et al. 2008). In December 2010, WHO endorsed the Xpert MTB/RIF automated molecular test for *M. tuberculosis* and rifampicin resistance (Boehme, Nabeta et al. 2010). However to diagnose pulmonary TB this molecular diagnostic is used to detect the presence of *M. tuberculosis* in sputum, thus limiting its application to those in whom a sputum or bronchial washings sample is available.

One of the diseases pulmonary TB is particularly difficult to distinguish from is sarcoidosis; requiring invasive procedures to determine between the differential diagnoses and thus leading to delay in treatment (Storla, Yimer et al. 2008).

Misdiagnosis of TB as sarcoidosis could lead to worsening disease outcome for the patient if treated incorrectly with the immunosuppressive therapy that is required for sarcoidosis and misdiagnosis of sarcoidosis as TB could lead to unpleasant side effects secondary to 6 months or more of multiple anti-tuberculous antibiotics.

Challenges in Sarcoidosis Diagnosis

The diagnosis of sarcoidosis is made by exclusion.

Diagnosis and clinical presentation

The diagnosis of sarcoidosis is complex as it can only be made by exclusion of other causes of granulomatous inflammation. The predominant disease sarcoidosis must be differentiated from is TB (Box 1). Currently only an invasive biopsy and other semi-invasive tests can help distinguish between these clinically analogous diseases.

Until quite recently a standard diagnostic test was the reaction to an intradermal injection of homogenates of human sarcoid tissue (spleen or lymph node), called the Kveim-Siltzbach reagent (Wyngarden 1988). Approximately 80% developed a granulomatous dermal inflammatory reaction several weeks after the injection (Munro and Mitchell 1987). However due to safety concerns of transmission of infections the reagent was discontinued in the UK in 1996.

On average, sarcoidosis patients have symptoms for more than three months and require three or more encounters with health care providers prior to diagnosis (Judson, Thompson et al. 2003)**.** Sarcoidosis patients presenting with pulmonary symptoms often have a further relative delay in the diagnosis of sarcoidosis as their symptoms are nonspecific therefore alternative diagnoses are often considered first. When assessing patients with suspected sarcoidosis no single test is specific enough or sensitive enough for diagnosis. Diagnosis should only be made by those with a specialist interest in sarcoidosis using compatible histological, radiological and clinical findings (Costabel and Hunninghake 1999). Patients present with a heterogeneous clinical picture, with pulmonary involvement occurring in over 90% and about 35% having extra-pulmonary disease (Baughman, Teirstein et al. 2001; Rizzato, Palmieri et al. 2004). Each affected organ, commonly lung, skin, lymph node or eye, is involved to a varying extent and degree. Approximately half of all cases are detected incidentally by pulmonary radiological abnormalities discovered in asymptomatic individuals having a routine chest radiograph (eg pre-operative work up or unrelated chest pains). Whilst the thorax is the most common site of disease, the skin is involved in at least 30% of patients and the eye in about 25% (Baughman, Teirstein et al. 2001).

Box 1

Pulmonary TB	Pulmonary Sarcoidosis
Thoracic radiology	Thoracic radiology
Cavities, opacities, lymphadenopathy	Opacities, lymphadenopathy
Respiratory symptoms	Respiratory symptoms
Cough, haemoptysis (blood)	Cough, dyspnoea
Systemic symptoms	Systemic symptoms
Drenching night sweats, weight loss	Fatigue (less common sweats, weight loss)
Biopsy of lung or lymph node	Biopsy of lung or lymph node
Necrotising granuloma's	Non-necrotising granuloma's
Microbiology on sputum/BAL	
Smear +ve (detects 60% of culture +ve)	
Culture +ve (only able to culture in 70%)	Diagnosis is made by compatible histological, radiological and clinical presentation & by elimination of other causes.
Diagnosis is made by culture of *M. tuberculosis*	

Pulmonary sarcoidosis is often assessed by three methods: thoracic radiological parameters, lung function tests and respiratory symptoms (Baughman, Teirstein et al. 2001). Pulmonary sarcoidosis affects either the thoracic lymph nodes, typically the hilar lymph nodes, the pulmonary parenchyma (lung tissue) or both the lymph nodes and the lung parenchyma. The disease elicits a range of presentations from those with no symptoms, the majority of patients, to those with severe and debilitating symptoms (Baughman, Teirstein et al. 2001). Pulmonary fibrosis is the commonest chronic phenomenon of sarcoidosis and occurs in 20 to 25% of patients (Iannuzzi, Rybicki et al. 2007).

Sarcoidosis activity and prognosis

The natural history and prognosis of sarcoidosis is protean. To add to the complexity there remains no consensus on how to reliably and pragmatically assess disease activity, disease severity or prognosis (Box 2). Furthermore disease activity may not correlate with disease severity or prognosis as active inflammation does not always indicate poor prognosis (WASOG 1999). For example, a patient with highly active disease at their first presentation may have an excellent prognosis and full recovery after only several months of treatment. Similarly a patient with irreversible pulmonary fibrosis and a poor prognosis may in fact have little or no on-going disease activity.

Each Patient Should Be Assessed For

Disease Activity

Disease Severity

Prognosis

Box 2

Although spontaneous resolution occurs in two-thirds of patients within 5 years it is not possible to predict this on an individual basis (WASOG 1999). The only widely acknowledged classification system is Scadding's criteria (Scadding 1961). Scadding's criteria solely use chest radiographs (Table 1) to classify the patients, therefore only accounts for the patient's pulmonary involvement. This classification system is unfortunately insufficient for reliable clinical decision making as it cannot guide physicians in making treatment decisions or reliably inform on prognosis.

Chest radiograph stage	*Radiological findings*	*Spontaneous resolution*
Stage 1	BHL*	75%
Stage 2	BHL and lung opacities	60%
Stage 3	Shrinking BHL and opacities	<30%
Stage 4	Lung fibrosis	None

Table 1. Scadding's criteria.

BHL = bilateral hilar lymphadenopathy

Standardising sarcoidosis phenotyping (Box 3) is imperative and may be helped by advances in genomic research and increased application of genetic profiling. The most common published classification schemes require clinical information that has been gathered over a period of time, often defining patients as either 'acute or chronic' or 'self-limited or progressive', but not considering disease activity on each presentation (Prasse, Katic et al. 2008; Lockstone, Sanderson et al. 2010). The most well described phenotype is Lofgren's syndrome, usually an acute presentation, typically associated with a good prognosis and spontaneous remission; but even Lofgren's syndrome has no uniform definition. As this study is taking a snap shot view of the host response a similar ‘snap shot’ approach was applied to clinically phenotyping the patients. Patients were phenotyped solely using their clinical features around the time of their blood sampling, irrespective of their disease severity, predicted prognosis or previous disease activity status. Therefore patients were classified purely as either those with active disease or non-active disease as defined at the time of the blood test.

Possible Sarcoidosis Classifications

Box 3

Active or Inactive*

Acute or Chronic

Progressive or Self-limited

Mild to Severe

Steroid responsive or non-responsive

Thoracic or extra-thoracic

Chest radiograph stages I-IV

Lofgren’s or not

**** = used in this study***

Disease activity assessment should reflect on-going persistent inflammation with evolving granuloma formation (WASOG 1994). Clinical findings thought to correlate with disease activity include: symptoms (WASOG 1999), elevated serum soluble IL-2 receptor (Keicho, Kitamura et al. 1990), serum angiotensin converting enzyme (Ainslie and Benatar 1985), serum neopterin (Homolka, Lorenz et al. 1992), serum hypergammaglobulinaemia (Mana, Salazar et al. 1996), blood lymphopenia (Morell, Levy et al. 2002; Sweiss, Salloum et al. 2010), bronchoalveolar lavage lymphocyte count (Leung, Brauner et al. 1998), change in chest radiographic disease (Keir and Wells 2010), presence of pulmonary nodules (Abehsera, Valeyre et al. 2000); (Wells 1998), gallium scan activity (Klech, Kohn et al. 1982), activity on positron emission tomography scans (Keijsers, Verzijlbergen et al. 2009), and changes in lung function test (Keir and Wells 2010) (Box 4).

Box 4

Possible Sarcoidosis Activity Markers

Serum ACE*

Chest radiograph stage*

Lung function changes

Soluble IL2 receptor

Neopterin

BAL lymphocyte count

Acute Phase Response Proteins

Respiratory symptoms*

Systemic symptoms

Gallium scan

CT findings

Serum IgG*

Serum lymphocyte count*

PET scan findings

Physician commenced treatment

**** = used in this study***

As all of these markers of disease activity are not specific for sarcoidosis each one can be altered in the presence of many other diseases. Thus there is no validated or established disease activity score, a reflection of the lack of a ‘gold standard’ test for assessing activity and progression of granulomatous inflammation specific to sarcoidosis. Classifications used frequently require the collation of knowledge on changes in clinical data, therefore preventing patient phenotyping without prolonged clinical assessment. Consequently studies apply a variety of different classification systems, including those based solely on the management plan of the practising physicians (Miyara, Amoura et al. 2006). However patient management maybe subjective and can vary between physicians and medical centres, for example whether to commence glucocorticoids or other immunosuppressive medications, the starting dose and the incremental dosing regimen.

For this study the patients were phenotyped into those with active or non-active disease, using our clinical classification system, to determine whether the patient’s transcriptional profiles correlated with disease activity. The classification system applied was based entirely on clinical evidence available from published literature and clinical variables that were available for the patients recruited at the various hospitals. The classification system used did not rely on progressive clinical information gathering and simply classed patients as either those with active disease or those with non-active disease, at the time of the blood test. The classification system also did not use detailed radiographic scores as due to the complexity of sarcoidosis thoracic radiological findings this would of required superior levels of inter-observer reliability across the many recruitment centres (Wasfi, Rose et al. 2006).

Histology of granulomatous inflammation

The role of granulomas is not always clear, are they purely protective for the host or do they contribute towards tissue pathology?

Granulomatous inflammation occurs in many diseases

Granuloma formation is fundamental to the immunopathogenesis of both sarcoidosis and tuberculosis, but is a relatively non-specific histological finding. A granuloma is a focal area of inflammation defined as a compact collection of cells of the monocyte lineage (macrophages, epithelioid cells, and multinucleated giant cells or fused epithelioid cells) with or without the presence of other inflammatory cells including lymphocytes (Adams 1976). It is commonly suggested that granulomas are part of the host's defence against exogenous and endogenous particles; the causative agent is walled off and sequestered by cells of macrophage lineage allowing it to be contained, if not destroyed altogether (Williams and Williams 1983). However many multisystem granulomatous disorders of unknown aetiology exist, such as sarcoidosis and Wegener's granulomatosis, in which it appears that granulomas have no protective function but instead are part of the disease pathology (Agostini, Adami et al. 2000). Most lung granulomas are associated with infection, particularly mycobacteria and fungal disease. The granulomatous lung diseases thought to be non-infectious are sarcoidosis, Wegener's granulomatosis, hypersensitivity pneumonitis, aspiration pneumonia, and talc granulomatosis (Mukhopadhyay and Gal 2010). The granulomas formed by infection tend to be necrotising and well formed, often in a random distribution.

TB granulomas

TB granulomas have been studied for over a century in particular in the guinea pig, which alongside the rabbit are thought to be the small animals that most closely resemble the immunopathological response found in humans infected with *M. tuberculosis* (McMurray 2001; Saunders and Orme 2008) . The first phase of the primary pulmonary lesion in guinea pigs is the influx of granulocytes, eosinophils and heterophils (neutrophil-equivalent cells), after which numerous macrophages and lymphocytes, with fewer granulocytes, coalesce to form the classical TB granuloma, before further expansion into the lung parenchyma and the formation of a central necrotic focus (Saunders and Orme 2008). Rabbits infected with highly virulent strains of *M. tuberculosis* can develop granulomas containing layers of macrophages, lymphocytes and fibroblasts surrounding a caseous necrotic centre, and can further mimic humans by developing cavities when granulomas are located near an airway (McMurray 2001; Saunders and Orme 2008). In recent years studies have shown increasing value for the non-human primate as an experimental model that can closely mimic the spectrum of human TB, although practical issues have ultimately limited their use in TB research. Lung histology from *M. tuberculosis* infected cynomolgus macaques, presenting with active TB, has revealed a variety of granuloma types not only across the macaques but also within each organ (Lin, Rodgers et al. 2009). Three main types have been described: the classical caseous granuloma, with central eosinophilic debris surrounded by macrophages and a layer of lymphocytes; the non-necrotising granuloma, with an internal compact core of macrophages and some neutrophils surrounded by a lymphocyte layer; the suppurative granuloma, with a central core of degenerative neutrophils surrounded by macrophages and multinucleated giant cells and an outer envelope of lymphocytes (Lin, Rodgers et al. 2009). In fact this variety of granuloma types was discovered in human post-mortem studies over 50 years ago, even in lesions of only 1mm^3, in patients considered to have ‘minimal pulmonary TB’ and who did not die from their disease (Medlar 1948).

Human TB granulomas are composed centrally of a mass of infected macrophages, stimulated macrophages that have differentiated into multinucleated giant cells, epithelioid cells and foamy macrophages loaded with lipid droplets, and neutrophils (Russell, Cardona et al. 2009). This inner accumulation of cells becomes surrounded by lymphocytes, largely CD4+ T cells but also CD8+ T cells and B-cells; as well as fibroblasts creating a peripheral fibrotic capsule (Peters and Ernst 2003). A variety of proinflammatory and inhibitory cytokines and chemokines, in addition to adhesion proteins, play key roles in the formation of granulomas. A study of lung tissue specimens from patients with multiple drug resistant TB found that the formation of granulomas required a minimal size of 0.1mm^3 (Ulrichs, Kosmiadi et al. 2004). They also reported the presence of lymphoid follicle-like structures in the peripheral margins of the granulomas, composed predominantly of B cells and some CD4+ and CD8+ T cells, surrounding infected macrophages. They concluded from their findings that mycobacteria can survive both within the granulomas, in the periphery of the granulomas and even further afield in apparently normal healthy parenchymal tissue (Ulrichs, Kosmiadi et al. 2004).

One of the classical features of human TB granulomas is the presence of a necrotic caseous core thought to be secondary to cell lysis and resulting in a central hypoxic, hostile environment (Tsai, Chakravarty et al. 2006). The caseous necrotic granulomas can then rupture releasing extracellular tubercle bacilli into the alveoli and airways, encouraging dissemination and infectivity. One theory is that in latent TB the bacilli reside in the central hypoxic zone in a dormant metabolically inactive state, but in active TB they are able to replicate in peripheral oxygenated areas (Barry, Boshoff et al. 2009). This raises the question whether the granulomas are in fact protecting the *M. tuberculosis.* Indeed the pathogen may be able to engineer a supportive environment for example through the manipulation of macrophage lipid metabolism (Russell, Cardona et al. 2009). In the zebrafish model it can be observed that intracellular mycobacteria induce recruitment of macrophages to early granulomas, suggesting the mycobacteria are using the host to facilitate the spread of infection (Davis and Ramakrishnan 2009). Although it should be remembered the zebrafish model uses *Mycobacterium marinum* and does not have an adaptive immune system, but this model could perhaps portray the response seen in TB patients who do not have an adequate adaptive immune response such as HIV co-infected individuals.

Sarcoidosis granulomas

In contrast to TB, sarcoidosis granulomas are non-necrotising, well-formed and track along the lymphatics, interlobular septa, bronchovascular bundles and pleura (Gerke and Hunninghake 2008). As the sarcoidosis granulomas mature, fibroblasts and collagen encase the ball of cells (Mitchell, Scadding et al. 1977). An integral part of sarcoidosis granulomas are epithelioid cells, differentiated macrophages with secretory and bactericidal capability. These epithelioid cells produce serum angiotensin converting enzyme (ACE), a commonly used surrogate marker of sarcoidosis disease activity as it is thought to reflect granuloma burden (Silverstein, Pertschuk et al. 1979). However this biomarker has poor specificity as it can be elevated in other diseases, particularly disorders with granulomatous inflammation such as TB (Ainslie and Benatar 1985), or reduced to within a normal range due to genetic variations (Biller, Zissel et al. 2006). Sarcoidosis granulomas are typically indistinguishable from the lesions in chronic beryllium disease, a granulomatous lung disorder caused by beryllium exposure and characterized by accumulation of beryllium-specific CD4+ T cells (Amicosante and Fontenot 2006). However unlike chronic beryllium disease, the main causative antigen(s) in sarcoidosis remain anonymous.

Immunology of tuberculosis

The protective and pathologic responses to M. tuberculosis are complex and multifaceted, involving many components of the immune system.

M. tuberculosis has adopted many unique features that have allowed it to successfully adapt to its often harsh, nutrient deficient host environment. The cell envelope is composed of two layers: an atypical cell wall, which is composed of mycolic acids, glycolipids and structural polymers, and the plasma membrane; both layers protect the bacterium from the host's immune response and ensure the bacterium's survival by importing nutrients and exporting products that interact with the host (Kaufmann and Rubin 2008). One of the cell wall's features, its resistance to dehydration, acids and alkalis, aids our ability to identify the bacilli during isolation under microscopy in samples such as sputum. This resistance to acids during staining of the bacilli can result in rapid identification of infected patients that are then subsequently labelled as 'smear positive'.

M. tuberculosis gains entry into the lung through aerosol inhalation where it is thought the first line of defence is the resident alveolar macrophages and recruited neutrophils (Eum, Kong et al. 2010). Pattern recognition receptors (PRRs) expressed on innate cells can mediate the uptake of bacteria including the tubercle bacilli, into the host's cells (Korbel, Schneider et al. 2008). It is recognised from the murine model that the Toll-like receptors (TLR)-2, TLR-4 and TLR-9, other PRRs such as C-type lectins (Mincle, mannose receptor and DC-SIGN), and their adaptor proteins, are all likely to play a critical but complex role in the antimycobacterial activity of macrophages and dendritic cells during *M. tuberculosis* infection (Edwards, Manickasingham et al. 2002; Ishikawa, Ishikawa et al. 2009; Dorhoi, Desel et al. 2010). In humans a link between TLR2 activation, vitamin D and mycobacterial killing has been clearly demonstrated in macrophages and monocytes (Liu, Stenger et al. 2006). It has been shown that once the pathogen is phagocytosed by the macrophage it has evolved several strategies to avoid the macrophage's intracellular killing mechanisms. These include inhibition of phagolysosome fusion, resistance to reactive nitrogen intermediates and inhibition of phagosome acidification (Armstrong and Hart 1971; Sturgill-Koszycki, Schlesinger et al. 1994; MacMicking, Xie et al. 1997; Cosma, Sherman et al. 2003). Initially the innate immune response continues to predominate as the activated macrophages produce a plethora of cytokines and chemokines to stimulate the migration of neutrophils, lymphocytes and mononuclear phagocytes (Ulrichs and Kaufmann 2006). In addition if the infected macrophage does not survive it may either die by necrosis, a traumatic cell death that potentially allows for further spread of the bacilli, or by apoptosis, where the plasma membrane remains intact and often confers a more protective outcome for the host (Behar, Martin et al. 2011). Therefore if apoptosis is prevented this can encourage the survival and growth of the bacilli as shown by the actions of a common virulent strain of *M.tuberculosis* in mice (H37Rv) that has the ability to inhibit the lipid mediator prostaglandin E2, a promoter of apoptosis (Divangahi, Desjardins et al. 2010). Detection of bacilli by the myeloid cells via PRRs and complement receptors, often results in the processing of mycobacterial antigens to enable the antigen presentating cells to activate T lymphocytes as key mediators of an acquired immune host response. Studies in mice have shown that T cell activation is likely to be initiated in the draining lymph nodes as a result of infected dendritic cells migrating from the lung (Wolf, Linas et al. 2007; Wolf, Desvignes et al. 2008). However activation of effector T cells appears to be slow, and what is more cannot be accelerated by adoptive transfer of antigen-specific T helper 1 (Th1) CD4+ cells, suggesting perhaps that initially post-infection the bacilli are able to hide or resist T cell antimycobacterial responses (Gallegos, Pamer et al. 2008). As the bacterium perseveres, a CD4+ T cell mediated response prevails while progressive remodelling of the site of infection with lymphocyte recruitment and further macrophage activation ultimately brings about the formation of granulomas (Cooper 2009). The naïve CD4+ T helper cells recognise peptides+ from the phagocytosed tubercle bacilli in association with MHC-class II molecules on the surface of the antigen presenting cells such as dendritic cells. This encounter drives the differentiation of naïve CD4+ T helper cells to Th1 cells by cytokines, including IL-12 and IL-18 derived from the antigen presenting cells, and IFN-γ produced from CD4+ T cells, CD8+ T cells and natural killer cells (Flynn and Chan 2001). Some of the key cytokines produced by the effector Th1 cells are IFN-γ, IL-2 and tumour necrosis factor- α (TNF-α). *M. tuberculosis* lipid antigens can also be processed and presented to unconventional T cells such as γδT cells and NKT cells (Tanaka, Morita et al. 1995; De Libero and Mori 2008).

Protective and pathogenic factors associated with human TB

It is often suggested that of those individuals that have been chronically exposed to *M.tuberculosis* only about 10-30% become infected, as evidenced by an acquired delayed sensitivity to *M. tuberculosis* proteins (Kassim, Zuber et al. 2000; North and Jung 2004). However studies in humans are obviously limited in their ability to prove exposure to a sufficient bacterial load (Fennelly, Jones-Lopez et al. 2012), determine if infection even occurred, or determine if the innate or the adaptive or both immune responses played a role in any protective responses. Even so it is speculated that both the innate and adaptive immune systems play a significant defensive part against prolonged infection with *M. tuberculosis.* An example of a potential innate response is a T-cell independent natural resistance that was suggested from a study by Cobat *et al*, where they were able to show asymptomatic individuals who had a negative delayed type hypersensitivity response from the TST were linked with a major chromosomal locus 11p14, in a highly endemic area in South Africa (Cobat, Gallant et al. 2009). The role of the adaptive immune response is far more evident due to the prolific prevalence of *M. tuberculosis* and HIV co-infection, where the reduced numbers of CD4+ T cells secondary to HIV infection facilitates mycobacterial activity and causes florid clinical disease (Corbett, Watt et al. 2003).

The cytokines IL-12 and IFN-γ have both been shown to be crucial in controlling *M. tuberculosis* infection. Although neither on their own are able to halt the pathogen, their presence is essential for protection (Cooper 2009). The importance of IL-12 and IFN-γ was initially demonstrated in experimental mouse models (Flynn, Chan et al. 1993; Cooper, Roberts et al. 1995). Subsequent studies of patients with autosomal inheritance of susceptibility to mycobacterial infection, including *M.tuberculosis*, have supported the animal findings and demonstrated that susceptibility can be caused by mutations in the genes for IL-12, STAT1 or the receptors for IFN-γ or IL-12 (Jouanguy, Altare et al. 1996; Altare, Durandy et al. 1998; Ottenhoff, Kumararatne et al. 1998; Boisson-Dupuis, El Baghdadi et al. 2011). A further demonstration of the complexity of the role of IFN-γ has been shown by the lack of success to treat adult TB, particularly MDR-TB, with IFN-γ in patients with no known genetic mutations (IFN-γ has been successfully used in patients with IL-12/IL-12R mutations). Studies in HIV-negative patients did initially show some promise with a transient decrease in the bacillary load however the overall efficacy of the therapy is questionable, even when therapy is received directly in to the lungs (Reljic 2007). Only one HIV-negative patient has been reported to have been successfully treated with adjuvant IL12 therapy (Greinert, Ernst et al. 2001). Other primary immunodeficiencies have also led to the discovery of single genes that are critical for antimycobacterial immunity. These include a mutation in the IRF8 gene, which is required for dendritic cell and monocyte survival, and a mutation in the CYBB gene, which encodes one of the phox subunits of NADPH oxidase required for respiratory burst in phagocytes (Bustamante, Arias et al. 2011; Hambleton, Salem et al. 2011).

Another cytokine demonstrated to have a crucial function is TNF-α, produced both by many immune cells including macrophages and stimulated T cells, it has been connected not only with immune protection but also with the formation of granulomas (Flynn and Chan 2001). Its job in controlling latent infection has been implicated by evidence of a five-fold increase in the rate of reactivation of *M. tuberculosis* infection occurring in individuals with Crohn’s disease or rheumatoid arthritis that were treated with anti-TNF-α or TNF-α receptor antibodies (Keane, Gershon et al. 2001; Gardam, Keystone et al. 2003; Long and Gardam 2003). When comparing a *M. tuberculosis*-specific CD4+ T cell response in the peripheral blood of active and latent TB patients, a significant TNF-α response was seen in active disease, although possibly reflecting an elevated degree of inflammation rather than protection (Harari, Rozot et al. 2011). From animal models its role in protection is clearly delineated in non-human primates (Lin, Myers et al. 2010) and in the past in the mouse model it has been suggested that TNF-α may help in the maintenance of the granulomas (Mohan, Scanga et al. 2001). This now appears unlikely as in the non-human primate model it has more recently been shown that the monkeys are still able to develop typical granulomas after anti-TNF-α treatment, although they remain unable to control the infection (Lin, Myers et al. 2010). In addition histology from three patients who reactivated TB after anti-TNF-α therapy were found to have normal granulomas present (Iliopoulos, Psathakis et al. 2006). Multifunctional CD4+ T cells secreting IFN-γ, TNF-α, and IL-2 in the lungs of mice have been proposed as correlates of protection after BCG immunisation (Forbes, Sander et al. 2008). Interestingly Harari *et al* showed an elevated number of multifunctional *M. tuberculosis*-specific CD4+ T cells in in the blood of latent individual’s compared to active TB patients (Harari, Rozot et al. 2011). However Kagina *et al* examined the blood of BCG vaccinated infants and were unable to find any specific cytokine profile of BCG-specific T cells, including multifunctional cells, which correlated with protection (lack of development of active TB within two years of vaccination) (Kagina, Abel et al. 2010).

Although the role of IFN-γ is well recognised, the role of type 1 IFN in TB is not as clearly documented. The recent human TB microarray study from O'Garra and collaborators revealed a correlation between disease severity and neutrophil driven type I IFN-inducible genes in the blood of patients with active TB (Berry, Graham et al. 2010). However exactly what role IFN-γ and type I IFNs play in human TB remains unclear. In the murine model the loss of the common type I IFN receptor (*ifnar1-/-*) or an over-abundance of type I IFNs (by a hyper-virulent *M.tuberculosis* strain) have shown that type I IFNs promotes *M.tuberculosis* infection (Manca, Tsenova et al. 2005). A recent study in *M.tuberculosis* infected human macrophages observed a suppression of the protective cytokine IL-1 by type I IFN, not seen in RD1-deficient strains (Novikov, Cardone et al. 2011). The role of the neutrophil in TB is also not well acknowledged although studies in patients with advanced and multiple drug resistant TB have shown the neutrophil to be the principal cell in untreated aspirates from cavities (Ulrichs, Kosmiadi et al. 2004), moreover the neutrophil was unexpectedly found to be the dominant *M.tuberculosis* infected cell type in both sputum and cavities (Eum, Kong et al. 2010). A vaccination study in mice has also demonstrated the potential for neutrophils to become infected by mycobacteria, and furthermore then act as carrier for the bacilli to the draining lymph nodes (Abadie, Badell et al. 2005). In addition there is evidence from patients with active TB of a correlation between worse outcome and peripheral neutrophilia, although whether this association is cause or effect remains unknown (Barnes, Leedom et al. 1988; Bandara, Bremner et al. 2008).

A cytokine suggested to modulate the host response in TB mouse models is IL-10. A deficiency in IL-10 results in a lower mycobacterial load, earlier Th1 response and an enhanced Th1 response in the mouse (Beamer, Flaherty et al. 2008; Cooper 2009; Redford, Boonstra et al. 2010; Redford, Murray et al. 2011). In patients with active TB, IL-10 has been associated with peripheral anergy (Boussiotis, Tsai et al. 2000), there is evidence of an increased production of IL-10 in pleural samples (Barnes, Lu et al. 1993) and neutralising antibodies to IL-10 resulted in increased IFN-γ in peripheral blood cells (Gong, Zhang et al. 1996). However, IL-10 polymorphism studies have added little knowledge as they show conflicting results, although perhaps just reflecting ethnic-specific genetic variations (Lopez-Maderuelo, Arnalich et al. 2003). A meta-analysis of IL-10 polymorphism studies revealed no statistical evidence of an association with active TB but did indicate a trend towards protection in association with certain IL-10 polymorphisms and pulmonary TB alone (Pacheco, Cardoso et al. 2008).

Another mechanism appearing to down modulate the immune response in patients with active TB is the cell surface signalling molecule programmed death-1 (PD-1) and its ligands (PDL-1, PDL-2). PD-1 is expressed by T cells in TB patients and stimulation of peripheral blood with sonicated *M. tuberculosis* up-regulated T cell expression of PDL-1, while blocking the PD-1/ligand system increased *M. tuberculosis*-specific IFNγ response (Jurado, Alvarez et al. 2008). Its involvement is further suggested by microarray analysis of active TB patient’s whole blood, which revealed a relative increased abundance compared to latent patients, but unexpectedly an association predominantly with neutrophils (McNab, Berry et al. 2011).

For over 60 years vitamin D supplementation has been suggested to aid mycobacterial killing (Charpy, Dowling et al. 1947). More recently it has been reported the odds of developing active TB are at least five-fold higher in vitamin D deficient individuals, either HIV-negative or HIV-positive (Martineau, Nhamoyebonde et al. 2011). Furthermore it has now been shown that vitamin D has the propensity to act as a key immune cofactor in both innate and adaptive antimycobacterial activities. In *M.tuberculosis* infected human macrophages mycobacterial growth was inhibited secondary to TLR2/1 stimulation in the presence of calcitriol (the active form of vitamin D, 1,25-dihyrdoxyvitamin) as this triggered the induction of the antimicrobial peptide cathelicidin (Liu, Stenger et al. 2006). Moreover cathelicidin production was significantly diminished in the serum of black patients with deficient calcidiol (which converts to calcitriol) levels, compared to white patients with sufficient calcidiol levels; production was then reversed in the black subjects with the addition of calcidiol (Liu, Stenger et al. 2006). A study by the same research group using a similar protocol also demonstrated the necessity of vitamin D for the adaptive response. *M.tuberculosis* infected human monocytes stimulated with IFN-γ in the presence of calcidiol-sufficient serum from white patients resulted in a reduction of viable bacilli, which was not seen in the presence of calcidiol-deficient serum from black patients. Although this *in vitro* data is promising for a therapeutic role for Vitamin D supplementation, a large randomised controlled trial of active TB patients showed that only patients with a vitamin D receptor polymorphism responded significantly and favourably to vitamin D supplementation in addition to standard antituberculous therapy (Martineau, Timms et al. 2011). Investigation of the effects of supplementation at an earlier stage of infection, in order to prevent those with latent infection from developing active TB, could be important.

Other human risk factors that have been linked with development of active disease include diabetes mellitus, alcohol excess and a smoking history, each with suggested biological plausibility relating to direct impairment of the host immune response (Bates, Khalakdina et al. 2007; Jeon and Murray 2008; Lonnroth, Williams et al. 2008).

Treatment and treatment monitoring of tuberculosis

Inadequate treatment and poor treatment monitoring leads to worsening disease, an increase in disease transmission, and spread of drug resistance.

TB treatment

In 1948 the British MRC conducted a landmark randomised controlled trial of streptomycin, the first drug to be successfully commenced for the partial treatment of pulmonary TB (Crofton 2006). A few years later in the early 1950s it was reported that isoniazid had activity against *M. tuberculosis* and by the mid-1960s rifampicin was added to the triple therapy of streptomycin, isoniazid, and pyrazinamide (Murray 2004). Because *M. tuberculosis* develops spontaneous, random, resistance mutations to streptomycin, ethambutol, isoniazid and rifampicin, each drug cannot be used alone (David 1970). These resistance mutations occur independently (where the highest risk of resistance is secondary to isoniazid) therefore the chances of any bacilli having spontaneous resistance to 3 or 4 drugs is extremely low (David 1970). In addition each drug works by a different mechanism and may complement each other. Isoniazid and ethambutol are used at bactericidal doses, rifampicin has additional sterilising activity, and pyrazinamide is bactericidal in acidic environments e.g. inside macrophages or areas of acute inflammation (Mitchison 1985). The optimal duration of treatment and combination of drugs was determined by numerous clinical trials conducted in the 1970s and 1980s by the British MRC, British Thoracic Association and Hong Kong Chest Service (ATS and CDC 2003). The standard course of antituberculous treatment still remains the 'short course' regimen which consists of six months of antibiotics. At the end of the course the patient is considered cured, although the global relapse rate in 2010 was approximately 4% (WHO 2010). If the bacterium are thought to be fully sensitive to first line drugs then treatment is started with isoniazid, rifampicin, ethambutol and pyrazinamide for two months (the intensive phase) followed by isoniazid and rifampicin for four months (the continuation phase) (WHO 2009). In the UK, the National Institute for Health and Clinical Excellence guidelines suggest ethambutol is only added if there is a possibility of drug resistance (NICE 2011). From numerous *in vitro* and *in vivo* studies it is observed that the intensive phase serves to rapidly kill those bacilli that are actively-multiplying, including both intracellular and extracellular organisms (Grosset 1980; Jain, Lamichhane et al. 2008). On the other hand, the continuation phase has a lower rate of killing due to the sterilizing activity against the persisting bacilli which are limited in number but undergo intermittent multiplication, and due to the reduced effectiveness of all the drugs on these bacilli therefore require a longer course of treatment to prevent relapses (Grosset 1980; Jain, Lamichhane et al. 2008).

TB treatment monitoring

After initiation of antituberculous treatment it is important that the patient's response is closely monitored. Monitoring allows observation for potential treatment side effects and most importantly for treatment interruptions and/or identifying if the patient is responding sufficiently (WHO 2009). Inadequate treatment commonly occurs due to poor patient compliance, lack of appropriate antibiotics, concurrent pathology or infection with drug resistant *M. tuberculosis* (WHO 2009). Inadequate treatment and poor treatment monitoring, causes worsening of an individual's disease, increased potential for disease spread, and an increased risk of the development and spread of drug resistant *M. tuberculosis*.

Currently the only validated and accepted biomarker of treatment success or failure is the 2-month sputum conversion test (Mitchison 1993). This requires a culture positive sputum sample prior to treatment and repeat sputum sample 2 months after treatment. If the repeat sample does not culture *M. tuberculosis* this suggests the patient has responded successfully to treatment. However there are many limitations to this test. Firstly many patients are not able to produce sputum samples, even prior to treatment health workers are unable to obtain samples from approximately 30% of patients in the USA and 50% of South African patients (CCDC 2007; WHO 2010). These difficulties in obtaining sputum are likely to be further exacerbated after successful treatment. Moreover patients who are unable to expectorate sputum at 2 months may be potentially incorrectly labelled as having a negative culture (Perrin, Lipman et al. 2007). In addition sputum culture is time consuming taking several weeks to grow the bacilli (Pfyffer, Cieslak et al. 1997), and results can be compromised by contamination (Small, McClenny et al. 1993). Furthermore, although sputum conversion is commonly used as a surrogate end point for treatment response in clinical trials evaluating new drugs, a systematic review and meta-analysis to assess its accuracy in predicting an individual's treatment failure revealed low sensitivity and only modest specificity (Horne, Royce et al. 2010; Wallis, Pai et al. 2010). While other biomarkers have also been trialled, including serum C-reactive protein, IFN-γ and neopterin, all have similarly shown poor sensitivity and specificity (Walzl, Ronacher et al. 2008). Chest radiographs are commonly used in the clinical setting as a marker of treatment response however they too have many limitations. Firstly chest radiographs generally improve slower than the clinical response and lack specificity as interpretation can be confounded by previous lung damage (Perrin, Lipman et al. 2007). Furthermore interpretation of radiographic changes in response to treatment has not yet been standardised, and the facilities are often not available in developing countries (Walzl, Ronacher et al. 2011).

There remains to date no available early biomarkers, before 2 months of treatment, correlating with treatment success or failure. TB treatment monitoring is a major challenge for global attempts to eradicate *M. tuberculosis* infection. So much so that in April 2010 the Center for Disease Control and National Institutes of Health brought together experts in the field and research scientists with the sole purpose of addressing this problem (Nahid, Saukkonen et al. 2011). Early biomarkers of treatment response are not only useful in clinical TB management but also play a significant role in clinical trials as surrogate markers of a pharmacological response. The preferred surrogate endpoints for TB treatment trials are early bactericidal activity, sputum culture and smear conversion rates (Jain, Lamichhane et al. 2008) (Nahid, Saukkonen et al. 2011). Early bactericidal activity (EBA) is the measure of the fall in viable colony forming units of *M. tuberculosis* in the sputum, and is a reliable measure of the loss of metabolically active *M. tuberculosis* during a 1-2 week course of therapy (Donald and Diacon 2008). However EBA has many problems as a surrogate marker including unknown correlation with the endpoint of treatment ‘cure’, high variability between patients, the lack of measurement of sterilising activity of bacilli persisters and the requirement for sputum samples (Jain, Lamichhane et al. 2008; Nahid, Saukkonen et al. 2011). In addition EBA is typically used to assess single agents, although a recent study has demonstrated feasibility with its use with multiple-agent combinations, interestingly response is often seen in two phases with the first two days showing the greatest rate of change in metabolic activity (Diacon, Dawson et al. 2012). Without surrogate endpoints the success of antituberculous treatment is commonly determined by treatment failure and the treatment relapse rate after 2 years, resulting in expensive, time-consuming and lengthy drug trials (Jain, Lamichhane et al. 2008). Treatment failure is defined as continuous positive sputum cultures while receiving an appropriate antibiotic treatment regimen; treatment relapse is defined as a patient who becomes culture-negative while on treatment but deteriorates later and becomes culture-positive after stopping treatment (ATS and CDC 2003). Most patients relapse within 6-12 months of completing treatment (Nunn, Phillips et al. 2010). These patients should be distinguished from patients who become re-infected with a different strain (van Rie, Warren et al. 1999).

Novel treatment monitoring biomarkers are greatly needed for both clinical management and drug development. Some areas that have been suggested include better radiological tools such as positron emission tomography, which may have better sensitivity and specificity than chest radiographs and computer tomography (Hofmeyr, Lau et al. 2007); better methods to measure the mycobacterial load such as the use of *M. tuberculosis* mRNA levels in sputum (Desjardin, Perkins et al. 1999); and ultimately the development of host genomic, proteomic and metabolomic tools (Walzl, Ronacher et al. 2008).

Aetiology and immunology of sarcoidosis

Little is known about either the cause or underlying immune mechanisms of sarcoidosis.

Aetiology

Because sarcoidosis usually affects the lungs, skin and eyes many airborne aetiological agents have been proposed. Although associations with exposures have been reported including mycobacterial antigens, insecticides and airborne substances post the World Trade Centre disaster, there has been no identification of a single predominate agent (Newman, Rose et al. 2004), (Izbicki, Chavko et al. 2007). One of the most widely proposed aetiological agents is *M.tuberculosis*. Although it has not been possible to culture *M. tuberculosis* in large studies, some small studies have shown evidence of mycobacterial DNA by PCR with prevalence varying between 0-50% and a meta-analysis suggesting a prevalence of 30% (Brown, Brett et al. 2003; Gupta, Agarwal et al. 2007). Furthermore multiple studies have been carried out to investigate the immune response, peripherally and at the site of disease, to a range of *M. tuberculosis* antigens: ESAT-6, *M. tuberculosis* catalase-peroxidase (MKatG), PPD and mycobacterial superoxide dismutase A, where responses were assessed by IFN-γ assays and flow cytometry (Song, Marzilli et al. 2005; Carlisle, Evans et al. 2007; Drake, Dhason et al. 2007; Allen, Evans et al. 2008; Chen, Wahlstrom et al. 2008; Oswald-Richter, Culver et al. 2009; Oswald-Richter, Sato et al. 2010; Oswald-Richter, Beachboard et al. 2010; Oswald-Richter, Beachboard et al. 2012). The patients in these studies had a range of sarcoidosis phenotypes, and controls were patients with other granulomatous diseases and/or healthy BCG unvaccinated participants. Taking into account the variations in participant phenotypes, type of samples and antigens used between studies, there was a significant immune response to *M. tuberculosis* antigens in at least a subset of sarcoidosis patients. Gupta *et al* carried out a systematic review of all MEDLINE studies since 1965 focussing on T and B cell responses to tubercular antigens in patients with sarcoidosis. They found a non-significant trend towards a mycobacterial T cell immune response in sarcoidosis patients and a significant trend towards a B cell response, compared to PPD-ve controls (Gupta, Agarwal et al. 2011). More promising though was a proteomics approach used to detect tissue antigens which identified antigenic bands with the same physicochemical properties as the Kveim-Siltzbach reagent. The enzyme *M. tuberculosis* catalase-peroxidase (mKatG) was identified as one of the tissue antigens that were present in significantly more patients with sarcoidosis than in the controls (Song, Marzilli et al. 2005). Given that this approach was not predicated on any specific hypothesis regarding aetiology, only the assumption that the antigen would be a poorly soluble protein within the granuloma, the recovery of a specific mycobacterial antigen provides support for a mycobacterial link to sarcoidosis granulomas. Nevertheless whatever the relationship is between *M. tuberculosis* and sarcoidosis it appears to be complicated. For example case reports of patients diagnosed with both TB and sarcoidosis are not common and nor is a prior diagnosis of TB a recognised risk factor for developing sarcoidosis (Rybicki, Iannuzzi et al. 2001). Furthermore the incidence of sarcoidosis is higher in the United States than in TB endemic countries (Jindal, Gupta et al. 2000), although this may just reflect increased awareness of sarcoidosis in the United States. Multiple studies of the components of the Kveim-Siltzbach reagent have failed to identify a responsible antigen but studies have demonstrated the Kveim reaction is characterised by CD4+ T cells with an oligoclonal TCR expansion consistent with an antigen specific host response (Moller 2007).

In summary, the aetiology of sarcoidosis is unknown, yet despite the lack of conclusive evidence in favour or against, one of the most commonly proposed hypotheses remains mycobacteria.

Immunological response of sarcoidosis

It is generally accepted that sarcoidosis is mediated by a MHC-restricted antigen driven process, based on observations such as an oligoclonal expansion of αβ T cells at the sites of disease and a restrictive repertoire of T cell receptors (Silver, Crystal et al. 1996; Grunewald, Wahlstrom et al. 2002). It has been proposed that the granuloma occurs as a consequence of a continuous exaggerated immune response against unknown antigen(s) capable of persisting at the site of disease, perhaps due to poor solubility and degradability (Agostini, Adami et al. 2000). It is thought there is both macrophage and T cell activation with CD4+ T cell differentiation into Th1 phenotype (Hunninghake and Crystal 1981; Agostini, Adami et al. 2000). The role of CD4+ T cells is intimated by the reactivation of sarcoidosis that can occur during treatment for HIV (Foulon, Wislez et al. 2004). At least early in the disease course it has been found that the dominant cytokine expression in the serum and lung are Th1 cytokines including IFN-γ, IL-12 and IL-2, and TNFα from stimulated macrophages (Agostini, Basso et al. 1998; Gerke and Hunninghake 2008). A recent study also reports elevated levels of T helper 17 cells in the peripheral blood and BAL of patients with active sarcoidosis (Facco, Cabrelle et al. 2011). Interestingly IFNα therapy is a well-documented risk factor for developing sarcoidosis, occurring in around 5% of hepatitis C patients treated with IFNα (Hoffmann, Jung et al. 1998), while IFNß has also been reported to induce sarcoidosis in several cases reports (Chakravarty, Harris et al. 2012). Consistent with the concept that sarcoidosis is caused by local stimuli most studies demonstrate a compartmentalised immune response more pronounced at the site of disease than in the blood (Wahlstrom, Katchar et al. 2001; Thillai, Eberhardt et al. 2012). However an elevated immune response can also be detected in the blood and an increase in soluble IL-2 receptor has been shown to be associated with increased disease activity (Grutters, Fellrath et al. 2003). The interaction of CD4+ T cells with antigen presenting cells initiates the formation and maintenance of the granulomas. The role of PRRs in sarcoidosis is unknown and studies looking for TLR polymorphisms as candidate susceptibility genes have been inconclusive (Schurmann, Kwiatkowski et al. 2008).

It is commonly observed that sarcoidosis patients develop peripheral anergy, in particular the lack of a dermal response to the TST (Demirkok, Basaranoglu et al. 2007). This phenomenon may be related to the reported accumulation of regulatory FOXP3+ T cells in patients with active sarcoidosis, in their BAL and peripheral blood, which could result in the suppression of IL-2 secretion and strongly inhibit T cell expansion and migration (Miyara, Amoura et al. 2006). Alternatively it may be related to an apparent reduced function of dendritic cells in peripheral blood of sarcoidosis patients, a finding also shown to be associated with increased disease activity (Mathew, Bauer et al. 2008). Cobat et al identified a non-MHC locus (in 5p15 region) linked to reduced tuberculin skin test reactivity in a TB endemic area (Cobat, Gallant et al. 2009). As this locus is also linked with sarcoidosis susceptibility, this association could relate to the peripheral anergy seen in sarcoidosis patients (Thompson, Rybicki et al. 2006).

About a quarter of sarcoidosis patients develop pulmonary fibrosis. The pathogenesis of this is not fully understood but it has been postulated that central to the pathogenesis is both the presence of matrix metalloproteinases (particularly MMP 8 and MMP 9), and possibly a shift from a predominant Th1 cytokine production to a more Th2 like environment (IL-4, IL-10, IL-13) (Henry, McMahon et al. 2002; Iannuzzi, Rybicki et al. 2007). However there is little published data on the underlying immunopathogenesis of pulmonary fibrosis related to sarcoidosis, therefore these hypotheses are not well supported by evidence.

In summary, the immune mechanisms that cause sarcoidosis are not well understood but it is proposed to begin with an antigenic stimulus, followed by macrophage and T-cell activation via a MHC Class II mediated pathway, which results in a milieu of Th1 and other cytokines.

Genetic susceptibility towards sarcoidosis

There is a strong link between sarcoidosis and genetic susceptibility based on both a tendency for familial clustering of the disease and an increased risk of sarcoidosis in family members (Rybicki, Iannuzzi et al. 2001). A number of candidate gene association studies and genomewide association studies have identified important genetic associations with sarcoidosis. These have predominantly examined HLA Class I and Class II genes linking both to disease susceptibility and prognosis (Grunewald 2010). Studies of non-HLA candidate genes including vitamin D have reported conflicting results (Iannuzzi and Rybicki 2007). The link between sarcoidosis and vitamin D is based on both the immunomodulatory effects of vitamin D, and suggestive epidemiological evidence that sarcoidosis is associated with vitamin D deficiency such as an increased incidence in spring, increased incidence in African Americans, increased prevalence in northern latitudes and lower prevalence nearer the equator (Gerke and Hunninghake 2008).

Interestingly in a cohort of over 100 Japanese patients, from a panel of 10 candidate genes including IFNγ and its receptors, an IFNα haplotype was found to be associated with susceptibility to sarcoidosis but not tuberculosis (Akahoshi, Ishihara et al. 2004). The authors also show the IFNα allele is associated with a higher IFNα production after in vitro stimulation by Sendai virus, therefore suggesting an increased endogenous production of IFNα may have predisposed these patients towards sarcoidosis.

Animal models of sarcoidosis

One of the major barriers to studying sarcoidosis is the lack of an accepted animal model. Early studies in the 1970s injected mice with sarcoidosis tissue homogenates, however results were inconsistent (Belcher and Reid 1975; Mitchell, Rees et al. 1976). More recently two mouse models have been proposed but due to the unknown aetiology and limited resemblance of human pathology these experimental models have not been widely accepted (Samokhin, Buhling et al. 2010; Swaisgood, Oswald-Richter et al. 2011).

Treatment of sarcoidosis

There are many challenges in the clinical decisions surrounding treatment. These difficulties are exacerbated by the lack of well-conducted trials and therefore reliance on only expert-driven evidence based guidelines.

While the treatment of TB is the commencement of antibacterial therapy for at least 6 months, the mainstay of sarcoidosis treatment is aimed at suppressing the inflammatory response, with the aim of reducing the burden of granulomas and preventing the development of fibrosis. Spontaneous remissions occur in 55- 90% of patients with Stage I disease, 40-70% of those with Stage II disease, 10-20% with Stage III disease, and 0% with Stage IV disease (WASOG 1999). Although spontaneous resolution is common, progressive lung disease occurs in approximately 25% of all cases and disabling organ failure in up to 10% of patients (Baughman 2004). Oral glucocorticoids are the first line of therapy and predominantly instituted due to their anti-inflammatory properties. They are thought to be capable of attenuating the granulomatous inflammation and slowing the development of fibrosis. However the challenge remains for many cases in deciding whether systemic treatment is appropriate in view of the known serious side effects. This difficulty is perhaps reflected by the wide range of patients, 20 - 70% across different studies, that physicians decide to start on systemic treatment (Baughman and Nunes 2012).

A Cochrane review (a systematic review of the highest standard in evidence based medicine) of oral glucocorticoids suggested they improve chest radiograph and respiratory symptoms (Paramothayan, Lasserson et al. 2005). However there was little evidence of an improvement in lung function, and limited data beyond two years to indicate whether they have any modifying effect on long-term pulmonary disease progression (Paramothayan, Lasserson et al. 2005). Unfortunately there is a lack of adequate trials in the treatment of sarcoidosis. There have been only 5 randomised controlled trials that satisfied the Cochrane systematic review criteria and moreover the trials were not adequate to distinguish between patients with differing radiological stages – the most commonly used clinical classification criteria. Immunological indications for the effect of glucocorticoids include a reversal of the elevated CD4:CD8 ratio (typically greater than 4:1) seen in the bronchoalveolar lavage (Winterbauer, Lammert et al. 1993). In addition after commencement of glucocorticoid treatment TNF levels also changed prognostically (Moodley, Dorasamy et al. 2000).

Both International and British guidelines advocate starting treatment for symptomatic patients, regardless of other clinical findings (WASOG 1999; Bradley, Branley et al. 2008). In asymptomatic patients with stage 0 or I radiological disease there is strong evidence that no treatment is required (Gibson, Prescott et al. 1996). Indeed 95% of patients with stage I disease will have a normal chest radiograph within 10 years (Nagai, Shigematsu et al. 1999). For asymptomatic patients with stable stage II-IV disease the current British guidelines recommend observation for development of symptoms, deteriorating radiological changes and deteriorating lung function (Bradley, Branley et al. 2008). Glucocorticoid dosing typically involves several or all of the following phases: (1) initial high doses to control inflammation; (2) tapering to a maintenance dose to lessen the risk of side effects (during this time steroid-sparing drugs maybe started); (3) continuing the maintenance dose for 6 - 24 months; (4) tapering the dose for complete steroid withdrawal. Data on long-term benefits, continuing oral glucocorticoids for longer than 2 years, still remains unclear. However many sarcoidosis specialists suggest that in some patients treatment should be continued to prevent relapses (Coker 2007). This view is apparent as over half of those started on systemic treatment continue treatment for more than 2 years (Baughman and Nunes 2012).

Some patients with pulmonary sarcoidosis cannot tolerate or do not respond to glucocorticoids (Paramothayan, Lasserson et al. 2006). Several alternative approaches have been introduced, such as the use of cytotoxic drugs e.g. methotrexate and azathioprine; however the efficacy of these therapies is restricted and each are associated with different toxicities (Paramothayan, Lasserson et al. 2006). Therapy to block TNF can be useful in refractory chronic sarcoidosis (Baughman, Drent et al. 2006; Rossman, Newman et al. 2006). Paradoxically, there are several reports in the medical literature describing the development of sarcoidosis in patients treated with TNF-alpha inhibitors for other diseases (Daien, Monnier et al. 2009; Clementine, Lyman et al. 2010). The antimalarial and anti-inflammatory drug hydroxychloroquine is often used as a second-line agent particularly in cutaneous sarcoidosis in part due to its relatively low side-effect profile.

In summary there clearly remains a significant need for improvements in the treatment of sarcoidosis. However without a better understanding of the disease process itself, better tools for stratifying patients who would benefit from treatment and an improved ability to monitor a patient's response, it is inevitable that improving sarcoidosis treatment is a complex and hard to achieve goal.

Brief summary on pneumonia and lung cancer

Pneumonia is typically an acute bacterial infection of the lungs therefore treatment is antibiotics and supportive management. The treatment for primary lung cancer, typically secondary to cigarette smoking, depends on the extent of the disease.

Community acquired pneumonia

Community acquired pneumonia is distinguished from hospital acquired pneumonia, which is an important distinction due to the likelihood of differing causal pathogens. Community acquired pneumonia is a common and potentially severe illness, that can be associated with substantial morbidity and mortality in adults, with up to 14% mortality in the UK which particularly effects the elderly (Lim, Baudouin et al. 2009). The most common cause of community acquired pneumonia worldwide is *Streptococcus pneumonia*. In the UK the top five bacterial causes of adult community acquire pneumonia are *Streptococcus pneumonia*, *Haemophilus influenza*, *Mycoplasma pneumonia*, *Chlamydophila psittaci* and gram-negative enteric bacilli; in addition viruses such as influenza A and B are also frequent aetiological agents (Lim, Baudouin et al. 2009). Although often the predisposing risk factor is not apparent, there are several variables that have been correlated with an increased risk, these include smoking, previous pneumonia, chronic lung disease and treated diabetes (Almirall, Bolibar et al. 1999).

The clinical definition of community acquired pneumonia can vary between studies and can depend on the setting in which the patient's diagnosis is made e.g. in a community setting with no radiology and microbiology facilities compared to a hospital setting with the latest available clinical tools. For this study the definition of pneumonia as defined in the British Thoracic Society (BTS) guidelines was applied for patients admitted to the hospital when a chest radiograph is available (Lim, Baudouin et al. 2009). In this scenario community acquired pneumonia is defined as the presence of symptoms and signs consistent with an acute lower respiratory tract infection e.g. cough, fever and new radiographic shadowing consistent with infection, and for which there is no other likely cause. In addition the pneumonia must be the prevailing reason for the hospital admission.

All patients on admission to hospital should be assessed to both confirm the diagnosis and to assess the severity of illness. Alongside clinical judgement the guidelines recommend using a severity score as pneumonia can cause a wide spectrum of disease from mild pneumonia that can be treated in the community to life-threatening and sometimes fatal disease (Lim, Baudouin et al. 2009). Severity scores help decide the patient's management and often can predict the likely prognosis. The Pneumonia Severity Index tool is a validated severity scoring tool but its practical use is limited due to the requirement of up to 20 measurable variables (Fine, Auble et al. 1997). CURB65 is a simpler tool that is widely accepted and only requires 5 easy to measure variables (Lim, van der Eerden et al. 2003). Treatment should then be guided by both clinical judgement and the severity score (Lim, Baudouin et al. 2009). Patients with low severity scores can potentially be treated in the community with oral antibiotics, but hospital admission and intravenous antibiotics should be considered for those with a moderate severity score and emergency hospital care with intravenous antibiotics should always be arranged for those with a high severity score (Lim, van der Eerden et al. 2003).

Lung cancer

The World Health Organisation classifies primary lung cancer broadly into four major histological types; three non-small cell carcinomas (NSCLC): adenocarcinoma, squamous cell carcinoma and large cell carcinoma; and small cell carcinoma (Travis, Brambilla et al. 2004). NSCLCs make up the vast majority of primary lung cancers. Cigarette smoking was shown to be a risk factor for developing lung cancer by Sir Richard Doll in 1950 and now is thought to account for about 90% of all lung cancers (Doll and Hill 1950; Dubey and Powell 2009). Cancer is associated with an inflammatory response but it is unclear whether the inflammation is the provocation or the consequence, for example there is increasing evidence that smoking encourages inflammation which leads to the development of lung cancer (O'Callaghan, O'Donnell et al. 2010). The treatment for NSCLCs is directed by the stage of the tumour and the patient’s ability to perform activities of daily living and their lung function status. Patients with early-stage NSCLC may be offered surgery with curative intent, later stage NSCLC cannot be cured by surgery therefore patients are only offered chemotherapy and/or radiotherapy (Lim, Baldwin et al. 2010; NICE 2011). In addition targeted agents such as epidermal growth factor receptor (EGFR) inhibitors or inhibitors of a kinase fusion oncogene can be suitable for patients with particular molecular and histological cancer features (Dienstmann, Martinez et al. 2011). Small cell carcinoma is treated with chemotherapy and/or radiotherapy (Lim, Baldwin et al. 2010; NICE 2011).

Gene expression profiling

Genomic signatures can serve as surrogates of clinical phenotypes. Integration of this information can provide new biological knowledge.

Background

For over a decade now gene expression profiling has been applied to human disease to improve both our comprehension and classification of the underlying molecular processes. Classification of samples (class discovery) is one of the most common uses of microarray (Stekel 2003). Microarray analysis can identify genes or groups of genes associated with a disease phenotype. The development of these genes into biomarkers can subsequently be used to facilitate diagnosis or prognosis relating to the natural history of the disease or after administration of therapy. This is most successfully established in the study of cancers. A landmark study was the use of microarray in 1999 to distinguish between the diseases acute myeloid leukaemia and acute lymphoblastic leukaemia (Golub, Slonim et al. 1999). This study shaped the way forward for a methodology that enables the discovery and prediction of disease classes independent of previous biological knowledge. More recently in breast cancer large completed and on-going phase III clinical trials show very promising results in using microarray to accurately predict prognosis and effectively direct treatment (van 't Veer, Dai et al. 2002; Bonnefoi, Underhill et al. 2009). Furthermore comparing gene expression (class comparison) of samples with different disease phenotypes can help elucidate potential biological functions for related or different gene expression patterns i.e. genes with similar expression patterns might be functionally related or working in the same pathway as co-expressed genes (Chaussabel, Quinn et al. 2008).

Microarray technology was introduced in the mid-1990s and has enabled expression analysis of thousands of genes at one time, enabling visualisation of complex gene expression patterns and perturbations of those patterns (Schena, Shalon et al. 1995). With the introduction of high-density platforms capable of incorporating tens of thousands of sequences and the sequencing work of the Human Genome Project, microarray technology can carry out measurements on every identified gene in the human genome. Microarray platforms are mainly produced by five manufacturers: Affymetrix, Applied Biosystems, Agilent, GE Healthcare, and Illumina. Most published microarray publications use either Affymetrix GeneChips or Illumina Sentrix BeadArrays, but comparisons between the platforms in fact indicate reasonably high agreement (Barnes, Freudenberg et al. 2005; Cheadle, Becker et al. 2007). Illumina technology, used in this study, has built arrays using the random self-assembly of microspheres (beads) on which 50-mer oligonucleotide probes are immobilized, onto a planar silica substrate (Illumina 2005). The beads spontaneously assemble, held by Van der Waals forces and hydrostatic interactions, within the walls of ordered microwells sketched into the silica substrate. This creates one of the highest density array platforms commercially available and was based on technological advances from the semiconductor manufacturing industry to build the millions of wells in highly ordered patterns (Illumina 2005). Affymetrix microarrays are also built on the same type of manufacturing technology use to build semiconductors. However unlike the spotted design of Illumina arrays Affymetrix synthesise their 25-mer oligonucleotides *in situ* using photolithographic synthesis to build the sequences across the silica substrate of the array (Barnes, Freudenberg et al. 2005).

The Illumina Sentrix BeadChip array used in this study targets more than 48,000 probes, which are derived primarily from the National Center for Biotechnology Information Reference Sequence (Illumina 2005). Because each probe contains hundreds of thousands of copies of the covalently attached oligonucleotide sequences, this generates an average 30-fold redundancy for each randomly generated sequence represented on the array. This random generation for the Illumina BeadChip therefore permits checks on quality control of technical replication that is not possible with the Affymetrix GeneChip (Barnes, Freudenberg et al. 2005). In the Affymetrix GeneChip multiple probes are assembled for each gene together with a control probe which has a one-base mismatch designed to allow detection of the background non-specific hybridisation (Barnes, Freudenberg et al. 2005). The Illumina probes were designed using a multi-step algorithm scoring of multiple parameters including: similarity to other genes, expressed sequence tag coverage, absence of highly repeated sequence in the genome, distance from 3' end of the transcript (Illumina 2005). In addition Illumina position negative control beads for each set of analytical probes to warrant measurement and thus subtraction of background non-specific hybridisation intensity.

The reliability of microarray technology to detect transcriptional differences representative of the original samples is also affected by the quality of the extracted RNA (Kim, Dix et al. 2007). Therefore the RNA integrity number (RIN) must satisfy a specific high standard to ensure accurate interpretation of transcriptional expression.

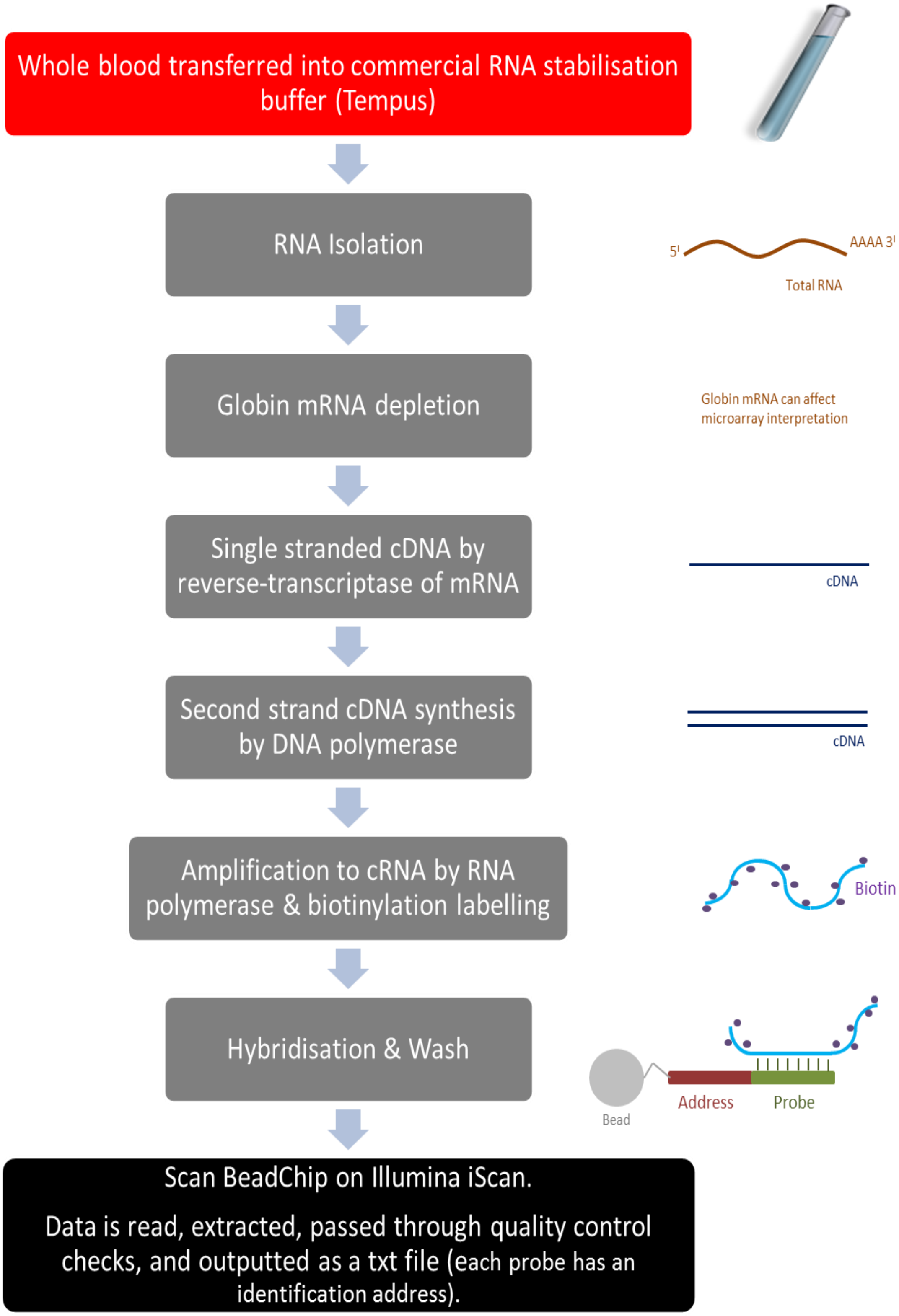

Whole blood transferred into commercial RNA stabilisation buffer (Tempus)
RNA Isolation
5' AAAA 3'
Total RNA
Globin mRNA depletion
Globin mRNA can affect microarray interpretation
Single stranded cDNA by reverse-transcriptase of mRNA
cDNA
Second strand cDNA synthesis by DNA polymerase
cDNA
Amplification to cRNA by RNA polymerase & biotinylation labelling
Biotin
Hybridisation & Wash
Bead
Address
Probe
Scan BeadChip on Illumina iScan.
Data is read, extracted, passed through quality control checks, and outputted as a txt file (each probe has an identification address).

Figure 1. Microarray processing steps from blood collection to data output.

Adapted from the Illumina protocol.

Microarray data analysis

Initially a significant problem with microarray data was technical reproducibility but with advances in technology this has been overcome. A major concern now lies in reproducibility of data interpretation, particularly in terms of the validity of the statistical analysis. The first step of analysis is to apply quality control checks on the raw data, to ensure the labelling, hybridisation and scanning occurred as planned (Figure 1). For example in Illumina Sentrix BeadChips control oligonucleotides are spiked into the hybridization solution such that performance of the controls can be checked, and specific housekeeping genes are compared to the background intensity values (Illumina 2005).

Microarray is not a quantitative tool, rather it measures changes in mRNA intensity values relative to a reference group therefore control samples must be included in the experiment (Chaussabel, Pascual et al. 2010). Results are often described as fold change in intensity levels because different genes are expressed at different levels and genes with the highest expression, often the ‘housekeeping’ genes, may not be the most relevant genes for that experiment (Ness 2006). Another advantage of fold change measurements is their ability to accentuate the changes in gene expression unlike total abundance of individual transcripts. The measurement of total abundance of transcripts are better measured using RNA sequencing rather than microarray (Pascual, Chaussabel et al. 2010). Fold change measurements are also comparable across different experiments and platforms because they are not influenced by the differences in raw values secondary to technical variations.

Microarray analysis of human samples

As mentioned above a common application of microarray in human studies is its use for class discovery (Stekel 2003). This involves grouping of samples with homogenous/similar expression profiles that may represent a particular disease, subgroup of disease, response to treatment or other clinical groupings the experiment is set out to discover (Peters 2008). This approach therefore requires the comparison of samples from different cohorts, for example patients with different diseases and a cohort of healthy controls. The study should be set up to compensate for the differences in inter-individual gene expression patterns across the cohorts, due to demographics such as age, ethnicity and gender (Whitney, Diehn et al. 2003; Eady, Wortley et al. 2005). This is because although determining biological variation in gene expression is the goal of microarray analysis, unwanted biological variation can mask the question being asked.

To determine if a transcriptional signature/profile, can differentiate between different groups of samples using an unbiased method requires an 'unsupervised' analytical approach (Berry, Graham et al. 2010; Pascual, Chaussabel et al. 2010). For an unsupervised analysis the gene expression profiles of all the samples are analysed blindly i.e. without *a priori* knowledge of clinical or demographic information. The first steps of the unsupervised analysis are the same as for any analysis protocol. Initially any transcripts that are not detected significantly above the background, as defined by negative control probes in an Illumina Sentrix BeadChip, must be removed (Illumina 2005). Typically more than half of the probes present on a microarray do not detect a signal for any of the samples in a given analysis (Pascual, Chaussabel et al. 2010). For Illumina probes a threshold level is set such that transcripts with low values e.g. less than 1 are given a value of 1, this is to compensate for the poor specificity of these low intensity values, otherwise in a fold change analysis these low level transcripts could appear far too significant. The next step is to transform the data using a logarithm to base 2 transformation (GeneSpring 2010). All microarray data is logarithmically transformed to reduce the effect of skewed data and allow a more equal influence of either very high or very low expression values (Peters 2008).The next step in the analysis is to normalise the intensity value of each gene to a specific value to minimize systematic non-biological differences and reveal true biological differences; sources of technical variation include quantities of RNA, quality of RNA, differences in hybridization between chips and differences between manufactured chips (Agilent 2010). This also creates an emphasis on the relative change in gene expression between samples rather than absolute values. For unsupervised analysis each gene is normalised to the median of that individual gene across all the samples (Berry, Graham et al. 2010). The median is selected, instead of the mean of a particular group such as the controls, to ensure the analysis is performed blindly i.e. unsupervised. The median value can be chosen rather than the mean to remove any assumptions that the gene expression values are normally distributed. This assumption is unlikely in human samples but for example in cell lines could be true.

Most microarray experiments are performed with the objective that only genes that change significantly between the cohorts will be relevant therefore genes that don't change in expression by a set amount (for example 2 fold up or down) from the median are discarded from the analysis (Pascual, Chaussabel et al. 2010). For an unsupervised analysis approach the median is chosen so that the identity of the samples remains blinded (i.e. the samples are not yet defined by any phenotype). In addition a level of acceptable number of samples to satisfy the fold change filter must be set. For example you could choose just one sample or 10% of the sample’s transcripts to fulfil the two-fold change around the median. Increasing the number can remove noise but equally may mistakenly remove genes of interest. If a 30% cut-off was set in a group of 10 samples of which 2 were of a different disease this filter could potentially exclude some relevant transcripts within that whole disease group. A caveat of this unsupervised approach is that it relies on the gene expression of each transcript to be fairly equally distributed across all the samples. For example if the study contains 30 samples and it is likely 20 of them will have many highly expressed genes and the other 10 will have much lower expression, then the median for most transcripts will lie within the 20 highly expressed samples. Therefore the type of samples chosen for the experiment requires thought regarding the study design prior to analysis.

The next step of the unsupervised analysis is to cluster the genes by similarity in their intensity values. To perform this there is a choice of clustering methods and various distance metrics. Distance metrics are methods of calculating the distance between the genes or samples. For example two commonly used metrics are Pearson’s correlation distance metric where the distance measured is influenced by the gene expression trend and Euclidean correlation distance metric where the distance measured is more influenced by gene expression magnitude (Quackenbush 2001; Agilent 2010). In GeneSpring 11, used in this study, there are three choices of Pearson’s correlation: absolute, uncentered and centered. For this study the uncentered method was chosen as this method separately recognises negative and positive correlations (positively correlated transcripts give values close to +1, negatively correlated close to -1 and unrelated close to 0). In addition a linkage criterion must be set; this specifies how the distance between the growing clusters of genes/samples are measured (Quackenbush 2001). In this study the average linkage was chosen as this calculates the average distance between members of two clusters (GeneSpring 2010). Other linkage options include complete (greatest distance between members), single (minimum distance between members), wards (distance based on the sum of squared errors around the mean). The average linkage is the most commonly used. Clustering of genes is typically performed by either hierarchical clustering or *k*-means clustering. Hierarchical clustering can be agglomerative, a bottom up approach that builds on the first two similar observations, or divisive, a top down approach that starts with all observations together and splits them iteratively (Peters 2008). This approach continues until a dendrogram (tree structure) of similar genes and dendrogram of similar samples can be mapped. Over-abundant genes are typically coloured as red and under-abundant blue, with no change in expression as yellow. A colour scale indicates the relative degree of normalised expression. The dendrogram allows visualisation of the most similar samples as sitting next to each other and the least similar as furthest away from each other. The height of the dendrogram branches is a specific distance measurement, calculated from the distance metric and linkage criteria, that correlates with the similarity of the genes/samples by their gene expression (Peters 2008).

Another common form of clustering is *k*-means clustering where observations are allocated into a fixed number of (*k*) clusters (Quackenbush 2001). Each cluster contains genes that are similar by application of a distance metric. To achieve this an average expression vector is set for each cluster, then using an iterative method genes are moved between clusters increasing the inter-cluster distances and decreasing the intra-cluster distances (Quackenbush 2001). A challenge with this form of clustering is deciding how many clusters are appropriate for the dataset and the number of iterations required.

An additional stage of statistical filtering is often added after the unsupervised analysis to improve the specificity of the generated transcript list. This part of the analysis is therefore now supervised as the statistical test can only be carried out knowing which samples it is comparing. In human samples as they are unlikely to follow a normal distribution it is difficult to make this assumption therefore non-parametric statistical tests may be more appropriate.

For supervised analysis used for class comparison some of the analysis steps are different. Firstly during the per-gene normalisation, the transcripts can be normalised to either the mean or the median of the controls. Secondly the fold change filter can be applied to compare one group to another e.g. two-fold change of the disease group to the mean of the control group for each transcript. The steps prior and after this do not depend on whether the analysis is unsupervised or supervised. Clustering can also be supervised; *k*-means is a form of supervised clustering as the samples are all assigned a phenotype for the algorithm; hierarchical clustering can be partly supervised by only performing unsupervised hierarchical clustering of the transcripts but not of the samples. The choice of the method applied depends on the question the analysis is trying to answer e.g. unsupervised analysis is suitable for class discovery while supervised analysis is more suitable for class comparison. For this study both approaches have been used: unsupervised analysis for the unbiased discovery of disease or treatment associations and supervised analysis for the identification of specific sets of genes differentially expressed between known groups.

The ideal data for statistical analysis has relatively few variables and many replicates, however in microarray experiments there are thousands of variables (transcripts) and often few replicates (samples) each with their own variables. Consequently usual statistical methods alone may have trouble dealing with the data (Ness 2006). Hence the application of multiple testing correction is vital as this controls for the huge number of statistical tests carried out for each experiment (Olson 2006; Dupuy and Simon 2007). For example if the standard *p*-value of p <0.05 is used this will allow 5% of genes to pass through by chance, which when analysing 40,000 transcripts could contribute to false identification of up to 1200 transcripts. Therefore the *p*-value is adjusted based on the number of tests performed, thus reducing the Type I error rate. Two commonly applied corrections are false discovery rate (FDR) e.g. Benjamini Hochberg, which controls for the expected frequency of false positives, and family wise error rate e.g. Bonferroni, which corrects for the chance of at least one false positive (therefore more stringent). A different statistical approach commonly used for expression analysis is called significance analysis of microarrays (SAM). This test was purely developed as traditional methods of multiple testing corrections were often too stringent for microarray data (Draghici 2012). The basic statistic used in SAM is based on the *t*-test but the analysis actually uses non-parametric statistics. Each gene is given a score based on the ‘relative difference’; this score includes the change in gene expression between conditions and the standard deviation of the change (Draghici 2012). A ‘fudge factor’ ensures the coefficient of variation is minimised such that genes with low expression levels are still considered. A permutation test is then used to assess the significance of each score and to estimate the false discovery rate. This is produced by calculating the ‘expected relative difference’ derived from controls generated by permutations of data (Draghici 2012). SAM is therefore more robust for the analysis of genes with low expression.

Another way of dealing with the large number of variables is to validate initial results. Validation should ideally occur in an independent cohort processed as an independent microarray experiment used exclusively for evaluating the original outcome (Olson 2006; Dupuy and Simon 2007). The most extensively applied and acknowledged method for validating results from a classification experiment is to use ‘training and test sets’; cross-validation is also an alternative and accepted method (Stekel 2003)**.** ‘Training sets’ are used to train the algorithm, where the algorithm is optimised to classify the training set data as best it can. The algorithm is then trialled on the ‘test set’ for independent confirmation of the success of the algorithm. As the test set is validating the training set findings it should satisfy the same experimental conditions as used in the training set.

In many circumstances it may be desired to test a set of genes for their ability to accurately predict the class membership (e.g. disease type) of a collection of samples. This involves methods of ‘class prediction’, where for example the disease-type of each sample is already known but we want to build a classifier (Stekel 2003). The machine learned algorithm support vector machines (SVM) is a frequently applied technique. The prediction model is built using the transcriptional signature from samples with known disease-types to predict the classification of a new collection of samples. The SVM algorithm maps samples in a theoretical n-dimensional space and tries to determine a best fit hyperplane that can separate the samples into the different disease-types (Draghici 2012). To prevent the model overfitting the predictive signature, the prediction model is then applied to the new collection of samples, such that the new samples are classified according to which side of the hyperplane they fall into. An advantage of SVM is that it can deal with samples that are intertwined as it transforms samples before setting the planes, however a disadvantage is that it requires the setting of numerous parameters to build the predictor (GeneSpring 2010).

Limitations of microarray analysis

A theoretical limitation of gene expression profiling is its assumption that changes in gene expression predict biological significance. Microarray analysis is typically carried out to only select genes with at least a 1.5 fold change between groups, but results have been published using a variety of different fold change filters including fold change of

1.25 (Nakaya, Wrammert et al. 2011). The filter intentionally excludes genes with low level expression differences that could play critical roles in the host response. Therefore it may be more appropriate to apply different fold change cut-offs to different genes. On the other hand some of the genes with large expression differences might not be associated with such a key role in the host response as we suppose. For example post-transcriptional or post-translational regulation could negate any correlation between mRNA levels and protein activity. In addition although the double filter, fold change and statistics, are intuitive and widely used there are inconsistencies between the two procedures; fold change assumes all genes share a common variance but the *t*-test assumes gene-specific variance which are opposing assumptions (Zhang and Cao 2009). Therefore it could perhaps be advocated that only a microarray-apposite statistical filter should be applied rather than the double filter (Zhang and Cao 2009). However the enormous wealth of microarray publications that have used relative changes in mRNA expression values to identify new disease mechanisms, therapeutic targets and biomarkers clearly demonstrates the advantage and robustness of the double filtering strategy. In addition the dependence of the *t*-test on variance has been shown to be the likely cause for disagreements of cross-platform assessments which run comparisons using *p*-values in contrast to the higher agreement found when comparisons performed using fold change analysis (Wilder, Kaisaki et al. 2009). In reality it is likely that different analysis approaches are suitable for different datasets and for answering different questions.

Limitations in microarray techniques have been well documented and although may occur only during the processing can consequently affect the analysis. Due to their nature microarrays are noisy, so that even if the identical experiment is carried out twice, after just the scanning and image processing many probes will be reported as different intensity values (Draghici 2012). Potential sources of errors include mRNA preparation, RNA processing, hybridisation, scanning and quantification of the pixels from the image (Draghici 2012). Many of these sources of error, particularly technical errors, have been greatly reduced over the last 10 years. However a major challenge in the use of microarray technology remains in the ability to determine whether differences in the data are technical or biological. Replication of findings is a critical part of the armament to defend findings obtained from microarray experiments (Stekel 2003; Olson 2006; Dupuy and Simon 2007; Draghici 2012). Genomics is not the only area that has problems with technical errors as many fields that work with massive amounts of data also suffer from the same difficulties, this can be partly dealt with by ensuring properly designed experiments and implication of standards (Nature, editorial 2012).

Our analysis process was to apply a logical and systematic approach to a sound experimental design, using appropriate analysis steps at each stage, thus hopefully considerably alleviating many potential sources of error. In addition we corroborated our findings in at least two independent cohorts of patients at each stage.

Application to clinical immunology

There are estimated to be around 23,000 genes in the human genomes, and entry of an antigen or pathogen into the body alters the expression of a substantial fraction of them as identified in multiple transcriptomic studies of inflammatory and infectious diseases (Bennett, Palucka et al. 2003; Griffiths, Shafi et al. 2005; Ramilo, Allman et al. 2007; Emamian, Leon et al. 2009; Nascimento, Braga-Neto et al. 2009). Well defined genomic signatures give us new insights into the complexity of the immune response, along with potential improvements in biomarkers to diagnose clinical disease phenotypes or as predictors of outcome (Haining and Wherry 2010). For example in the autoimmune diseases SLE and systemic onset juvenile idiopathic arthritis, microarray technology has led not only to the identification of pathogenic pathways, potential diagnostic and prognostic biomarkers, but also to new therapeutic strategies (Bennett, Palucka et al. 2003) (Allantaz, Chaussabel et al. 2007; Pascual, Chaussabel et al.). In infectious diseases gene profiling has been able to discriminate between clinical forms of disease from the same pathogen, without a priori clinical information, and led to powerful insights into the regulation of the host response (Bleharski, Li et al. 2003; Berry, Graham et al. 2010; Tattermusch, Skinner et al. 2012).

Furthermore a genomic approach was able to identify immune correlates of protection for the highly effective yellow fever vaccine (Querec, Akondy et al. 2009). Prior to this study there was little global knowledge of the immune mechanisms inducing such an effective response; the distinct transcriptional signature that was found not only revealed a global picture of the immune response but this approach could also give an insight into understanding vaccine non-responders (Querec, Akondy et al. 2009).

Transcriptional profiling of peripheral blood

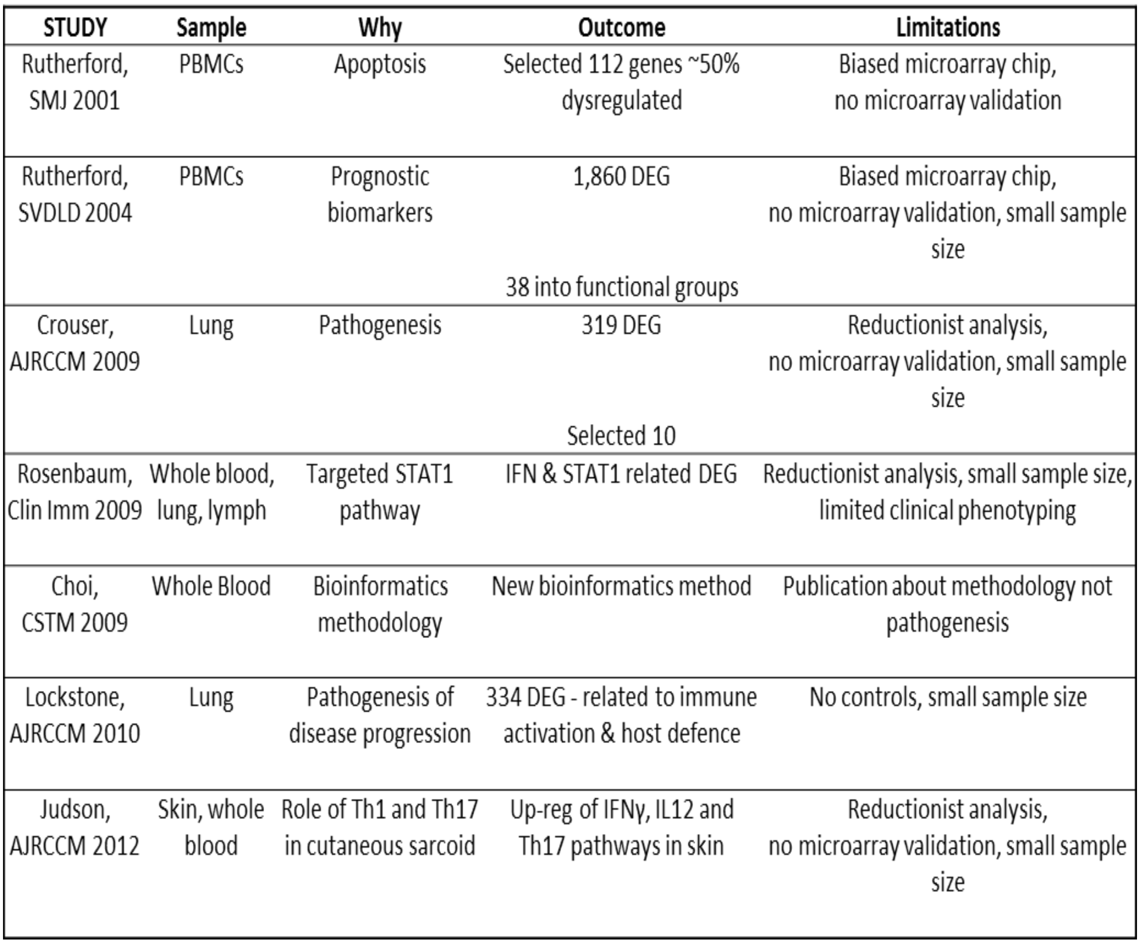
Peripheral blood has the capacity to reflect pathological and immunological changes in the body, and identification of disease associated alterations can be determined by a blood transcriptional signature (Mohr and Liew 2007). Because blood interacts with every organ and tissue in the body it is an effective means for approaching the complexity of systems biology. In the past most studies used peripheral blood mononuclear cells (PMBCs) however it is now recognised that microarray can be effectively performed using whole blood, thus including neutrophils an important cellular source when looking at inflammatory and infectious diseases. Demonstrating this concept a neutrophil driven interferon (IFN)-inducible signature was revealed by the whole blood microarray approach designed to fully characterise the immune response to *M. tuberculosis* carried out recently by O'Garra and collaborators (Berry, Graham et al. 2010).

