An integrated framework modelling susceptibility to tuberculosis in homogeneous and admixed populations

by

Zoe Zerihun Gebremariam

Thesis presented in partial fulfilment of the requirements for the degree of Master of Science

at Stellenbosch University

Department of Mathematical Sciences,
Mathematics Division,
University of Stellenbosch,
Private Bag X1, Matieland 7602, South Africa.

Supervisor: Dr. Gaston K. Mazandu

December 2016

Declaration

By submitting this thesis electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Signature: Z.Z Gebremariam .....

Date: December 2016 .....

Abstract

An integrated framework modelling susceptibility to tuberculosis in homogeneous and admixed populations

Z.Z Gebremariam

Department of Mathematical Sciences,
Mathematics Division,
University of Stellenbosch,
Private Bag X1, Matieland 7602, South Africa.

Thesis: MSc

March 2016

In spite of the wide variety of anti-tuberculosis drugs, tuberculosis (TB), caused by *mycobacterium tuberculosis* (MTB), is the second leading infectious disease after *Human Immunodeficiency Virus* (HIV) or Acquired Immunodeficiency Syndrome (AIDS), and one of the leading causes of human death from infectious diseases, especially in Sub-Saharan Africa. Approximately one-third of the world population are latently infected with MTB, of which, 10 % progress to active TB. Obstacles in TB control include lengthy treatment regimens of more than 6 months, drug resistance, lack of an effective vaccine and limited knowledge and incomplete information about factors that trigger the progression of an MTB infection to disease. Moreover, the association of TB and HIV or AIDS has also promoted all of the conditions of an explosive increase in TB incidence and prevalence. Several studies suggest that host genetic factors also affect susceptibility and resistance to TB. Genome wide association study (GWAS) provides a way of examining many common variants in different populations to see if any variant is associated with a trait by searching for small variations, called single nucleotide polymorphisms (SNPs). However, it is well known that GWAS alone is insufficient to elucidate the genetic structure of a complex disease and may lead to non conclusive results. In this thesis, we use a post association analysis, which has been suggested as a new paradigm to GWAS, to elucidate and analyze human genetic susceptibility in relation to the infecting MTB by combining association signals from GWAS and available functional and comparative genomics information for human and MTB. We have identified 6 disease associated genes for the admixed

South Africa coloured (SAC) population and 8 disease associated genes for the homogeneous Ghana-Gambia population. We used a graph-based approach to establish a relationship between these different disease associated genes and front-line drug targets in relation to MTB. Furthermore, we performed Gene Ontology (GO) process and pathway enrichment analyses. These yielded sub-networks, enriched processes and pathways that may play critical role in TB immunogenicity and pathogenesis. We also investigated ancestry-specific TB risk in the SAC population and results revealed that the African Khomani (Sub-Kalahari San) ancestry highly contributes to disease risk in this population observed to be highly susceptible to TB. Several studies have been conducted on identifying candidate genes conferring risk susceptibility to TB. However, most of these studies only analysed relationships between these genes and the host system. Here, we have also considered the pathogen system, thus combining host, pathogen and host-pathogen protein-protein functional interactions to examine relationships between host TB susceptibility and pathogenesis. Furthermore we perform functional relationships between identified candidate genes and front-line drug targets based on these functional networks. This may enhance our understanding about TB susceptibility and pathogenesis, and enhance research for TB drug and vaccine development.

Uittreksel

'N geïntegreerde raamwerk modellering vatbaarheid vir tuberkulose in homogene en vermeng bevolkings

Z.Z Gebremariam

Departement Wiskundige Wetenskappe,
Universiteit van Stellenbosch,
Privaatsak X1, Matieland 7602, Suid Afrika.

Tesis: MSc

Maart 2016

Ten spyte van die wye verskeidenheid van anti-tuberkulose dwelms, tuberkulose (TB), wat veroorsaak word deur *Mycobacterium tuberculosis* (MTB), is die tweede grootste aansteeklike siektes ná *Menslike Immuniteitsgebrekvirus* (MIV) of Verworwe Immuniteitsgebreksindroom (VIGS), en een van die grootste oorsake van menslike dood van aansteeklike siektes, veral in Sub-Sahara Afrika. Ongeveer een derde van die wêreld se bevolking is sluimerend besmet is met MTB, waarvan, 10 % vordering aktiewe TB. Struikelblokke in TB beheer sluit in langbehandelingsregimes van meer as 6 maande, weerstand teen die medikasie, 'n gebrek aan 'n doeltreffende entstof en beperkte kennis en onvolledige inligting oor faktore wat die verloop van 'n MTB infeksie teen siektes veroorsaak. Daarbenewens het die vereniging van TB en MIV of vigs ook bevorder al die voorwaardes van 'n plofbare toename in TB voorkoms en die voorkoms. Verskeie studies dui daarop dat gasheer genetiese faktore ook 'n invloed vatbaarheid en weerstand teen TB. Genoom wye assosiasie studie (GWAS) bied 'n manier om die behandeling van baie algemene variante in verskillende bevolkings om te sien of enige variant is wat verband hou met 'n eienskap deur te soek vir klein variasies, genoem enkele nukleotied polimorfismes (SNPs). Dit is egter bekend dat GWAS alleen onvoldoende is om die genetiese struktuur van 'n komplekse siekte toe te lig en kan lei tot nie afdoende resultate. In hierdie tesis, gebruik ons 'n post vereniging analise, wat as 'n nuwe paradigma te GWAS het voorgestel, om toe te lig en te ontleed menslike genetiese vatbaarheid met betrekking tot die besmet MTB deur die kombinasie van assosiasie seine van GWAS en beskikbaar funksionele en vergelykende genomika inligting vir menslike en MTB. Ons het 6 siekte geassosieer gene

vir die venmeng Suid-Afrika gekleurde (SAC) bevolking en 8 siekte geassosieer gene vir die homogene Ghana-Gambië bevolking geïdentifiseer. Ons gebruik 'n grafiek gebaseerde benadering tot 'n verhouding tussen die verskillende siektes wat verband hou gene en tussen siekte gene en front-line dwelm teikens te stel met betrekking tot MTB. Verder het ons uitgevoer Gene Ontologie (GO) proses en pad verryking ontleed. Hierdie opgelewer sub-netwerke, verryk prosesse en roetes wat kritieke rol kan speel in die TB immunogenisiteit en patogenese. Ons ondersoek ook afkoms spesifieke TB risiko in die SAC bevolking en resultate het getoon dat die Afrikaanse Khomani (Sub-Kalahari San) afkoms hoogs dra by tot siekte risiko in hierdie bevolking waargeneem hoogs vatbaar vir TB te wees. Verskeie studies is gedoen op die identifisering van kandidaat gene wat die risiko vatbaarheid vir TB verleen. Maar die meeste van hierdie studies het net ontleed verhoudings tussen hierdie gene en die gasheer stelsel. Hier het ons ook gekyk na die patogeen stelsel, dus die kombinasie van gasheer, patogene en gasheer-patogeen proteïen-proteïen funksionele interaksies om verhoudings tussen gasheer TB vatbaarheid en patogenese is oorweeg. Verder voer ons funksionele verwantskappe tussen geïdentifiseer kandidaat gene en voorste lyn dwelm mikpunte gebaseer op hierdie funksionele netwerke. Dit kan ons begrip oor TB vatbaarheid en patogenese verbeter, en verbeter navorsing vir TB dwelm en entstof ontwikkeling.

Acknowledgements

First of all I give my thanks to my God. Then I would greatly like to thank my supervisor Dr. Gaston K. Mazandu for giving me this chance to work with him and for sacrificing his time and energy patiently to help me reach this point. I would also like to thank AIMS-South Africa with all the staff. This MSc research would not have been possible without the financial support of the Canadian Government via the International Development Research Center (IDRC) through the African Institute for Mathematical Sciences - Next Einstein Initiative (AIMS-NEI). Finally, I would like to express my sincere gratitude to the following people: Dr. Wilfred Ndifon, Dr. Simukai Utete, and Rene January for their valuable comments, encouragement, and help.

Dedications

To my lovely family

Contents

List of Figures

List of Tables

Abreviations

TB Tuberculosis

MTB Mycobacterium tuberculosis

GWAS Genome Wide Association Studies

BioGrid Biological General Repository for Interaction Database

DIP Database of Interacting Proteins

STRING Search Tool for the Retrieval of Interacting Genes

InterPro Integrated documentation resources for protein families, domains and functional sites

SNP Single Nucleotide Polymorphism

DOT Direct Observed Treatment

BCG Bacille Calmette-Guerin

DNA deoxyribonuclic acid

PPIs Protein-Protein Interactions

GOA Gene Ontology Annotations

GO Gene Ontology

DAG Directed Acyclic Graph

MF Molecular Function

CC Cellular Component

BP Biological Process

BMA Best Match Average

SAC South African Coloured

DNA Deoxyribonuclic Acid

WHO World Health Organization

BLAST Basic Local Alignment Search Tool

LAP Local Ancestry Proportion

LAI Local Ancestry Inference

BC Biological Process

BF Biological Function

CC Cellular Component

IC Information Content

YRI Yoruba in Ibadan, Nigeria

KHS Khomani (Sub-Kalahari San)

CEU Caucasian Western European

GIH Gujarati Indian

CHS Southern Han Chinese

Chapter 1

Introduction

Tuberculosis (TB) is an infectious disease caused by a microbial pathogen called Micobacterium tuberculosis (MTB). TB is one of the leading causes of human death from infectious diseases (WHO). Since the discovery of TB, more than 100 years ago, numerous efforts have been made in attempt to control the disease, including the development of anti-TB vaccine and drugs. However, in spite of all these efforts TB remains a public health challenge. According to the World Health Organization (WHO), in 2013 only, 9 million were infected with TB, and 1.5 million people died from TB.

There is a number of factors making TB control implementation difficult. The main one rendering even the front line drugs ineffective is emergence of drug resistant TB, which is caused by inconstant adherence to treatment. For instance, multi-drug resistant TB (MDR-TB) is caused by inappropriate use of anti-TB drugs, in which case an infected individual does not respond to standard treatments. The other factor that contributes to TB control inefficiency is the synergy between TB and HIV. TB is still the leading killer of people living with HIV (WHO).

MTB spreads through air, and anyone exposed to it is at risk. Anyone can get infected with MTB in different ways. There is a number of factors associated with the exposed individual susceptibility to MTB infection and pathogenesis, which depends on the immune status of the infected host and the virulence of the infecting pathogen. A case-control study in West Africa found out the following host-related and environment-related risk factors that play a role in the development of tuberculosis: male sex, HIV infection, smoking, single/widowed/divorced marital status, history of asthma, adult crowding, family history of TB, and renting the house (Lienhardt *et al.*, 2005). Though there is a number of host environmental risk factors, many individuals progress to TB without any identifiable risk factors. This suggests that host genetics variation may influence susceptibility to disease.

One of following three events occur to an individual who is exposed to MTB: resists the infection, becomes infected but shows no clinical signs of the disease, or progresses from mild to severe disease. The occurrence of an outcome depends on the interaction of environmental factors and the genetic make up of both host and pathogen. Maliarik and Iannuzzi (2003) has presented evidence that genetic factors influence the outcome of exposure to MTB and emphasized the host genetic make up pointing out the fact that, of those exposed to MTB in a given similar environment, about 25 percent become infected and from those only 10 percent develop clinical disease.

1.1 Literature review

1.1.1 Mycobacterium strain variation

Mycobacterium tuberculosis, the pathogen which causes TB, was discovered in 1882 by Robert Koch, a German physician and bacteriologist (Arisoa, 2012). MTB belongs to the MTB complex which mainly infects human. Mycobacterium bovis, which predominantly causes tuberculosis in cattle may also infect human (Cousins *et al.*, 2003). MTB is genetically diverse and this genetic diversity may lead to significant phenotypic differences between clinical isolates.

Mutation and recombination lead to DNA sequence variation, which may result in genetic variants, considered as the outcome of natural selection and random genetic drift. These evolutionary forces play an important role in generating bacterial strain variation. This suggests strain variations may be an indication of selective pressure, which possibly alter genes for adaptation to the environment during infection and transmission, influencing pathogenesis and immunity. Thus, these variations are reflected on the genotype and intracellular lifestyle differences between different strains, mapping to the strain's virulence and disease phenotype. MTB exhibits very little genomic sequence diversity compared to other bacteria, and most genetic variability that has been detected is associated with transposable elements and drug resistance phenotype (Kubica *et al.*, 1972; Parish and Brown, 2009). At the whole genomic level, using genomic deletions to type strains and strain lineages there exists 875 strains and these are classified in six main strain lineages (Parish and Brown, 2009).

The global population structure of MTB consists of the six main strain lineages associated with particular geographic regions (Parish and Brown, 2009): the East-Asian strain lineage is most frequent in East Asia, Russia and South Africa. The East-Africa-Indian strain lineage mainly occurs on the Indian subcontinent and in East Africa. The Euro-American strain lineage

dominates in Europe and the Americans. The West-African strain lineages, commonly called M.africanum, occur almost exclusively in West Africa, and the Indo-Oceanic lineage around the Indian Ocean. Most research in the pathogenesis and imunology of TB has been performed using the laboratory strains H37Rv (virulent), H37Ra (attenuated), and the clinical strain CDC1551 and all of these strains belong to the Euro-American strain linage.

The two laboratory MTB strains H37Rv and H37Ra were discovered after the discovery of the H37 strain in 1905, considered to be the ancestor of avirulent and virulent colony forms. The (rough) virulent variant form was designated by H37Rv and the avirulent form by H37Ra (Kubica *et al.*, 1972). H37 strain and the clinical strain CDC1551 are derived from common parental strain as shown in Figure 1.1 (Mazandu, 2010). In this work, we use the clinical strain CDC1551 to analyse TB disease outcome.

Strain CDC1551 or "Oshkosh" was isolated in an outbreak that occured in a rural community on the border of Tennessee Kentucky during the mid-1990s from a 21-year old male clothing factory worker in US (Parish and Brown, 2009; Arisoa, 2012). The CDC1551 strain is highly infectious compared to the virulent strain H37Rv and has more immunoreactivity than H37Rv and other clinical strains (Fleischmann *et al.*, 2002).

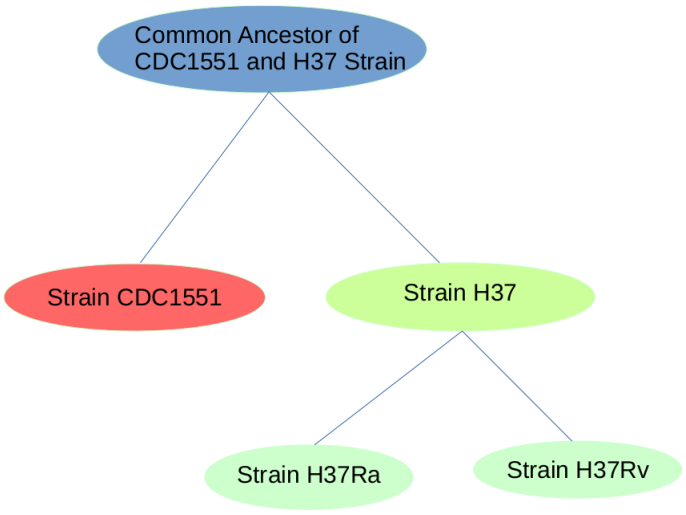


Figure 1.1: Schematic diagram depicting evolution of three MTB strains, H37Ra, H37Rv and CDC1551 (Mazandu, 2010)

1.1.2 Genetics and TB susceptibility

Most of the information about the human body is contained in the genetic matter called chromosome. The chromosome contains tightly coiled strands of DNA, a molecule that encodes the genetic instructions used in the development and functioning of the body. These genetic instruction codes play a major role to the host susceptibility to or outcome of a disease. As a result of the advancement in high-throughout biology technologies, it is now possible to study the whole genome of individuals with and without TB, allowing scientists to identify small genetic regions, which may cause increased TB susceptibility or resistance.

MTB infection occurs in every part of the world. One third of the world population has TB, but only ten percent of those who get infected with MTB will develop clinically active disease. This indicates that TB pathogenesis differs considerably between individuals, leaving a high percentage of individuals infected ($\approx$ 90%) with MTB worldwide non-infectious. This suggests that host susceptibility is an important risk factor with a strong genetic factor determining the outcome of infection. Bellamy (1998) shows that there are more reasons for the development of the disease beyond environmental factors and the pathogen virulence. It has been shown also that, though it is difficult to identify all tuberculosis susceptibility genes, there is convincing evidence that host genetic factors are important in determining the outcome of infection.