Transcriptional profiling in the study of pulmonary TB

The aforementioned comprehensive unbiased study of TB patients, established through unsupervised data mining, a robust transcriptional signature for active TB, in individuals from both intermediate and high burden countries (Berry, Graham et al. 2010). 42 samples (13 active, 17 latent, 12 controls) were included in the initial cohort (training set) from which a distinct 393 transcript signature was revealed to be associated with active TB patients. This signature was validated in both a UK test set (21 active, 21 latent, 12 controls) and a South Africa validation set (20 active, 31 latent). Using the 'weighted molecular distance to health', a distance metric algorithm that links gene expression changes to a chosen clinical classification of disease severity (Pankla, Buddhisa et al. 2009), it was shown that the signature correlated with lung radiographic extent of disease and was diminished with antituberculous treatment. They then compared the microarray data from their cohorts to microarray data from other disease cohorts, these cohorts were not recruited for their study but the expression data was acquired by the same microarray processing (arrays and facilities). To compare all the data from the different cohorts they used a statistical approach called 'analysis of significance' (Chaussabel, Allman et al. 2005). This approach compares the study disease to its own controls to obtain *p*-values for relatively over/under expressed genes, and then repeats this for the other diseases, a list of genes for the study disease is obtained by selecting only those genes significant for the study disease and not the other diseases (Chaussabel, Allman et al. 2005; Allantaz, Chaussabel et al. 2007). Applying this analytical approach Berry *et al.* 2010, additionally obtained an 86 transcript signature that discriminated active TB from patients with SLE, Still's disease, group A *Streptococcus* infection and *Staphylococcus* infection. This 86 transcript signature also diminished after successful antituberculous treatment. Discrimination from other diseases was also demonstrated by the use of a modular data-mining strategy (Chaussabel, Quinn et al. 2008). The modules are genes determined from gene expression profiles from cohorts of 8 different diseases. The modules of similarly expressed genes were extracted from all 8 cohorts by a complex algorithm involving *k*-means clustering to identify the sets of genes, after which the genes were functionally annotated by unbiased literature data mining (Chaussabel, Quinn et al. 2008). The modular approach works under the premise that co-expressed genes are likely to be co-regulated and contribute to a common biological function. Importantly the modules are data-driven sets of genes rather than sets of genes thought to be functionally related from mining published literature. Using the modular analysis and computer software Ingenuity Pathway Analysis (IPA), Berry *et al.* 2010, were able to show the type I and IFN-γ signalling molecules were dominant in the transcriptional signature of active TB. The active TB 393-transcript signature also may have exposed those latent individuals who will develop active TB as 10-20% of the latent TB samples displayed a similar signature to the active TB samples. This percentage is comparable to the expected frequency of progression from latent TB to active TB (Berry, Graham et al. 2010). A longitudinal transcriptional profiling study of latent TB patients is currently being planned to identify prognostic biomarkers and investigate the underlying heterogeneity of latent TB. The Berry *et al.* 2010, study was the first whole genome unbiased comprehensive expression profiling study in human TB and evidently demonstrated the possible gains from microarray studies, such as new knowledge of immunopathogenesis and new potential credible diagnostic and prognostic biomarkers. Earlier blood expression profiling studies in TB patients did not show such comprehensive findings or feasible transcriptional signatures for active TB (Jacobsen, Repsilber et al. 2007; Mistry, Cliff et al. 2007; Maertzdorf, Repsilber et al. 2011). This was most likely due to a reductionist approach to the microarray analysis e.g. selecting only 5 genes as a diagnostic biomarker between active TB and latent TB (Maertzdorf, Repsilber et al. 2011), in addition to other limitations such as inadequate patient selection e.g. patients already on antituberculous treatment, small study numbers and the use of custom microarrays not containing the full human genome (Jacobsen, Repsilber et al. 2007; Mistry, Cliff et al. 2007).

Transcriptional profiling in the study of sarcoidosis

Initial studies in sarcoidosis only examined individual or small sets of genes. In 1996 a restricted mRNA differential display study looked at bronchoalveolar lavage (BAL) cells from 18 sarcoidosis patients and 8 patients with other disparate lung diseases (Wiwien, Hiyama et al. 1996). Three PCR products were consistently detected for sarcoidosis, including CD44 and TNF-α, however these genes are known not to be specific for sarcoidosis. Years later a broader study was performed using mRNA differential display to differentiate between granuloma-associated alteration of gene expression, by examining BAL of sarcoidosis, TB and healthy controls (Gaede, Mamat et al. 2004). Applying this unbiased approach Gaede *et al* were able to amplify 2,498 PCR products from the three cohorts. Analysis revealed a differential regulation of 6.5% of genes from both diseases compared to the controls and a concordance of 1.8% of genes between the diseases. Although a limited study the findings were encouraging for future unbiased gene expression studies in TB and sarcoidosis.



STUDY	Sample	Why	Outcome	Limitations
Rutherford, SMJ 2001	PBMCs	Apoptosis	Selected 112 genes ~50% dysregulated	Biased microarray chip, no microarray validation
Rutherford, SVDLD 2004	PBMCs	Prognostic biomarkers	1,860 DEG 38 into functional groups	Biased microarray chip, no microarray validation, small sample size
Crouser, AJRCCM 2009	Lung	Pathogenesis	319 DEG Selected 10	Reductionist analysis, no microarray validation, small sample size
Rosenbaum, Clin Imm 2009	Whole blood, lung, lymph	Targeted STAT1 pathway	IFN & STAT1 related DEG	Reductionist analysis, small sample size, limited clinical phenotyping
Choi, CSTM 2009	Whole Blood	Bioinformatics methodology	New bioinformatics method	Publication about methodology not pathogenesis
Lockstone, AJRCCM 2010	Lung	Pathogenesis of disease progression	334 DEG - related to immune activation & host defence	No controls, small sample size
Judson, AJRCCM 2012	Skin, whole blood	Role of Th1 and Th17 in cutaneous sarcoid	Up-reg of IFNγ, IL12 and Th17 pathways in skin	Reductionist analysis, no microarray validation, small sample size

Table 2. Microarray studies of sarcoidosis patients

DEG = differentially expressed genes.

Rutherford *et al* were the first to use the newer microarray techniques; however this study was carried out 10 years before the full annotation of the human genome, and used a custom microarray containing only 12,626 genes (Rutherford, Kehren et al. 2001) (Table 2). They compared PBMCs from 12 controls and 12 sarcoidosis patients before treatment, in 2 subgroups of progressive and self-limited disease. The results published were intentionally biased only examining 112 genes, to focus on the role of apoptosis. However they were unable to prove or disprove apoptosis related patterns, likely to be influenced by the small numbers of patients in each subgroup. A few years later they published further results from the same study, this time focussing on the antigen processing/presentation and T cell activation in respect to the outcome of sarcoidosis (Rutherford, Staedtler et al. 2004) (Table 2). They identified 729 differentially expressed genes between controls and patients. Known genes were grouped according to the functions they were interested in. This left them with examination of only 35 genes, of which solely 6 were associated with the disease subtypes progressive or self-limited. Unfortunately there was no validation of their findings although this study did illustrate the prospect of developing a discriminatory gene set from peripheral blood to help differentiate the different sarcoidosis phenotypes.

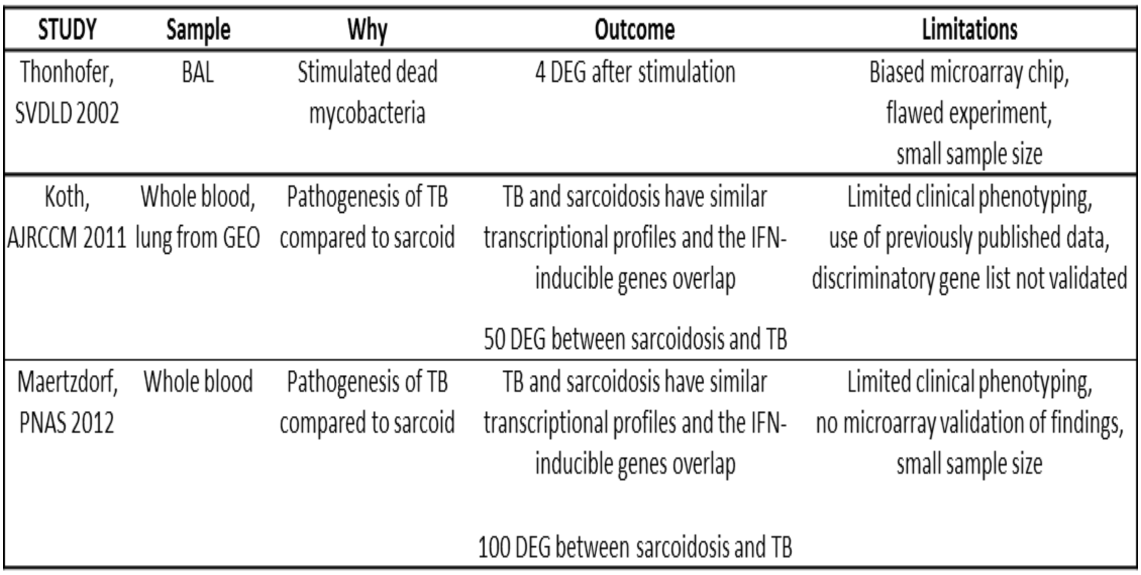
In this respect recent published findings, from our collaborators Ling-pei Ho's lab at the Weatherall Institute of Molecular Medicine, Oxford, compared lung biopsies from progressive-fibrotic pulmonary sarcoidosis to patients with self-limiting sarcoidosis, with the aim of defining discriminating genes to our improve understanding of the underlying pathogenesis (Lockstone, Sanderson et al. 2010) (Table 2). Samples of granulomatous tissue were taken during disease activation and before starting treatment. Patients were then clinically phenotyped 2 years later into either progressive-fibrotic (4 patients) or self-limiting (4 patients). 334 genes were differentially expressed using a whole human genome array; most of the genes were up-regulated in the progressive-fibrotic group. To help categorise the vast amount of information they used the literature-driven Gene Set Enrichment Analysis (GSEA). GSEA are pre-specified sets of genes grouped together by published data on their biological processes, chromosomal location or regulation (Subramanian, Tamayo et al. 2005). This analysis revealed enrichment in the progressive-fibrotic group for genes related to host immune activation including leukocyte differentiation/activation and cytokine production, and cell life including cell proliferation/cycle/apoptosis. The GSEA results were validated in a separate analysis of 7 patients. Limitations to this study were the lack of controls (disease-free lung biopsies) and the small number of patients.

Crouser *et al* also examined lung biopsies of untreated sarcoidosis patients by whole human genome microarray, this time comparing gene expression to healthy controls with the goal of providing insights into the pathogenesis (Crouser, Culver et al. 2009) (Table 2). However they took a reductionist approach and after identifying a few genes of interest (MMP-12 and ADAMDEC1) from an initial cohort of 6 participants per group they validated their findings by RT-PCR in a larger cohort. Their microarray analysis discovered 319 genes differentially expressed from the controls, the significantly over-represented genes were associated with a Th1 immune response. The study had several limitations including the small study size, varying clinical presentations, reductionist analysis approach and the lack of validation.

Judson *et al* examined the transcriptional profiles of cutaneous sarcoidosis using skin biopsies from well characterised sarcoidosis patients compared to unaffected skin biopsies from the same patient and skin biopsies from healthy controls (Judson, Marchell et al. 2012) (Table 2). Their study was focussed on examining Th1 and Th17 genes. They also compared their findings to whole blood profiles from the same patients. Unsupervised hierarchical clustering was able to clearly distinguish samples from the sarcoidosis-affected skin biopsies from the unaffected biopsies which were again distinguished from the healthy biopsies. Using IPA they found a significant association with the IFN-signalling pathway and differentially expressed genes in both the skin biopsies and whole blood. They also found over-abundance of activated macrophage proteins and inflammatory related proteins associated with the skin biopsies. The main focus of their paper was the discovery of an over-abundance of IL-23 and IL-21, validated in the same cohort by RT-PCR but not seen in the patient’s serum. This led to their conclusion of novel findings of potential activation of the Th17 pathway in cutaneous sarcoidosis. Unfortunately the study did not have a comparison group of patients with another cutaneous disease – in particular psoriasis, a skin disease with proven Th17 involvement (Papp, Leonardi et al. 2012), would have been of interest. Furthermore the significance of the Th17 pathway in relation to the other differentially expressed genes was not reported.

Rosenbaum *et al* used whole human genome microarray to address the hypothesis that sarcoidosis patients will have characteristic transcriptional profiles in whole blood and tissue (Rosenbaum, Pasadhika et al. 2009) (Table 2). However they then focussed their analysis solely on genes associated with STAT1. Their reasoning for this biased analytical approach was both due to the very high and consistent expression of the STAT1 related genes, with no comment on the statistical expression of other genes, and due to the critical role STAT1 plays in the inflammatory response (Rosenbaum, Pasadhika et al. 2009). Peripheral whole blood was taken from 12 untreated sarcoidosis patients, half of whom had uveitis secondary to sarcoidosis; and 12 healthy controls. They found 1039 over-abundant transcripts and 872 under-abundant transcripts differentiating sarcoidosis patients from the controls. This study was not set up to find discriminating genes between sarcoidosis phenotypes, and did not comment on this. However a heatmap of the over-abundant and under-abundant transcripts suggests most of the patients with uveitis could be visually distinguished from the controls. The hypothesis of the experiment was proven as several associated interferon and STAT-1 transcripts were significantly over-abundant in the peripheral blood and then validated in lung and lymph node samples by microarray. Microarray data from the same research group was published earlier in the same year from peripheral whole blood samples of patients with sarcoidosis compared to patients with ankylosing spondylitis (Table 2). It would appear to be the same sarcoidosis patients used in both studies. This study was carried out predominantly to assess a new bioinformatics method of clustering (Choi, Sharma et al. 2009).

Transcriptional profiling in the study of active TB and sarcoidosis



STUDY	Sample	Why	Outcome	Limitations
Thonhofer, SVDLD 2002	BAL	Stimulated dead mycobacteria	4 DEG after stimulation	Biased microarray chip, flawed experiment, small sample size
Koth, AJRCCM 2011	Whole blood, lung from GEO	Pathogenesis of TB compared to sarcoid	TB and sarcoidosis have similar transcriptional profiles and the IFN-inducible genes overlap 50 DEG between sarcoidosis and TB	Limited clinical phenotyping, use of previously published data, discriminatory gene list not validated
Maertzdorf, PNAS 2012	Whole blood	Pathogenesis of TB compared to sarcoid	TB and sarcoidosis have similar transcriptional profiles and the IFN-inducible genes overlap 100 DEG between sarcoidosis and TB	Limited clinical phenotyping, no microarray validation of findings, small sample size

Table 3. Microarray studies comparing patients with active TB and sarcoidosis

DEG = differentially expressed genes.

The first sarcoidosis gene expression study was designed specifically to test the hypothesis that mycobacterial antigens trigger an autoimmune response in genetically predisposed populations (Thonhofer, Maercker et al. 2002) (Table 3). Bronchoalveolar lavage cells from 6 untreated sarcoidosis patients were stimulated with dead *Mycobacterium avium* (Thonhofer, Maercker et al. 2002). They used other granulomatous diseases, 2 TB and 3 hypersensitivity pneumonitis, as the control groups. 1,500 probes from a whole human genome array had altered differential expression after stimulation but only 2 known genes (and 2 expressed sequence tags of unknown function) were exclusively differentiated in sarcoidosis samples. The methodology shown in the article is ambiguous, the sample numbers small and only the 2 genes of interest were validated (by reverse hybridization). Overall it appears doubtful that there would be only 4 differing genes between these diseases in response to the stimulation.

Two publications in the last year have compared human whole genome blood transcriptional profiles of patients with sarcoidosis to patients with TB (Table 3). Koth *et al.* 2011, recruited 38 sarcoidosis patients and 20 healthy controls and analysed their raw data alongside deposited raw data from different microarray platforms from two other sarcoidosis cohorts (blood and lymph node), one cohort of hypersensitivity pneumonitis patients (blood) and data from all the patients included in the Berry *et al*. 2010, study (the three TB cohorts, paediatric SLE, adult SLE, staphylococcal and streptococcal infections) (Koth, Solberg et al. 2011). Their principal finding was the similarity of blood gene expression of the sarcoidosis and TB patients, of which the dominant pathway in both diseases was the IFN-inducible genes. At the time of publication we had also found the same results from the training set in this study. Although Koth *et al.* 2011, identified a set of genes that could discriminate between their collection of sarcoidosis and tuberculosis patients they did not test the ability of these genes to discriminate between such patients in an independent cohort. In addition although the paper was focussed on sarcoidosis there was an insufficient amount of clinical information available about the sarcoidosis patients. For example they state there was no difference in expression profiles between those patients taking systemic glucocorticoids and those who were on no treatment, but they do not report the glucocorticoid dosages and as such they could be equivalent to normal endogenous levels. Furthermore they described a significant association between sarcoidosis expression profiles and a clinical phenotype they describe as severe sarcoidosis. This phenotyping was extremely limited (two lung function parameters FEV_1 and FVC) and what is more these are not evidenced based and not commonly recognised or reported in clinical practice as markers of sarcoidosis severity. Lastly they compare the sarcoidosis profiles to tuberculosis and hypersensitivity pneumonitis as three diseases with granulomatous inflammation. Although hypersensitivity pneumonitis can be a granulomatous disease this depends on the stage of the disease and therefore this is not a true comparison of granulomatous disease unless biopsies were taken, again this was not reported.

The second publication comparing blood expression profiles of patients with sarcoidosis (18 subjects) and active tuberculosis (9 subjects) as well as controls (17 subjects) also ascertained that the overlap between the two diseases was dominated by IFN-inducible genes (Maertzdorf, Weiner et al. 2012) (Table 3). Although their preliminary expression findings are compatible to the Koth *et al.* 2011 study, regrettably they neglected to confirm any findings in an independent cohort. Therefore the discriminative power of the 100 probes they describe as distinguishing TB from sarcoidosis has a potentially high error rate. It is also notable that their selection of 100 probes had no overlapping genes in common with the transcript list generated from the Koth *et al.* 2011 study. Maertzdorf's *et al* study reveals even less clinical information about the sarcoidosis patients than the Koth *et al* 2011 study, including no mention of any therapy the patients may have commenced or not; therefore it is difficult to gain an understanding of the spread of sarcoidosis patients that participated. Interestingly they examined microRNA expression in parallel with the gene expression profiling and multiplex cytokine analysis. The microRNA expression also presented a highly comparable pattern in both sarcoidosis and TB with only 4 microRNAs significantly differentiated between the diseases. Again there was no validation of these findings. Their cytokine analysis revealed an increase in pro-inflammatory proteins in the TB patients relative to the sarcoidosis patients but did not reveal any correlation between protein and gene expression patterns.

In summary there have been several sarcoidosis gene expression studies and two comparing sarcoidosis to TB. Although many of the studies have limitations they demonstrate the potential of gene expression studies as a methodology to improve our understanding of the biology underpinning sarcoidosis and as a prospective clinical tool to improve diagnosis. Furthermore the latest two studies suggest sarcoidosis and TB have similar whole blood transcriptional profiles.

Transcriptional profiling in other respiratory diseases

Peripheral blood transcriptional profiles have been identified in adults and children with viral respiratory infections, in both a laboratory setting and in the community (Thach, Agan et al. 2005; Ramilo, Allman et al. 2007; Zaas, Chen et al. 2009). Ramilo *et al* were able to define a specific 35 gene signature to differentiate between children with influenza A and those with either *Escherichia coli* or *Streptococcus pneumoniae*, where those with *Streptococcus pneumoniae* had a predominant respiratory illness (Ramilo, Allman et al. 2007). Their signature had 95% accuracy in discriminating power in an independent cohort. Notably children with viral infections revealed a prominent type 1 IFN profile while a third of the children with bacterial infection also displayed elevated levels of IFN-related genes. The authors speculate this may have been due to an undiagnosed or preceding viral infection. However we have been unable to find any published studies that have surveyed adults with community acquired bacterial pneumonia, which may represent a more complex group than children with a higher likelihood of potential co-existing illnesses.

Idiopathic interstitial pneumonias is the term used for a diverse group of diffuse parenchymal lung diseases that predominantly result in lung fibrosis but the aetiology often remains unknown (ATS 2002). Sarcoidosis, idiopathic pulmonary fibrosis (IPF) and hypersensitivity pneumonitis, all fall under the umbrella of idiopathic interstitial pneumonias. Hypersensitivity pneumonitis is caused by exposure to an inhaled antigen and when chronic can result in granulomatous inflammation. Gene expression profiling of lung biopsies has been used to try and classify the idiopathic interstitial pneumonias and understand the underlying mechanisms (Selman, Pardo et al. 2006) (Yang, Burch et al. 2007). Selman *et al* compared lung biopsies from 15 patients with hypersensitivity pneumonitis and 12 patients with IPF; the gene expression profile of hypersensitivity pneumonitis lung samples were enriched for genes that are functionally associated with inflammation, T-cell activation and immune responses, whereas the IPF profile was characterised by tissue remodeling, epithelial, and myofibroblast genes (Selman, Pardo et al. 2006). Intriguingly Lockstone *et al* compared their sarcoidosis lung biopsy transcriptional profiles for the progressive-fibrotic patients to data from the Selman *et al* study and unexpectedly found a similarity with hypersensitivity pneumonitis but not IPF profiles (Lockstone, Sanderson et al. 2010).

Lung cancer heterogeneity is well documented and reflected in its broad variety of histological subtypes, of which different histological types require different treatment modalities. Microarray hierarchical clustering analysis has been applied in numerous lung cancer studies of tissue biopsies to help improve disease classification with encouraging results (Garber, Troyanskaya et al. 2001; Gordon, Jensen et al. 2002; Yamagata, Shyr et al. 2003). Although a large number of studies have also claimed to have found potential prognostic signatures it appears that the majority of them had serious flaws in their design and analysis (Subramanian and Simon 2010). Therefore it seems unlike the validated and now clinically available 70-gene breast cancer-prognostic signature (van 't Veer, Dai et al. 2002), we are still a long way from identifying a clinically applicable prognostic signature for lung cancer. All these studies have used tissue biopsies for their sampling. Few solid tumour studies have used peripheral whole blood as the source for gene expression profiles.

Summary of the caveats of previous microarray studies

To our knowledge this study is the first study comparing whole genome transcriptional profiles of untreated active pulmonary TB patients to pulmonary sarcoidosis and other respiratory diseases. Although there is a reasonable collection of literature on gene expression studies in sarcoidosis patients only a few used the whole human genome, most did not validate their findings and there was no overall conformity in the study aims. In addition little acknowledgement has been made of the clinical heterogeneity of sarcoidosis and the impact this may have on individual transcriptional profiles.

Study Objectives

The main goal of our study was to use a broad unbiased microarray approach to develop a better understanding of the host responses that are underlying tuberculosis. The aim was to compare untreated tuberculosis to the untreated respiratory diseases sarcoidosis, pneumonia and lung cancer; and in addition examine tuberculosis before, during and after successful treatment, as well as compared sarcoidosis and pneumonia patients before and after receiving treatment. Alongside any discoveries relating to immunopathology this approach has the potential to elicit clinically applicable diagnostic and treatment monitoring biomarkers.

Although the application of microarray is often thought not to be hypothesis driven, rather a descriptive data driven approach, it can be very fruitful when predetermined focussed questions are outlined prior to data analysis.

The questions addressed in this analysis were:

- Are the peripheral blood transcriptional profiles in patients with TB the same or different to those in patients with sarcoidosis, pneumonia or lung cancer?
- Are the transcriptional responses of patients with sarcoidosis heterogeneous?
- What are the biological functions of the sets of genes associated with each disease or with more than one disease?
- Is there are specific TB-related transcriptional signature distinguishing TB from other diseases. If so do these genes have known biological significance and could they be used to discriminate TB from the other respiratory diseases and controls?
- Do the transcriptional profiles change in response to antituberculous treatment?
- Do the transcriptional profiles change in response to immunosuppressive treatment for sarcoidosis or antibacterial treatment for pneumonia – are they changes similar to each disease?

METHODS

METHODS

Patients and Healthy Controls

Ethics

Ethical approval was gained from the National Research Ethics Service. All subsequent changes to protocols were approved on an individual basis. Ethical approval was also gained for each NHS Trust that participants were from: Royal Free Hospital NHS Trust, Oxford Radcliffe Hospital NHS Trust, Barnet and Chase Farm NHS Trust, Imperial College Healthcare NHS Trust. Ethical approval was gained separately by the research groups/physicians in South Africa and France for the treated South African TB patients and for the patients recruited by the Lyon Collaborative Clinical Network (lung cancer and TB patients) or recruited in the Avicenne Hospital in Paris (sarcoidosis patients).

Subject recruitment and eligibility of the patients

This was an observational prospective case-control study. Subjects were recruited if considered to have pulmonary TB, pulmonary sarcoidosis, community acquired bacterial pneumonia, primary lung cancer or a healthy individual with no significant disease or exposure to mycobacteria. All patients were recruited consecutively over time. On initial recruitment all subjects gave informed written consent, answered relevant clinical questions, donated a blood sample and gave permission to access their hospital records. Some patients were also bled at further time points after commencing treatment.

Samples were only included for microarray analysis if the participants satisfied specific inclusion and exclusion criteria. For Pulmonary TB patients inclusion criteria was *M. tuberculosis* culture confirmed in either sputum or bronchoalveolar lavage. Samples were collected before patients' commenced antituberculous therapy. TB patients and controls were excluded if they had any significant medical past history. Pulmonary sarcoidosis patients had to have granulomatous inflammation on biopsy, as well as compatible clinical and radiological findings diagnosed by the specialist whose clinic they were attending. Sarcoidosis patients were only included if they had radiology imaging (chest radiograph or computed tomography scan) showing evidence of pulmonary involvement within 6 months of the blood test. Nearly all patients were included prior to commencing any immunosuppressive treatment for their sarcoidosis, or at least 6 weeks after stopping any immunosuppressive treatment. Two were recruited who were already on low dose treatment for their sarcoidosis (one patient was on 5mg prednisolone daily and one was on 200mg hydroxychoroquine daily). Sarcoidosis patients and controls were excluded if they had any significant medical past history. Community acquired pneumonia patients were only included if they had symptoms and signs consistent with an acute lower respiratory tract infection e.g. cough or fever, new radiographic shadowing consistent with infection and for which there was no other likely cause, the presenting illness was the primary reason for hospital admission and the hospital admission was managed as a community acquire pneumonia. Lung cancer patients were included if their histological and radiological features indicated their cancer was a primary lung cancer. Cancer patients were excluded if they had received any treatment for their cancer, had another cancer or any significant respiratory illness. All patients and controls were excluded if they were immunosuppressed other than by the study diseases.

Pulmonary TB patients were recruited mostly from the Royal Free Hospital but some were recruited from the Lyon Collaborative Network and from St Mary's Hospital. Pulmonary and latent TB patients recruited for the TB treatment section (chapter 8) were recruited from the Ubuntu HIV/TB clinic in South Africa by the local

TB research team under the global supervision of Dr Rob Wilkinson, or from the Royal Free Hospital. Sarcoidosis patients were recruited from Royal Free Hospital, John Radcliffe Hospital, Barnet Hospital, St Mary's Hospital and L'hospital Avicenne. The sarcoidosis patients from John Radcliffe Hospital were recruited by the specialist respiratory registrars Dr Yvonne West and Dr Anjali Cranshaw; from Lyon Collaborative Network some TB and all but one of the lung cancer patients were recruited by Mitra Saadatian; a few of the TB patients were recruited from St Mary's Hospital by Dr Matthew Berry (the patients were of white ethnicity and necessary to try and balance ethnicities between all the disease groups); some of the validation set sarcoidosis patients were recruited at L'Hôpital Avicenne by Dr Dianne Bouvry. In January 2011 the research nurse Fotini Rozakeas joined Anne O'Garra's research team to help with recruitment, I trained and supervised her throughout the study. Fotini Rozakeas recruited all the community acquired pneumonia patients, some of the sarcoidosis, TB and one lung cancer patients. I recruited all other TB and sarcoidosis patients (78 patients that were included in this study).

Subject recruitment and eligibility of the healthy controls

Healthy controls had no major medical illnesses and had no known exposure to mycobacteria as evidenced by their history, IGRA +/- TST results. Healthy controls were recruited from NIMR, Hammersmith Hospital and Royal Free Hospital by Fotini Rozakeas and me.

A crucial aspect in the recruitment of healthy controls was to ensure their age, gender and ethnicity matched the recruited patients, and also that they were not displaying any current symptoms of illness including coryzal symptoms. Therefore they were recruited alongside the patients to ensure matching as much as possible. The importance of this was demonstrated by a small study we carried out in 2009. This study was designed to determine any blood transcriptional affects due to the tuberculin skin test (TST). Twenty voluntary participants were bled before and at several time points after the TST (1 day, 3 days, 7 days and 29 days). Unsupervised analysis and unsupervised hierarchical clustering revealed that inter-individual transcriptional differences outweighed any transcriptional changes that may have occurred due to the tuberculin skin test, as each of the twenty patients clustered together including their time course profiles within their sub-cluster (Figure 2). Therefore this experiment demonstrated the TST does not significantly affect the blood transcriptional response, even though the TST does affect the transcriptional response when measured in the skin at the site of the injection (Tomlinson, Cashmore et al. 2011). In addition this experiment demonstrated important factors relating to the main study. Firstly the unsupervised clustering of the blood transcriptional profiles showed that inter-individual differences influenced the clustering over any intra-individual differences that may have occurred over the 29 days. This is helpful information for any longitudinal studies including the study of patients before, during and after treatment. Secondly over-abundance of genes was seen for the male sex genes and for two participants who had during their time course noted mild coryzal symptoms. As these variables could influence the clustering and influence the apparent differentially expressed genes they should therefore be minimised as much as possible to remove any bias from the analysis.

Sample and data collection

Venesection for whole blood samples was performed on every participant and collected in tubes containing the reagent Tempus that lyses cells and stabilises RNA. Clinical and demographic information was collected on each participant. All samples were pseudoanonymised. Results of the blood tests, chest radiographs, CT scans and lung function tests were obtained, once formally reported at the hospital, as part of the patient's routine medical care.

Clinical classification of sarcoidosis patients into active and non-active

Sarcoidosis patients were classified according to the clinical classification criteria (Figure 3). This criteria is based on evidence published in the literature that are thought to link particular clinical parameters with disease activity and on the commonly available test results from the hospitals patients were recruited from.

Experimental processing

Interferon Gamma Release Assays

Most patients had blood taken for the QuantiFERON-Gold In Tube ELISA (Cellestis, Carnegia, Australia). The assay was performed according to the manufacturer's instructions.

Serum Collection

1-2ml of blood was collected into serum clot activator tubes (2ml vacutainer tubes Becton Dickinson). Tubes were centrifuged at 2000g for 5 minutes at room temperature and the serum portion extracted and frozen at –80ºC pending protein analysis.

RNA extraction, amplification and hybridisation

An experienced laboratory technician Dr Chris Graham processed the samples by the following methods and throughout the processing I assisted with every step for the training set samples. RNA was isolated and purified from 1.5ml of whole blood using protocols developed in collaboration with the Baylor Institute of Immunological Research (BIIR). 2.5µg of isolated total RNA was globin reduced using the GLOBINclear 96-well format kit (Ambion). This is required to ensure that the large quantities of globin mRNA transcripts present from the red blood cells do not affect the sensitivity of the microarray. 200 - 250ng of globin-reduced RNA was used to prepare biotinylated, amplified RNA targets (cRNA) using the Illumina CustomPrep RNA amplification kit (Ambion). Total RNA, globin-reduced RNA and cRNA integrity was assessed using an Agilent 2100 Bioanalyzer (Agilent Technologies) and RNA yield was assessed using a Nanodrop 1000 spectrophotometer (Thermo Fisher Scientific Inc). A technician at BIIR labelled then hybridized the cRNA to Illumina HT-12 version 4 BeadChip arrays (Illumina Inc). The BeadChip arrays were scanned on an Illumina iScan (Illumina Inc) which generated the signal intensity values. Background is defined as the average signal intensity estimated from the negative control bead types and is subtracted prior to array normalisation.

Microarray and statistical analysis

Detection from background, threshold value, logarithmic transformation and normalisation

Raw data, expression filtering and statistical filtering were processed using GeneSpring GX version 11.5 (Agilent Technologies). The following was applied to all analyses. After background subtraction each probe was attributed a flag to denote its signal intensity detection *p*-value. Flags were used to filter out probe sets that did not result in a ‘present’ call in at least 10% of the samples, where the ‘present’ lower cut off = 0.99. Signal values were then set to a threshold level. For all analysis this threshold was set to

10 (to reduce the noise from low level intensity values) but for the analysis of the TB treated patients (Chapter 7) the value was set to 1 (the technical variation in this microarray was much lower therefore a lower threshold value could be afforded). All data was then log2 transformed (to reduce effects from skewed data that is common place in microarray) and per-chip normalised using 75th percentile shift algorithm. Illumina probes often generate many low expression values which have a lower specificity than the higher values therefore this normalisation algorithm was chosen to reduce reliance on these values. Next per-gene normalisation was applied by dividing each transcript either by the median intensity of all transcripts (part of the unsupervised analysis protocol) or by the median intensity of the latent TB transcripts (part of the supervised analysis protocol).

Quality control

To ensure there is no technical variation causing unexpected skewing of the raw data quality control was performed. Two different checks were performed; firstly principal component analysis (PCA) on the unfiltered data and secondly unsupervised hierarchical clustering on data only filtered by unsupervised analysis. These two checks were applied to allow identification of samples that were likely to be outliers. For example PCA can be used to screen all samples at once for obvious outliers in variation of the main component of variation derived in the PCA. In addition PCA can also be used to screen, for example, all TB samples simultaneously. This ensures genuine outlying samples are identified, not just samples that appear different due to their disease. For all outliers that were identified both their RNA quality and clinical data were checked to ensure the reason for apparent outliers was due to technical variation or human error, not due to biological variation.

Quantifying changes in expression

After per-gene normalisation a fold change filter was applied. In most analysis the filter was set to include only transcripts that had at least two-fold changes from the median and were present in at least 10% of the samples. In the South African TB treatment analysis transcripts had to satisfy a three-fold expression filter in 12 of the 15 training set matched untreated and 6-month treated samples. This supervised analysis was introduced to derive a more specific list of genes that change in response to treatment.

Statistical analysis of microarray data

To improve the specificity of the analysis a statistical filter was applied after the expression filter. All data was analysed using non-parametric tests as the genes tested did not have a Gaussian distribution. For two groups either Mann-Whitney or the less stringent Significance Analysis of Microarray was applied. For more than two groups Kruskal-Wallis (non-parametric equivalent of analysis of variance) was used. Either the False Discovery Rate (FDR) Benjamini-Hochberg (BH) or the more stringent Bonferroni multiple testing corrections was also applied to all analysis.

Choosing how to perform the initial steps of data analysis

The beginning steps of the data analysis were based on the recommendations from GeneSpring (Agilent Technologies) for Illumina gene expression analysis. Prior microarray analysis carried out by Dr O’Garra’s research group used a slightly different approach, based on the analysis strategies used at Baylor Institute of Immunological Research (Berry, Graham et al. 2010). The vast literature of studies applying different analysis approaches demonstrates that to analyse large complex data sets there is no ‘correct analysis approach’ however standardisation of strategies would be of great benefit for pragmatic interpretation between experiments (Quackenbush 2001; Nature, editorial. 2012). To determine the effects of these two different strategies some comparative analysis using the datasets from the Berry *et al*. 2010 study was carried out (Figure 4). The first step is to normalise the samples, the different normalisation approaches particularly affected the unsupervised analysis as this uses a cut-off filter around the median of all transcripts – the median of each transcript however can change according to the normalisation applied. The next steps are to set a threshold level and then filter out transcripts that are not significantly different from the background intensity values, the two filtering processes had little effect on the outcome and the difference was most likely related to a cut-off of <0.99 compared to ≤0.99 (Figure 5b). The threshold step had a much more significant effect on the outcome as a lower threshold allows many more genes through the unsupervised analysis filter around the median (Figure 5c). The lower threshold may include more noisy genes but the higher threshold may disregard genes of interest. For data with less technical variation it may be better to set a lower threshold to ensure fewer falsely negative transcripts. After the additional statistical filter many of the changes, in terms of number of transcripts, have less of a proportional difference. The main changes are secondary to the threshold value and the normalisation of samples (Figure 5d & Table 4). However the crucial question is not really how many genes are different but are the functional pathways changed by the analysis. The Microarray Quality Control Project by the United States Food and Drug Administration set out to specifically address this issue of the reliability of microarray data and in response to this published a paper comparing multiple different analysis approaches for the same experiments (Shi, Campbell et al. 2010). Their findings were that regardless of the analysis strategies applied the transcriptional profiles, although somewhat differing at the gene level, revealed remarkably similar biological properties suggesting that microarray does appear to be reliable at the biological level. Furthermore a robust dataset should not have widely varying biological outcomes from the data when applying different analyses. Therefore with this in mind the different transcript lists generated using slightly different initial analysis steps by IPA were compared (Ingenuity Systems). To simplify this comparison the IFN-signalling pathway, a pathway known to be significantly represented in each dataset, was used to assess the results. The IFN-signalling pathway remained highly significant by all forms of analysis in the three main cohorts but not in the SA Validation and SA Baseline cohorts (Table 5). This was most likely due to the very stringent multiple testing corrections (Bonferroni) applied to these two cohorts to reduce the transcript lists to a size parallel to the others. The stringency of the correction results in many more false negative genes and the number of genes in the IFN-signalling pathway is only 36 therefore the loss of 2 or 3 genes can make a huge impact on the statistical significance.

In summary the analysis strategy is a choice that should be based on the quality of the raw data, the type of samples and the questions to be addressed. Importantly though there is no single correct strategy but it should be used consistently throughout the experiment and any comparative experiments. The use of different approaches to determine functional associations that subsequently show agreement will add to the confidence of any findings. For this reason a number of different tools (modular analysis, IPA, gene ranking, Venn diagrams and comparing disease to disease) were applied to try and find biological links with the transcript lists (Figure 6).

Clustering

Hierarchical clustering was unsupervised for both the transcripts and samples, clustered only by gene expression and sample similarity without knowledge of sample identity, or partially supervised, intentional clustering of the samples by particular parameters e.g. time course. For all clustering the distance metric Pearson's uncentered correlation with average linkage was applied, so that clustering was driven by the trend in expression of the transcripts. The hierarchical transcriptional clustering was visualised using a heatmap with a vertical dendrogram indicating the most similar transcripts and horizontal dendrogram indicating the most similar samples. A colour bar indicates the normalised intensity values where yellow is zero (no different from the median if samples are normalised to the median), red is high (over-abundant relative to the median) and blue is low (under-abundant relative to the median). For *k*-means clustering the distance metric Pearson's uncentered was applied with *k*=20 with 200 iterations.

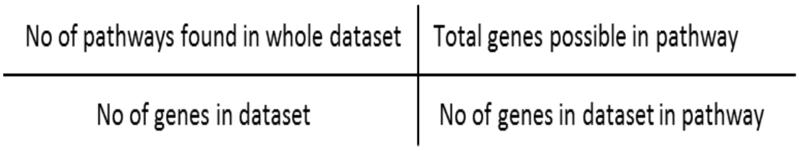
Class prediction

The machine learning algorithm support vector machine was performed in GeneSpring 11.5. The training set was used to build the prediction model and the test, validation and other datasets were used to run the model. When raw data was obtained from different platforms the model was built again for that platform. The model was built using sample classifiers 'TB' or 'not TB'. The kernel type used was linear, maximum iterations 100,000, cost 100, ratio 1 and validation type N-fold where N=3 with 10 repeats.

Ingenuity Pathway Analysis

IPA (Ingenuity Systems) software package was used to help elucidate gene pathways using a top-down approach. The repository underlying IPA is called

Ingenuity® Knowledge Base (IngenuitySystems 2012). This database is created from millions of individually modelled relationships from diseases to tissues to genes. It is built by manually abstracting and curating much of the biomedical literature, an expert review process, with findings added weekly. IPA includes a gene expression specific database which consists of genes associated with particular biological and functional pathways. In our study only gene expression canonical pathway analysis was carried out. Two forms of IPA pathway analysis was performed; discovery of significant pathways for each disease at one time and discovery of significant pathways for each disease in comparison to the other diseases. The significance of each pathway was calculated by IPA using Fisher's exact test with a Benjamini Hochberg multiple testing correction.



No of pathways found in whole dataset	Total genes possible in pathway
No of genes in dataset	No of genes in dataset in pathway

The number of genes present in that pathway from your dataset was also provided.

Weighted molecular distance to health and temporal molecular response

Weighted molecular distance to health (MDTH) is an algorithm used to determine the degree of perturbation of expression of a group of samples compared to a set of controls (Pankla, Buddhisa et al. 2009). Berry *et al*. 2010, demonstrated a significant correlation between MDTH and the severity of active TB as measured by the radiographic extent of disease (Berry, Graham et al. 2010). MDTH is calculated by the number of transcripts per sample that differ by more than two standard deviations from the mean of the controls. The MDTH score for each sample is then a function of both the number of deviating transcripts and the amount of standard deviations they deviate by.

Temporal molecular response is an algorithm devised to calculate the change in a transcriptional profile over time. Unlike MDTH it does not require a control cohort as it uses the first time point (baseline) as the comparator profile. Another advantage over MDTH is that it is more sensitive and specific to changes in longitudinal analysis; this is in part because it does not rely on a control cohort that can have variable profile heterogeneity. The temporal molecular response was calculated for a particular transcript list for each individual patient. The raw intensity transcript values in the transcript list were consecutively compared at each time point to the baseline (pre-treatment). The numbers of transcripts that were at least two-fold up or two-fold down from the baseline were added together for each time point. This sum was then divided by the total number of transcripts in the transcript list to calculate a percentage score for each time point. This generated a percentage score of change at each time point compared to the baseline, where the baseline always remains zero (no change from itself). To allow for two-fold changes from zero any baseline raw transcript intensity values of zero were converted to 10^{-20}. The MDTH and temporal molecular response scores were calculated using Microsoft Excel 2010. GraphPad Prism version 5 for Windows was used to generate the graphs. Fixed effects longitudinal data regression models were used to determine *p*-values using Stata Statistical Software: Release 9. (College Station, TX: StataCorp LP). This statistical method allows regression analysis even when there are missing data points but does not create dummy variables.

Modular analysis

The modules are sets of functionally related genes derived from true biological conditions. These co-expressed genes were determined, using an algorithm based around *k*-means clustering, from gene expression profiles from patients in cohorts of different diseases (Chaussabel, Quinn et al. 2008). Subsequent to the clustering the genes and hence modules functional meaning were annotated using unbiased literature profiling (IPA, Pubmed and iHOP databases). The modules used in this study are modified modules, different from those used by Berry *et al*. 2010, (Berry, Graham et al. 2010). These modules were generated using the Illumina platform, from whole blood gene expression profiles and from patients with 9 different diseases. The older modules were generated using the Affymetrix platform, from PBMC gene expression profiles and from patients with 8 different diseases. In addition the output of these modified modules is per patient, as opposed to an average for the disease in the older modules (Guiducci, Gong et al. 2010). Module colour intensity represents the relative amount of over-abundance (red) or under-abundance (blue) compared to the controls (p<0.05), no colour indicates insignificant change in expression compared to the controls (p>0.05).

Four-set Venn diagram

As it is only possible to create 3-set Venn diagrams in GeneSpring 11.5, the 4-set Venn diagram was created using Venny (Oliveros 2007).

Patient randomisation

Patients were randomised using a computer algorithm for randomisation (Haahr 1988). As the expected accuracy of the signature was high the cohort was divided so approximately half the samples were included the training set and half were included in the test set (Dobbin and Simon 2011).

Translating probes across different microarray platforms

Probe/transcript lists were translated using the NIH Database for Annotation, Visualization and Integrated Discovery (DAVID). When possible for probes or genes that translation was not recognised by DAVID, annotation and conversion to Illumina probes in GeneSpring was carried out manually.

Clinical data statistical analysis

Univariate and multivariate regression analysis, and the chi-squared analysis were calculated using STATA9 Data Analysis and Statistical Software. To prevent listwise deletion due to missing data points in the multivariate regression analysis, dummy variable adjustment was used.

Figures for methods

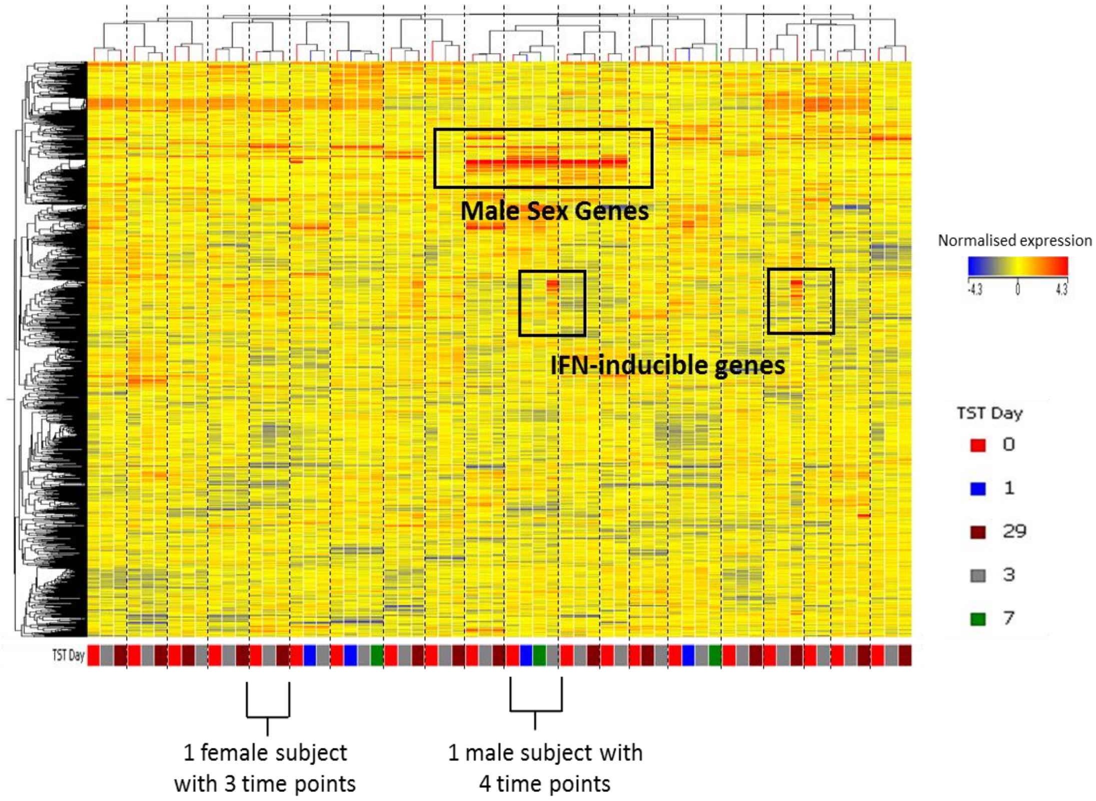


Figure 2. Unsupervised hierarchical clustering demonstrating that inter-individual transcriptional differences outweigh the intra-individual differences

RNA was extracted from whole blood from 20 healthy volunteers before and at the time points shown after the TST. The RNA was hybridised to Illumina HT 12 V3 microarray chips. The 1,064 transcripts shown in the heatmap were generated by unsupervised analysis; filtered firstly by their detection compared to background intensity ($p<0.01$) and then by a two-fold filter from the median in ≥ 10% of the samples. Unsupervised hierarchical clustering was then applied to these transcripts. Each row represents a transcript and each column represents a sample. The transcripts and samples were clustered by Pearson uncentered distance metric with average linkage. The vertical dendrogram shows the clustering of the transcripts and the horizontal dendrogram shows the samples clustering. The relative abundance of the normalised transcripts is indicated by the colour scale. The coloured bar at the bottom of the heatmap indicates the group the sample belongs to, as shown in the legend. The horizontal dendrogram divides into 20 main clusters representing the 20 different subjects in the experiment. Subjects clustered according to over-abundance of ribosomal variants (top left hand red genes), the male sex genes and other unclear factors. Subjects only clustered by intra-individual differences, i.e. the longitudinal time course of 29 days, within their own individual clusters. Transcriptional profiles are therefore influenced more by inter-individual differences than by either the affects from the tuberculin skin test or from the time course.

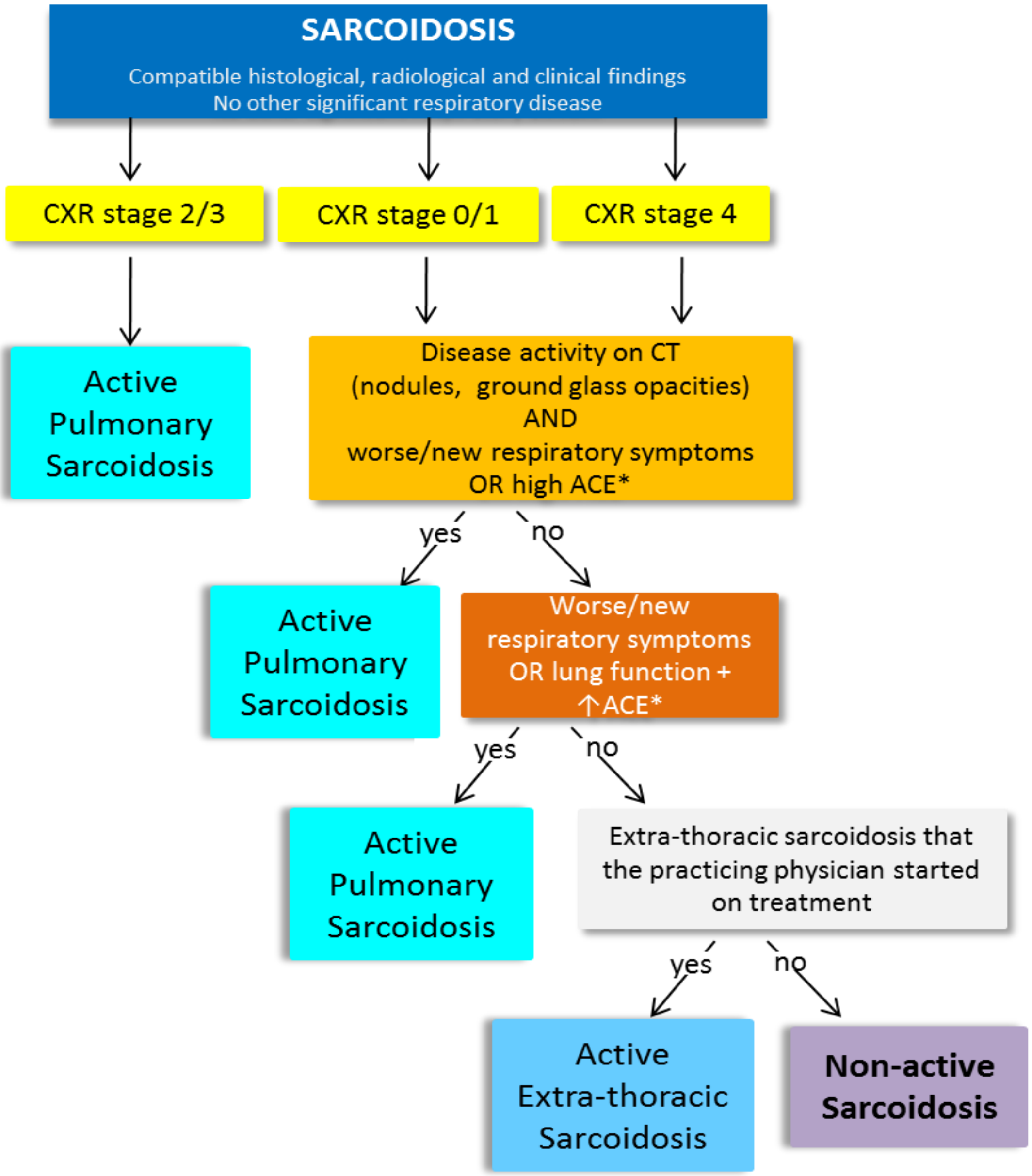

SARCOIDOSIS
Compatible histological, radiological and clinical findings
No other significant respiratory disease
CXR stage 2/3
CXR stage 0/1
CXR stage 4
Active Pulmonary Sarcoidosis
Disease activity on CT (nodules, ground glass opacities) AND worse/new respiratory symptoms OR high ACE*
yes
no
Active Pulmonary Sarcoidosis
Worse/new respiratory symptoms OR lung function + ↑ACE*
yes
no
Active Pulmonary Sarcoidosis
Extra-thoracic sarcoidosis that the practicing physician started on treatment
yes
no
Active Extra-thoracic Sarcoidosis
Non-active Sarcoidosis

Figure 3. Flow diagram of the clinical classification for sarcoidosis patients.

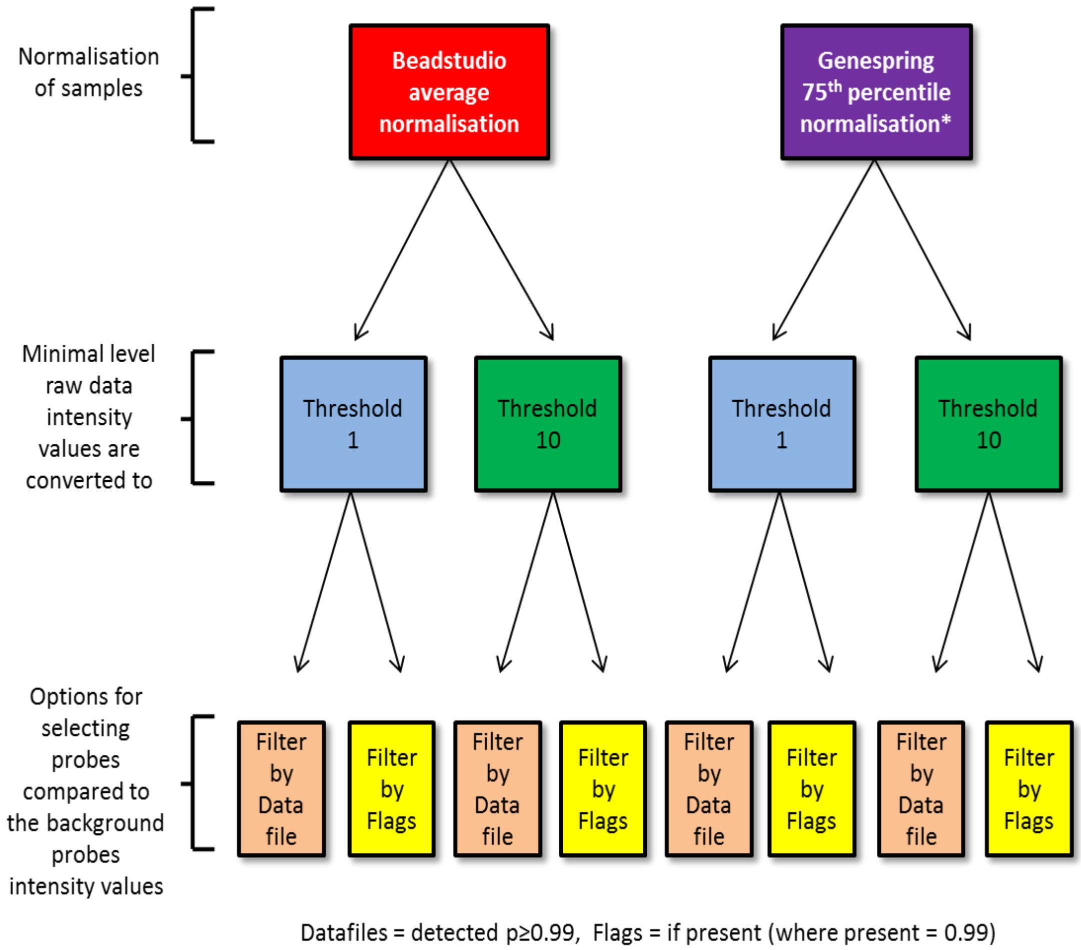
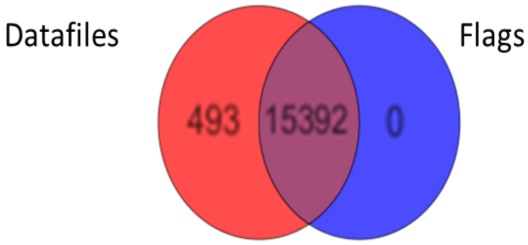

Normalisation of samples
Beadstudio average normalisation
Genespring 75th percentile normalisation*
Minimal level raw data intensity values are converted to
Threshold 1
Threshold 10
Threshold 1
Threshold 10
Options for selecting probes compared to the background probes intensity values
Filter by Data file
Filter by Flags
Filter by Data file
Filter by Flags
Filter by Data file
Filter by Flags
Filter by Data file
Filter by Flags
Datafiles = detected p≥0.99, Flags = if present (where present = 0.99)

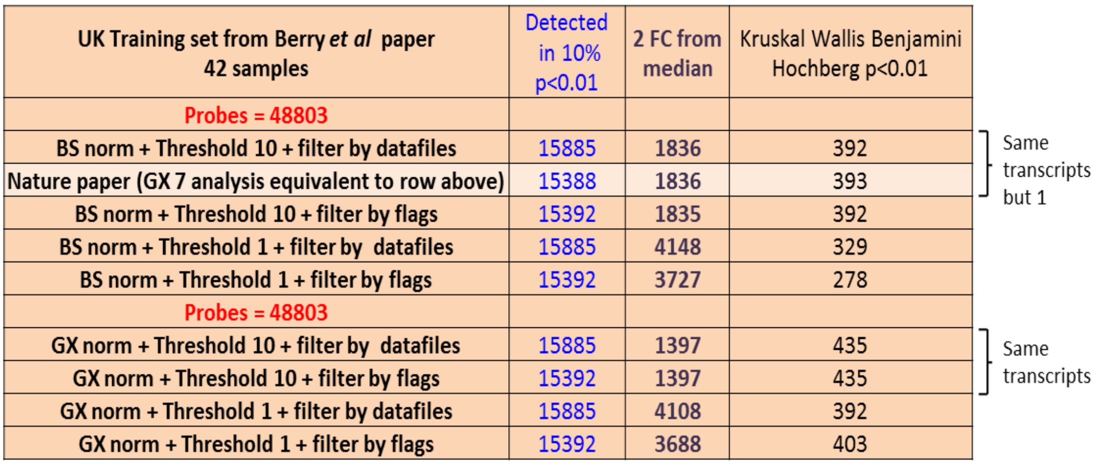
Figure 4. Comparing analysis strategies: classical Baylor Institute of Immunological Research approach versus GeneSpring advised approach.

The GeneSpring approach is recommended for Illumina data as there are many probes with low intensity values which are not as reliable as the higher values therefore they recommend to use a normalisation based around a higher percentile, such as 75th. However they also recommend using a threshold value of 1 to ensure no loss of data interpretation at the low lying intensity values. After threshold application the data is filtered to remove transcripts that are not significantly different from the background hybridisation intensity. Background subtraction is the average bead type intensity of each analytical probe minus the average intensity from the negative control beads for that probe. The filtering analysis can be performed in two ways in GeneSpring by ‘datafiles’ or by ‘flags’. Each probe is given a detection 1-(p-value) in Genome Studio. In addition a ‘flag’ or ‘call’ can also be set in GeneSpring for each Illumina probe to correlate with the detection value. A filter is then applied either by datafiles of ≥0.99, which translates to 1-(≥0.99) = p<0.01. Or a filter is set according to the ‘present flags’ where present =0.99, also equivalent to p<0.01.

a) Different filtering against background probes



b) Results of fold change around the median



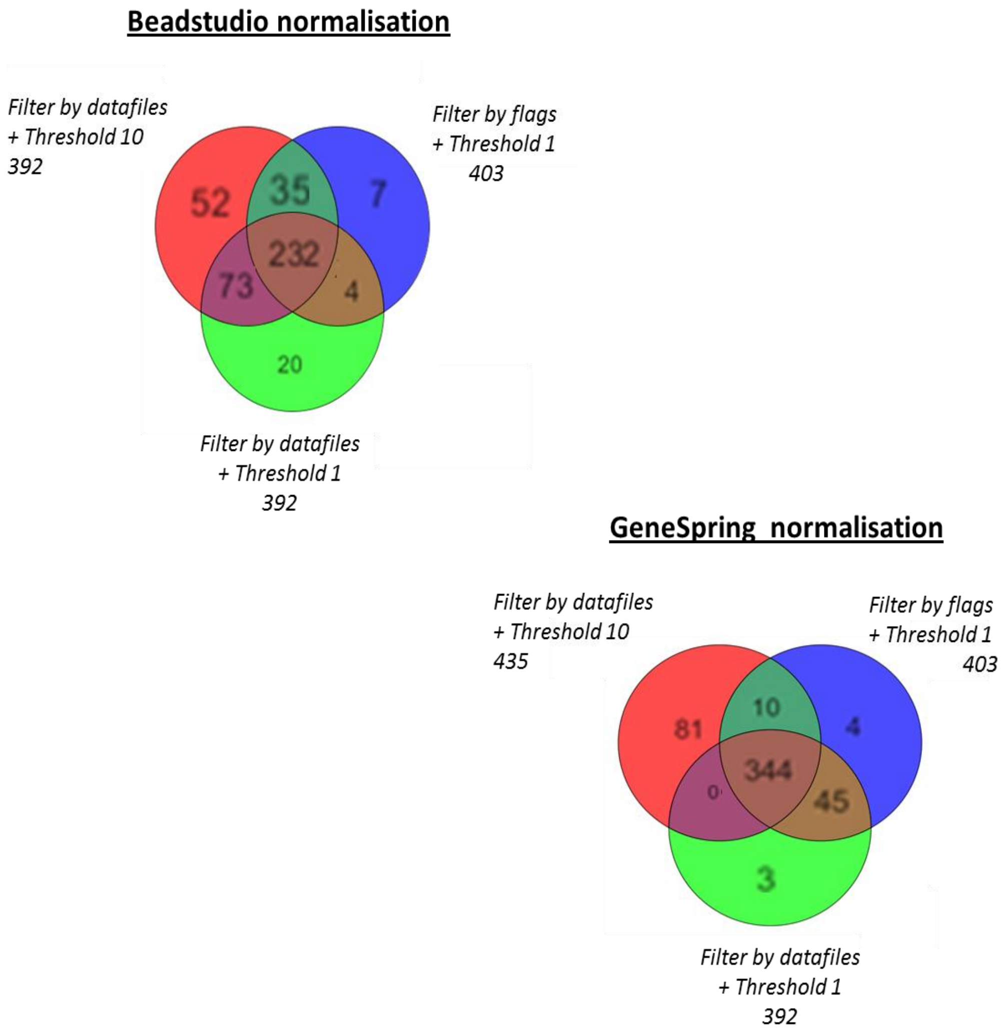
UK Training set from Berry *et al* paper 42 samples	Detected in 10% p<0.01	2 FC from median	Kruskal Wallis Benjamini Hochberg p<0.01
Probes = 48803			
BS norm + Threshold 10 + filter by datafiles	15885	1836	392
Nature paper (GX 7 analysis equivalent to row above)	15388	1836	393
BS norm + Threshold 10 + filter by flags	15392	1835	392
BS norm + Threshold 1 + filter by datafiles	15885	4148	329
BS norm + Threshold 1 + filter by flags	15392	3727	278
Probes = 48803			
GX norm + Threshold 10 + filter by datafiles	15885	1397	435
GX norm + Threshold 10 + filter by flags	15392	1397	435
GX norm + Threshold 1 + filter by datafiles	15885	4108	392
GX norm + Threshold 1 + filter by flags	15392	3688	403

Same transcripts but 1

Same transcripts

Figure 5. Using the training set from the Berry *et al* paper to compare the different strategies (a)Filtering from background (b) Fold change around the median (c) Statistical filtering (see next page)

5 Normalisation does not affect the number of genes after detection from background probes as would be expected. It does affect the fold change filter based around the median as the median for each transcript could be different depending on how the samples are normalised. (a) Both filter using *p*≤0.99 however when filtering by datafiles there is a fault in GeneSpring where it does not recognised 'at least' 10% of samples as ≥10% but sees it as 10% of samples. (b) Filtering around the median after threshold of 1 includes many more genes than after threshold of 10. The extra genes are at the low intensity value as the filter cut-off is based on a 2 fold change. This may therefore include false positive genes but equally could include relevant genes.
BS = Beadstudio, GX = GeneSpring, FC = fold change.

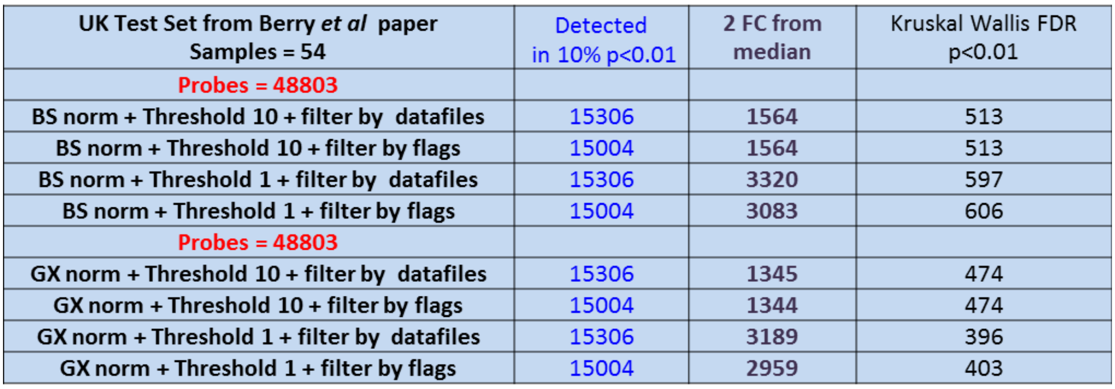


Beadstudio normalisation

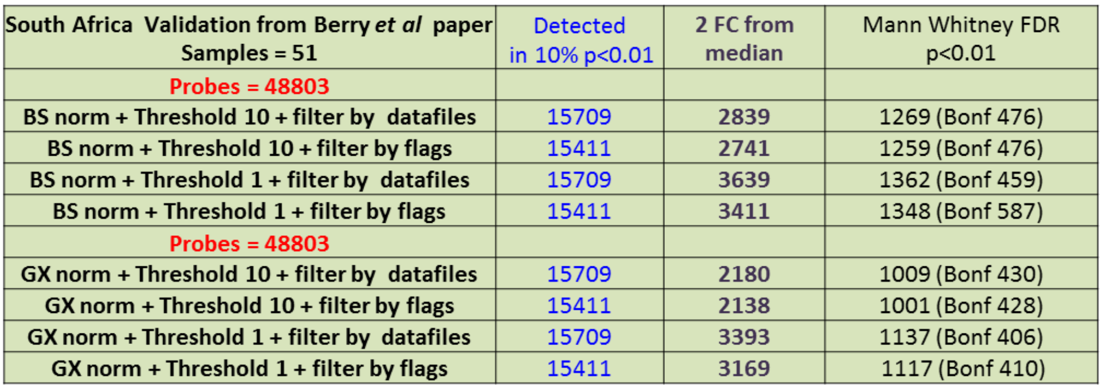
GeneSpring normalisation

Figure 5c. Comparing strategies – effects after statistical filtering

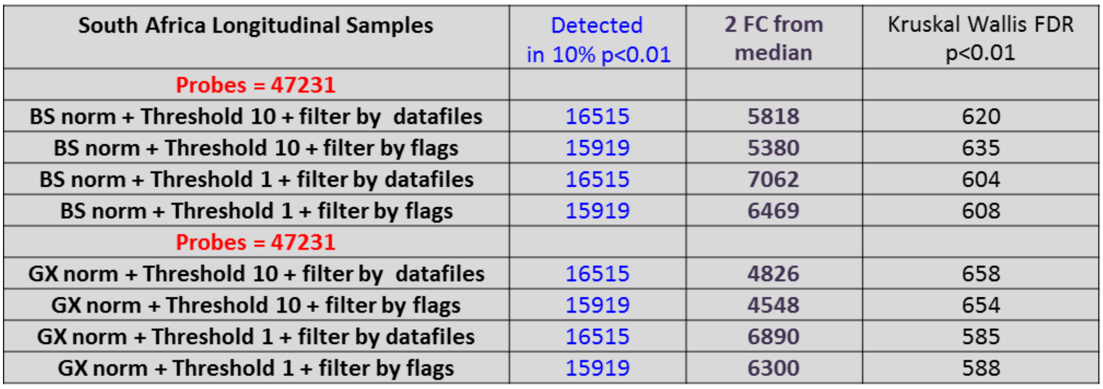
After statistical filtering the differences in transcripts are minimised but the main differences remain secondary to the threshold value. Although the higher threshold value reduces the genes that filter through the 2 fold change around the median, after statistical filtering there are fewer genes when tested in this cohort. This could be explained by the multiple testing correction that is applied – the more genes the more stringent the correction as it is calculated in relation to the number of genes.



UK Test Set from Berry *et al* paper Samples = 54	Detected in 10% p<0.01	2 FC from median	Kruskal Wallis FDR p<0.01
Probes = 48803			
BS norm + Threshold 10 + filter by datafiles	15306	1564	513
BS norm + Threshold 10 + filter by flags	15004	1564	513
BS norm + Threshold 1 + filter by datafiles	15306	3320	597
BS norm + Threshold 1 + filter by flags	15004	3083	606
Probes = 48803			
GX norm + Threshold 10 + filter by datafiles	15306	1345	474
GX norm + Threshold 10 + filter by flags	15004	1344	474
GX norm + Threshold 1 + filter by datafiles	15306	3189	396
GX norm + Threshold 1 + filter by flags	15004	2959	403



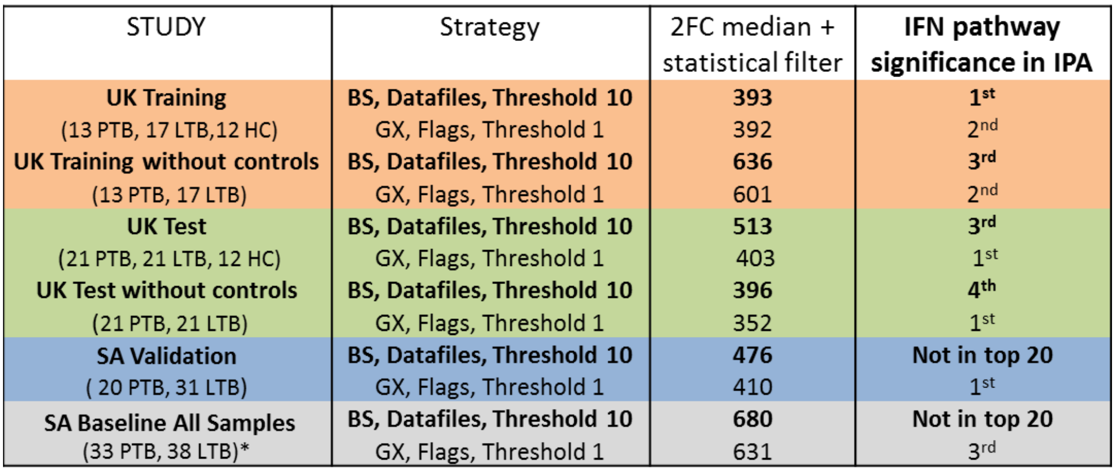
South Africa Validation from Berry *et al* paper Samples = 51	Detected in 10% p<0.01	2 FC from median	Mann Whitney FDR p<0.01
Probes = 48803			
BS norm + Threshold 10 + filter by datafiles	15709	2839	1269 (Bonf 476)
BS norm + Threshold 10 + filter by flags	15411	2741	1259 (Bonf 476)
BS norm + Threshold 1 + filter by datafiles	15709	3639	1362 (Bonf 459)
BS norm + Threshold 1 + filter by flags	15411	3411	1348 (Bonf 587)
Probes = 48803			
GX norm + Threshold 10 + filter by datafiles	15709	2180	1009 (Bonf 430)
GX norm + Threshold 10 + filter by flags	15411	2138	1001 (Bonf 428)
GX norm + Threshold 1 + filter by datafiles	15709	3393	1137 (Bonf 406)
GX norm + Threshold 1 + filter by flags	15411	3169	1117 (Bonf 410)



South Africa Longitudinal Samples	Detected in 10% p<0.01	2 FC from median	Kruskal Wallis FDR p<0.01
Probes = 47231			
BS norm + Threshold 10 + filter by datafiles	16515	5818	620
BS norm + Threshold 10 + filter by flags	15919	5380	635
BS norm + Threshold 1 + filter by datafiles	16515	7062	604
BS norm + Threshold 1 + filter by flags	15919	6469	608
Probes = 47231			
GX norm + Threshold 10 + filter by datafiles	16515	4826	658
GX norm + Threshold 10 + filter by flags	15919	4548	654
GX norm + Threshold 1 + filter by datafiles	16515	6890	585
GX norm + Threshold 1 + filter by flags	15919	6300	588

Table 4. The observed findings are also seen when comparing strategies in the three other datasets.

BS = Beadstudio, GX = GeneSpring, FC = fold change, Bonf = Bonferroni multiple test correction, FDR = false discovery rate.



STUDY	Strategy	2FC median + statistical filter	IFN pathway significance in IPA
UK Training	**BS, Datafiles, Threshold 10**	**393**	**1st**
(13 PTB, 17 LTB,12 HC)	GX, Flags, Threshold 1	392	2nd
UK Training without controls	**BS, Datafiles, Threshold 10**	**636**	**3rd**
(13 PTB, 17 LTB)	GX, Flags, Threshold 1	601	2nd
UK Test	**BS, Datafiles, Threshold 10**	**513**	**3rd**
(21 PTB, 21 LTB, 12 HC)	GX, Flags, Threshold 1	403	1st
UK Test without controls	**BS, Datafiles, Threshold 10**	**396**	**4th**
(21 PTB, 21 LTB)	GX, Flags, Threshold 1	352	1st
SA Validation	**BS, Datafiles, Threshold 10**	**476**	**Not in top 20**
(20 PTB, 31 LTB)	GX, Flags, Threshold 1	410	1st
SA Baseline All Samples	**BS, Datafiles, Threshold 10**	**680**	**Not in top 20**
(33 PTB, 38 LTB)*	GX, Flags, Threshold 1	631	3rd

Table 5. Comparing the effect of the two extremes of the different analysis strategies using Ingenuity Pathway Analysis.

Although the number of genes may not differ greatly between the analysis strategies for the same cohort this can affect the IPA pathways that are shown as the most significant. For example it would be expected that IFN signalling pathway would be highly significant in all these datasets but clearly by using a more stringent multiple testing correction this has altered the significance of the IFN signalling pathway. This will be due to the small number (36 genes) present in that pathway therefore changes of a few changes can greatly affect the significance.

In summary from this comparison of analyses it would seem the choice of the strategy can affect the results and could be considered based on the quality of the raw data, type of samples used and questions to be addressed from the experiment. However there is no single ‘correct’ analysis strategy. When searching for biological patterns that are associated with the resulting transcript lists it could be of benefit to use several different data mining approaches to ensure the findings are consistent.

BS = Beadstudio, GX = GeneSpring, FC = fold change, SA = South Africa, PTB = pulmonary TB, LTB = latent TB, IPA = Ingenuity Pathway Analysis, IFN = interferon.

RESULTS

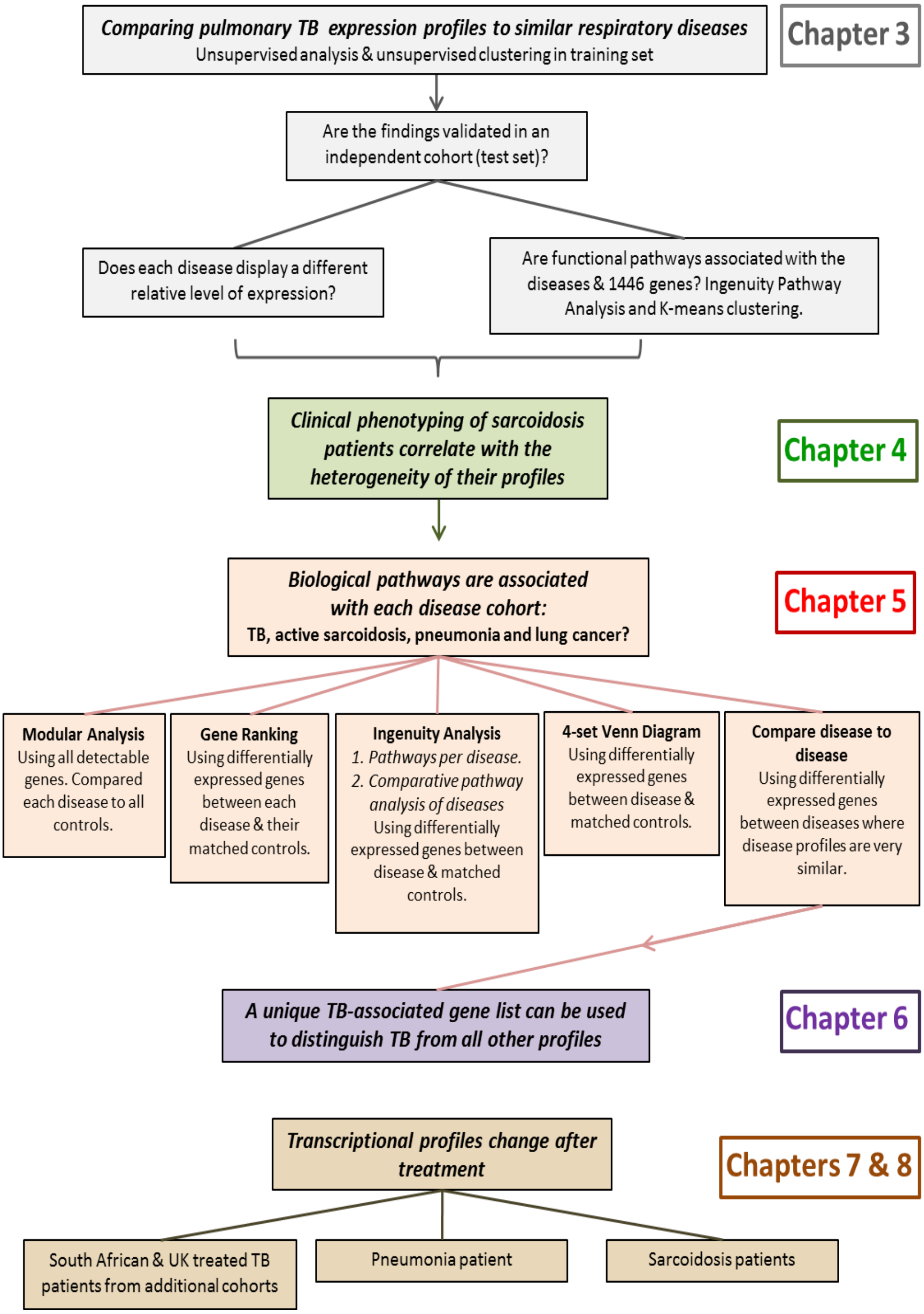

Comparing pulmonary TB expression profiles to similar respiratory diseases
Unsupervised analysis & unsupervised clustering in training set
Chapter 3
Are the findings validated in an independent cohort (test set)?
Does each disease display a different relative level of expression?
Are functional pathways associated with the diseases & 1446 genes? Ingenuity Pathway Analysis and K-means clustering.
Clinical phenotyping of sarcoidosis patients correlate with the heterogeneity of their profiles
Chapter 4
Biological pathways are associated with each disease cohort:
TB, active sarcoidosis, pneumonia and lung cancer?
Chapter 5
Modular Analysis Using all detectable genes. Compared each disease to all controls.
Gene Ranking Using differentially expressed genes between each disease & their matched controls.
Ingenuity Analysis *1. Pathways per disease.* *2. Comparative pathway analysis of diseases* Using differentially expressed genes between disease & matched controls.
4-set Venn Diagram Using differentially expressed genes between disease & matched controls.
Compare disease to disease Using differentially expressed genes between diseases where disease profiles are very similar.
A unique TB-associated gene list can be used to distinguish TB from all other profiles
Chapter 6
Transcriptional profiles change after treatment
Chapters 7 & 8
South African & UK treated TB patients from additional cohorts
Pneumonia patient
Sarcoidosis patients

Figure 6. Flow diagram of the results chapter to explain the analysis strategy of the whole study.

Chapter 3

Comparing pulmonary tuberculosis blood gene expression profiles to similar respiratory diseases

Chapter 3: Comparing pulmonary tuberculosis blood gene expression profiles to similar respiratory diseases

Introduction

TB is a complex infectious disease and in the face of numerous years of investigation there are still many host-pathogen immunological aspects we have little knowledge about or do not fully understand. A systems biology approach can evaluate a far broader framework than traditional reductionist methods by use of large complex datasets to identify key networks of interactions and new functional associations (Young, Stark et al. 2008). In addition a side product, or the main objective, can be the discovery of potential surrogate markers of diagnosis, prognosis and/or disease monitoring. Earlier blood transcriptional studies in TB have uncovered previously underappreciated roles of Type I IFN and demonstrated the power of transcriptional signatures to discriminate between grades of TB severity and discriminate between active TB and other infectious and inflammatory diseases (Berry, Graham et al. 2010), including between active TB and sarcoidosis (Koth, Solberg et al. 2011; Maertzdorf, Weiner et al. 2012). Furthermore the Berry *et al*. 2010, study showed potential for transcriptional signatures to help guide our understanding of the suspected clinical and immunological heterogeneity of latent TB.

This study is using the clinical similarity of other respiratory diseases to try to formulate distinct biological patterns associated with TB compared to the other diseases. For this reason particular similar diseases were chosen: pulmonary sarcoidosis – another respiratory granulomatous disease, community acquired pneumonia – another respiratory infectious disease, and primary lung cancer – another pulmonary inflammatory disease. All three diseases are also differential diagnoses for pulmonary TB (Campbell and Bah-Sow 2006). For example of three patients recruited initially as active TB in this study, one had sarcoidosis, one had pneumonia and one had lung cancer, all only determined after further invasive diagnostic tests. Therefore the possibility of identifying a transcriptional signature that can differentiate TB from these other respiratory diseases could be of great clinical value to prevent delays in the diagnosis of TB (Rodger, Jaffar et al. 2003), in addition to providing new information regarding the host factors underlying pathogenesis.

Results

Study Recruitment for the Training Set

TB patients were recruited from August 2009 to October 2010 (Royal Free Hospital) (Figure 7). Twenty three TB patients were recruited, three were excluded as the diagnosis of TB was not confirmed (lung cancer, emphysema and culture negative), and three were excluded due to the subsequent diagnosis of additional diseases (B-cell lymphoma and hepatitis B) (Figure 7). A further four samples were not included but instead four samples from white TB patients which had been previously recruited were used (Berry, Graham et al. 2010). This was to reduce the ethnicity bias of the TB patients compared to the sarcoidosis patients.

Sarcoidosis patients were recruited from October 2009 until October 2010 (Royal Free Hospital and Oxford). Forty one sarcoidosis patients were recruited, seven were excluded as their radiology imaging performed after the blood test was normal, three had a confounding respiratory disease (moderate asthma) and three had no biopsy or a normal tissue histology (Figure 7). One patient was included although he was on glucocorticoids, a very low dose (prednisolone 5mg daily), as he had severe active sarcoidosis but had previously refused therapeutic doses.

Pneumonia patients were recruited from May to October 2011 (Royal Free Hospital). Twelve were recruited of whom four were excluded (normal chest radiograph, wrong diagnosis or pneumonia resolved) (Figure 7). Lung cancer patients were recruited from June 2010 to May 2011 (Lyon, France). Eight primary lung cancer patients were recruited (Figure 7).

Healthy controls were recruited from April 2009 to November 2011 (NIMR and Royal Free Hospital) (Figure 8). Forty five were recruited of whom seven were excluded due to positive Quantiferon ELISA results.

Study Recruitment for the Test Set

Thirty two sarcoidosis patients were recruited (St Mary's Hospital, Barnet Hospital and Royal Free Hospital) (Figure 9). Six were excluded, one had a co-existing respiratory disease, three had a normal or no biopsy and two were in remission. Thirteen TB patients were recruited (Royal Free Hospital and Lyon, France). One was excluded due to a negative *M.tuberculosis* culture. Ten pneumonia patients were recruited (Royal Free Hospital). Three were excluded, two had no radiological evidence of pneumonia and one was given the wrong diagnosis. Eight lung cancer patients were recruited (Lyon, France). Sixty one healthy controls were recruited (NIMR and Royal Free Hospital) (Figure 10). Eight were excluded due to positive Quantiferon ELISA results.

Study Recruitment for the Validation Set

Twenty six sarcoidosis patients were recruited (St Mary's Hospital, Barnet Hospital, Royal Free Hospital and Paris, France) (Figure 11). Two were excluded, one had a co-existing respiratory disease and one did not have up-to-date abnormal thoracic radiology. Ten TB patients were recruited (Royal Free Hospital and Lyon, France). One was excluded due to the subsequent diagnosis of an additional disease. Twenty seven healthy controls were recruited (NIMR and Royal Free Hospital). Two were excluded due to positive Quantiferon ELISA results.