In relation to the linkage between TB and host genetics (Maliarik and Iannuzzi, 2003) black populations have higher rates of tuberculosis and are also more likely to develop the more fulminate forms of the disease. Geographically South Africa has the highest occurrence of TB and Western Europe the lowest (Maliarik and Iannuzzi, 2003). Although these racial differences and incidence variation may result from environmental and socioeconomic factors, there is evidence that the difference is strongly influenced by genetic factors.

Stead *et al.* (1990) found that, among over 25 000 tuberculin negative nursing home residents, black subjects were twice as likely to become infected with tuberculosis as white subjects living in the same environment (Bellamy, 1998). The other evidence that shows that genetics factors are important in tuberculosis susceptibility is the twin studies (Schurr, 2011). It has been found that there is a much higher concordance for diseases among homozygous twins than dizygous twins. This suggests that even within ethnic groups, host genetic factors exert a major influence on tuberculosis susceptibility.

Admixed mapping, the process of finding areas of the genome that harbour genetic variants that increase risk of developing a disease, has been used to discover disease susceptibility genes. Due to the existence of population admixture, this process is highly dependent on accuracy of a local ancestry inference (LAI) per individual across their genome (Daya *et al.*, 2014*a*).

When admixture happens between two or more previously isolated population groups, recombination occurs and results in chromosomes that are a chunk of ancestry blocks derived from different source populations. Local ancestry inference (LAI) is used to determine the bounds of these segments and to assign the most probable source ancestries to them. This can be done using statistical techniques given the genetic data of an admixed individual and their source population.

The South African coloured (SAC) population is an example of an admixed population with a mixed ancestry whose genomes consist of five different population. These five populations are Yoruba in Ibadan (YRI : 33 %), Khomani SAN (KHS : 31 %), Caucasian Western European (CEU : 16 %), Gujarati Indian (GIH : 13 %), and Southern Han Chinese (CHE : 7 %) (Chimusa *et al.*, 2013). SAC population has a high incidence of TB and is ideally suited to the discovery of TB susceptibility genetic variants and their probable ethnic origins. Daya *et al.* (2014*a*), in their study on the admixed SAC population, have shown that African ancestry is associated with higher risk of TB infection, whereas European and Asian ancestries are protective.

1.1.3 Pharmacogenetics and anti-TB drugs

After the discovery of TB, significant efforts have been made in drug discovery and administration against TB. The first antibiotic agent for treating TB, streptomycin was discovered in 1943. Subsequently, due to the appearance of an MTB drug resistance strain, other drugs were added. The cure rate increased and antibiotic resistance decreased when the two antituberclosis agents, thiacetazone and paraaminosalicylic were introduced and either of them were given in combination with streptomycin. In 1951 isoniazid was introduced for worldwide use, then pyrazinamide (1952), cycloserine (1952), ethionamide (1956), rifampin (1957), and ethambutol (1962) followed (Keshavjee and Farmer, 2012).

Though new clinically improved drugs were developed, (Maliarik and Iannuzzi, 2003) every new drug led to the selection of mutations conferring resistance to it and using a single drug led to drug resistance. This has resulted in the introduction of multi-drug treatment. Through a series of multicountry clinical trials, led by the British Medical Research Council, a four-drug regimen was recommended for use in patients with newly

diagnosed tuberculosis (Keshavjee and Farmer, 2012). The four core first-line drugs which are particularly used to treat an active TB patient who has not taken any TB drug treatment are isoniazid, rifampicin, pyrazinamide, and ethambutol.

Currently, these four drugs called first- or front-line drugs are used to treat TB through Direct Observed Treatment (DOT), a control strategy implemented by the World Health Organization (WHO) in order to control TB globally. Moreover, there is a vaccine, called Bacille Calmette-Guerin or BCG, used to prevent TB. This vaccine is generally given to infants and children, but in some cases it is only recommended for individuals with specific criteria and in agreement with a TB expert (http://www.cdc.gov/tb/topic/vaccines/).

A drug taken has different responses due to different factors, such as environmental factors and genetic differences. It is likely that individuals who are subjected to the same drug respond differently. This inter-individual drug response variability is one of the challenges in drug development. Generally, drug response can be classified using two criteria: efficacy and toxicity. These interindividual drug response differences are higher among individuals belonging to the same population than within an individual at different times (Ramachandran and Swaminathan, 2012).

Single Nucleotide Polymorphisms (SNPs), the most common genetic variations among people, are key in determining individual's disease susceptibility and drug response. SNPs can occur anywhere along the genome and most SNPs occurring in non-coding or non-regulatory regions of the genome are functionally silent and has no effect. But some SNPs that are found in coding regions may alter protein or gene product structure, leading to disease susceptibility or variation in drug response. The altered genes or proteins that are responsible for drug response are called pharmacogenes.

Pharmacogenetics studies to identify variation in drug response have provided ample examples of causal relations between genotypes and drug response to account for phenotype variations of clinical importance in drug therapy. Pharmacogenetics studies how genetics affects drug response and it revolutionized drug development to personal level based on genetic make up (Alwi, 2005).

1.1.4 Protein-protein interactions

There are 100 trillion cells in the human body and inside each human cell nucleus there are 23 pairs of chromosomes that come from each parent (http://www.thehumangenome.co.uk/THE_HUMAN_GENOME/Primer.html). Each chromosome is made up of two coiled double helix shaped strands of

deoxyribonuclic acid (DNA). These two strands of DNA are composed of a sequence of four bases called nucleotides: adenine (A), guanine (G), cytosine (C), and thymine (T), and the human genome contains approximately 3 billion nucleotides. A gene is a locus or region of DNA that contains code for making proteins which are the building blocks of life.

The human genome project started in 1984 and completed in 2003 has produced the whole human genome (http://www.genome.gov/). The microbial genome sequencing projects yielded complete genome sequence of crucial microbial pathogens of humans, animals and plants (Mazandu and Mulder, 2011*a*). The complete genome of the MTB clinical strain (CDC1551) have been sequenced (Kinsella *et al.*, 2003). As a result of the availability of complete genome sequences of different organisms, the complex properties of living organisms can be studied at the system level. These systems are made up of molecules, which interact among themselves to yield the complex properties of living things.

Proteins or gene products are responsible for most of biological functions in a body. Very often proteins do not work in isolation, they interact directly or indirectly with each other in different processes and pathways to perform their functions. Hence, studying protein-protein interactions (PPIs) is important to understand how proteins function at the systems level. This can help identify pathways and elucidate proteins that play a major role in disease outcome, pathogenesis and drug response.

Protein interactions include physical and functional interactions (Yellaboina *et al.*, 2007). Physical interactions are interactions that involve physical contact between proteins. On the other hand, functional interactions do not necessarily involve direct physical contact, but it refers to the mechanism through which a protein participates in cell functions (Mazandu and Mulder, 2011*a*).

Generally, protein-protein interactions can be detected experimentally or by computational analysis. Functional interactions, can be retrieved from biological knowledge such as coexpression data from microarray analysis. Physical interactions, on the other hand, can be detected using direct experimental techniques, such as pull-down assays, co-immunoprecipitation, or tandem affinity purification coupled to mass spectrometry (Yellaboina *et al.*, 2007).

Protein interactions can be modelled as a network, referred to as a protein-protein functional network or interactoms. The network structure can be represented using mathematical objects called graphs consisting of nodes or vertices and edges (Wagner, 2003). In a protein interaction, graph or network nodes

or vertices are proteins and edges or links represent pairwise interactions or functional relationships within an organism. In this work, integrated protein-protein interaction (PPI) networks are built using PPIs from different data sources. We used functional protein-protein interactions for human and MTB, and human-MTB interaction network.

1.2 Thesis rationale and objectives

Populations of different ancestry may differ in disease susceptibility and drug response. Host and pathogen genetics play a major role in the susceptibility or outcome of the disease in the host. As pointed out previously, some studies have shown that black populations are more susceptible (**?**) and in South Africa the admixed South African Coloured (SAC) population residing in the Western Cape have a high incidence of TB (Chimusa *et al.*, 2014). In addition, it has also been found that there is a positive correlation between African San ancestry and TB susceptibility, and negative correlations with European and Asian ancestries (Daya *et al.*, 2014*b*; Chimusa *et al.*, 2014).

Host susceptibility is an important risk factor to the progression from MTB infection to active disease, but factors that govern this progression are not well understood. We aim at developing a model for analysing human genetic susceptibility in relation to the MTB system and identify genes, biological processes, and potential pathways involved in TB susceptibility. We check whether there is correlation between genome-wide association studies (GWAS) candidate genes and previously identified drug targets by combining association signals from GWAS and available functional and comparative genomic information for humans and MTB. In addition, we predict interactions between humans and the bacterial pathogen influencing TB outcome. This contributes to advancing research for TB drug and vaccine design, and thus might ultimately improve disease diagnosis and prevention.

A graph-based model is developed using protein-protein functional interaction of the organisms. We analyse human genetic susceptibility in relation to the infecting mycobacterium tuberculosis (MTB) system. The correlation between GWAS candidate genes and previously identified drug targets are analysed at system level using protein-protein interaction. The main objective of this research is to elucidate the relationship between human and pathogen in TB disease outcome in association with front-line drug targets at the system level. In this work, high-throughput biological data of the MTB clinical strain (CDC1551) and human genetic data of an admixed South African Coloured (SAC) population and the homogeneous Ghana-Gambia population are used. In summary this project proposes a systems level based model to:

1 Discover possible novel risk genes by combining moderate GWAS signals and relationship between theses genes in association with front-line drug targets;

2 Identify enriched biological processes and pathways in which these risk genes are involved;

3 Discover ethnic differences in disease risk and investigate ancestry-specific disease in the context of the SAC population;

4 Determine human-pathogen interactions influencing different phenotypes or infection outcomes.

1.3 Project outline

The rest of this thesis is organized as follows: In chapter two, we describe different databases used to retrieve the dataset used in this study and discuss scoring schemes of protein-protein functional interactions from each data source. Chapter three provides the details on the integrated scoring scheme used to produce unified networks integrating interactions from different databases. We also discuss about topological properties of networks and network centrality measures which numerically characterize the importance of proteins in and general features of the network. Moreover, we describe clustering method to identify sub-graphs, approaches used to elucidate significant processes and pathways implicated in the disease and for combining effects of different SNPs and ancestry contribution at gene level. Chapter four presents and discusses results obtained by applying different methods. We conclude this thesis in chapter five, summarizing different results obtained and potential future work.

Chapter 2

Exploring different sources of datasets used

There has been an exponential increase in biological data for several model organisms, including human, animals, plants and their crucial microbial pathogens, as results of high-throughput biology technologies and bioinformatics scanning approaches. The use of computational methods and algorithms have enabled the extraction of information concerning complex organization and relatedness of these genomes, including gene content and relationships between these genes, as well as their sizes and other essential features. These biological datasets are stored in public repositories and often freely available to the research community. These include the international Haplotype Map (HapMap) Phase 3 at http://www.hapmap.org, the 1,000 Genomes Project (http://www.1000genomes.org/), the Universal Protein (UniProt) and the European Bioinformatics Institutes (EBI) resources (http://www.ebi.ac.uk/). In this thesis, we use a systems level analyses integrating genotype data, protein-protein interactions, other functional, genomics and pharmaceutical data into a unified framework to identify disease-related genes and enriched processes and pathways in which they are involved.

2.1 Retrieving GWAS and protein target datasets

Genome-wide association study (GWAS) (Jia *et al.*, 2011) examines many common variants in different populations to check whether any variant (genotype) is associated with a trait (phenotype) by searching for small variations, called single nucleotide polymorphisms (SNPs), also referred to as variants or alleles. There have been many successful GWAS (Welter *et al.*, 2014), but detecting variants that have low disease risk is still a challenge.

This is mostly due to the fact that GWAS is a single-marker testing model (Chimusa *et al.*, 2015), which may fail to identify genetic variants with low or moderate risk, which could not meet the standard genome-wide significance threshold of 5.00e-08, thus yielding an increased number of false negatives. In the context of complex diseases, such as TB, where multiple genetic and the environmental factors contribute to the disease outcome through gene-gene and gene-environment interactions (Zhang *et al.*, 2014), it is essential to combine the effects of all SNPs within genes in order to increase the likelihood of identifying disease genes showing weak genetic effects or having strong epistatic effects.

For this study, genetic data are extracted from different literature sources. The genetic data includes the set of human SNPs with their p-values and corresponding genes. SNPs associated with TB, p-values and ancestry contribution for the South African Coloured (SAC) population are taken from Chimusa *et al.* (2014). TB associated SNPs for the homogeneous Ghana-Gambia population are retrieved from Thye *et al.* (2010). We use a post-GWAS meta-analysis techniques to combine the effect of different SNPs within a gene under consideration in order to prioritize essential genes (Crombie and Davies, 2009; Begum *et al.*, 2012).

Moreover, we investigate how front-line drug targets and disease associated genes interact in the system. Human target proteins or enzymes metabolizing TB front-line drugs are collected from the drug bank database (`http://www.drugbank.ca/drugs`) and one common target protein (P11473) is added from the Guide to Pharmacology database at `http://www.guidetopharmacology.org/`.

Table 2.1: TB first line drugs and their target proteins

Target	Gene name	Description	Drug
P11712	CYP2C9	Cytochrome P450 2C9	Rifampin, Isoniazid
P05177	CYP1A2	Cytochrome P450 1A2	Rifampin, Isoniazid, Pyrazinamide
P10632	CYP2C8	Cytochrome P450 2C8	Rifampin, Isoniazid
P08684	CYP3A4	Cytochrome P450 3A4	Rifampin, Isoniazid, Pyrazinamide
P20813	CYP2B6	Cytochrome P450 2B6	Rifampin
P22309	CYP2B6	Cytochrome P450 2B6	Rifampin
P33261	CYP2C19	Cytochrome P450 2C19	Rifampin
P11509	CYP2A6	Cytochrome P450 2A6	Rifampin , Isoniazid
P05181	CYP2E1	Cytochrome P450 2E1	Rifampin , Isoniazid
Q9HB55	CYP3A43	Cytochrome P450 3A43	Rifampin
P20815	CYP3A5	Cytochrome P450 3A5	Rifampin
P24462	CYP3A7	Cytochrome P450 3A7	Rifampin
Q02928	CYP4A11	Cytochrome P450 4A11	Rifampin
O75469	NR1I2	Nuclear receptor subfamily 1 group I member 2	Rifampin
P11473	VDR	Vitamin D3 receptor	Rifampin, Isoniazid, Pyrazinamide, Ethambutol
P10635	CYP2D6	Cytochrome P450 2D6	Isoniazid
P11245	NAT2	Arylamine N-acetyltransferase 2	Isoniazid
P47989	XDH	Xanthine dehydrogenase/oxidase	Pyrazinamide
Q06278	AOX1	Aldehyde oxidase	Pyrazinamide

2.2 Identification of protein-protein functional interactions

In order to produce the protein-protein functional network, data are collected from different databases. In addition to the PPI data directly downloaded from different freely available online databases, protein-protein functional interactions are also predicted from protein sequence similarity and conserved protein signature matches.

The protein-protein interaction datasets are downloaded from STRING, BioGRID, DIP and IntAct databases. Additional protein functional interactions are predicted using protein sequence similarity and conserved protein signature matches (shared domain) data derived from UniProt and InterPro databases, respectively (Table 2.2). Unless specified explicitly, throughout this thesis sequence data refers to sequence similarity and shared domain data.

STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) is a freely accessible online database of known and predicted protein interactions (Szklarczyk *et al.*, 2015). Its data is derived from experiments, public literature collections, and computational prediction methods based on domain fusion, gene fusion, gene neighbourhood, homology and phylogenetic profiling. It provides full coverage and accesses to experiential and predicted interactions between proteins in more than 1000 organisms.

BioGRID (Biological General Repository for Interaction Datasets) is a freely

accessible database of physical and genetic interactions. Interactions stored in this database are compiled through comprehensive curation efforts. The current version (3.4.129) holds over 830,000 interactions curated from both high-throughput datasets and individual focused studies, as derived from over 55,000 publications in the primary literature (http://www.thebiogrid.org).

DIP (Database of Interacting Proteins) stores experimentally determined protein interactions. It combines information from a variety of sources to generate a single, consistent set of protein-protein interactions. The data stored within the DIP database were curated, both, manually by expert curators and also automatically using computational approaches that utilize the knowledge about the protein-protein interaction networks extracted from the most reliable, core subset of the DIP data. Currently the database contains 27883 proteins, 749 organisms, and 79646 interactions (http://dip.doe-mbi.ucla.edu).

IntAct is a freely available online database that contains molecular interaction data derived from literature curation or direct data depositions by expert curators. It also contains valuable tools that can be used to search for, analyze and graphically display protein interaction data from a wide variety of species. IntAct currently contains 82491 proteins, 351399 interactions of different species of which 39.1% is for human, these data get updated whenever a new molecular interaction has been submitted (http://www.ebi.ac.uk/intact).