RNA integrity and quality control

Inadequate RNA integrity is a common cause of technical difficulties that result in inaccurate interpretation of microarray data. This is usually measured using a combination of different features of the electrophorectic trace of the RNA such that an algorithm generates a numerical classification of the RNA from 1-10 where 10 correlates with the highest integrity. In this study the total RNA had an average RNA integrity number (RIN) of 7.9 and globin-reduced RNA had an average RIN of 7.9. RNA quality was therefore sufficient for microarray processing using Illumina technology at the core facility at Baylor Institute of Immunological Research. Typically RIN above 7 is adequate for microarray, less than 7 may need further validation. In total there were 13 samples (5% of total 243 samples) that were technical outliers by both PCA and unsupervised hierarchical clustering and which were therefore excluded from the analysis. As studies do not have to publish any data on samples that were excluded prior to data analysis it is difficult to gauge how many outliers you may expect from a certain sample size.

Demographics of the training set

Sixteen TB patients, twenty five sarcoidosis, eight pneumonia, eight cancer and thirty eight controls were included in the final training set data analysis (Table 6). Gender, age and ethnicity were well matched between the sarcoidosis, TB and controls (mean age was within 10 years different, percentage of males was within 15% different and percentage of ethnicities was within 20% different across the three cohorts). The pneumonia and cancer patients were slightly older, more male and predominantly white as would be expected in the UK and France. Hence healthy controls were matched to the patient's demographics to the best of our ability. Many more controls were recruited than needed so that each disease group could have its own control group with near-perfect matching for the individual disease-control analysis (see Chapter 5 functional analysis).

Demographics of the test and validation set were similar to the test set

The gender and ethnicity matching of the TB patients in the test set was not quite as good as in the training set due to the lower number of white TB patients (Table 7). However the matching for the other diseases and controls was reasonably good (percentage of males was within 27% different and percentage of ethnicities was within 37% different across the three cohorts). Again the pneumonia and cancer patients were predominantly white patients. The validation set only contains TB, sarcoidosis and controls, therefore it was possible to obtain a good match for gender, ethnicity and age (Table 8).

Clinical characteristics of the training set

Eleven of the sixteen TB patients were smear positive (Table 9). The patients all presented with typical symptoms: TB patients had cough, sweats and weight loss, sarcoidosis patients had cough, dyspnoea and fatigue, the pneumonia patients had cough, dyspnoea and fevers, and the cancer patients had cough, dyspnoea and weight loss (Tables 9-11). The TB, pneumonia and cancer patients had a significantly higher number of symptoms than the sarcoidosis patients (2.5 compared to 1, *t*-test $p<0.05$)

(Tables 9-11). The sarcoidosis patients commonly presented with no symptoms (57%) but symptoms were more likely to be chronic with a far longer average presentation compared to the other respiratory diseases (4.5 years). The serum inflammatory marker C-reactive protein and the peripheral neutrophil counts were on average much higher in TB, pneumonia and cancer than in the sarcoidosis patients, with the highest values in the pneumonia patients (CRP & neutrophil count respectively: TB 77mg/L & 6x10^9/L; sarcoidosis 6mg/L & 3.7x10^9/L; pneumonia 270mg/L & 12.3x10^9/L; cancer 63mg/L & 3.7x10^9/L) (Tables 9-11). Sarcoidosis had the largest proportion of patients with a lymphocyte count below 1x10^9/L. Healthy controls had an average lymphocyte count of 1.9 x 10^9/L and neutrophil count of 2.9 x 10^9/L.

The majority of TB patients presented with a density on their chest radiograph with 8 patients presenting with cavities (Tables 9-11, Figure 12). Most sarcoidosis patients had thoracic lymph node enlargement; 58% had stage I, 23% had stage II, 12% had stage 3 and 8% had stage IV. Pneumonia patients all presented with consolidation, of which one third had definite multilobar consolidation (Table 10). The radiological feature of consolidation is due to filling of the alveolar spaces and can occupy varying distributions; multilobar refers to the consolidation affecting more than one lung lobe. Cancer patients had a mean staging value between Stage 3a and 3b disease, therefore their radiological results suggest on average their disease had a relatively poor prognosis and currently involved the lung parenchyma and local lymph nodes, but had not yet resulted in distant metastases.

Clinical characteristics of the test and validation sets were similar to the training set

The clinical characteristics of the test and validation sets were not significantly different from the training set, except for the number of smear positive TB patients in the training set, which was significantly more than in the test and validation sets (*p* <0.05) (Tables 12-15). In addition the lung cancer patients had more severe disease in the test set than in the training set, with more patients having a higher stage of disease which is consistent with the presence of distant metastases (Stage 4) on their radiology imaging (Fishers exact *p*<0.05) (Table 15).

Unsupervised analysis of the training set revealed differences between the controls and disease cohorts

To determine if the transcriptional profiles from patients with different respiratory diseases are similar or distinct we applied an unsupervised analysis approach followed by unsupervised hierarchical clustering of the transcriptional profiles. RNA was first extracted from whole blood and then processed for microarray. After quality control, an unsupervised analysis approach was applied using GeneSpring 11 to sixteen TB, twenty-five sarcoidosis, eight pneumonia and eight cancer patients' gene expression profiles. 3,422 transcripts satisfied the detection filter (non-parametric Illumina specific method *p*<0.01 compared to background) and the expression filter (two-fold change from the median) (Figure 13). Unsupervised hierarchical clustering of the 3,422 transcripts and samples revealed two main clusters in the horizontal dendrogram, as demonstrated by the addition of the dotted line on the heatmap (Figure 13). One of the clusters contained nearly all the control samples. The other cluster contained nearly all the patient samples. In the main cluster containing most of the patients' transcriptional profiles, the transcriptional profiles from the pneumonia and cancer patients clustered together and the transcriptional profiles from the TB and sarcoidosis patients cluster together. Notably the sarcoidosis profiles were far more heterogeneous than any of the other diseases. Therefore the controls' transcriptional profiles were very similar to each other, the TB and sarcoidosis patients' transcriptional profiles were very similar to each other and the pneumonia and cancer patients' transcriptional profiles were also very similar to each other.

Unsupervised analysis & statistical filtering of the training set accentuated the differences between disease profiles

Next an additional statistical filter was applied to the analysis to reduce the transcript list to a more specific set of genes. The statistical filter Kruskal Wallis with Benjamini Hochberg correction (p<0.01) was applied. This analysis generated 1,446 differentially expressed transcripts (Figure 14). Unsupervised hierarchical clustering of the 1,466 transcripts and samples again revealed two main clusters; one contained most of the control profiles and the other contained most of the patients' profiles. The patients cluster could be further split into two sub-clusters, one contained several sarcoidosis patients and the other contained a branch with the pneumonia and cancer profiles and a branch with the TB and sarcoidosis profiles. Therefore the unsupervised analysis with an additional statistical filter demonstrated a comparable clustering pattern of the patients and controls as seen by the unsupervised analysis alone.

Validating the clustering in the test set

To validate the findings observed with the training set an independent cohort of patients were recruited and processed for microarray, the test set. The 1,446 transcripts derived from the training set (see figure 14) were then applied to the test set and again unsupervised hierarchical clustering of the test set transcripts and samples was performed (Figure 15). The same clustering pattern was seen as observed in the training set. The transcriptional profiles of the controls clustered away from the transcriptional profiles of the patients. Within the main cluster of patients' transcriptional profiles the pneumonia and cancer profiles tightly clustered away from the TB and sarcoidosis profiles (Figure 15). Again the transcriptional profiles from the sarcoidosis patients displayed the largest spread.

To further validate the observations from the 1,446 transcripts applied to both the training and the test set, the same analysis approach that was applied to the training set (identical unsupervised analysis and statistical analysis), was then applied to the test set. This analysis resulted in 1,070 differentially expressed transcripts (Figure 16). Unsupervised hierarchical clustering was then performed using the 1,070 transcripts in the test set (Figure 16). The clustering again demonstrated a similar pattern as was observed with the 1,466 transcripts in the training set and in the test set. The TB and sarcoidosis samples clustered together but separately from the cluster of cancer and pneumonia samples.

A Venn diagram was then used to compare the two sets of transcripts, 1,466 from the training set and 1,070 from the test set. The Venn diagram demonstrated that a large proportion of the transcripts were similar between the two transcript lists (Figure 17). The variance in the actual number of transcripts obtained by the same analysis (1,466 versus 1,077) was most likely related to the unsupervised analysis as this is based on a fold change cut-off around the median, where any change in the median could affect the outcome. Because the training and test set contained different numbers of samples, with non-identical numbers of control samples and disease samples in each group, it would be anticipated that the median for each transcript would not be the same in the training set and test set.

Neither gender nor ethnicity appear to influence the clustering

To rule out any impact of ethnicity or gender in the clustering, the 1,446 transcripts were applied again to the training set and subjected to unsupervised hierarchical clustering, with only white patients (Figure 18a) or male patients (Figure 18b) included in the clustering. The clustering again demonstrated very similar clustering to that seen when all samples were included, demonstrating neither ethnicity nor gender had a significant affect. This was particularly important to determine as both the pneumonia and cancer patients were dominated by white males.

Weighted molecular distance to health used to reflect disease activity

Weighted molecular distance to health (MDTH) is an algorithm used to determine the degree of perturbation of expression of a group of samples compared to a set of controls (Pankla, Buddhisa et al. 2009). Berry *et al* 2010 demonstrated a significant correlation between MDTH and the severity of active TB (Berry, Graham et al. 2010), therefore we applied the algorithm in this study as a surrogate marker of disease activity. MDTH of each disease group in both the training and test set revealed the highest average score in the pneumonia patients, followed by the TB patients (Figure 19). The sarcoidosis and cancer patients displayed scores more towards the level of the controls (Figure 19). The MDTH scores could be thought of as matching the clinical presentation of these diseases, where the cancer and sarcoidosis patients had a more chronic illness with lower inflammatory blood markers than pneumonia or TB (Tables 9-13).

Ingenuity pathway analysis of the 1,446 transcripts in the training set

IPA (Ingenuity Systems) is a database of genes that are associated with particular biological and functional pathways, created by manually abstracting and curating the biomedical literature. In our study only gene expression canonical pathway analysis was carried out. IPA of the three main gene clusters in the 1,446 transcripts revealed distinct functional pathways associated with the different disease groups (Figure 20). The cancer and pneumonia patients had a relative over-abundance of genes (at the top of the heatmap) associated with multiple signalling pathways associated with inflammation e.g. IL-8, TLR and IL-10 signalling pathways; whereas most of the TB and sarcoidosis patients had an under-abundance of these pathways. TB and most sarcoidosis profiles instead were associated with over-abundance of the IFN-signalling and other immune pathways (Figure 21), of which many of the genes within the immune pathways were also IFN-inducible. All the diseases, particularly the pneumonia and TB patients, had a relative under-abundance of many T and B cell related pathways (Figure 20).

To determine if the over-abundant transcripts within the IFN-signalling pathway were related to Type I or Type II IFN's the transcripts were overlaid on the IPA IFN-signalling pathway (Figure 21). This demonstrated that both Type I and Type II IFN-inducible genes were over-abundant. Furthermore when examining the annotation of each transcript on the heatmap in the middle section of the 1,446 transcripts, which were over-abundant in the TB and most sarcoidosis patients, the IFN-inducible genes were clearly the highest number of transcripts present e.g. STAT1, STAT2, IRF7, OAS1, GBP5, IFIT3, MX1, CXCL10 (Figures 22a-c). In contrast when inspecting the band of genes that were highly over-abundant in most of the pneumonia patients, these were clearly dominated by neutrophil anti-microbial genes e.g. CAMP, DEFA4, DEFA1, ELANE, BPI, MPO (Figure 23). These genes were also over-abundant in some of the TB patients but not to the same extent.

Therefore by applying the IPA analysis and by close inspection of the transcripts in each gene cluster, it could be observed that TB and sarcoidosis were associated with an over-abundance of the IFN-signalling and other immune pathways (Figure 20-22), the pneumonia and cancer patients were associated with an over-abundance of many signalling pathways related to inflammation (Figure 20), and all the diseases were associated with under-abundance of T and B cell pathways (Figure 20).

k-means clustering of the 1,446 transcripts in the training set

To support the findings from the hierarchical clustering another common form of clustering, *k*-means clustering, was also applied. *k*-means clustering allows the user more control over the clusters but is therefore a more biased form of clustering. The training set 1,446 transcripts were divided into 10 clusters based on an iterative process related to the number of transcript clusters on the vertical dendrogram of the hierarchical heatmap (as seen in Figure 14). The expression profiles within each of the 10 clusters were shown to be associated with particular disease(s) relative to the control group (Figure 24). Clusters 1 and 6 were over-abundant in TB and less so in sarcoidosis, both were associated with IPA pathways related to the IFN genes such as interferon signalling, the antigen presentation pathway and the role of PRRs in bacteria and viruses (Figure 24). Clusters 2 and 3 were under-abundant in most diseases and associated with T and B cell IPA pathways, such as TCR signalling and CTLA4 signalling in cytotoxic T lymphocytes. Clusters 5 and 7 were over-abundant in pneumonia and associated with IPA signalling pathways such as p38 MAPK signalling and HIF1α signalling. Clusters 8-11 did not reveal any significant association with IPA pathways. The *k*-means clustering therefore showed robust correlation with the hierarchical clustering.

Discussion

The main objective of this prospective observational case-control study was to discover the differences and similarities of the underlying immunopathogenesis of the granulomatous diseases pulmonary TB (caused by *M. tuberculosis* infection) and pulmonary sarcoidosis (cause unknown), the respiratory infectious disease community acquired pneumonia (typically caused by bacteria) and the inflammatory respiratory disease primary lung cancer (predominantly caused by smoking). Furthermore molecular characterisation of these diseases could potentially be used to aid in the common difficulties associated with their clinical management. For example, due to the similarity of the initial clinical presentation of pulmonary TB and pulmonary sarcoidosis these two diseases are often difficult to differentiate diagnostically. In addition sarcoidosis is a clinically heterogeneous disease with a lack of robust clinical phenotyping, which results in less effective clinical decision making. To achieve a comprehensive molecular knowledge and comparison of these diseases a 'bottom-up' approach was taken by applying microarray technology, a broad unbiased gene expression survey, to peripheral whole blood of patients with these four diseases as compared to matched healthy individuals. Unsupervised analysis and clustering of the transcriptional profiles demonstrated that TB and sarcoidosis revealed very similar transcriptional profiles, which differed from the similar transcriptional profiles from the pneumonia and cancer patients. The TB and sarcoidosis profiles were found to be significantly associated with IFN-inducible genes, while the pneumonia and cancer profiles were significantly associated with signalling pathways associated with inflammation. These transcriptional signatures could assist in our understanding of the underlying disease mechanisms and have potential as diagnostic biomarkers. This holistic genomic approach is in keeping with the objectives from the recent NHLBI workshop ‘Genomic Medicine and Lung Disease’ (Center, Schwartz et al. 2012).

Similarities and differences of the clinical features of the four respiratory diseases

All four disease groups presented with very similar respiratory symptoms of cough and dyspnoea (breathlessness), and frequently the same systemic features such as weight loss and fevers (Tables 9-11). In addition their radiological features were often indistinguishable (Tables 9-11, Figure 12). However some variables were different between the diseases, including the length of time their symptoms were present for and their blood inflammation markers, C-reactive protein (CRP) and differential cell blood counts. Theses clinical tests were chosen as comparators across all four diseases as they are routinely measured in most British (and French) hospitals. CRP, an acute phase protein, is consistently used as a measurement of inflammation in many acute and chronic conditions including infection, cancer, sarcoidosis, cardiovascular disease and rheumatic disorders (Windgassen, Funtowicz et al. 2011) (Drent, Wirnsberger et al. 1999). However, CRP is both nonspecific and insensitive, therefore greatly limiting its use as a biomarker particularly in TB, sarcoidosis and cancer (Oremek, Sauer-Eppel et al. 2007; Walzl, Ronacher et al. 2008). For this study a combination of the known diagnosis, serum CRP level, neutrophil count, lymphocyte count and length of illness possibly provides an indication to some differences between these diseases but certainly does not negate the need for a comprehensive characterisation of the disease specific host immune responses.

Comparing the two infectious diseases pulmonary TB and pneumonia, TB patients typically had a more prolonged illness than the pneumonia patients, but a significantly lower CRP and neutrophil count (p<0.05, data not shown) (Tables 9, 10, 11 & 14). This is in agreement with a previous study addressing CRP and procalcitonin levels in TB and pneumonia patients from a low TB burden area (Kang, Kwon et al. 2009). When comparing the granulomatous diseases, the sarcoidosis patients often had far more prolonged but mild symptoms compared to the TB patients, with CRP and neutrophil counts within the normal limits (Tables 9, 11 & 12). In addition their radiological appearance tended to involve the thoracic lymph nodes rather than the lung parenchyma, whereas the lung parenchyma was typically involved in the TB patients. Notably the sarcoidosis patients were clinically heterogeneous in comparison with TB and the other diseases, with some sarcoidosis patients presenting with an analogous clinical phenotype to the TB patients, while others were asymptomatic and diagnosis only discovered due to investigation for another reason. In fact the two granulomatous diseases are so well recognised to be alike that it is often suggested that they are the same disease irrespective of their possible aetiological link (Gupta, Agarwal et al. 2012). The cancer patients also had elevated CRP and neutrophil counts, which could represent an added infection due to the relative immunosuppression from the cancer, or could be secondary to the disease itself (Oremek, Sauer-Eppel et al. 2007).

Distinct clustering pattern of the diseases are observed in the training set and validated in the test set

By applying an unbiased approach, with unsupervised analysis and unsupervised hierarchical clustering, it could be seen there was a distinction in the clustering of the different disease groups (Figure 13). The control profiles clustered separately from the patient profiles, and within the patients the TB and sarcoidosis profiles clustered together but separately from the pneumonia and cancer profiles. The transcript list (3,422 transcripts) derived completely from unsupervised analysis was further refined by adding a statistical filter to the analysis (1,446 transcripts, Figure 14). After unsupervised hierarchical clustering the same pattern was seen as before but with a sharper contrast in the sub-clusters visualised with the aid of the horizontal dendrogram (Figure 14). Furthermore it was validated in the test set by two different strategies, firstly to use the same 1,446 transcripts for the clustering (Figure 15), and secondly to use the same analytical approach to derive a different transcript list (1,070 transcripts) from the test set prior to the same clustering algorithm (Figure 16). Both these strategies resulted in similar clustering of the disease groups and healthy controls. The large crossover in overlapping genes (704 transcripts) between the training and test set further strengthens the robustness of these results (Figure 17).

The similarity of TB and sarcoidosis patients' whole blood gene expression profiles has been demonstrated in previous studies (Koth, Solberg et al. 2011; Maertzdorf, Weiner et al. 2012). However this is the first study to compare TB, sarcoidosis and other similar respiratory diseases. In addition the two earlier studies did not comment on the diversity of their sarcoidosis profiles, which is evident in all three cohorts (training, test and validation) in this study. The distinct clustering of the pneumonia and cancer profiles together is in part related to their dissimilarity from the granulomatous diseases and controls, not just purely their apparent similarity towards each other. Because the ethnicity and gender of the training set pneumonia and cancer patients were different from the other patients this could have prejudiced the clustering pattern, as it is well established that gender and ethnicity can both affect gene expression profiles (Whitney, Diehn et al. 2003; Eady, Wortley et al. 2005). Therefore to determine that gender and ethnicity were not biasing the results the clustering of the 1,446 transcripts was repeated with only white patients/controls and only males. Both heatmaps clearly demonstrated the same clustering pattern as before, suggesting any influence ethnicity or gender may have on the gene expression is outweighed by the influence of the diseases processes (Figure 18).

Weighted molecular distance to health can be used to reflect disease activity

The weighted molecular distance to health (MDTH) is a quantification of the transcriptional perturbation between one set of samples and a set of controls (Pankla, Buddhisa et al. 2009). Berry *et al.* 2010, demonstrated a strong correlation between the MDTH and the degree of pulmonary TB activity as evidenced by their extent of radiological disease (Berry, Graham et al. 2010). MDTH could therefore have a potential role as a surrogate marker of disease activity in TB. Following on from this concept, MDTH was also shown to be strongly correlated with the clinical phenotype of another infectious disease, HTLV-1. Infected patients who exhibited no clinical indication of infection had a significantly lower MDTH than infected patients who had consequently developed a neurodegenerative disorder (Tattermusch, Skinner et al. 2012). To this end the MDTH algorithm was applied to the four disease groups and demonstrated that pneumonia and TB had the highest scores, while cancer and sarcoidosis had the lowest scores. This is perhaps in keeping with the likely clinical presentation as the pneumonia and TB patients had higher serum inflammatory markers and larger number of symptoms than the sarcoidosis patients. However the cancer patients also had high serum inflammatory markers and number of symptoms but did not have such a high MDTH score. Therefore already this potential surrogate marker for disease activity is informing us more than we are able to delineate from standard clinical data and tests. On the other hand this may reflect a lack of appropriate clinical data obtained from the patients, as pneumonia and TB patients can often appear generally more unwell than lung cancer patients on initial presentation.

Distinct functional pathways derived from the 1,446 transcripts were associated with different diseases

The 1,446 transcripts were clustered into three main clusters of genes (Figure 20). The middle cluster contained genes that were relatively over-abundant in the TB patients and most of the sarcoidosis patients. The pathway most significantly associated with these genes was the IFN-signalling pathway; the next most significant ones were immune pathways containing a high number of IFN-inducible genes (Figures 20 & 21). The full list of transcripts within the middle cluster is revealed (Figures 22a-c), and as would be expected the list of genes is dominated by IFN-inducible genes, particularly the most over-abundant genes. These findings are in keeping with two earlier publications that identified the similarity of TB and sarcoidosis profiles, as they too found both diseases induce a large number of over-abundant IFN-inducible genes (Koth, Solberg et al. 2011; Maertzdorf, Weiner et al. 2012). The Koth et al. 2011, study compared sarcoidosis patients they had recruited to the data publicly available from other studies including the Berry *et al*, Lockstone *et al* and Rosenbaum *et al* studies (Rosenbaum, Pasadhika et al. 2009; Berry, Graham et al. 2010; Lockstone, Sanderson et al. 2010). Kaufmann *et al* recruited a new small cohort of TB and sarcoidosis patients but did not validate their finding in an independent cohort. Therefore this current study validates and confirms the finding that TB and sarcoidosis have similar transcriptional profiles. Furthermore this study demonstrates this finding with larger patient numbers, additional similar respiratory diseases and additional data mining strategies (see chapter 5).

The top cluster of the 1,446 transcripts contained genes that were relatively over-abundant in most of the pneumonia and cancer patients, and that were associated with IPA pathways related to inflammation. There is a general consensus that inflammation plays a role in cancer, whether it is cause or effect, including primary lung cancer where there is increasing evidence for example that smoking promotes lung inflammation (O'Callaghan, O'Donnell et al. 2010). It has long been recognised that pneumonia resulting from infection causes an acute lung inflammation such that even glucocorticoids, thought to reduce the inflammatory response, may in some cases expedite resolution (Chen, Li et al. 2011). Furthermore the inflammatory response associated with pneumonia is not just compartmentalised to the lungs but has also been described in the blood (Fernandez-Serrano, Dorca et al. 2003).

The bottom cluster of the 1,446 transcripts contained mostly relatively under-abundant genes associated with all the diseases as opposed to their relative over-abundance in the healthy controls. These genes showed significant correlation with many B and T cell IPA pathways. Again this finding is perhaps predictable as reduced numbers of T and B cells in the blood of active TB patients, sarcoidosis patients and bacterial infection have previously been demonstrated by flow cytometry analysis (Ardura, Banchereau et al. 2009; Berry, Graham et al. 2010; Sweiss, Salloum et al. 2010). The cause remains unresolved but could be due to preferential migration of the immune cells to the site of disease or cell death as a consequence of the pathogenesis. The percentage of lymphocytes found in the bronchoalveolar lavage has been shown to be higher in sarcoidosis and active TB patients than in healthy controls, possibly suggesting preferential migration (Hoheisel, Tabak et al. 1994).

A band of highly over-abundant transcripts was observed in the top cluster of the 1,446 transcripts (Figure 23) and was found to be driven by neutrophil-antimicrobial genes e.g. CAMP, LCN2 and DEFA1. The genes were over-abundant in the majority of the pneumonia patients and to some extent in several of the TB patients. A correlation of neutrophil genes and infectious bacterial diseases is not surprising but helps to demonstrate that blood gene expression profiling of these diseases nicely parallels previous understanding of the immune host response.

To confirm the sets of correlated genes derived by unsupervised hierarchical clustering a different form of clustering was applied, *k*-means clustering. 10 clusters were selected and found similar IPA pathways as was determined by the hierarchical clustering. *k*-means clustering found an association with the relative over-abundance of genes in TB and sarcoidosis e.g. the IFN-signalling pathway; over-abundance of signalling pathways in pneumonia and cancer e.g. IL-10 signalling; and an under-abundance of the T and B cell pathways in all the diseases e.g. TCR signalling.

Chapter Summary

An unbiased survey of the human transcriptome revealed distinct clustering of pulmonary TB and pulmonary sarcoidosis expression profiles compared to a distinct clustering of pneumonia and lung cancer expression profiles. TB and sarcoidosis showed a robust similarity of their molecular and functional phenotypes that was dominated by the IFN-inducible genes (further detailed in chapter 5). However TB had a more active transcriptional response than sarcoidosis, reflecting the clinical phenotypes. Pneumonia and lung cancer also showed a robust similarity of their molecular and functional phenotypes, which appeared dominated by inflammation genes (further detailed in chapter 5). The transcriptional profiles from the pneumonia patients and some TB patients were also dominated by neutrophil genes. All the disease groups were associated with an under-abundance of T and B cell pathways. These biological processes are likely explained by a combination of changes in cellular numbers of discrete immune cell populations as well as changes in gene expression in individual cell populations as described in Berry *et al.* 2010. This constitutes the scope of on-going work.

To further determine the different biological pathways underlying each disease it was first necessary to clarify the clinical and transcriptional phenotypes resulting in the heterogeneity of the transcriptional profiles from the sarcoidosis samples. The majority of the transcriptional profiles from the sarcoidosis samples clustered with the TB patients but some of the sarcoidosis profiles clustered with the controls. Understanding this heterogeneity is the focus of chapter 4. Once the disease phenotypes were established the transcriptional profiles could then be used for data mining as described in chapter 5.

Figures for chapter 3

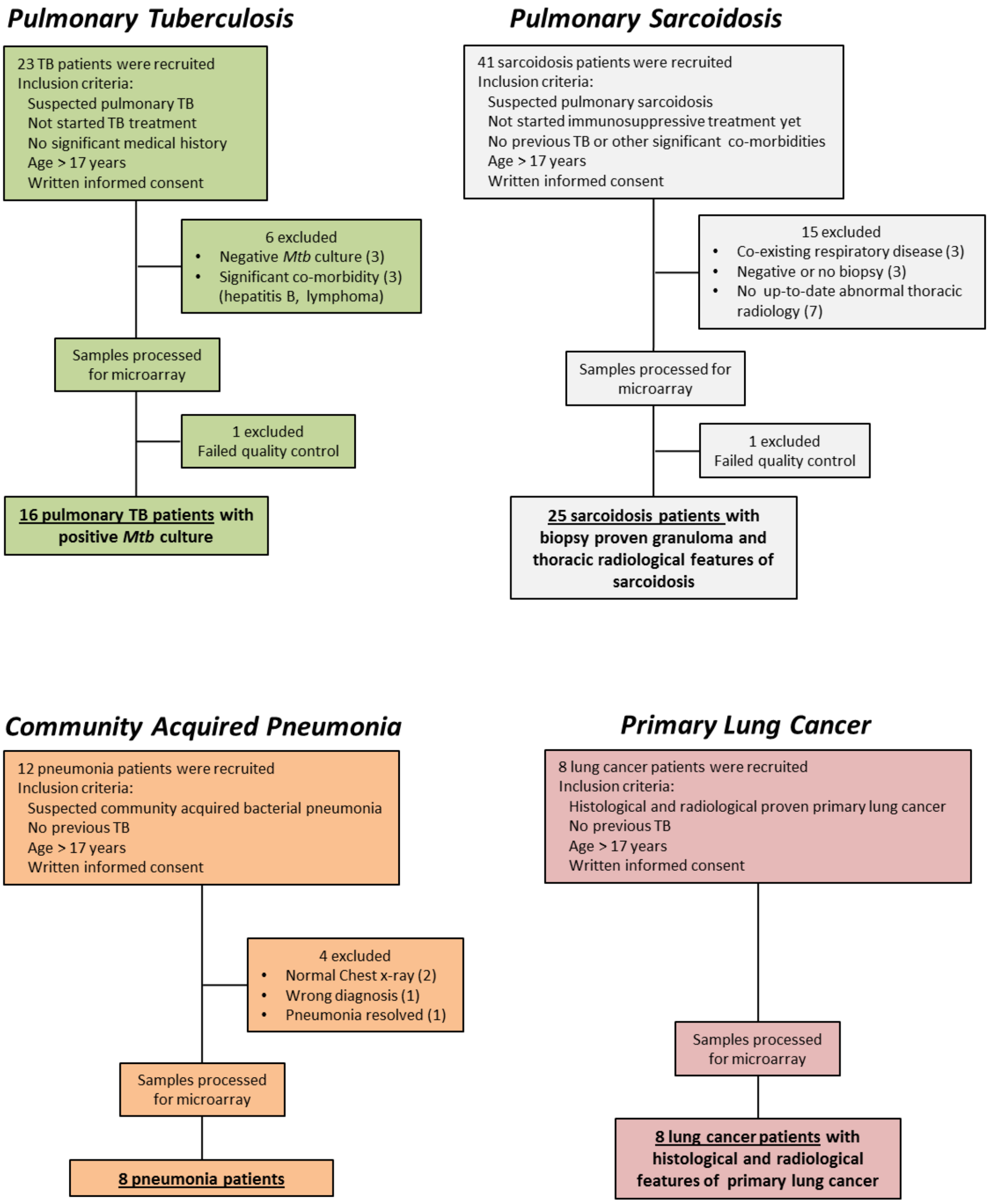

Pulmonary Tuberculosis
23 TB patients were recruited
Inclusion criteria:
Suspected pulmonary TB
Not started TB treatment
No significant medical history
Age > 17 years
Written informed consent
6 excluded
• Negative Mtb culture (3)
• Significant co-morbidity (3) (hepatitis B, lymphoma)
Samples processed for microarray
1 excluded
Failed quality control
16 pulmonary TB patients with positive Mtb culture
Pulmonary Sarcoidosis
41 sarcoidosis patients were recruited
Inclusion criteria:
Suspected pulmonary sarcoidosis
Not started immunosuppressive treatment yet
No previous TB or other significant co-morbidities
Age > 17 years
Written informed consent
15 excluded
• Co-existing respiratory disease (3)
• Negative or no biopsy (3)
• No up-to-date abnormal thoracic radiology (7)
Samples processed for microarray
1 excluded
Failed quality control
25 sarcoidosis patients with biopsy proven granuloma and thoracic radiological features of sarcoidosis
Community Acquired Pneumonia
12 pneumonia patients were recruited
Inclusion criteria:
Suspected community acquired bacterial pneumonia
No previous TB
Age > 17 years
Written informed consent
4 excluded
• Normal Chest x-ray (2)
• Wrong diagnosis (1)
• Pneumonia resolved (1)
Samples processed for microarray
8 pneumonia patients
Primary Lung Cancer
8 lung cancer patients were recruited
Inclusion criteria:
Histological and radiological proven primary lung cancer
No previous TB
Age > 17 years
Written informed consent
Samples processed for microarray
8 lung cancer patients with histological and radiological features of primary lung cancer

Figure 7. Recruitment of respiratory patients for the TRAINING set.

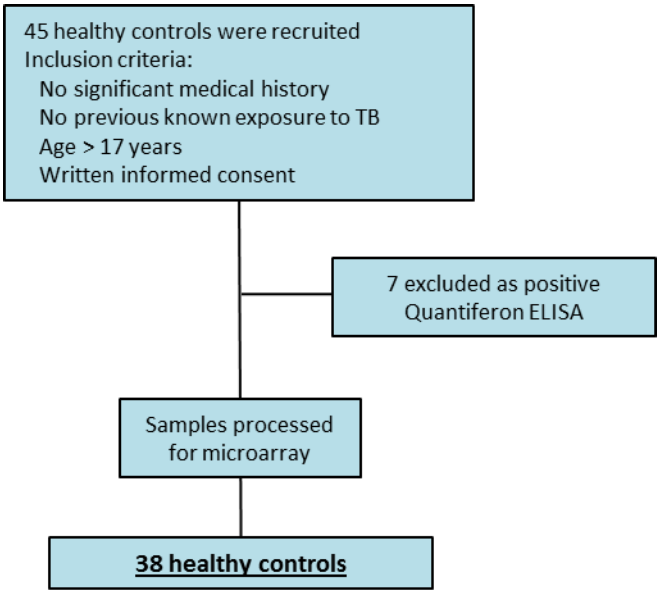

45 healthy controls were recruited
Inclusion criteria:
No significant medical history
No previous known exposure to TB
Age > 17 years
Written informed consent
7 excluded as positive Quantiferon ELISA
Samples processed for microarray
38 healthy controls

Figure 8. Recruitment of healthy controls for the TRAINING set.

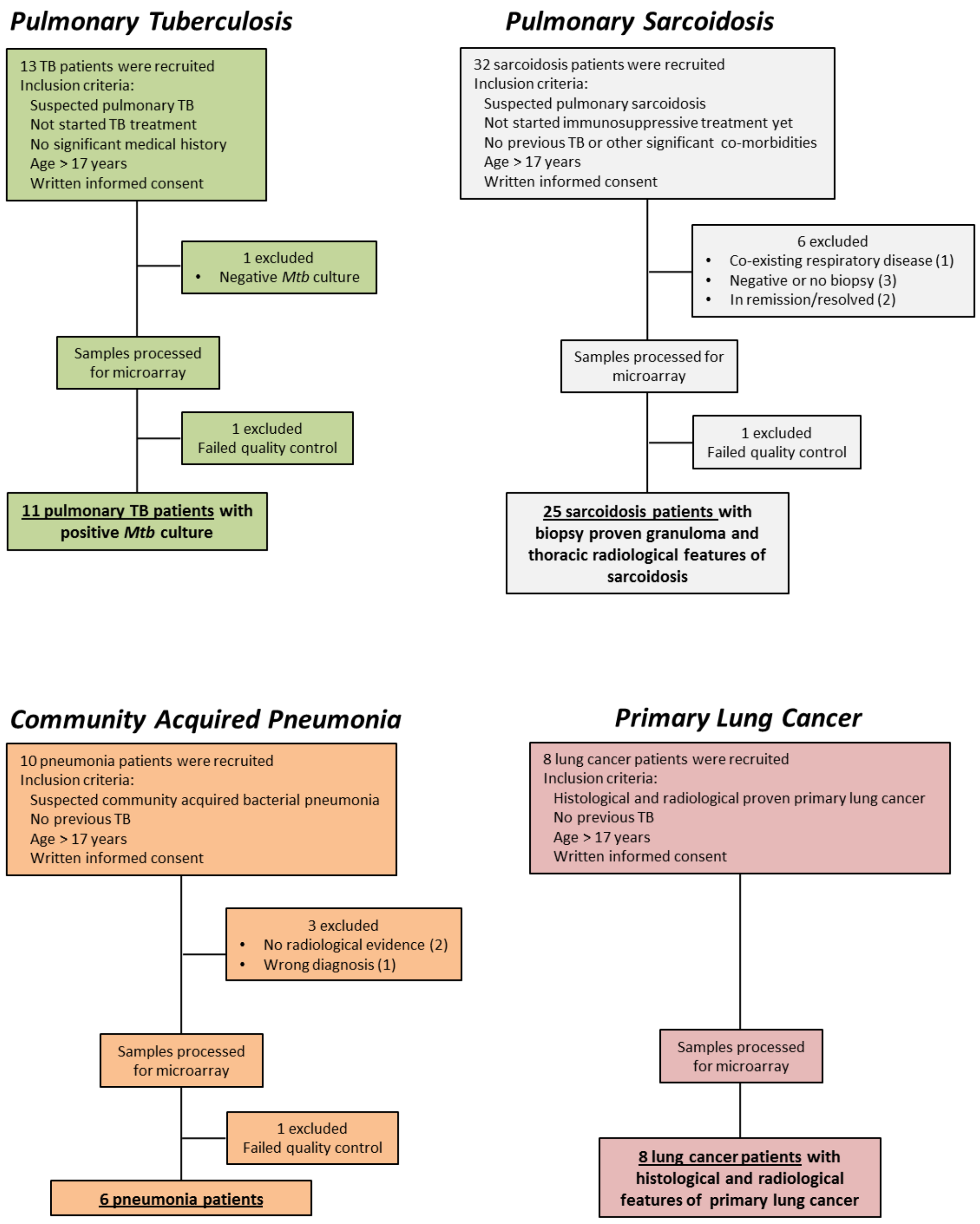

Pulmonary Tuberculosis
13 TB patients were recruited
Inclusion criteria:
Suspected pulmonary TB
Not started TB treatment
No significant medical history
Age > 17 years
Written informed consent
1 excluded
• Negative *Mtb* culture
Samples processed for microarray
1 excluded
Failed quality control
11 pulmonary TB patients with positive *Mtb* culture
Pulmonary Sarcoidosis
32 sarcoidosis patients were recruited
Inclusion criteria:
Suspected pulmonary sarcoidosis
Not started immunosuppressive treatment yet
No previous TB or other significant co-morbidities
Age > 17 years
Written informed consent
6 excluded
• Co-existing respiratory disease (1)
• Negative or no biopsy (3)
• In remission/resolved (2)
Samples processed for microarray
1 excluded
Failed quality control
25 sarcoidosis patients with biopsy proven granuloma and thoracic radiological features of sarcoidosis
Community Acquired Pneumonia
10 pneumonia patients were recruited
Inclusion criteria:
Suspected community acquired bacterial pneumonia
No previous TB
Age > 17 years
Written informed consent
3 excluded
• No radiological evidence (2)
• Wrong diagnosis (1)
Samples processed for microarray
1 excluded
Failed quality control
6 pneumonia patients
Primary Lung Cancer
8 lung cancer patients were recruited
Inclusion criteria:
Histological and radiological proven primary lung cancer
No previous TB
Age > 17 years
Written informed consent
Samples processed for microarray
8 lung cancer patients with histological and radiological features of primary lung cancer

Figure 9. Recruitment of respiratory patients for the TEST set.

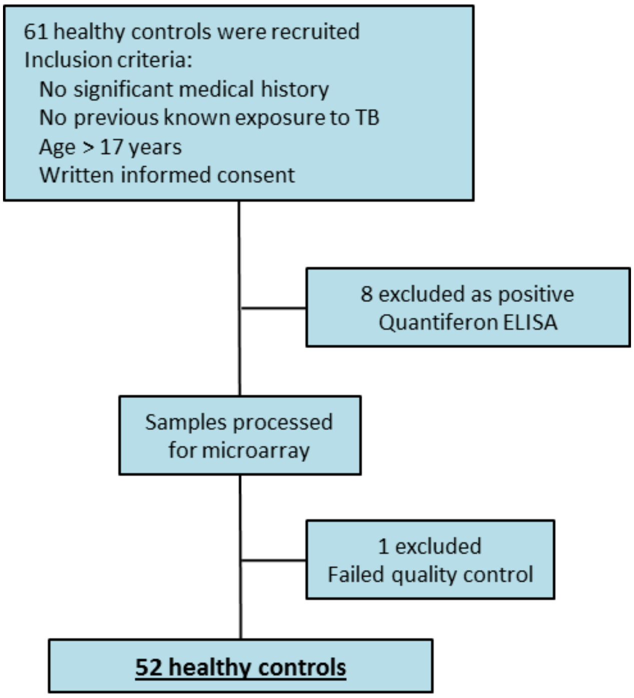

61 healthy controls were recruited
Inclusion criteria:
No significant medical history
No previous known exposure to TB
Age > 17 years
Written informed consent
8 excluded as positive Quantiferon ELISA
Samples processed for microarray
1 excluded Failed quality control
52 healthy controls

Figure 10. Recruitment of healthy controls for the TEST set.

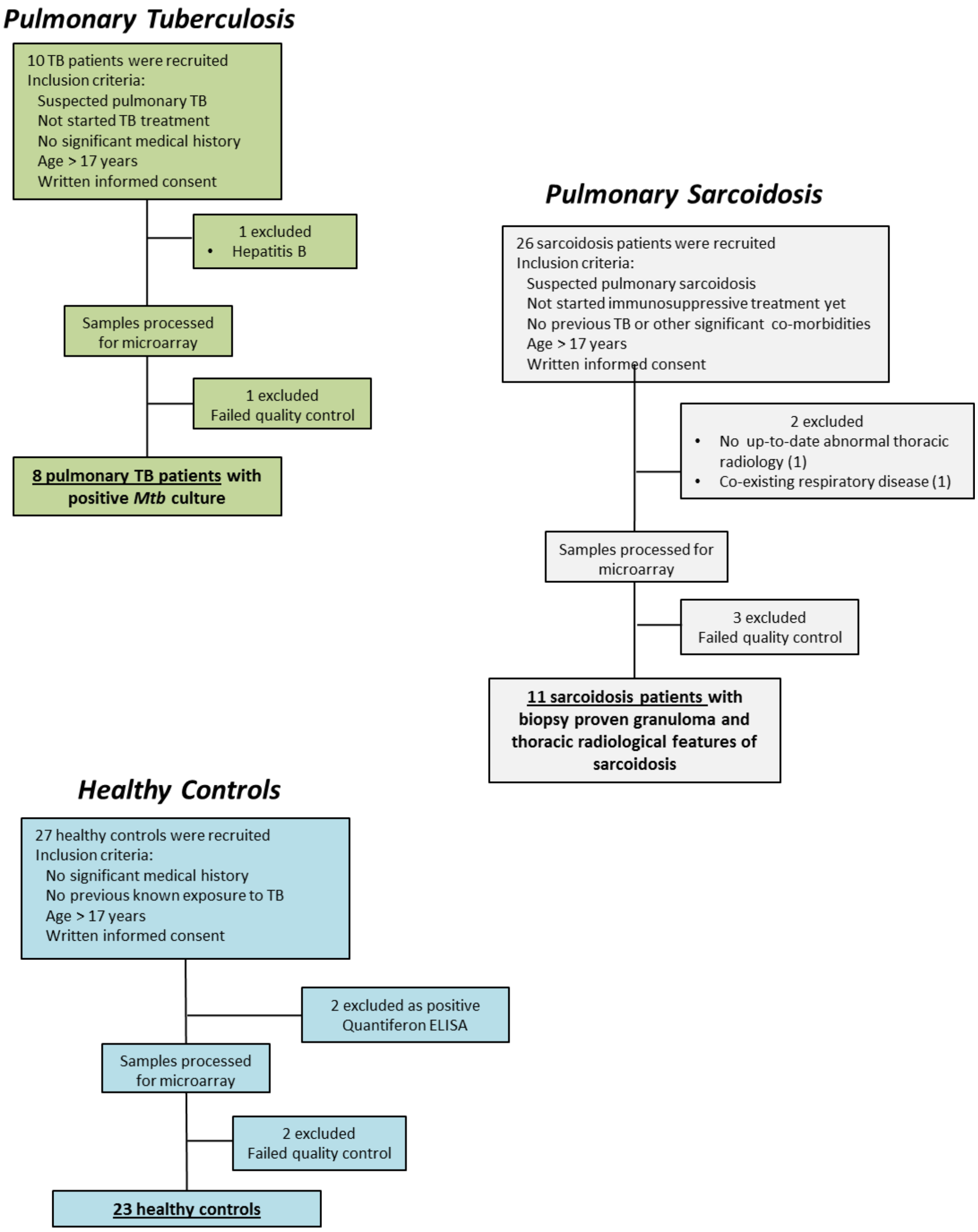
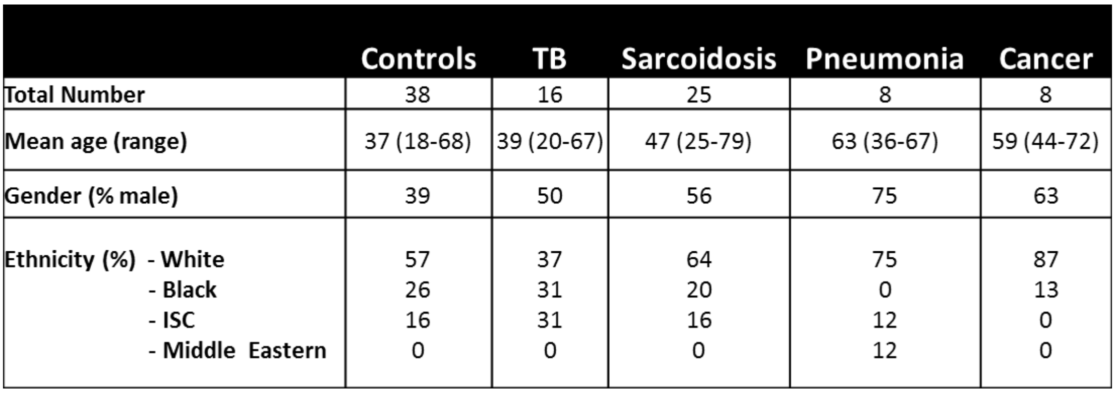
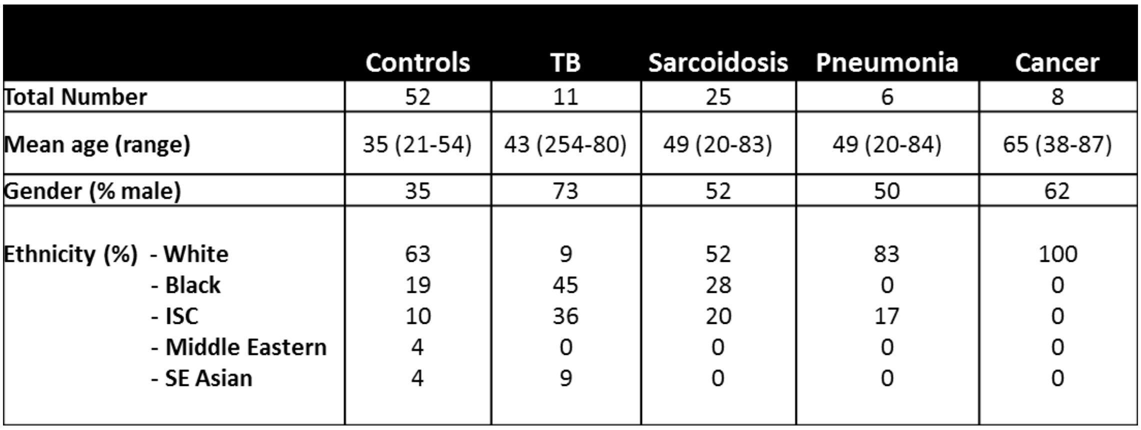

Pulmonary Tuberculosis
10 TB patients were recruited
Inclusion criteria:
Suspected pulmonary TB
Not started TB treatment
No significant medical history
Age > 17 years
Written informed consent
1 excluded
• Hepatitis B
Samples processed for microarray
1 excluded
Failed quality control
8 pulmonary TB patients with positive Mtb culture
Pulmonary Sarcoidosis
26 sarcoidosis patients were recruited
Inclusion criteria:
Suspected pulmonary sarcoidosis
Not started immunosuppressive treatment yet
No previous TB or other significant co-morbidities
Age > 17 years
Written informed consent
2 excluded
• No up-to-date abnormal thoracic radiology (1)
• Co-existing respiratory disease (1)
Samples processed for microarray
3 excluded
Failed quality control
11 sarcoidosis patients with biopsy proven granuloma and thoracic radiological features of sarcoidosis
Healthy Controls
27 healthy controls were recruited
Inclusion criteria:
No significant medical history
No previous known exposure to TB
Age > 17 years
Written informed consent
2 excluded as positive Quantiferon ELISA
Samples processed for microarray
2 excluded
Failed quality control
23 healthy controls

Figure 11. Recruitment of the TB, sarcoidosis and healthy controls for the VALIDATION set.



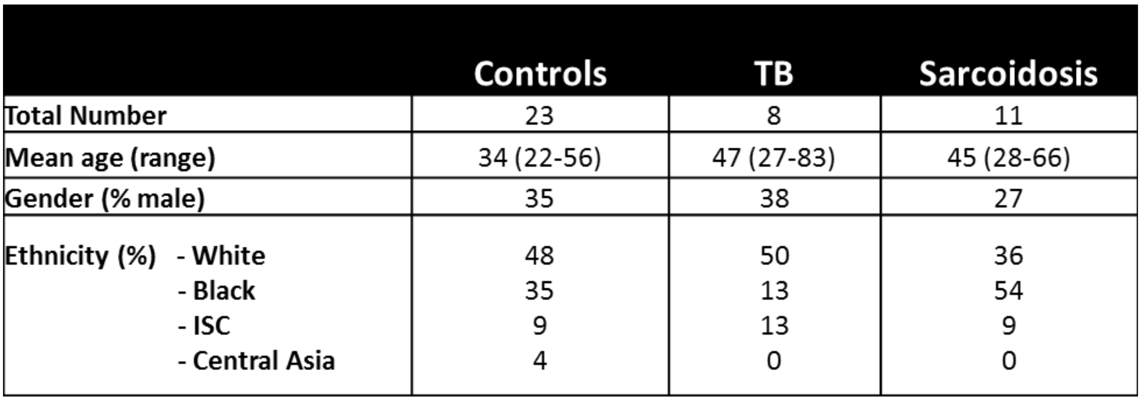
	Controls	TB	Sarcoidosis	Pneumonia	Cancer
Total Number	38	16	25	8	8
Mean age (range)	37 (18-68)	39 (20-67)	47 (25-79)	63 (36-67)	59 (44-72)
Gender (% male)	39	50	56	75	63
Ethnicity (%) - White	57	37	64	75	87
- Black	26	31	20	0	13
- ISC	16	31	16	12	0
- Middle Eastern	0	0	0	12	0

Table 6. Demographics of TRAINING set.



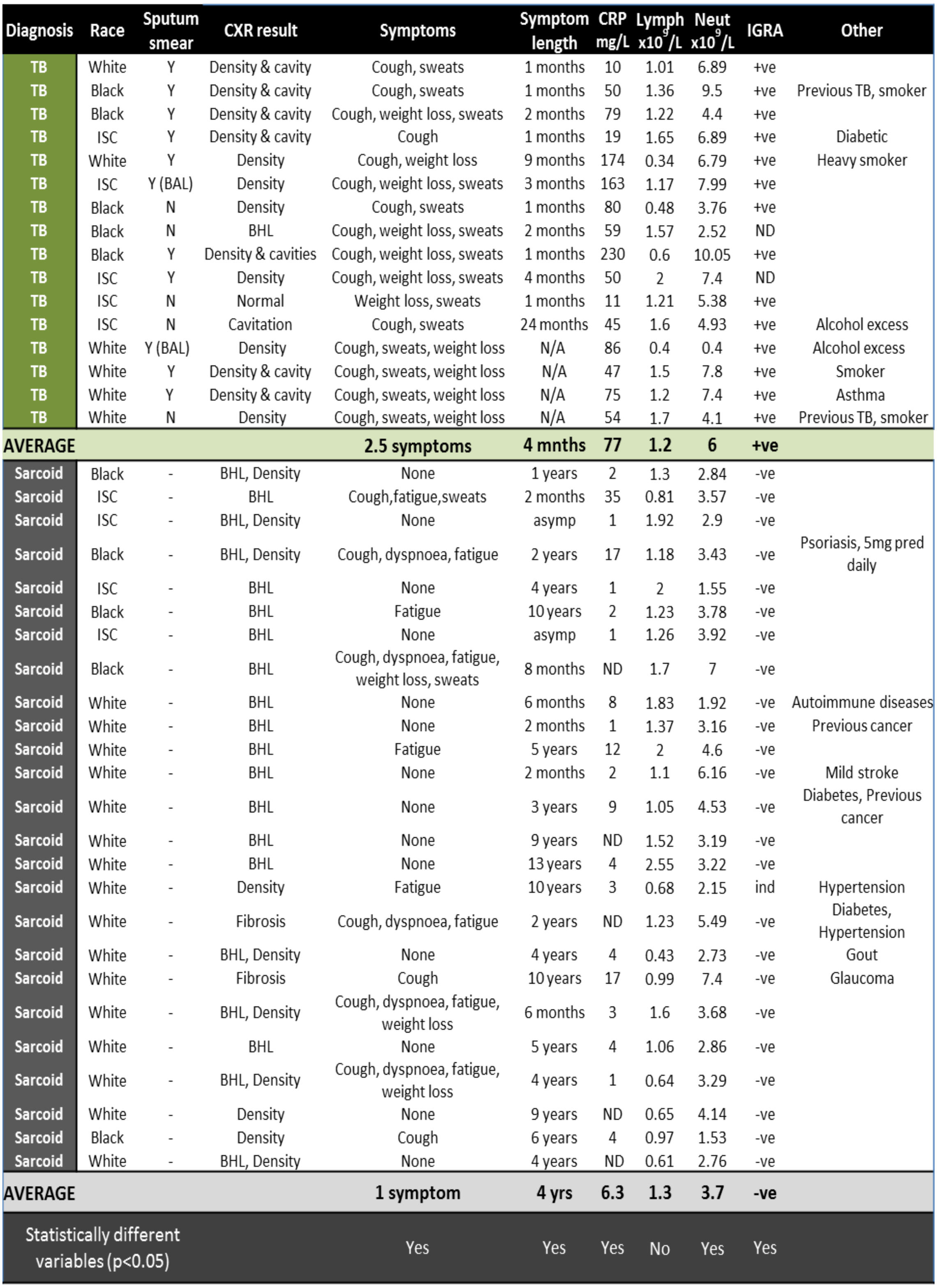
	Controls	TB	Sarcoidosis	Pneumonia	Cancer
Total Number	52	11	25	6	8
Mean age (range)	35 (21-54)	43 (254-80)	49 (20-83)	49 (20-84)	65 (38-87)
Gender (% male)	35	73	52	50	62
Ethnicity (%) - White	63	9	52	83	100
- Black	19	45	28	0	0
- ISC	10	36	20	17	0
- Middle Eastern	4	0	0	0	0
- SE Asian	4	9	0	0	0

Table 7. Demographics of the TEST set.



	Controls	TB	Sarcoidosis
Total Number	23	8	11
Mean age (range)	34 (22-56)	47 (27-83)	45 (28-66)
Gender (% male)	35	38	27
Ethnicity (%) - White	48	50	36
- Black	35	13	54
- ISC	9	13	9
- Central Asia	4	0	0

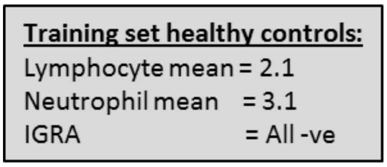
Table 8. Demographics of VALIDATION set.



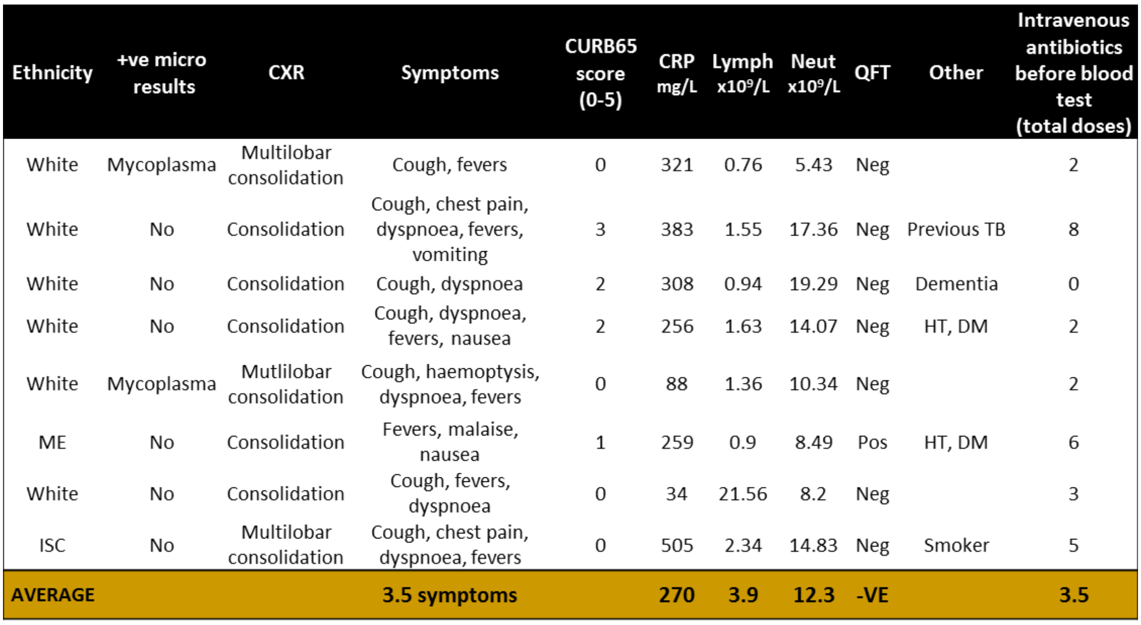
Diagnosis	Race	Sputum smear	CXR result	Symptoms	Symptom length	CRP mg/L	Lymph x10^{9}/L	Neut x10^{9}/L	IGRA	Other
TB	White	Y	Density & cavity	Cough, sweats	1 months	10	1.01	6.89	+ve	
TB	Black	Y	Density & cavity	Cough, sweats	1 months	50	1.36	9.5	+ve	Previous TB, smoker
TB	Black	Y	Density & cavity	Cough, weight loss, sweats	2 months	79	1.22	4.4	+ve	
TB	ISC	Y	Density & cavity	Cough	1 months	19	1.65	6.89	+ve	Diabetic
TB	White	Y	Density	Cough, weight loss	9 months	174	0.34	6.79	+ve	Heavy smoker
TB	ISC	Y (BAL)	Density	Cough, weight loss, sweats	3 months	163	1.17	7.99	+ve	
TB	Black	N	Density	Cough, sweats	1 months	80	0.48	3.76	+ve	
TB	Black	N	BHL	Cough, weight loss, sweats	2 months	59	1.57	2.52	ND	
TB	Black	Y	Density & cavities	Cough, weight loss, sweats	1 months	230	0.6	10.05	+ve	
TB	ISC	Y	Density	Cough, weight loss, sweats	4 months	50	2	7.4	ND	
TB	ISC	N	Normal	Weight loss, sweats	1 months	11	1.21	5.38	+ve	
TB	ISC	N	Cavitation	Cough, sweats	24 months	45	1.6	4.93	+ve	Alcohol excess
TB	White	Y (BAL)	Density	Cough, sweats, weight loss	N/A	86	0.4	0.4	+ve	Alcohol excess
TB	White	Y	Density & cavity	Cough, sweats, weight loss	N/A	47	1.5	7.8	+ve	Smoker
TB	White	Y	Density & cavity	Cough, sweats, weight loss	N/A	75	1.2	7.4	+ve	Asthma
TB	White	N	Density	Cough, sweats, weight loss	N/A	54	1.7	4.1	+ve	Previous TB, smoker
AVERAGE				2.5 symptoms	4 mnths	77	1.2	6	+ve	
Sarcoid	Black	-	BHL, Density	None	1 years	2	1.3	2.84	-ve	
Sarcoid	ISC	-	BHL	Cough,fatigue,sweats	2 months	35	0.81	3.57	-ve	
Sarcoid	ISC	-	BHL, Density	None	asymp	1	1.92	2.9	-ve	
Sarcoid	Black	-	BHL, Density	Cough, dyspnoea, fatigue	2 years	17	1.18	3.43	-ve	Psoriasis, 5mg pred daily
Sarcoid	ISC	-	BHL	None	4 years	1	2	1.55	-ve	
Sarcoid	Black	-	BHL	Fatigue	10 years	2	1.23	3.78	-ve	
Sarcoid	ISC	-	BHL	None	asymp	1	1.26	3.92	-ve	
Sarcoid	Black	-	BHL	Cough, dyspnoea, fatigue, weight loss, sweats	8 months	ND	1.7	7	-ve	
Sarcoid	White	-	BHL	None	6 months	8	1.83	1.92	-ve	Autoimmune diseases
Sarcoid	White	-	BHL	None	2 months	1	1.37	3.16	-ve	Previous cancer
Sarcoid	White	-	BHL	Fatigue	5 years	12	2	4.6	-ve	
Sarcoid	White	-	BHL	None	2 months	2	1.1	6.16	-ve	Mild stroke
Sarcoid	White	-	BHL	None	3 years	9	1.05	4.53	-ve	Diabetes, Previous cancer
Sarcoid	White	-	BHL	None	9 years	ND	1.52	3.19	-ve	
Sarcoid	White	-	BHL	None	13 years	4	2.55	3.22	-ve	
Sarcoid	White	-	Density	Fatigue	10 years	3	0.68	2.15	ind	Hypertension
Sarcoid	White	-	Fibrosis	Cough, dyspnoea, fatigue	2 years	ND	1.23	5.49	-ve	Diabetes, Hypertension
Sarcoid	White	-	BHL, Density	None	4 years	4	0.43	2.73	-ve	Gout
Sarcoid	White	-	Fibrosis	Cough	10 years	17	0.99	7.4	-ve	Glaucoma
Sarcoid	White	-	BHL, Density	Cough, dyspnoea, fatigue, weight loss	6 months	3	1.6	3.68	-ve	
Sarcoid	White	-	BHL	None	5 years	4	1.06	2.86	-ve	
Sarcoid	White	-	BHL, Density	Cough, dyspnoea, fatigue, weight loss	4 years	1	0.64	3.29	-ve	
Sarcoid	White	-	Density	None	9 years	ND	0.65	4.14	-ve	
Sarcoid	Black	-	Density	Cough	6 years	4	0.97	1.53	-ve	
Sarcoid	White	-	BHL, Density	None	4 years	ND	0.61	2.76	-ve	
AVERAGE				1 symptom	4 yrs	6.3	1.3	3.7	-ve	
Statistically different variables (p<0.05)				Yes	Yes	Yes	No	Yes	Yes	

Table 9. Comparable clinical variables of the TB and sarcoidosis patients in the training set.

BAL = bronchoalveolar lavage, IGRA = IFN gama-release assay, Lymph = lymphocyte count, BHL = bilateral hilar lymphadenopathy, Neut = neutrophil count, CXR = chest x-ray, ISC = Indian subcontinent, CRP = C-reactive protein, Ind = indeterminate, ND = not done, N/A = not available, pred = prednisolone, Dyspnoea = breathlessness.

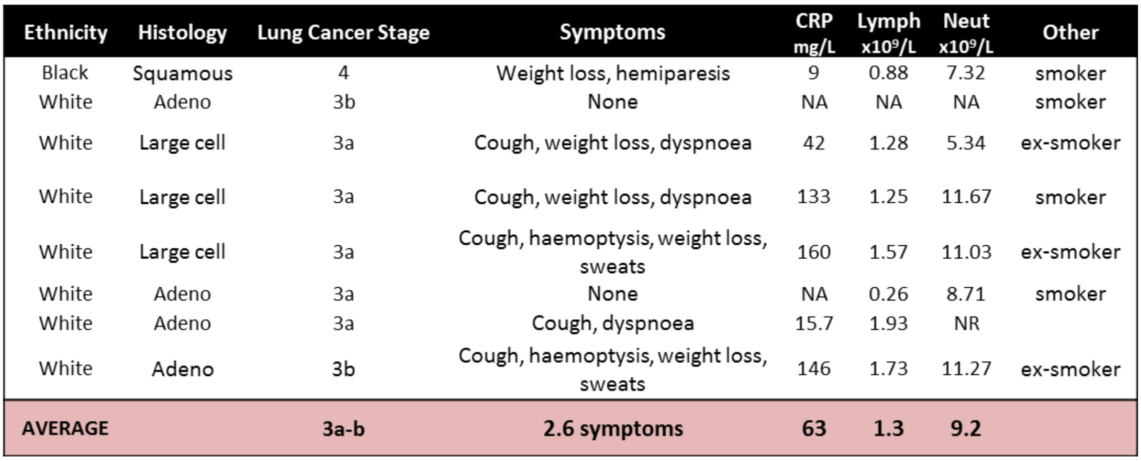


Training set healthy controls:
Lymphocyte mean = 2.1
Neutrophil mean = 3.1
IGRA = All -ve



Ethnicity	+ve micro results	CXR	Symptoms	CURB65 score (0-5)	CRP mg/L	Lymph x10^9/L	Neut x10^9/L	QFT	Other	Intravenous antibiotics before blood test (total doses)
White	Mycoplasma	Multilobar consolidation	Cough, fevers	0	321	0.76	5.43	Neg		2
White	No	Consolidation	Cough, chest pain, dyspnoea, fevers, vomiting	3	383	1.55	17.36	Neg	Previous TB	8
White	No	Consolidation	Cough, dyspnoea	2	308	0.94	19.29	Neg	Dementia	0
White	No	Consolidation	Cough, dyspnoea, fevers, nausea	2	256	1.63	14.07	Neg	HT, DM	2
White	Mycoplasma	Mutlilobar consolidation	Cough, haemoptysis, dyspnoea, fevers	0	88	1.36	10.34	Neg		2
ME	No	Consolidation	Fevers, malaise, nausea	1	259	0.9	8.49	Pos	HT, DM	6
White	No	Consolidation	Cough, fevers, dyspnoea	0	34	21.56	8.2	Neg		3
ISC	No	Multilobar consolidation	Cough, chest pain, dyspnoea, fevers	0	505	2.34	14.83	Neg	Smoker	5
AVERAGE			**3.5 symptoms**		**270**	**3.9**	**12.3**	**-VE**		**3.5**

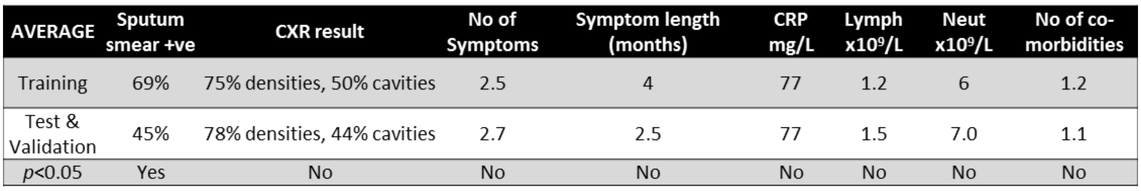
Table 10. Clinical variables of the community acquired pneumonia patients in the training set.



Ethnicity	Histology	Lung Cancer Stage	Symptoms	CRP mg/L	Lymph x10^9/L	Neut x10^9/L	Other
Black	Squamous	4	Weight loss, hemiparesis	9	0.88	7.32	smoker
White	Adeno	3b	None	NA	NA	NA	smoker
White	Large cell	3a	Cough, weight loss, dyspnoea	42	1.28	5.34	ex-smoker
White	Large cell	3a	Cough, weight loss, dyspnoea	133	1.25	11.67	smoker
White	Large cell	3a	Cough, haemoptysis, weight loss, sweats	160	1.57	11.03	ex-smoker
White	Adeno	3a	None	NA	0.26	8.71	smoker
White	Adeno	3a	Cough, dyspnoea	15.7	1.93	NR	
White	Adeno	3b	Cough, haemoptysis, weight loss, sweats	146	1.73	11.27	ex-smoker
AVERAGE		**3a-b**	**2.6 symptoms**	**63**	**1.3**	**9.2**	

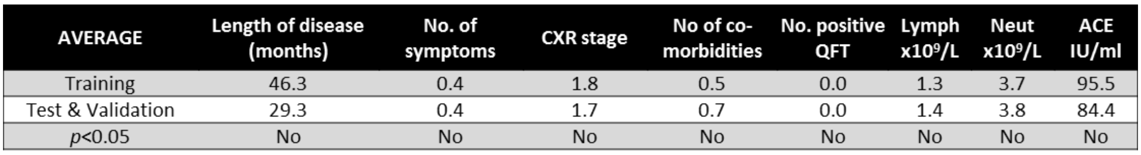
Table 11. Clinical variables of the lung cancer patients in the training set.

Dyspnoea = breathlessness. Haemoptysis = coughing up blood. CURB65 score = pneumonia severity score where 5 is the most severe. HT = hypertension. DM = hypertension. Adeno = adenocarcinoma.CXR = chest x-ray. CRP = C-reactive protein. ME = middle eastern.



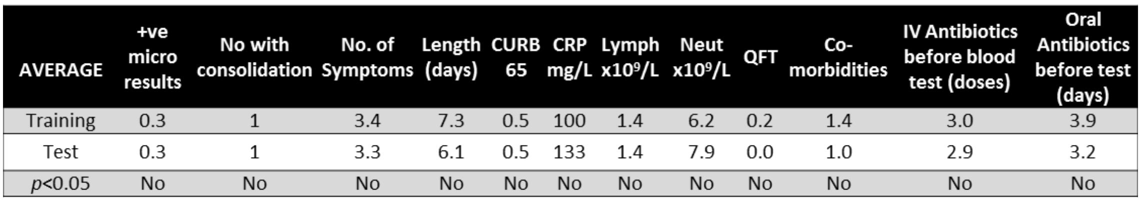
AVERAGE	Sputum smear +ve	CXR result	No of Symptoms	Symptom length (months)	CRP mg/L	Lymph x10⁹/L	Neut x10⁹/L	No of co-morbidities
Training	69%	75% densities, 50% cavities	2.5	4	77	1.2	6	1.2
Test & Validation	45%	78% densities, 44% cavities	2.7	2.5	77	1.5	7.0	1.1
p<0.05	Yes	No	No	No	No	No	No	No

Table 12. Comparing clinical characteristics of all three datasets – TB patients



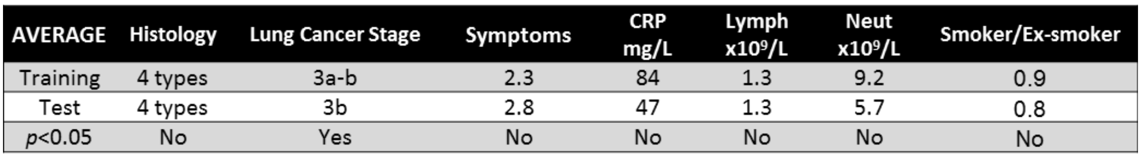
AVERAGE	Length of disease (months)	No. of symptoms	CXR stage	No of co-morbidities	No. positive QFT	Lymph x10⁹/L	Neut x10⁹/L	ACE IU/ml
Training	46.3	0.4	1.8	0.5	0.0	1.3	3.7	95.5
Test & Validation	29.3	0.4	1.7	0.7	0.0	1.4	3.8	84.4
p<0.05	No	No	No	No	No	No	No	No

Table 13. Comparing clinical characteristics of all three datasets – sarcoidosis patients



AVERAGE	+ve micro results	No with consolidation	No. of Symptoms	Length (days)	CURB 65	CRP mg/L	Lymph x10⁹/L	Neut x10⁹/L	QFT	Co-morbidities	IV Antibiotics before blood test (doses)	Oral Antibiotics before test (days)
Training	0.3	1	3.4	7.3	0.5	100	1.4	6.2	0.2	1.4	3.0	3.9
Test	0.3	1	3.3	6.1	0.5	133	1.4	7.9	0.0	1.0	2.9	3.2
p<0.05	No	No	No	No	No	No	No	No	No	No	No	No

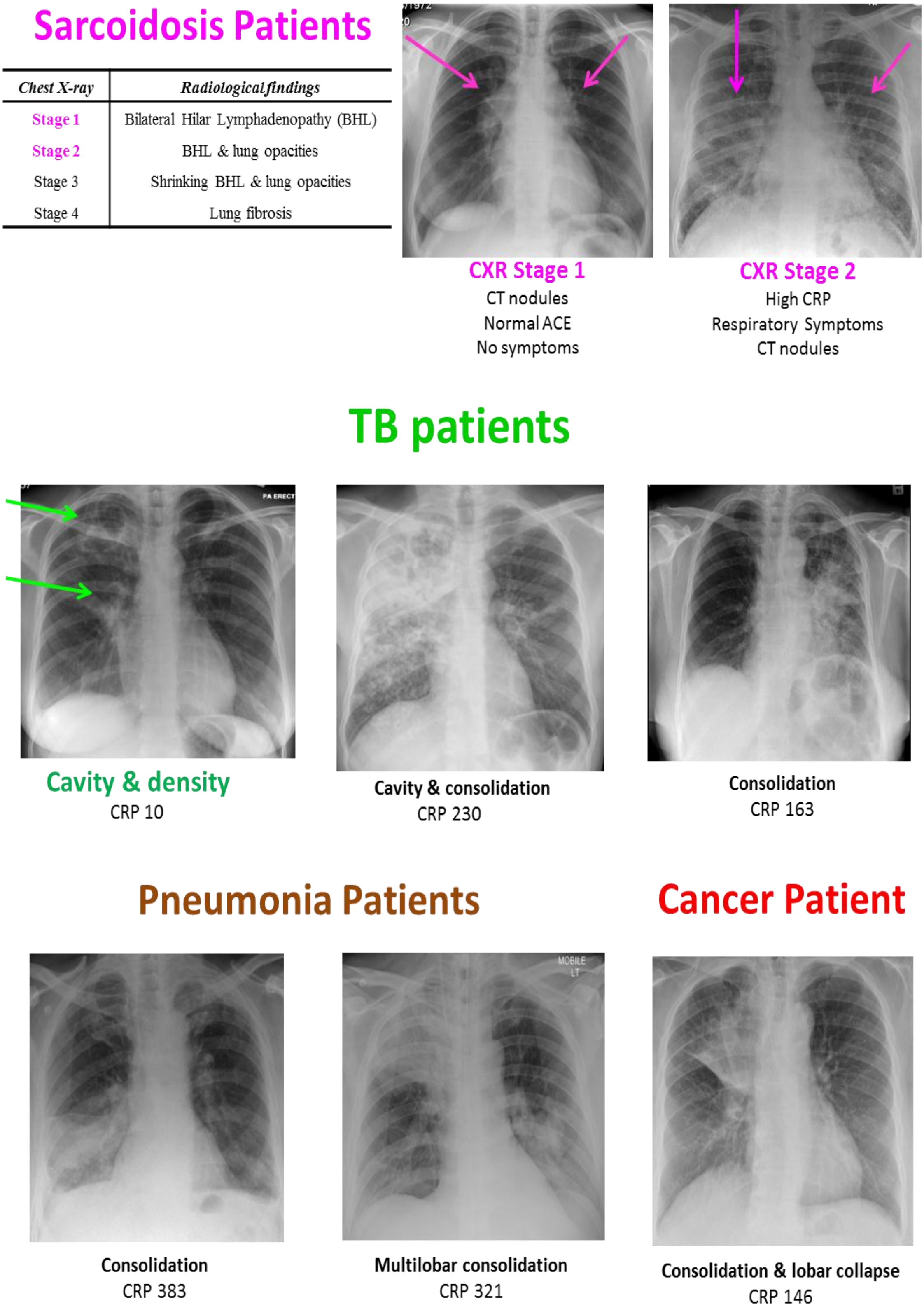
Table 14. Comparing clinical characteristics of all there datasets – pneumonia patients



AVERAGE	Histology	Lung Cancer Stage	Symptoms	CRP mg/L	Lymph x10⁹/L	Neut x10⁹/L	Smoker/Ex-smoker
Training	4 types	3a-b	2.3	84	1.3	9.2	0.9
Test	4 types	3b	2.8	47	1.3	5.7	0.8
p<0.05	No	Yes	No	No	No	No	No

Table 15. Comparing clinical characteristics of all three datasets – lung cancer patients

Each table shows an average value and if there is a statistical difference between the training and the test/validation sets. No = no difference (*p*<0.05). The two variables that were significantly different between the training set and test/validation set were the percentage of sputum smear positive TB patients and the stage of lung cancer. There were more smear positive patients in the training set and the test set contained patients with a higher average lung cancer stage.



Sarcoidosis Patients

Chest X-ray	*Radiological findings*
Stage 1	Bilateral Hilar Lymphadenopathy (BHL)
Stage 2	BHL & lung opacities
Stage 3	Shrinking BHL & lung opacities
Stage 4	Lung fibrosis

CXR Stage 1
CT nodules
Normal ACE
No symptoms

CXR Stage 2
High CRP
Respiratory Symptoms
CT nodules

TB patients

Cavity & density
CRP 10

Cavity & consolidation
CRP 230

Consolidation
CRP 163

Pneumonia Patients

Consolidation
CRP 383

Multilobar consolidation
CRP 321

Cancer Patient

Consolidation & lobar collapse
CRP 146

Figure 12. Example chest radiographs from the training set.

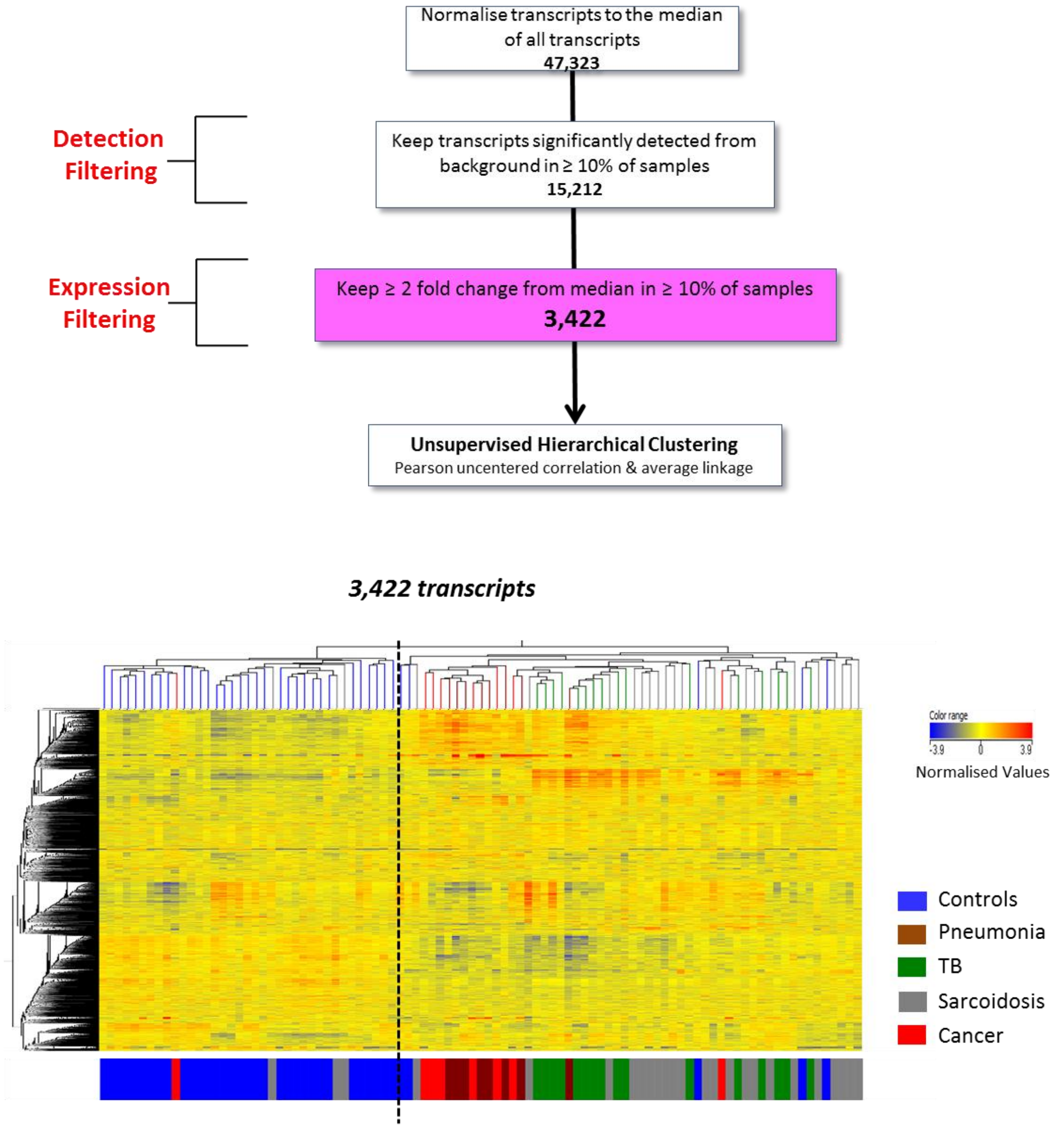


Figure 13. Unsupervised analysis of the training set.

RNA was extracted from whole blood from 16 TB, 25 sarcoidosis, 8 pneumonia, 8 cancer patients and 38 healthy controls. The RNA was hybridised to Illumina HT 12 V4 microarray chips. The 3,422 transcripts shown in the heatmap were generated by unsupervised analysis; filtered firstly by their detection compared to background intensity ($p<0.01$) and then by a two-fold filter from the median in ≥ 10% of the samples, as shown in the flow diagram. Unsupervised hierarchical clustering was then applied to these transcripts. Each row represents a transcript and each column represents a sample. The transcripts and samples were clustered by Pearson uncentered distance metric with average linkage. The vertical dendrogram shows the clustering of the transcripts and the horizontal dendrogram shows the samples clustering. The relative abundance of the normalised transcripts is indicated by the colour scale. The coloured bar at the bottom of the heatmap indicates the group the sample belongs to, as shown in the legend to the right of the heatmap. The dotted line on the heatmap was added manually to aid visualisation of the two main clusters.

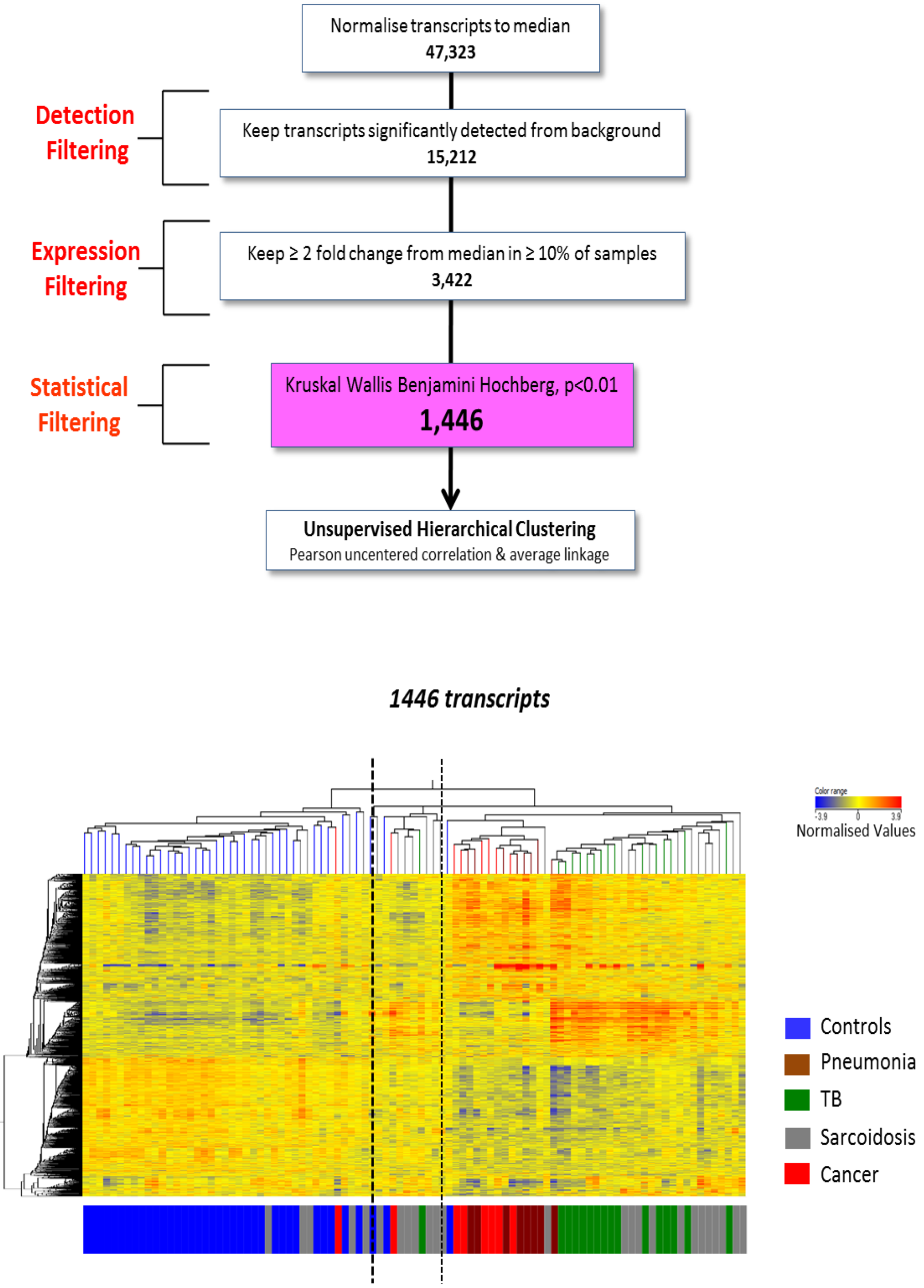


Figure 14. Unsupervised analysis and statistical filtering of the training set.