InterPro is a public data resource for protein families, domains or protein signatures and functional sites (http://www.ebi.ac.uk/interpro). InterPro provides functional analysis of protein sequences by classifying them in to families and predicting domains and important sites. InterPro integrates signatures from different databases in to a single searchable resource and uses these signatures to classify proteins.

UniProt (Universal Protein Resource) is a freely accessible database for protein sequence data and functional annotation data of proteins (http://www.uniprot.org). It is composed of four components; the UniProt Knowledgebase (UniProtKB), the UniProt Reference Clusters (UniRef), the UniProt Archive (UniParc), and the UniProt Metagenomic and Environmental Sequences (UniMES) databases.

Table 2.2: Data source databases

Database	Description	Data type	Reference
STRING	Search Tool for the Retrieval of Interacting Genes/Proteins	Pretreated protein interaction	http://string-db.org
BioGRID	Biological General Repository for Interaction Database	Physical and genetic interactions	http://www.thebiogrid.org
DIP	Database of Interacting Proteins	Protein interactions	http://dip.doe-mbi.ucla.edu
IntAct	Molecular Interactions	Experimentally determined protein interactions	http://www.ebi.ac.uk/intact
UniProt	Universal Protein Resource	Protein sequence data	http://www.uniprot.org
InterPro	Integrated documentation resources for protein families, domains and functional sites	Protein signature or shared domain	http://www.ebi.ac.uk/interpro
Drug Bank	A unique bioinformatics and cheminformatics resource that combines detailed drug data with comprehensive drug target	Drugs and drug targets	http://www.drugbank.ca/drugs
GOA	Gene Ontology annotation	Protein annotations	http://geneontology.org/
IUPHAR/BPS	Guide to PHARMACOLOGY	Drugs and Drug targets	http://www.guidetopharmacology.org/
KEGG	Kyoto Encyclopedia of Genes and Genomes	Genomes, biological pathways, diseases, drugs	http://www.kegg.jp/

2.3 Scoring protein-protein functional interactions

As pointed out previously, functional interactions are retrieved from multiple sources, including the STRING database (Szklarczyk *et al.*, 2015) and protein-protein interaction datasets described in Table 2.2. We mapped different protein identifiers from different sources to UniProt accession numbers for human (homo sapiens) and MTB strain CDC1551 downloaded from the UniProt database (UniProt Consortium, 2015). Each dataset source produces its protein-protein functional network configuration of its own, which raises the issue of accuracy of each network configuration, especially as these interactions are retrieved from inaccurate and noisy data produced by high-throughput biology experiments. This issue is alleviated by assigning a confidence or reliability score to each functional interaction, which represents the likelihood of the occurrence of the interaction under consideration and quantifies our confidence level in this functional interaction.

2.3.1 Scoring interactions from sequence data

For this we downloaded protein sequence data for human (reviewed) and Mycobacterium tuberculosis clinical strain CDC1551 from the UniProt database using their taxonomy numbers 9606 and 83331, respectively. A FASTA (canonical) file, this is the file containing protein sequences for the organism or strain under consideration is downloaded. For each organism, we also downloaded a tab-separated file, which is a protein reference file for the organism or strain under consideration. Similarly, protein signature or shared domain data is downloaded from the InterPro database.

2.3.1.1 Computing sequence similarity based confidence score

Basic Local Alignment Search Tool (Altschul *et al.*, 1990, 1997), referred to as BLAST, is used to retrieve sequence similarity data used to derive protein-protein functional interactions. BLAST is an algorithm for sequence similarity searching and sequence comparison. It aligns two sequences and outputs alignments which produce high alignment bit score and calculates the statistical significance of matches. The availability of genome sequences has enabled analyses of genes and their products within an organism and their comparisons across different organisms through the identification of similarity between sequences. Protein-protein functional interactions are predicted from these sequence similarity scores by assuming that two protein sequences which are significantly similar are evolutionary linked and might thus share similar functions at the molecular levels or participate to the same biological process or act in the same pathway without direct physical contact.

Using protein sequence data, a local Blast database for each organism is created and each organism proteome is Blasted against itself. The Blast result for each organism is cleaned, and a file containing only protein pairs and Blast scores is produced. The link reliability or confidence score of a protein pair (p, q), $\mathcal{S}_{\text{seq}}(p, q)$, is calculated using alignment bit scores produced from BLAST as suggested in (Mazandu and Mulder, 2011*b*) and given by:

$$\mathcal{S}_{\text{seq}}(p, q) = \frac{\mathbb{A}(p, q)}{2 \times \max\{S(p, p), S(q, q)\}} \tag{2.3.1}$$

where $\mathbb{A}(p, q) = S(p, q) + S(q, p)$ is the bit score obtained by aligning the protein sequence q against the protein sequence p and p against q and quantifies their conserved biological features during evolution (Bastien *et al.*, 2005; Bastien and Maréchal, 2008) with $S(p, q)$ the BLAST bit score resulting from aligning the protein sequence q against protein sequence p. This bit score $\mathbb{A}(p, q)$ reflects the amount of information shared by these two protein sequences due to their common origin and parallel evolution under similar selective pressure. This shared information is normalized by dividing it with the maximum possible relative entropy produced by aligning these protein sequences, which is $2 \times \max\{S(p, p), S(q, q)\}$, in order to correct the bias which may be yielded by an unpredictable increase of bit score. Thus, formula (2.3.1) produced normalized confidence score (value range between 0 and 1) which only depends on the two protein sequences under consideration and measures how the protein sequence 'p' is able to predict the protein sequence 'q' and vice versa, and includes the case where no similarity is identified between two protein sequences, producing the confidence score of 0.

2.3.1.2 Computing shared domain-based confidence score

The InterPro signature data downloaded from the InterPro database is processed with the protein reference file, the tab separated file, and we produced a file that contains proteins with their InterPro signature for each organism. Finally, we computed the confidence or reliability score, $\mathcal{S}_{\text{dom}}(p,q)$, between pairwise proteins p and q, using an information-theory based model suggested in (Mazandu and Mulder, 2011*b*), as follows:

$$\mathcal{S}_{\text{dom}}(p,q) = 1 - H_2(h)\,/bit \tag{2.3.2}$$

with $H_2(h)$ the binary entropy function quantifying the uncertainty associated with the number s of common InterPro signatures hits, given by

$$H_2(h) = -h\log_2(h) - (1-h)\log_2(1-h) \tag{2.3.3}$$

where the confidence level function $h \equiv h(s,\sigma,\beta)$ is given by

$$h(s,\sigma,\beta) = \phi\left(\frac{s^{\beta}}{\sigma}\right) \tag{2.3.4}$$

with ϕ the cumulative probability distribution of the standard normal distribution defined as follows:

$$\phi(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}\exp\left(-\frac{z^2}{2}\right)dz \tag{2.3.5}$$

σ the standard deviation and $\beta \geq 0.5$ the calibration control parameter, reflecting the impact of the confidence level for the InterPro signature dataset.

2.3.2 Scoring interactions from other datasets

Other human and MTB protein-protein functional interactions are mainly extracted from the STRING database (Szklarczyk *et al.*, 2015) using the organism taxonomy number, accessed on 25 October, 2015. This database contains predicted and known protein-protein interactions derived from genomic context, text mining, information from pathway databases and biological experiments. These protein-protein functional interactions are used with their confidence or reliability scores as defined by the STRING system. For the human functional network, more protein functional interaction data are also derived from the Biological General Repository for Interaction Datasets (BioGRID) (Chatr-Aryamontri *et al.*, 2013), expert-curated and experimentally determined PPI from the Database of Interacting Proteins (DIP) (Salwinski *et al.*, 2004) and the IntAct database. These interactions are assumed to be of reasonable quality and a fixed confidence score of 0.85 is assigned to each predicted interactions.

2.3.3 Scoring human-MTB protein-protein functional interactions

Human-MTB protein-protein functional interactions are derived from manual curation of the literature and predicted using the interologs model based on human and MTB functional networks. Interologs are conserved interactions between a pair of proteins which have interacting orthologs in another organism. The interaction between proteins X and Y in one species is referred to as interologs of the interaction X' and Y' in another species if X' and Y' are orthologs (have common ancestor) of X and Y respectively (Arisoa, 2012). In order to infer these interologs, interaction datasets are collected from manually verified interactions between human and bacterial proteins from the Host-Pathogen Interaction database (HPIDB) (Kumar and Nanduri, 2010) and the Pathosystems Resource Integration Center (PATRIC) (Snyder *et al.*, 2007), and intra-species interacting pairwise proteins from DIP, MINT and IntAct. Protein orthologs were retrieved Ensembl BioMart at http://www.ensembl.org/biomart/ and interologs predicted based on the premise that orthologs of interacting proteins also interact. These interactions are assigned a score of 0.60 as they are assumed to be of high quality. Note that these functional interactions are complemented by functional interactions from sequence data, more specifically interactions predicted from protein sequence similarity and shared domains between proteins from the InterPro database.

2.4 Gene Ontology annotation and pathway datasets

Cells are functional units of life and each protein or gene product in a cell contributes to different biological functions by collaborating in pathways and processes, and interacting with the cellular environment in order to promote the cell's growth and function (Mazandu and Mulder, 2011*a*; Mulder *et al.*, 2014). Generally, proteins have six primary functions in our body; repair and maintenance, energy production, hormone creation, chemical reaction enzymes, and transportation and storage of molecules. Thus, performing functional analyses of protein sets is useful to understand the biological phenomena underlying a given protein set by identifying enriched processes and pathways in which these proteins are involved. In case where these proteins or genes are implicated in the disease outcome, this analysis may enable the identification of essential processes and pathways involved in the disease. Understanding these processes and pathways can contribute to the development of effective therapy which considers underlying causes of disease and minimizes side effects. Biological process and pathway information are found in bioinformatics resources, such as the Gene Ontology (GO) and Gene Ontology Annotation (GOA) (Huntley *et al.*, 2015), and the Kyoto

Encyclopedia of Genes and Genomes (Kanehisa and Goto, 2000), referred to as KEGG, databases.

GO is designed as a directed acyclic graph (DAG) in which each node is a biological term describing genes and proteins in any organism, and produces a well adapted platform to computationally process data at the functional level (Mazandu and Mulder, 2013*a*). GO has been widely adopted and successfully deployed in several biological and biomedical applications, ranging from theoretical to experimental and computational biology (Mazandu and Mulder, 2013*b*). Currently, more than 4.2×10^7 proteins (see GOA UniProt version 152 at http://www.ebi.ac.uk/GOA/uniprot_release, released on 13 February, 2016) are already annotated with GO terms and this dataset is integrated into the GOA database under the GOA-UniProt project (Huntley *et al.*, 2015), mapping different annotated proteins from UniProt knowledge-base (UniProtKB) to their GO annotations. It has been suggested that incorporating the GO structure in GO annotation-based protein analyses has significantly contributed to the improved outcomes of protein functional analyses (Mazandu and Mulder, 2013*b*,*a*). Thus, several GO semantic similarity measures (Mazandu and Mulder, 2013*b*; Mazandu *et al.*, 2015) have been proposed in recent years and have enabled the integration of biological knowledge embedded in the GO structure into different biological analyses.

In this study, we use the GO biological process data based semantic similarity model built on the GO-universal metric (Mazandu and Mulder, 2012*a*) to identify enriched biological processes for a given set of proteins. The complete set of GO data and protein-GO term associations are downloaded from the GO and GOA (version 148, released on 11 November, 2015) databases, accessed on the 16th November, 2015. Finally, for pathway enrichment analyses, literature was mined and pathway dataset was extracted from KEGG. The Protein Interaction Network Viewer tool (PINV) is used to visualize interactions of interest.

Chapter 3

Integrative model for analyzing susceptibility to tuberculosis

As pointed out previously, GWAS as a single marker-based model is very limited and yields a number of false negatives as several markers often fall below the cut-off level. Genes often interact to perform some biological function in a cell which can lead to a specific phenotype. It is likely that many markers or genes singularly with low or moderate risk may interact to produce a significant combined effect for complex diseases, such as TB. This suggests that analyzing genes at the systems level based on protein-protein functional interaction networks may help understand better the etiology of a disease and elucidate genetic factors influencing the disease pathogenesis. Mathematically, protein-protein functional interaction networks are represented by undirected graphs with proteins as nodes and functional interactions (connections) between proteins as edges or links. In this study, we use a model that integrates GWAS data from a given population (admixed or homogeneous), with the human and MTB protein-protein interaction network to predict sets of genes that interact to investigate predisposition to a disease for individuals in a given population.

In this chapter, we provided details on the integrated scoring scheme used to produce unified networks integrating interaction datasets from a variety of sources. We also discuss network centrality measures to score the relevance of proteins in the network and examine other topological properties of the biological networks. These biological networks are often modular in nature, indicating that some proteins in the network are more essential or central than others. Finally, we describe a sub-graph finding (clustering) algorithm that is used to identify key sub-graphs associated with disease risk and methods used to elucidate the most significant processes and pathways implicated in the disease and for combining effects of different SNPs and ancestry contribution within each gene. Throughout this chapter, $G = (N, L)$ represents the undirected graph where N is the set of interacting proteins (nodes) and L is the

set of functional interactions (links or connections) between proteins in the system.

3.1 Building unified networks and centrality measures

Protein-protein interaction datasets are derived from different sources. Depending on the source, these interactions are scored as these datasets are often noisy and unreliable. For a given interaction, this confidence score is simply the probability that the interaction occurs and this depends on data source and technology used. Thus, each interaction dataset produces a graph with weighted relationships between each protein pair. Integrating these different datasets into a unified network increase coverage and reduce the likelihood of a false negative. Thus, an integrative scoring scheme is necessary to produce an integrated protein-protein functional interaction data from a variety of sources to generate a complete interaction network. In this thesis, we generate the human, MTB strain CDC1551 and human-MTB protein-protein functional interaction networks as summarized in Figure 3.1.

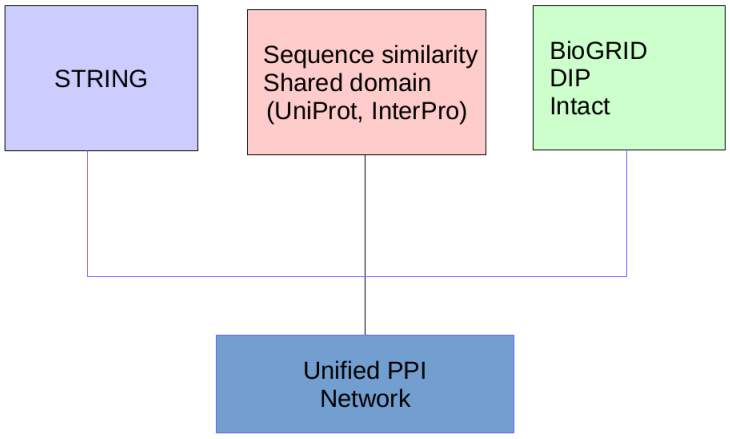


Figure 3.1: **Summary of different protein-protein functional interaction datasets.** Integration of protein-protein functional interactions derived from different sources into a unified functional network

.

3.1.1 Integrative interaction scoring function and effectiveness

The reliability or confidence score of an interaction between proteins p and q quantifies how reliable is this specific interaction and represents the probability that this interaction occurs. Assume that n different sources were used to predict this interaction and let $\overline{E_{pq}}$ be an event indicating that the functional interaction between proteins p and q could not be inferred from any of these n sources under consideration, that is:

$$\overline{E_{pq}} = \bigcap_{j=1}^{n} \overline{E_{pq}^{j}} \tag{3.1.1}$$

with $\overline{E_{pq}^{s}}$ the event indicating that the functional interaction could not be retrieved using the source s. Under the assumption that sources are independent, the probability $\mathbb{P}\left(\overline{E_{pq}}\right)$ of the event $\overline{E_{pq}}$ is given by:

$$\begin{aligned} \mathbb{P}\left(\overline{E_{pq}}\right) &= \mathbb{P}\left(\bigcap_{j=1}^{n} \overline{E_{pq}^{j}}\right) \\ &= \prod_{j=1}^{n} \mathbb{P}\left(\overline{E_{pq}^{j}}\right) \\ &= \prod_{j=1}^{n} \left(1 - \mathbb{P}\left(E_{pq}^{j}\right)\right) \end{aligned} \tag{3.1.2}$$

where E_{pq}^{j} is the event indicating that the functional interaction is retrieved using the source s and thus $\mathbb{P}\left(E_{pq}^{j}\right) = s_{pq}^{j}$ with s_{pq}^{j} the confidence score of a functional association between p and q predicted using the source j. Thus, the combined confidence score $\mathcal{S}_{pq}$ for interacting proteins p and q, which is the probability of the event E_{pq}, which indicates that the functional interaction between proteins p and q can be inferred from at least one of the sources, contrary to $\overline{E_{pq}}$, is given by:

$$\begin{aligned} \mathcal{S}_{pq} &= \mathbb{P}\left(E_{pq}\right) \\ &= 1 - \mathbb{P}\left(\overline{E_{pq}}\right) \\ &= 1 - \prod_{j=1}^{n} \left(1 - \mathbb{P}\left(E_{pq}^{j}\right)\right). \end{aligned} \tag{3.1.3}$$

It follows that:

$$\mathcal{S}_{pq} = 1 - \prod_{j=1}^{n} \left(1 - s_{pq}^{j}\right) \tag{3.1.4}$$

under the assumption of independency.