The transcripts were derived as shown in the top of this figure. The layout of the heatmap is as described in figure 13.

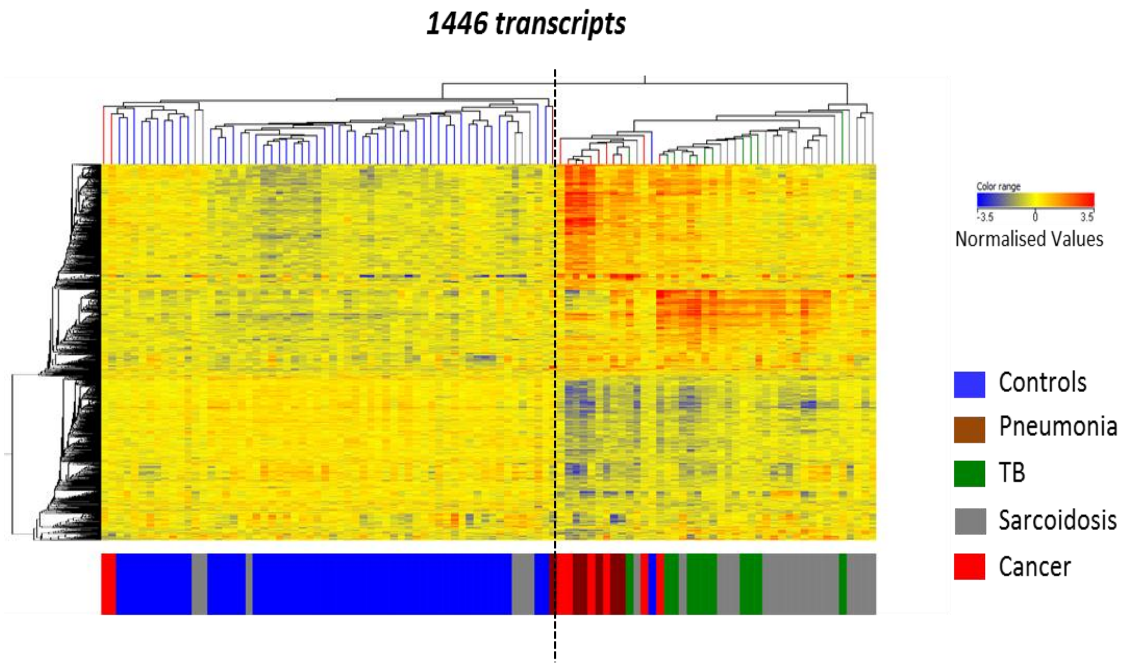


Figure 15. Testing the 1,446 transcripts derived from the training set in the TEST set.

The transcripts were derived as shown in the top of this figure. The layout of the heatmap is as described in figure 13.

Applying the 1,446 transcripts to both the training and test set demonstrated

- The controls clustered separately from the majority of the patients
- Within the cluster containing the majority of the patients, TB and most of the sarcoidosis profiles clustered together
- Some sarcoidosis patients cluster with the controls
- Within the cluster containing the majority of the patients, pneumonia and cancer profiles clustered together

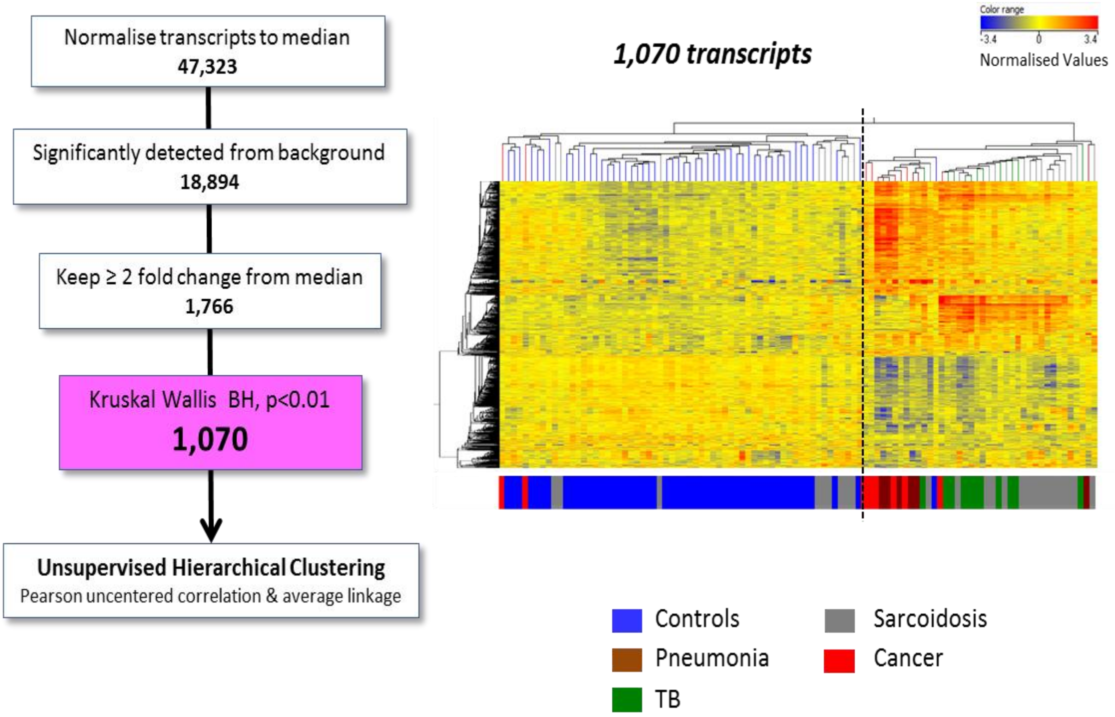


Figure 16. Same analysis approach (unsupervised analysis and statistical filtering) was applied to the test set as had been used in the training set.

The transcripts were derived as shown in the left of this figure. The layout of the heatmap is as described in figure 13.

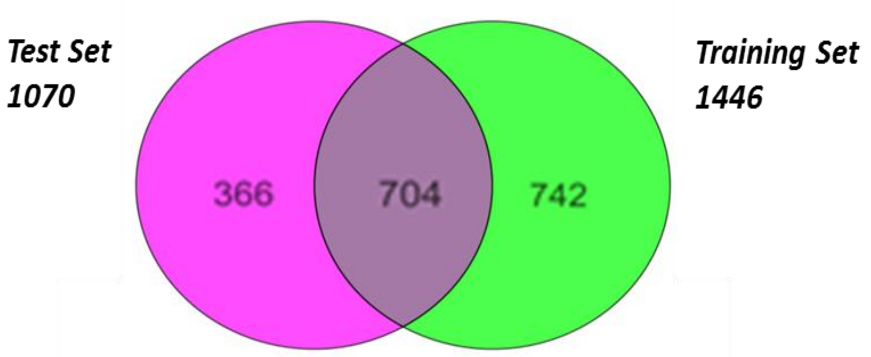


Figure 17. Venn diagram demonstrates there a large proportion of overlapping genes in the training and test set derived from the same analysis.

The unsupervised analysis uses a fold change cut-off around the median of all transcripts. Therefore any changes in the median will affect the number of transcripts derived, as has occurred here.

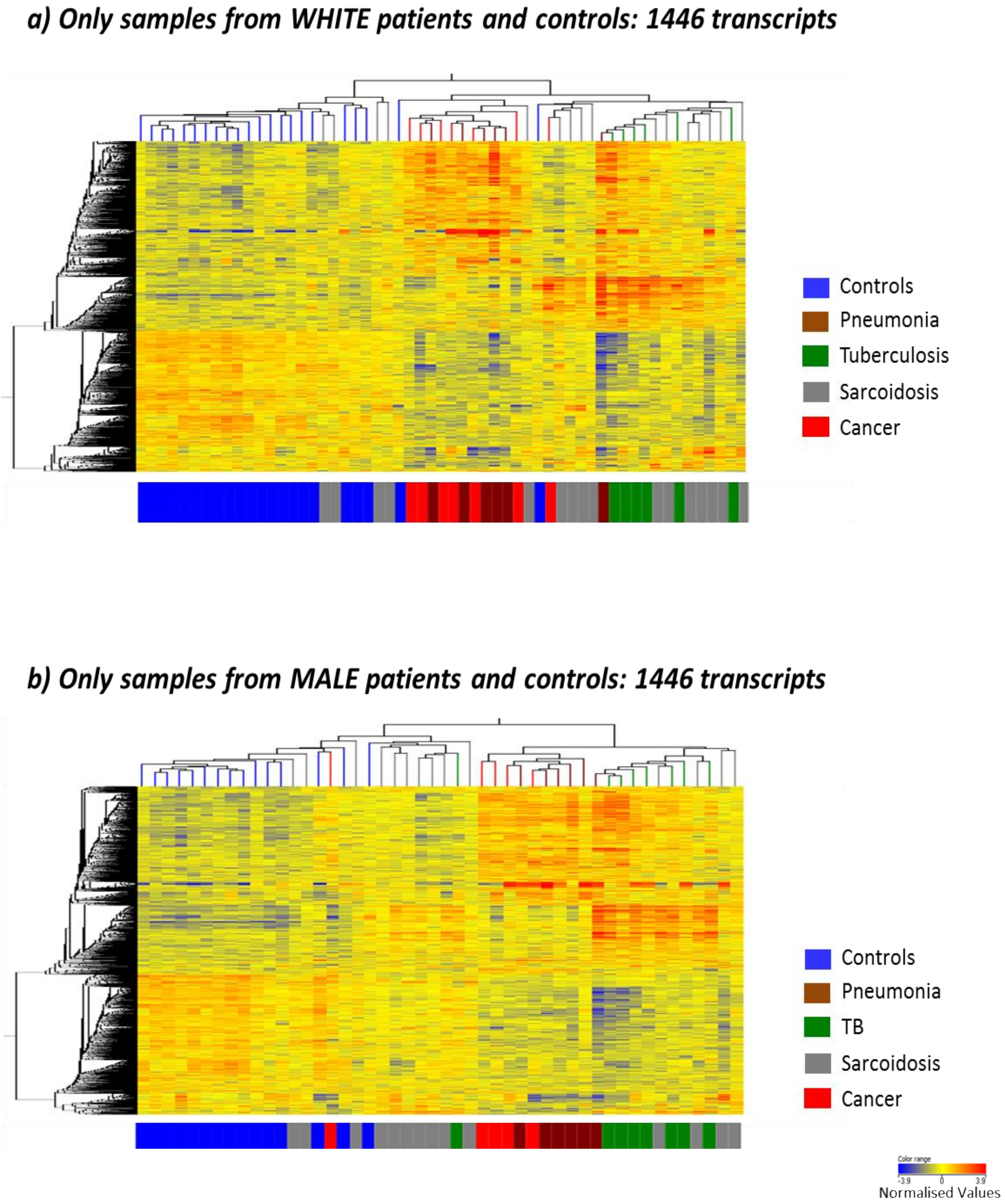
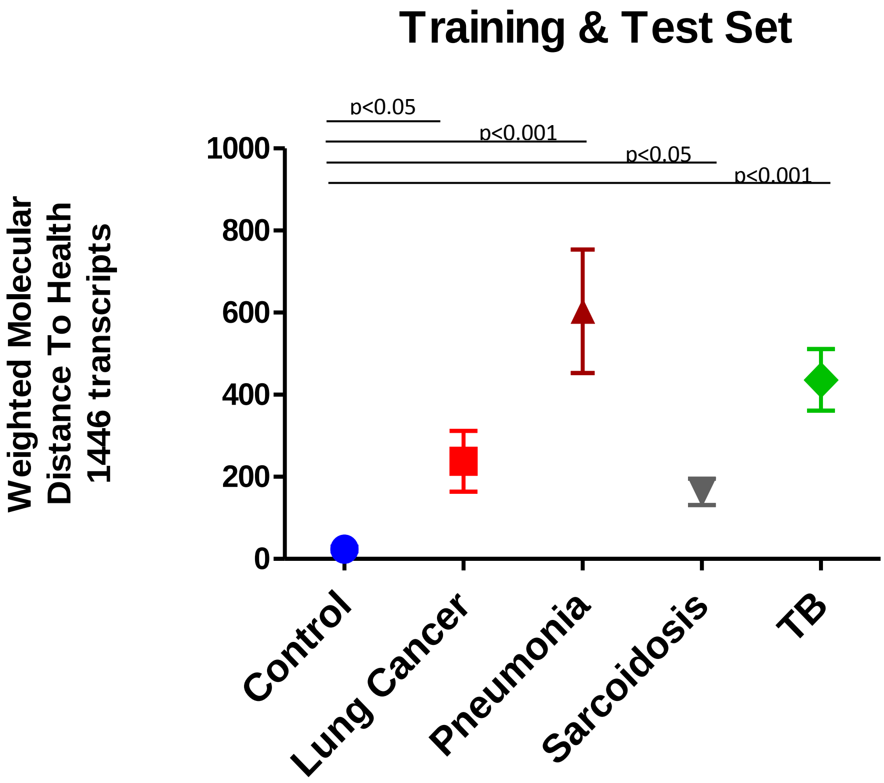


Figure 18. Clustering of samples is not influenced by differences in ethnicity or gender.

To ensure the clustering was not led by ethnicity of gender, unsupervised hierarchical clustering of the 1,446 transcripts was performed again on the training set but only including the patients and controls of the same ethnicity (white) or same gender (male). The transcripts were derived as shown in figure 14. The layout of the heatmap is as described in figure 13.



Weighted molecular distance to health measures the magnitude of transcriptional perturbation compared to the controls.

Figure 19. Molecular distance to health used to reflect disease activity.

The graph displays the mean, SEM and *p* values by ANOVA with Tukey‘s multiple comparison test.

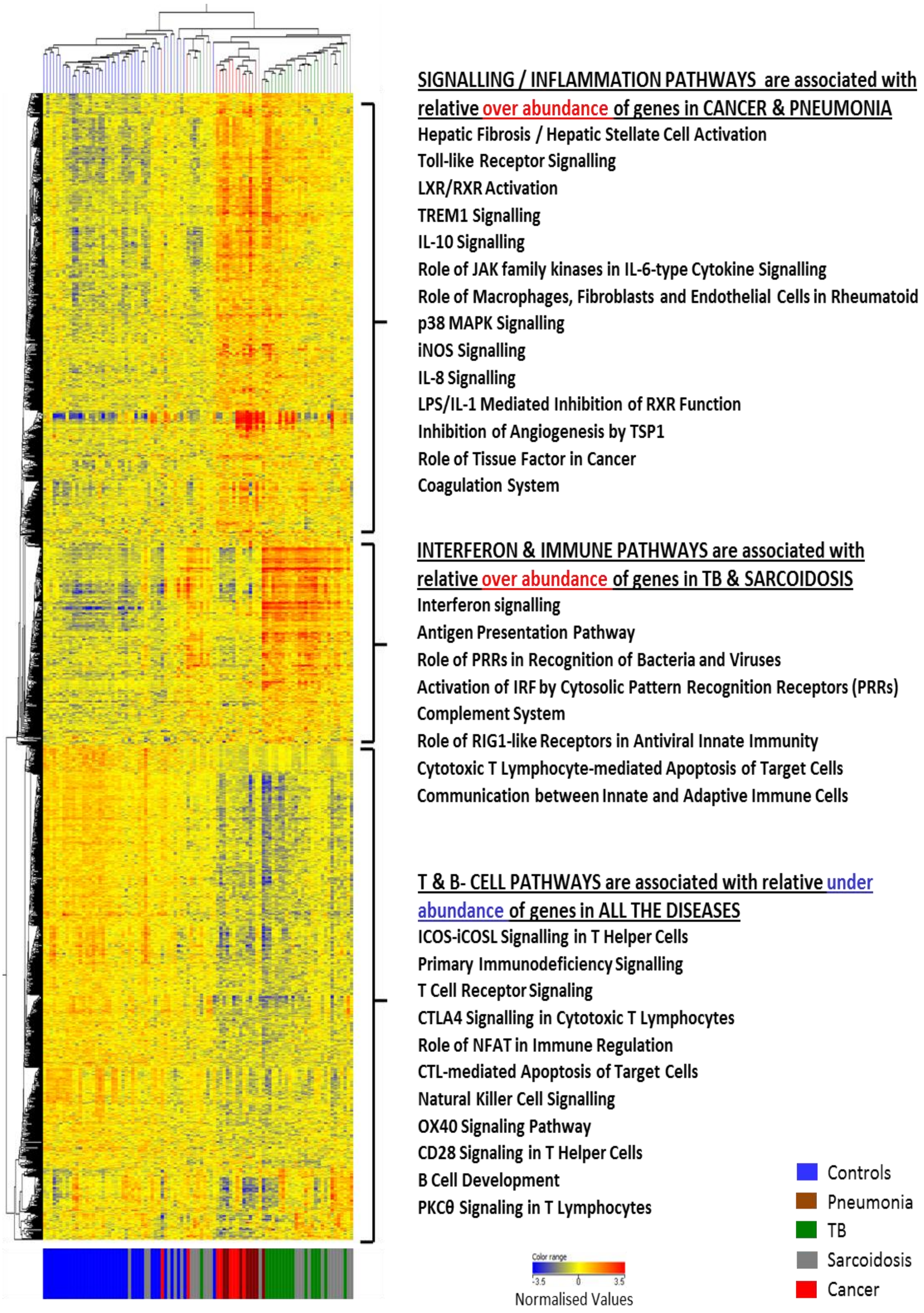


Figure 20. Ingenuity Pathway Analysis of the 1,446 transcripts derived from the training set.

IPA analysis was used to classify significantly associated pathways for each of the three main gene clusters identified by the vertical dendrogram. Only significant pathways are shown (Fishers exact Benjamini Hochberg *p*<0.05).

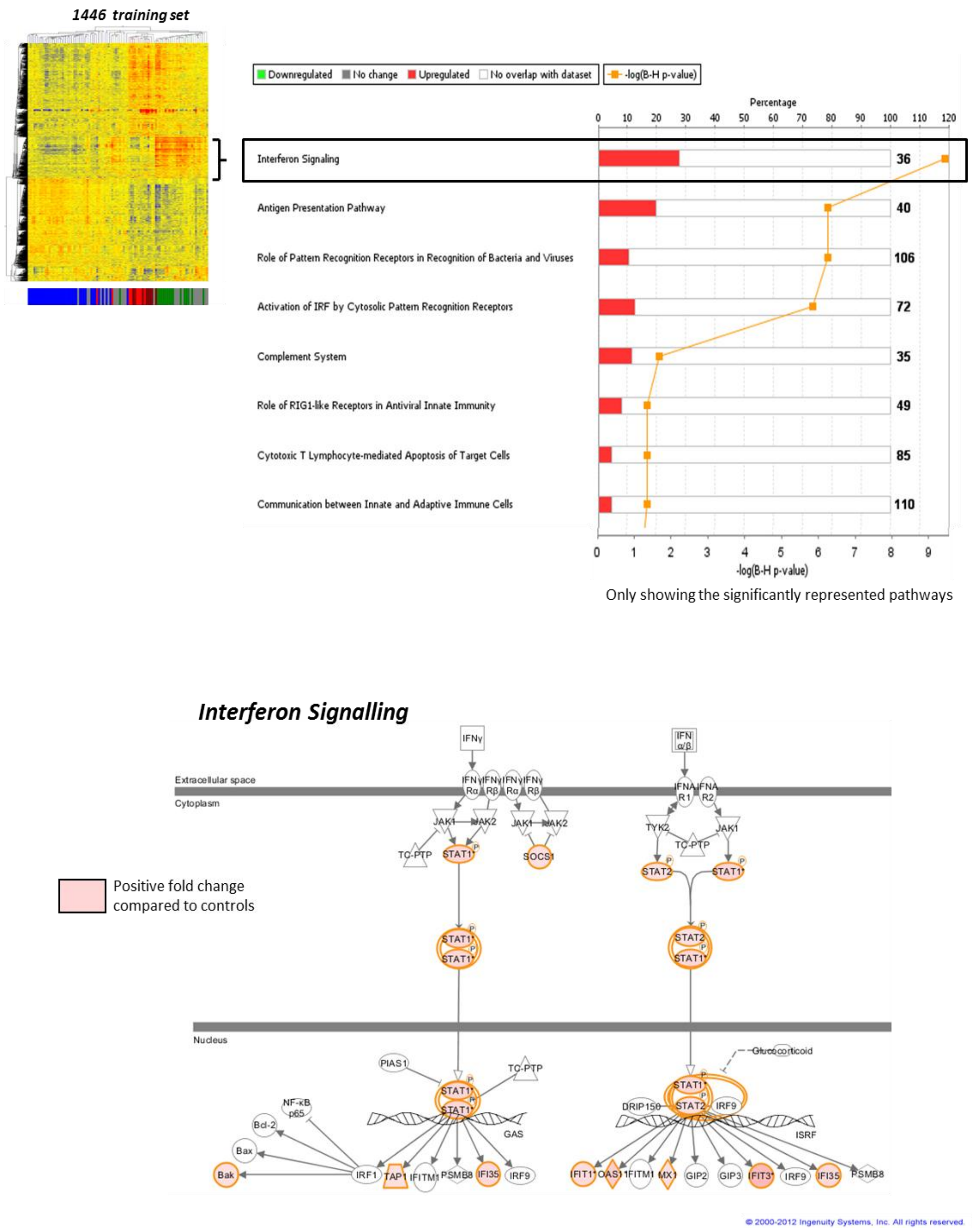


Figure 21. Ingenuity Pathway Analysis of the middle gene cluster of the 1,446 heatmap.

The graph of IPA pathways displays the percentage of genes present in that pathway along the top axis (red if upregulated as per the legend) and the log (Benjamini Hochberg *p* value) by the orange line and x-axis. The numbers at the end of each pathway along the right hand side indicate the total number of possible genes in each pathway. The genes from the cluster that were significantly associated with the IFN-signalling pathway are shown in the cartoon below the graph. They are coloured according to the fold change relative to the controls, red represents a positive fold change.

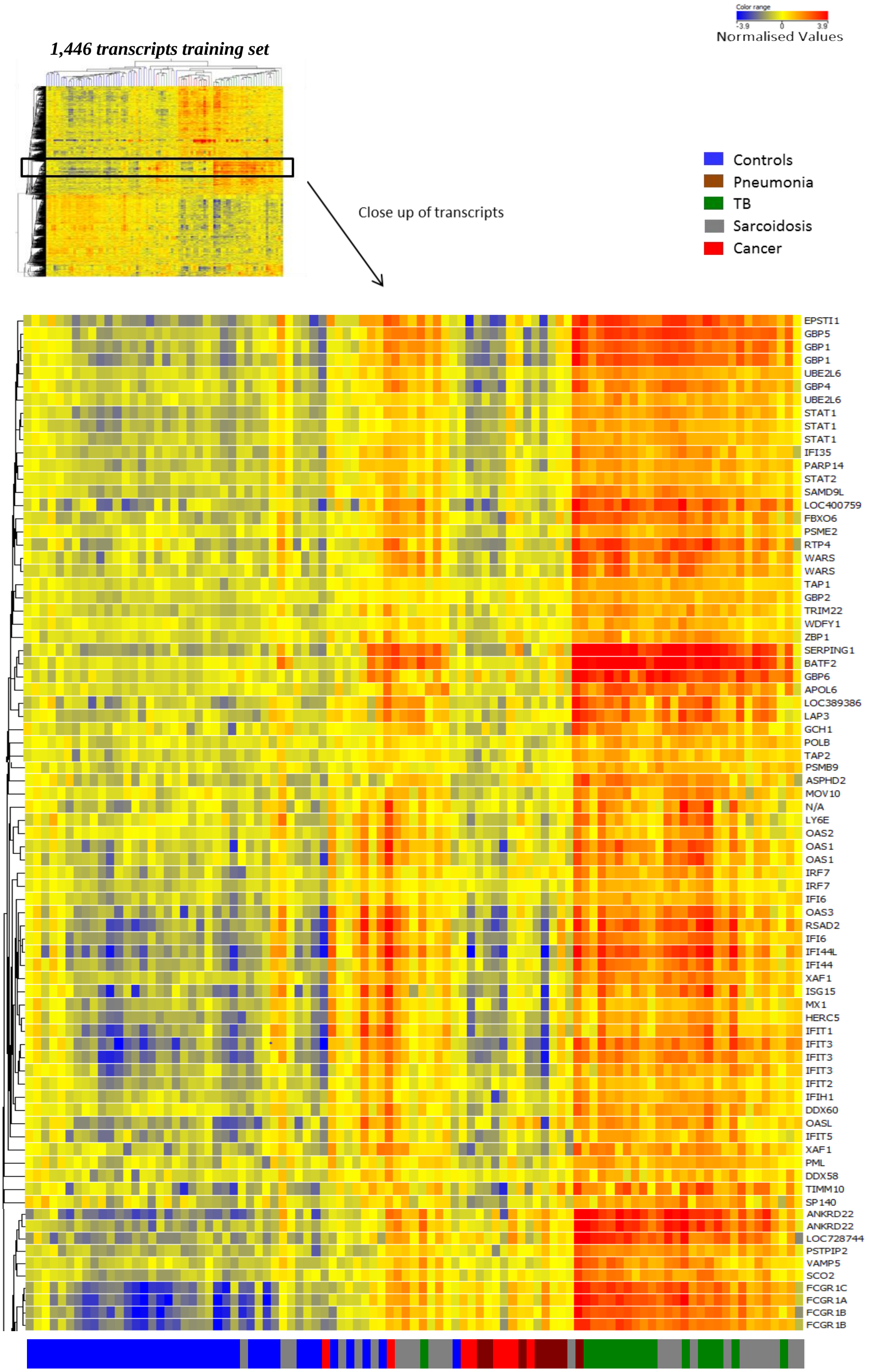


Figure 22a. Close up of the annotation of each transcript in the middle gene cluster of the 1,446 heatmap – top third

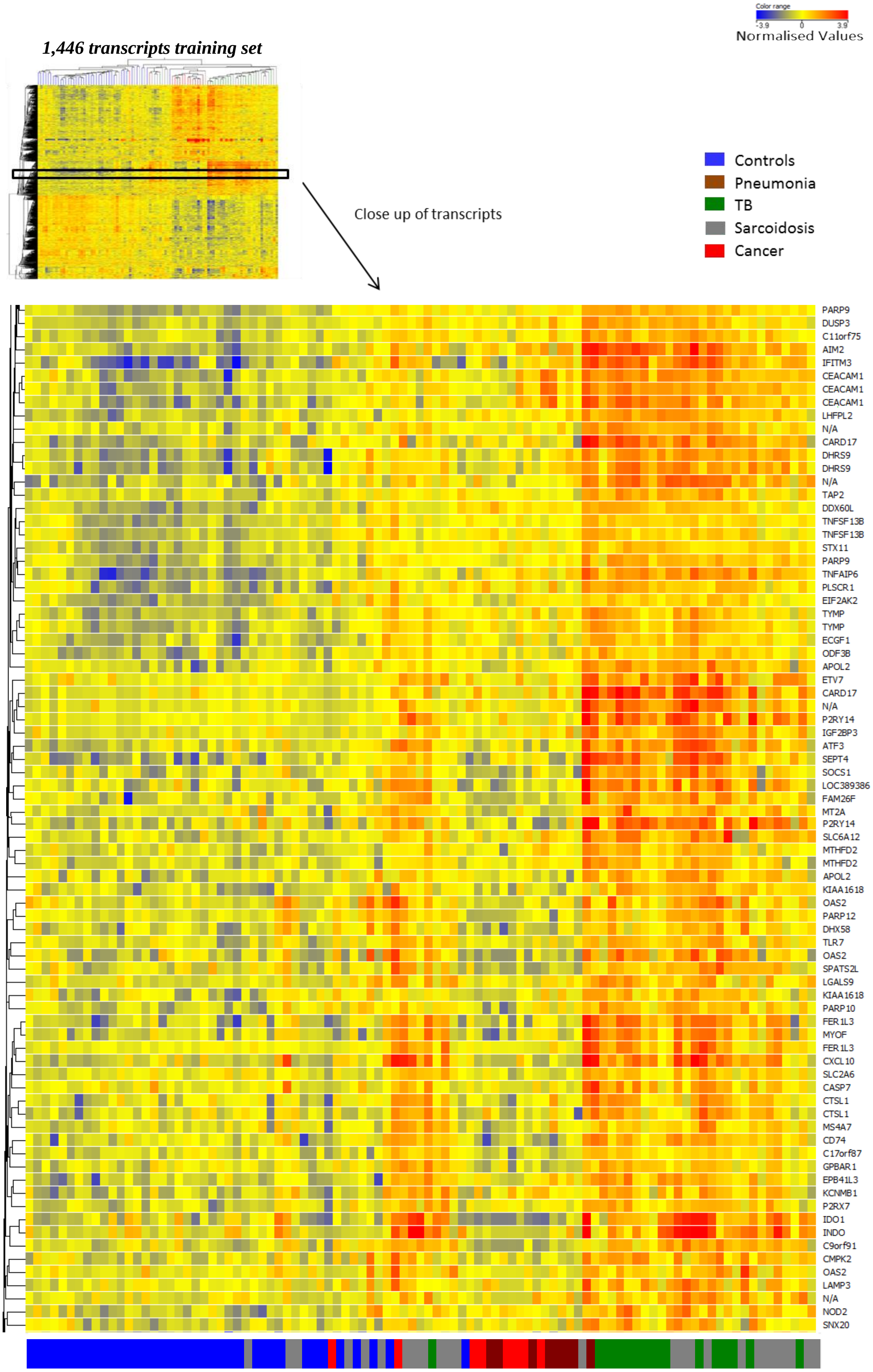


Figure 22b. Close up of the annotation of each transcript in the middle gene cluster of the 1,446 heatmap – middle third

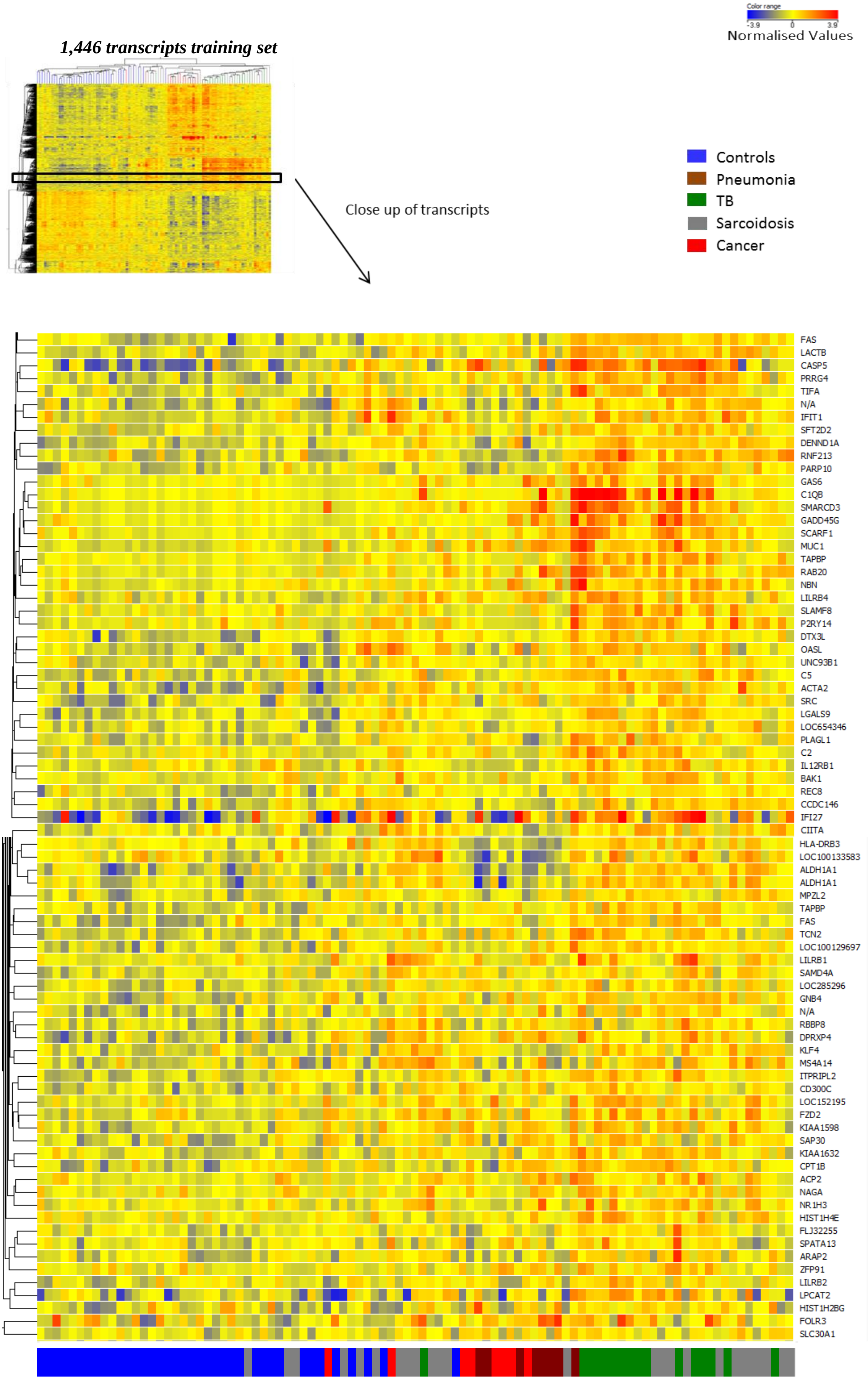
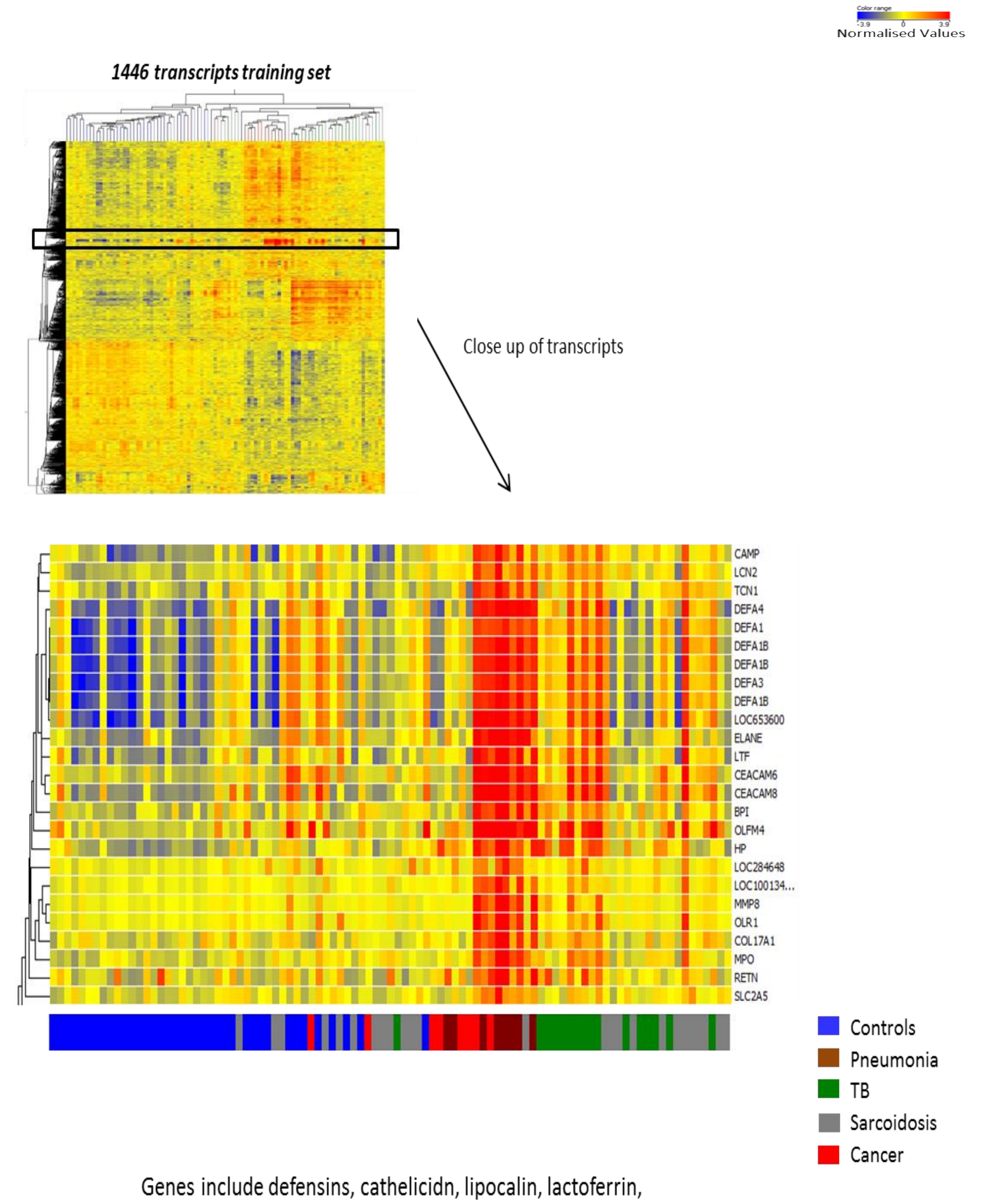


Figure 22c. Close up of the annotation of each transcript in the middle gene cluster of the 1,446 heatmap – bottom third



Genes include defensins, cathelicidn, lipocalin, lactoferrin, neutrophil elastase, myeloperoxidase, MMP8 and bactericidal permeability-increasing protein.

Figure 23. Identification & annotation of a cluster of genes that are highly over-abundant in most pneumonia patients.

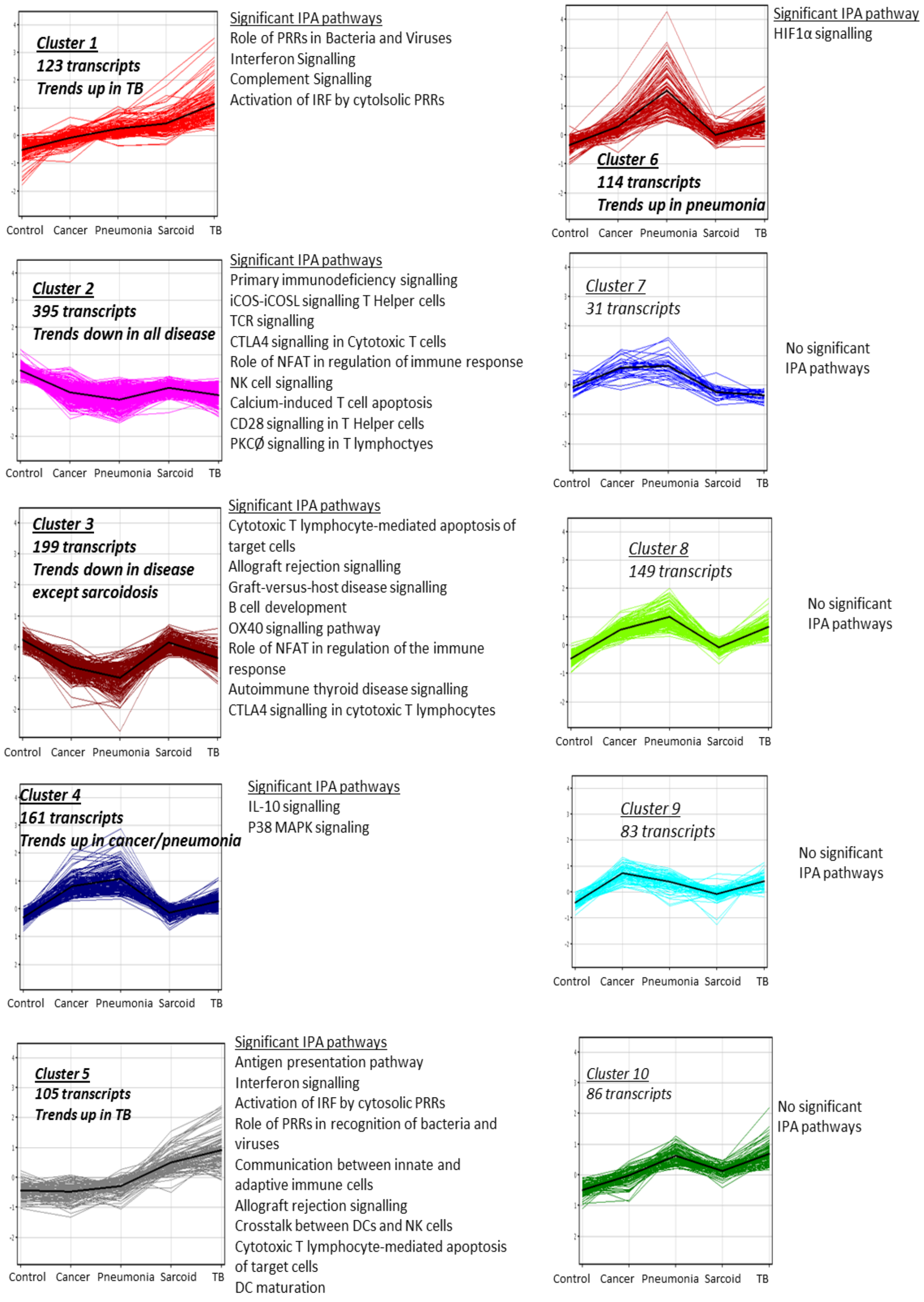


Figure 24. *k*-means clustering of the 1,446 transcripts in the training set.

The training set 1,446 transcripts were divided into 10 clusters by *k*-means clustering. The expression profiles within each of the 10 clusters are associated with particular disease(s) as indicated along the x-axis of each cluster. The y-axis indicates the normalised expression value.

Chapter 4

Clinical phenotyping of sarcoidosis patients correlates with the heterogeneity of the profiles

Chapter 4: Clinical phenotyping of sarcoidosis patients correlates with the heterogeneity of the profiles

Introduction

To gain a better appreciation of the underlying biological pathways associated with each respiratory disease group it was essential to first understand the heterogeneity of the sarcoidosis patients' transcriptional profiles as some patients clustered with the controls while the majority clustered with the other patients. This was achieved by correlating the clinical phenotyping of the sarcoidosis patients with their transcriptional profiles.

Although many attempts have been reported in the literature to clinically phenotype sarcoidosis into different categories unfortunately there remains only one accepted and globally acknowledged classification. The Scadding's chest radiograph criteria, originally formulated in the 1960's, is still widely applied in both the clinical setting and in publications (Scadding 1961). However it is insufficient for clinical decision making and furthermore is often not reproducible between different physicians and radiologists (Thillai, Eberhardt et al. 2012).

The majority of published classification schemes use clinical information that has been collated over time, to allow patients to be defined as having either 'acute or chronic disease', or 'self-limited or progressive', but these classifications do not consider disease activity at a single time point (Prasse, Katic et al. 2008; Lockstone, Sanderson et al. 2010). This type of assessment therefore prevents any clinical phenotyping without a prolonged clinical assessment over the course of time. As this study is taking a snap shot view of the host response a similar approach was applied to clinically phenotyping the patients. Patients were characterised solely using the clinical features assessed around the time of their blood sampling, irrespective of their disease severity, predicted prognosis or previous disease activity status. Therefore patients were classified purely as either those with active disease or non-active disease, as defined at the time of the blood test, where disease activity is thought to reflect granulomatous inflammation (WASOG 1994). Many clinical findings have been shown to correlate with disease activity, some of which were available from routine tests at the hospitals the patients were recruited from e.g. symptoms (WASOG 1999), serum angiotensin converting enzyme (ACE) (Ainslie and Benatar 1985), blood lymphopenia (Morell, Levy et al. 2002; Sweiss, Salloum et al. 2010), presence of pulmonary nodules (Abehsera, Valeyre et al. 2000) and change in lung function test (Keir and Wells 2010). None of these are specific markers of sarcoidosis disease activity but used in conjunction may offer some discriminatory value. In addition patients should be classified not just by their disease activity but also on the basis of their expected prognosis and by their current severity of illness. However these assessments are beyond the scope of the clinical data collected for this study.

For this study a clinical classification criteria was specifically designed (Figure 3) due to the lack of availability of any validated classification system specifically for disease activity and the lack of a classification system that did not require prolonged follow-up of the patient. The classification system used in our study was devised from (a) evidence-based clinical variables shown to correlate with disease activity and (b) clinical variables which were also available for most of the recruited patients. This classification system should therefore be applicable in routine clinical care. However a major caveat of the classification system is its dependence on recent thoracic radiology tests, where the detailed results available from a high-resolution computed tomography scan (HRCT) permitted the most accurate classification. Although most physicians regularly requested a chest radiograph for the patients they did not often request a HRCT due to its appreciable level of radiation. In addition the classification system was formulated for pulmonary sarcoidosis as a classification system for sarcoidosis involving other organs would be considerably more complex. Therefore to take into account those patients with active disease in other organs the practising physicians own acumen was used to define the patient's classification (Figure 3). It is common practice in sarcoidosis studies to use the management plan of the practising physician to define the patient's clinical outcome (Miyara, Amoura et al. 2006; Baughman, Nagai et al. 2011).

Results

Sarcoidosis patient's clinical variables and classification

After applying our classification system 17 of the 25 training set sarcoidosis patients were classified as active pulmonary sarcoidosis (Table 16a). In the test set 13 of the 25 sarcoidosis patients were classified as active pulmonary sarcoidosis and 2 were classified as active extra-thoracic sarcoidosis (Table 16b). In the validation set 5 of the 11 patients were classified as active pulmonary sarcoidosis and one as active extra-thoracic sarcoidosis (Table 17). Most of the patients (57%) had a chest radiograph stage I, 18% had stage II, 11% had stage III and 11% had stage IV, with just one patient who had stage 0. Just under a third of patients had not had a recent (within 6 months of the blood profile) HRCT performed. Of those that did three-quarters had features of active disease in their lung parenchyma and only one-quarter did not. By far the most common feature of active sarcoidosis were pulmonary nodules; the other feature sometimes seen was ground glass opacification, a non-specific finding due to a hazy opacity within the parenchyma that unlike consolidation does not obscure vessels and bronchi. A quarter of patients had lymphopenia (defined as <1x10^9/L). In approximately one third of all patients sarcoidosis was thought to be affecting more than one organ, particularly the skin, joints and eyes. Over half of the patients had abnormal lung function parameters, albeit many of them were only mildly abnormal, this was most commonly due to low gas diffusion (TLCO % predicted <80) rather than low lung volumes (FVC % predicted <80) or raised FEV1/FVC ratio which is indicative of a restrictive lung pathology. The majority of the patients had biopsies taken from either their mediastinal lymph nodes or from their lung parenchyma. Patients were offered treatment in 39% of cases; three patients declined treatment while all others accepted. The number of patients offered treatment by the practising physician was 23% less than the number of patients classified as having active disease. Only two patients classified as having non-active sarcoidosis were offered treatment.

There were equal numbers of males and females between those classified as active and non-active (Table 18). There was more disparity across the ethnicities, as the majority of non-active patients but just under half of the active patients were of white ethnicity.

Clustering of the sarcoidosis patients in the training set compares well to their clinical classifications of active or non-active sarcoidosis

To determine if our clinical classification system of the sarcoidosis patients correlated with the clustering of the patient's expression profiles the same analysis approach was applied as had been used earlier (Figure 14, 1446 transcripts). However the difference in the analysis was instead of treating the sarcoidosis patients as one group they were divided into either an active sarcoidosis sub-group or non-active sarcoidosis sub-group using our clinical classification system. Therefore the analysis was performed with six different groups of patients/controls: TB, active sarcoidosis, non-active sarcoidosis, cancer, pneumonia and controls. The analysis approach used was exactly the same as before, unsupervised analysis (two-fold change from the median) followed by statistical filtering (Kruskal Wallis Benjamini Hochberg p<0.01). This generated 1,396 transcripts (Figure 25), instead of the 1,446 transcripts generated when sarcoidosis was treated as just one group. Applying the same unsupervised hierarchical clustering algorithm it could be seen that all but one of the active sarcoidosis profiles clustered with the majority of the other patients, of which the TB and active sarcoidosis profiles formed the closest cluster. In addition all but two of the non-active sarcoidosis profiles clustered with the controls. As would be expected clustering of the Tb, pneumonia and cancer profiles were not affected by the newly derived transcript list (Figure 25).

These results therefore demonstrated that of the two main clusters in the heatmap the active sarcoidosis profiles clustered with the TB profiles while the non-active sarcoidosis profiles clustered with the controls profiles.

Findings in the training set were validated in the test and validation sets

Next the same transcript list as above, the 1396 transcripts derived from the training set, were used to verify the robustness of the clustering of the two different sub-groups of sarcoidosis patients, by performing unsupervised hierarchical clustering of the test set and validation set samples (Figure 26). In the test set only one active sarcoidosis patient did not cluster with the majority of the other patients (Figure 26). In addition in the validation set all of the active sarcoidosis patients clustered with the TB patients (Figure 26). However in the test set 44%, and in the validation set 40%, of the non-active sarcoidosis patients clustered within the main patients cluster, away from the controls. All the non-active patients that did cluster with the other patients could be found towards the edge of the cluster. Therefore their transcriptional profiles were the least similar compared to the other patients' profiles.

These results verify the observation seen in the training set that in the test set and validation set the active sarcoidosis profiles again clustered with the TB profiles. In addition most of the non-active sarcoidosis profiles again clustered with the controls' profiles.

Individual clinical variables were not as effective as predicting clustering as the clinical classification system

To determine any potential predictive clinical features for the clustering of the sarcoidosis patients' statistical tests were used to help define relationships between the clusters and each clinical variable. First it was established for the training and test set patients whether they fell into the cluster with the patients or the cluster with the controls according to the unsupervised hierarchical clustering of the 1,446 list (Figure 14). The 1,446 transcript list rather than the 1,396 transcript list was used. This is because although both lists were derived by the same analysis the generation of the 1,466 transcript list was not biased by the clinical classifications of the sarcoidosis patients. To generate the largest power to test the association of the clinical variables with the clustering, both the training and test set were used (total of 50 sarcoidosis patients where 11 clustered with the controls and 39 clustered with the patients). For categorical and ordinal variables Pearson chi-squared test for significance was used (Table 19), and for continuous variables logistic regression was used (Table 20). Some variables were converted to both categorical and continuous e.g. serum ACE of 30 IU/L was either 'normal' (categorical: normal or high where >55 IU/L = high) or left as the unit of 30 (continuous). This was important as the cut-off for variables such as ACE level was arbitrarily developed and subsequently accepted clinically, the cut-off > 55 IU/L was chosen as it is applied in most of the hospitals the patients were recruited from.

Variables that were significantly associated with the clustering of sarcoidosis patients were blood lymphocyte count, commencing treatment, active changes on HRCT, and the clinical classification system (*p*<0.05, Table 19). The clinical classification system had the most statistically significant correlation with the clustering (*p*<0.0001). The odds of a sarcoidosis patient clustering with the TB patients were only slightly increased for two variables, a rise in a unit of serum ACE or a fall of 1x10^7/L of the blood lymphocyte count (*p*<0.05, Table 20). Whereas the odds of a sarcoidosis patients clustering with the other patients were 17.4 times higher for patients classified as active sarcoidosis than those classified as non-active sarcoidosis (*p*<0.002, Table 20). No other variables were significantly associated with sarcoidosis patients' cluster predictions.

This analysis suggested that single clinical variables were less effective at predicting the sarcoidosis clustering than the clinical classification system. Therefore to determine if a prediction model containing more than one clinical variable could be effective multivariate regression analysis was performed, using the most significant variables from the univariate analysis. However all three multivariate models had lower significances in relation to the clinical classification system (Table 21).

To determine if the clinical classification could have improved predictive abilities, additional clinical variables were tested by multivariable regression. Only variables with high significance by univariate analysis and variables not already included in the classification system were tested. However, the only possible additional variable, the blood lymphocyte count, reduced the predictive value of the classification system (Table 22).

In summary from the regression analysis of the clinical variables it could be clearly seen that our clinical classification system had the highest predictive potential for identifying which sarcoidosis patients clustered with the controls and which clustered with the patients.

Discussion

Sarcoidosis is a disease with a protean presentation (Baughman, Teirstein et al. 2001) (WASOG 1999). However there is a paucity of validated assessment tools to aid patient management in respect to its variable clinical phenotypes. Standardising sarcoidosis phenotyping is imperative and may be helped by advances in genomic research and increased application of genetic profiling. This study has provided an example of how this could be achieved. By applying a plausible clinical classification to a cohort of patients it was possible to correlate their clinical phenotype to their blood gene expression profiles attained by unbiased analysis. Lockstone *et al* also demonstrated the feasibility of this approach in expression profiles of lung biopsies from sarcoidosis patients, classifying patients either into progressive-fibrotic or self-limited (Lockstone, Sanderson et al. 2010). However the advantage of a blood transcriptome over the transcriptome from lung biopsies is that it is non-invasive and may diagnose disease involved in other organs that was missed on initial presentation.

Classification and clinical features of the sarcoidosis patients

The majority of the sarcoidosis patients were categorised by the clinical classification as having active pulmonary disease but most of these patients had a chest radiograph of only stage I. Scadding’s stage I is the presence of bilateral hilar lymphadenopathy and no lung parenchymal involvement. It may therefore seem surprising that these patients were classified as having active disease as according to treatment guidelines stage I disease does not require treatment unless symptomatic and only after further investigation (Bradley, Branley et al. 2008; Baughman and Nunes 2012). The treatment guidelines may explain the discordance (23% difference) between the numbers of patients classified as active sarcoidosis compared to the number of patients offered treatment. However it is interesting that this part of the guidelines is predominantly based on the results of just one study, a 1996 British Thoracic Society study that followed a particular protocol and cohort, a process that may not be applicable to all sarcoidosis patients (Gibson, Prescott et al. 1996). However the benefits of treatment in symptomatic stage I disease may actually be due to the high chance of spontaneous resolution without treatment in patients with stage I radiology (WASOG 1999). This is reflected in a systematic review, with subgroup analysis of the different Scadding stages, which found treatment was only beneficial in those patients with stage II-III (Paramothayan and Jones 2002). Therefore although these patients may have active disease the current treatments available are not of much advantage over the host response. Another reason for this apparent disparity between those you might have expected to have ‘active’ pulmonary disease and the number that classified as active pulmonary sarcoidosis is related to the substantial dependency on HRCT findings in the classification system. While it is recognised that grading of chest radiographs by the Scadding system is not that reliable (Thillai, Eberhardt et al. 2012), interpretations of HRCT scans should be more sensitive. The evidence of active disease on HRCT in our cohort included the features, nodules and ground glass opacities, which often cannot be seen on a chest radiograph. Pulmonary nodules have been shown to correlate with disease activity and furthermore are invariably reversible (Wells 1998; Abehsera, Valeyre et al. 2000). The inclusion criteria for ‘disease activity on CT’ was purposely broad and non-specific. More tightly controlled definitions for HRCT evidence of active disease could have been specified, such as only including peribronchovascular or subpleural nodules, a common finding in sarcoidosis, rather than the presence of nodules in any distribution (WASOG 1999). However this broad criterion was chosen for two main reasons; firstly it negated the need for detailed interpretation which can be highly variable (Erdal, Crouser et al. 2012), secondly although there are some pattern distributions of nodules that are typical for sarcoidosis numerous atypical patterns have also been described (Marchiori, Zanetti et al. 2011). However as the classification system relies fairly heavily on radiology findings, some of the patients categorised as non-active who did not have a recent HRCT, may have been incorrectly classified.

In the sarcoidosis and TB blood gene expression study by Koth et al. 2011, they identified a significant association with the blood profiles and a phenotype they described as sarcoidosis severity (Koth, Solberg et al. 2011). However the lung function parameters they used for this were FEV1 (volume of forcibly expired air in 1 second) and/or FVC (total volume of forcibly expired air), which are two parameters commonly used to assess for obstructive or restrictive lung disease. These two parameters alone or used together are not a common or evidenced based measure of sarcoidosis severity (Keir and Wells 2010). From their data it appeared that only FEV1 was statistically significantly different between their low and high severity groups therefore only FEV1, a measure of airways obstruction, was really driving their phenotyping. However airway obstruction is found in over half of sarcoidosis patients (Iannuzzi, Rybicki et al. 2007), and a static FEV1 result has not been shown to be linked with severity (Wasfi, Rose et al. 2006). Therefore this leaves the question whether the two sarcoidosis subgroups they described purely as a consequence of statistical analysis of their clinical data (FEV1 & FVC), carries any real clinical interpretation. It is true many studies have used lung function parameters to reflect sarcoidosis pulmonary involvement but they apply a combination of variables relating to both gas diffusion and lung volumes (e.g. TLCO, total lung capacity and FVC) and do not include FEV1 (Erdal, Crouser et al. 2012) (Zappala, Desai et al. 2011). Even these standard measures of sarcoidosis lung function are confounded by the heterogeneity of sarcoidosis pulmonary function (Keir and Wells

2010). In this study there was no significant association with gas diffusion variables (TLCO and KCO) or FVC, however many of the patients did not have recent lung function tests performed (Table 16 & 17). Moreover the trend in lung function is reported to be far more relevant than static results (Zappala, Desai et al. 2011).

Although our study contained small patient numbers for epidemiological conclusions, it is worth noting that there was a disparity between ethnicity, where there was a larger percentage of active patients who were black than non-active patients, and vice versa for white patients (Table 19). This is perhaps predictable as studies of different ethnic groups in London have previously shown that black patients are more likely to have a more severe and extensive disease than white patients (Edmondstone and Wilson 1985). However the diversity in ethnicity did not influence the transcriptional signature.

Clustering of the classified sarcoidosis patients

From the unsupervised hierarchical clustering of the 1,446 transcripts, including all sarcoidosis patients as one group, it could be clearly seen that the sarcoidosis patients were more heterogeneous than the other groups of diseases in both the training set and the test set (Figures 14 & 15). Interestingly this was not a characteristic of the sarcoidosis patients that either of three earlier blood transcriptional sarcoidosis studies commented on (Rosenbaum, Pasadhika et al. 2009; Koth, Solberg et al. 2011; Maertzdorf, Weiner et al. 2012). This could be because unsupervised hierarchical clustering was not performed (only supervised hierarchical clustering of supervised analysis (Rosenbaum, Pasadhika et al. 2009)) or due to biased study populations, as neither study published relevant clinical details such as length of illness or current symptoms. The findings from our study are in keeping with the known clinical heterogeneity of sarcoidosis patients (Baughman, Teirstein et al. 2001) (WASOG 1999).

After applying the clinical classification system to phenotype the sarcoidosis patients the same unsupervised analysis and statistical filtering approach was applied as used earlier in this study. By unsupervised hierarchical clustering it could be visualised in the training set that the patients designated as having active sarcoidosis clustered with the other patients, in particular the TB patients (Figure 25). This clustering was subsequently verified in both the test set and the validation set (Figure 26). However a few of the patients designated as non-active sarcoidosis clustered more towards the active sarcoidosis (2/8 in the training set, 4/9 in the test set, 2/5 validation set), although these profiles were on the edge of the clusters indicating a weaker affinity towards the group. There are three possible reasons why some of the non-active sarcoidosis patients appeared to have transcriptional profiles more similar to the active sarcoidosis patients and other patients. Firstly, even if all variables were available the clinical classification system may not be sensitive enough to detect all signs of active disease. Secondly, nearly a third of the cohort had not had a recent HRCT scan; therefore one of the major criteria could not be included. Thirdly, although a third of patients had disease knowingly affecting more than one organ, which is classical of sarcoidosis presentations (Baughman, Teirstein et al. 2001), some of the patients may have had disease in other organs that was undiagnosed and active. Typically physicians do not pursue a diagnosis that requires invasive tests unless the patient's symptoms are suggestive but even active pulmonary sarcoidosis is often asymptomatic (Table 9). Therefore it is possible this blood transcriptional profile may provide a means of detecting disease activity without invasive techniques.

Multiple clinical variables were tested for their correlation with the clustering of sarcoidosis patients (Tables 19-21). The significant variables were disease activity on HRCT, blood lymphocyte count, serum ACE and commencing treatment. All these variables have previously also been shown to correlate with disease activity (Ainslie and Benatar 1985; Abehsera, Valeyre et al. 2000; Morell, Levy et al. 2002; Baughman, Nagai et al. 2011). However univariate and multivariate analysis of the numerous available clinical variables was not able to improve upon the predictive value of the clinical classification system for the clustering of the sarcoidosis patients. Although this clinical classification system was devised according to the available clinical variables that had literature based evidence of their correlation to disease activity, it has not been validated in a prior clinical cohort nor reviewed by multiple clinical experts in the field. Nevertheless it provides a reasonable explanation for the heterogeneous transcriptional profiles seen in the sarcoidosis patients. Interestingly Minshall *et al* in 1997 also showed a distinct mRNA cytokine profile that differed between patients with active sarcoidosis and patients with non-active sarcoidosis (Minshall, Tsicopoulos et al. 1997). Although the study was only able to measure 9 different cytokines, they found an elevated level of IL2, IL12 and IFNγ in the active patients compared to the non-active patients.

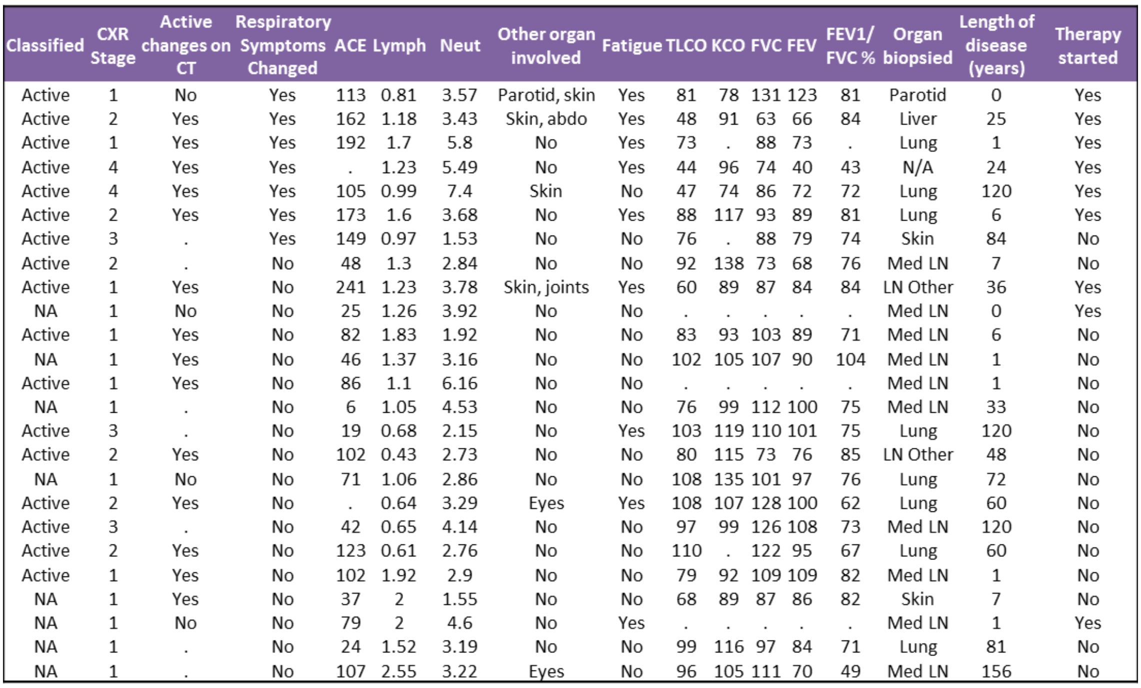
Therefore our study shows that existing disease activity (a possible surrogate marker for granulomatous inflammation) is reflected in the whole blood transcriptional profile, and that pulmonary sarcoidosis patients with active disease have both a very similar clinical and molecular phenotype to pulmonary TB patients.

Chapter Summary

Although pulmonary sarcoidosis is a very complex disease the whole blood transcriptional profiles from the sarcoidosis patients appeared to correlate well with the heterogeneity of their clinical features. This suggests transcriptional profiling could play a role in the classification of sarcoidosis patients thus enabling a targeted clinical management plan that is currently hard to achieve.

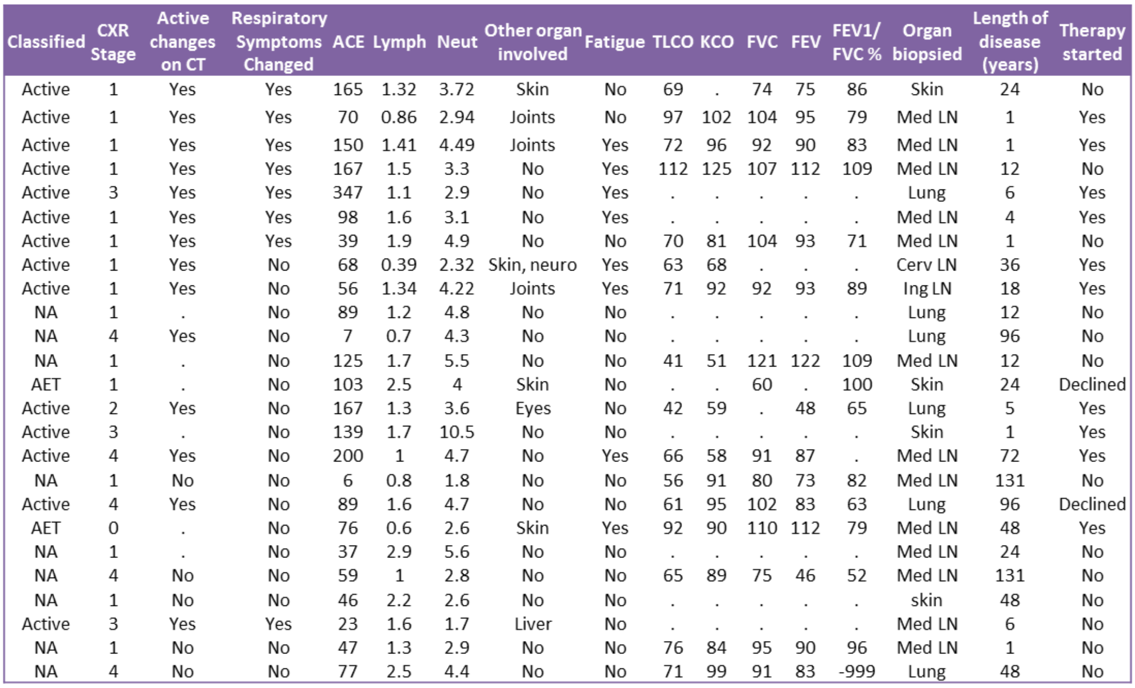
Figures for chapter 4

a) Training set



Classified	CXR Stage	Active changes on CT	Respiratory Symptoms Changed	ACE	Lymph	Neut	Other organ involved	Fatigue	TLCO	KCO	FVC	FEV	FEV1/ FVC %	Organ biopsied	Length of disease (years)	Therapy started
Active	1	No	Yes	113	0.81	3.57	Parotid, skin	Yes	81	78	131	123	81	Parotid	0	Yes
Active	2	Yes	Yes	162	1.18	3.43	Skin, abdo	Yes	48	91	63	66	84	Liver	25	Yes
Active	1	Yes	Yes	192	1.7	5.8	No	Yes	73	.	88	73	.	Lung	1	Yes
Active	4	Yes	Yes	.	1.23	5.49	No	Yes	44	96	74	40	43	N/A	24	Yes
Active	4	Yes	Yes	105	0.99	7.4	Skin	No	47	74	86	72	72	Lung	120	Yes
Active	2	Yes	Yes	173	1.6	3.68	No	Yes	88	117	93	89	81	Lung	6	Yes
Active	3	.	Yes	149	0.97	1.53	No	No	76	.	88	79	74	Skin	84	No
Active	2	.	No	48	1.3	2.84	No	No	92	138	73	68	76	Med LN	7	No
Active	1	Yes	No	241	1.23	3.78	Skin, joints	Yes	60	89	87	84	84	LN Other	36	Yes
NA	1	No	No	25	1.26	3.92	No	No	.	.	.	.	.	Med LN	0	Yes
Active	1	Yes	No	82	1.83	1.92	No	No	83	93	103	89	71	Med LN	6	No
NA	1	Yes	No	46	1.37	3.16	No	No	102	105	107	90	104	Med LN	1	No
Active	1	Yes	No	86	1.1	6.16	No	No	.	.	.	.	.	Med LN	1	No
NA	1	.	No	6	1.05	4.53	No	No	76	99	112	100	75	Med LN	33	No
Active	3	.	No	19	0.68	2.15	No	Yes	103	119	110	101	75	Lung	120	No
Active	2	Yes	No	102	0.43	2.73	No	No	80	115	73	76	85	LN Other	48	No
NA	1	No	No	71	1.06	2.86	No	No	108	135	101	97	76	Lung	72	No
Active	2	Yes	No	.	0.64	3.29	Eyes	Yes	108	107	128	100	62	Lung	60	No
Active	3	.	No	42	0.65	4.14	No	No	97	99	126	108	73	Med LN	120	No
Active	2	Yes	No	123	0.61	2.76	No	No	110	.	122	95	67	Lung	60	No
Active	1	Yes	No	102	1.92	2.9	No	No	79	92	109	109	82	Med LN	1	No
NA	1	Yes	No	37	2	1.55	No	No	68	89	87	86	82	Skin	7	No
NA	1	No	No	79	2	4.6	No	Yes	.	.	.	.	.	Med LN	1	Yes
NA	1	.	No	24	1.52	3.19	No	No	99	116	97	84	71	Lung	81	No
NA	1	.	No	107	2.55	3.22	Eyes	No	96	105	111	70	49	Med LN	156	No

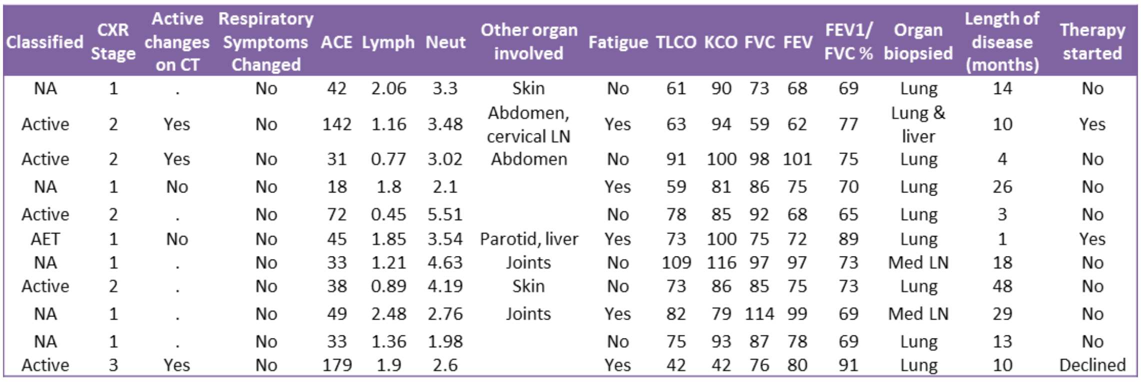
b) Test set



Classified	CXR Stage	Active changes on CT	Respiratory Symptoms Changed	ACE	Lymph	Neut	Other organ involved	Fatigue	TLCO	KCO	FVC	FEV	FEV1/ FVC %	Organ biopsied	Length of disease (years)	Therapy started
Active	1	Yes	Yes	165	1.32	3.72	Skin	No	69	.	74	75	86	Skin	24	No
Active	1	Yes	Yes	70	0.86	2.94	Joints	No	97	102	104	95	79	Med LN	1	Yes
Active	1	Yes	Yes	150	1.41	4.49	Joints	Yes	72	96	92	90	83	Med LN	1	Yes
Active	1	Yes	Yes	167	1.5	3.3	No	Yes	112	125	107	112	109	Med LN	12	No
Active	3	Yes	Yes	347	1.1	2.9	No	Yes	.	.	.	.	.	Lung	6	Yes
Active	1	Yes	Yes	98	1.6	3.1	No	Yes	.	.	.	.	.	Med LN	4	Yes
Active	1	Yes	Yes	39	1.9	4.9	No	No	70	81	104	93	71	Med LN	1	No
Active	1	Yes	No	68	0.39	2.32	Skin, neuro	Yes	63	68	.	.	.	Cerv LN	36	Yes
Active	1	Yes	No	56	1.34	4.22	Joints	Yes	71	92	92	93	89	Ing LN	18	Yes
NA	1	.	No	89	1.2	4.8	No	No	.	.	.	.	.	Lung	12	No
NA	4	Yes	No	7	0.7	4.3	No	No	.	.	.	.	.	Lung	96	No
NA	1	.	No	125	1.7	5.5	No	No	41	51	121	122	109	Med LN	12	No
AET	1	.	No	103	2.5	4	Skin	No	.	.	60	.	100	Skin	24	Declined
Active	2	Yes	No	167	1.3	3.6	Eyes	No	42	59	.	48	65	Lung	5	Yes
Active	3	.	No	139	1.7	10.5	No	No	.	.	.	.	.	Skin	1	Yes
Active	4	Yes	No	200	1	4.7	No	Yes	66	58	91	87	.	Med LN	72	Yes
NA	1	No	No	6	0.8	1.8	No	No	56	91	80	73	82	Med LN	131	No
Active	4	Yes	No	89	1.6	4.7	No	No	61	95	102	83	63	Lung	96	Declined
AET	0	.	No	76	0.6	2.6	Skin	Yes	92	90	110	112	79	Med LN	48	Yes
NA	1	.	No	37	2.9	5.6	No	No	.	.	.	.	.	Med LN	24	No
NA	4	No	No	59	1	2.8	No	No	65	89	75	46	52	Med LN	131	No
NA	1	No	No	46	2.2	2.6	No	No	.	.	.	.	.	skin	48	No
Active	3	Yes	Yes	23	1.6	1.7	Liver	No	.	.	.	.	.	Med LN	6	No
NA	1	No	No	47	1.3	2.9	No	No	76	84	95	90	96	Med LN	1	No
NA	4	No	No	77	2.5	4.4	No	No	71	99	91	83	-999	Lung	48	No

Table 16. Clinical variables and sarcoidosis classification of every patient in the training and test sets

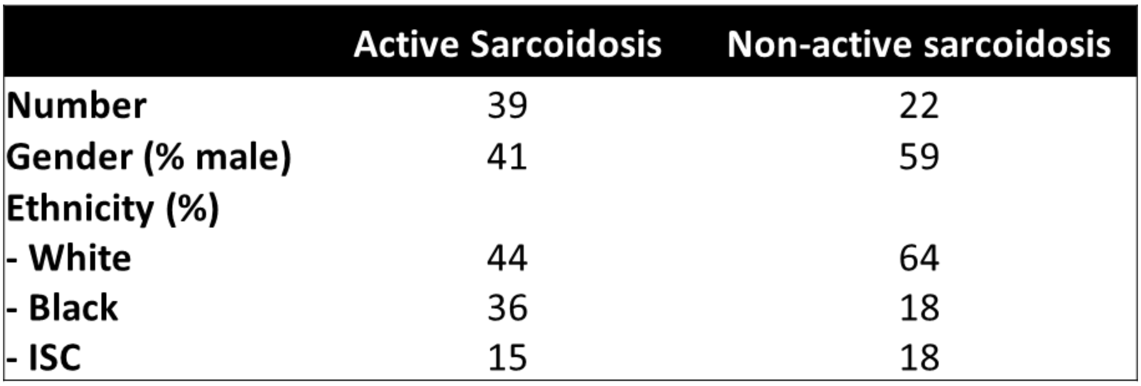
CXR = chest radiograph, CT = computer tomography, ACE = angiotensin converting enzyme, Lymph = lymphocyte count, Neut = neutrophil count, TLCO = transfer factor for carbon monoxide, KCO = transfer coefficient, FVC = forced vital capacity, FEV1 = forced expiratory volume in 1 second, Abdo = abdomen, LN = lymph node, Med = mediastinal, NA = non-active sarcoidosis, AET = active extra-thoracic sarcoidosis, Neuro = neurological disease.



Classified	CXR Stage	Active changes on CT	Respiratory Symptoms Changed	ACE	Lymph	Neut	Other organ involved	Fatigue	TLCO	KCO	FVC	FEV	FEV1/ FVC %	Organ biopsied	Length of disease (months)	Therapy started
NA	1	.	No	42	2.06	3.3	Skin	No	61	90	73	68	69	Lung	14	No
Active	2	Yes	No	142	1.16	3.48	Abdomen, cervical LN	Yes	63	94	59	62	77	Lung & liver	10	Yes
Active	2	Yes	No	31	0.77	3.02	Abdomen	No	91	100	98	101	75	Lung	4	No
NA	1	No	No	18	1.8	2.1		Yes	59	81	86	75	70	Lung	26	No
Active	2	.	No	72	0.45	5.51		No	78	85	92	68	65	Lung	3	No
AET	1	No	No	45	1.85	3.54	Parotid, liver	Yes	73	100	75	72	89	Lung	1	Yes
NA	1	.	No	33	1.21	4.63	Joints	No	109	116	97	97	73	Med LN	18	No
Active	2	.	No	38	0.89	4.19	Skin	No	73	86	85	75	73	Lung	48	No
NA	1	.	No	49	2.48	2.76	Joints	Yes	82	79	114	99	69	Med LN	29	No
NA	1	.	No	33	1.36	1.98		No	75	93	87	78	69	Lung	13	No
Active	3	Yes	No	179	1.9	2.6		Yes	42	42	76	80	91	Lung	10	Declined

Table 17. Clinical variables and sarcoidosis classification of every patient in the validation set.

CXR = chest radiograph, CT = computer tomography, ACE = angiotensin converting enzyme, Lymph = lymphocyte count, Neut = neutrophil count, TLCO = transfer factor for carbon monoxide, KCO = transfer coefficient, FVC = forced vital capacity, FEV1 = forced expiratory volume in 1 second, Abdo = abdomen, LN = lymph node, Med = mediastinal, NA = non-active sarcoidosis, AET = active extra-thoracic sarcoidosis, Neuro = neurological disease.



	Active Sarcoidosis	Non-active sarcoidosis
Number	39	22
Gender (% male)	41	59
Ethnicity (%)		
- White	44	64
- Black	36	18
- ISC	15	18

Table 18. Ethnicity and gender of the sarcoidosis patients divided into their clinical classifications of active or non-active.

ISC = Indian subcontinent.

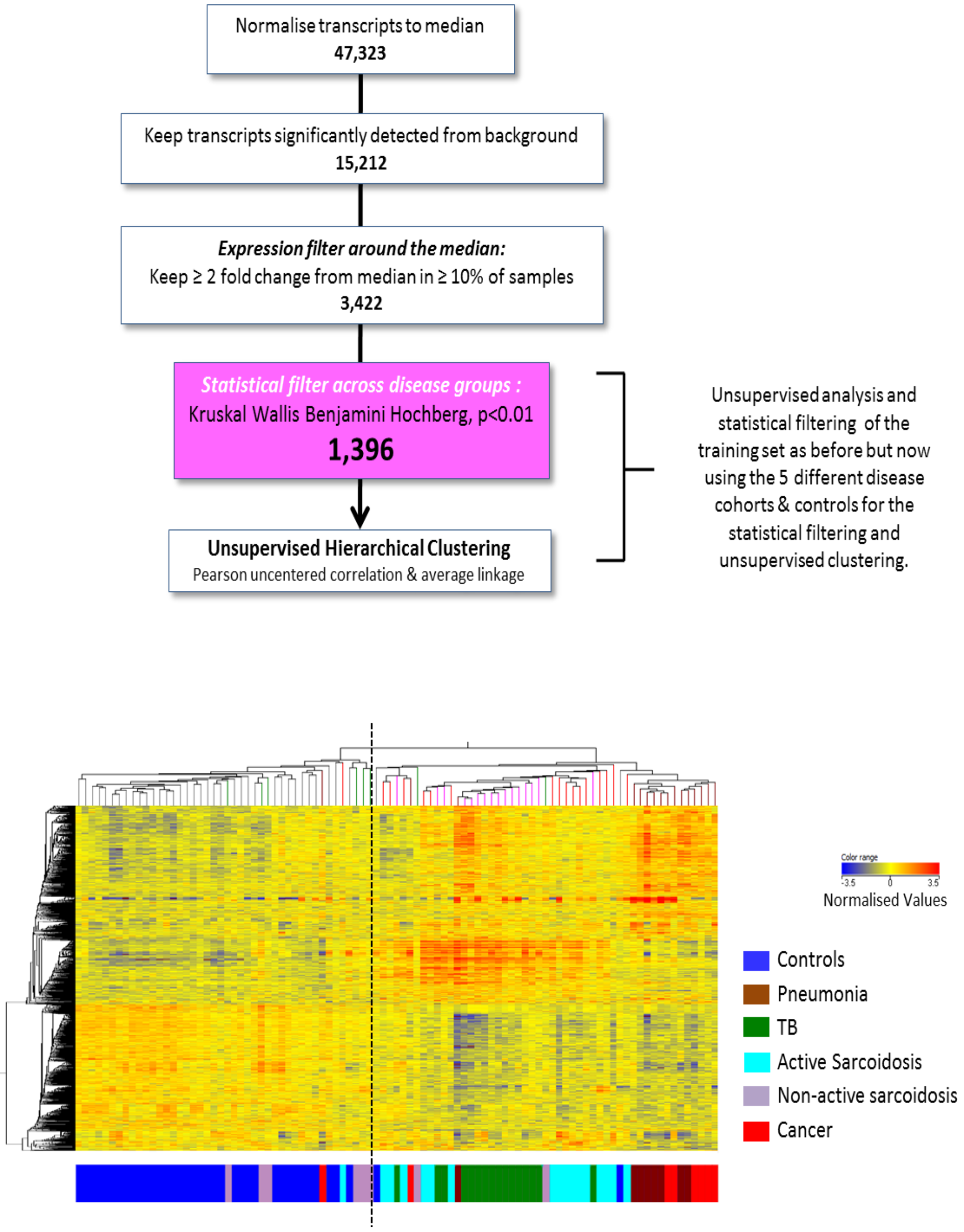
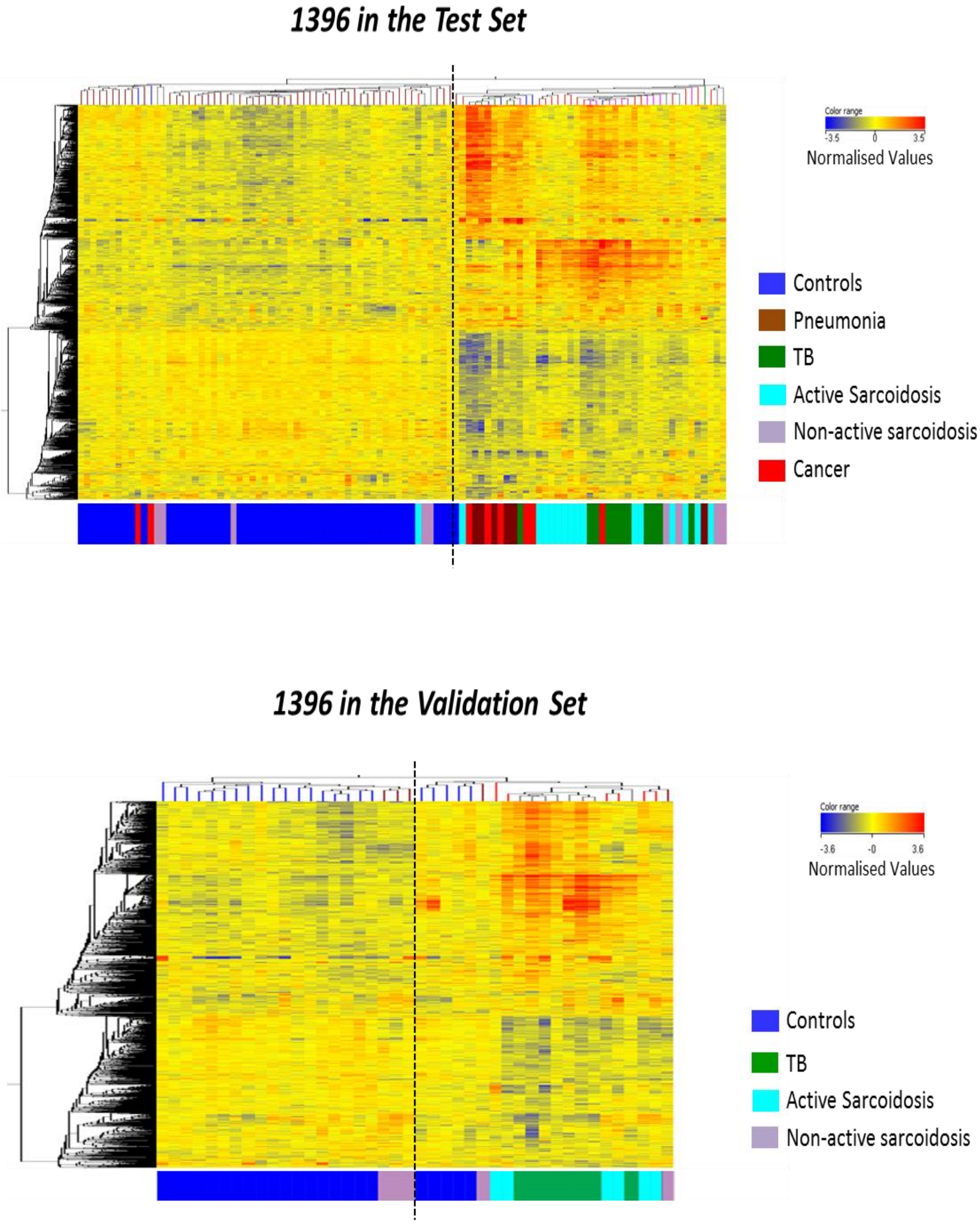


Figure 25. Clustering of the training set sarcoidosis patients correlates well with the clinical classification system.

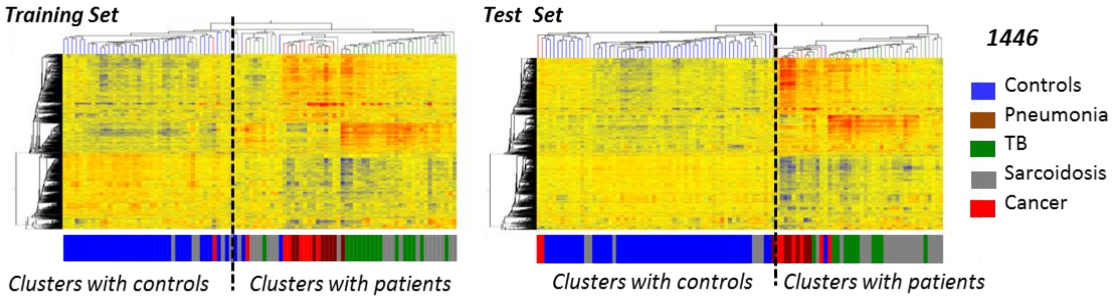
The same unsupervised analysis and statistical filtering was applied as used earlier, but this time the sarcoidosis patients were divided by the clinical classifications system into either active or non-active sarcoidosis, see flow diagram at top of figure. The layout of the heatmap is as described in figure 13.


1396 in the Test Set

1396 in the Validation Set

Figure 26. The 1,396 transcripts derived from the training set also show revealed the active sarcoidosis cluster with the TB profiles in the test and validation set.

The same 1,396 transcripts as derived in figure 25 were applied to two independent cohorts, the test and validation set. Unsupervised hierarchical clustering was then performed in each dataset to determine the clustering of the two sub-groups of sarcoidosis patients. The layout of the heatmap is as described in figure 13.



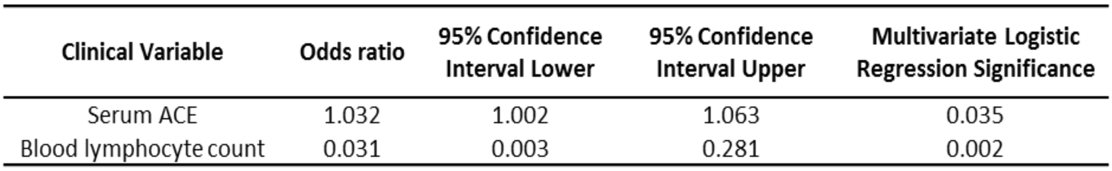
Clinical Variable	Pearson Chi-Square significance
Ethnicity	0.871
FVC (low or normal)	0.823
TLCO (low or normal)	0.791
CXR stage	0.548
Gender	0.468
Involves other organs	0.412
Biopsy site	0.318
Respiratory symptoms (yes or no)	0.266
Past medical history	0.158
CXR stage (2/3 or 1/4)	0.148
Fatigue	0.100
Serum ACE (high or normal, >55IU/L = high)	0.089
Blood lymphocyte count (low or normal, <1x10^9/L = low)	0.026
Commenced treatment	0.025
CT active changes (yes or no)	0.018
Clinical classification system (active or non-active)	**0.000**

Table 19. Individual clinical variables are not as effective at predicting sarcoidosis clustering as the clinical classification system – categorical variables.

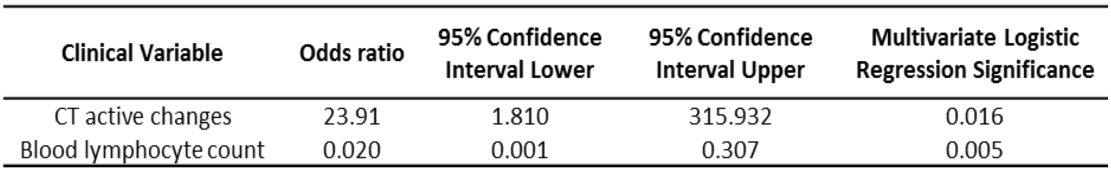
Clinical Variable	Odds ratio	95% Confidence Interval Lower	95% Confidence Interval Upper	Univariate Logistic Regression Significance
TLCO % predicted	0.995	0.955	1.037	0.826
KCO % predicted	1.000	0.970	1.030	0.770
FVC % predicted	1.001	0.967	1.060	0.740
Length of disease	0.995	0.981	1.010	0.530
Age (years)	1.020	0.970	1.010	0.394
FEV1 % predicted	1.019	0.976	1.060	0.383
FEV1/FVC % predicted	1.000	0.999	1.001	0.274
Blood neutrophil count	1.480	0.826	2.671	0.186
Serum ACE	1.017	1.000	1.033	0.039
Blood lymphocyte count	0.050	0.008	0.327	0.002
Clinical classification system	**17.438**	**3.129**	**97.189**	**0.001**

Table 20. Individual clinical variables are not as effective at predicting sarcoidosis clustering as the clinical classification system – continuous variables.

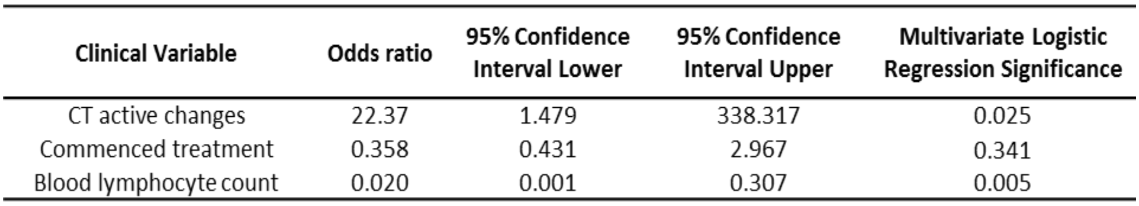
To determine any predictive clinical variables the patients were categorised as either clustering with controls or patients using the 1,446 transcript list in both the training and test set above. The statistical significance of each variable was determined using either Pearson chi-squared for the categorical variables or logistic regression for the continuous variables. All the samples in the training and test set were used to determine any statistical association between the clinical variables and the clustering pattern (11 sarcoidosis samples clustered with the controls and 39 clustered with the patients).



Clinical Variable	Odds ratio	95% Confidence Interval Lower	95% Confidence Interval Upper	Multivariate Logistic Regression Significance
Serum ACE	1.032	1.002	1.063	0.035
Blood lymphocyte count	0.031	0.003	0.281	0.002



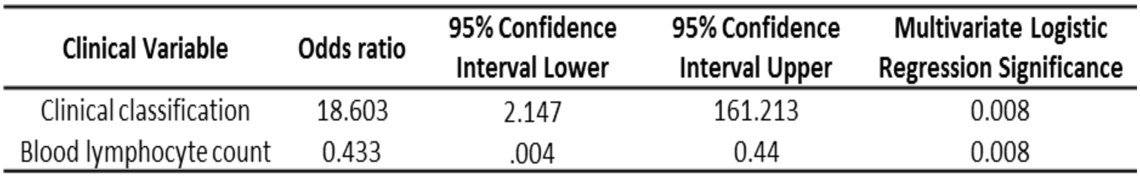
Clinical Variable	Odds ratio	95% Confidence Interval Lower	95% Confidence Interval Upper	Multivariate Logistic Regression Significance
CT active changes	23.91	1.810	315.932	0.016
Blood lymphocyte count	0.020	0.001	0.307	0.005



Clinical Variable	Odds ratio	95% Confidence Interval Lower	95% Confidence Interval Upper	Multivariate Logistic Regression Significance
CT active changes	22.37	1.479	338.317	0.025
Commenced treatment	0.358	0.431	2.967	0.341
Blood lymphocyte count	0.020	0.001	0.307	0.005

Table 21. Multiple clinical variables were also not as effective at predicting sarcoidosis transcriptional clustering as the clinical classification system.

Multivariate regression analysis was performed with variables that appeared the most significant from the single variable analysis.



Clinical Variable	Odds ratio	95% Confidence Interval Lower	95% Confidence Interval Upper	Multivariate Logistic Regression Significance
Clinical classification	18.603	2.147	161.213	0.008
Blood lymphocyte count	0.433	.004	0.44	0.008

Table 22. Adding single clinical variables to the clinical classification system reduced the predictive value of the model.

To determine if the clinical classification system as a prediction model could be improved multivariate regression analysis was performed with the most significant variables from the single variable analysis that had not already been used in the classification system

Chapter 5

Biological patterns of genes are associated with each disease group

Chapter 5: Biological patterns of genes are associated with each disease group

Introduction

Having classified the sarcoidosis patients into two groups: active and non-active, the analysis was continued to examine the differences in the biological pathways between the five different disease groups: TB, active sarcoidosis, non-active sarcoidosis, lung cancer and pneumonia. Using blood gene expression to gain a better understanding or new knowledge about the underlying disease mechanisms has led to many important and novel discoveries (Bleharski, Li et al. 2003; Berry, Graham et al. 2010; Tattermusch, Skinner et al. 2012). Previous publications have shown particular pathways are associated with both TB and sarcoidosis, especially IFN-inducible driven pathways (Berry, Graham et al. 2010; Koth, Solberg et al. 2011; Maertzdorf, Weiner et al. 2012). This study therefore aimed to discover if the same pathways are significant for many respiratory diseases or if there are distinct patterns of genes linked with each of the diseases. Five different data mining strategies were applied to elucidate the functional gene patterns (Figure 6) to ensure there was a strong consistency between results. This is because from previous experience slightly different transcript lists may result in slightly different pathways, for example by IPA (Table 5), therefore consistency between analyses provides much better guidance as to the likely ‘true’ biological findings. In addition different data mining strategies employ diverse techniques for linking functional annotations to sets of genes. Applying different processes should have the advantage of broadening the search for biological links. Previous publications have specifically addressed the issue of the reliability of microarray analysis and demonstrated when comparing multiple different strategies for the same experiments that the results were the same at a biological level although not necessarily at the gene level (Shi, Campbell et al. 2010). This implies analysis of a dataset should not have widely varying outcomes when applying different analysis approaches. All five data mining strategies (modular analysis, gene ranking, IPA, Venn-diagram, comparing disease to disease rather than to controls) are accepted and commonly applied approaches, that have been applied in preceding aforementioned publications. The modular analysis was designed by Chaussabel *et al*, at Baylor Institute for Immunological Research (Chaussabel, Quinn et al. 2008). Each module contains a set of co-expressed genes determined from gene expression profiles from cohorts of different diseases. The modules of similarly expressed genes were extracted by a complex algorithm involving *k*-means clustering. Each module was assigned a biological functional by unbiased literature data mining. Therefore the advantage of the modular analysis is that it is data-driven rather than literature-driven, which is the case for IPA. This study uses a modified version of the modules that were devised from whole blood expression profiles from patients with nine different diseases on an Illumina platform, in addition the output is per patient instead of an average for each disease (Guiducci, Gong et al. 2010). Apart from for the modular analysis where all detectable genes were used, for the other functional analyses each disease group was compared to a matched control group – containing controls with the same ethnicity and gender mix as the disease group. This was particularly important for the cancer and pneumonia groups which had a smaller sample size therefore this reduced any confounding variables.

Results

Modular analysis

The modules are a data-driven analysis tool that can be used for the biological interpretation of gene expression data obtained from different diseases compared to data obtained from a group of controls. The modules are data-driven sets of biologically Even at first glance it could be seen from this analysis that the TB and active sarcoidosis patients appeared alike, while the pneumonia and cancer patients appeared alike to each other but distinct from TB and sarcoidosis (Figure 28). In addition the non-active sarcoidosis patients, apart from a low over-abundance in the interferon modules, did not appear very different from the controls. The key module associated with over-abundance in the TB and active sarcoidosis patients were the three IFN modules (Figure 28). In a few of the TB and active sarcoidosis patients a small percentage of genes were also found to be over-abundant in some of the other modules: the myeloid related modules, inflammation modules and DC/apoptosis module. This is in contrast to the cancer and pneumonia patients where a much higher percentage of genes in all the inflammation modules were over-abundant in all the patients except one cancer patient (Figure 28). In the pneumonia patients a large number of genes in the neutrophil module were also over-abundant and a lower number of genes were over-abundant in the cell death and DC/apoptosis module. A few pneumonia patients also showed a significant over-abundance of genes in the IFN modules, but to a lesser degree than the TB and active sarcoidosis patients (Figure 28). In all the diseases there was under-abundance in the T cell module and in many of the patients in of the B cells modules. The cytotoxic/NK module was under-abundant in many of the TB and pneumonia patients.

By plotting graphs of the average module score per disease of the three disease defining modules it can be seen that TB had the highest percentage of genes in the IFN modules, followed by active sarcoidosis. Whereas pneumonia followed by cancer had the highest percentage of genes in the inflammation modules and pneumonia had the highest percentage of genes in the neutrophil module (Figure 29). Calculating the neutrophil module score for each patient in both the training and test set, there was a strong correlation between the score and the blood neutrophil count ($p<0.05$, data not shown) (Tables 9-11).

Gene ranking

To determine the differentially expressed genes with the highest expression abundance for each disease in its own right and relative to the other diseases, gene ranking was applied. Each disease was first matched to a set of controls containing the same percentage of gender and ethnic groups. Differentially expressed genes were then generated by comparing each disease group in turn to its own matched controls. Genes were called differentially expressed if they satisfied a 1.5 fold change cut-off from the controls and were statistically significant (Mann Whitney unpaired, Benjamini Hochberg $p<0.01$). The numbers of differentially expressed transcripts were 2,524 for TB, 1,391 for active sarcoidosis, 2,801 for pneumonia and 1,626 for lung cancer. Non-active sarcoidosis had no differentially expressed genes using this analysis. The differentially expressed genes were then ranked according to their fold change compared to their matched controls; the top 50 over-abundant genes for each disease are shown (Figure 30). Many IFN-inducible genes were found to be the most over-abundant genes in TB and active sarcoidosis including FCGR1, SERPING1, IFITM3, IFI44L, all the GBPs, IFIT3, AIM2, ISG15, IFI27, WARS, IFI44, CXCL10, OAS1, STAT1 and IFI6 (DeYoung, Ray et al. 1997; Bennett, Palucka et al. 2003; Martens and Howard 2006; Moran, Duke et al. 2007; Berry, Graham et al. 2010). However the degree of fold change for each gene was considerably greater in TB than sarcoidosis. Many neutrophil anti-microbial genes were found to be the most over-abundant genes in pneumonia (Figure 30). All the top ranking genes in cancer were expressed at lower levels than the top genes in the other diseases. The most significantly expressed gene in cancer was ARG1 (arginase 1) which is known to be related to alternatively activated macrophages that are associated with cancer (M2) (Mantovani, Sozzani et al. 2002). All four diseases had over-abundance of FCGR1A, B and C within the top 50 ranking genes.

Ingenuity Analysis Pathways

Next we applied an additional data mining strategy, IPA analysis, to further identify and verify biological pathways associated with each disease. Applying the same differentially expressed genes as was used for the gene ranking, it was found that each disease had a different set of highly significant IPA pathways (Benjamini Hochberg, $p<0.05$, Figure 31). IFN-signalling was in the top 5 significant pathways for TB and active sarcoidosis. The other significant pathways for these granulomatous diseases were also mostly immune related pathways, and contained a number of IFN-inducible genes. The significant pathways in lung cancer were mixed expression of T cell and NK cell signalling. The dominant pathway in pneumonia was under-abundance of the EIF2 signalling pathway, associated with mRNA translation and protein secretion.

‘Comparison IPA’ analysis involves looking at all four diseases simultaneously allowing identification of the pathways that are most significant in one disease compared to the others. The top four significant pathways by this analysis were EIF2 signalling – driven by pneumonia; IFN signalling – driven by TB and active sarcoidosis; the role of PRRs in recognition of bacteria and viruses – driven by TB; and the antigen presentation pathway – driven by both TB and active sarcoidosis (Figures 32 & 33). Both the significance of the pathway relative to each disease (Figure 32) and the percentage of genes in each pathway relative to each disease are shown (Figure 33). The significance of each pathway approximately parallels the percentage of genes in that pathway such that the disease with the highest significance tends to also be the disease with the highest number of perturbed genes relative to the controls in that pathway.

Apart from the top four significant pathways by ‘comparison IPA’ analysis, many other significant pathways were more associated with one disease than another (Tables 24-26). TB correlated the most with over-abundance of other immune pathways including DC maturation, crosstalk between DCs and NK cells, and communication between innate and adaptive cells, notably many molecules within these pathways were also present in the IFN-signalling pathway (Table 24). Active sarcoidosis was not associated more significantly than any of the other diseases with any of the pathways. The foremost significant pathways associated with pneumonia were under-abundance of the EIF2 signalling and three T-cell pathways, and over-abundance of the apoptosis signalling pathway (Table 25). The main significant pathways associated with cancer, relative to the other diseases, were mixed expression of the NK signalling pathway, CTLA4 signalling and hepatocyte growth factor (HGF) signalling (Table 26).

The differentially expressed genes for each disease were overlaid on the IFN-signalling pathway such that a comparison of the abundance of the IFN-signalling genes between the diseases could be easily visualised (Figure 34). This comparison demonstrated that most of the genes were over-abundant in TB and active sarcoidosis, compared to very few in cancer and pneumonia (Figure 34). In TB and active sarcoidosis both IFN Type I and Type II pathways were involved whereas only the IFN-γ receptors were involved in pneumonia and lung cancer.

The EIF2 signalling pathway is shown for the pneumonia patients to demonstrate that most of the genes involved in mRNA translation/protein secretion are under-abundant (Figure 35). Some of the signalling genes (MAPK, PIK3, RAF) that are over-abundant are involved in signalling in many other pathways. To further elucidate the expression of the genes involved in protein translation all the eukaryotic translation initiation factors (EIFs) and ribosomal proteins that were at least 1.3 fold change from the controls were selected (Figure 36). Although pneumonia had the lowest relative expression values, the other diseases also had low expression levels relative to the controls. The key players in the unfolded protein response (UPR), an endoplasmic stress pathway intrinsically involved with regulating protein synthesis, were also selected (Table 27). These genes were predominantly either not expressed relative to the controls, or under-abundant, in all the diseases. Only ATF6 was over-abundant. Again expression was more marked in pneumonia than the other diseases.

4-set Venn diagram

To determine unique disease-related transcripts a 4-set Venn diagram was used to assess the differentially expressed genes from each disease compared to their matched controls. The same differentially expressed genes as used for the IPA and gene ranking were inserted into a 4-set Venn diagram. This allowed easy visualisation of the number of overlapping and unique transcripts (Figure 37). The ratio of unique-disease related transcripts across the diseases matched the same trend as was seen by the MDTH scores, where pneumonia had the largest number, followed by TB, and cancer and active sarcoidosis had the smallest number of unique-disease related transcripts (Figure 19, 37 & 38). IPA was applied for each set of unique transcripts and the overlapping transcripts (Figure 39). Due to the low number of genes in most of the unique disease-related transcripts there were no significant pathways for active sarcoidosis, TB or cancer but there were significant pathways for pneumonia after, applying a multiple testing correction (Benjamini Hochberg *p*>0.05). However the pathways that were significant without multiple testing correction were found to be in accordance with the pathways attained by the earlier IPA analysis. The overlapping 375 transcripts at the centre of the Venn diagram (grey segment) were associated with under-abundance of T cell pathways, also in accordance with pathways attained earlier using different data mining strategies.

Comparing similar diseases to each other not the controls

From all the previous functional analysis the diseases TB and active sarcoidosis profiles had revealed very similar pathways; therefore the diseases were directly compared to each other to help uncover the dissimilarities between them. The differentially expressed genes (144 transcripts) were obtained by the less stringent statistical test, significance of microarray (SAM), followed by a 1.5 fold change between the two diseases (Figure 40). Unsupervised hierarchical clustering of the 144 transcripts demonstrated this transcript list was able to distinguish the TB patients from the sarcoidosis patients, and also from all the patients and controls. Many of the 144 transcripts are known to be IFN-inducible (Figure 41) but due to the small number of genes no IPA pathways were found to be significantly associated with the transcripts, although IFN-signalling was the top pathway identified without multiple testing correction.

From all the previous functional analyses, the diseases pneumonia and lung cancer have also revealed very similar pathways, therefore again they were directly compared to each other. Both groups had similar ethnicities, gender and age distribution. The differentially expressed genes (1,165 transcripts) were obtained using a 1.5 fold change cut-off and SAM (False discover rate, *q*<0.05). Of the top 50 overexpressed genes in the pneumonia patients, 46% were related to innate cells, in particular neutrophils, while 10% were related to apoptosis (Figure 42). In the top 50 overexpressed genes in the cancer patients, 40% were related to cellular processes such as cell differentiation, signalling and protein ubiquitination and 10% of genes were related to transcriptional regulation (Figure 42).

Validation of the functional gene patterns in the test set

The same functional analysis methods used for the training set were also applied to the test set (Figure 43). The modular analysis was consistent with the training set and showed over-abundance of the IFN modules in the TB and active sarcoidosis patients, while there was again a dominant over-abundance of the inflammation modules in the cancer and pneumonia patients. T and B cells were also under-abundant in all diseases. The neutrophil and DC/apoptosis modules were, as found in the training set, the most over-abundant in the pneumonia patients. When just looking at the top ranking genes, again similar patterns were observed; many IFN-inducible genes were associated with TB and active sarcoidosis, and neutrophil genes were associated with pneumonia patients (data not shown). The top genes in the cancer patients' differentially expressed transcript list were once more the lowest relatively expressed genes, and included ARG1 in the top three (data not shown). IPA analysis also generated comparable pathways as were seen in the training set. IFN-signalling was highly significant for TB and active sarcoidosis, along with the other immune pathways seen in the training set. As before EIF2 signalling was most significantly associated with pneumonia. The same signalling pathways, NK cells, CTLA4, and HGF signalling were all significantly associated with cancer.

Discussion

By performing pathway analysis of the differentially expressed genes for each disease group separately distinct biological patterns of genes were uncovered and shown to be associated with TB, active sarcoidosis, community acquired pneumonia and primary lung cancer. This was consistently achieved through several different processes therefore adding strength to the findings.

Active tuberculosis

By all the functional analyses applied, the IFN-inducible genes were shown to be over-abundant in the TB patients as compared to the controls. Furthermore their relative expression was greater in the TB patients than in the other three similar respiratory diseases. The modular analysis revealed TB was significantly related to many of the genes present in the IFN modules (Figure 28 & 29), the gene ranking showed many of the top 50 over-abundant genes were IFN-inducible (Figure 30), IPA analysis demonstrated IFN-signalling was the most significant pathway for TB (Figure 31) and by 'comparison IPA' analysis IFN-signalling was more significantly associated with TB than the other diseases (Figures 32-33). Lastly the Venn diagram proved that even the unique TB genes were dominated by IFN-inducible genes (Figure 39). Berry *et al.* 2010, had previously also shown in a different cohort of TB patients, the IFN-signalling IPA pathway was the most significant, and using a slightly older version of the modules that a large percentage of the genes in the IFN module were significantly associated with TB (Berry, Graham et al. 2010). In addition other more recent whole blood gene expression studies have also confirmed the dominant presence of interferon signalling in active TB (Lesho, Forestiero et al. 2011; Maertzdorf, Ota et al. 2011; Maertzdorf, Repsilber et al. 2011). Berry *et al*. 2010, also found over-abundance in a large percentage of the genes within the myeloid modules that were present in the older version of the modules used in their study. The older myeloid modules are no longer identically defined the same as in the newer modules; however many of the myeloid related modules in the newer modules were also significantly related and over-abundant in the TB patients. Other functional relationships found to be more significant for TB than the other diseases included over-abundance of host immune response pathways such as pattern recognition receptors, dendritic cell related pathways, antigen presentation, communication between innate and adaptive cells (Figure 32 & Table 24). The association of TB with the activation of these particular biological pathways is reflective of our current understanding of the underlying host response towards *M. tuberculosis*.

Active sarcoidosis

Like TB, by all the data mining strategies the IFN-inducible genes were found to be over-abundant in the active sarcoidosis patients compared to the controls, although the absolute numbers and level of expression of the IFN-inducible genes were significantly lower than in TB (Figure 29). Analysis revealed active sarcoidosis was significantly related to many of the genes present in the IFN modules (Figure 28 & 29), the gene ranking showed many of the top 50 over-abundant genes were IFN-inducible (Figure 30), IPA analysis demonstrated IFN-signalling was one of the most significant pathway (Figure 31) and by 'comparison IPA' analysis IFN-signalling was found to be highly significant compared to pneumonia and cancer (Figures 32-33). However the Venn diagram did not find many of the unique active sarcoidosis genes were IFN-inducible (Figure 39). Two former studies have correspondingly found comparable results, as they both described IFN-signalling IPA pathway to be significantly associated with TB and sarcoidosis (Koth, Solberg et al. 2011; Maertzdorf, Repsilber et al. 2011). However neither study commented on the absolute numbers or fold change values for the sarcoidosis patients relative to the TB patients. Our current study identified lower numbers of IFN-inducible genes and lower expression values in the active sarcoidosis patients compared to the TB patients (Figure 29 & 30). Moreover on direct comparison between the two diseases, several IFN-inducible genes were found to be significantly over-abundant by at least 1.5 fold in the TB patients relative to the active sarcoidosis patients (Figure 41). Active sarcoidosis, like TB, showed many other host immune response pathways such as pattern recognition receptors and antigen presentation to be significantly over-abundant. The closeness of the biological pathways that are found to be involved in both active pulmonary TB and active pulmonary sarcoidosis by blood gene expression analysis strengthens the implication that the underlying immunological processes have much in common (Gerke and Hunninghake 2008). Active sarcoidosis was not significantly associated with particular pathways more than the other diseases by ‘comparison IPA’ analysis. This may have been related to the lower number of differentially expressed genes for the sarcoidosis patients (1391) compared to TB patients (2524) or pneumonia patients (2801) (Figure 19).

Non-active sarcoidosis

Non-active sarcoidosis patients revealed no differentially expressed genes when compared to the controls. This was undoubtedly related to its quiet transcriptional profile relative to the controls and in part the small sample size. Although interestingly for such a quiescent disease the modular analysis demonstrated a significant over-abundance of a small number of genes in the IFN modules (Figure 28).

Community acquired pneumonia

Pneumonia is an infection of the respiratory tract that results in an acute inflammation of the lungs, although the inflammation is not just localised to the site of disease but can also occur systemically (Fernandez-Serrano, Dorca et al. 2003; Windgassen, Funtowicz et al. 2011). Implication of a systemic inflammatory response was supported by both the modular and IPA pathway analysis (Figures 20 & 28). Inflammation can occur in response to numerous stimuli, including infectious agents, and is a very broad category encompassing huge numbers of molecules of which some are pro-inflammatory, anti-inflammatory or can be both depending on the situation. In keeping with an over-abundant inflammatory response secondary to an infectious bacterial disease, pneumonia contained the highest percentage of genes in the neutrophil module (Figure 29) and many neutrophil genes related to their antimicrobial activity appeared in the top ranking differentially expressed genes (Figure 30). Moreover when comparing pneumonia to lung cancer, nearly half of the top over-abundant genes were associated with the innate immune response including neutrophil-related genes (Figure 42). Four of the pneumonia patients also had a small percentage of significantly over-abundant genes in the IFN module (Figure 28). This may have been part of the primary immune response to a bacterial pneumonia or may be reflecting a viral component to their infection. This finding has also been observed previously in a paediatric cohort of patients with *streptococcal pneumonia* infection (Ramilo, Allman et al. 2007). Although none of the pneumonia patients in this study reported symptoms classical for viral infections, it is known that symptoms are a poor indicator of the causal agent (Farr, Kaiser et al. 1989).

‘Comparison IPA’ analysis and the unique-disease genes generated by the Venn diagram led to the discovery of the EIF2 signalling, mTOR signalling and regulation of EIF4 and P70S6K signalling pathways as being far more significant and under-abundant in pneumonia than in the other diseases (Figures 32, 33, 39 & Table 25). All these pathways enclose multiple genes relating to mRNA translation and protein secretion, predominantly the eukaryotic initiation factors (EIFs) and ribosomal proteins – including genes encoding for components of both the 40S and 60S subunits. Nearly all the relevant genes were under-abundant and many more were significantly associated with pneumonia than in the other diseases (Figure 35 & 36). A reasonable explanation for this could be the preferential migration of protein making cells to the site of infection/inflammation and a simultaneous preservation of energy at sites away from the source of infection/inflammation. In agreement with this theory the unfolded protein response (UPR) also appeared to be dampened or non-existent in the blood profiles of pneumonia patients (Table 27), while apoptosis signalling was significantly over-abundant compared to the other diseases (Table 25 & Figure 28).

Messenger RNA translation is a highly regulated process that is coordinated by signalling from the endoplasmic reticulum (ER) to regulate the assembly, speed and accuracy of the folding of proteins, such that only properly folded proteins leave the ER to reach the cell surface (Kaufman 2004). The UPR is an ER stress pathway that safeguards cells from the accumulation of misfolded proteins that can occur at times of cellular stress; it is activated in numerous disease processes including diabetes, cancer and neurodegenerative disorders (Walter and Ron 2011). The UPR has two outcomes, involving at least three mechanisms, either it will restore the cell’s homeostasis by stopping translation or if the stress remains unmitigated within a certain time limit it causes the cell to apoptose (Figure 27) (Walter and Ron 2011). The balance between the body’s profit and loss secondary to the UPR may depend on the underlying pathology. For example it induces detrimental effects in diabetes where excessive demand for insulin on pancreatic cells results in apoptosis, also in viral infections it appears the virus can manipulate the UPR to assists in its own replication by the ER (Walter and Ron 2011). Translation regulation pathways are critical for certain immunological functions including dendritic cell activation by pathogens, antigen processing, cytokine production and differentiation of T cells (Pierre 2009). Although the role of the UPR has been inferred in many viral infections, it has not been widely associated with bacterial infections but a recent study confirmed activation of the UPR in a mouse cell line infected with *Listeria monocytogenes*, which led to improved antimicrobial killing by ER-stress induced apoptosis (Pillich, Loose et al. 2012). Furthermore in macrophages isolated from granulomas from active TB patients and *M.tuberculosis* infected mice, ER stress markers were found to be up-regulated in conjunction with an abundance of apoptotic cells (Seimon, Kim et al. 2010). While these studies examined the cells at the site of infection, this current study is looking at the peripheral blood away from the site of infection, which may explain the under-abundance observed. The cells in the blood maybe conserving energy by reducing protein translation while rapid and excessive protein translation is taking place in response to the acute infection/inflammation occurring in the lung.

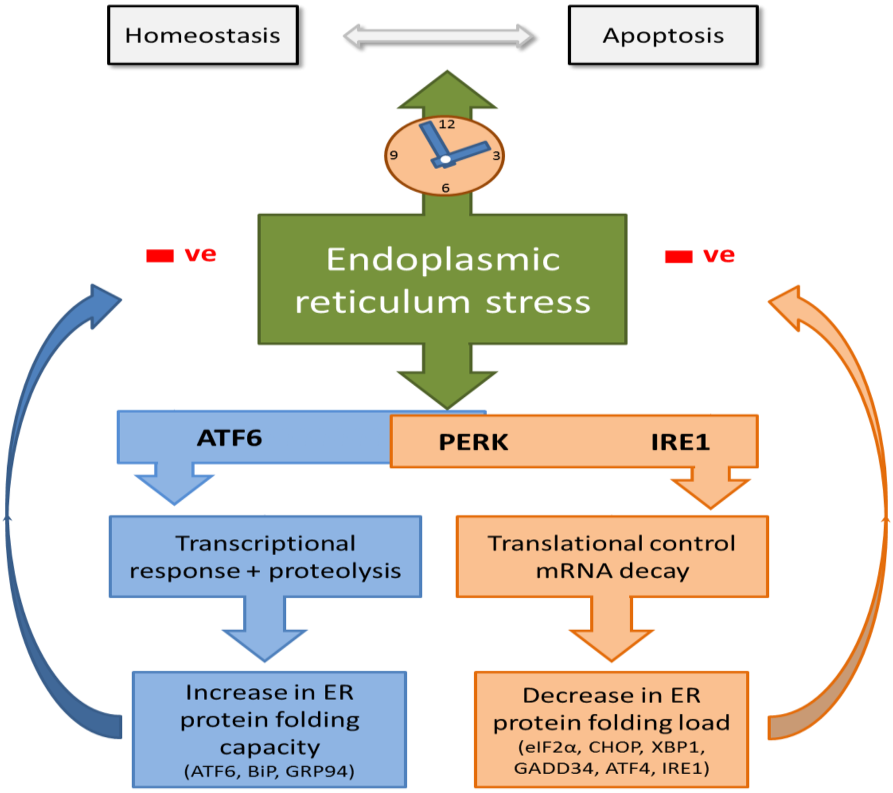

Homeostasis
Apoptosis
12
9
3
6
Endoplasmic reticulum stress
− ve
− ve
ATF6
PERK
IRE1
Transcriptional response + proteolysis
Translational control mRNA decay
Increase in ER protein folding capacity (ATF6, BiP, GRP94)
Decrease in ER protein folding load (eIF2α, CHOP, XBP1, GADD34, ATF4, IRE1)

Figure 27. The unfolded protein response.

Adapted from Walter *et al* Science, 2011.

Primary lung cancer

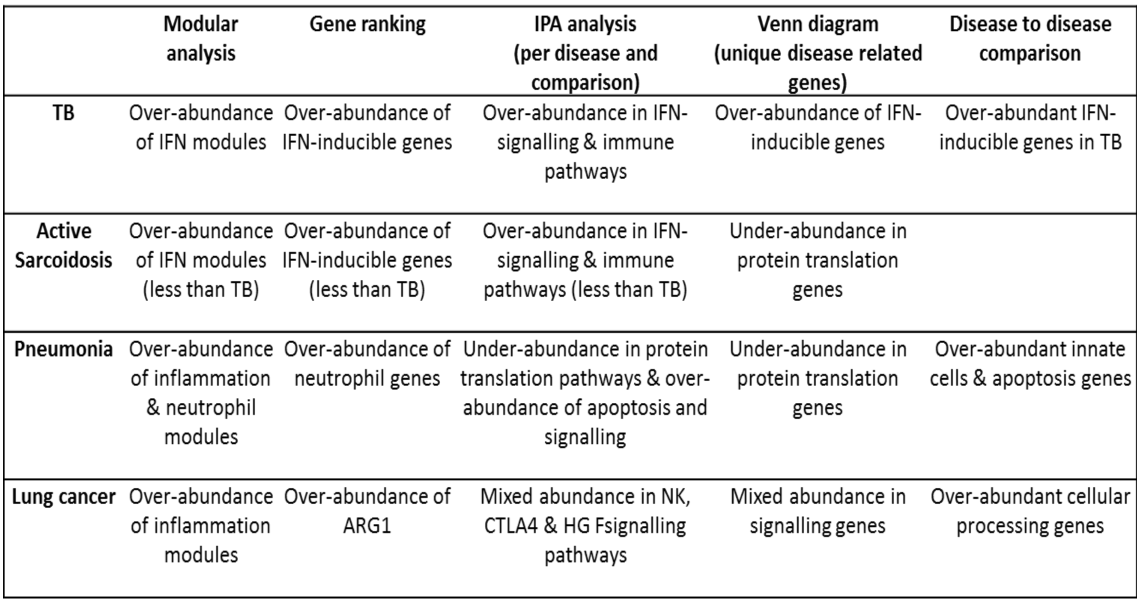
Cancer, like pneumonia, showed an association with the inflammation pathways (Figures 20, 28 & 29). Inflammation has been shown to play a part in cancer whether it is the stimulus or pathological outcome, for example in primary lung cancer there is increasing evidence that smoking encourages lung inflammation (O'Callaghan, O'Donnell et al. 2010). Interestingly one of the top over-abundant genes in cancer was ARG1 (arginase 1) which is known to be associated with the alternatively activated macrophages (M2) (Mantovani, Sozzani et al. 2002). It has been suggested that M2 macrophages are involved in the prevention of the adaptive immune response and enhancement of pro-tumour inflammation pathways, resulting in the promotion of cancer progression and metastasis (Mantovani, Sozzani et al. 2002). From the ‘comparison IPA’ analysis three pathways were identified as more significantly associated with cancer than the other diseases: NK cell signalling, CTLA4 signalling in cytotoxic T lymphocytes and HGF signalling (Table 26). Since the 1980s clinical trials have used NK cell-based immunotherapies against cancer, although their efficacy has on the whole been poor they still remain a promising tool due to their ability to migrate towards inflammation sites and kill target cells without previous activation (Zamai, Ponti et al. 2007). In contrast CTLA4 (cytotoxic T lymphocyte-associated antigen 4) is a T cell receptor that down-regulates the T cell response and CTLA4-blockade is already an established treatment against malignant melanoma and in clinical trials for prostate cancer (Kwek, Cha et al. 2012). Hepatocyte growth factor receptor (HGF) and its receptor (tyrosine kinase MET) are integrally involved in cell survival and migration, with cancer cells using these functions to their advantage for invasion and metastasis (Gherardi, Birchmeier et al. 2012). In non-small cell lung cancer aberrant activity is correlated with poor prognosis and in addition resistance to the EGFR inhibitors can occur through MET signalling. HGF-MET inhibitors have displayed good efficacy in Phase III trials for lung cancer, as well as showing benefits for patients with resistance to EGFR treatment (Gherardi, Birchmeier et al. 2012).

All diseases

All the diseases showed under-abundance of the modules and IPA pathways relating to the T and B cells (Figures 20, 28 and 39). Reduced numbers of T and B cells in the blood of active TB patients, sarcoidosis patients and bacterial infection have previously been demonstrated by flow cytometry analysis (Ardura, Banchereau et al. 2009; Berry, Graham et al. 2010; Sweiss, Salloum et al. 2010). This could be due to preferential migration of the immune cells to the site of disease or cell death as a consequence of the pathogenesis. The percentage of lymphocytes found in the bronchoalveolar lavage from sarcoidosis and active TB patients is higher than in healthy controls, possibly suggesting a preferential migration (Hoheisel, Tabak et al. 1994).

Chapter Summary

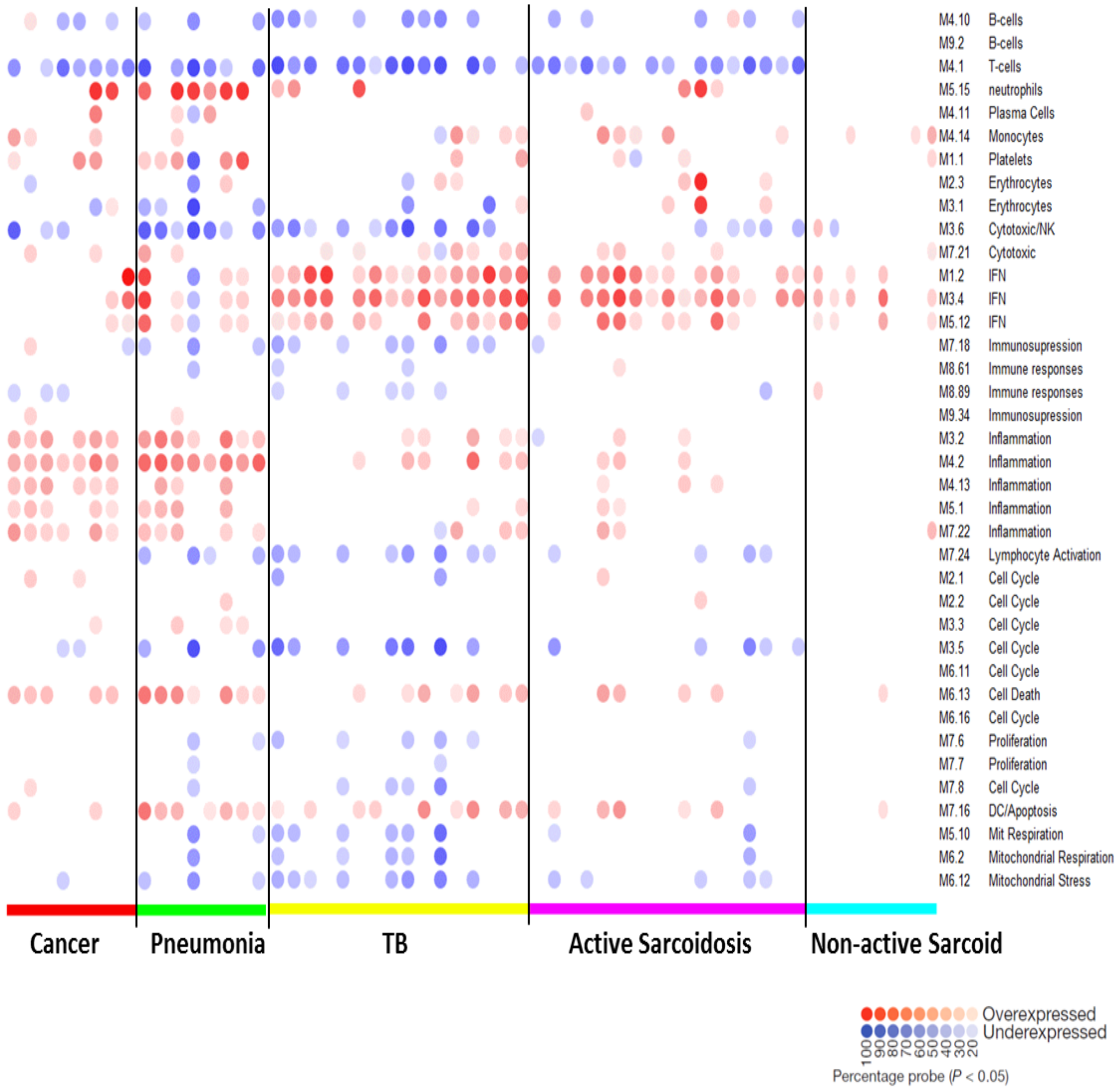
Through several data mining strategies distinct biological pathways were allocated from the differentially expressed genes for each of the four similar respiratory diseases (Table 23). TB and active sarcoidosis were significantly associated with IFN-inducible genes, as shown previously. However a novel finding identified was the increased number and higher expression level of the IFN-inducible genes in TB compared to sarcoidosis. Pneumonia was significantly associated with an over-abundance of inflammatory, neutrophil antimicrobial and apoptosis genes, and an under-abundance of protein translation genes. Lung cancer was associated with an over-abundance of inflammatory genes and an alteration in the abundance of three signalling pathways (NK, CTLA4 and HGF) known to be therapeutic targets. These findings add to the accumulating evidence of the value of blood expression profiling in understanding pathogenesis of diseases.



	Modular analysis	**Gene ranking**	**IPA analysis (per disease and comparison)**	**Venn diagram (unique disease related genes)**	**Disease to disease comparison**
TB	Over-abundance of IFN modules	Over-abundance of IFN-inducible genes	Over-abundance in IFN-signalling & immune pathways	Over-abundance of IFN-inducible genes	Over-abundant IFN-inducible genes in TB
Active Sarcoidosis	Over-abundance of IFN modules (less than TB)	Over-abundance of IFN-inducible genes (less than TB)	Over-abundance in IFN-signalling & immune pathways (less than TB)	Under-abundance in protein translation genes	
Pneumonia	Over-abundance of inflammation & neutrophil modules	Over-abundance of neutrophil genes	Under-abundance in protein translation pathways & over-abundance of apoptosis and signalling	Under-abundance in protein translation genes	Over-abundant innate cells & apoptosis genes
Lung cancer	Over-abundance of inflammation modules	Over-abundance of ARG1	Mixed abundance in NK, CTLA4 & HG Fsignalling pathways	Mixed abundance in signalling genes	Over-abundant cellular processing genes

Table 23. Summary of significant findings from the data mining strategies.

Figures for chapter 5



Modules were derived from clusters of transcriptionally co-regulated genes that were identified from a large dataset of patient's blood expression profiles from nine different diseases. Each module has a functional theme.

Figure 28. Modular analysis reveals functional similarities and differences between the diseases in the training set.

The 15,212 transcripts that were significantly detected compared to the background intensity (see figure 14) were applied to the modular analysis (*p*<0.01). Each module is coloured according to the number of expressed genes relative to the controls, such that red is over-abundant and blue under-abundant while no colour represents the genes are not significantly different from the controls (*p*<0.05). The deeper red or blue colour correlates with a higher percentage of genes in that module found to be significantly different from the controls.

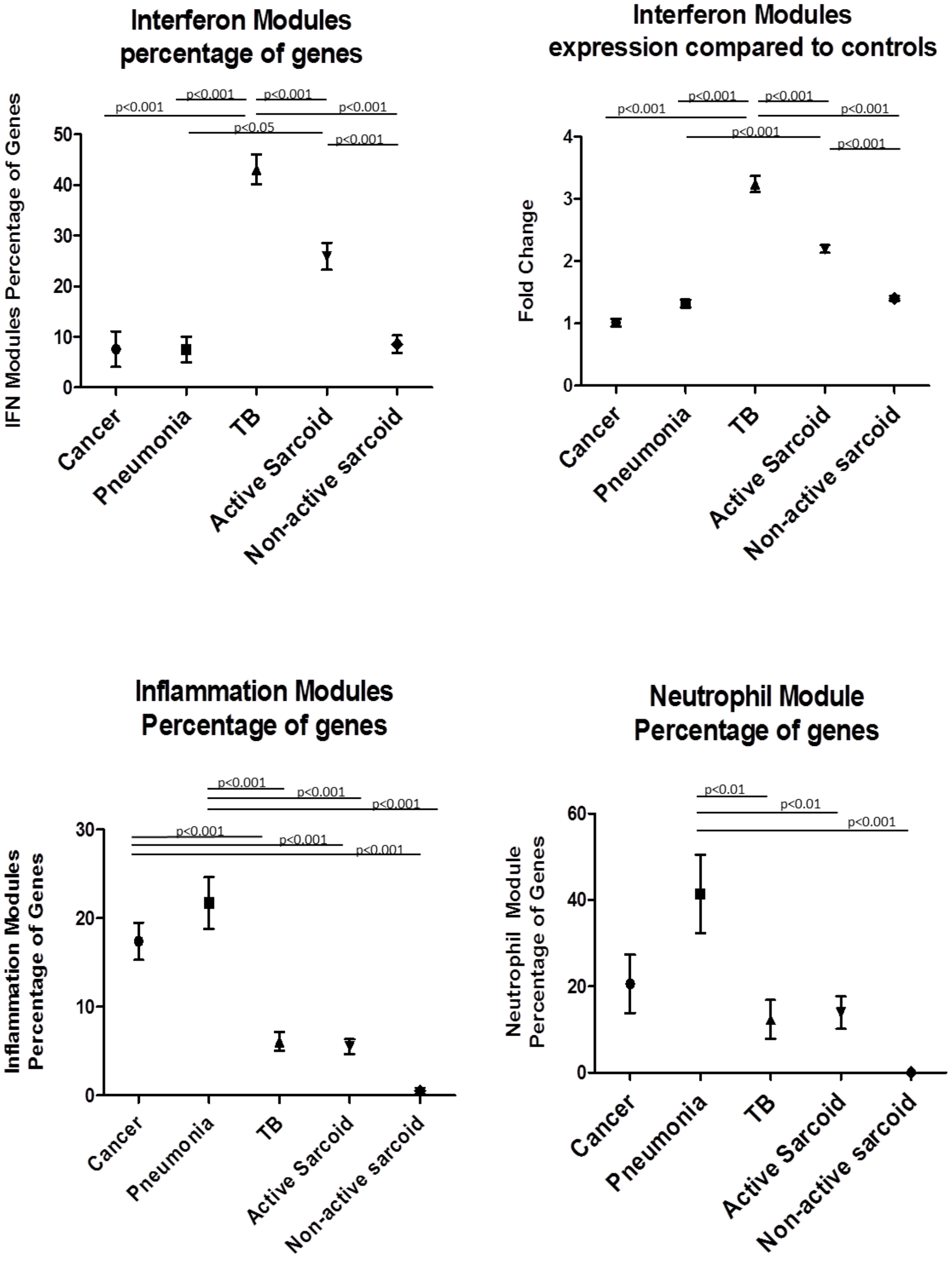
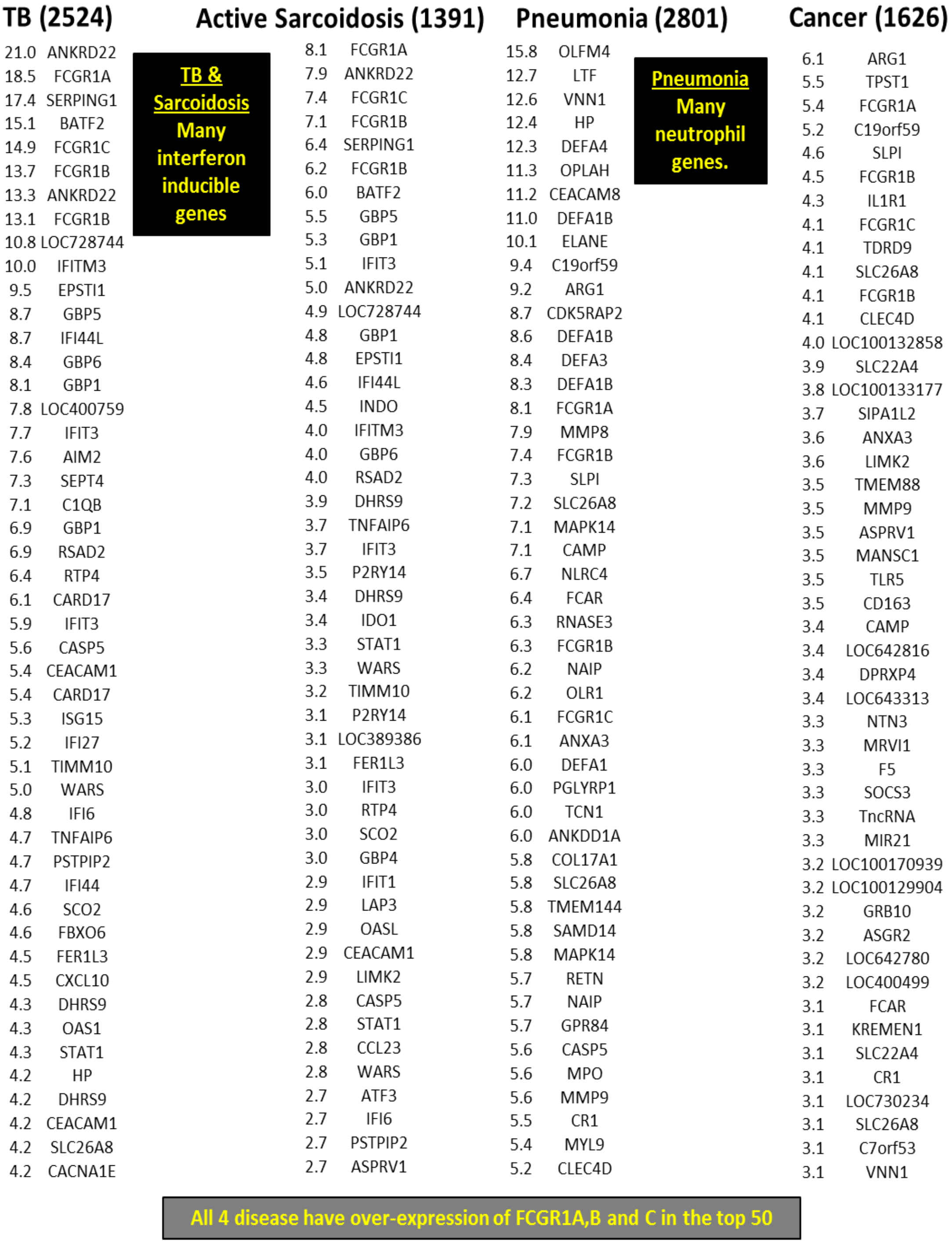


Figure 29. Percentage of over-abundant genes for each of the key modules; interferon modules, inflammation modules and the neutrophil module.

The graphs display the mean, SEM and *p* values from ANOVA with Tukey's multiple comparison test.



TB (2524)		Active Sarcoidosis (1391)		Pneumonia (2801)		Cancer (1626)	
21.0	ANKRD22	8.1	FCGR1A	15.8	OLFM4	6.1	ARG1
18.5	FCGR1A	7.9	ANKRD22	12.7	LTF	5.5	TPST1
17.4	SERPING1	7.4	FCGR1C	12.6	VNN1	5.4	FCGR1A
15.1	BATF2	7.1	FCGR1B	12.4	HP	5.2	C19orf59
14.9	FCGR1C	6.4	SERPING1	12.3	DEFA4	4.6	SLPI
13.7	FCGR1B	6.2	FCGR1B	11.3	OPLAH	4.5	FCGR1B
13.3	ANKRD22	6.0	BATF2	11.2	CEACAM8	4.3	IL1R1
13.1	FCGR1B	5.5	GBP5	11.0	DEFA1B	4.1	FCGR1C
10.8	LOC728744	5.3	GBP1	10.1	ELANE	4.1	TDRD9
10.0	IFITM3	5.1	IFIT3	9.4	C19orf59	4.1	SLC26A8
9.5	EPSTI1	5.0	ANKRD22	9.2	ARG1	4.1	FCGR1B
8.7	GBP5	4.9	LOC728744	8.7	CDK5RAP2	4.1	CLEC4D
8.7	IFI44L	4.8	GBP1	8.6	DEFA1B	4.0	LOC100132858
8.4	GBP6	4.8	EPSTI1	8.4	DEFA3	3.9	SLC22A4
8.1	GBP1	4.6	IFI44L	8.3	DEFA1B	3.8	LOC100133177
7.8	LOC400759	4.5	INDO	8.1	FCGR1A	3.7	SIPA1L2
7.7	IFIT3	4.0	IFITM3	7.9	MMP8	3.6	ANXA3
7.6	AIM2	4.0	GBP6	7.4	FCGR1B	3.6	LIMK2
7.3	SEPT4	4.0	RSAD2	7.3	SLPI	3.5	TMEM88
7.1	C1QB	3.9	DHRS9	7.2	SLC26A8	3.5	MMP9
6.9	GBP1	3.7	TNFAIP6	7.1	MAPK14	3.5	ASPRV1
6.9	RSAD2	3.7	IFIT3	7.1	CAMP	3.5	MANSC1
6.4	RTP4	3.5	P2RY14	6.7	NLRC4	3.5	TLR5
6.1	CARD17	3.4	DHRS9	6.4	FCAR	3.5	CD163
5.9	IFIT3	3.4	IDO1	6.3	RNASE3	3.4	CAMP
5.6	CASP5	3.3	STAT1	6.3	FCGR1B	3.4	LOC642816
5.4	CEACAM1	3.3	WARS	6.2	NAIP	3.4	DPRXP4
5.4	CARD17	3.2	TIMM10	6.2	OLR1	3.4	LOC643313
5.3	ISG15	3.1	P2RY14	6.1	FCGR1C	3.3	NTN3
5.2	IFI27	3.1	LOC389386	6.1	ANXA3	3.3	MRVI1
5.1	TIMM10	3.1	FER1L3	6.0	DEFA1	3.3	F5
5.0	WARS	3.0	IFIT3	6.0	PGLYRP1	3.3	SOCS3
4.8	IFI6	3.0	RTP4	6.0	TCN1	3.3	TncRNA
4.7	TNFAIP6	3.0	SCO2	6.0	ANKDD1A	3.3	MIR21
4.7	PSTPIP2	3.0	GBP4	5.8	COL17A1	3.2	LOC100170939
4.7	IFI44	2.9	IFIT1	5.8	SLC26A8	3.2	LOC100129904
4.6	SCO2	2.9	LAP3	5.8	TMEM144	3.2	GRB10
4.6	FBXO6	2.9	OASL	5.8	SAMD14	3.2	ASGR2
4.5	FER1L3	2.9	CEACAM1	5.8	MAPK14	3.2	LOC642780
4.5	CXCL10	2.9	LIMK2	5.7	RETN	3.2	LOC400499
4.3	DHRS9	2.8	CASP5	5.7	NAIP	3.1	FCAR
4.3	OAS1	2.8	STAT1	5.7	GPR84	3.1	KREMEN1
4.3	STAT1	2.8	CCL23	5.6	CASP5	3.1	SLC22A4
4.2	HP	2.8	WARS	5.6	MPO	3.1	CR1
4.2	DHRS9	2.7	ATF3	5.6	MMP9	3.1	LOC730234
4.2	CEACAM1	2.7	IFI6	5.5	CR1	3.1	SLC26A8
4.2	SLC26A8	2.7	PSTPIP2	5.4	MYL9	3.1	C7orf53
4.2	CACNA1E	2.7	ASPRV1	5.2	CLEC4D	3.1	VNN1

TB & Sarcoidosis Many interferon inducible genes

Pneumonia Many neutrophil genes.

All 4 disease have over-expression of FCGR1A,B and C in the top 50

Figure 30. The top 50 differentially expressed genes for each disease demonstrates the dominance of the interferon inducible genes in both TB and sarcoidosis.

Differentially expressed genes between each disease group and their matched-controls, by ethnicity and gender, were derived by applying a detection filtering (*P*<0.01 compared to the background), expression filter (1.5-fold change compared to the mean of the matched controls) and then a statistical filter (Mann Whitney unpaired Benjamini Hochberg *p*<0.01). The differentially expressed transcripts obtained are shown in brackets next to the disease name at the top of the figure.

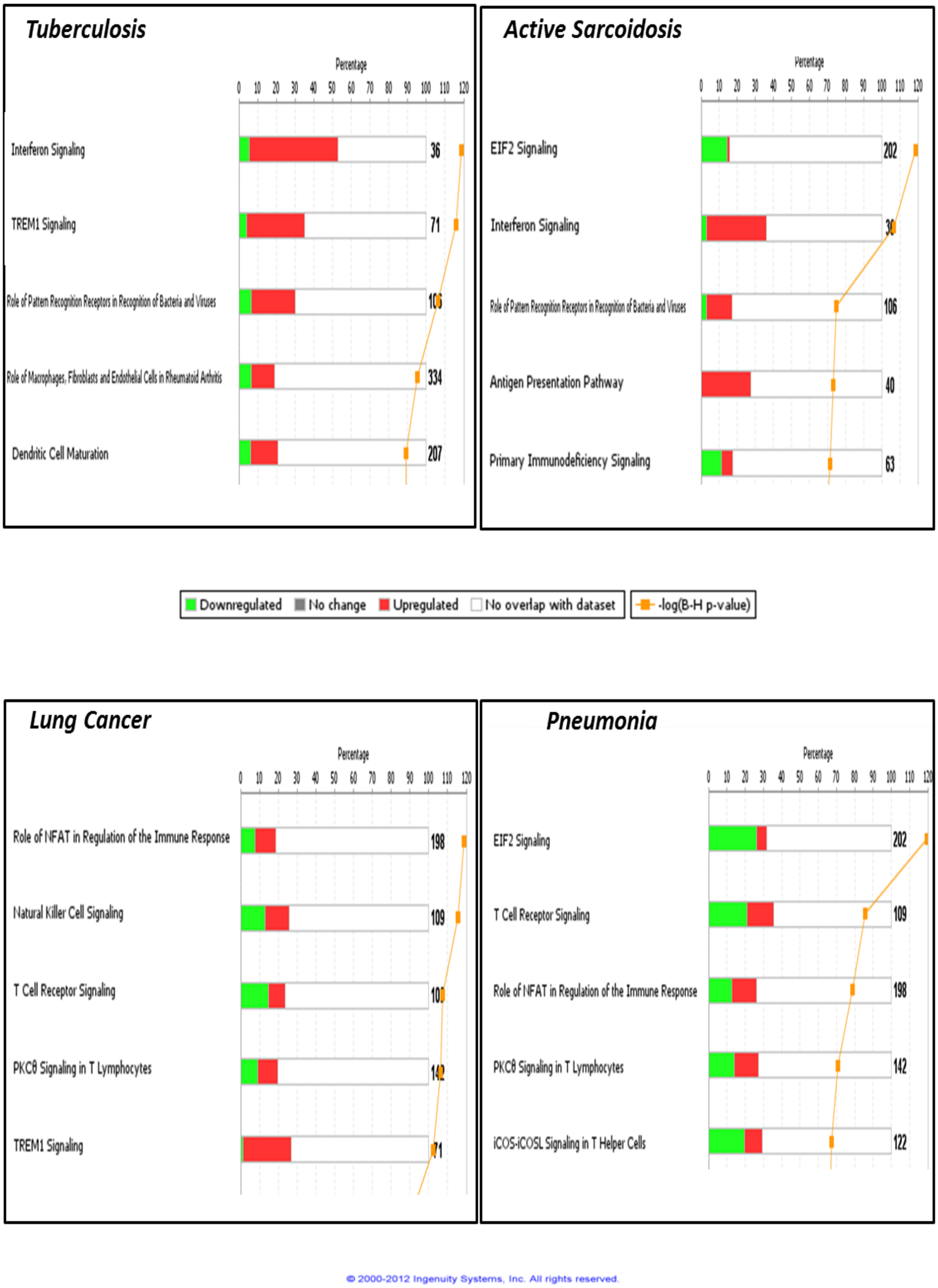


Figure 31. Ingenuity Pathway Analysis for each disease showing the top 5 significant pathways.

IPA analysis was used to determine pathways that were significantly associated with the differentially expressed genes each disease group compared to their matched-controls (Fishers exact Benjamini Hochberg *p*<0.05). Red indicates upregulated and green indicates downregulated. The differentially expressed genes were derived as described in figure 30.

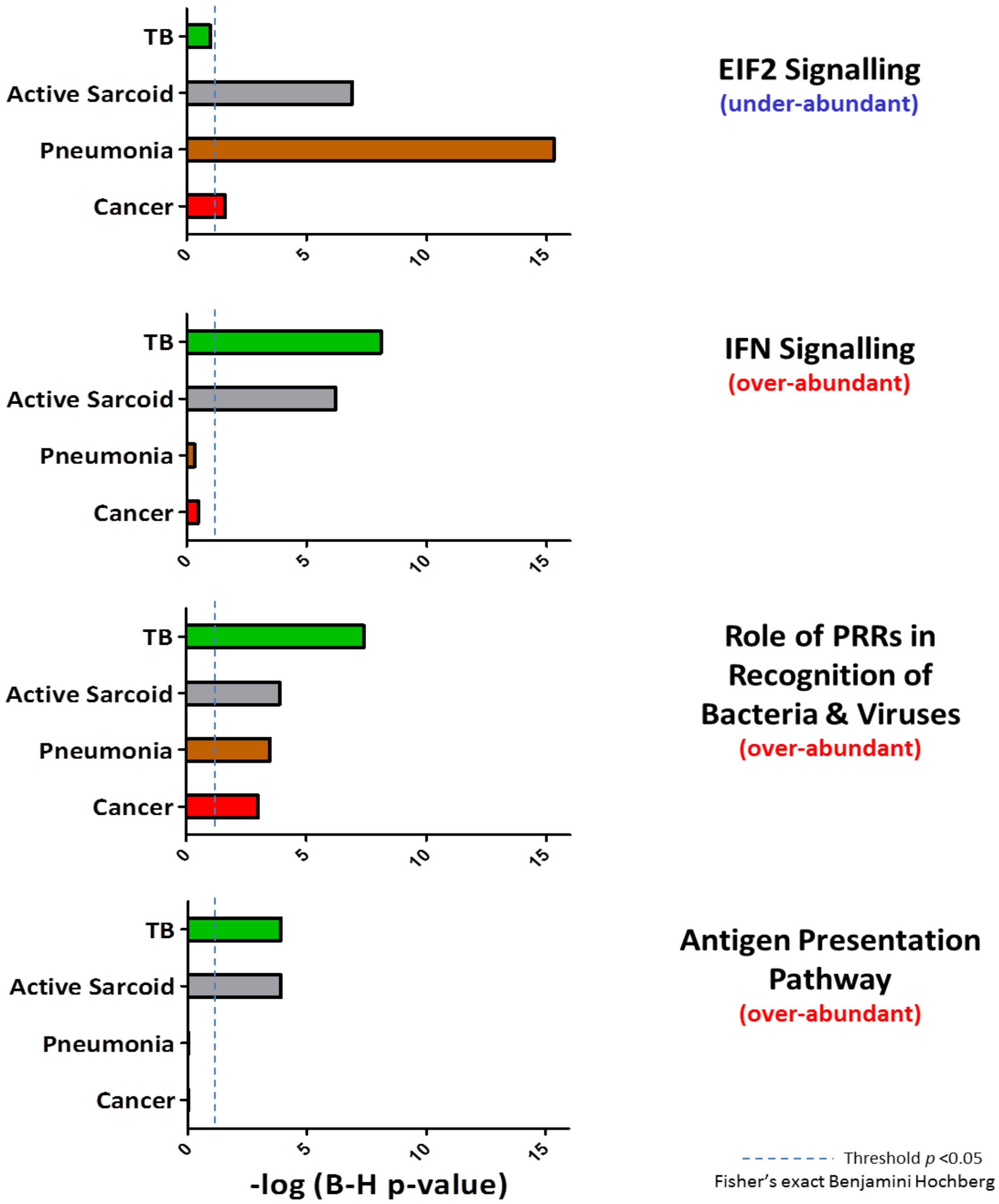


Figure 32. Comparison Ingenuity Pathway Analysis of all the diseases showing the top 4 significant pathways – displaying the *p* value for each pathway.

'Comparative IPA' analysis was used to determine pathways that were significantly associated with the differentially expressed genes for each disease group compared to the other disease groups (Fishers exact Benjamini Hochberg *p*<0.05).The graphs display the log *p* value for each pathway for each disease group. The dotted line is a threshold level of significance set at *p*<0.05. The differentially expressed genes were derived as described in figure 30.

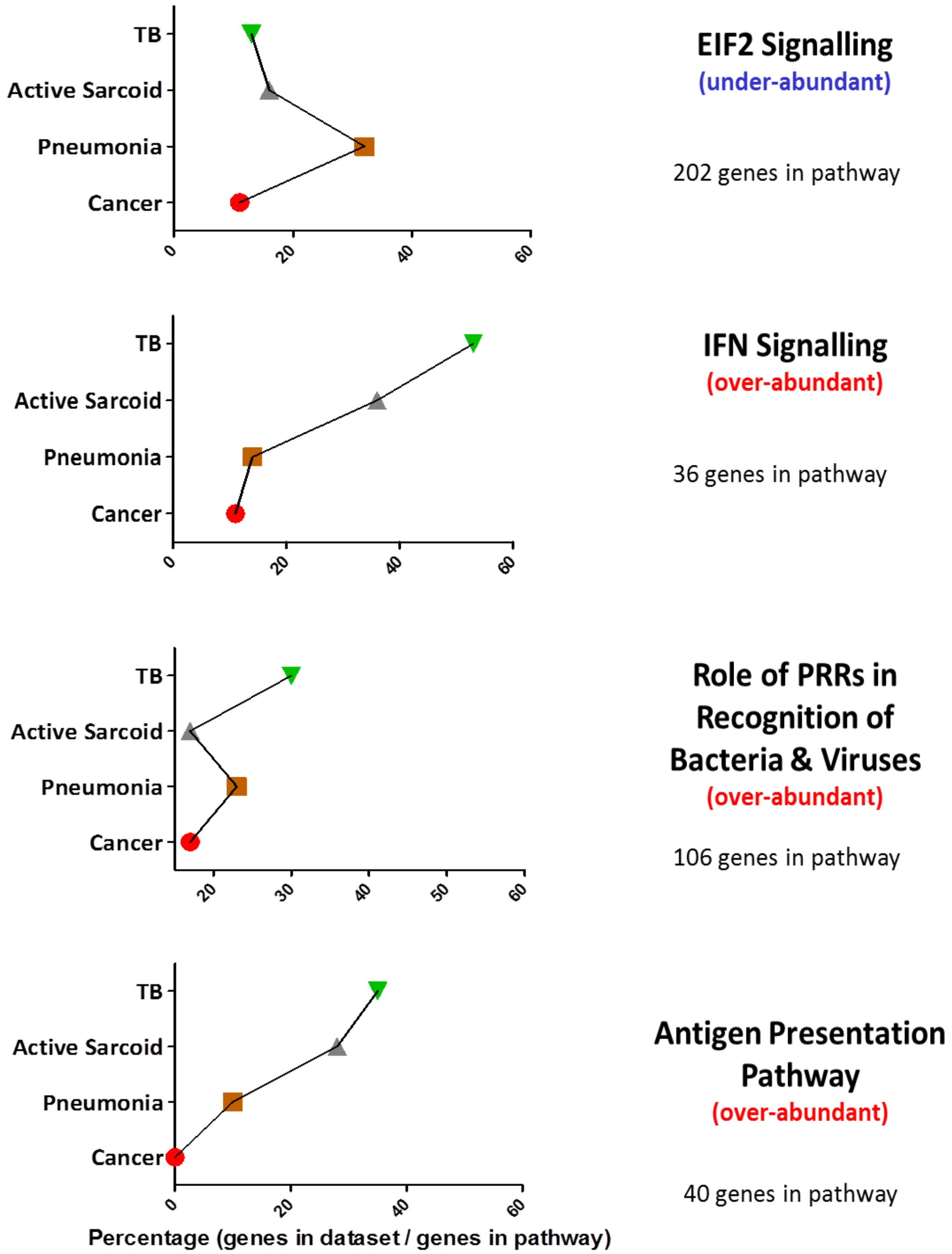
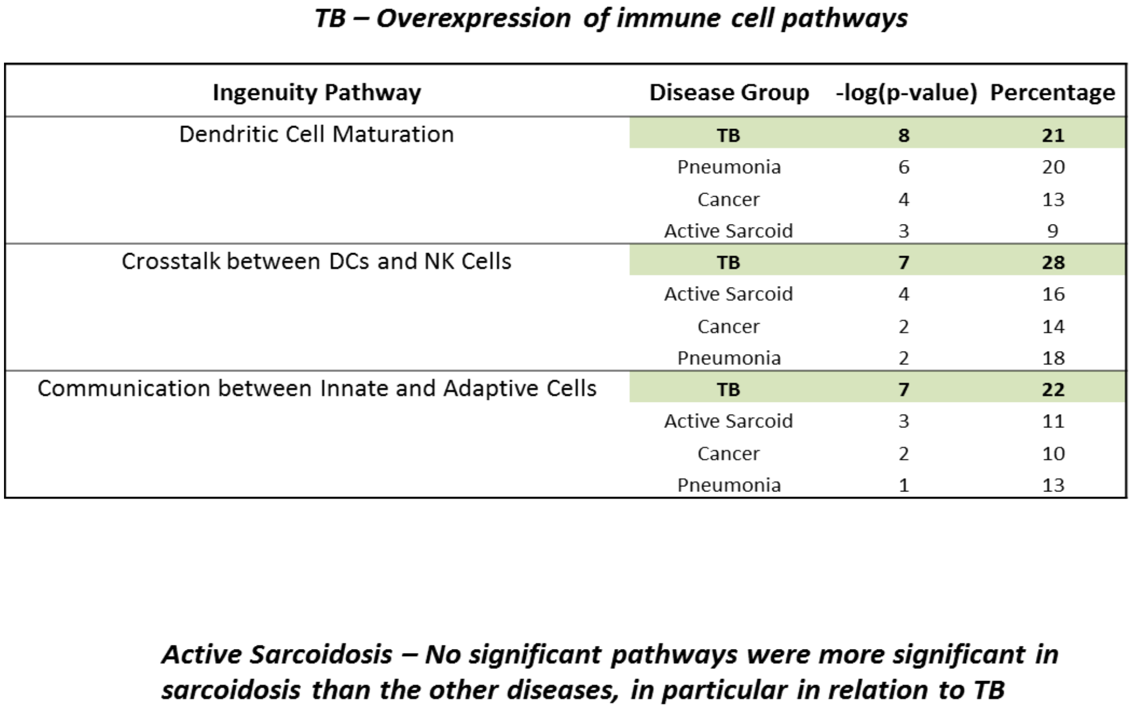


Figure 33. Comparison Ingenuity Pathway Analysis of all the diseases showing the top 4 significant pathways – displaying the percentage of genes for each disease.

‘Comparative IPA’ analysis was used to determine pathways that were significantly associated with the differentially expressed genes for each disease group compared to the other disease groups (Fishers exact Benjamini Hochberg *p*<0.05).The graphs display the percentage of genes present in the pathway for each disease group. The differentially expressed genes were derived as described in figure 30.



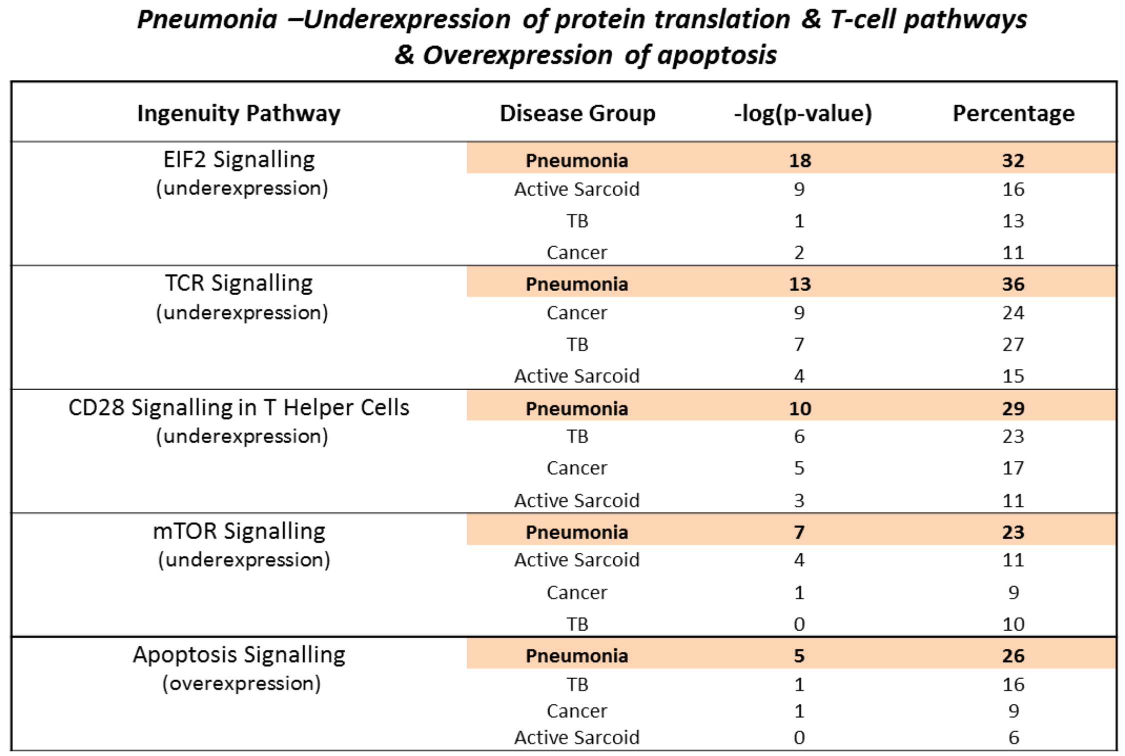
TB – Overexpression of immune cell pathways

Ingenuity Pathway	Disease Group	-log(p-value)	Percentage
Dendritic Cell Maturation	**TB**	**8**	**21**
	Pneumonia	6	20
	Cancer	4	13
	Active Sarcoid	3	9
Crosstalk between DCs and NK Cells	**TB**	**7**	**28**
	Active Sarcoid	4	16
	Cancer	2	14
	Pneumonia	2	18
Communication between Innate and Adaptive Cells	**TB**	**7**	**22**
	Active Sarcoid	3	11
	Cancer	2	10
	Pneumonia	1	13

Active Sarcoidosis – No significant pathways were more significant in sarcoidosis than the other diseases, in particular in relation to TB

Table 24. Comparison Ingenuity Pathway Analysis can identify dominant pathways for each disease relative to the other three diseases (1) TB

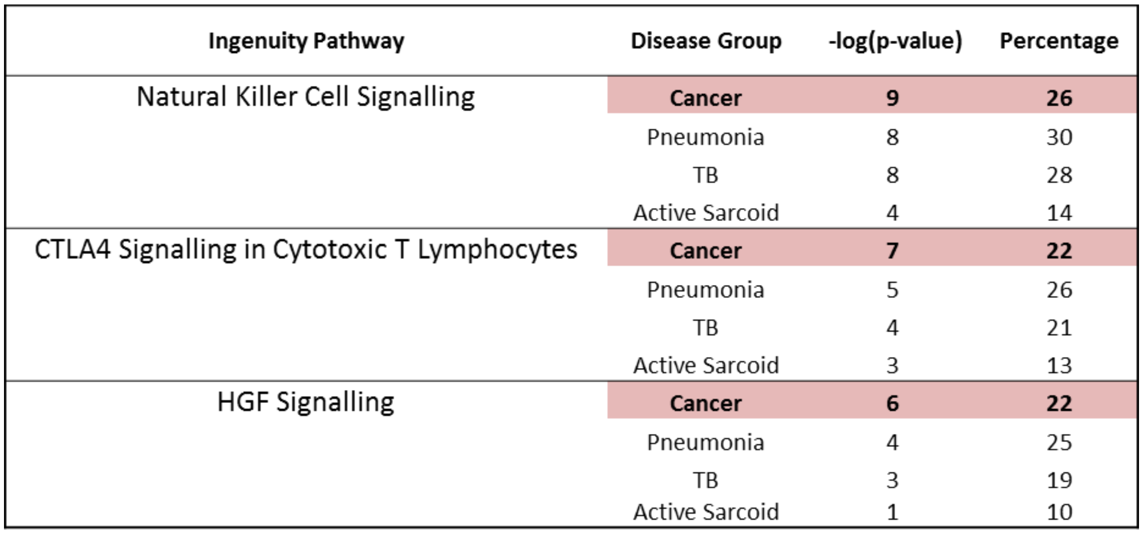
‘Comparative IPA’ was applied to determine which pathways were more significantly associated with TB than the other diseases, relative to the controls. The log(p-value) and percentage of genes present in that pathway are shown. The differentially expressed genes were derived as described in figure 30.



Pneumonia –Underexpression of protein translation & T-cell pathways & Overexpression of apoptosis

Ingenuity Pathway	Disease Group	-log(p-value)	Percentage
EIF2 Signalling (underexpression)	**Pneumonia**	**18**	**32**
	Active Sarcoid	9	16
	TB	1	13
	Cancer	2	11
TCR Signalling (underexpression)	**Pneumonia**	**13**	**36**
	Cancer	9	24
	TB	7	27
	Active Sarcoid	4	15
CD28 Signalling in T Helper Cells (underexpression)	**Pneumonia**	**10**	**29**
	TB	6	23
	Cancer	5	17
	Active Sarcoid	3	11
mTOR Signalling (underexpression)	**Pneumonia**	**7**	**23**
	Active Sarcoid	4	11
	Cancer	1	9
	TB	0	10
Apoptosis Signalling (overexpression)	**Pneumonia**	**5**	**26**
	TB	1	16
	Cancer	1	9
	Active Sarcoid	0	6

Table 25. Comparison Ingenuity Pathway Analysis can identify dominant pathways for each disease relative to the other three diseases (2) Pneumonia



Ingenuity Pathway	Disease Group	-log(p-value)	Percentage
Natural Killer Cell Signalling	**Cancer**	**9**	**26**
	Pneumonia	8	30
	TB	8	28
	Active Sarcoid	4	14
CTLA4 Signalling in Cytotoxic T Lymphocytes	**Cancer**	**7**	**22**
	Pneumonia	5	26
	TB	4	21
	Active Sarcoid	3	13
HGF Signalling	**Cancer**	**6**	**22**
	Pneumonia	4	25
	TB	3	19
	Active Sarcoid	1	10

Table 26. Comparison Ingenuity Pathway Analysis can identify dominant pathways for each disease relative to the other three diseases (3) Cancer

'Comparative IPA' was applied to determine which pathways were more significantly associated with pneumonia (table 25) or cancer (table 26) than the other diseases, relative to the controls. The log(p-value) and percentage of genes present in that pathway are shown. The differentially expressed genes were derived as described in figure 30.

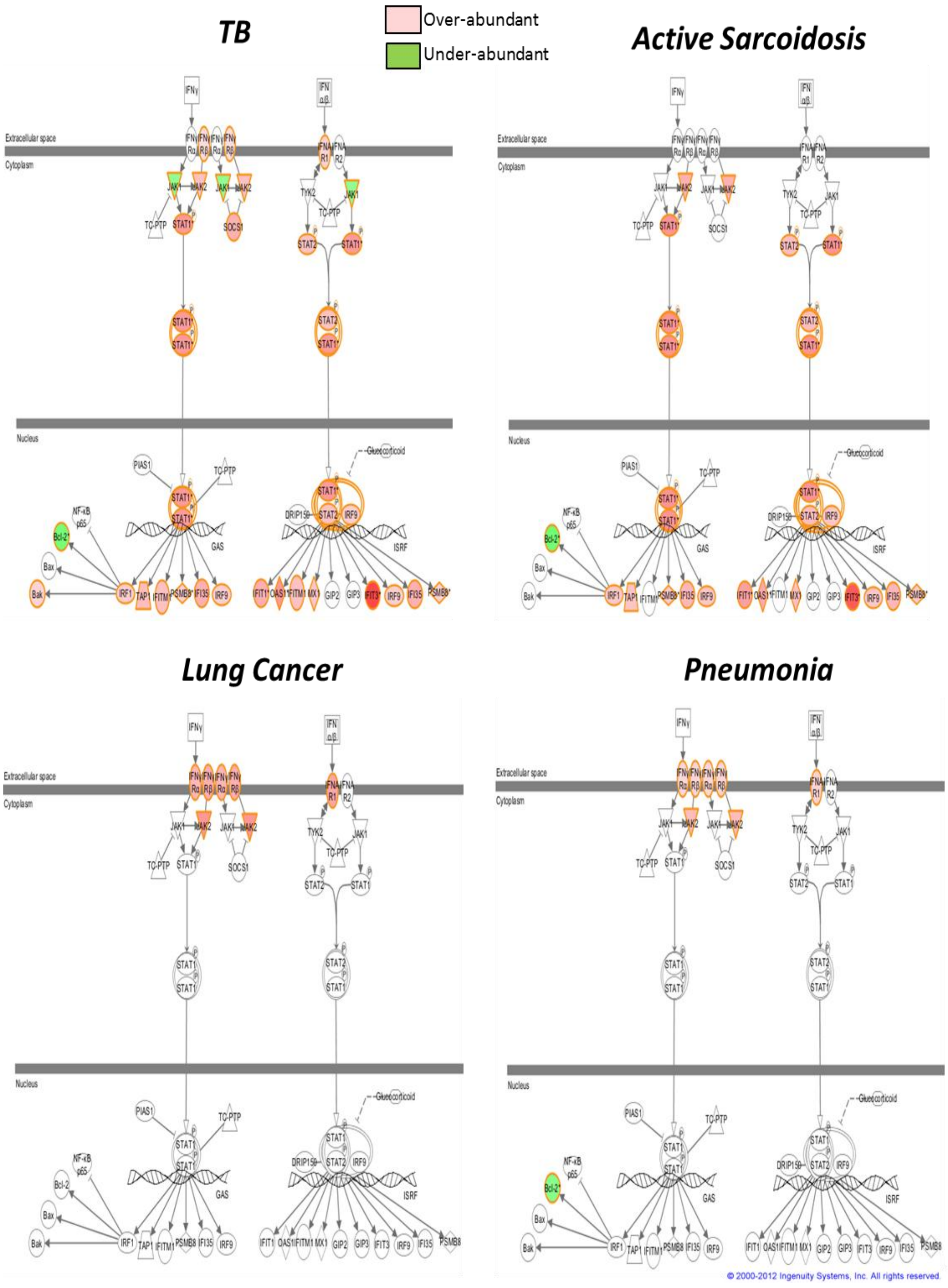


Figure 34. Interferon signalling IPA pathway for each disease compared to their controls.