Finally, one may choose to use other scoring functions, such as minimum (min), maximum (max) and average (mean) of different confidence scores, however, these produce biased combined or unified scores. Let us assume that out of $n = 5$ different data sources for human, the functional interaction between proteins p and q was predicted from 2 sources out of 5 with confidence scores of 0.200 and 0.130. So, for any other source, the confidence score is assumed to be 0, and it follows that:
– Using the min function, we get $\mathcal{S}_{pq} = \min\{0.00, 0.00, 0.00, 0.200, 0.130\}$, which implies that $\mathcal{S}_{pq} = 0.00$, indicating that the confidence score is 0 and this interaction will be ignored in different analyses whereas it was predicted by two different sources.
– Using max and mean, the combined confidence score, $\mathcal{S}_{pq}$, is equal to 0.200 and 0.066, respectively. The max function does not reflect the fact that the functional interaction was predicted from two different sources and the mean function reduces our confidence level. Intuitively, as this interaction was predicted by two different sources, one expects its confidence level to increase, but instead it is decreasing. This suggests that these scoring functions are not in agreement with what can be expected and show biases by underestimating combined interaction scores in the final network. On the other hand, using the scoring function in equation (3.1.4) as used in this study, we have $\mathcal{S}_{pq} = 0.304$, showing more realistic combined confidence score compared to other scoring functions, and is in agreement with what one would expect.

In the context of this study, the combined confidence score values are categorized into three different confidence levels: low (score < 0.3), medium ($0.3 \leq$ score ≤ 0.7), and high (score > 0.7). Interactions with scores lower than 0.3 are considered to be low confidence, interactions with scores range from 0.3 to 0.7 are classified as medium scored, and interactions with score greater than 0.7 are said to be high confidence interactions. In order to minimize the number of false positives and produce a reliable unified protein-protein functional interaction networks, we only consider interactions in the medium or high confidence categories and those which are predicted by at least two different sources.

3.1.2 Network centrality measures

Network centrality measures are used to numerically characterize the importance of proteins in the network. We use these measures to examine the criticality of a protein in a given network. These centrality measures include degree or connectivity, betweenness, closeness, and eigenvector centrality. Note that a path between two proteins $p_0, p_\imath \in N$ in a protein-protein functional network $G(N, L)$ is a sequence of adjacent proteins $p_0, p_1, \ldots, p_{\imath-1}, p_\imath \in N$ leading from p_0 to $p_\imath$. The number of edges in a path is called path length measuring the

distance between two proteins in networks. The mean or characteristic path length of a graph G is the average path length of shortest paths between all pairs of proteins (Barabasi and Oltvai, 2004), and when a network has a low mean path length, the network is said to satisfy*a small world property*.

3.1.2.1 Degree centrality

Given a protein v in a network, the degree centrality $C_d(v) = deg(v)$ of v is defined as the number of other proteins it interacts with. The degree of a protein node v tells us the number of links the protein has to other proteins and it is given by Mazandu and Mulder (2011*a*)

$$deg(v) = \sum_{u \in N} \delta(v, u), \tag{3.1.5}$$

where

$$\delta(v,u) = \begin{cases} 1 & \text{if the protein } u \text{ is functionally linked to the protein } v, \\ 0 & \text{otherwise.} \end{cases}$$

$deg(v)$ is the number of proteins interacting with v. Degree centrality of a node is used to characterize the importance of the node in the network. A protein which has many functional connections is said to be a key protein as it may has contributions to many important processes in the system (Mazandu and Mulder, 2011*a*).

3.1.2.2 Closeness centrality

The closeness centrality $C_c(v)$ of a protein v in a connected graph G is the inverse of its status, that is the inverse of the average shortest distance to all other proteins connected to it. For a given network, the normalized closeness centrality of a protein v in the network is given by Mazandu and Mulder (2011*a*):

$$C_c(v) = \frac{n_c - 1}{(L_c - 1) \times S(v)}, \tag{3.1.6}$$

where $|N_v| = n_c$ is the number of proteins in the connected component of the graph containing the protein node, L_c is the number of functional links in the connected component, and $S(v)$ is the status of v relative to its connected component, which is the average shortest distance to all other proteins connected to v, given by:

$$S(v) = \frac{1}{n_c - 1} \sum_{u \in N_v} \gamma_{vu}, \tag{3.1.7}$$

where N_v is the set of proteins interacting with v, $n_c = |N_v|$ is the number of nodes in N_v and γ_{vu} is the shortest path length between v and u in the

network.

The closeness measure indicates the importance of a protein in such a way that an important protein node is typically close to and can communicate quickly with, the other proteins in the network.

3.1.2.3 Betweenness centrality

This centrality measure reveals the importance of a protein in the system when it is not highly connected. It shows the importance of a protein for transmission of information between other two node proteins in the network. This measure can be used to identify non-hub proteins that are important in the system or to classify hubs according to their positions in the network (Bonacich, 2007).

The betweenness centrality of the protein v in a functional network is defined as the sum of the ratio of the number of shortest paths passing through the protein to the total number of shortest paths in the functional network. The betweenness centrality, $C_b(v)$, of the protein v is given by

$$C_b(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}, \tag{3.1.8}$$

where s and t are proteins in N different from v, $\sigma_{st}(v)$ denotes the number of shortest paths from protein s to protein t passing through v, and σ_{st} is the number of shortest paths from s to t.

The normalized betweenness value ($0 \leq C_b(v) \leq 1$) of the protein v is obtained by dividing by the number of protein pairs excluding v

$$C_b(v) = \frac{2}{(n-1)(n-2)} \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}, \tag{3.1.9}$$

$n = |N|$ is the number of proteins in the network.

3.1.2.4 Eigenvector centrality

The eigenvector centrality measure assigns weight or importance to proteins based on their functional connections. Proteins functionally connected to important proteins are more influential than proteins connected to less important proteins (Mazandu and Mulder, 2011*a*). Let C_i be the contribution or weight of a protein i in the network. (Bonacich, 2007) Let $A = (a_{ij})_{1 \leq i,j \leq n}$ be the adjacency matrix of the network, where $a_{ij} = 1$ when proteins i and j are functionally connected and $a_{ij} = 0$ if they are not connected. The eigenvector

centrality C_i of a protein i is defined as

$$C_i = \frac{1}{\lambda} \sum_{j \in N} a_{ij} C_j, \tag{3.1.10}$$

where λ is the largest eigenvalue of A. That is, $AC = \lambda C$ where $C = (c_1, \ldots c_n)^T$ is the eigenvector of A associated with the eigenvalue λ, the transpose of a vector that defines the contribution of each protein in the network.

In this measure the importance of a protein depends on the quality of its neighbours rather than the number of its interaction. A protein with high degree centrality may has less eigenvector centrality value than a protein connected to few, but important nodes and it may has high contribution to the functioning and survival of the system.

3.1.3 Degree Distribution of proteins in the functional network

The connectivity or degree distribution, $\mathbb{P}(k)$, is the probability that a selected protein is connected to k proteins (of degree k) in the network. $\mathbb{P}(k)$ is approximated by the frequency of occurrence of protein of degree k in the network, given by

$$\mathbb{P}(k) = \frac{n_k}{n} \tag{3.1.11}$$

where n_k is the number of nodes with degree k, $k = 1, 2, 3, \ldots$ and n the total number of nodes in the network. Networks can be categorized in two classes based on their degree distribution (Zhang, 2009), namely random and scale-free networks.