The differentially expressed genes from each disease, derived as described in figure 30, were overlaid onto the IPA IFN-signalling pathway. The pink genes are over-abundant and blue genes are under-abundant relative to the controls.

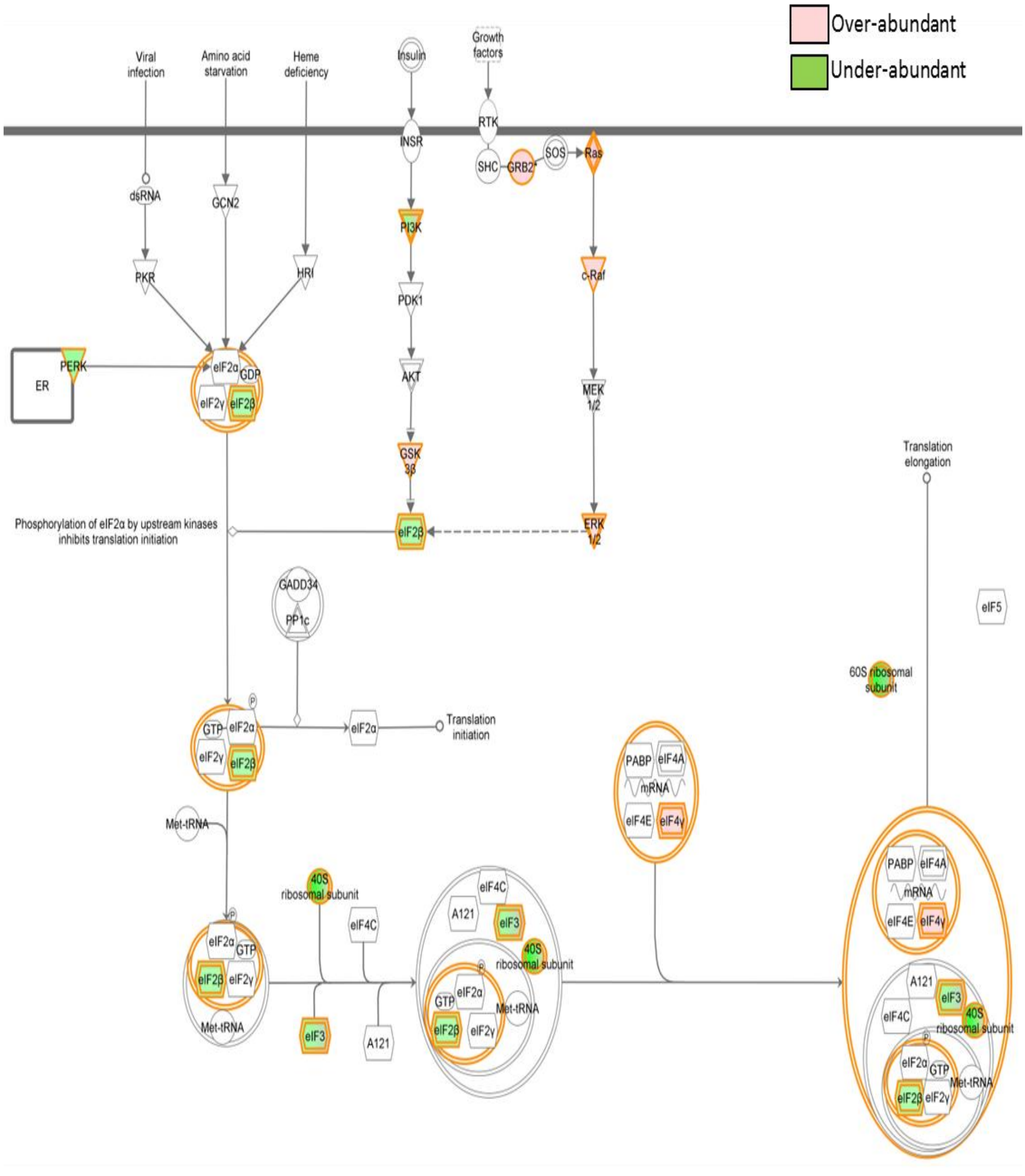


Figure 35. EIF2 signalling pathway in pneumonia demonstrating that the genes involved with protein translation are mostly under-abundant

The differentially expressed genes from pneumonia, derived as described in figure 30, were overlaid onto the IPA IFN-signalling pathway. The pink genes are over-abundant and blue genes are under-abundant relative to the controls.

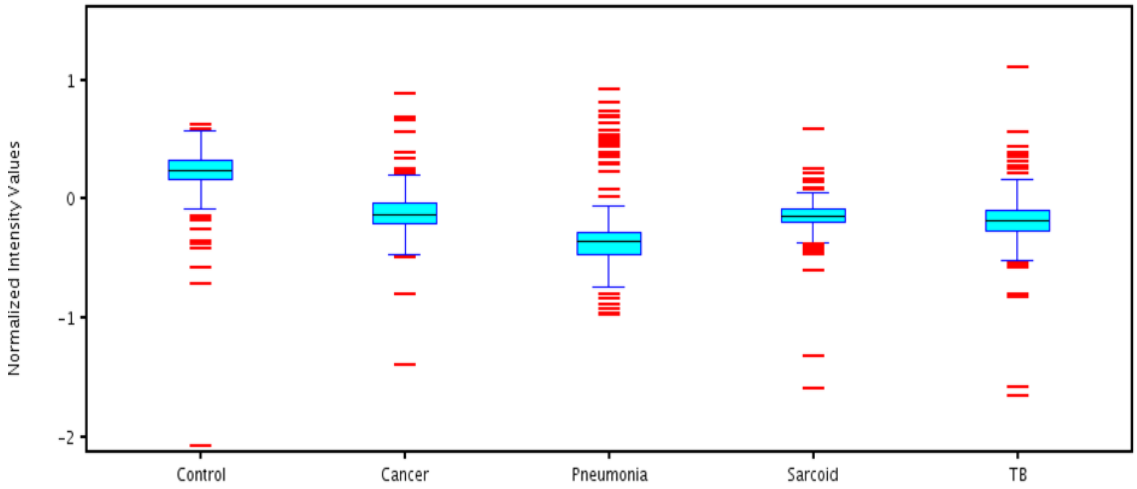


Figure 36. Under-abundance of protein translation genes are found in all diseases but the largest number occurs in the pneumonia patients

Genes were selected as related to protein translation, either eukaryotic translation initiation factors or ribosomal proteins (582 transcripts), and >1.3 fold change from the controls (251 transcripts). The graph displays box plots of the normalised intensity values of each disease group. Box plots show the median, 25th and 75th interquartile and outliers in red.

Symbol	Fold change	**Fold change**	Fold change	Fold change
	Cancer vs Control	**Pneumonia vs Control**	Sarcoid vs Control	TB vs Control
ATF6B	LFC	**1.4**	LFC	LFC
ATF6A	1.5	**1.8**	LFC	1.6
IRE1	LFC	**LFC**	LFC	-1.5
ATF4	1.4	**LFC**	LFC	LFC
XBP1	LFC	**LFC**	LFC	-1.3
GADD34	LFC	**LFC**	LFC	LFC
BiP	LFC	**LFC**	LFC	LFC
GRP94	LFC	**LFC**	LFC	LFC
PERK (eIF2α)	LFC	**-1.8**	LFC	LFC
CHOP	-1.2	**-1.3**	-1.1	-1.3
ABCE1	-1.5	**-2.0**	-1.4	-1.8

Table 27. Predominant under-abundance, or no change, in unfolded protein response genes was found by a targeted analysis of related genes in pneumonia and the other diseases.

Genes were selected as related to the unfolded protein response. LFC = fold change less than 1.3.

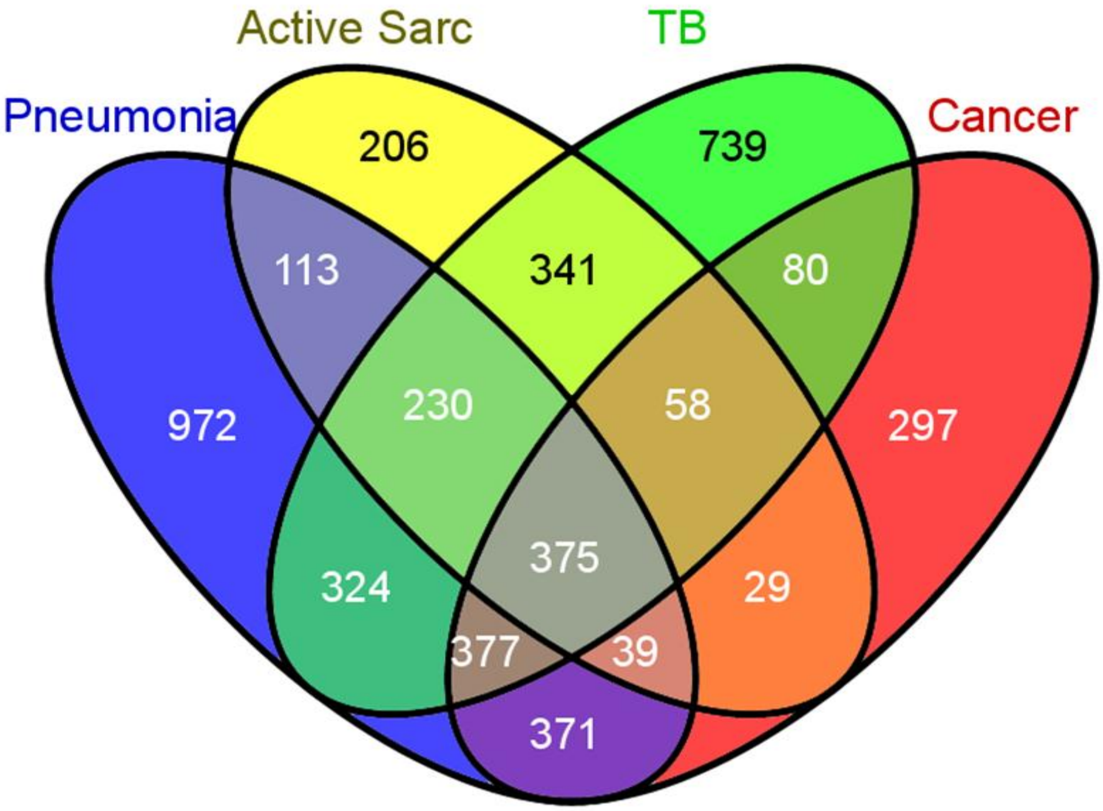


Figure 37. Venn diagram displays the overlapping of the differentially expressed genes and reveals unique-disease related genes.

The differentially expressed genes were derived as described in figure 30.

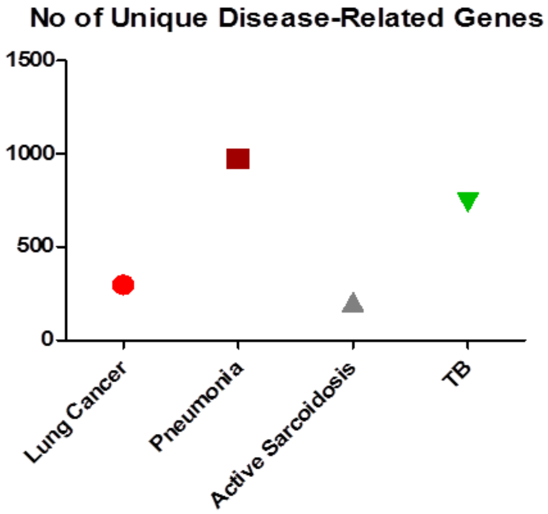
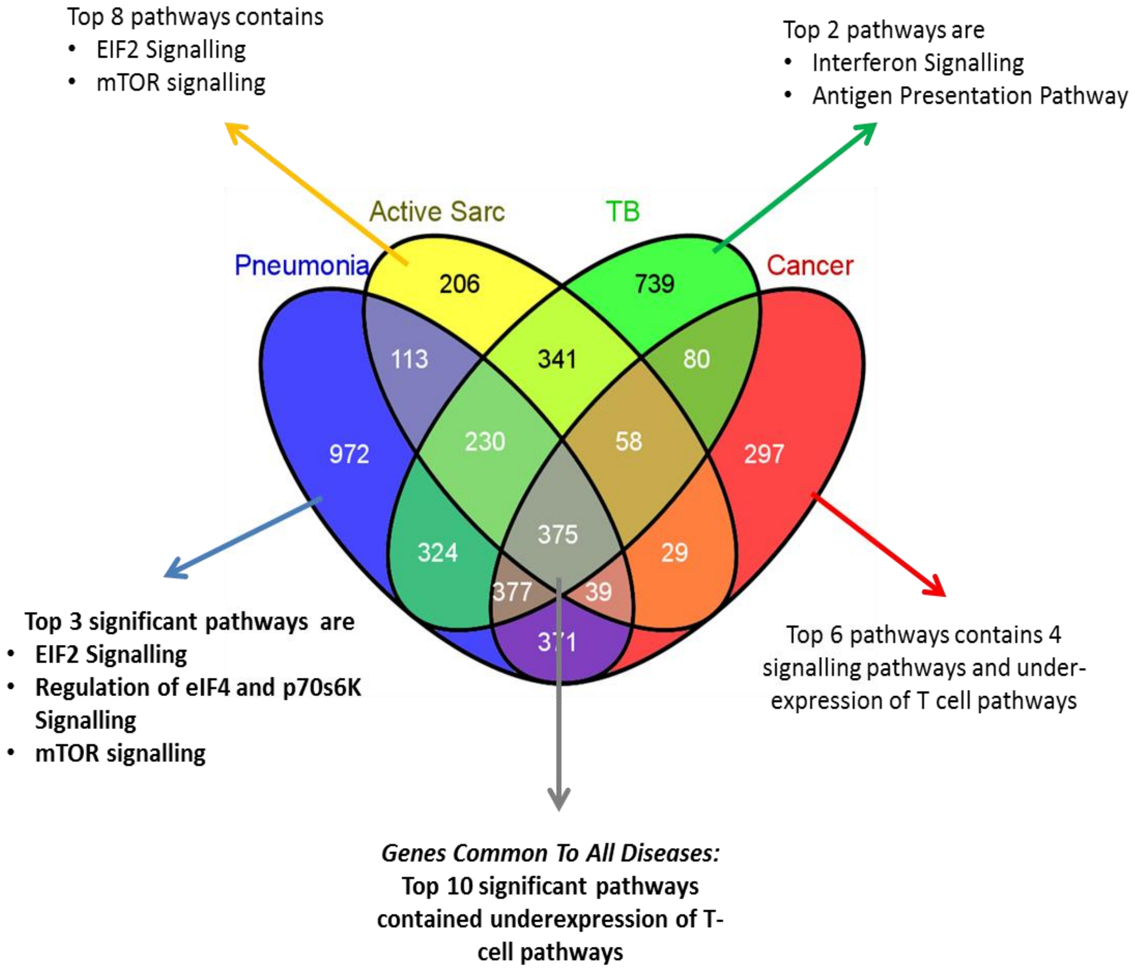


Figure 38. Comparing the number of unique disease-related genes generated by the 4-set Venn diagram

The graph displays the total number of unique disease-related transcripts as derived from the Venn diagram above.



Top 8 pathways contains
- EIF2 Signalling
- mTOR signalling

Top 2 pathways are
- Interferon Signalling
- Antigen Presentation Pathway

Top 3 significant pathways are
- **EIF2 Signalling**
- **Regulation of eIF4 and p70s6K Signalling**
- **mTOR signalling**

Top 6 pathways contains 4 signalling pathways and under-expression of T cell pathways

Genes Common To All Diseases:
Top 10 significant pathways contained underexpression of T-cell pathways

Figure 39. Unique disease-related genes each show unique functional differences by Ingenuity Pathway Analysis.

The unique-disease related transcripts from the venn diagram were applied to IPA analysis. Only the 972 pneumonia and 375 common transcripts were found to have significant pathways (Fishers Exact Benjamini Hochberg p<0.01), these pathways are shown in bold. However without multiple testing correction each disease showed unique functional associations with their unique-disease related genes.

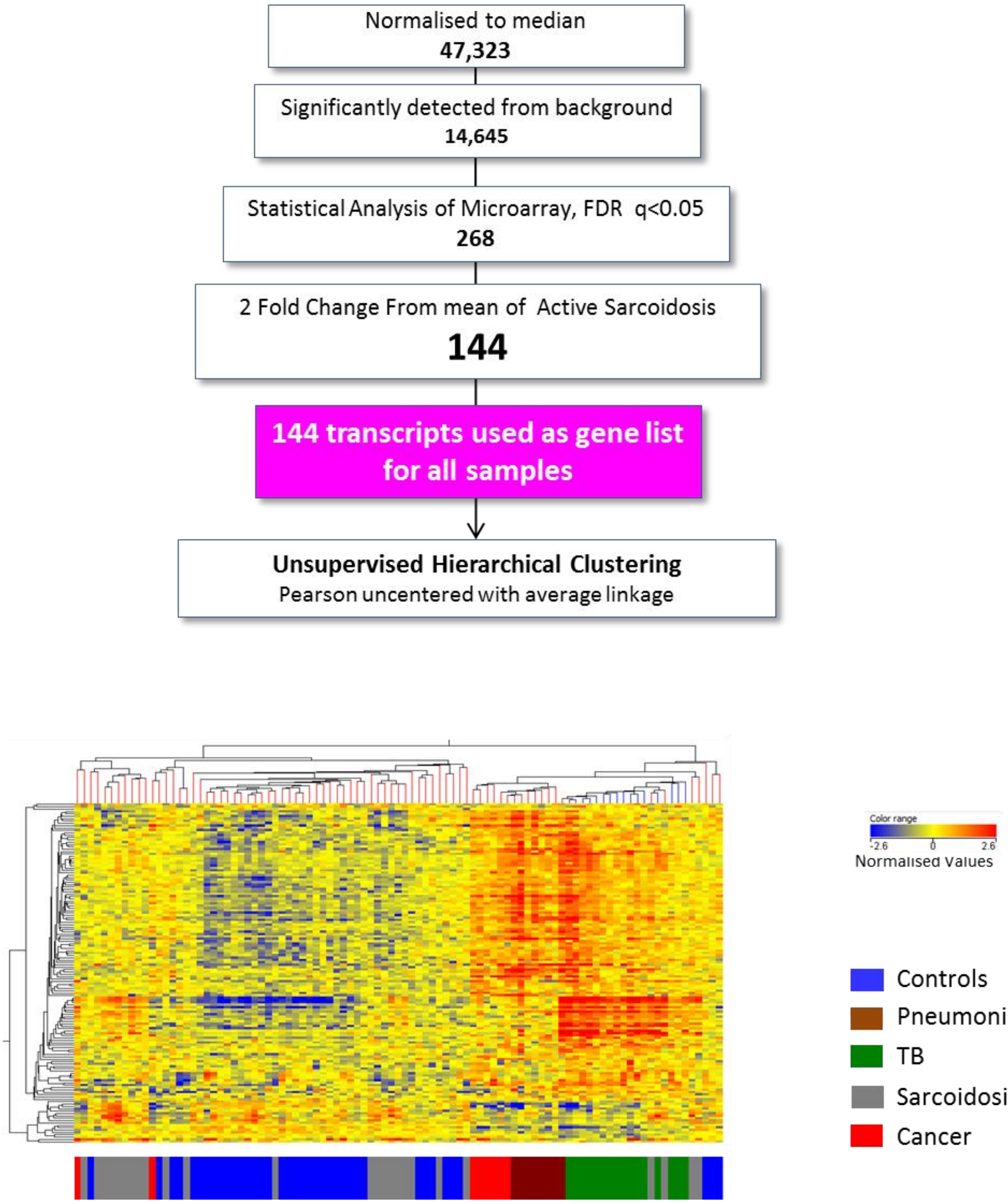
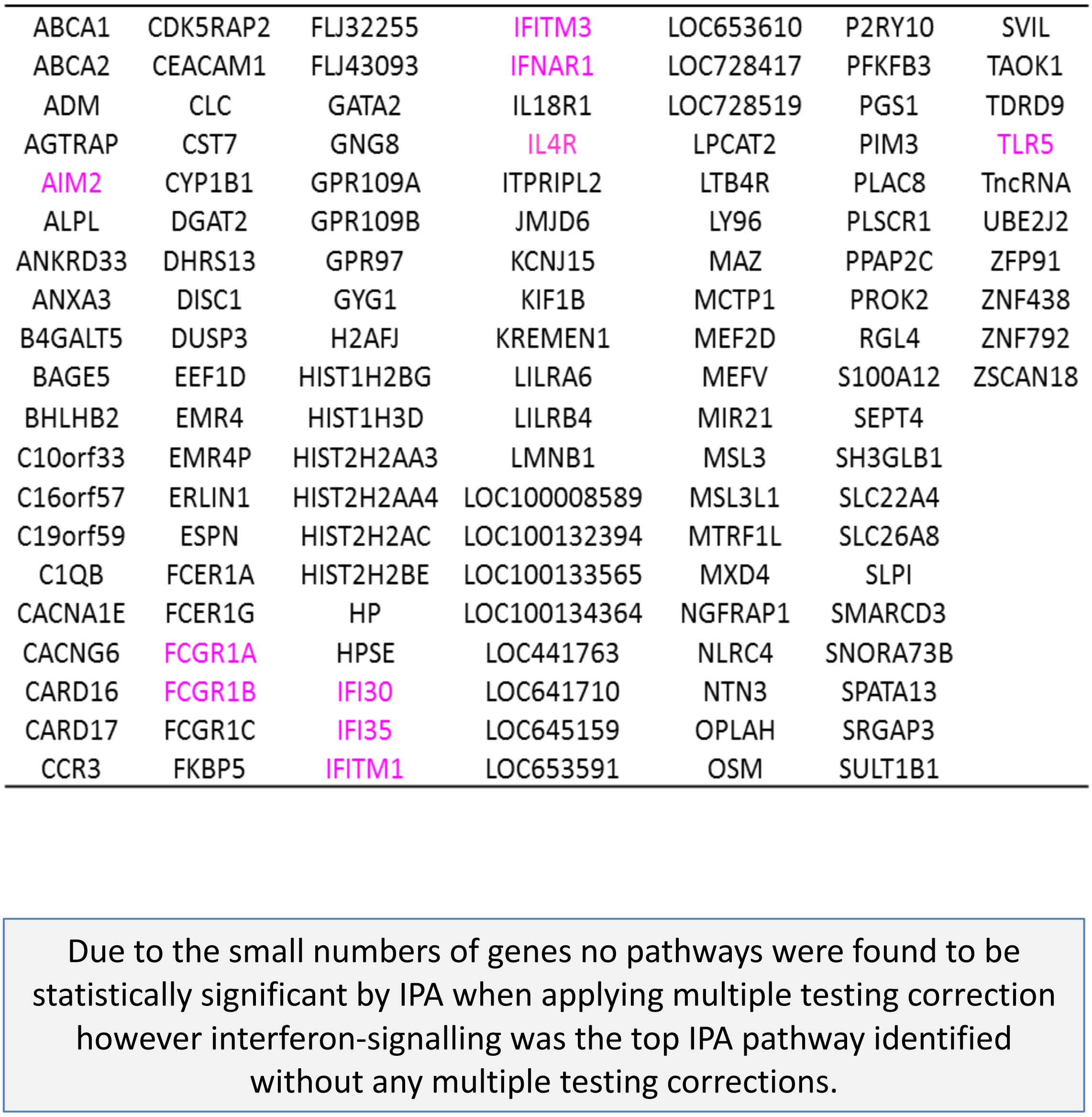


Figure 40. Supervised analysis was able to identify the differentially expressed genes between TB and active sarcoidosis.

As TB and active sarcoidosis are so similar a supervised analysis was carried out to expose those transcripts that are differentially expressed between the two granulomatous diseases. A less stringent and supervised analysis was necessary to derive these genes due to the high similarity between them, as shown in the flow diagram in this figure. The layout of the heatmap is as described in figure 13.



ABCA1
ABCA2
ADM
AGTRAP
AIM2
ALPL
ANKRD33
ANXA3
B4GALT5
BAGE5
BHLHB2
C10orf33
C16orf57
C19orf59
C1QB
CACNA1E
CACNG6
CARD16
CARD17
CCR3
CDK5RAP2
CEACAM1
CLC
CST7
CYP1B1
DGAT2
DHRS13
DISC1
DUSP3
EEF1D
EMR4
EMR4P
ERLIN1
ESPN
FCER1A
FCER1G
FCGR1A
FCGR1B
FCGR1C
FKBP5
FLJ32255
FLJ43093
GATA2
GNG8
GPR109A
GPR109B
GPR97
GYG1
H2AFJ
HIST1H2BG
HIST1H3D
HIST2H2AA3
HIST2H2AA4
HIST2H2AC
HIST2H2BE
HP
HPSE
IFI30
IFI35
IFITM1
IFITM3
IFNAR1
IL18R1
IL4R
ITPRIPL2
JMJD6
KCNJ15
KIF1B
KREMEN1
LILRA6
LILRB4
LMNB1
LOC100008589
LOC100132394
LOC100133565
LOC100134364
LOC441763
LOC641710
LOC645159
LOC653591
LOC653610
LOC728417
LOC728519
LPCAT2
LTB4R
LY96
MAZ
MCTP1
MEF2D
MEFV
MIR21
MSL3
MSL3L1
MTRF1L
MXD4
NGFRAP1
NLRC4
NTN3
OPLAH
OSM
P2RY10
PFKFB3
PGS1
PIM3
PLAC8
PLSCR1
PPAP2C
PROK2
RGL4
S100A12
SEPT4
SH3GLB1
SLC22A4
SLC26A8
SLPI
SMARCD3
SNORA73B
SPATA13
SRGAP3
SULT1B1
SVIL
TAOK1
TDRD9
TLR5
TncRNA
UBE2J2
ZFP91
ZNF438
ZNF792
ZSCAN18

Due to the small numbers of genes no pathways were found to be statistically significant by IPA when applying multiple testing correction however interferon-signalling was the top IPA pathway identified without any multiple testing corrections.

Figure 41. The 144 transcripts were differentially expressed between TB and active sarcoidosis.

The 144 transcripts translated to 132 genes as several genes were represented by more than one probe on the chip. Several of the 132 genes shown are known to be interferon-inducible (pink) and were all over-expressed in TB compared to active sarcoidosis.

Pneumonia vs Cancer
46% genes related to innate cells
(neutrophils, eosinophils, histamine)
10% genes related to apoptosis

DEFA4
OLFM4
CEACAM8
LTF
ELANE
MMP8
DEFA1B
VNN1
DEFA3
IFI27
COL17A1
CEACAM6
TCN1
OLR1
OPLAH
CDK5RAP2
GPR84
BPI
CTSG
DEFA1
TACSTD2
HP
RETN
CD24
AZU1
HIST1H2BG
UPB1
BEX1
TMTC1
CD177
MS4A3
CDC20
RNASE3
ANKRD55
TIMM10
TMEM176A
MPO
EMR1
LCN2
C6orf25
TXNDC5
MAPK14
CACNA1E
ABP1
HBE1
CAMP
TMEM158
RAB20
RAP1GAP

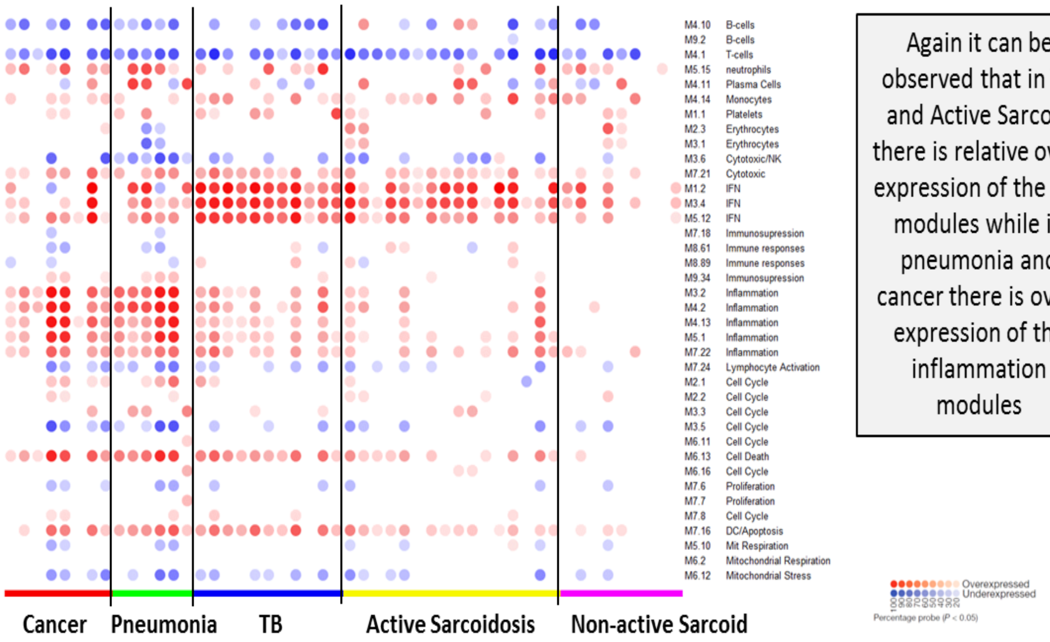
Cancer vs Pneumonia
40% related to cellular processes
(cell differentiation, signalling, protein ubiquitination)
10% genes related to transcriptional regulation

C4BPA
HEMGN
TRIM10
RUNDC3A
SNORD13
PRKAR1A
SF1
ZNF683
TSPAN5
PRR5
RPGRIP1
VWCE
HECW2
PON2
MRVI1
PRR5
KIF27
TMEM107
MYOM2
PDZK1IP1
MBNL3
NOV
BOAT
LOC253039
HLA-DQA1
FBXO9
CBLB
MARCH8
GNLY
EPHA4
LOC100133678
LOC100130905
DISC1
PID1
NHS
VENTX
PTGDR
ZNF23
TESC
COL9A2
LOC642073
LOC196549
BIRC3
LOC643733
IL1RAP
HLA-DRB1
BDKRB2
CREBZF

Figure 42. The top 50 over-abundant differentially expressed genes between pneumonia and cancer.

From analysis of all diseases simultaneously or each disease compared to matched controls, pneumonia and cancer have many similarities. To elucidate the differences the diseases were directly compared to each other. Transcripts were derived by a 1.5 fold change filter between each other and then by a statistical filter using significance analysis of microarray, FDR q<0.05. Total of 1165 transcripts were derived, only the top 50 are shown.

Modular Analysis



Again it can be observed that in TB and Active Sarcoid there is relative over-expression of the IFN modules while in pneumonia and cancer there is over-expression of the inflammation modules

Top Ranking Genes by abundant expression levels relative to the controls

As was observed in the training set in TB and Active Sarcoid the most abundant genes by expression relative to their matched controls contained many IFN-inducible genes while pneumonia contained many neutrophil genes.

Ingenuity Pathway Analysis

As was observed in the training set:

- Relative to the other diseases IFN-signalling pathway was highly significant for TB and Active Sarcoid
- Antigen presentation and Role of PRRs in bacteria/viruses were the most significant in TB then in Active Sarcoid
- Relative to the other diseases EIF2 signalling was highly significant for pneumonia
- In cancer again there was a dominance of significant signalling pathways including NK signalling, CTLA4 signalling in cytotoxic T cells and HGF signalling

Figure 43. Functional gene patterns associated with each disease are validated in the test set.

Chapter 6

144 tuberculosis-specific transcripts can distinguish tuberculosis from all other diseases

Chapter 6: 144 tuberculosis-specific transcripts can distinguish tuberculosis from all other diseases

Introduction

In individuals with classical symptoms of pulmonary TB for whom the diagnosis is suspected, confirmation can only occur by culture of *M. tuberculosis* from sputum or bronchoalveolar washings. There are many challenges attached to making this diagnosis. Firstly culture confirmation can take up to 6 weeks (Pfyffer, Cieslak et al. 1997); although sputum microscopy smear is more commonly and easily attained it only detects 60% of culture positive pulmonary *M. tuberculosis* (Young, Perkins et al. 2008). Secondly it is often not possible for the patient to expectorate therefore an invasive procedure is required to obtain bronchoalveolar washings (Tamura, Shimada et al.). For example in the USA only 70% of pulmonary TB is diagnosed by bacterial culture, the rest by an observed response to antituberculous treatment (CCDC 2007). While in South Africa only 50% of pulmonary TB is culture-confirmed (WHO 2010), presumably due to the reduced impetus to obtain samples, the lack of microbiology facilities and the lack of facilities to perform invasive procedures. Promisingly a relatively new development in TB diagnosis is the use of the WHO endorsed Xpert MTB/RIF automated PCR test to detect *M. tuberculosis* and common drug resistance strains (Taegtmeyer, Beeching et al. 2008; Boehme, Nabeta et al. 2010). Unfortunately Xpert MTB/RIF also requires sputum or bronchoalveolar washing samples. The difficulty in diagnosing pulmonary TB is the ability to distinguish it from the other common differential diagnoses including sarcoidosis, pneumonia and lung cancer. This difficulty may lead to the use of invasive procedure and delays in treatment (Storla, Yimer et al. 2008). Besides if there is a misdiagnosis, for example sarcoidosis instead of pulmonary TB, this could result in worsening pathology due to the commencement of immunosuppressive therapy, and vice versa could result in unwarranted side effects from six months of the antibiotics. Therefore a biomarker capable of distinguishing active TB from other similar diseases that is achievable from a simple blood test could offer substantial clinical value. In the previous chapter 144 transcripts were demonstrated to be differentially expressed between TB and active sarcoidosis (Figure 40). Although this study did not set out to define a biomarker for TB diagnosis because these two diseases are the most comparable by clinical and molecular profiles it seemed a reasonable hypothesis that this 144-transcript list could be applied to distinguish active TB profiles from other similar diseases.

Results

A set of 144 TB-specific transcripts had good sensitivity and specificity in independent cohorts

Class prediction is a supervised learning method where the algorithm learns from samples with a known phenotype (e.g. the training set) to establish a prediction rule to classify new samples (e.g. the test and validation sets). The machine learning algorithm support vector machines (SVM) is a frequently applied technique in class prediction. The prediction model was built using the training set and then run in the test and validation sets. The classification of each sample can be used to calculate the sensitivity and specificity for each dataset of samples. Samples were classifies as either a TB sample or a 'non-TB' sample (i.e. controls, sarcoidosis, pneumonia or lung cancer).

The model was also built again, using identical settings, in the Maertzdorf *et al.* 2012 dataset and the Berry *et al.* 2010 dataset as it was not possible in GeneSpring 11.5 to run the model in these datasets due to the conversion of 144 Illumina HT12 V4 transcripts to Agilent probes for the Maertzdorf *et al.* 2012 dataset, and conversion to

Illumina HT Version 3 for the Berry *et al.* 2010 dataset. Only 131 transcripts out of the 144 were present in the older Illumina V3 chip therefore the model was built with a reduced number of transcripts from the Berry *et al.* 2010 dataset.

The 144 transcripts showed a high sensitivity (≥ 82%) in the training set, test set, validation set and the Maertzdorf dataset (Figure 44). The sensitivity was slightly lower (74%) when built using the Berry dataset. This could be related to the loss of 13 transcripts when converting between the different Illumina chips. The specificity was above 90% for all datasets (Figure 44).

The 144 transcripts had superior accuracy to the two published transcript lists

Two publications have previously reported genes lists able to distinguish active TB patients from sarcoidosis patients. Koth *et al* found 50 discriminating genes and Maertzdorf *et al* found 100 discriminating transcripts (Koth, Solberg et al. 2011; Maertzdorf, Weiner et al. 2012). Both these transcript lists were applied to the three datasets (training, test and validation). The Maertzdorf 100 Agilent probes only translated to 76 genes after conversion to genes recognised by the DAVID converter and manual annotation. The Maertzdorf 100 transcript (76 gene list) had a poor sensitivity in the training (56%) and test set (45%) and but was reasonable in the validation set (75%), all were below their reported sensitivity and specificity in their own dataset (88%, 97% respectively, (Maertzdorf, Weiner et al. 2012)) (Figure 44). The Koth 50 genes, developed using an Affymetrix platform, had reasonable sensitivity in the training set (75%) but poor in both the test set (45%) and the validation set (50%). Koth *et al* did not test their own transcript list for its ability to discriminate between their sarcoidosis and TB patients nor to distinguish between the sarcoidosis and TB samples accumulated from the Berry *et al.* 2010 study (Koth, Solberg et al. 2011).

Therefore these results show that the 144 transcript list derived from our study, compared to the two recently published discriminating transcript lists, had the highest and most consistent sensitivity and specificity for discriminating between the TB samples and all the other samples (including samples from different diseases and healthy controls), from different cohorts, different research studies, and across different microarray platforms.

Few transcripts overlapped between the discriminating transcripts lists

Only one gene was overlapping between the Maertzdorf *et al* transcript list and the 144 list (Figure 45). None were overlapping between the Maertzdorf *et al* and Koth *et al* transcript list. Six genes were overlapping between the smaller Koth *et al* transcript list and the 144 transcript list.

Discussion

The 144 TB-specific transcript list had good sensitivity and specificity and had superiority over previously published discriminating transcript lists

The 144 list had good sensitivity and specificity in the training set. However the 144 list was built using this dataset therefore to ensure there was no overfitting of the prediction model it was then run on two independent cohorts, the test set and the validation set. The model continued to show good sensitivity (82% test, 88% validation) and specificity (91% test, 92% validation) for both the test and validation sets (Figure 44). For a more rigorous assessment the 144 list was also tested by external validation using a dataset (Maertzdorf *et al.* 2012) collected by a different research team (Stefan Kaufmann *et al*), from a different research institution (Max Planck), recruited in a different country (Germany) and run on a different platform (Agilent) (Maertzdorf, Weiner et al. 2012). External validation by means of a dataset from an entirely different study is said to afford the ultimate conclusive evidence that a model is valid (Taylor, Ankerst et al. 2008). The 144 transcript list was originally derived from comparing only pulmonary TB and active pulmonary sarcoidosis. It is not surprising that it has such good accuracy for distinguishing TB also from the pneumonia, lung cancer and control samples due to the divergence of their profiles as visualised by unsupervised hierarchical clustering from both the granulomatous diseases. It is pertinent to bear this in mind when comparing the 144 transcript list with the two other transcript lists because the authors only claimed their lists to be able to distinguish TB from sarcoidosis (Koth, Solberg et al. 2011; Maertzdorf, Weiner et al. 2012). Neither of the published transcript lists offered such good consistent sensitivity and specificity across all tested datasets as the 144 list (Figure 44). The inability to use all 100 probes from the Maertzdorf *et al.* 2012 study may have affected the accuracy of the transcript list, however, because many of the probes in the Maertzdorf *et al.* 2012 study are not recognised genes their value as a possible surrogate marker is unclear. The Maertzdorf *et al.* 2012 publication did not include validation of their transcript list in any independent dataset. Furthermore the Koth *et al.* 2011 study did not attempt to test their transcript list in either their own dataset or in an independent dataset. This lack of validation may explain why so few genes overlapped between the three transcript lists, although there were six overlapping genes with the Koth *et al.* 2011 transcript list and the 144 transcript list – which may be due to the much larger sample size and power of the Koth *et al.* 2011 study compared to the Maertzdorf *et al.* 2012 study. Interestingly six of the over-expressed genes in the Maertzdorf *et al.* transcript list were neutrophil genes, which could suggest that the TB patients had more severe TB in their study.

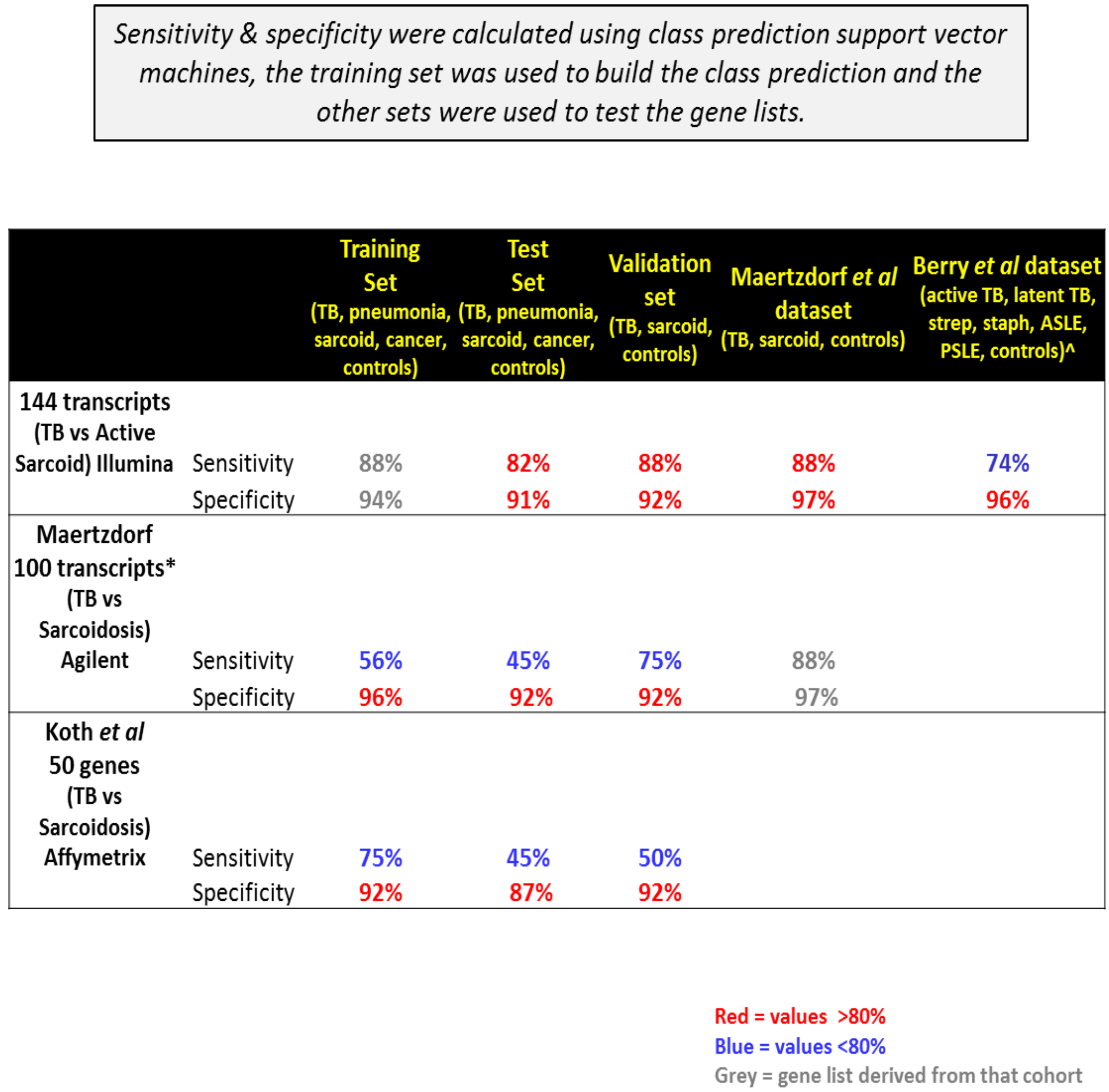
Validation was also attempted in the very large dataset from the Berry *et al.* 2010 paper, which included expression profiles from patients with active TB, latent TB, adult SLE, paediatric SLE, staphylococcal infections, streptococcal infections and healthy controls (Berry, Graham et al. 2010). The model showed reasonable sensitivity of 74% considering it was not designed to discriminate between all these diseases and 13 of the transcripts were missing from the 144 list due to the lack of annotation in the older Illumina HT V3 chip.

The high specificity achieved from all datasets and transcript lists was undoubtedly due to the very low Type I error rate attributable to the small prevalence of true TB samples within each dataset. For this reason the positive predictive values (PPV) for the transcript lists were fairly low, although they were higher in the cohorts with an increased TB prevalence, while the negative predictive values (NPV) were high (data not shown). PPV and NPV, unlike sensitivity and specificity, are influenced by the prevalence of a disease (Lalkhen and McCluskey 2008). In our study there are far fewer true TB (e.g. 16 TB samples in the training set) than true non-TB samples (e.g. 25 sarcoidosis, 8 pneumonia, 8 cancer and 38 control samples in the training set), but in a clinical setting this would not be the case. A potential diagnostic biomarker would not be tested on every respiratory patient, only those with a high suspicion for active TB. However in this case a diagnostic biomarker with a high NPV is worthwhile as the misdiagnosis of active TB and thus subsequent 6 months treatment with antibiotics could have a detrimental impact. Overall these findings suggest transcriptional signatures – similar to the 144 transcript list - have promise as additional diagnostic biomarkers for pulmonary TB if satisfying rigorous testing and validation in large populations.

Chapter Summary

A TB-specific transcript list containing 144 transcripts showed good sensitivity and specificity for distinguishing pulmonary TB patients from patients with pulmonary sarcoidosis, pneumonia, primary lung cancer, latent TB, SLE, Stills, streptococcal infections, staphylococcal infections and healthy controls. This 144 transcript list (132 genes) also appeared to demonstrate superior predictive sensitivity over two previously published transcript lists. These findings potentially implicate the application of blood transcriptional signatures as biomarkers to aid in the diagnosis of pulmonary TB.

Figures for chapter 6



Sensitivity & specificity were calculated using class prediction support vector machines, the training set was used to build the class prediction and the other sets were used to test the gene lists.

		Training Set (TB, pneumonia, sarcoid, cancer, controls)	Test Set (TB, pneumonia, sarcoid, cancer, controls)	Validation set (TB, sarcoid, controls)	Maertzdorf *et al* dataset (TB, sarcoid, controls)	Berry *et al* dataset (active TB, latent TB, strep, staph, ASLE, PSLE, controls)^
144 transcripts (TB vs Active Sarcoid) Illumina	Sensitivity	88%	82%	88%	88%	74%
	Specificity	94%	91%	92%	97%	96%
Maertzdorf 100 transcripts* (TB vs Sarcoidosis) Agilent	Sensitivity	56%	45%	75%	88%	
	Specificity	96%	92%	92%	97%	
Koth *et al* 50 genes (TB vs Sarcoidosis) Affymetrix	Sensitivity	75%	45%	50%		
	Specificity	92%	87%	92%		

Red = values >80%
Blue = values <80%
Grey = gene list derived from that cohort

Figure 44. Testing the 144 TB-specific transcript list in independent cohorts.

The 144 transcripts that were derived from the training set were verified in the test and validation sets as well as the Maertzdorf *et al* dataset and Berry *et al* dataset. The 100 transcripts from the Maertzdorf *et al* dataset and 50 genes from the Koth *et al* dataset were tested in all our cohorts (training, test and validation).The sensitivity (number of true positives/ true positives + false negatives) and specificity (number of true negatives/ true negatives + false positives) was calculated for each dataset as shown in the table. 13 of the 144 transcripts were not recognised by the older Illumina HT V3 chip used for the Berry *et al* datasets in 2009. Only 76 of the Agilent probes used in the Maertzdorf *et al* dataset were recognised as genes in the NIH Database for Annotation, Visualization and Integrated Discovery (DAVID). ASLE = adult SLE, PSLE = paediatric SLE. Values above 80% are in red and below 80% in blue. Grey values represent datasets where the transcript list tested was also derived from that dataset.

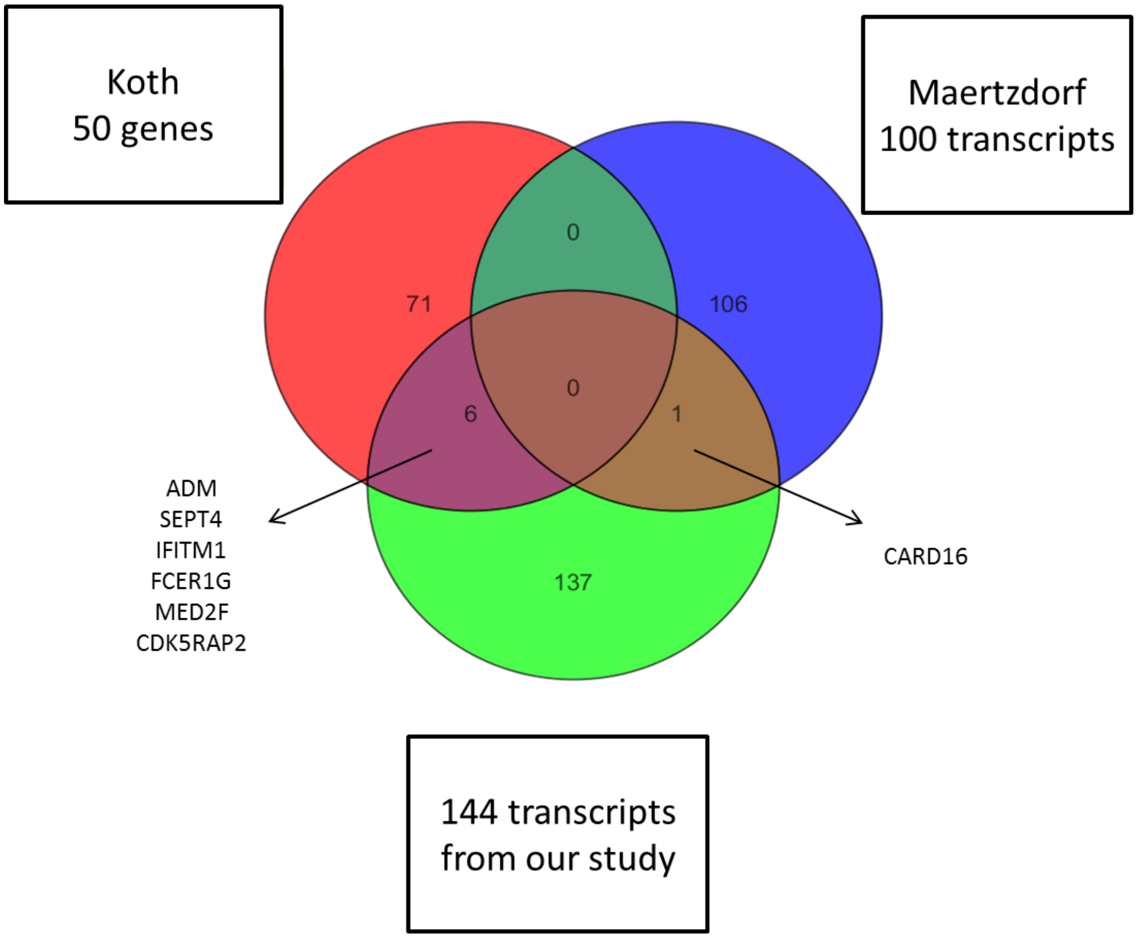


Figure 45. Comparing transcript lists derived to distinguish TB from sarcoidosis profiles.

The venn diagram was used to compare the 144 Illumina transcripts with the Koth *et al* 50 genes (translates to 77 Illumina probes) and the Maertzdorf *et al* 100 Agilent probes (translates to 76 known genes and 107 Illumina probes).

Chapter 7

Comparing blood transcriptional responses before, during and after tuberculosis treatment

Chapter 7: Comparing blood transcriptional responses before, during and after tuberculosis treatment

Introduction

After the diagnosis of TB there are no available early biomarkers correlating with treatment success, resulting in significant delay in assessing treatment response. Currently conversion to negative culture after two months of treatment is the only accepted biomarker (Mitchison 1993). However a systematic review and meta-analysis of sputum conversion revealed low sensitivity and modest specificity for the prediction of treatment failure (Horne, Royce et al. 2010). Chest radiographs are commonly used to assess response but are not universally available and assessment is difficult to standardise (Walzl, Ronacher et al. 2011). This lack of effective treatment monitoring can lead to the development and spread of drug-resistant TB, which is mainly, caused by non-adherence or inappropriate drug regimens, with a detrimental impact on global TB control. Earlier blood biomarkers correlating with treatment response would improve monitoring of individual patient treatment responses without the need for sputum production, and may also permit stratification of patients requiring differing treatment regimens. Furthermore early biomarkers may aid in drug development. Berry *et al.* 2010, demonstrated that their UK active TB transcriptional signature diminished after two months of successful treatment in eight patients and reverted to that of healthy individuals after completing treatment (Berry, Graham et al. 2010). Our current study looked at an earlier time point of two weeks as this could offer a clinical advantage if able to act as a surrogate marker of treatment success. In addition our study examined the gene expression data mostly from TB patients from South Africa, one of the top high-burden TB countries.

Results

Participants demographics and characteristics

Twenty-nine active TB patients were included from South Africa and 8 active TB patients from the UK (Figure 46). None of the TB patients relapsed within 1 year and all were discharged from the program as cured. The 29 South African patients were sampled at: pre-treatment (29/29 patients), 2 weeks (25/29 patients), 2 months (24/29 patients), 6 months (25/29 patients) and 12 months (29/29 patients) after initiation of treatment. Only one of the South African TB patients was smear negative but 50% of the UK TB patients were smear negative (Tables 28 & 29). All South African patients were of the same ethnicity and race, but the UK patients were a more diverse ethnicity mix (Tables 28 & 29). In South Africa it is not routine to perform chest radiographs due to the lack of facilities therefore only one had a chest radiograph, but all UK patients had a chest radiograph of which seven were abnormal, the eighth patient had an abnormal HRCT scan. Thirty-eight South African latent individuals were also sampled as asymptomatic controls as in the area of South Africa the patients were recruited from there is such a high exposure to *M. tuberculosis* that a control is someone with evidence of exposure to TB. All latent TB patients were IGRA positive.

A change in transcriptional response is detectable after 2 weeks of antituberculous treatment

To determine whether an active TB blood transcriptional signature was perturbed upon treatment, gene expression profiles of significantly detectable genes without further filtering, were examined in the 29 active TB patients before, during (2 weeks and 2 months), at the end of (6 months), and after treatment (12 months) (Figure 47). By plotting the expression profiles of the 15,837 detectable transcripts along a time scaled x-axis, a marked change was readily observed after 2 weeks of treatment (Figure 47).

This suggested a change in transcriptional response as early as two weeks after treatment initiation.

To refine the transcript list capable of demonstrating a transcriptional change in response to treatment we initially derived a South African active TB transcriptional signature. The South Africa active TB 664-transcript signature was derived from applying a two-fold change filter to the mean of the transcripts in the latent TB profiles and a further statistical filter (Mann Whitney unpaired, Bonferroni $p<0.01$) (Figure 48). When this South Africa active TB 664-transcript signature was applied to the treated South Africa cohort, a marked and rapid change in the transcriptional response was again observed as early as two weeks, which then continued through two and six months, after treatment initiation (Figure 49).

The Transcriptional Response Changes Significantly at 2 Weeks

To appreciate if the changes visualised on the heatmap were significant changes the MDTH algorithm was applied to the 664-transcripts as this generates a quantitative score for the degree of transcriptional perturbation in a disease cohort relative to the controls (Pankla, Buddhisa et al. 2009). The median MDTH of the active TB 664-transcript signature decreased significantly at two weeks onwards, compared to the median pre-treatment MDTH (Figure 50A).

In addition we devised a novel algorithm called temporal molecular response. This algorithm calculates the change in a transcriptional profile over time. Unlike MDTH it does not require a control cohort as it uses the pre-treatment time point as the comparator profile. It is also more sensitive to changes in longitudinal analysis; in part because it does not rely on a control cohort that can have variable profile heterogeneity.

For a given signature the temporal molecular response was determined by measuring the transcriptional perturbation between two time points, and expressing this value as a percentage of the total number of transcripts constituting the signature (see methods for details). The mean temporal molecular response calculated for the active TB 664-transcript signature revealed a statistically significant change in the transcriptional response at 2 weeks after treatment initiation (Figure 50B). This continued to change between 2 weeks and 2 months, and between 2 weeks and 6 months, after treatment initiation. The magnitude of the patient's temporal molecular response during treatment did not correlate with the magnitude of their untreated transcriptional signature, as measured by MDTH ($p<0.01$) (Figure 50C).

Deriving a treatment specific transcriptional signature

Although the active TB 664-transcript signature was shown to change significantly in response to successful treatment we wished to derive a more specific treatment-related signature. To determine this the South Africa cohort was randomised into two groups of patients, 15 patients into a training set and 14 patients into a test set. The signature was then derived from the training set and validated in the test set. A three-fold filter between the untreated samples and the mean of their paired 6-month samples was applied, where transcripts had to satisfy the filter in 12 of the 15 patients, followed by a statistical filter (Mann Whitney paired, Benjamini Hochberg $p<0.01$). This generated 320 transcripts as significantly differentially expressed between the untreated active TB training set samples and their paired 6-month treated samples (Figure 51). The treatment specific 320-transcript signature was shown to rapidly and significantly change at two weeks onwards after treatment initiation, in the active TB training set (Figure 51A &

B). This was validated in the active TB Test Set (Figure 51C & D). In both cohorts the change in the temporal molecular response was significant at 2 weeks post-treatment.

However analysis by both algorithms, MDTH and temporal molecular response, were not able to show significant changes between two months and six months. Therefore to establish whether any significant changes occurred from two months onwards, each of the profiles from each time point were compared to the latent TB profiles. 96 transcripts were significantly differentially expressed between two months and latent TB (Mann Whitney paired Benjamini Hochberg $p<0.01$, data not shown). As expected no genes were significantly differentially expressed between 6 months & 12 months, or 6 months & latent TB, or 12 months & latent TB (Mann Whitney paired Benjamini Hochberg $p>0.01$).

These results show the 320-transcript treatment specific signature and the 664-active TB signature changed significantly over time in response to treatment, as early as two weeks after treatment initiation and onwards at two months. The transcriptional profiles however appeared not to change significantly at the end of treatment onwards (at the 6 and 12 month time points) compared to the latent controls.

Comparing the genes lists from the active TB and treatment specific signatures

A Venn diagram was then used to compare the similarity of the two signatures: the 664-transcript active TB and 320-transcript treatment specific signatures. The treatment specific signature contained 74% of genes present in the active TB signature (Figure 52). IPA of the active TB 664-transcript signature demonstrated a highly significant over-representation of IFN-signalling genes including Type I and Type II IFN (Figure 52). IPA of the 320 transcripts indicated the most significantly represented pathways were related to the innate immune pathways, encompassing genes related to complement and Toll-like receptors.

Measuring an individual patient's transcriptional response to antituberculous treatment

To determine if the treatment response could be measured on an individual patient basis the temporal molecular response was applied to each patient individually. Each patient's discrete 320-transcript treatment specific response was visualised first in the heatmap and then quantified by the temporal molecular response in the training set (Figure 53) and in the test set (Figure 54). All 29 patients in the active TB treated cohort had a rapid and early positive temporal response after two weeks of treatment. Interestingly, not all the individual transcriptional responses were identical as demonstrated by the quantitative scoring provided by the temporal molecular responses. Some of the patients also had a slight increase in their temporal molecular response and in the heatmap at 12 months.

Further validation of the 2 week treatment transcriptional response

To determine whether the significant change in the treatment specific 320-transcript signature that we had demonstrated in a South African cohort was also applicable to patients in an intermediate burden setting, we tested the signature in a UK cohort. As observed in the South African cohort the signature was rapidly and significantly diminished from two weeks post-treatment initiation (Figure 55). The changes in the blood transcriptional response could also be clearly quantified in individual patients by the temporal molecular response. The significant transcriptional blood change correlated with successful treatment of patients as assessed after 6 months by radiographic and clinical parameters (data not shown).

For additional validation that the active-TB transcriptional signatures showed significant changes as early as two weeks after treatment initiation, the active TB signatures (393 and 86 transcript signatures) from the Berry *et al.* 2010, study were also used in the treated South African cohort (Berry, Graham et al. 2010). Both signatures again significantly diminished after two weeks treatment (Figure 56). In addition the TB-specific 144 transcript signature, derived from comparing pulmonary TB to active pulmonary sarcoidosis as described in chapter 5, also showed significant changes as early as 2 weeks (Figure 57).

Discussion

TB treatment monitoring is a major challenge in attempts to eradicate *M.tuberculosis* infection. In April 2010 the Centers for Disease Control and National Institutes of Health brought together experts in the field and research scientists with the sole purpose of addressing this problem (Nahid, Saukkonen et al. 2011). Poor treatment monitoring, and hence inadequate treatment, leads to worsening of a patient's disease, increasing the potential for disease spread and the risk of developing drug resistant mycobacteria. Currently the two-month sputum culture conversion is the only biomarker of successful TB treatment (Mitchison 1993). However it is time consuming, taking several weeks to grow the bacilli and results can be compromised by contamination. Moreover patients who have clinically improved may be unable to expectorate sputum at two months and potentially incorrectly labelled as having a negative culture (Perrin, Lipman et al. 2007). Furthermore, although sputum conversion is commonly used as a surrogate end point for treatment response in clinical trials evaluating new drugs, a systematic review and meta-analysis to assess its accuracy in predicting an individual's treatment failure revealed low sensitivity and only modest specificity (Horne, Royce et al. 2010; Wallis, Pai et al. 2010). While other biomarkers have also been trialled, including C-reactive protein, IFNγ and neopterin, all have similarly shown poor sensitivity and specificity (Walzl, Ronacher et al. 2008). Chest radiographs are commonly used in the clinical setting as a marker of treatment response but they generally improve slower than the clinical response and lack specificity as interpretation can be confounded by previous lung damage (Perrin, Lipman et al. 2007). Moreover interpretation of radiographic changes in response to treatment has not yet been standardised, and the facilities are not always available in developing countries (Walzl, Ronacher et al. 2011). Therefore there is clearly a need for early and easily detectable biomarkers for treatment monitoring, capable of potentially identifying poor responses due to drug resistance or lack of treatment adherence, and available for patients unable to produce sputum. In this chapter we have shown and validated that two signatures, an active TB signature and a treatment specific signature, both significantly diminish after just two weeks of treatment. In addition the transcriptional response to antituberculous treatment could also be individually quantified for each patient. Together, these findings suggest that blood transcriptional signatures could be used as early surrogate biomarkers of a successful treatment response, in both the clinical setting and in drug development.

Study participants

The South African patients involved in this part of the study were active and latent TB patients recruited from a high burden area in South Africa. Khayelitsha, is a large peri-urban African township in Cape Town which has over 1000 TB notifications annually. From the clinical data available it can be observed the South African pulmonary TB patients had more severe disease than the UK patients, with higher bacilli loads and higher number of presenting symptoms (Tables 28-29). The is likely to explain the greater transcript normalised intensity values that can be seen when comparing the heatmaps from the South African patients (range +4.5 to -4.5) and the UK patients (+3.8 to -3.8) (Figures 51 & 55).

The transcriptional response changes significantly at 2 weeks

Berry *et al.* 2010, previously demonstrated in a small number of patients that blood transcriptional signatures in UK active TB patients diminished after two months of antituberculous treatment (Berry, Graham et al. 2010). Two other studies have also described relevant treatment related transcriptional differences. Mistry *et al* found that patients who had completed a course of antituberculous treatment displayed similar expression profiles to a latent TB group, but they did not examine any patients during their antituberculous treatment course (Mistry, Cliff et al. 2007). Joosten *et al* showed in a small number of samples that their active TB gene set diminished after two months of antituberculous treatment, however they did not examine any patients at earlier time points (Joosten, Goeman et al. 2012).

From this current study it can be seen a significant blood transcriptional response to antituberculous treatment occurs as early as two weeks (Figures 47-57). This early transcriptional response could be as a consequence of the observed rapid and high killing capacity of antimycobacterial antibiotics leading to a substantial reduction in mycobacterial load (Jindani, Aber et al. 1980; Gumbo, Louie et al. 2007; de Steenwinkel, de Knegt et al. 2010). Although the signatures derived may not be completely specific for active TB, since clinically similar diseases such as sarcoidosis show common transcripts (Koth, Solberg et al. 2011), demonstration of a response to antimycobacterial therapy, could help resolve this overlap. Furthermore the TB-specific signature (144 transcripts) derived from comparing pulmonary TB to active pulmonary sarcoidosis also significantly diminished with treatment (Figure 57).

The treatment specific 320-transcript signature also had many genes in common with the active TB 664-transcript signature (Figure 52). This overlap of genes is highly suggestive that this study will help guide future development of a subset of genes that most accurately correlates with a patient's response to antituberculous treatment, acting as a surrogate marker of treatment failure or success. Both derived signatures, the South African 664-transcript active TB signature and the treatment specific 320-transcript signature, were dominated by IFN signalling and innate immune response genes (Figure 52). These findings are in agreement with earlier gene expression studies in TB (Berry, Graham et al. 2010; Maertzdorf, Repsilber et al. 2011) and findings from the earlier cohorts described in this study (chapters 3 & 5).

Two algorithms can demonstrate significance of the change in transcriptional profiles

It has previously been shown that MDTH positively correlates with the severity of active pulmonary TB, as defined by the radiological extent of disease (Berry, Graham et al. 2010). However the 'temporal molecular response' offers a potential advantage in the clinical setting, as it allows an individual assessment of each patient's expression change and does not require a reference control group. Both algorithms clearly demonstrated the significant change that occurs at least two weeks after commencing treatment (Figure 50A & B). In addition there was no correlation between the pre-treatment MDTH and the two-week or two-month temporal molecular response (Figure 50C). This suggested a patient's untreated transcriptional signature is not predictive of the patient's treatment response.

The transcriptional response continues to change after 2 weeks

A further problem in the management of TB is the extended length of treatment, requiring a minimum duration of six months. However the treatment duration required for maximum efficacy and preventing resistance, has not been fully established. The ability therefore to stratify patients into groups requiring shorter or longer treatment durations, particularly in resource limited settings, could be of value in improving patient compliance and reducing treatment related side effects. It can be seen from the individual temporal molecular responses that some patient's transcriptional response appeared to plateau before six months (Figures 53-55**)** suggesting blood transcriptional signatures may could aid in patient stratification for treatment with differing regimen lengths. It can also be observed some of the South African's transcriptional responses appear to have increased at twelve months relative to their six month score, although their profiles at twelve months resembled the latent profiles (Figures 53 & 54). One could postulate this is a consequence of stopping the broad spectrum antibiotics. For example, either the patient has since acquired or re-reactivated a bacterial infection between the six month and the twelve month time point (no patients were thought to be positive for *M. tuberculosis* infection) or this signature is secondary to antibiotic-induced changes to the gut microbiome.

This robust correlation of a significant change in transcriptional response to successfully antituberculous treatment occurring between different host populations, likely different *M. tuberculosis* strains, diverse environments and microarray analysis strategies indicates that blood transcriptomics have great potential to be developed into treatment monitoring biomarkers.

Chapter Summary

A whole blood active TB 664-transcript signature and a treatment specific 320-transcript signature significantly changed in active TB patients after just 2 weeks of initiation of clinically successful antituberculous treatment. The significant change in the treatment-specific signature was observed in patients from the high TB-burden setting of South Africa and from the intermediate TB-burden setting of London. Both the active-TB and treatment-specific transcriptional signatures were dominated by IFN-signalling and innate immune response genes. The transcriptional response to antituberculous treatment could be individually quantified for each patient. Together, these findings suggest that blood transcriptional signatures could be used as early surrogate biomarkers of a successful treatment response, in both the clinical setting and in drug development.

Figures for chapter 7

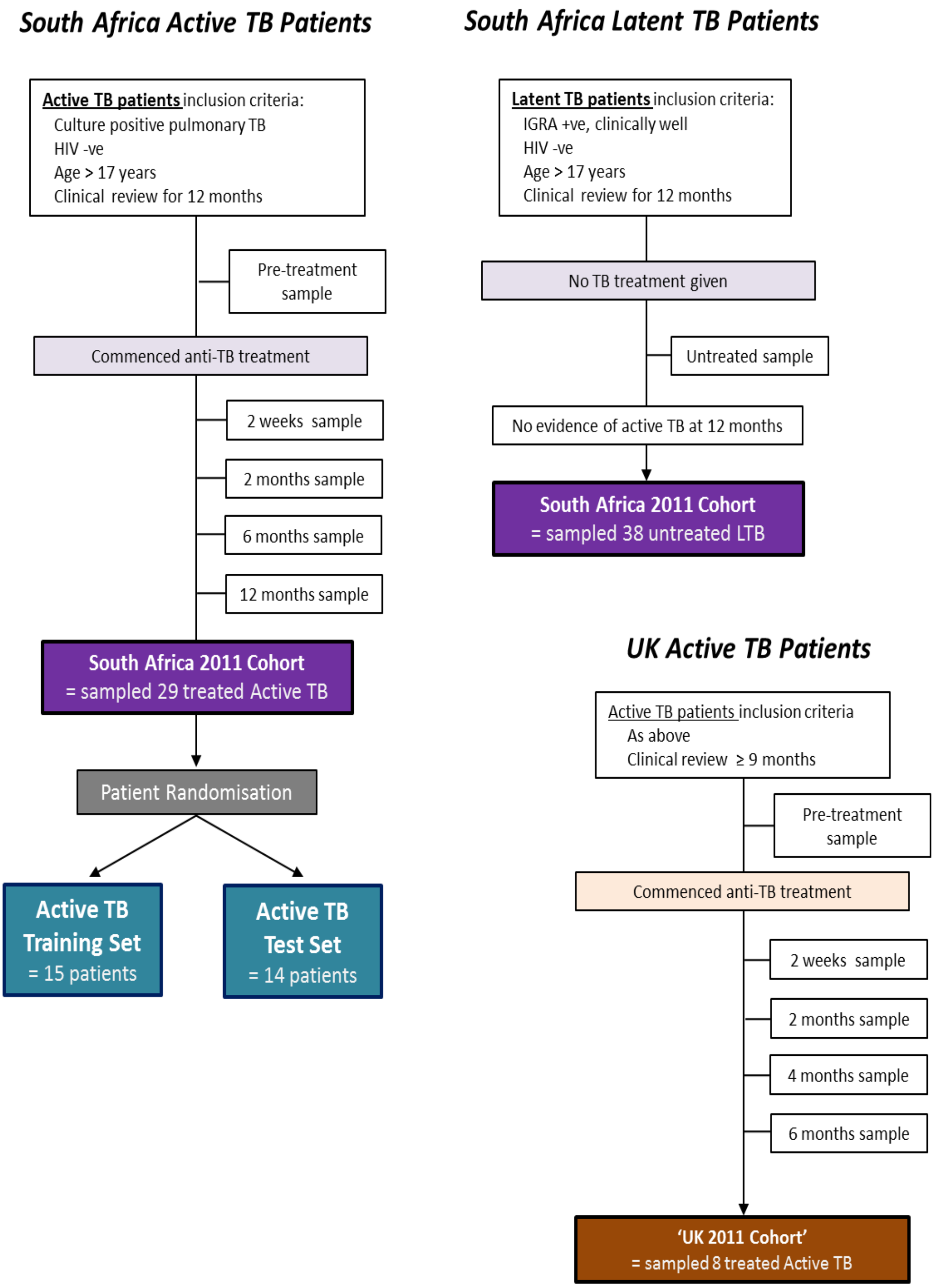
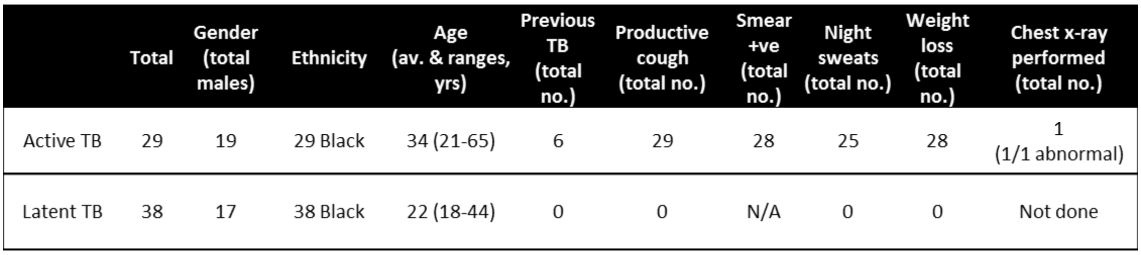
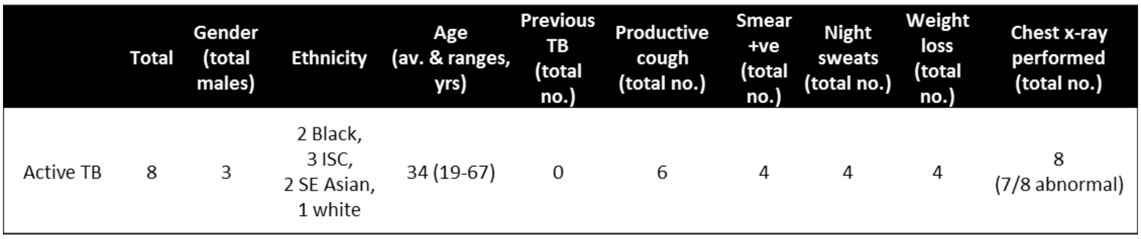

South Africa Active TB Patients
Active TB patients inclusion criteria:
Culture positive pulmonary TB
HIV -ve
Age > 17 years
Clinical review for 12 months
Pre-treatment sample
Commenced anti-TB treatment
2 weeks sample
2 months sample
6 months sample
12 months sample
South Africa 2011 Cohort
= sampled 29 treated Active TB
Patient Randomisation
Active TB Training Set
= 15 patients
Active TB Test Set
= 14 patients
South Africa Latent TB Patients
Latent TB patients inclusion criteria:
IGRA +ve, clinically well
HIV -ve
Age > 17 years
Clinical review for 12 months
No TB treatment given
Untreated sample
No evidence of active TB at 12 months
South Africa 2011 Cohort
= sampled 38 untreated LTB
UK Active TB Patients
Active TB patients inclusion criteria
As above
Clinical review ≥ 9 months
Pre-treatment sample
Commenced anti-TB treatment
2 weeks sample
2 months sample
4 months sample
6 months sample
'UK 2011 Cohort'
= sampled 8 treated Active TB

Figure 46. Recruitment of South Africa and UK TB patients before, during and after treatment.



	Total	Gender (total males)	Ethnicity	Age (av. & ranges, yrs)	Previous TB (total no.)	Productive cough (total no.)	Smear +ve (total no.)	Night sweats (total no.)	Weight loss (total no.)	Chest x-ray performed (total no.)
Active TB	29	19	29 Black	34 (21-65)	6	29	28	25	28	1 (1/1 abnormal)
Latent TB	38	17	38 Black	22 (18-44)	0	0	N/A	0	0	Not done

Table 28. South African treated active TB patients and untreated latent TB patients.



	Total	Gender (total males)	Ethnicity	Age (av. & ranges, yrs)	Previous TB (total no.)	Productive cough (total no.)	Smear +ve (total no.)	Night sweats (total no.)	Weight loss (total no.)	Chest x-ray performed (total no.)
Active TB	8	3	2 Black, 3 ISC, 2 SE Asian, 1 white	34 (19-67)	0	6	4	4	4	8 (7/8 abnormal)

Table 29. UK treated active TB patients.

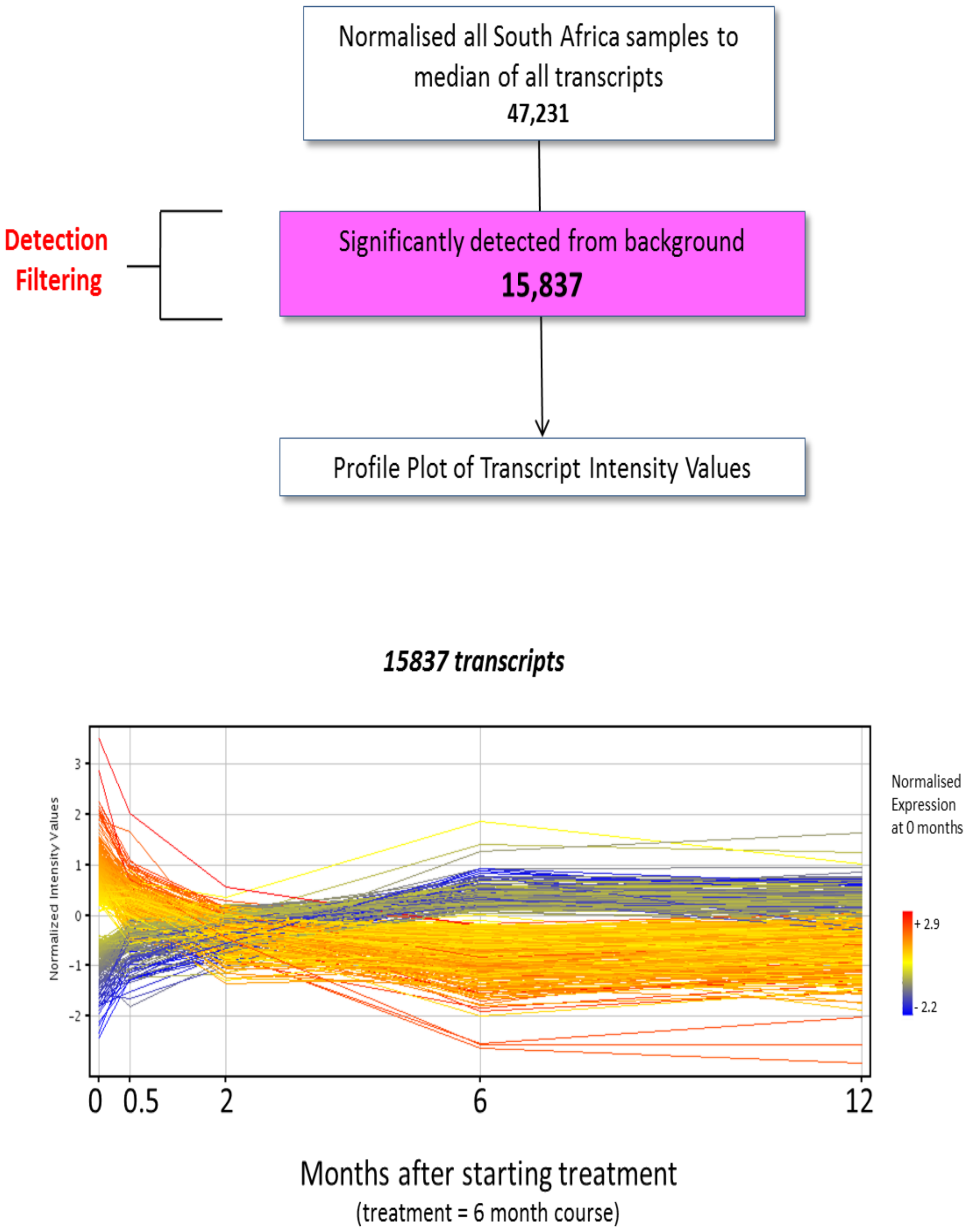
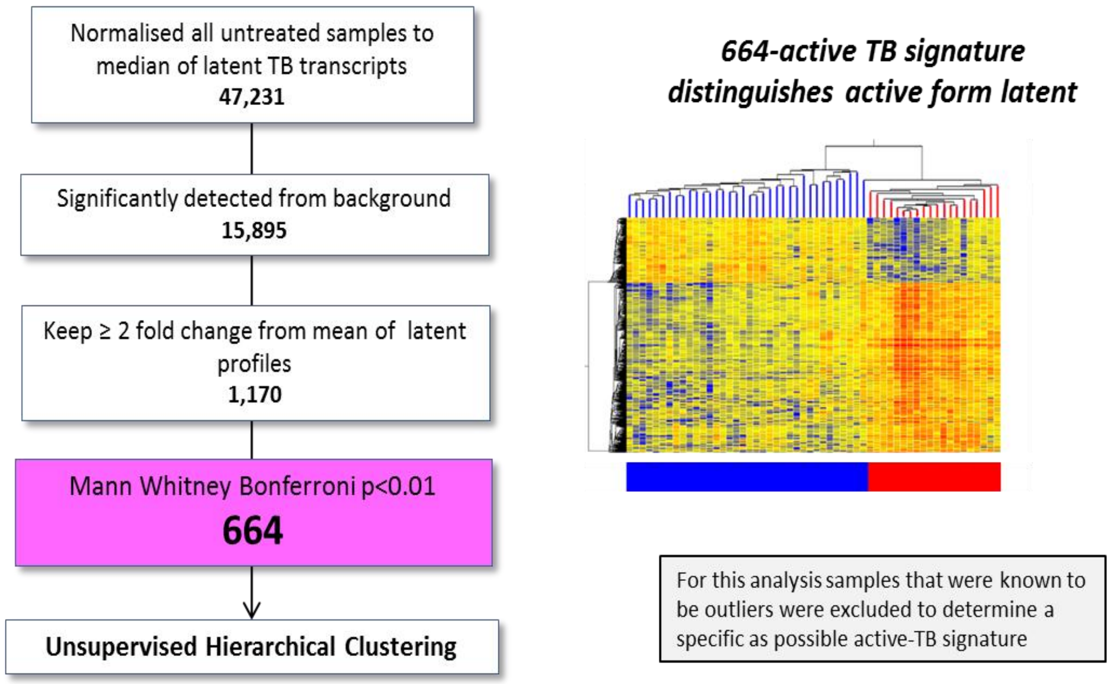


Figure 47. Individual patient’s transcriptional response occurs at variable rates south africa training set.

The transcripts were derived as shown at the top of this figure. The profile plot displays the expression normalised intensity values over time where the red profiles indicate a high normalised value at baseline and blue a low normalised value (as indicated by the colour bar).


Normalised all untreated samples to median of latent TB transcripts
47,231
Significantly detected from background
15,895
Keep ≥ 2 fold change from mean of latent profiles
1,170
Mann Whitney Bonferroni p<0.01
664
Unsupervised Hierarchical Clustering
664-active TB signature distinguishes active form latent

For this analysis samples that were known to be outliers were excluded to determine a specific as possible active-TB signature

Figure 48. To refine the transcript list capable of demonstrating a transcriptional change in response to treatment 2 analysis strategies were used (1) the South Africa active TB signature changes in response to treatment.

The transcripts were derived as shown in the left of this figure. The layout of the heatmap is as described in figure 13.

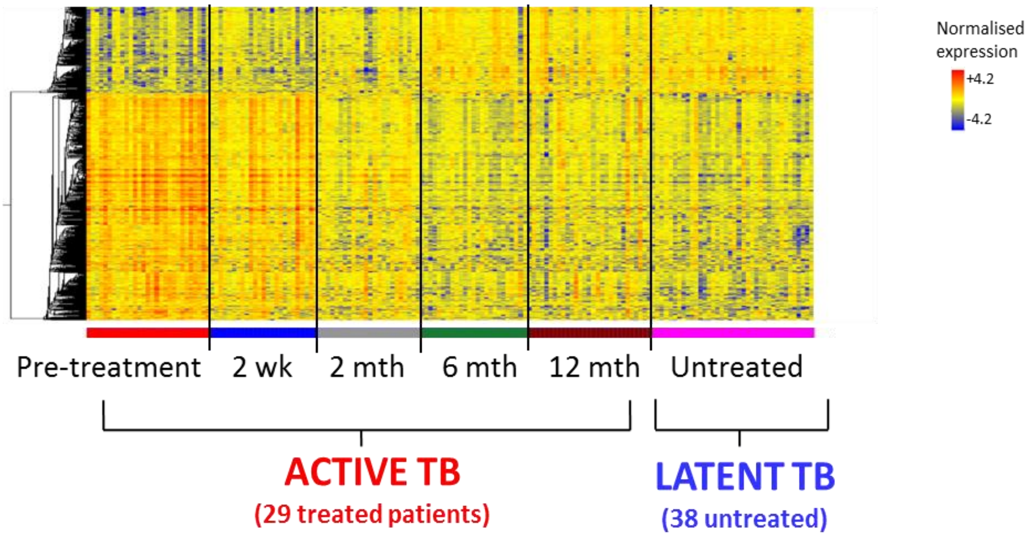
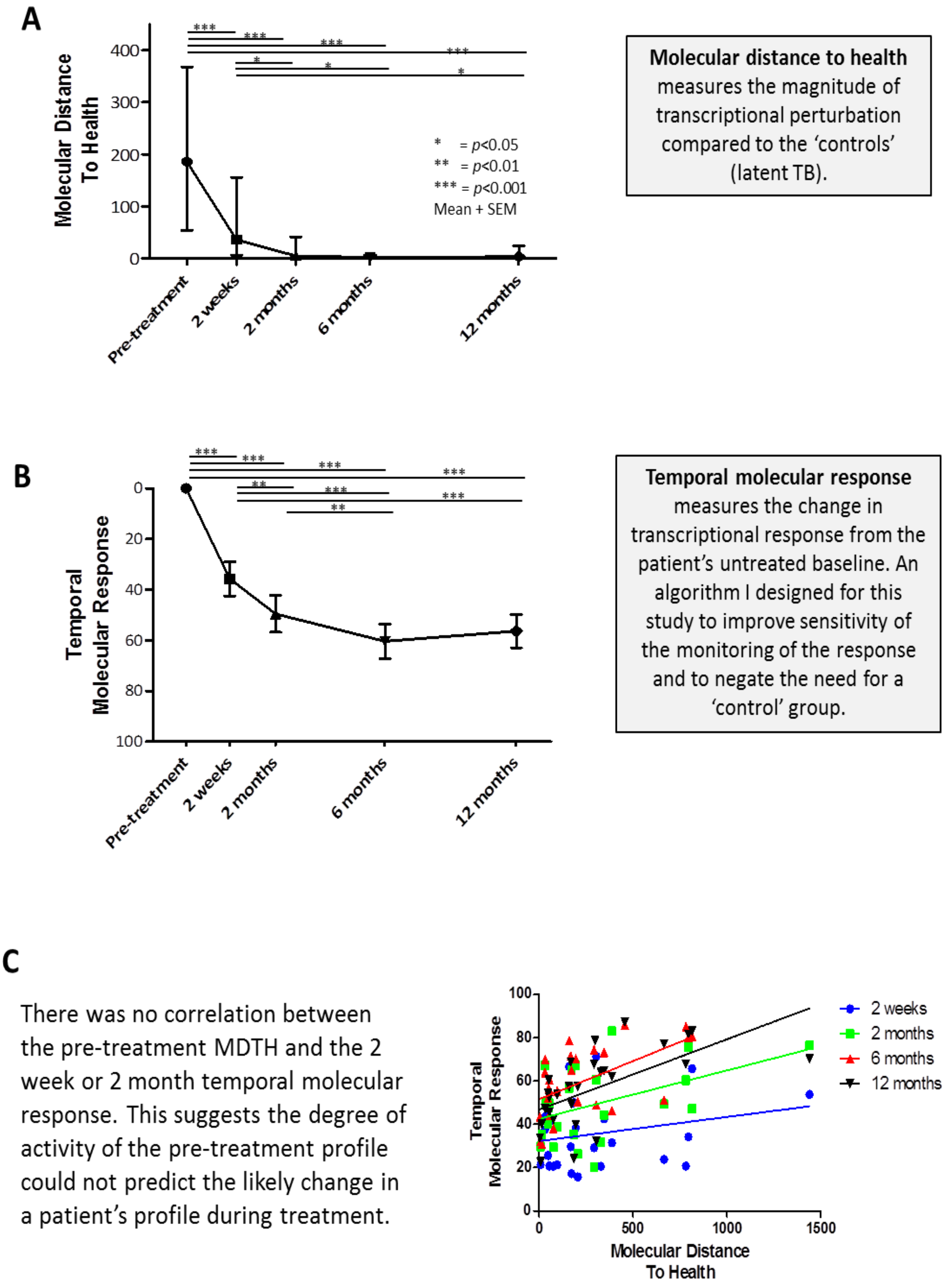


Figure 49. The 664 signature is applied to all patients before, during and after treatment and demonstrates a change in the signature at 2 weeks onwards.

The layout of the heatmap is as described in figure 13.



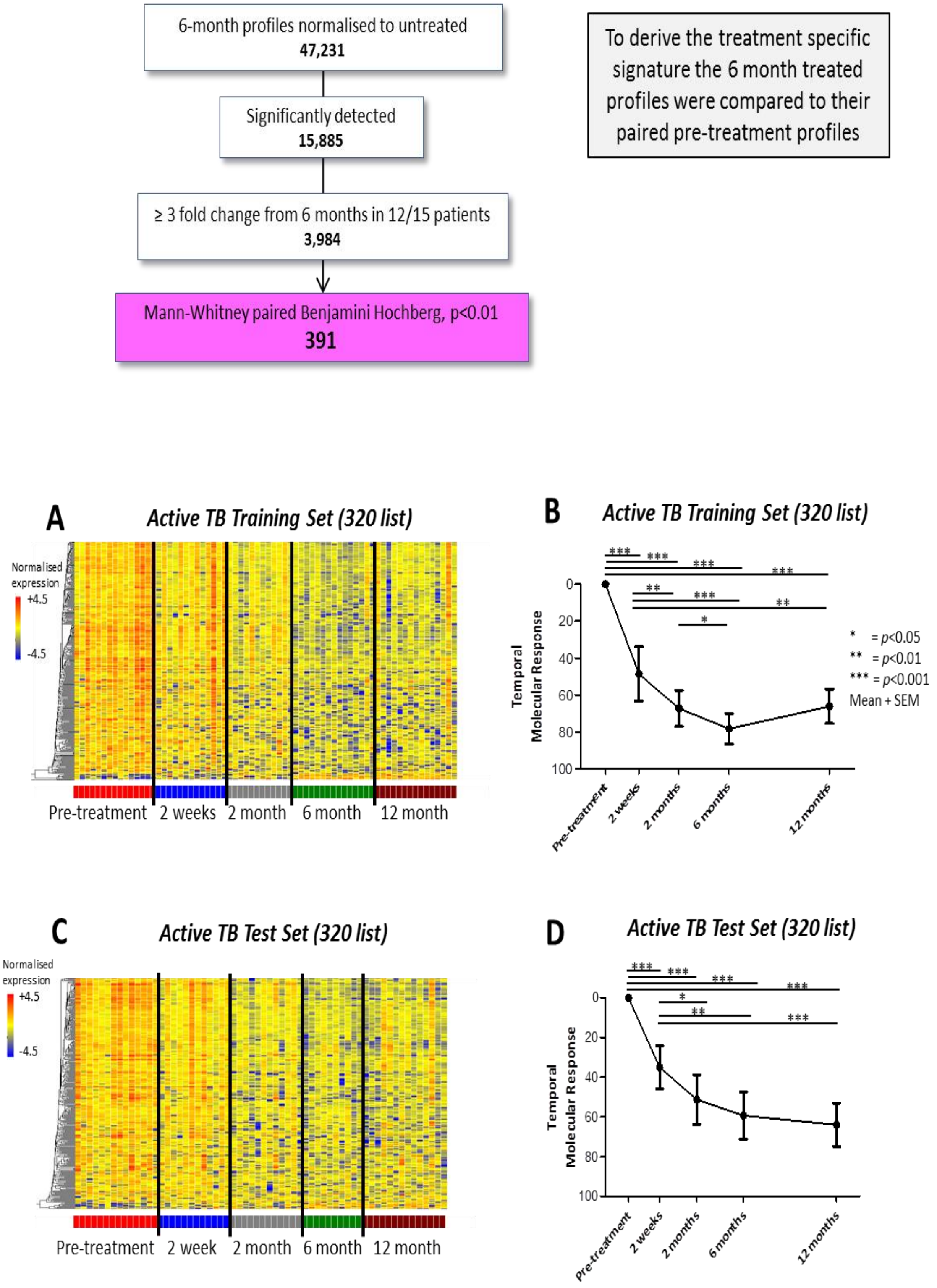
Molecular distance to health measures the magnitude of transcriptional perturbation compared to the 'controls' (latent TB).

Temporal molecular response measures the change in transcriptional response from the patient's untreated baseline. An algorithm I designed for this study to improve sensitivity of the monitoring of the response and to negate the need for a 'control' group.

There was no correlation between the pre-treatment MDTH and the 2 week or 2 month temporal molecular response. This suggests the degree of activity of the pre-treatment profile could not predict the likely change in a patient's profile during treatment.

Figure 50. The change in the signature of the 664 active-TB signature is statistically significant at 2 weeks as shown by both MDTH and the temporal molecular response.

The graphs A and B display the mean, SEM, *p* value from longitudinal regression analysis with fixed effects. Graph C displays linear regression best fit slopes.



6-month profiles normalised to untreated
47,231

Significantly detected
15,885

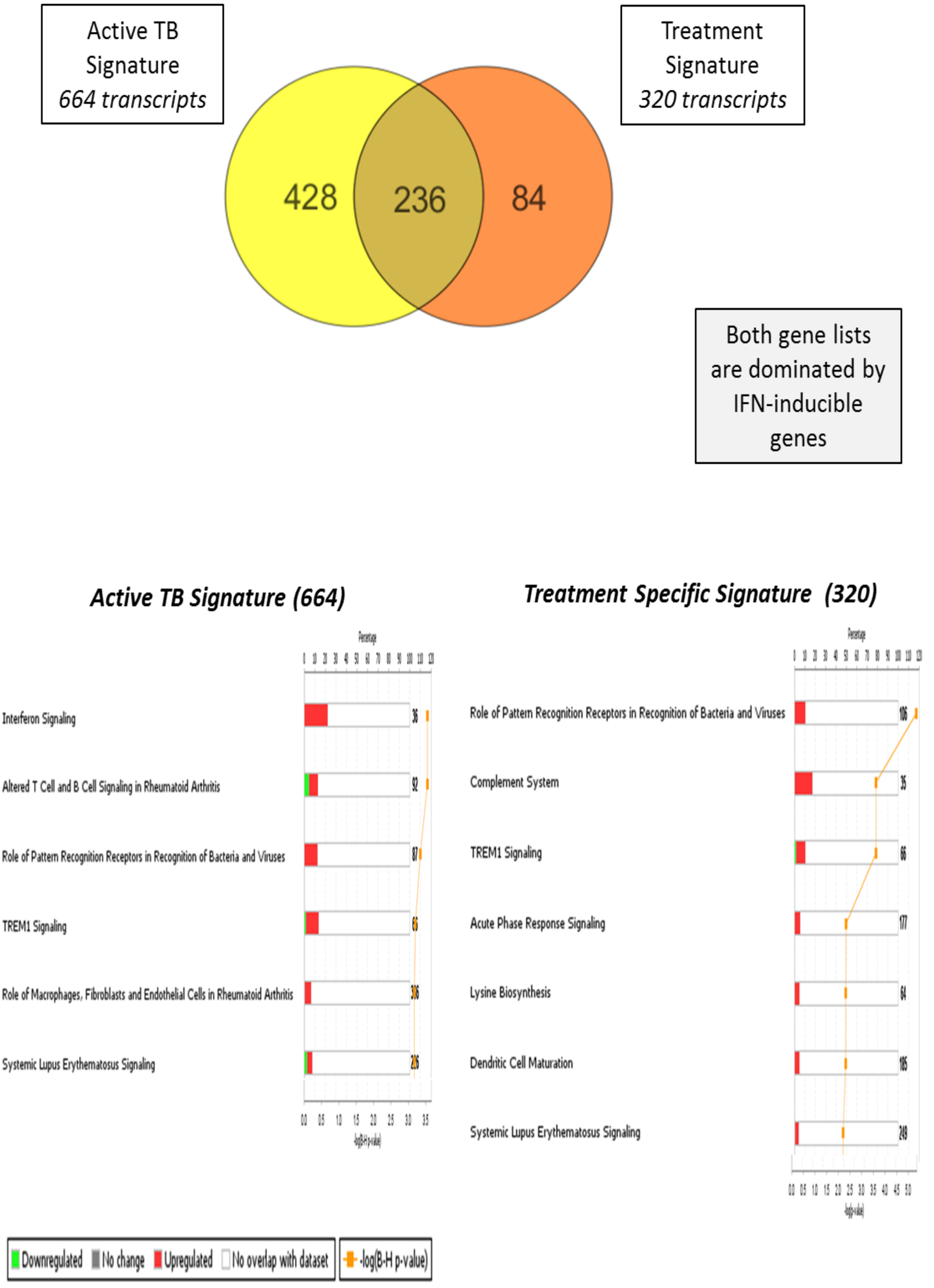
≥ 3 fold change from 6 months in 12/15 patients
3,984

Mann-Whitney paired Benjamini Hochberg, p<0.01
391

To derive the treatment specific signature the 6 month treated profiles were compared to their paired pre-treatment profiles

Figure 51. To refine the transcript list demonstrating a transcriptional response to treatment two strategies were used (2) Treatment specific signature also significantly changes at 2 weeks.

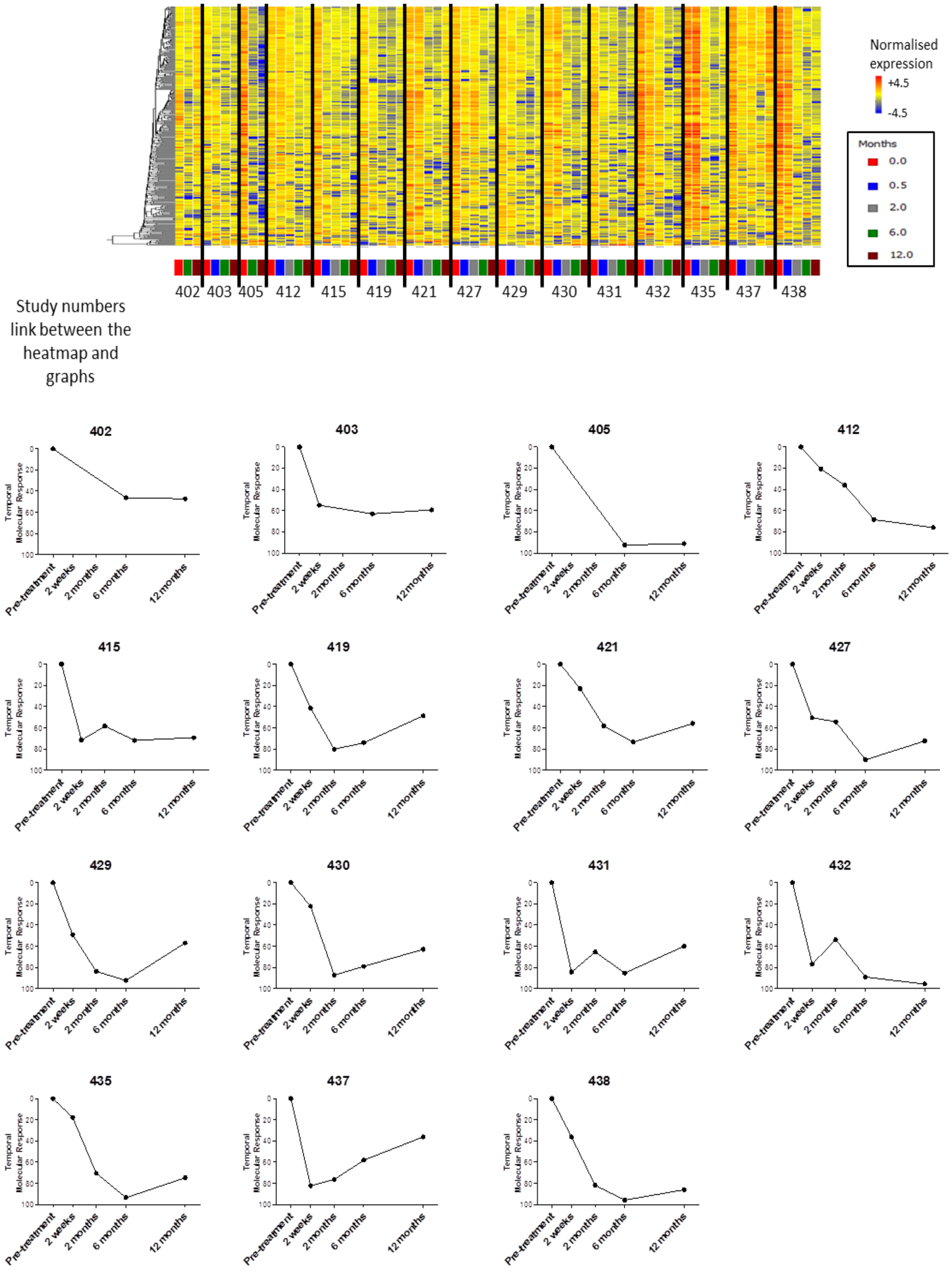
The transcripts were derived as shown at the top of this figure. The layout of the heatmap is as described in figure 13. The graphs B and D display the mean, SEM, and *p* value from longitudinal regression analysis with fixed effects.



© 2000-2012 Ingenuity Systems, Inc. All rights reserved.

Figure 52. Many genes overlap in the active TB signature and the treatment specific signature.

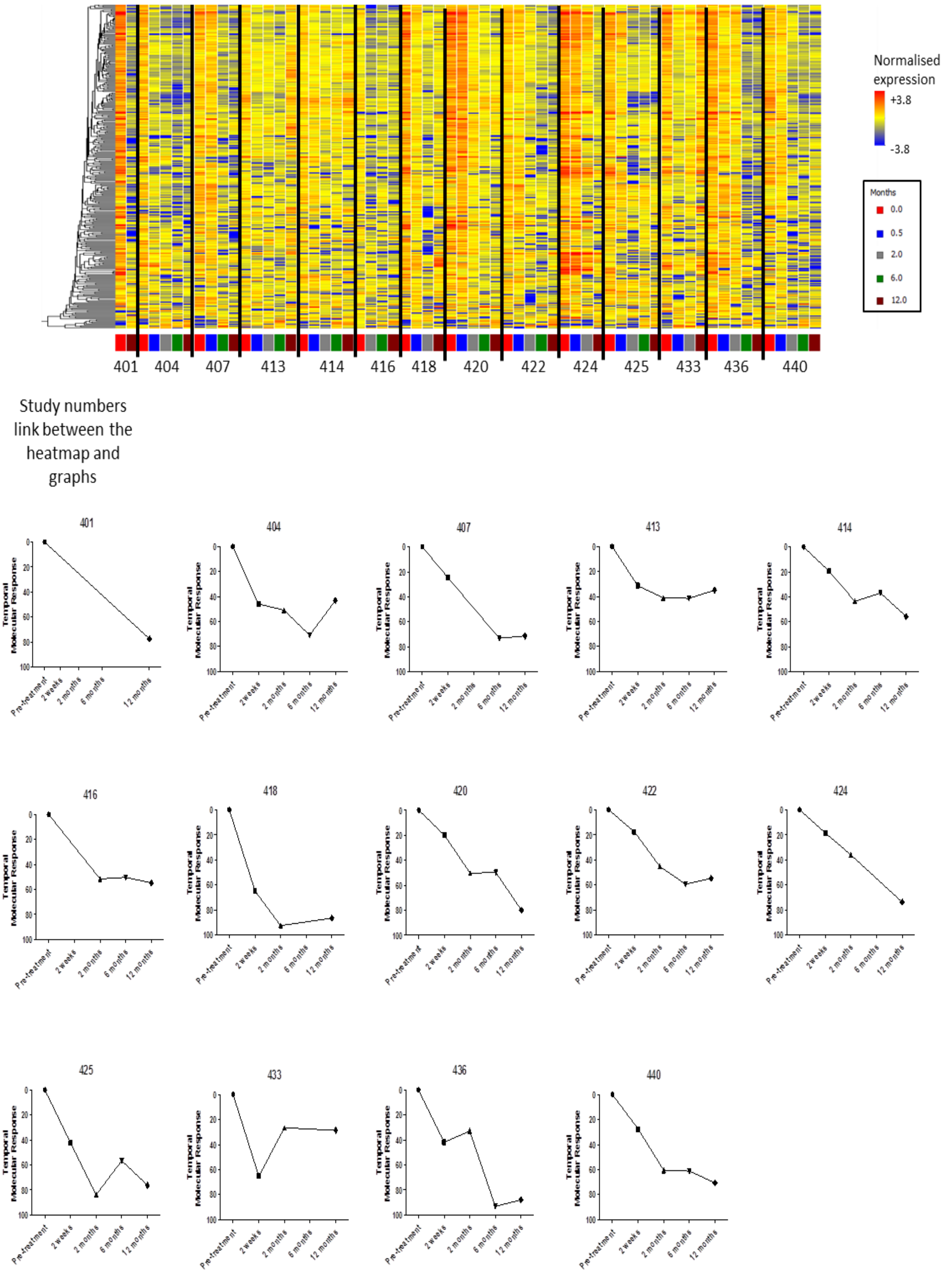
The layout of the IPA pathways are as described in figure 21.



Study numbers link between the heatmap and graphs

Figure 53. Individual patient's transcriptional response occurred at variable rates in the South Africa training set.

Both the heatmap and graphs show each patient's 320 treatment response profile over time. The transcripts were derived as shown in figure 51. The layout of the heatmap is as described in figure 13.



Study numbers link between the heatmap and graphs

Figure 54. Individual patient's transcriptional response occurred at variable rates in the South Africa test set.

Both the heatmap and graphs show each patient's 320 treatment response profile over time. The transcripts were derived as shown in figure 51. The layout of the heatmap is as described in figure 13.

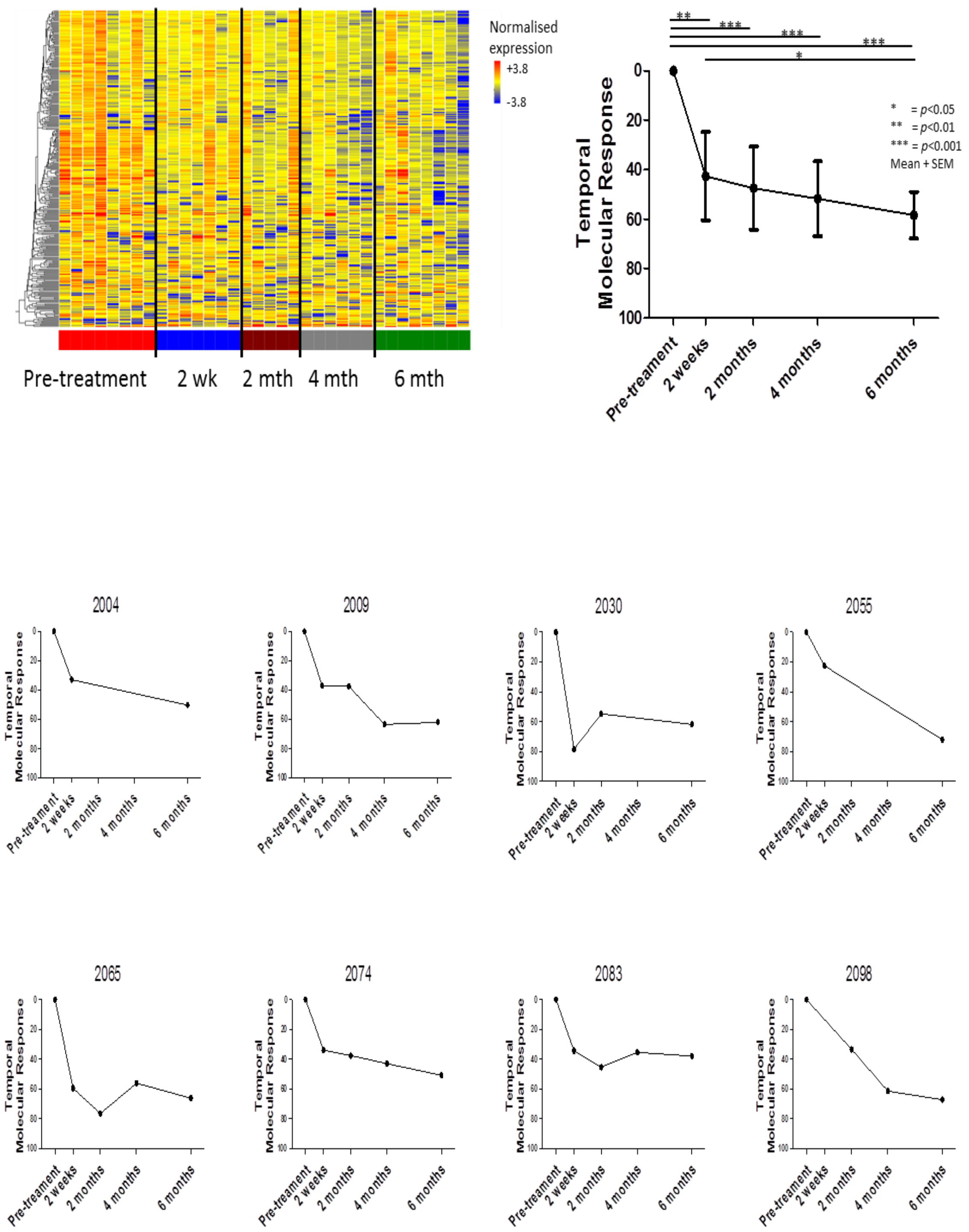


Figure 55. The changes in the treatment specific signature were validated in an independent UK cohort.

Both the heatmap and graphs show each patient's 320 treatment response profile over time. The transcripts were derived as shown in figure 51. The layout of the heatmap is as described in figure 13. The cumulative data graph display the mean, SEM, and *p* value from the longitudinal regression analysis with fixed effects.

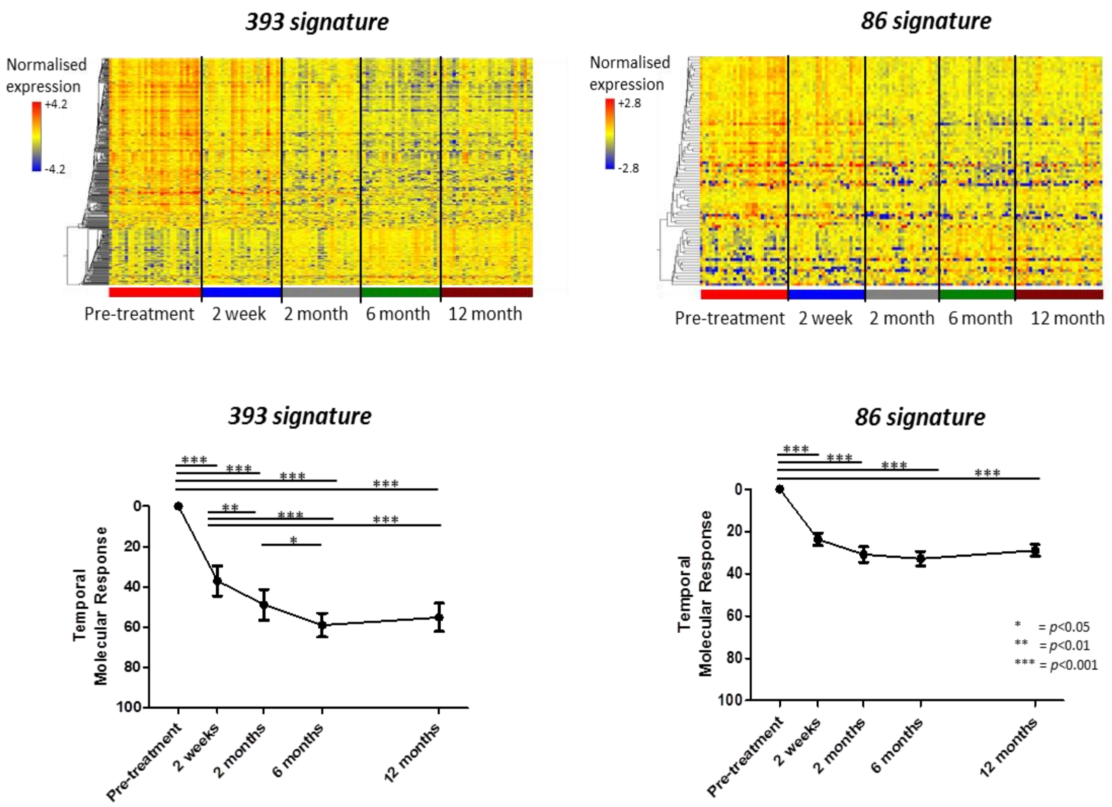


Figure 56. The Berry *et al* active TB signatures also change significantly in response to successful treatment.

The graphs display the mean, SEM, and *p* value from the longitudinal regression analysis with fixed effects. The layout of the heatmap is as described in figure 13.

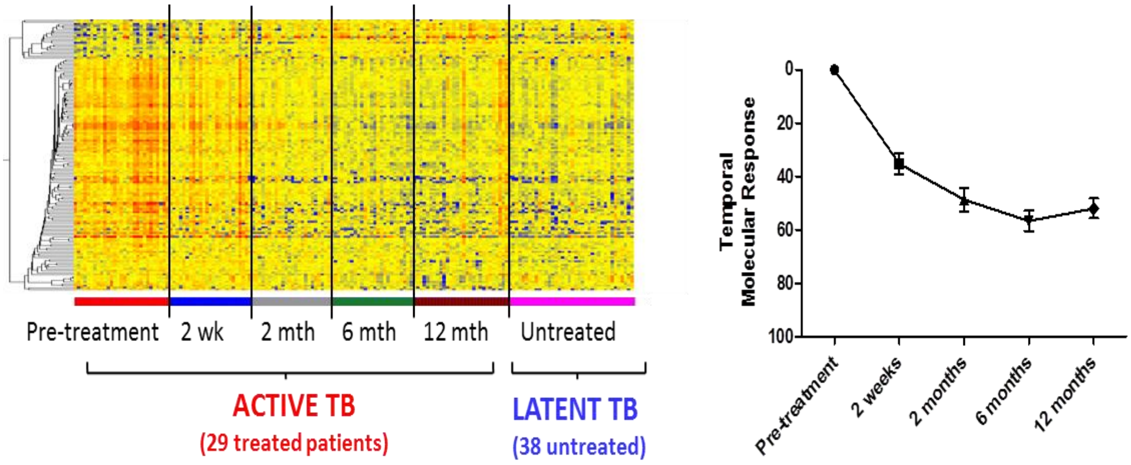


Figure 57. The 144 TB-specific signature identified by comparing TB to active sarcoidosis also significantly changed at 2 weeks of treatment.

The transcripts were derived as shown in figure 40. The layout of the heatmap is as described in figure 13.

Chapter 8

Transcriptional profiles change during and after treatment of sarcoidosis and pneumonia

Chapter 8: Transcriptional profiles change during and after treatment in sarcoidosis and pneumonia

Introduction

There are many decisions concerning the treatment of pulmonary sarcoidosis patients including who should receive it, when to start, which medications to use – which ones work and which ones have unacceptable side-effects, which tests can assess a treatment response, how to know if there is a response, how can relapses be prevented and how long to continue treatment? The lack of corroborated answers is due to a number of factors including a paucity of acceptable randomised controlled trials, the complexity of the underlying pathology, the clinical heterogeneity and the uncertain effectiveness of immunosuppression which is the current mainstay of sarcoidosis treatment (Paramothayan and Jones 2002; Baughman 2003; Judson 2003; Paramothayan, Lasserson et al. 2005; Coker 2007; Baughman and Nunes 2012). It is still not clear how effective immunosuppression is because many patients have self-limited disease or spontaneous resolution, and defined criteria for the monitoring of a treatment response is not standardised (WASOG 1999; Bradley, Branley et al. 2008). A better understanding therefore of what the treatment is doing at a molecular level, in parallel with clinical parameters, could help us understand how the treatment is exerting its effects and may help in our understanding of the underlying immunopathogenesis of sarcoidosis.

All but one of the sarcoidosis patients in this study were treated with glucocorticoids after initial recruitment. Glucocorticoids are frequently used in inflammatory and autoimmune conditions due to their potent anti-inflammatory actions and immune modulatory effects. Glucocorticoids can effect target cell transcriptional regulation and protein synthesis by several different mechanisms. Firstly they bind to the ubiquitous cytoplasmic glucocorticoid receptors (GCR), DNA-binding transcription factors, which once activated thereby release chaperone molecules such that the GCR can translocate to the nucleus (McColl, Michlewska et al. 2007). GCRs can then transactivate (increase transcription of anti-inflammatory genes) or transrepress (inhibit pro-inflammatory genes) through specific binding to positive or negative glucocorticoid-responsive elements in the promoter region of glucocorticoid-responsive genes (McColl, Michlewska et al. 2007). In addition the GCR-complex can cause transrepression in a hormone dependent manner through modulation of the activity of transcription factors such as NK-κB or AP-1 (De Bosscher, Vanden Berghe et al. 2003). Transcriptional regulation is then achieved through modification of core histones, structural remodelling and DNA methylation (Biddie, Conway-Campbell et al. 2012).

This part of our study is focussing only on the sarcoidosis patients before they commenced treatment and again while they were receiving it. To add a valuable dimension to the analysis the transcriptional profiles of community acquired pneumonia patients before and after their curative antibiotics were also examined, as well as compared to the transcriptional changes that were seen in the successfully treated TB patients.

Results

Sarcoidosis profiles changed only if they had shown a good clinical response to treatment

To investigate the effect of treatment on the transcriptional profiles of the sarcoidosis patients an unsupervised analysis approach followed by unsupervised hierarchical clustering was performed on the seven sarcoidosis patients who were followed up after commencement of treatment. Unsupervised analysis generated 5,223 transcripts (Figure 58). Unsupervised hierarchical clustering of the transcripts and samples only revealed a difference in transcript abundance in the seven treated sarcoidosis patients if they had a good clinical response to treatment (Figure 58). The decision to commence treatment was made at the 'first clinic visit' and subsequent clinic visits were then labelled sequentially. Four of the patients only had blood samples taken at the second clinic visit, three patients had samples also taken at the third clinic visit (Figure 59). A clinical response to treatment was defined retrospectively by the practising physician's management at the second and third clinic appointments. If the physician increased the treatment this was interpreted as a lack of adequate response to treatment, but if the treatment was reduced or maintained this was interpreted as an adequate response. The sarcoidosis treated patients were seen by three independent sarcoidosis specialists and at the time each physician was not aware that their decision was to be included as part of the data collection for this study. Five patients were thought not to be responding adequately by their practising physician (Figure 59a). Two of these patients subsequently did respond well after altering their treatment (patient no. 3 and 5). For the other three patients samples were not available on their third visit (patients no. 2, 4 & 7), one of the three inadequately responding patients was receiving hydroxychloroquine alone. During the second visit patient no. 5 was labelled as an inadequate responder because she was thought to be still symptomatic from her sarcoidosis. Interestingly the physician documented in her clinical notes that the symptoms were either due to the sarcoidosis or to the prednisolone. But, due to the lack of clarity, the physician decided to increment her prednisolone by a small dose. By the third visit after reviewing her blood tests and radiology, which had been requested on her second visit, it was established her symptoms were actually due to the glucocorticoids not her sarcoidosis.

She subsequently was weaned on to a much lower dose of prednisolone and improved thereafter. Her profile clustered with the treated patients who showed a good clinical response not with those who had an inadequate response (Figures 58 & 59). Using unsupervised hierarchical clustering adequate responders were shown to cluster away from inadequate responders and the pre-treatment samples (Figure 58) and by supervised hierarchical clustering grouped into; pre-treatment, good responders and inadequate responders (Figure 59a) or grouped per person over time (Figure 59b). Undoubtedly a change in the transcript abundance correlated with a clinical response to treatment of sarcoidosis.

Good-responders to sarcoidosis treatment clustered separately from the pre-treatment sarcoidosis samples, untreated TB samples and the controls

Next we wished to determine if the profiles of the treated sarcoidosis patients with a good clinical response were similar to the profiles of controls or similar to profiles of the patients. Therefore we carried out unsupervised analysis and unsupervised hierarchical clustering of all the pre-treatment sarcoidosis patients, treated sarcoidosis patients, untreated TB patients and controls. Unsupervised analysis of the test set untreated TB, pre-treatment sarcoidosis, treated sarcoidosis and controls samples generated 2,077 transcripts (Figure 60). Unsupervised hierarchical clustering revealed a distinct cluster of the treated sarcoidosis patients who showed a good clinical response (Figure 60). This distinct cluster of good-response sarcoidosis patients fell within the main cluster with all the patients, not with the controls, however it was separate from the pre-treatment sarcoidosis patients and from the inadequately-treated sarcoidosis patients. The untreated TB and pre-treatment sarcoidosis patients clustered together as previously demonstrated in chapters 3 and 5. Within this cluster of untreated TB and sarcoidosis samples also lay the samples from the inadequately treated sarcoidosis patients. The patient who was originally misdiagnosed as steroid-unresponsive clustered with the adequately responding profiles (bright pink mark on the profile label bar, Figure 60). All but one of the controls clustered in a separate main cluster away from all the patients, regardless of treatment status. The test set was used as all the sarcoidosis pre- and post-treatment samples were processed and run for microarray with the test set samples.

Therefore these results demonstrated that the sarcoidosis patients with a good response to treatment showed a distinct transcriptional signature from the inadequately treated sarcoidosis patients and from the pre-treated sarcoidosis patients.

A good treatment response in sarcoidosis appears to induce an active transcriptional change with an over-abundance of many genes

From the heatmaps it can be seen numerous genes were highly over-abundant in the good-response sarcoidosis treatment group while some genes were highly under-abundant, compared to all the other patients and controls (Figure 60). A few of the good-responders also showed under-abundance of a subgroup of genes with normalised intensity values akin to the healthy controls. These genes include the IFN-inducible genes GBP1, GBP5, GBP6, STAT1, STAT2, IFI35, IRF1, TAP1, FCGR1A, FCGR1B and FCGR1C. Modular analysis was applied to determine and compare the biological functions of the different sarcoidosis treatment outcomes. By this analysis it could be seen there was an over-abundance of inflammation, cell death and DC/apoptosis genes in the good-responders compared to the inadequate-responders and pre-treatment samples, relative to the controls (Figure 61). The highly over-abundant inflammatory genes included IL1R2, IL1RAP, IL18RAP and DUSP1 genes. However there was little change in the IFN modules in response to clinically effective treatment. There was under-abundance of lymphocyte activation, B cells, monocytes and mitochondrial functions. Both the heatmap and modular analysis demonstrated an active transcriptional response in certain genes in the patients with a good treatment response.

Cured pneumonia patients clustered separately from the untreated pneumonia and have profiles parallel to the healthy controls

Five pneumonia patients were followed up 6 weeks after hospital discharge. All five completed their prescribed antibiotics and were diagnosed as cured by their practising physician. All had complete or near-complete resolution of radiological changes, symptoms and CRP <5 mg/L. Unsupervised analysis (7,806) and unsupervised hierarchical clustering of just the pre-treatment and post-treatment pneumonia samples clearly illustrated the samples sub-dividing into two clusters, containing pre-treatment and post-treatment profiles (Figure 62). The 1,446 transcript list derived from the initial unsupervised analysis and statistical filtering of all the untreated training set profiles (Figure 14), were used to demonstrate by unsupervised hierarchical clustering that the same pneumonia profiles after treatment now appeared within the main cluster with the controls (Figure 63). The post-treatment pneumonia samples are pink on the sample label bar and the pre-treatment profiles obtained from the same patients are as before brown on the sample label bar (Figure 63). The training set was used as all the pneumonia pre- and post-treatment samples were processed and run for microarray with the training set samples.

These results show there was no transcriptional difference between the post-treatment pneumonia samples and the controls.

MDTH of TB, sarcoidosis and pneumonia changed significantly after treatment

MDTH has previously been shown to be associated with disease activity therefore the algorithm was applied to determine the effect of treatment on the patients' transcriptional profiles. After treatment both the TB and pneumonia MDTH scores significantly decreased such that their scores became synonymous with the healthy controls (Figure 65b & c). However the treatment responsive sarcoidosis significantly increased their MDTH, in keeping with an active transcriptional response after treatment as described earlier and indicating a likely glucocorticoid response (Figure 65a). The sarcoidosis patients who were commenced on treatment but did not respond clinically (inadequate responders) showed no significant change from the untreated sarcoidosis (Figure 65a).

Discussion

Sarcoidosis transcriptional profiles correlated strongly with the patient's clinical response to systemic glucocorticoids

To prevent any bias in the analysis of the effects of treatment on sarcoidosis patients a completely unsupervised analysis approach (therefore only with the fold change from the median and without any statistical filtering) was applied. This was followed by unsupervised hierarchical clustering using the interpretation of 'treated with immunosuppression' or 'pre-treatment'. The analysis therefore provides a totally unbiased answer as to the clustering of the patients. However it was perhaps not an anticipated finding that some patients who were started on immunosuppression should look like the patients who were not commenced on any treatment. But on closer inspection of the clinical data of the samples it can be seen the patients instead distinctly cluster into (1) those that had a good treatment response and (2) those who did not respond satisfactorily or all the pre-treatment samples (Figure 50 - 51). Therefore it seems the transcriptional profiles correlated with a good response to glucocorticoid treatment rather than just whether the patient was receiving glucocorticoids. As expected adding a further level of analysis with a statistical filter to compare the two groups, ‘pre-treatment/inadequately treated’ compared to ‘good response’, replicated the clustering performed with pure unsupervised analysis (data not shown).

Of particular interest was the patient who was thought initially not to be responding to treatment on their second clinic visit but on their third clinic visit was determined to be fully responsive - as her transcriptional profile post-treatment was always typical of a patient who was clinically responding well even at the second clinic visit. This suggests transcriptional profiling may be useful as a clinical tool in aiding decision making regarding treatment response in clinically challenging situations.

There were three patients who did not appear to respond well either clinically or by extrapolating interpretation of their transcriptional profiles. One of the three patients was only receiving high dose hydroxychloroquine. However the evidence for measurable efficacy of single therapy hydroxychloroquine in pulmonary sarcoidosis is poor (Bradley, Branley et al. 2008), which may explain the lack of a response. The other two patients were receiving prednisolone 20mg daily. There could be several reasons for their lack of response including an insufficient prednisolone dose (as demonstrated by patients 3 & 5 who responded well after increasing their treatment), poor patient compliance or possibly glucocorticoid resistant disease. Steroid-resistance in treating sarcoidosis is a common reason for instituting alternative immunosuppressive therapy such as methotrexate or azathioprine (Paramothayan, Lasserson et al. 2006), steroid-resistance is also an established phenomenon recognised in other respiratory diseases e.g. asthma (Schwartz, Lowell et al. 1968). Although this study was not designed to determine a biomarker of treatment response, one could speculate there is a pragmatic potential for transcriptional profiles to act as surrogate markers of treatment response in sarcoidosis. This would be of great clinical value as there is currently little consensus on how to comprehensively assess treatment response in pulmonary sarcoidosis.

The transcriptional response to successful treatment of sarcoidosis was associated with over-abundance of the inflammatory response

An additional finding was that the sarcoidosis treatment-responsive patients, all of whom were treated with glucocorticoids, showed an over-abundance of many of the genes, particularly the inflammatory genes (Figure 58 & 59). The sarcoidosis patients with a good treatment response remained in the main cluster containing all the patients rather than the main cluster containing the controls (Figure 60), although they clustered distinctly from untreated patients within the main patient cluster. It might be expected that a reduction in the inflammatory response would occur in patients receiving glucocorticoids; however our study appeared to show the opposite. There are a number of reasons this may have occurred. Firstly most sarcoidosis patients are treated with curative intent for many months with immunosuppressive therapy where over 50% of patients are treated for longer than 2 years (Baughman and Nunes 2012). The average length of time the patients in this cohort were sampled at was 14 weeks after treatment was commenced (Figure 59). Therefore the patients were only sampled in the middle of their treatment regimen, which may explain their on-going active transcriptional response trying to reverse the underlying disease processes. In addition this may explain the lack of significant down-regulation of the IFN modules in all the patients (Figure 61), although from the heatmap it can be seen some of the treatment-responsive patients did show an under-abundance of many IFN-inducible genes, e.g. STAT1, STAT2, many

GBPs and IRF1 (data not shown), compared to the pre-treatment patients (Figure 60). Perhaps if these patients were sampled again in 1-2 years' time their transcriptional profiles would appear much more quiescent, resembling the non-active sarcoidosis profiles. Interestingly it has been shown that glucocorticoids while able to suppress the NF-κB pathway in many cells in SLE exert no effect on the NF-κB pathway in pDCs, thus allowing the continued secretion of IFNα and intimating the reason for the reduced glucocorticoid sensitivity seen in SLE (Guiducci, Gong et al. 2010). Indeed glucocorticoid resistance has been reported in several other diseases including asthma, rheumatoid arthritis, acute lymphocytic leukaemia and ulcerative colitis (Biddie, Conway-Campbell et al. 2012). The underlying mechanisms resulting in resistance in these diseases may therefore also explain the seemingly partial or negligible response seen in many sarcoidosis patients (Paramothayan, Lasserson et al. 2006). Although in this study there was over-abundance of the inflammation modules in the treatment-responsive patients compared to the untreated/inadequate responders, some of the over-abundant genes were anti-inflammatory genes including IL1R2, IL1RAP, IL18RAP, DUSP1, FOS, IκBα and MAPK1 (fold change >2, Figure 61 or data not shown) (McColl, Michlewska et al. 2007; Shipp, Lee et al. 2010; Veenbergen, Smeets et al. 2010). Therefore although there is an over-abundance of inflammation genes as defined by the inflammation modules, many of these inflammatory genes may be involved in anti-inflammatory processes. In addition some of the inflammation genes over-abundant in the module are members of the TNF super family (TNFRSF10B, TNFRSF10C, TNFRSF1A and TNFSF13B) with known roles in apoptosis. Furthermore the glucocorticoids may be applying their anti-inflammatory effects at the post-transcriptional level, as has been described in previous studies (De Bosscher, Vanden Berghe et al. 2003).

Modular analysis of the sarcoidosis transcriptional response to glucocorticoids was similar to that seen in glucocorticoid-treated SLE patients

Many of the other glucocorticoid effects seen by modular analysis in this study were also seen by the same modular analysis in the Guiducci *et al* study of glucocorticoids in SLE patients (Guiducci, Gong et al. 2010). Although they did not publish all the modules, including the inflammation modules and apoptosis module, they found similar differences between their treated and untreated patients as in this study for many of the other modules. These similarities included no change in the IFN module (suggested to be related to the glucocorticoid-unresponsive pDCs in SLE), under-abundance of the B cells, monocytes and mitochondrial related modules (Figure 61). Another module not shown in their paper but shown to be over-abundant in this study was the DC/apoptosis module (Figure 53). This is in keeping with the knowledge that glucocorticoids induce apoptosis in many cells including neutrophils, eosinophils, thymocytes and pDCs (Boor, Metselaar et al. 2006; McColl, Michlewska et al. 2007).

Unlike this study an earlier sarcoidosis study surprisingly found no difference in the transcriptome of patients either receiving or not receiving systemic glucocorticoids

Although there was an unexpected significant inflammatory response and transcriptional activity in the treatment-responsive patients, the other glucocorticoid-induced responses were in keeping with previous studies, possibly suggesting a unique glucocorticoid-related inflammatory process occurs within the peripheral blood of sarcoidosis patients that has not been seen in other inflammatory diseases. Only one other sarcoidosis blood transcriptome study compared patients receiving systemic glucocorticoids to those not receiving treatment (Koth, Solberg et al. 2011). This study reported no difference in the treated or untreated patient’s transcriptional response. In view of our knowledge of the action of glucocorticoids and their effects this would not have been anticipated and could be explained by at least three factors. Firstly they do not document in their publication the dose of prednisolone the patients were taking and as sarcoidosis patients are often maintained on very low doses e.g. 5mg prednisolone, without this information it is difficult to interpret the data. Secondly the patients were two separate groups – not the same patients assessed before and after treatment, therefore the treated group may have originally had more active disease requiring treatment which on receiving treatment had become more similar to those untreated patients. Thirdly it is possible all 12 of the treated patients in their cohort were inadequately responding to the glucocorticoids, especially as treatment response can be difficult to assess, therefore these patients showing no difference in their profiles would match the inadequate responders in our study.

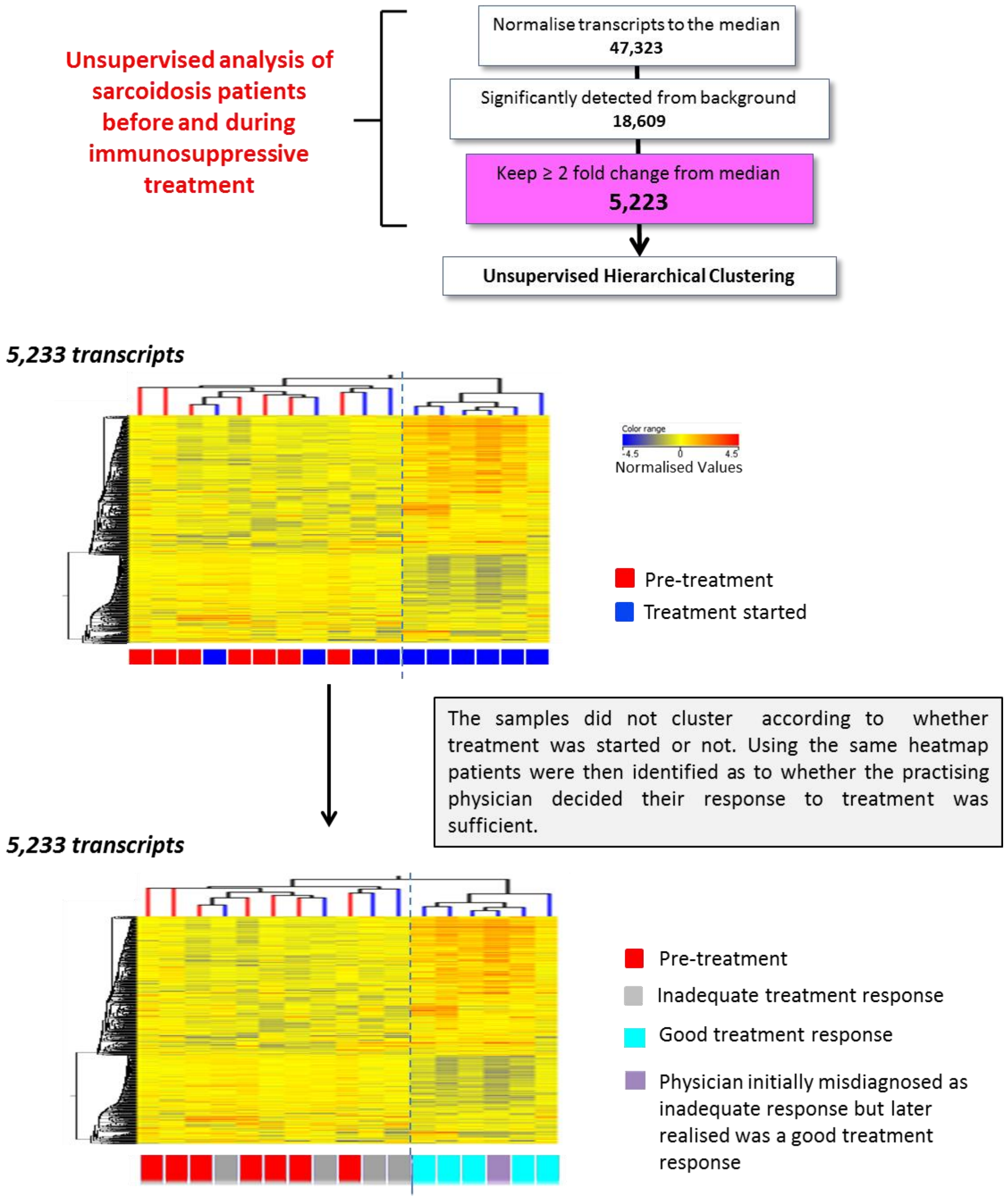
The transcriptional response of successfully treated pneumonia and pulmonary TB patients resemble healthy controls

The transcriptional profiles of the five cured pneumonia patients all returned to a transcriptional pattern identical to that of the controls. This could be seen on the heatmap (Figure 63), by the functional modular analysis (Figure 64) and by their MDTH score (Figure 65). As described in chapter 7 the cured TB patients also resembled the controls (Figures 49, 50 & 57). The increase in transcriptional activity of the treatment-responsive sarcoidosis patients was further illuminated by comparing the MDTH of all three diseases, before and during/after treatment (Figure 65). The treatment-responsive sarcoidosis patients showed a significant rise in transcriptional activity contrasting the significant fall in transcriptional activity seen in the cured TB and pneumonia patients. Interestingly the MDTH scores of the sarcoidosis inadequate-responders paralleled the clustering visualised in the heatmaps, in that there was no difference between them and the untreated sarcoidosis profiles. Two of the sarcoidosis patients only displayed a change in their transcriptional response after increasing their prednisolone dose (Figure 59). Thus you could postulate the change in transcriptional activity is either due to a glucocorticoid dose response or due to glucocorticoid-induced effects on the sarcoidosis immunopathology rather than a direct effect of glucocorticoids.

Chapter Summary

Sarcoidosis patients receiving immunosuppressive treatment showed a distinct change in their transcriptional signature only if they achieved a good clinical response to treatment. This change appeared to be an active transcriptional response with a prevailing over-abundance of inflammatory genes which included many anti-inflammatory genes, over-abundance of apoptosis genes and under-abundance of certain leukocyte cell types. This response contrasted with the transcriptional changes seen after curative antibiotic treatment of pulmonary TB patients and community acquired pneumonia patients, as these patient’s profiles were comparable with the controls after successful completion of treatment.

Figures for chapter 8

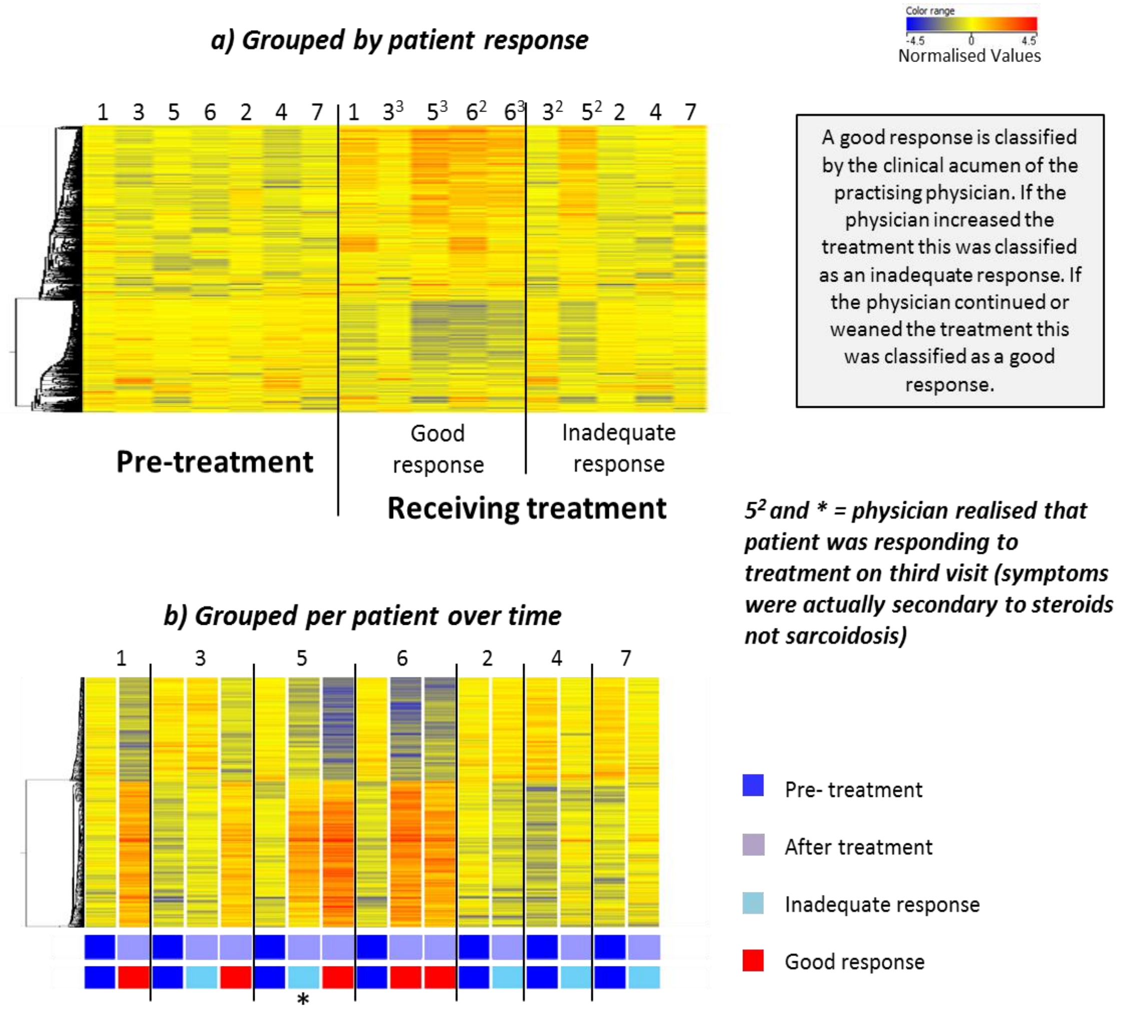


The samples did not cluster according to whether treatment was started or not. Using the same heatmap patients were then identified as to whether the practising physician decided their response to treatment was sufficient.

Figure 58. Unsupervised analysis and clustering revealed that treated sarcoidosis patients show a change in transcriptional signature only if they had a good clinical response to treatment.

Transcripts were derived as shown at the top of the figure. The same unsupervised clustered heatmap of the 5,233 transcripts is displayed twice. The top heatmap reveals whether the samples were from patients before or after treatment. The bottom heatmap reveals whether the samples were from patients who had clinically responded to treatment or not. The layout of the heatmaps is as described in figure 13.

ID	Weeks	Treatment before second profile	Treatment Increased on second visit
1	16	Prednisolone 40mg	No
2	6	Hydroxychloroquine 800mg	Yes
3^2	15	Mycophenolate 1000mg + Pred 5mg	Yes
3^3	52	Mycophenolate 1500mg, methylprednisolone 3 pulses 250mg IV + Pred 15mg	No
4	9	Prednisolone 20mg	Yes
5^2	3	Prednisolone 25mg	Yes
5^3	15	Prednisolone 30mg	No
6^2	3	Prednisolone 25mg	No
6^3	16	Prednisolone 20mg	No
7	9	Prednisolone 20mg	Yes



A good response is classified by the clinical acumen of the practising physician. If the physician increased the treatment this was classified as an inadequate response. If the physician continued or weaned the treatment this was classified as a good response.

*5^2 and * = physician realised that patient was responding to treatment on third visit (symptoms were actually secondary to steroids not sarcoidosis)*

Figure 59. Showing same 5,233 transcripts but patients now clustered (a) by their treatment response and (b) per patient.

The 5,233 transcripts used in the heatmaps were derived as shown in figure 58. The heatmaps were grouped by (a) the patient response or (b) each individual patient. Each patient is denoted a number in the table and top heatmap. A superscript number indicates the clinical visit number if there was more than one visit e.g. 3^2 indicates the sample was taken on their second clinic visit and 3^3 indicates the sample was taken on the third clinic visit.

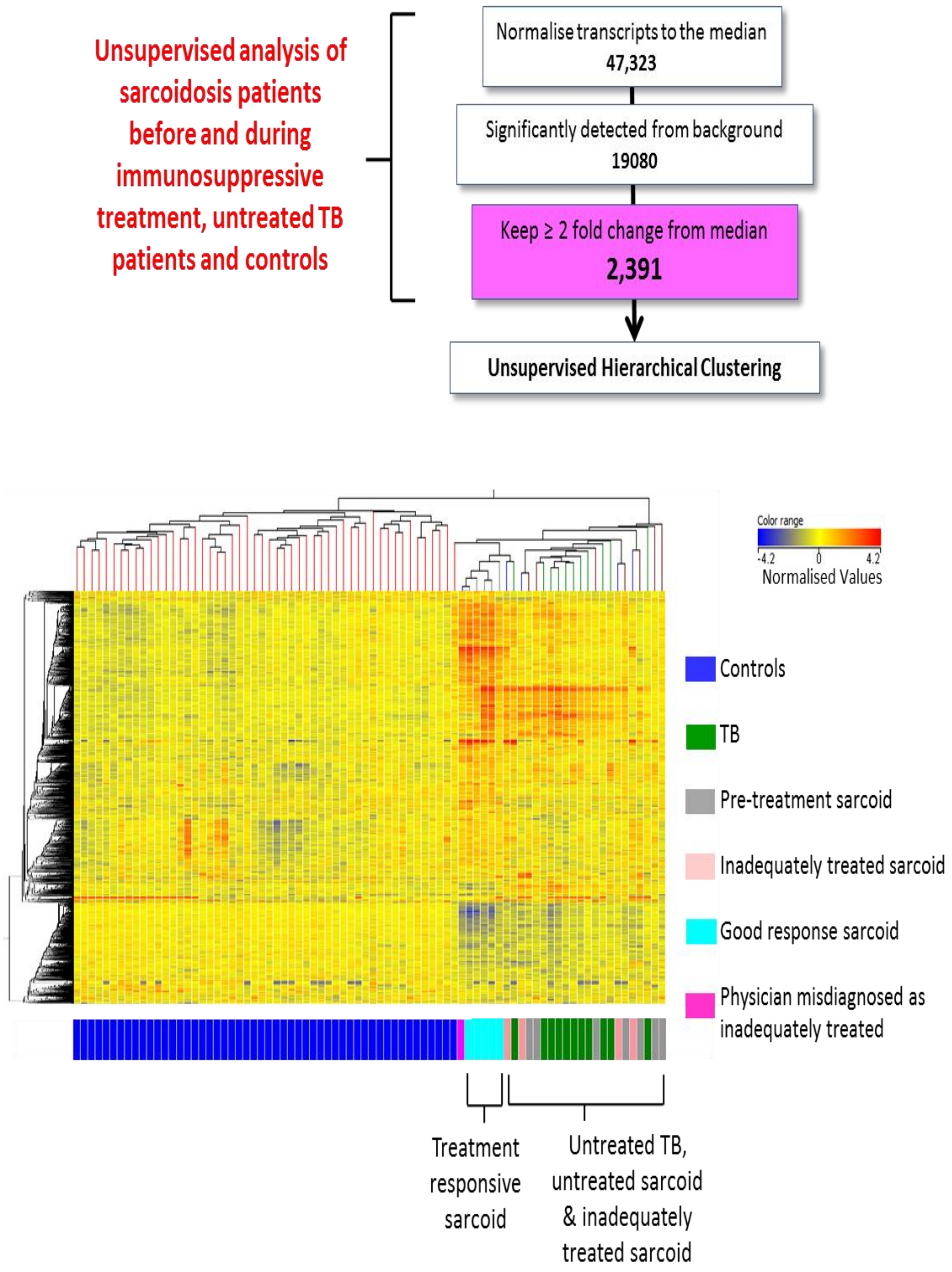


Figure 60. Sarcoidosis patients responding adequately to treatment clustered separately from the untreated sarcoidosis & TB patients.

Transcripts were derived as shown at the top of the figure. The layout of the heatmap is as described in figure 13.

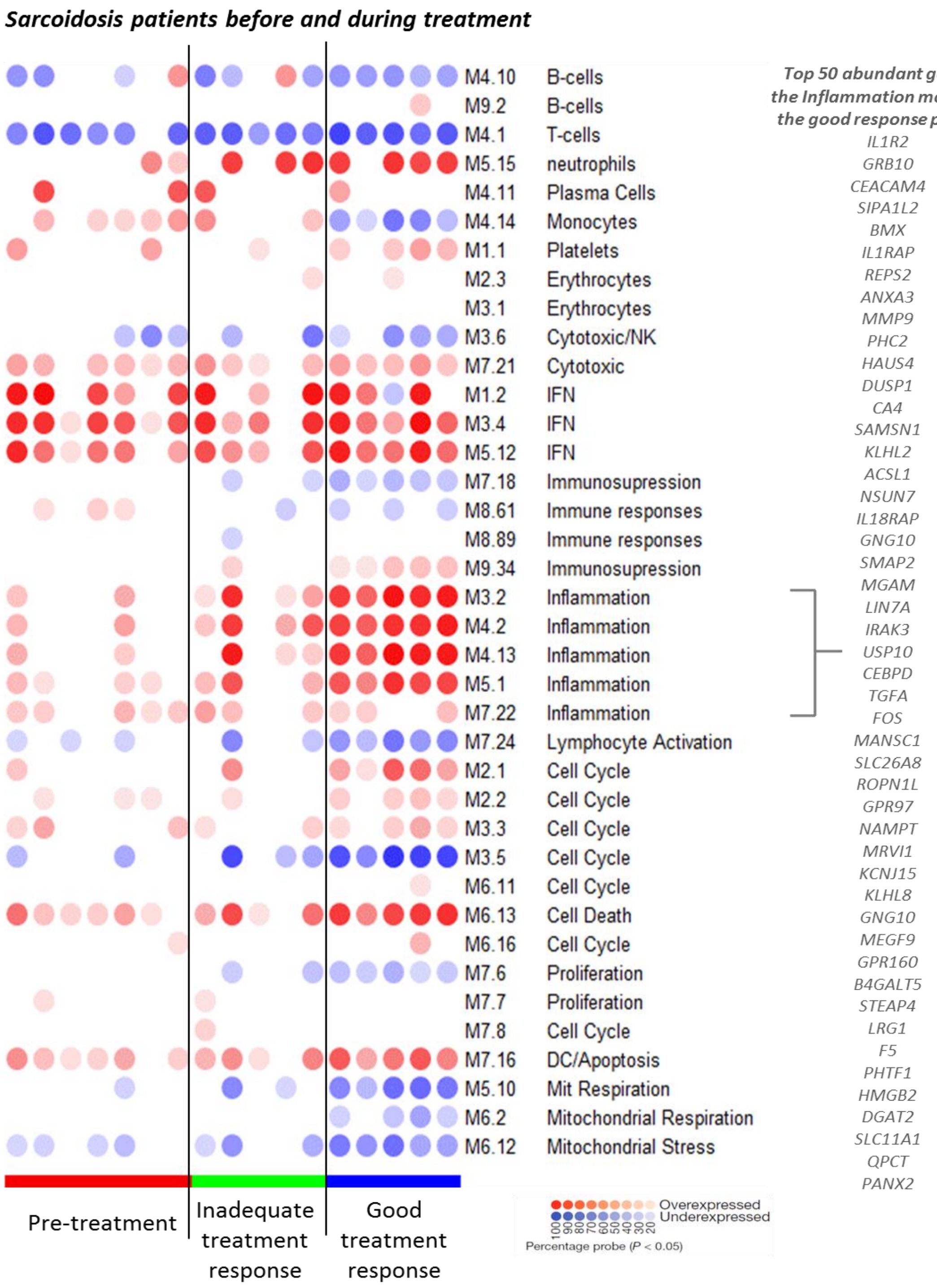
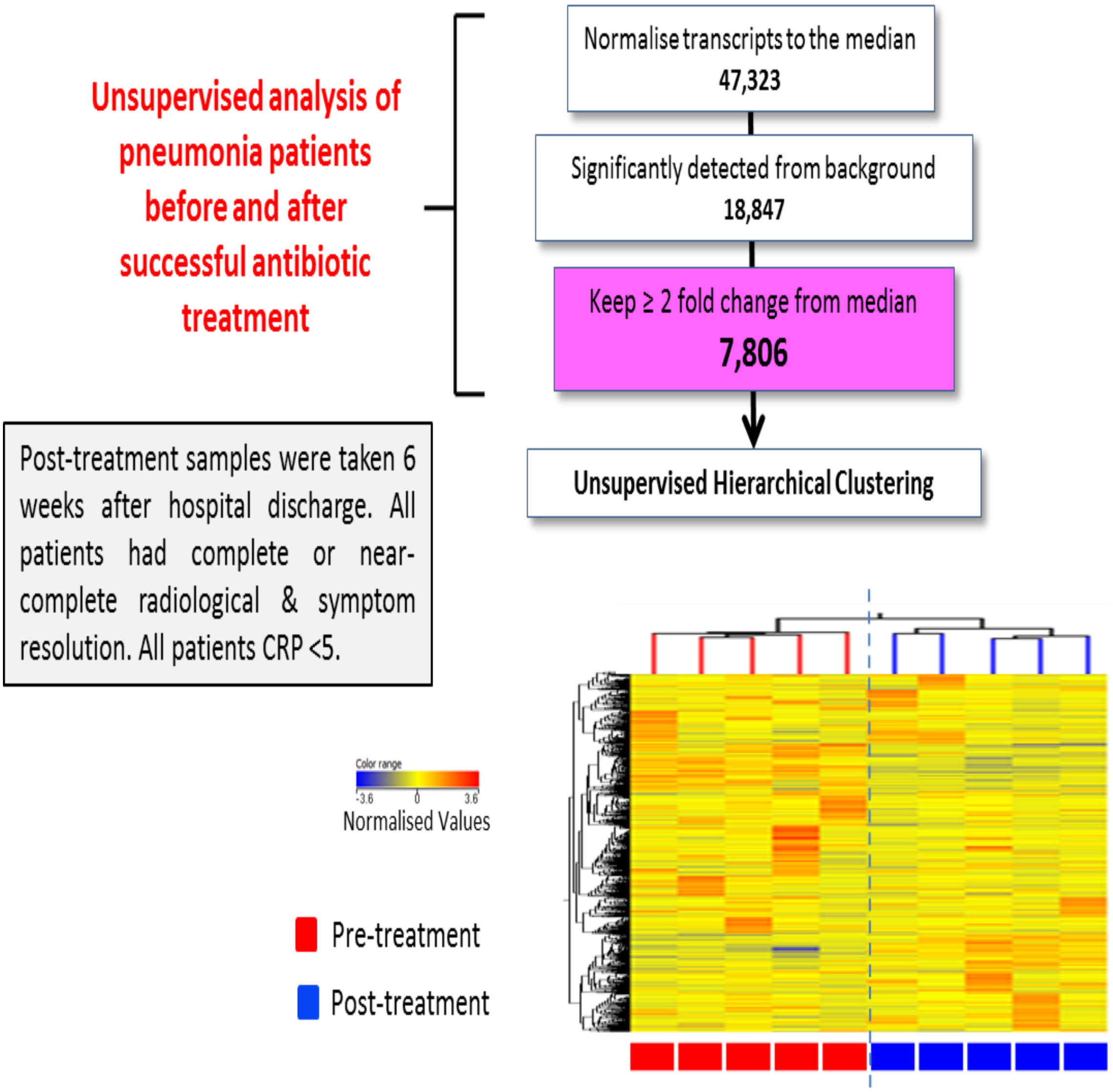


Figure 61. A good response to sarcoidosis treatment appears to be associated with inflammatory genes.

Only transcripts that were significantly detected compared to the background intensity (>15,000) were applied to the modular analysis (p<0.01). Modules with a red dot contain genes that were significantly over-expressed in that patient compared to the controls, a blue dot represents genes that were significantly under-expressed, no dot indicates no significant change in expression compared to the controls (p<0.05). The shade of the colour indicates the percentage of genes in that module that are significantly expressed as shown by the colour legend at the bottom of the figure.



Unsupervised analysis of pneumonia patients before and after successful antibiotic treatment

Normalise transcripts to the median 47,323

Significantly detected from background 18,847

Keep ≥ 2 fold change from median 7,806

Unsupervised Hierarchical Clustering

Post-treatment samples were taken 6 weeks after hospital discharge. All patients had complete or near-complete radiological & symptom resolution. All patients CRP <5.

Figure 62. Unsupervised analysis and clustering demonstrated that successfully treated pneumonia profiles cluster separately from untreated pneumonia profiles.

Transcripts were derived by unsupervised analysis as shown at the top of the figure from the six patients' before and after treatment. Unsupervised hierarchical clustering of the 7,806 transcripts was then performed. The layout of the heatmap is as described in figure 13.

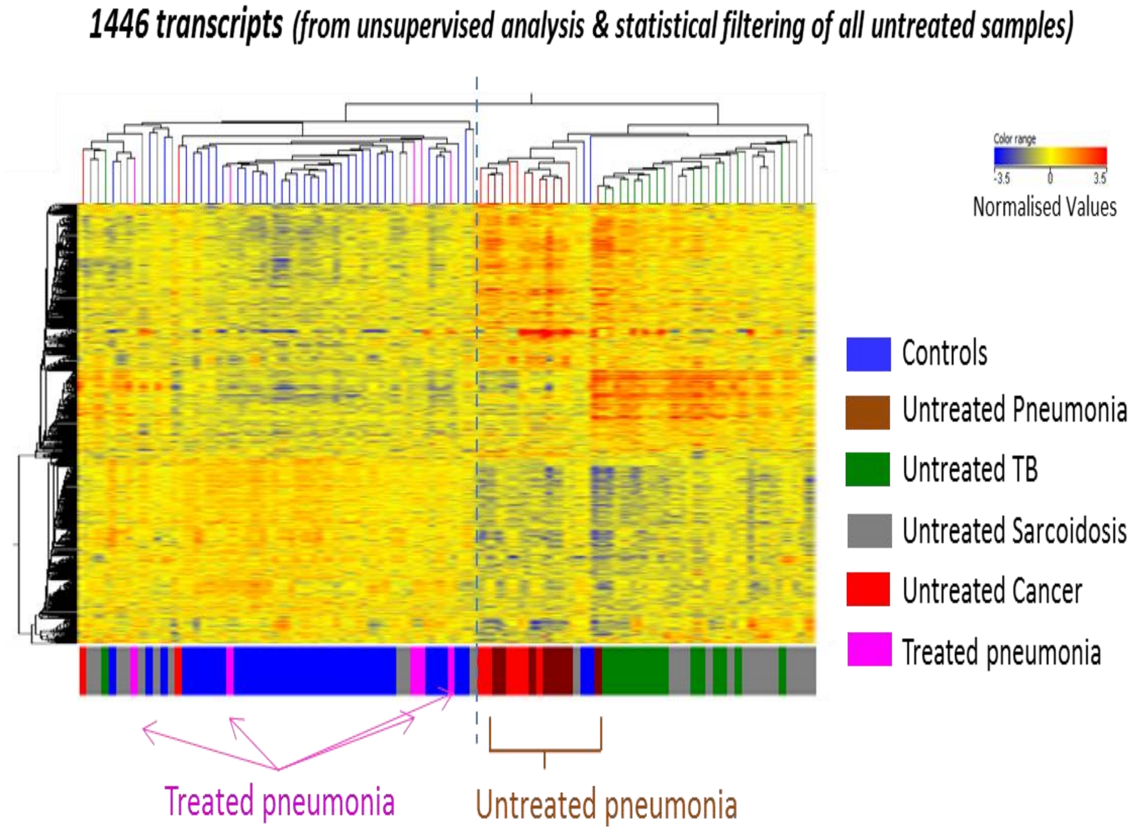


Figure 63. After successful treatment pneumonia profiles cluster with the controls.

The 1,466 transcripts that were derived originally from all the untreated trainings set samples by unsupervised analysis and statistical analysis (see figure 14) were applied again to the same cohort but with the addition of the six post-treatment pneumonia samples. Unsupervised analysis was then performed using the 1,446 transcripts and all the untreated samples as well as the six treated pneumonia samples. The layout of the heatmap is as described in figure 13.

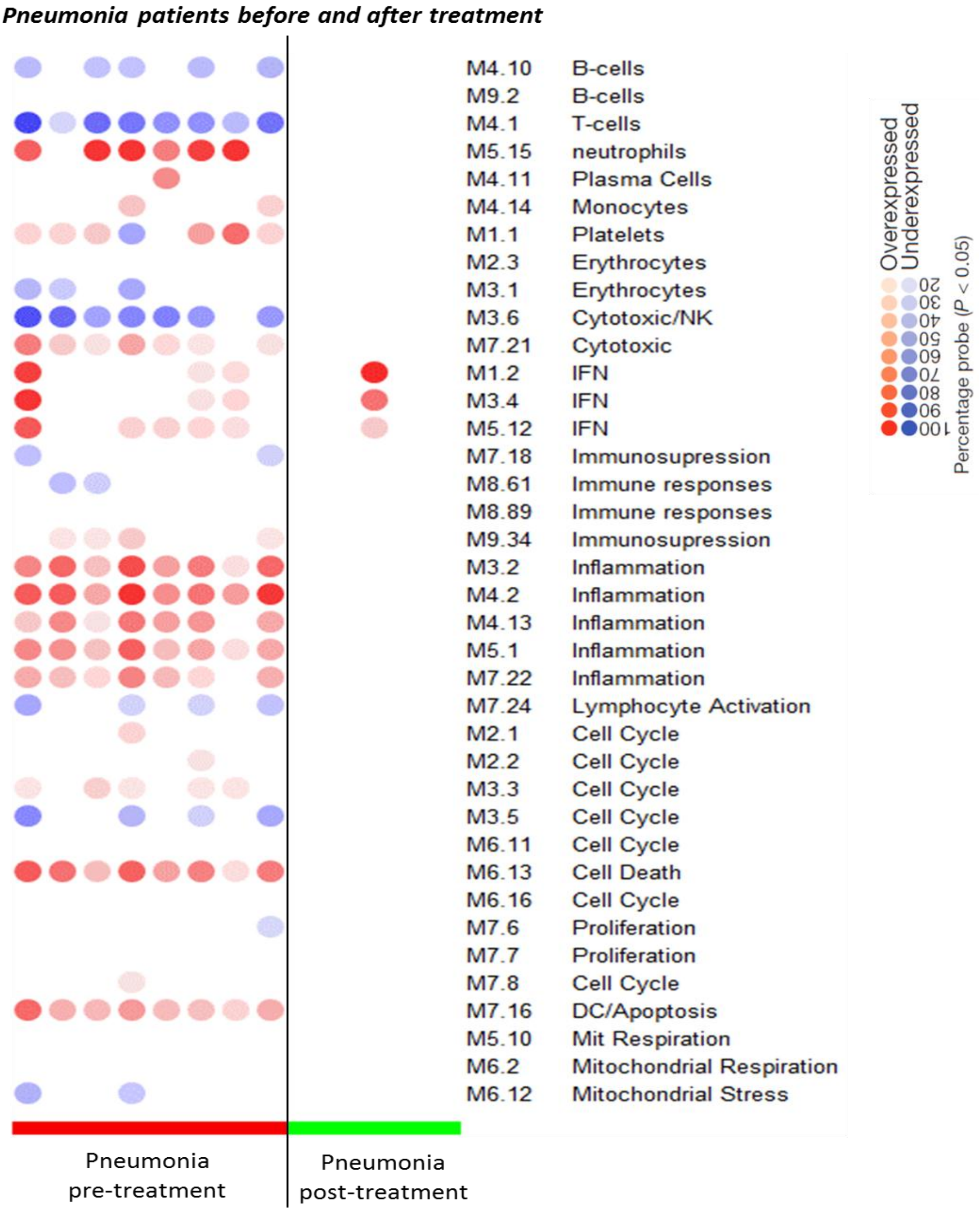


Figure 64. Cured pneumonia patients showed a significant change in the modular analysis towards the controls.

Only transcripts that were significantly detected compared to the background intensity (>15,000) were applied to the modular analysis (p<0.01). Modules with a red dot contain genes that were significantly over-expressed in that patient compared to the controls, a blue dot represents genes that were significantly under-expressed, no dot indicates no significant change in expression compared to the controls (p<0.05). The shade of the colour indicates the percentage of genes in that module that are significantly expressed as shown by the colour legend at the bottom of the figure.

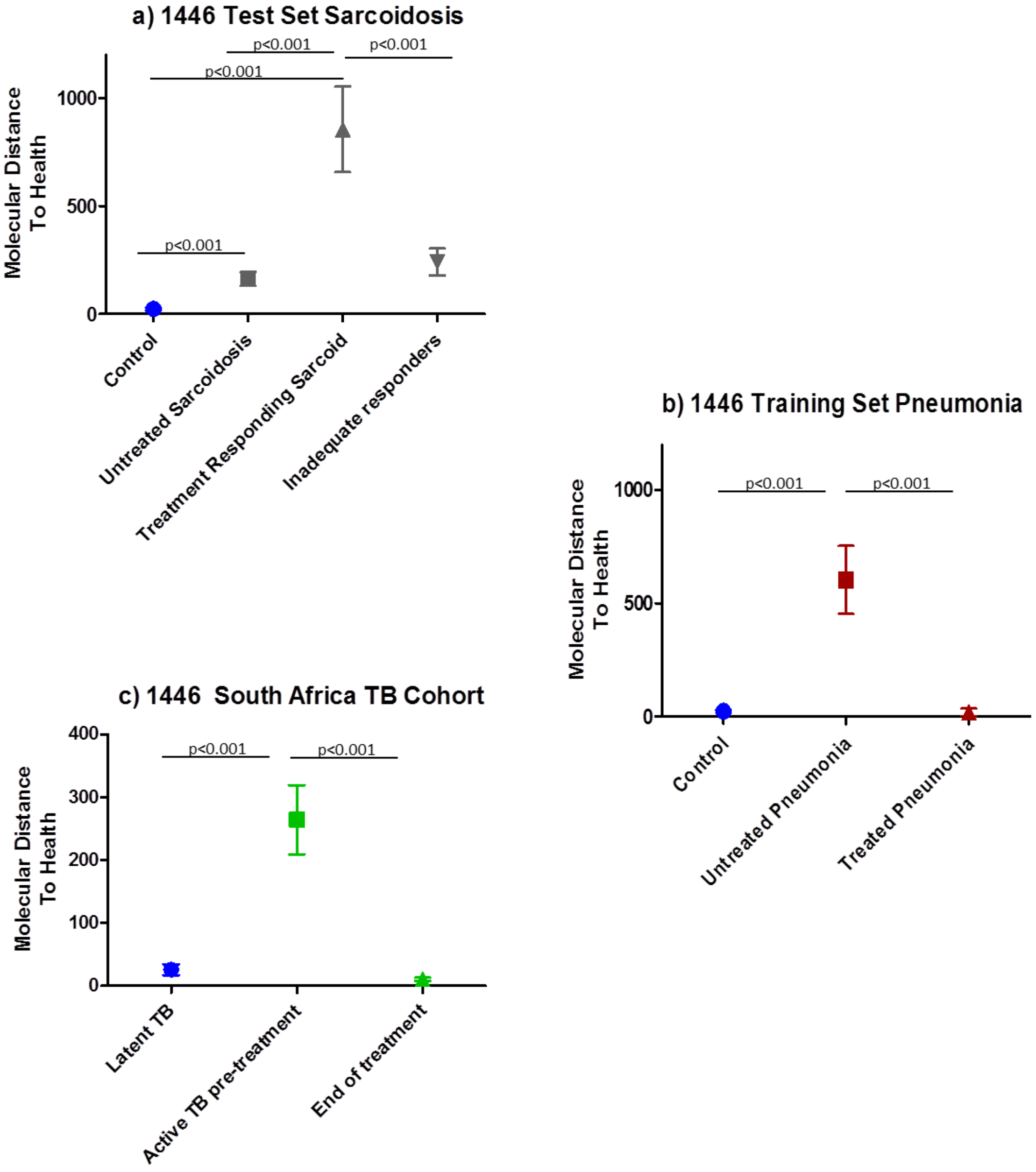


Figure 65. Comparing MDTH in each disease before and after treatment using the same 1446 transcripts.

MDTH algorithm was applied to each disease group to assess disease activity in the untreated samples compared to the post-treatment samples. The same transcript list was applied to all the disease groups for a fair comparison. The 1,446 transcript list originally derived from all the untreated samples was used (see figure 14). The graphs displays mean, SEM and *p* values from ANOVA with Tukey's multiple comparison test.

Chapter 9

Summary and future perspectives

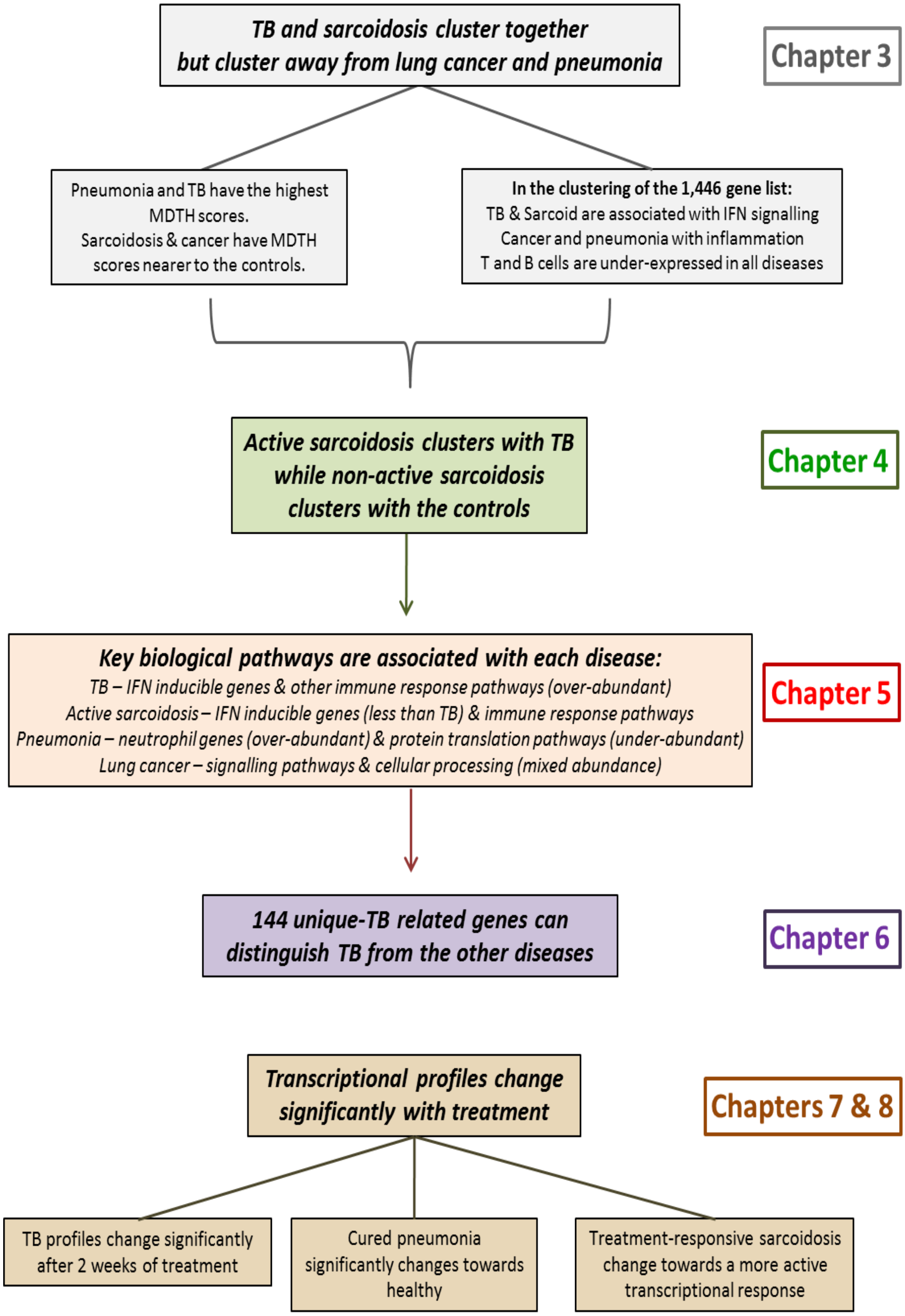

TB and sarcoidosis cluster together but cluster away from lung cancer and pneumonia
Chapter 3
Pneumonia and TB have the highest MDTH scores. Sarcoidosis & cancer have MDTH scores nearer to the controls.
In the clustering of the 1,446 gene list:
TB & Sarcoid are associated with IFN signalling
Cancer and pneumonia with inflammation
T and B cells are under-expressed in all diseases
Active sarcoidosis clusters with TB while non-active sarcoidosis clusters with the controls
Chapter 4
Key biological pathways are associated with each disease:
TB – IFN inducible genes & other immune response pathways (over-abundant)
Active sarcoidosis – IFN inducible genes (less than TB) & immune response pathways
Pneumonia – neutrophil genes (over-abundant) & protein translation pathways (under-abundant)
Lung cancer – signalling pathways & cellular processing (mixed abundance)
Chapter 5
144 unique-TB related genes can distinguish TB from the other diseases
Chapter 6
Transcriptional profiles change significantly with treatment
Chapters 7 & 8
TB profiles change significantly after 2 weeks of treatment
Cured pneumonia significantly changes towards healthy
Treatment-responsive sarcoidosis change towards a more active transcriptional response

Figure 66. Flow diagram of the results chapters summarising the main findings.

Chapter 9: Summary and future perspectives

Summary of the results

To our knowledge this is the first study to compare the blood transcriptional profiles of patients with pulmonary TB and the other similar respiratory diseases pulmonary sarcoidosis, community acquired pneumonia and primary lung cancer (Figure 66). The two clinically and pathologically analogous granulomatous diseases TB and sarcoidosis had very similar but non-identical molecular and biological characteristics that were distinct from the molecular and functional characteristics of the clinically similar respiratory diseases pneumonia and cancer. However it was possible to identify a unique set of TB-related genes that could differentiate TB from all the other profiles. The TB, pneumonia and sarcoidosis patients showed a significant transcriptional response after receiving potentially curative treatment. Their response varied depending on the disease.

Clinically similar diseases TB and sarcoidosis had comparable blood transcriptional signatures, distinct from pneumonia and lung cancer

Unsupervised analysis and statistical filtering generated 1,446 differentially expressed transcripts across all the training set samples (Figure 14). Samples clustered into controls and patients, with TB and sarcoidosis profiles clustering distinctly from the cancer and pneumonia profiles. These clustering configurations were validated in a test set and were independent of ethnicity and gender (Figure 15, 16 & 18). The diseases TB and pneumonia showed the highest transcriptional activity, as evidenced by their MDTH scores in both the training and test set, while sarcoidosis and cancer appeared to be more transcriptionally quiescent diseases with scores approaching the controls (Figure 19). Functional analysis by IPA of the 1,446 transcripts found an association between the IFN-signalling pathway and other immune response pathways with both TB and sarcoidosis; whereas pneumonia and cancer had significant associations with the inflammation and signalling pathways (Figure 20). All the diseases correlated with an under-abundance of T and B cell pathways. Two prior studies have also found TB and sarcoidosis transcriptional profiles to be very similar and to correlate with over-abundance of IFN-inducible genes (Koth, Solberg et al. 2011; Maertzdorf, Weiner et al. 2012).

Sarcoidosis transcriptional profiles were heterogeneous in the training, test and validation set and appeared to form at least two subgroups, those clustering with controls and those clustering with the other patients, particularly the TB patients' profiles (Figures 25 & 26). A complex clinical classification system was applied to divide sarcoidosis patients into either having active disease or non-active disease. This classification showed significant clustering-prediction abilities, more so than any single clinical variable, or combinations of the twenty-five different variables (Table 19–20). There were three previous sarcoidosis blood transcriptional profiling papers, however none of these papers commented on the known clinical heterogeneity of sarcoidosis or on the relationship of acknowledged sarcoidosis clinical variables with their expression profiles. This may be because two of the three papers did not apply unsupervised analysis or unsupervised hierarchical clustering to visualise their transcriptional profiles instead narrowing down their analysis to genes of interest for the investigators, thus their approach was more biased and led more by their pre-considered hypotheses (Rosenbaum, Pasadhika et al. 2009; Koth, Solberg et al. 2011; Maertzdorf, Weiner et al. 2012). However Rosenbaum *et al* did apply supervised hierarchical clustering to their 564 transcripts differentially expressed of the sarcoidosis patients compared to the controls (fold difference ≥ two and q ≤0.05), from which it can be clearly visualised some of the sarcoidosis patients do have profiles corresponding to the controls. This observation suggests that their cohort also had heterogeneous transcriptional profiles, although this was not acknowledged in their publication.

Functional & biological characteristics were associated with each of the four active diseases

Modular analysis displayed a distinct pattern of significant over-abundance of the IFN modules for the TB and active sarcoidosis patients and a minor over-abundance in the IFN modules in the non-active sarcoidosis (Figure 28). However the percentage of IFN genes over-abundant in the TB patients was significantly more than in the active sarcoidosis patients and therefore all sarcoidosis patients (Figure 29). By IPA comparison analysis the significance and the number of genes present in the IFN-signalling pathway were both higher in the TB than the active sarcoidosis patients (Figures 32 & 33). However, the two previous papers that demonstrated an association with both TB and sarcoidosis and the IFN-inducible genes, did not appear to have assessed the differences between the number of genes in the two diseases (Koth, Solberg et al. 2011; Maertzdorf, Weiner et al. 2012). In addition it can be seen from the top 50 ranked genes that the IFN-inducible genes over-abundant in the TB patients have a much higher fold difference from the controls than when over-abundant in the active sarcoidosis patients (Figure 30), and in the unique disease related genes determined from the Venn diagram it is only in the TB-related genes that the IFN-signalling pathway was significant (Figure 39). Other pathways associated with both granulomatous diseases were immune response pathways such as the role of PRRs in recognition of bacteria and viruses and the antigen presentation pathway (Figure 32).

The pneumonia and cancer patients were associated with very different biological sets of genes, in particular inflammation genes, as seen in the modular analysis (Figures 28 & 29). These global biological findings are in keeping with the known underlying disease mechanisms acknowledged to occur in both pneumonia and primary lung cancer (Fernandez-Serrano, Dorca et al. 2003; O'Callaghan, O'Donnell et al. 2010). However inflammation is a very broad category encompassing a large number of genes, therefore the degree of similarity of the inflammatory response between these two diseases needs to be further defined. Pneumonia was also significantly correlated with over-abundance of the neutrophil module, while many of the top 50 ranked differentially expressed genes compared to the controls were also neutrophil anti-microbial genes (Figures 28-30), matching the high peripheral neutrophil blood count (Table 10).

‘Comparison IPA’ analysis revealed a very significant association of pneumonia with the under-abundance of protein translation pathways EIF2, and mTOR signalling (Figures 32, 33 & Table 25). This finding was also seen by investigating the unique disease-related genes acquired from the Venn diagram (Figure 39). Furthermore targeted analysis of protein translation and unfolded protein response genes also identified an under-abundance, relative to the controls and relative to the other diseases (Figures 35, 36 & Table 27), along with a dominant association of over-abundance of the IPA apoptosis pathway (Table 25). Past reports have noted an up-regulation of UPR related genes in two different bacterial infections, in cells infected with *Listeria monocytogenes* and in macrophages from human TB granulomas (Seimon, Kim et al. 2010; Pillich, Loose et al. 2012), which is in contrast to the under-abundance of UPR genes found in our study. However our study has focussed only on the peripheral blood away from the site of the disease. Therefore a potential reason for this disparity in regulation could be due to preservation of cellular energy at sites away from the source of infection/inflammation (i.e. peripheral blood) while there is a preferential sequestration of protein making cells at the site of infection/inflammation (i.e. the lungs).

‘Comparison IPA’ also revealed three pathways as more significantly associated with cancer than the other diseases: NK cell signalling, CTLA4 signalling in cytotoxic T lymphocytes and HGF signalling (Table 26). Interestingly all three pathways have been involved in the therapeutic management against cancer. NK cell-based immunotherapies have had little success in humans but remain promising due to the suggested ability of NK cellls to migrate towards inflammation and kill target cells without previous activation (Zamai, Ponti et al. 2007). Anti-CTLA4 antibodies have already been approved for use in patients with metastatic or unresectable melanoma and is in clinical trials for prostate cancer (Kwek, Cha et al. 2012). In lung cancer HGF-MET inhibitors have shown good efficacy in Phase III trials and in patients with resistance to EGFR-therapy (Gherardi, Birchmeier et al. 2012).

The set of 144 transcripts could differentiate TB patients from the other diseases

Although this study was not designed to obtain a biomarker, the 144 transcripts (132 genes) that were differentially expressed between TB and active sarcoidosis, were also able to distinguish TB from all other respiratory diseases with good accuracy (Figure 44). This was achieved in the training set, test set, validation set and external datasets, demonstrating the robustness of the transcript list. The 144-transcript list also revealed better sensitivity than two previously published transcript lists that could differentiate TB from sarcoidosis (Koth, Solberg et al. 2011; Maertzdorf, Weiner et al. 2012).

Transcriptional profiles changed significantly during and after treatment

Active TB patients showed a significant change during and after antituberculous treatment in a derived active TB transcriptional signature, a derived treatment specific signature and the 144 TB-specific transcriptional signature (Figures 50, 51 & 57). The transcriptional change occurred as early as 2 weeks after treatment was initiated. The transcriptional response could be quantified in each patient using the novel algorithm 'temporal molecular response' (Figures 53-55). Post treatment signatures showed insignificant changes compared to the latent TB controls and revealed comparable MDTH scores to the latent TB controls (Figure 65c). A similar significant reduction in the transcriptional response towards the healthy controls could also be observed in pneumonia patients after administration of a curative course of antibiotics (Figures 62, 63 & 65b). However while both the TB and pneumonia patients' transcriptional response returned to a state equivalent to the controls, sarcoidosis patients treated with glucocorticoids displayed a very different transcriptional response to the healthy controls. Interestingly transcriptional changes could only be visualised in sarcoidosis patients who had a good clinical response to treatment, not just all patients who were receiving glucocorticoids (Figures 58-60). Interestingly the expression profiles of patients who did respond clinically to glucocorticoids showed an increase in transcriptional activity. This increase was dominated by inflammation genes, of which at least some are recognised to have anti-inflammatory roles. Other changes in the gene expression profiles secondary to the glucocorticoids were in keeping with previously published findings, such as under-abundance of T cells and monocytes but over-abundance of apoptosis genes, and little change in the IFN modules (Figure 61) (Guiducci, Gong et al. 2010).

Future perspectives

Compare the expression profiles found in the whole blood to those distinct for each of the different cell types

A change in the blood transcriptional response could reflect changes in all cells or changes in gene expression only in discrete cell populations. The cell populations of the whole blood were different for each disease (Tables 9-11) and different cell types played a significant role in the transcriptional signatures as evidenced particularly by the modular analysis (Figure 28). To establish the influence each cell type and their total number were having on the signature would be best determined by comparing across the expression profiles of each cell population. For example the percentage of genes in the neutrophil module for all patients significantly correlated with their peripheral neutrophil count, yet from Berry *et al.* 2010 study and other studies, we known the neutrophil plays an important role in patients with TB and should not be dismissed (Barnes, Leedom et al. 1988; Berry, Graham et al. 2010; Eum, Kong et al. 2010). It would be particularly exciting to explore the differences and similarities between the different cell populations in the TB and sarcoidosis patients, in view of their similar whole blood transcriptional profiles. Are the IFN-inducible genes as prominent in the neutrophils in sarcoidosis as they are in TB or are they more dominant in the lymphocytes? There are several methodologies available for trying to identify the contributions of different cell types to the total expression, such as statistical-based deconvolution methods or the use of a meta-analysis of expression profiles from different cell types (Shen-Orr, Tibshirani et al. 2010; Nakaya, Wrammert et al. 2011). However none of these methods can truly replicate the robustness of the data from processing each of the individual cell types separately. Whole blood of six TB patients, ten sarcoidosis patients and seven healthy controls was separated into the different cell compositions: PBMCs, neutrophils, CD4+, CD8+, CD14+ and B cells. Unfortunately due to a technical delay the gene expression data was not available at the time of writing this thesis but should be available for analysis very soon.

Compare the blood gene expression data to protein

In parallel to the gene expression profiles, serum was collected from the patients and will be analysed using a multiplex panel of cytokines. These results will be available soon and interesting to compare to the gene expression data for each disease, to compared across the diseases, and compare before and after treatment. For example the mRNA data may not reflect the protein data due to post-transcriptional and post-translational regulations or due to other biological factors such as the rate of degradability of the measured proteins. However this additional knowledge should further help build a picture of the underlying disease mechanisms that are the same or differ between the diseases.

Blood transcriptional profiles as potential biomarkers

This study has demonstrated the proof-of-principle that blood transcriptional profiles can act as surrogate markers for disease diagnosis and treatment monitoring. This study has alluded to the possibility of using blood transcriptional signatures as biomarkers for differentiating active TB patients from patients with other clinically similar respiratory diseases, differentiating different clinical phenotypes of sarcoidosis to help guide clinical management, identifying a successful treatment response to antituberculosis treatment earlier than any currently available tool and differentiating a successful sarcoidosis treatment response from an inadequate response. The use of a commercially available whole genome microarray platform together with broadly available bioinformatics analyses programmes should allow rapid validation in subsequent larger studies. Subsequent studies could also include additional cohorts such as extending the number of respiratory diseases or including other non-respiratory diseases such as extra-pulmonary TB and extra-thoracic sarcoidosis. For the TB treatment monitoring it would be vital to include a cohort of patients with MDR-TB and also HIV/TB co-infected cohorts. This study focussed on TB patients who are not co-infected with HIV, as they represent the majority of patients infected with *M. tuberculosis*. WHO 2010 reports that of the 1.4 million deaths, three-quarters were not known to be co-infected with HIV (WHO 2010). Blood transcriptional signatures for use in the management of TB have great potential for development as blood biomarkers for clinical use and could be measured in the field using a polymerase chain reaction assay, similar to the WHO endorsed GeneXpert MTB/RIF test already in use for TB diagnostics in both developing and developed countries. However a blood host biomarker, based on our transcriptional signature, would have advantages over the GeneXpert test since it would not require sputum. In sarcoidosis treatment studies besides a much larger number of patients, it would also be of benefit to have subgroups of patients receiving different immunosuppressive treatments and crucial to follow-up their response at the end of treatment and many years subsequent due to the relapse rates.

Development of biomarkers would not only require validation of the findings in larger and more diverse cohorts but would also require biostatistics analysis tools to optimize the minimum number of genes required. The optimal gene signature would likewise need to be tested using different technology platforms.

Understanding the protective human host immune response to M. tuberculosis

Another very important aspect in improving our understanding of TB, which requires future study, is a better understanding of why some individuals develop symptomatic clinical disease, active TB, after infection with *M.tuberculosis* while others appear not to develop clinical disease, including those with latent infection. Blood transcriptional profiles could help identify the molecular characterisation of a protected individual from an individual who develops active TB. Preliminary work towards this goal was achieved in Anne O'Garra's laboratory by stimulating whole blood, from patients with active TB, latent TB and BCG vaccinated healthy controls, with the mycobacterial antigens ESAT-6 and PPD. The RNA was prepared and processed for microarray. Analysis of the microarray data identifed a specific signature of mycobacterial exposure, demonstrating the feasibility of the approach. However no strong transcriptional responses were found to be associated with the latent TB patients alone. This is most likely related to one of more of the following: the length of stimulation (20 hours), the use of whole blood rather than PBMCs, the limitations of the stimulations added and the obscuring of more subtle immune responses by the strong transcriptional response detectable in untreated active TB patients. An expansion of this work, including conditions able to address the potential issues above, is currently on-going to try and answer these important host immune response questions.

Comparing and contrasting TB animal models to humans using the blood transcriptome

Experimentation in patients with TB has limited capacity. Hence, it is necessary to develop experimental animal models closely resembling human disease, in order to investigate pathways of pathogenesis and to test novel therapeutic strategies for disease intervention. While the murine model is an excellent model there are factors that may hypothetically unduly affect the proficiency of the murine model to mimic infected humans for example there are many different mice strains and *M. tuberculosis* strains. One method to help clarify the global differences and similarities between

M.tuberculosis infected animal models and humans, is to compare whole blood transcriptional signatures in humans and mice. Members of O'Garra's research team have run multiple mouse experiments for whole blood microarray analysis and then compared these to whole blood microarray data from patients with active TB. Preliminary comparative mouse and human data has already shown interesting results, with many further experiments on-going and planned to elucidate important questions such as comparing the role of IFN-inducible genes between the two species.

Conclusion

This broad human whole-genome study has provided new insight into the parallels and differences of the molecular signatures of four similar respiratory diseases, pulmonary tuberculosis, pulmonary sarcoidosis, community acquired pneumonia and primary lung cancer. The findings have unveiled new biological knowledge about their disease mechanisms and revealed prospective pragmatic biomarkers for disease diagnosis and treatment monitoring.

REFERENCES

Abadie, V., E. Badell, et al. (2005). "Neutrophils rapidly migrate via lymphatics after Mycobacterium bovis BCG intradermal vaccination and shuttle live bacilli to the draining lymph nodes." Blood **106**(5): 1843-1850.

Abehsera, M., D. Valeyre, et al. (2000). "Sarcoidosis with pulmonary fibrosis: CT patterns and correlation with pulmonary function." AJR Am J Roentgenol **174**(6): 1751-1757.

Adams, D. O. (1976). "The granulomatous inflammatory response. A review." Am J Pathol **84**(1): 164-192.

Agilent (2010) "GeneSpring 11.5 Manual."

Agostini, C., F. Adami, et al. (2000). "New pathogenetic insights into the sarcoid granuloma." Curr Opin Rheumatol **12**(1): 71-76.

Agostini, C., U. Basso, et al. (1998). "Cells and molecules involved in the development of sarcoid granuloma." J Clin Immunol **18**(3): 184-192.

Ainslie, G. M. and S. R. Benatar (1985). "Serum angiotensin converting enzyme in sarcoidosis: sensitivity and specificity in diagnosis: correlations with disease activity, duration, extra-thoracic involvement, radiographic type and therapy." Q J Med **55**(218): 253-270.

Akahoshi, M., M. Ishihara, et al. (2004). "Association between IFNA genotype and the risk of sarcoidosis." Hum Genet **114**(5): 503-509.

Allantaz, F., D. Chaussabel, et al. (2007). "Blood leukocyte microarrays to diagnose systemic onset juvenile idiopathic arthritis and follow the response to IL-1 blockade." J Exp Med **204**(9): 2131-2144.

Allen, S. S., W. Evans, et al. (2008). "Superoxide dismutase A antigens derived from molecular analysis of sarcoidosis granulomas elicit systemic Th-1 immune responses." Respir Res **9**: 36.

Almirall, J., I. Bolibar, et al. (1999). "Risk factors for community-acquired pneumonia in adults: a population-based case-control study." Eur Respir J **13**(2): 349-355.

Altare, F., A. Durandy, et al. (1998). "Impairment of mycobacterial immunity in human interleukin-12 receptor deficiency." Science **280**(5368): 1432-1435.

Amicosante, M. and A. P. Fontenot (2006). "T cell recognition in chronic beryllium disease." Clin Immunol **121**(2): 134-143.

Anderson, S. R., H. Maguire, et al. (2007). "Tuberculosis in London: a decade and a half of no decline [corrected]." Thorax **62**(2): 162-167.

Ardura, M. I., R. Banchereau, et al. (2009). "Enhanced monocyte response and decreased central memory T cells in children with invasive Staphylococcus aureus infections." PLoS One **4**(5): e5446.

Armstrong, J. A. and P. D. Hart (1971). "Response of cultured macrophages to Mycobacterium tuberculosis, with observations on fusion of lysosomes with phagosomes." J Exp Med **134**(3 Pt 1): 713-740.

ATS (2000). "Diagnostic Standards and Classification of Tuberculosis in Adults and Children. This official statement of the American Thoracic Society and the Centers for Disease Control and Prevention was adopted by the ATS Board of Directors, July 1999. This statement was endorsed by the Council of the Infectious Disease Society of America, September 1999." Am J Respir Crit Care Med **161**(4 Pt 1): 1376-1395.

ATS (2002). "American Thoracic Society/European Respiratory Society International Multidisciplinary Consensus Classification of the Idiopathic Interstitial

Pneumonias." American Journal of Respiratory and Critical Care Medicine **165**: 277-304.

ATS and CDC (2003). "Centers for Disease Control and Prevention. Treatment of Tuberculosis, American Thoracic Society, CDC, and Infectious Diseases Society of America. ." MMWR **52(No. RR-11)**.

Bandara, A., S. Bremner, et al. (2008). "Neutrophilia in tuberculosis." Thorax Supplement **63**: A111-A117.

Barnes, M., J. Freudenberg, et al. (2005). "Experimental comparison and cross-validation of the Affymetrix and Illumina gene expression analysis platforms." Nucleic Acids Res **33**(18): 5914-5923.

Barnes, P. F., J. M. Leedom, et al. (1988). "Predictors of short-term prognosis in patients with pulmonary tuberculosis." J Infect Dis **158**(2): 366-371.

Barnes, P. F., S. Lu, et al. (1993). "Cytokine production at the site of disease in human tuberculosis." Infect Immun **61**(8): 3482-3489.

Barry, C. E., 3rd, H. I. Boshoff, et al. (2009). "The spectrum of latent tuberculosis: rethinking the biology and intervention strategies." Nat Rev Microbiol **7**(12): 845-855.

Bates, M. N., A. Khalakdina, et al. (2007). "Risk of tuberculosis from exposure to tobacco smoke: a systematic review and meta-analysis." Arch Intern Med **167**(4): 335-342.

Baughman, A. L. (2003). "Are corticosteroids the drug of choice for chronic sarcoidosis? The con position." 69th Ann Mtng Amer Coll Chest Phys **October 25-30**.

Baughman, R. P. (2004). "Pulmonary sarcoidosis." Clin Chest Med **25**(3): 521-530, vi.

Baughman, R. P., M. Drent, et al. (2006). "Infliximab therapy in patients with chronic sarcoidosis and pulmonary involvement." Am J Respir Crit Care Med **174**(7): 795-802.

Baughman, R. P., S. Nagai, et al. (2011). "Defining the clinical outcome status (COS) in sarcoidosis: results of WASOG Task Force." Sarcoidosis Vasc Diffuse Lung Dis **28**(1): 56-64.

Baughman, R. P. and H. Nunes (2012). "Therapy for sarcoidosis: evidence-based recommendations." Expert Rev Clin Immunol **8**(1): 95-103.

Baughman, R. P., A. S. Teirstein, et al. (2001). "Clinical characteristics of patients in a case control study of sarcoidosis." Am J Respir Crit Care Med **164**(10 Pt 1): 1885-1889.

Beamer, G. L., D. K. Flaherty, et al. (2008). "Interleukin-10 promotes Mycobacterium tuberculosis disease progression in CBA/J mice." Journal of immunology **181**(8): 5545-5550.

Behar, S. M., C. J. Martin, et al. (2011). "Apoptosis is an innate defense function of macrophages against Mycobacterium tuberculosis." Mucosal Immunol **4**(3): 279-287.

Belcher, R. W. and J. D. Reid (1975). "Sarcoid granulomas in CBA/J mice. Histologic response after inoculation with sarcoid and nonsarcoid tissue homogenates." Arch Pathol **99**(5): 283-285.

Bennett, L., A. K. Palucka, et al. (2003). "Interferon and granulopoiesis signatures in systemic lupus erythematosus blood." J Exp Med **197**(6): 711-723.

Berry, M. P., C. M. Graham, et al. (2010). "An interferon-inducible neutrophil-driven blood transcriptional signature in human tuberculosis." Nature **466**(7309): 973-977.

Biddie, S. C., B. L. Conway-Campbell, et al. (2012). "Dynamic regulation of glucocorticoid signalling in health and disease." Rheumatology (Oxford) **51**(3): 403-412.

Biller, H., G. Zissel, et al. (2006). "Genotype-corrected reference values for serum angiotensin-converting enzyme." Eur Respir J **28**(6): 1085-1090.

Bleharski, J. R., H. Li, et al. (2003). "Use of genetic profiling in leprosy to discriminate clinical forms of the disease." Science **301**(5639): 1527-1530.

Boehme, C. C., P. Nabeta, et al. (2010). "Rapid molecular detection of tuberculosis and rifampin resistance." N Engl J Med **363**(11): 1005-1015.

Boisson-Dupuis, S., J. El Baghdadi, et al. (2011). "IL-12Rbeta1 deficiency in two of fifty children with severe tuberculosis from Iran, Morocco, and Turkey." PLoS One **6**(4): e18524.

Bonnefoi, H., C. Underhill, et al. (2009). "Predictive signatures for chemotherapy sensitivity in breast cancer: are they ready for use in the clinic?" Eur J Cancer **45**(10): 1733-1743.

Boor, P. P., H. J. Metselaar, et al. (2006). "Prednisolone suppresses the function and promotes apoptosis of plasmacytoid dendritic cells." Am J Transplant **6**(10): 2332-2341.

Boussiotis, V. A., E. Y. Tsai, et al. (2000). "IL-10-producing T cells suppress immune responses in anergic tuberculosis patients." J Clin Invest **105**(9): 1317-1325.

Bradley, B., H. M. Branley, et al. (2008). "Interstitial lung disease guideline: the British Thoracic Society in collaboration with the Thoracic Society of Australia and New Zealand and the Irish Thoracic Society." Thorax **63 Suppl 5**: v1-58.

Bray, F., J. S. Ren, et al. (2012). "Global estimates of cancer prevalence for 27 sites in the adult population in 2008." Int J Cancer.

Brown, S. T., I. Brett, et al. (2003). "Recovery of cell wall-deficient organisms from blood does not distinguish between patients with sarcoidosis and control subjects." Chest **123**(2): 413-417.

Bustamante, J., A. A. Arias, et al. (2011). "Germline CYBB mutations that selectively affect macrophages in kindreds with X-linked predisposition to tuberculous mycobacterial disease." Nat Immunol **12**(3): 213-221.

Campbell, I. A. and O. Bah-Sow (2006). "Pulmonary tuberculosis: diagnosis and treatment." BMJ **332**(7551): 1194-1197.

CancerStats (2012) "http://info.cancerresearchuk.org/cancerstats/keyfacts/lung-cancer/." Cancer Research UK.

Cardoso, F. L., P. R. Antas, et al. (2002). "T-cell responses to the Mycobacterium tuberculosis-specific antigen ESAT-6 in Brazilian tuberculosis patients." Infect Immun **70**(12): 6707-6714.

Carlisle, J., W. Evans, et al. (2007). "Multiple Mycobacterium antigens induce interferon-gamma production from sarcoidosis peripheral blood mononuclear cells." Clin Exp Immunol **150**(3): 460-468.

CCDC (2007). "Center for Communicable Disease Control and Prevention. Reported Tuberculosis in the United States, 2007. C. U.S. Department of Health and Human Services. Atlanta, GA.".

Center, D. M., D. A. Schwartz, et al. (2012). "Genomic Medicine and Lung Diseases: NHLBI Workshop." Am J Respir Crit Care Med.

Chakravarty, S. D., M. E. Harris, et al. (2012). "Sarcoidosis Triggered by Interferon-Beta Treatment of Multiple Sclerosis: A Case Report and Focused Literature Review." Semin Arthritis Rheum.

Charpy, J., G. B. Dowling, et al. (1947). "Vitamin D in cutaneous tuberculosis." Lancet **2**(6472): 398.

Chaussabel, D., W. Allman, et al. (2005). "Analysis of significance patterns identifies ubiquitous and disease-specific gene-expression signatures in patient peripheral blood leukocytes." Ann N Y Acad Sci **1062**: 146-154.

Chaussabel, D., V. Pascual, et al. (2010). "Assessing the human immune system through blood transcriptomics." BMC Biol **8**: 84.

Chaussabel, D., C. Quinn, et al. (2008). "A modular analysis framework for blood genomics studies: application to systemic lupus erythematosus." Immunity **29**(1): 150-164.

Cheadle, C., K. G. Becker, et al. (2007). "A rapid method for microarray cross platform comparisons using gene expression signatures." Mol Cell Probes **21**(1): 35-46.

Chen, E. S., J. Wahlstrom, et al. (2008). "T cell responses to mycobacterial catalase-peroxidase profile a pathogenic antigen in systemic sarcoidosis." J Immunol **181**(12): 8784-8796.

Chen, Y., K. Li, et al. (2011). "Corticosteroids for pneumonia." Cochrane Database Syst Rev(3): CD007720.

Choi, D., S. M. Sharma, et al. (2009). "Application of Biostatistics and Bioinformatics Tools to Identify Putative Transcription Factor-Gene Regulatory Network of Ankylosing Spondylitis and Sarcoidosis." Commun Stat Theory Methods **38**(18): 3326-3338.

Clementine, R. R., J. Lyman, et al. (2010). "Tumor necrosis factor-alpha antagonist-induced sarcoidosis." J Clin Rheumatol **16**(6): 274-279.

Cobat, A., C. J. Gallant, et al. (2009). "Two loci control tuberculin skin test reactivity in an area hyperendemic for tuberculosis." J Exp Med **206**(12): 2583-2591.

Coker, R. K. (2007). "Guidelines for the use of corticosteroids in the treatment of pulmonary sarcoidosis." Drugs **67**(8): 1139-1147.

Comstock, G. W., V. T. Livesay, et al. (1974). "The prognosis of a positive tuberculin reaction in childhood and adolescence." Am J Epidemiol **99**(2): 131-138.

Cooper, A. M. (2009). "Cell-mediated immune responses in tuberculosis." Annu Rev Immunol **27**: 393-422.

Cooper, A. M., A. D. Roberts, et al. (1995). "The role of interleukin-12 in acquired immunity to Mycobacterium tuberculosis infection." Immunology **84**(3): 423-432.

Corbett, E. L., C. J. Watt, et al. (2003). "The growing burden of tuberculosis: global trends and interactions with the HIV epidemic." Arch Intern Med **163**(9): 1009-1021.

Cosma, C. L., D. R. Sherman, et al. (2003). "The secret lives of the pathogenic mycobacteria." Annu Rev Microbiol **57**: 641-676.

Costabel, U. and G. W. Hunninghake (1999). "ATS/ERS/WASOG statement on sarcoidosis. Sarcoidosis Statement Committee. American Thoracic Society. European Respiratory Society. World Association for Sarcoidosis and Other Granulomatous Disorders." Eur Respir J **14**(4): 735-737.

Crofton, J. (2006). "The MRC randomized trial of streptomycin and its legacy: a view from the clinical front line." J R Soc Med **99**(10): 531-534.

Crouser, E. D., D. A. Culver, et al. (2009). "Gene expression profiling identifies MMP-12 and ADAMDEC1 as potential pathogenic mediators of pulmonary sarcoidosis." Am J Respir Crit Care Med **179**(10): 929-938.

Daien, C. I., A. Monnier, et al. (2009). "Sarcoid-like granulomatosis in patients treated with tumor necrosis factor blockers: 10 cases." Rheumatology (Oxford) **48**(8): 883-886.

David, H. L. (1970). "Probability distribution of drug-resistant mutants in unselected populations of Mycobacterium tuberculosis." Appl Microbiol **20**(5): 810-814.

Davis, J. M. and L. Ramakrishnan (2009). "The role of the granuloma in expansion and dissemination of early tuberculous infection." Cell **136**(1): 37-49.

De Bosscher, K., W. Vanden Berghe, et al. (2003). "The interplay between the glucocorticoid receptor and nuclear factor-kappaB or activator protein-1: molecular mechanisms for gene repression." Endocr Rev **24**(4): 488-522.

De Libero, G. and L. Mori (2008). "How T cells get grip on lipid antigens." Curr Opin Immunol **20**(1): 96-104.

de Steenwinkel, J. E., G. J. de Knegt, et al. (2010). "Time-kill kinetics of anti-tuberculosis drugs, and emergence of resistance, in relation to metabolic activity of Mycobacterium tuberculosis." J Antimicrob Chemother **65**(12): 2582-2589.

Demirkok, S. S., M. Basaranoglu, et al. (2007). "Seasonality of the onset of symptoms, tuberculin test anergy and Kveim positive reaction in a large cohort of patients with sarcoidosis." Respirology **12**(4): 591-593.

Desjardin, L. E., M. D. Perkins, et al. (1999). "Measurement of sputum Mycobacterium tuberculosis messenger RNA as a surrogate for response to chemotherapy." Am J Respir Crit Care Med **160**(1): 203-210.

DeYoung, K. L., M. E. Ray, et al. (1997). "Cloning a novel member of the human interferon-inducible gene family associated with control of tumorigenicity in a model of human melanoma." Oncogene **15**(4): 453-457.

Diacon, A. H., R. Dawson, et al. (2012). "14-day bactericidal activity of PA-824, bedaquiline, pyrazinamide, and moxifloxacin combinations: a randomised trial." Lancet.

Dienstmann, R., P. Martinez, et al. (2011). "Personalizing therapy with targeted agents in non-small cell lung cancer." Oncotarget **2**(3): 165-177.

Divangahi, M., D. Desjardins, et al. (2010). "Eicosanoid pathways regulate adaptive immunity to Mycobacterium tuberculosis." Nat Immunol **11**(8): 751-758.

Dobbin, K. K. and R. M. Simon (2011). "Optimally splitting cases for training and testing high dimensional classifiers." BMC Med Genomics **4**: 31.

Doll, R. and A. B. Hill (1950). "Smoking and carcinoma of the lung; preliminary report." Br Med J **2**(4682): 739-748.

Donald, P. R. and A. H. Diacon (2008). "The early bactericidal activity of anti-tuberculosis drugs: a literature review." Tuberculosis (Edinb) **88 Suppl 1**: S75-83.

Dorhoi, A., C. Desel, et al. (2010). "The adaptor molecule CARD9 is essential for tuberculosis control." J Exp Med **207**(4): 777-792.

Draghici, S. (2012). Statistics and data analysis for microarrays using R and bioconductor.

Drake, W. P., M. S. Dhason, et al. (2007). "Cellular recognition of Mycobacterium tuberculosis ESAT-6 and KatG peptides in systemic sarcoidosis." Infect Immun **75**(1): 527-530.

Drent, M., R. M. Wirnsberger, et al. (1999). "Association of fatigue with an acute phase response in sarcoidosis." Eur Respir J **13**(4): 718-722.

Dubey, S. and C. A. Powell (2009). "Update in lung cancer 2008." Am J Respir Crit Care Med **179**(10): 860-868.

Dupuy, A. and R. M. Simon (2007). "Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting." J Natl Cancer Inst **99**(2): 147-157.

Dye, C., S. Scheele, et al. (1999). "Consensus statement. Global burden of tuberculosis: estimated incidence, prevalence, and mortality by country. WHO Global Surveillance and Monitoring Project." JAMA **282**(7): 677-686.

Eady, J. J., G. M. Wortley, et al. (2005). "Variation in gene expression profiles of peripheral blood mononuclear cells from healthy volunteers." Physiol Genomics **22**(3): 402-411.

Edmondstone, W. M. and A. G. Wilson (1985). "Sarcoidosis in Caucasians, Blacks and Asians in London." Br J Dis Chest **79**(1): 27-36.

Edwards, A. D., S. P. Manickasingham, et al. (2002). "Microbial recognition via Toll-like receptor-dependent and -independent pathways determines the cytokine response of murine dendritic cell subsets to CD40 triggering." J Immunol **169**(7): 3652-3660.

Emamian, E. S., J. M. Leon, et al. (2009). "Peripheral blood gene expression profiling in Sjogren's syndrome." Genes Immun **10**(4): 285-296.

Erdal, B. S., E. D. Crouser, et al. (2012). "Quantitative Computerized Two-Point Correlation Analysis of Lung CT Scans Correlates with Pulmonary Function in Pulmonary Sarcoidosis." Chest.

Eum, S. Y., J. H. Kong, et al. (2010). "Neutrophils are the predominant infected phagocytic cells in the airways of patients with active pulmonary TB." Chest **137**(1): 122-128.

Facco, M., A. Cabrelle, et al. (2011). "Sarcoidosis is a Th1/Th17 multisystem disorder." Thorax **66**: 144e150.

Farr, B. M., D. L. Kaiser, et al. (1989). "Prediction of microbial aetiology at admission to hospital for pneumonia from the presenting clinical features. British Thoracic Society Pneumonia Research Subcommittee." Thorax **44**(12): 1031-1035.

Fennelly, K. P., E. C. Jones-Lopez, et al. (2012). "Variability of Infectious Aerosols Produced During Coughing by Patients with Pulmonary Tuberculosis." Am J Respir Crit Care Med.

Fernandez-Serrano, S., J. Dorca, et al. (2003). "Molecular inflammatory responses measured in blood of patients with severe community-acquired pneumonia." Clin Diagn Lab Immunol **10**(5): 813-820.

Fine, M. J., T. E. Auble, et al. (1997). "A prediction rule to identify low-risk patients with community-acquired pneumonia." N Engl J Med **336**(4): 243-250.

Flynn, J. L. and J. Chan (2001). "Immunology of tuberculosis." Annu Rev Immunol **19**: 93-129.

Flynn, J. L., J. Chan, et al. (1993). "An essential role for interferon gamma in resistance to Mycobacterium tuberculosis infection." J Exp Med **178**(6): 2249-2254.

Forbes, E. K., C. Sander, et al. (2008). "Multifunctional, high-level cytokine-producing Th1 cells in the lung, but not spleen, correlate with protection against Mycobacterium tuberculosis aerosol challenge in mice." J Immunol **181**(7): 4955-4964.

Foulon, G., M. Wislez, et al. (2004). "Sarcoidosis in HIV-infected patients in the era of highly active antiretroviral therapy." Clin Infect Dis **38**(3): 418-425.

Gaede, K. I., U. Mamat, et al. (2004). "Differential gene expression pattern in alveolar macrophages of patients with sarcoidosis and tuberculosis." J Mol Med **82**(3): 206-210.

Gallegos, A. M., E. G. Pamer, et al. (2008). "Delayed protection by ESAT-6-specific effector CD4+ T cells after airborne M. tuberculosis infection." The Journal of experimental medicine **205**(10): 2359-2368.

Garber, M. E., O. G. Troyanskaya, et al. (2001). "Diversity of gene expression in adenocarcinoma of the lung." Proc Natl Acad Sci U S A **98**(24): 13784-13789.

Gardam, M. A., E. C. Keystone, et al. (2003). "Anti-tumour necrosis factor agents and tuberculosis risk: mechanisms of action and clinical management." Lancet Infect Dis **3**(3): 148-155.

GeneSpring (2010). "GeneSpring Manual." Agilent Technologies Inc. .

Gerke, A. K. and G. Hunninghake (2008). "The immunology of sarcoidosis." Clin Chest Med **29**(3): 379-390, vii.

Gherardi, E., W. Birchmeier, et al. (2012). "Targeting MET in cancer: rationale and progress." Nat Rev Cancer **12**(2): 89-103.

Gibson, G. J., R. J. Prescott, et al. (1996). "British Thoracic Society Sarcoidosis study: effects of long term corticosteroid treatment." Thorax **51**(3): 238-247.

Golub, T. R., D. K. Slonim, et al. (1999). "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring." Science **286**(5439): 531-537.

Gong, J. H., M. Zhang, et al. (1996). "Interleukin-10 downregulates Mycobacterium tuberculosis-induced Th1 responses and CTLA-4 expression." Infection and Immunity **64**(3): 913-918.

Gordon, G. J., R. V. Jensen, et al. (2002). "Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma." Cancer Res **62**(17): 4963-4967.

Greenaway, C., D. Menzies, et al. (2002). "Delay in diagnosis among hospitalized patients with active tuberculosis--predictors and outcomes." Am J Respir Crit Care Med **165**(7): 927-933.

Greinert, U., M. Ernst, et al. (2001). "Interleukin-12 as successful adjuvant in tuberculosis treatment." The European respiratory journal : official journal of the European Society for Clinical Respiratory Physiology **17**(5): 1049-1051.

Gribbin, J., R. B. Hubbard, et al. (2006). "Incidence and mortality of idiopathic pulmonary fibrosis and sarcoidosis in the UK." Thorax **61**(11): 980-985.

Griffiths, M. J., M. J. Shafi, et al. (2005). "Genomewide analysis of the host response to malaria in Kenyan children." J Infect Dis **191**(10): 1599-1611.

Grosset, J. (1980). "Bacteriologic basis of short-course chemotherapy for tuberculosis." Clin Chest Med **1**(2): 231-241.

Grunewald, J. (2010). "Review: role of genetics in susceptibility and outcome of sarcoidosis." Semin Respir Crit Care Med **31**(4): 380-389.

Grunewald, J., J. Wahlstrom, et al. (2002). "Lung restricted T cell receptor AV2S3+ CD4+ T cell expansions in sarcoidosis patients with a shared HLA-DRbeta chain conformation." Thorax **57**(4): 348-352.

Grutters, J. C., J. M. Fellrath, et al. (2003). "Serum soluble interleukin-2 receptor measurement in patients with sarcoidosis: a clinical evaluation." Chest **124**(1): 186-195.

Guest, J. F. and A. Morris (1997). "Community-acquired pneumonia: the annual cost to the National Health Service in the UK." Eur Respir J **10**(7): 1530-1534.

Guiducci, C., M. Gong, et al. (2010). "TLR recognition of self nucleic acids hampers glucocorticoid activity in lupus." Nature **465**(7300): 937-941.

Gumbo, T., A. Louie, et al. (2007). "Isoniazid bactericidal activity and resistance emergence: integrating pharmacodynamics and pharmacogenomics to predict efficacy in different ethnic populations." Antimicrob Agents Chemother **51**(7): 2329-2336.

Gupta, D., R. Agarwal, et al. (2007). "Molecular evidence for the role of mycobacteria in sarcoidosis: a meta-analysis." Eur Respir J **30**(3): 508-516.

Gupta, D., R. Agarwal, et al. (2012). "Sarcoidosis and tuberculosis: the same disease with different manifestations or similar manifestations of different disorders." Curr Opin Pulm Med.

Gupta, D., R. Agarwal, et al. (2011). "Immune Responses to Mycobacterial Antigens in Sarcoidosis: A Systematic Review." Ind J Chest Dis **53**: 41-49.

Haahr, M. (1988) "random.org." School of Computer Science and Statistics, Trinity College, Dublin.

Haining, W. N. and E. J. Wherry (2010). "Integrating genomic signatures for immunologic discovery." Immunity **32**(2): 152-161.

Hambleton, S., S. Salem, et al. (2011). "IRF8 mutations and human dendritic-cell immunodeficiency." N Engl J Med **365**(2): 127-138.

Harari, A., V. Rozot, et al. (2011). "Dominant TNF-alpha(+) Mycobacterium tuberculosis-specific CD4(+) T cell responses discriminate between latent infection and active disease." Nat Med.

Henry, M. T., K. McMahon, et al. (2002). "Matrix metalloproteinases and tissue inhibitor of metalloproteinase-1 in sarcoidosis and IPF." Eur Respir J **20**(5): 1220-1227.

Hirsch, C. S., Z. Toossi, et al. (1999). "Depressed T-cell interferon-gamma responses in pulmonary tuberculosis: analysis of underlying mechanisms and modulation with therapy." J Infect Dis **180**(6): 2069-2073.

Hoffmann, R. M., M. C. Jung, et al. (1998). "Sarcoidosis associated with interferon-alpha therapy for chronic hepatitis C." J Hepatol **28**(6): 1058-1063.

Hofmeyr, A., W. F. Lau, et al. (2007). "Mycobacterium tuberculosis infection in patients with cancer, the role of 18-fluorodeoxyglucose positron emission tomography for diagnosis and monitoring treatment response." Tuberculosis (Edinb) **87**(5): 459-463.

Hoheisel, G. B., L. Tabak, et al. (1994). "Bronchoalveolar lavage cytology and immunocytology in pulmonary tuberculosis." Am J Respir Crit Care Med **149**(2 Pt 1): 460-463.

Homolka, J., J. Lorenz, et al. (1992). "Evaluation of soluble CD 14 and neopterin as serum parameters of the inflammatory activity of pulmonary sarcoidosis." Clin Investig **70**(10): 909-916.

Horne, D. J., S. E. Royce, et al. (2010). "Sputum monitoring during tuberculosis treatment for predicting outcome: systematic review and meta-analysis." Lancet Infect Dis **10**(6): 387-394.

HPA (2010). "Report of tuberculosis surveillance in the UK. Health Protection Agency."

Hunninghake, G. W. and R. G. Crystal (1981). "Pulmonary sarcoidosis: a disorder mediated by excess helper T-lymphocyte activity at sites of disease activity." N Engl J Med **305**(8): 429-434.

Iannuzzi, M. C. and B. A. Rybicki (2007). "Genetics of sarcoidosis: candidate genes and genome scans." Proc Am Thorac Soc **4**(1): 108-116.

Iannuzzi, M. C., B. A. Rybicki, et al. (2007). "Sarcoidosis." N Engl J Med **357**(21): 2153-2165.

Iliopoulos, A., K. Psathakis, et al. (2006). "Tuberculosis and granuloma formation in patients receiving anti-TNF therapy." Int J Tuberc Lung Dis **10**(5): 588-590.

Illumina (2005). Illumina Gene Expression Profiling Technical Bulletin.

IngenuitySystems (2012) "Ingenuity Knowledge Base." © 2012 Ingenuity Systems, Inc.

Ishikawa, E., T. Ishikawa, et al. (2009). "Direct recognition of the mycobacterial glycolipid, trehalose dimycolate, by C-type lectin Mincle." J Exp Med **206**(13): 2879-2888.

Izbicki, G., R. Chavko, et al. (2007). "World Trade Center "sarcoid-like" granulomatous pulmonary disease in New York City Fire Department rescue workers." Chest **131**(5): 1414-1423.

Jacobsen, M., D. Repsilber, et al. (2007). "Candidate biomarkers for discrimination between infection and disease caused by Mycobacterium tuberculosis." J Mol Med **85**(6): 613-621.

Jain, S. K., G. Lamichhane, et al. (2008). "Antibiotic Treatment of Tuberculosis:Old Problems, New Solutions." Microbe **3**(6): 286-292.

Janssens, J. P., P. Roux-Lombard, et al. (2007). "Quantitative scoring of an interferon-gamma assay for differentiating active from latent tuberculosis." Eur Respir J **30**(4): 722-728.

Jeon, C. Y. and M. B. Murray (2008). "Diabetes mellitus increases the risk of active tuberculosis: a systematic review of 13 observational studies." PLoS Med **5**(7): e152.

Jindal, S. K., D. Gupta, et al. (2000). "Sarcoidosis in developing countries." Curr Opin Pulm Med **6**(5): 448-454.

Jindani, A., V. R. Aber, et al. (1980). "The early bactericidal activity of drugs in patients with pulmonary tuberculosis." Am Rev Respir Dis **121**(6): 939-949.

Joosten, S. A., J. J. Goeman, et al. (2012). "Identification of biomarkers for tuberculosis disease using a novel dual-color RT-MLPA assay." Genes Immun **13**(1): 71-82.

Jouanguy, E., F. Altare, et al. (1996). "Interferon-gamma-receptor deficiency in an infant with fatal bacille Calmette-Guerin infection." N Engl J Med **335**(26): 1956-1961.

Judson, M. A. (2003). "Are corticosteroids the drug of choice for chronic sarcoidosis? The pro position." 69th Ann Mtng Amer Coll Chest Phys **October 25-30**.

Judson, M. A., R. M. Marchell, et al. (2012). "Molecular profiling and gene expression analysis in cutaneous sarcoidosis: the role of interleukin-12, interleukin-23, and the T-helper 17 pathway." J Am Acad Dermatol **66**(6): 901-910, 910 e901-902.

Judson, M. A., B. W. Thompson, et al. (2003). "The diagnostic pathway to sarcoidosis." Chest **123**(2): 406-412.

Jurado, J. O., I. B. Alvarez, et al. (2008). "Programmed death (PD)-1:PD-ligand 1/PD-ligand 2 pathway inhibits T cell effector functions during human tuberculosis." J Immunol **181**(1): 116-125.

Kagina, B. M., B. Abel, et al. (2010). "Specific T cell frequency and cytokine expression profile do not correlate with protection against tuberculosis after bacillus Calmette-Guerin vaccination of newborns." Am J Respir Crit Care Med **182**(8): 1073-1079.

Kang, Y. A., S. Y. Kwon, et al. (2009). "Role of C-reactive protein and procalcitonin in differentiation of tuberculosis from bacterial community acquired pneumonia." Korean J Intern Med **24**(4): 337-342.

Kassim, S., P. Zuber, et al. (2000). "Tuberculin skin testing to assess the occupational risk of Mycobacterium tuberculosis infection among health care workers in Abidjan, Cote d'Ivoire." Int J Tuberc Lung Dis **4**(4): 321-326.

Kaufman, R. J. (2004). "Regulation of mRNA translation by protein folding in the endoplasmic reticulum." Trends Biochem Sci **29**(3): 152-158.

Kaufmann, S. H. and E. Rubin (2008). Molecular Biology and Biochemistry. Handbook of Tuberculosis.

Keane, J., S. Gershon, et al. (2001). "Tuberculosis associated with infliximab, a tumor necrosis factor alpha-neutralizing agent." N Engl J Med **345**(15): 1098-1104.

Keicho, N., K. Kitamura, et al. (1990). "Serum concentration of soluble interleukin-2 receptor as a sensitive parameter of disease activity in sarcoidosis." Chest **98**(5): 1125-1129.

Keijsers, R. G., F. J. Verzijlbergen, et al. (2009). "18F-FDG PET, genotype-corrected ACE and sIL-2R in newly diagnosed sarcoidosis." Eur J Nucl Med Mol Imaging **36**(7): 1131-1137.

Keir, G. and A. U. Wells (2010). "Assessing pulmonary disease and response to therapy: which test?" Semin Respir Crit Care Med **31**(4): 409-418.

Kim, S. J., D. J. Dix, et al. (2007). "Effects of storage, RNA extraction, genechip type, and donor sex on gene expression profiling of human whole blood." Clin Chem **53**(6): 1038-1045.

Klech, H., H. Kohn, et al. (1982). "Assessment of activity in Sarcoidosis. Sensitivity and specificity of 67Gallium scintigraphy, serum ACE levels, chest roentgenography, and blood lymphocyte subpopulations." Chest **82**(6): 732-738.

Korbel, D. S., B. E. Schneider, et al. (2008). "Innate immunity in tuberculosis: myths and truth." Microbes Infect **10**(9): 995-1004.

Koth, L. L., O. D. Solberg, et al. (2011). "Sarcoidosis blood transcriptome reflects lung inflammation and overlaps with tuberculosis." Am J Respir Crit Care Med **184**(10): 1153-1163.

Kwek, S. S., E. Cha, et al. (2012). "Unmasking the immune recognition of prostate cancer with CTLA4 blockade." Nat Rev Cancer **12**(4): 289-297.

Lalkhen, A. G. and A. McCluskey (2008). "Clinical tests: sensitivity and specificity." Contin Edu Anaesth Crit Care **8**(6): 221-223.

Lee, E. and R. S. Holzman (2002). "Evolution and current use of the tuberculin test." Clin Infect Dis **34**(3): 365-370.

Lesho, E., F. J. Forestiero, et al. (2011). "Transcriptional responses of host peripheral blood cells to tuberculosis infection." Tuberculosis (Edinb) **91**(5): 390-399.

Leung, A. N., M. W. Brauner, et al. (1998). "Sarcoidosis activity: correlation of HRCT findings with those of 67Ga scanning, bronchoalveolar lavage, and serum angiotensin-converting enzyme assay." J Comput Assist Tomogr **22**(2): 229-234.

Lim, E., D. Baldwin, et al. (2010). "Guidelines on the radical management of patients with lung cancer." Thorax **65 Suppl 3**: iii1-27.

Lim, W. S., S. V. Baudouin, et al. (2009). "BTS guidelines for the management of community acquired pneumonia in adults: update 2009." Thorax **64 Suppl 3**: iii1-55.

Lim, W. S., M. M. van der Eerden, et al. (2003). "Defining community acquired pneumonia severity on presentation to hospital: an international derivation and validation study." Thorax **58**(5): 377-382.

Lin, P. L., A. Myers, et al. (2010). "Tumor necrosis factor neutralization results in disseminated disease in acute and latent Mycobacterium tuberculosis infection with normal granuloma structure in a cynomolgus macaque model." Arthritis Rheum **62**(2): 340-350.

Lin, P. L., M. Rodgers, et al. (2009). "Quantitative comparison of active and latent tuberculosis in the cynomolgus macaque model." Infect Immun **77**(10): 4631-4642.

Liu, P. T., S. Stenger, et al. (2006). "Toll-like receptor triggering of a vitamin D-mediated human antimicrobial response." Science **311**(5768): 1770-1773.

Lockstone, H. E., S. Sanderson, et al. (2010). "Gene set analysis of lung samples provides insight into pathogenesis of progressive, fibrotic pulmonary sarcoidosis." Am J Respir Crit Care Med **181**(12): 1367-1375.

Long, R. and M. Gardam (2003). "Tumour necrosis factor-alpha inhibitors and the reactivation of latent tuberculosis infection." CMAJ **168**(9): 1153-1156.

Lonnroth, K., B. G. Williams, et al. (2008). "Alcohol use as a risk factor for tuberculosis - a systematic review." BMC Public Health **8**: 289.

Lopez-Maderuelo, D., F. Arnalich, et al. (2003). "Interferon-gamma and interleukin-10 gene polymorphisms in pulmonary tuberculosis." Am J Respir Crit Care Med **167**(7): 970-975.

MacMicking, J., Q. W. Xie, et al. (1997). "Nitric oxide and macrophage function." Annu Rev Immunol **15**: 323-350.

Maertzdorf, J., M. Ota, et al. (2011). "Functional correlations of pathogenesis-driven gene expression signatures in tuberculosis." PLoS One **6**(10): e26938.

Maertzdorf, J., D. Repsilber, et al. (2011). "Human gene expression profiles of susceptibility and resistance in tuberculosis." Genes Immun **12**(1): 15-22.

Maertzdorf, J., J. Weiner, 3rd, et al. (2012). "Common patterns and disease-related signatures in tuberculosis and sarcoidosis." Proc Natl Acad Sci U S A **109**(20): 7853-7858.

Mahairas, G. G., P. J. Sabo, et al. (1996). "Molecular analysis of genetic differences between Mycobacterium bovis BCG and virulent M. bovis." J Bacteriol **178**(5): 1274-1282.

Mana, J., A. Salazar, et al. (1996). "Are the pulmonary function tests and the markers of activity helpful to establish the prognosis of sarcoidosis?" Respiration **63**(5): 298-303.

Manca, C., L. Tsenova, et al. (2005). "Hypervirulent M. tuberculosis W/Beijing strains upregulate type I IFNs and increase expression of negative regulators of the Jak-Stat pathway." J Interferon Cytokine Res **25**(11): 694-701.

Mantovani, A., S. Sozzani, et al. (2002). "Macrophage polarization: tumor-associated macrophages as a paradigm for polarized M2 mononuclear phagocytes." Trends in immunology **23**(11): 549-555.

Marchiori, E., G. Zanetti, et al. (2011). "Atypical distribution of small nodules on high resolution CT studies: patterns and differentials." Respir Med **105**(9): 1263-1267.

Martens, S. and J. Howard (2006). "The interferon-inducible GTPases." Annu Rev Cell Dev Biol **22**: 559-589.

Martineau, A. R., S. Nhamoyebonde, et al. (2011). "Reciprocal seasonal variation in vitamin D status and tuberculosis notifications in Cape Town, South Africa." Proc Natl Acad Sci U S A **108**(47): 19013-19017.

Martineau, A. R., P. M. Timms, et al. (2011). "High-dose vitamin D(3) during intensive-phase antimicrobial treatment of pulmonary tuberculosis: a double-blind randomised controlled trial." Lancet **377**(9761): 242-250.

Mathew, S., K. L. Bauer, et al. (2008). "The anergic state in sarcoidosis is associated with diminished dendritic cell function." J Immunol **181**(1): 746-755.

McColl, A., S. Michlewska, et al. (2007). "Effects of glucocorticoids on apoptosis and clearance of apoptotic cells." ScientificWorldJournal **7**: 1165-1181.

McMurray, D. N. (2001). "Disease model: pulmonary tuberculosis." Trends Mol Med **7**(3): 135-137.

McNab, F. W., M. P. Berry, et al. (2011). "Programmed death ligand 1 is over-expressed by neutrophils in the blood of patients with active tuberculosis." Eur J Immunol.

Medlar, E. M. (1948). "The pathogenesis of minimal pulmonary tuberculosis; a study of 1,225 necropsies in cases of sudden and unexpected death." Am Rev Tuberc **58**(6): 583-611.

Minshall, E. M., A. Tsicopoulos, et al. (1997). "Cytokine mRNA gene expression in active and nonactive pulmonary sarcoidosis." Eur Respir J **10**(9): 2034-2039.

Mistry, R., J. M. Cliff, et al. (2007). "Gene-expression patterns in whole blood identify subjects at risk for recurrent tuberculosis." J Infect Dis **195**(3): 357-365.

Mitchell, D. N., R. J. Rees, et al. (1976). "Transmissible agents from human sarcoid and Crohn's disease tissues." Lancet **2**(7989): 761-765.

Mitchell, D. N., J. G. Scadding, et al. (1977). "Sarcoidosis: histopathological definition and clinical diagnosis." J Clin Pathol **30**(5): 395-408.

Mitchison, D. A. (1985). "The action of antituberculosis drugs in short-course chemotherapy." Tubercle **66**(3): 219-225.

Mitchison, D. A. (1993). "Assessment of new sterilizing drugs for treating pulmonary tuberculosis by culture at 2 months." Am Rev Respir Dis **147**(4): 1062-1063.

Miyara, M., Z. Amoura, et al. (2006). "The immune paradox of sarcoidosis and regulatory T cells." J Exp Med **203**(2): 359-370.

Mohan, V. P., C. A. Scanga, et al. (2001). "Effects of tumor necrosis factor alpha on host immune response in chronic persistent tuberculosis: possible role for limiting pathology." Infect Immun **69**(3): 1847-1855.

Mohr, S. and C. C. Liew (2007). "The peripheral-blood transcriptome: new insights into disease and risk assessment." Trends Mol Med **13**(10): 422-432.

Moller, D. R. (2007). "Potential etiologic agents in sarcoidosis." Proc Am Thorac Soc **4**(5): 465-468.

Moodley, Y. P., T. Dorasamy, et al. (2000). "Correlation of CD4:CD8 ratio and tumour necrosis factor (TNF)alpha levels in induced sputum with bronchoalveolar lavage fluid in pulmonary sarcoidosis." Thorax **55**(8): 696-699.

Moran, L. B., D. C. Duke, et al. (2007). "The microglial gene regulatory network activated by interferon-gamma." J Neuroimmunol **183**(1-2): 1-6.

Morell, F., G. Levy, et al. (2002). "Delayed cutaneous hypersensitivity tests and lymphopenia as activity markers in sarcoidosis." Chest **121**(4): 1239-1244.

Mukhopadhyay, S. and A. A. Gal (2010). "Granulomatous lung disease: an approach to the differential diagnosis." Arch Pathol Lab Med **134**(5): 667-690.

Munro, C. S. and D. N. Mitchell (1987). "The K veim response: still useful, still a puzzle." Thorax **42**(5): 321-331.

Murray, J. F. (2004). "A century of tuberculosis." Am J Respir Crit Care Med **169**(11): 1181-1186.

Nagai, S., M. Shigematsu, et al. (1999). "Clinical courses and prognoses of pulmonary sarcoidosis." Curr Opin Pulm Med **5**(5): 293-298.

Nahid, P., J. Saukkonen, et al. (2011). "Tuberculosis Biomarker and Surrogate Endpoint Research Roadmap." Am J Respir Crit Care Med **184**(8): 972-979.

Nahid, P., J. Saukkonen, et al. (2011). "CDC/NIH Workshop. Tuberculosis biomarker and surrogate endpoint research roadmap." Am J Respir Crit Care Med **184**(8): 972-979.

Nakaya, H. I., J. Wrammert, et al. (2011). "Systems biology of vaccination for seasonal influenza in humans." Nat Immunol **12**(8): 786-795.

Nascimento, E. J., U. Braga-Neto, et al. (2009). "Gene expression profiling during early acute febrile stage of dengue infection can predict the disease outcome." PLoS One **4**(11): e7892.

Nash, D. R. and J. E. Douglass (1980). "Anergy in active pulmonary tuberculosis. A comparison between positive and negative reactors and an evaluation of 5 TU and 250 TU skin test doses." Chest **77**(1): 32-37.

Nature, editorial, et al. (2012). "Error prone." Nature **487**(7408): 406.

Ness, S. A. (2006). "Basic microarray analysis: strategies for successful experiments." Methods Mol Biol **316**: 13-33.

Newman, L. S., C. S. Rose, et al. (2004). "A case control etiologic study of sarcoidosis: environmental and occupational risk factors." Am J Respir Crit Care Med **170**(12): 1324-1330.

NICE (2011). "CG117 Tuberculosis: NICE guideline " National Institute for Health and Clinical Excellence.

NICE (2011). "Lung cancer: The diagnosis and treatment of lung cancer. ." National Institute for Health and Clinical Excellence.

North, R. J. and Y. J. Jung (2004). "Immunity to tuberculosis." Annu Rev Immunol **22**: 599-623.

Novikov, A., M. Cardone, et al. (2011). "Mycobacterium tuberculosis triggers host type I IFN signaling to regulate IL-1beta production in human macrophages." J Immunol **187**(5): 2540-2547.

Nunn, A. J., P. P. Phillips, et al. (2010). "Timing of relapse in short-course chemotherapy trials for tuberculosis." Int J Tuberc Lung Dis **14**(2): 241-242.

O'Callaghan, D. S., D. O'Donnell, et al. (2010). "The role of inflammation in the pathogenesis of non-small cell lung cancer." J Thorac Oncol **5**(12): 2024-2036.

O'Garra, A. B., W.J., Ed. (2008). Cytokines in Tuberculosis. Handbook of Tuberculosis.

Oliveros, J. C. (2007) "Venny." An interactive tool for comparing lists with Venn Diagrams.

Olson, N. E. (2006). "The microarray data analysis process: from raw data to biological significance." NeuroRx **3**(3): 373-383.

Oremek, G. M., H. Sauer-Eppel, et al. (2007). "Value of tumour and inflammatory markers in lung cancer." Anticancer Res **27**(4A): 1911-1915.

Oswald-Richter, K., H. Sato, et al. (2010). "Mycobacterial ESAT-6 and katG are recognized by sarcoidosis CD4+ T cells when presented by the American sarcoidosis susceptibility allele, DRB1*1101." J Clin Immunol **30**(1): 157-166.

Oswald-Richter, K. A., D. C. Beachboard, et al. (2012). "Dual Analysis for Mycobacteria and Propionibacteria in Sarcoidosis BAL." J Clin Immunol.

Oswald-Richter, K. A., D. C. Beachboard, et al. (2010). "Multiple mycobacterial antigens are targets of the adaptive immune response in pulmonary sarcoidosis." Respir Res **11**: 161.

Oswald-Richter, K. A., D. A. Culver, et al. (2009). "Cellular responses to mycobacterial antigens are present in bronchoalveolar lavage fluid used in the diagnosis of sarcoidosis." Infect Immun **77**(9): 3740-3748.

Ottenhoff, T. H., D. Kumararatne, et al. (1998). "Novel human immunodeficiencies reveal the essential role of type-I cytokines in immunity to intracellular bacteria." Immunol Today **19**(11): 491-494.

Pacheco, A. G., C. C. Cardoso, et al. (2008). "IFNG +874T/A, IL10 -1082G/A and TNF -308G/A polymorphisms in association with tuberculosis susceptibility: a meta-analysis study." Hum Genet **123**(5): 477-484.

Pai, M., A. Zwerling, et al. (2008). "Systematic review: T-cell-based assays for the diagnosis of latent tuberculosis infection: an update." Ann Intern Med **149**(3): 177-184.

Pankla, R., S. Buddhisa, et al. (2009). "Genomic transcriptional profiling identifies a candidate blood biomarker signature for the diagnosis of septicemic melioidosis." Genome Biol **10**(11): R127.

Papp, K. A., C. Leonardi, et al. (2012). "Brodalumab, an anti-interleukin-17-receptor antibody for psoriasis." N Engl J Med **366**(13): 1181-1189.

Paramothayan, N. S., T. J. Lasserson, et al. (2005). "Corticosteroids for pulmonary sarcoidosis." Cochrane Database Syst Rev(2): CD001114.

Paramothayan, S. and P. W. Jones (2002). "Corticosteroid therapy in pulmonary sarcoidosis: a systematic review." JAMA **287**(10): 1301-1307.

Paramothayan, S., T. J. Lasserson, et al. (2006). "Immunosuppressive and cytotoxic therapy for pulmonary sarcoidosis." Cochrane Database Syst Rev(3): CD003536.

Paramothayan, S., T. J. Lasserson, et al. (2006). "Immunosuppressive and cytotoxic therapy for pulmonary sarcoidosis." Cochrane Database Syst Rev **3**: CD003536.

Pascual, V., D. Chaussabel, et al. (2010). "A genomic approach to human autoimmune diseases." Annu Rev Immunol **28**: 535-571.

Pathan, A. A., K. A. Wilkinson, et al. (2001). "Direct ex vivo analysis of antigen-specific IFN-gamma-secreting CD4 T cells in Mycobacterium tuberculosis-infected individuals: associations with clinical disease state and effect of treatment." J Immunol **167**(9): 5217-5225.

Perrin, F. M., M. C. Lipman, et al. (2007). "Biomarkers of treatment response in clinical trials of novel antituberculosis agents." Lancet Infect Dis **7**(7): 481-490.

Peters, J. K. (2008). Introduction to Microarray Bioinformatics. Netherlands.

Peters, W. and J. D. Ernst (2003). "Mechanisms of cell recruitment in the immune response to Mycobacterium tuberculosis." Microbes Infect **5**(2): 151-158.

Pfyffer, G. E., C. Cieslak, et al. (1997). "Rapid detection of mycobacteria in clinical specimens by using the automated BACTEC 9000 MB system and comparison with radiometric and solid-culture systems." J Clin Microbiol **35**(9): 2229-2234.

Pierre, P. (2009). "Immunity and the regulation of protein synthesis: surprising connections." Curr Opin Immunol **21**(1): 70-77.

Pillich, H., M. Loose, et al. (2012). "Activation of the unfolded protein response by Listeria monocytogenes." Cell Microbiol **14**(6): 949-964.

Prasse, A., C. Katic, et al. (2008). "Phenotyping sarcoidosis from a pulmonary perspective." Am J Respir Crit Care Med **177**(3): 330-336.

Quackenbush, J. (2001). "Computational analysis of microarray data." Nat Rev Genet **2**(6): 418-427.

Querec, T. D., R. S. Akondy, et al. (2009). "Systems biology approach predicts immunogenicity of the yellow fever vaccine in humans." Nat Immunol **10**(1): 116-125.

Ramilo, O., W. Allman, et al. (2007). "Gene expression patterns in blood leukocytes discriminate patients with acute infections." Blood **109**(5): 2066-2077.

Redford, P. S., A. Boonstra, et al. (2010). "Enhanced protection to Mycobacterium tuberculosis infection in IL-10-deficient mice is accompanied by early and enhanced Th1 responses in the lung." Eur J Immunol **40**(8): 2200-2210.

Redford, P. S., P. J. Murray, et al. (2011). "The role of IL-10 in immune regulation during M. tuberculosis infection." Mucosal Immunol **4**(3): 261-270.

Reljic, R. (2007). "IFN-gamma therapy of tuberculosis and related infections." Journal of interferon & cytokine research : the official journal of the International Society for Interferon and Cytokine Research **27**(5): 353-364.

Richeldi, L. (2006). "An update on the diagnosis of tuberculosis infection." Am J Respir Crit Care Med **174**(7): 736-742.

Rizzato, G., G. Palmieri, et al. (2004). "The organ-specific extrapulmonary presentation of sarcoidosis: a frequent occurrence but a challenge to an early diagnosis. A 3-year-long prospective observational study." Sarcoidosis Vasc Diffuse Lung Dis **21**(2): 119-126.

Rodger, A., S. Jaffar, et al. (2003). "Delay in the diagnosis of pulmonary tuberculosis, London, 1998-2000: analysis of surveillance data." BMJ **326**(7395): 909-910.

Rosenbaum, J. T., S. Pasadhika, et al. (2009). "Hypothesis: sarcoidosis is a STAT1-mediated disease." Clin Immunol **132**(2): 174-183.

Rossman, M. D., L. S. Newman, et al. (2006). "A double-blinded, randomized, placebo-controlled trial of infliximab in subjects with active pulmonary sarcoidosis." Sarcoidosis Vasc Diffuse Lung Dis **23**(3): 201-208.

Russell, D. G., P. J. Cardona, et al. (2009). "Foamy macrophages and the progression of the human tuberculosis granuloma." Nat Immunol **10**(9): 943-948.

Rutherford, R. M., J. Kehren, et al. (2001). "Functional genomics in sarcoidosis--reduced or increased apoptosis?" Swiss Med Wkly **131**(31-32): 459-470.

Rutherford, R. M., F. Staedtler, et al. (2004). "Functional genomics and prognosis in sarcoidosis--the critical role of antigen presentation." Sarcoidosis Vasc Diffuse Lung Dis **21**(1): 10-18.

Rybicki, B. A., M. C. Iannuzzi, et al. (2001). "Familial aggregation of sarcoidosis. A case-control etiologic study of sarcoidosis (ACCESS)." Am J Respir Crit Care Med **164**(11): 2085-2091.

Rybicki, B. A., M. Major, et al. (1997). "Racial differences in sarcoidosis incidence: a 5-year study in a health maintenance organization." Am J Epidemiol **145**(3): 234-241.

Samokhin, A. O., F. Buhling, et al. (2010). "ApoE-deficient mice on cholate-containing high-fat diet reveal a pathology similar to lung sarcoidosis." Am J Pathol **176**(3): 1148-1156.

Sarrazin, H., K. Wilkinson, et al. (2009). "Association between tuberculin skin test reactivity, the memory CD4 cell subset, and circulating FoxP3-expressing cells in HIV-infected persons." J Infect Dis **199**: 702-710.

Saunders, B. M. and I. M. Orme (2008). Immunopathology of Tuberculosis. Handbook of Tuberculosis.

Scadding, J. G. (1961). "Prognosis of intrathoracic sarcoidosis in England. A review of 136 cases after five years' observation." Br Med J **2**(5261): 1165-1172.

Schena, M., D. Shalon, et al. (1995). "Quantitative monitoring of gene expression patterns with a complementary DNA microarray." Science **270**(5235): 467-470.

Schurmann, M., R. Kwiatkowski, et al. (2008). "Study of Toll-like receptor gene loci in sarcoidosis." Clin Exp Immunol **152**(3): 423-431.

Schwartz, H. J., F. C. Lowell, et al. (1968). "Steroid resistance in bronchial asthma." Ann Intern Med **69**(3): 493-499.

Seimon, T. A., M. J. Kim, et al. (2010). "Induction of ER stress in macrophages of tuberculosis granulomas." PLoS One **5**(9): e12772.

Selman, M., A. Pardo, et al. (2006). "Gene expression profiles distinguish idiopathic pulmonary fibrosis from hypersensitivity pneumonitis." Am J Respir Crit Care Med **173**(2): 188-198.

Shen-Orr, S. S., R. Tibshirani, et al. (2010). "Cell type-specific gene expression differences in complex tissues." Nat Methods **7**(4): 287-289.

Shi, L., G. Campbell, et al. (2010). "The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models." Nat Biotechnol **28**(8): 827-838.

Shipp, L. E., J. V. Lee, et al. (2010). "Transcriptional regulation of human dual specificity protein phosphatase 1 (DUSP1) gene by glucocorticoids." PLoS One **5**(10): e13754.

Silver, R. F., R. G. Crystal, et al. (1996). "Limited heterogeneity of biased T-cell receptor V beta gene usage in lung but not blood T cells in active pulmonary sarcoidosis." Immunology **88**(4): 516-523.

Silverstein, E., L. P. Pertschuk, et al. (1979). "Immunofluorescent localization of angiotensin converting enzyme in epithelioid and giant cells of sarcoidosis granulomas." Proc Natl Acad Sci U S A **76**(12): 6646-6648.

Small, P. M., N. B. McClenny, et al. (1993). "Molecular strain typing of Mycobacterium tuberculosis to confirm cross-contamination in the mycobacteriology laboratory and modification of procedures to minimize occurrence of false-positive cultures." J Clin Microbiol **31**(7): 1677-1682.

Song, Z., L. Marzilli, et al. (2005). "Mycobacterial catalase-peroxidase is a tissue antigen and target of the adaptive immune response in systemic sarcoidosis." J Exp Med **201**(5): 755-767.

Stekel, D. (2003). Microarray Bioinformatics, Cambridge.

Storla, D. G., S. Yimer, et al. (2008). "A systematic review of delay in the diagnosis and treatment of tuberculosis." BMC Public Health **8**: 15.

Sturgill-Koszycki, S., P. H. Schlesinger, et al. (1994). "Lack of acidification in Mycobacterium phagosomes produced by exclusion of the vesicular proton-ATPase." Science **263**(5147): 678-681.

Subramanian, A., P. Tamayo, et al. (2005). "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles." Proc Natl Acad Sci U S A **102**(43): 15545-15550.

Subramanian, J. and R. Simon (2010). "Gene expression-based prognostic signatures in lung cancer: ready for clinical use?" J Natl Cancer Inst **102**(7): 464-474.

Swaisgood, C. M., K. Oswald-Richter, et al. (2011). "Development of a sarcoidosis murine lung granuloma model using Mycobacterium superoxide dismutase A peptide." Am J Respir Cell Mol Biol **44**(2): 166-174.

Sweiss, N. J., R. Salloum, et al. (2010). "Significant CD4, CD8, and CD19 lymphopenia in peripheral blood of sarcoidosis patients correlates with severe disease manifestations." PLoS One **5**(2): e9088.

Taegtmeyer, M., N. J. Beeching, et al. (2008). "The clinical impact of nucleic acid amplification tests on the diagnosis and management of tuberculosis in a British hospital." Thorax **63**(4): 317-321.

Tamura, A., M. Shimada, et al. "The value of fiberoptic bronchoscopy in culture-positive pulmonary tuberculosis patients whose pre-bronchoscopic sputum specimens were negative both for smear and PCR analyses." Intern Med **49**(2): 95-102.

Tamura, A., M. Shimada, et al. (2010). "The value of fiberoptic bronchoscopy in culture-positive pulmonary tuberculosis patients whose pre-bronchoscopic sputum specimens were negative both for smear and PCR analyses." Intern Med **49**(2): 95-102.

Tanaka, Y., C. T. Morita, et al. (1995). "Natural and synthetic non-peptide antigens recognized by human gamma delta T cells." Nature **375**(6527): 155-158.

Tattermusch, S., J. A. Skinner, et al. (2012). "Systems biology approaches reveal a specific interferon-inducible signature in HTLV-1 associated myelopathy." PLoS pathogens **8**(1): e1002480.

Taylor, J. M., D. P. Ankerst, et al. (2008). "Validation of biomarker-based risk prediction models." Clin Cancer Res **14**(19): 5977-5983.

Thach, D. C., B. K. Agan, et al. (2005). "Surveillance of transcriptomes in basic military trainees with normal, febrile respiratory illness, and convalescent phenotypes." Genes Immun **6**(7): 588-595.

Thillai, M., C. Eberhardt, et al. (2012). "Sarcoidosis and tuberculosis cytokine profiles: indistinguishable in bronchoalveolar lavage but different in blood." PLoS One **7**(7): e38083.

Thomas, K. W. and G. W. Hunninghake (2003). "Sarcoidosis." JAMA **289**(24): 3300-3303.

Thompson, C. L., B. A. Rybicki, et al. (2006). "Reduction of sample heterogeneity through use of population substructure: an example from a population of African American families with sarcoidosis." Am J Hum Genet **79**(4): 606-613.

Thonhofer, R., C. Maercker, et al. (2002). "Expression of sarcoidosis related genes in lung lavage cells." Sarcoidosis Vasc Diffuse Lung Dis **19**(1): 59-65.

Tomlinson, G. S., T. J. Cashmore, et al. (2011). "Transcriptional profiling of innate and adaptive human immune responses to mycobacteria in the tuberculin skin test." Eur J Immunol **41**(11): 3253-3260.

Travis, W. D., E. Brambilla, et al., Eds. (2004). Pathology and Genetics of Tumours of the Lung, Pleura, Thymus and Heart. World Health Organisation Classification of Tumours., Lyon: IARC Press.

Tsai, M. C., S. Chakravarty, et al. (2006). "Characterization of the tuberculous granuloma in murine and human lungs: cellular composition and relative tissue oxygen tension." Cell Microbiol **8**(2): 218-232.

Ulrichs, T. and S. H. Kaufmann (2006). "New insights into the function of granulomas in human tuberculosis." J Pathol **208**(2): 261-269.

Ulrichs, T., G. A. Kosmiadi, et al. (2004). "Human tuberculous granulomas induce peripheral lymphoid follicle-like structures to orchestrate local host defence in the lung." J Pathol **204**(2): 217-228.

van 't Veer, L. J., H. Dai, et al. (2002). "Gene expression profiling predicts clinical outcome of breast cancer." Nature **415**(6871): 530-536.

van Rie, A., R. Warren, et al. (1999). "Exogenous reinfection as a cause of recurrent tuberculosis after curative treatment." N Engl J Med **341**(16): 1174-1179.

Veenbergen, S., R. L. Smeets, et al. (2010). "The natural soluble form of IL-18 receptor beta exacerbates collagen-induced arthritis via modulation of T-cell immune responses." Ann Rheum Dis **69**(1): 276-283.

Vekemans, J., C. Lienhardt, et al. (2001). "Tuberculosis contacts but not patients have higher gamma interferon responses to ESAT-6 than do community controls in The Gambia." Infect Immun **69**(10): 6554-6557.

Vukmanovic-Stejic, M. (2006). "Mantoux Test as a model for a secondary immune response in humans." Immunol Lett **107**: 93-101.

Vynnycky, E. and P. E. Fine (2000). "Lifetime risks, incubation period, and serial interval of tuberculosis." Am J Epidemiol **152**(3): 247-263.

Wahlstrom, J., K. Katchar, et al. (2001). "Analysis of intracellular cytokines in CD4+ and CD8+ lung and blood T cells in sarcoidosis." Am J Respir Crit Care Med **163**(1): 115-121.

Wallis, R. S., M. Pai, et al. (2010). "Biomarkers and diagnostics for tuberculosis: progress, needs, and translation into practice." Lancet **375**(9729): 1920-1937.

Walter, P. and D. Ron (2011). "The unfolded protein response: from stress pathway to homeostatic regulation." Science **334**(6059): 1081-1086.

Walzl, G., K. Ronacher, et al. (2008). "Biomarkers for TB treatment response: challenges and future strategies." J Infect **57**(2): 103-109.

Walzl, G., K. Ronacher, et al. (2011). "Immunological biomarkers of tuberculosis." Nat Rev Immunol **11**(5): 343-354.

Wasfi, Y. S., C. S. Rose, et al. (2006). "A new tool to assess sarcoidosis severity." Chest **129**(5): 1234-1245.

WASOG (1994). "Consensus conference: activity of sarcoidosis. Third WASOG meeting, Los Angeles, USA, September 8-11, 1993." Eur Respir J **7**(3): 624-627.

WASOG (1999). "Statement on sarcoidosis. Joint Statement of the American Thoracic Society (ATS), the European Respiratory Society (ERS) and the World Association of Sarcoidosis and Other Granulomatous Disorders (WASOG) adopted by the ATS Board of Directors and by the ERS Executive Committee, February 1999." Am J Respir Crit Care Med **160**(2): 736-755.

Wells, A. (1998). "High resolution computed tomography in sarcoidosis: a clinical perspective." Sarcoidosis Vasc Diffuse Lung Dis **15**(2): 140-146.

Whitney, A. R., M. Diehn, et al. (2003). "Individuality and variation in gene expression patterns in human blood." Proc Natl Acad Sci U S A **100**(4): 1896-1901.

WHO (2009). "Treatment of Tuberculosis Guidelines. World Health Organisation." **Fourth Edition**.

WHO (2010). "Global tuberculosis control. World Health Organisation.".

Wilder, S. P., P. J. Kaisaki, et al. (2009). "Comparative analysis of methods for gene transcription profiling data derived from different microarray technologies in rat and mouse models of diabetes." BMC Genomics **10**: 63.

Williams, G. T. and W. J. Williams (1983). "Granulomatous inflammation--a review." J Clin Pathol **36**(7): 723-733.

Windgassen, E. B., L. Funtowicz, et al. (2011). "C-reactive protein and high-sensitivity C-reactive protein: an update for clinicians." Postgrad Med **123**(1): 114-119.

Winterbauer, R. H., J. Lammert, et al. (1993). "Bronchoalveolar lavage cell populations in the diagnosis of sarcoidosis." Chest **104**(2): 352-361.

Wiwien, H. W., K. Hiyama, et al. (1996). "Differential display of messenger RNA expressed in bronchoalveolar lavage cells in pulmonary sarcoidosis patients." Hiroshima J Med Sci **45**(1): 1-10.

Wolf, A. J., L. Desvignes, et al. (2008). "Initiation of the adaptive immune response to Mycobacterium tuberculosis depends on antigen production in the local lymph node, not the lungs." J Exp Med **205**(1): 105-115.

Wolf, A. J., B. Linas, et al. (2007). "Mycobacterium tuberculosis infects dendritic cells with high frequency and impairs their function in vivo." J Immunol **179**(4): 2509-2519.

Wyngarden, J. B. (1988). Cecil Textbook of Medicine.

Yamagata, N., Y. Shyr, et al. (2003). "A training-testing approach to the molecular classification of resected non-small cell lung cancer." Clin Cancer Res **9**(13): 4695-4704.

Yang, I. V., L. H. Burch, et al. (2007). "Gene expression profiling of familial and sporadic interstitial pneumonia." Am J Respir Crit Care Med **175**(1): 45-54.

Young, D., J. Stark, et al. (2008). "Systems biology of persistent infection: tuberculosis as a case study." Nat Rev Microbiol **6**(7): 520-528.

Young, D. B., M. D. Perkins, et al. (2008). "Confronting the scientific obstacles to global control of tuberculosis." J Clin Invest **118**(4): 1255-1265.

Zaas, A. K., M. Chen, et al. (2009). "Gene expression signatures diagnose influenza and other symptomatic respiratory viral infections in humans." Cell Host Microbe **6**(3): 207-217.

Zamai, L., C. Ponti, et al. (2007). "NK cells and cancer." J Immunol **178**(7): 4011-4016.

Zappala, C. J., S. R. Desai, et al. (2011). "Optimal scoring of serial change on chest radiography in sarcoidosis." Sarcoidosis Vasc Diffuse Lung Dis **28**(2): 130-138.

Zhang, S. and J. Cao (2009). "A close examination of double filtering with fold change and T test in microarray analysis." BMC Bioinformatics **10**: 402.