Random networks have a degree distribution, $\mathbb{P}(k)$, that follows a poisson distribution. This network is homogeneous, in that most proteins have roughly the same number of links. These networks are characterized by a $\mathbb{P}(k)$ that peaks at an average $< k >$ and decays exponentially for large k (Albert *et al.*, 2000). Scale-free networks, in contrast, are heterogeneous. That is, they have few nodes with many interactions and many nodes with few interactions. The number of links of a given node approximates a power law model, for which $\mathbb{P}(k)$ decays as a power-law, that is

$$\mathbb{P}(k) \sim k^{-\gamma}, \tag{3.1.12}$$

where the power exponent γ is a constant characteristic of the network. This distribution is independent of the number of nodes, thus the networks are said to be *scale free*. Albert *et al.* (2000) indicated that the probability that a protein has a large number of links, i.e., the degree larger than the mean

degree of all proteins in the network ($k >><< k >$), is small in scale-free networks and highly connected proteins (hubs) are essential and play significant role in scale-free networks.

Recent studies have shown that many biological networks are scale-free. Thus their degree distribution approximates power law, indicating that most proteins participate in only a few interactions, while a small set of hubs participate in many interactions (Zhang, 2009; Wagner, 2003; Albert *et al.*, 2000). It is observed that networks that have scale-free topological properties have a higher error tolerance or resistance to random node failure. The ability of their nodes to communicate being unaffected even by unrealistically high failure rates (Albert *et al.*, 2000). But they are very vulnerable to the selection and removal of the few nodes that play a major role in maintaining the network's connectivity. Albert *et al.* (2000) demonstrated the difference between the random and scale-free networks as shown below in Figure 3.2. Both networks have 130

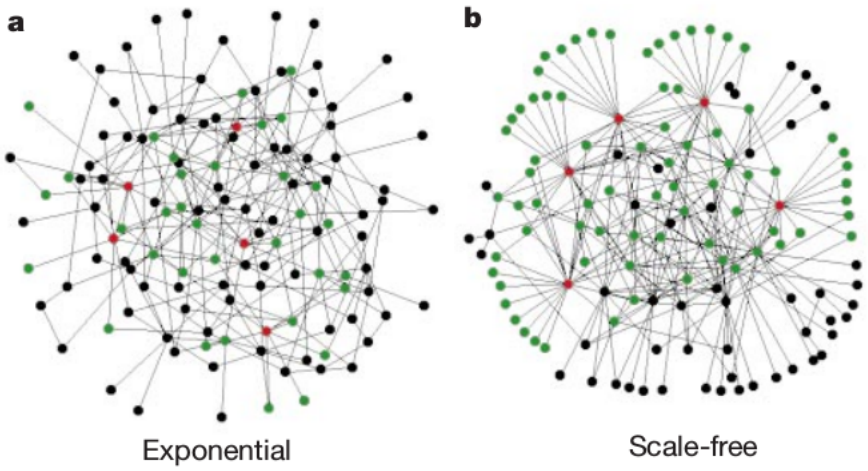


Figure 3.2: Graphical illustration of the difference between an exponential and a scale-free network (Albert *et al.*, 2000)

nodes and 215 links ($< k >= 3.3$). The five red nodes are the nodes with the highest number of links and the green nodes are their first neighbours. In the exponential network only 27% of the nodes are reached by the the five red nodes, but in the scale-free network 60% of the nodes are reached showing the importance of the most connected nodes. This shows that in these networks, nodes that are highly connected are important in the maintenance of the system.

3.1.4 Identifying network key proteins

The three degree centrality measures; degree, betweenness, and closeness centrality values are used to determine the key proteins or hubs in the functional network, in which case, these centrality values reach a certain threshold. For degree centrality measure, the threshold is mean degree centrality, mean($\mathcal{C}_d$). For betweenness measure, the threshold is the total number of shortest paths expected in the network, which is given by $n \times \pi$, where n is the number of nodes in the network and π is the average shortest path length. The cut-off for closeness centrality measure is $\frac{1}{\pi}$, the inverse of the average shortest path length. A protein v in a network is a key protein if $\mathcal{C}_d(v) \geq \text{mean}(\mathcal{C}_d)$, $\mathcal{C}_b \geq n \times \pi$ and $\mathcal{C}_c(v) \geq \frac{1}{\pi}$.

3.2 Network proteins clustering

Classifying proteins in human and pathogen functional networks can help understand biological mechanisms underlying the system in the study, we identify the densely connected sub-networks or communities of functional network. We perform functional analysis of clusters or communities containing key proteins and important candidate proteins of the two population groups under consideration. This may enable the discovery of hidden information in the complex network, such as the difference between the admixed and homogeneous population regarding susceptibility to disease. There exist different community detection methods for identifying community structures in complex networks with the quality of the communities being measured by their 'modularity' (Steinhaeuser and Chawla, 2010; Blondel *et al.*, 2008). Here we used the Blondel *et al.* (2008) method, which was shown to perform better than other methods (Blondel *et al.*, 2008), to partition different functional networks in order to detect connected sub-networks/communities.

3.3 Combining p-values at gene level

GWAS is not sufficient in itself to reveal the genetic structure of complex diseases considering the multiple genetic and environmental factors contributing to development of a complex disease, such as TB. Thus, considering the combined effects of genes by detecting genetic signals beyond single gene polymorphisms provides increased potential to fully characterize the susceptible genes and the genetic structure of complex diseases.

In the unified human network, GWAS SNPs and their p-values retrieved from literature are mapped to their corresponding genes in the network. We combined SNP effects at gene level using their p-values from the genome wide association studies (GWAS). We used the Liptak-Stouffer Test to assess the

significance of genes disease association and to discover possible novel risk genes in the network (Chimusa *et al.*, 2015). Stouffer's test changes the p-values in to standard normal distribution and performs a z-test. Under assumption that p_i $(i = 1, 2, \ldots, n)$ are independent p-values of SNPs associated with a gene and that these p-values follow a standard normal distribution, let ϕ be the cumulative distribution function of standard normal distribution $(N(0, 1)$, i.e, with mean 0 and standard deviation 1) defined in equation 2.3.5, and $q_i = \phi^{-1}(1-p_i)$ quantiles derived from $1-p_i$. Then, the Liptak's combined p-value statistic, C^p, is given by

$$C^p = \frac{\sum_{i=1}^{n} q_i}{\sqrt{n}}$$

which follows the standard normal distribution $N(0, 1)$ and the related combined p-value is $p^* = 1 - \phi\,(C^p)$.

3.4 Combining local ancestry at gene level in admixed population

For the admixed South African Coloured (SAC) population, we estimated the local ancestry contribution for each of the source population at gene level. Then we investigated ancestry-specific disease risk in the population to identify the ancestries that are correlated to the disease susceptibility. We combined locus specific ancestry contribution of SNPs within a gene to estimate gene specific ancestry contribution. The gene-specific ancestry from the specific ancestral population is estimated using the maximum likelihood approach.

For a given gene j. Let ϕ_{mk} be the average locus-specific ancestry of SNP $m \in \{1, 2, \ldots, n\}$ from the k^{th} ancestral population associated with gene $j, (j = 1, 2, \ldots, J)$. Consider each $\phi_{1k}, \phi_{2k}, \ldots, \phi_{nk}$ as n independent and identically distributed observations. Assume that each observation ϕ_{mk} can be approximated as a normal distribution under natural drift with mean 0 and empirical variance V_{mk} and inverse variance W_{mk}. Let μ_{jk} be the unknown true gene-specific ancestry of gene j from the k^{th} ancestral population. We estimate the maximum average likelihood of the observations (Chimusa *et al.*, 2015), $\hat{\mu}_{jk}$ by solving the following system of equation equation

$$\frac{\partial L}{\partial \mu_{jk}}(\mu_{jk}, V_{mk}; \phi_{1k}, \phi_{2k}, \ldots, \phi_{nk}) = 0$$

with $L(\mu_{jk}, V_{mk}; \phi_{1k}, \phi_{2k}, \ldots, \phi_{nk})$ the likelihood function given by

$$L(\mu_{jk}, V_{mk}; \phi_{1k}, \phi_{2k}, \ldots, \phi_{nk}) = \prod_{m=1}^{n} \frac{1}{\sqrt{2\pi V_{mk}}} \exp\left(-\frac{(\phi_{mk} - \mu_{jk})^2}{2V_{mk}}\right)$$

and the maximum likelihood estimate, $\hat{\mu}_{jk}$ is a weighted sum of the SNPs locus-specific ancestry using precision (inverse-variance) weights divided by the sum of the precision, given by

$$\hat{\mu}_{jk} = \frac{\sum_{m=1}^{n} W_{mk}\phi_{mk}}{\sum_{m=1}^{n} W_{mk}} \tag{3.4.1}$$

with variance, $\hat{V}_{jk}$, given by:

$$\hat{V}_{jk} = \frac{1}{\sum_{m=1}^{n} W_{mk}}. \tag{3.4.2}$$

Given a gene or protein j in the network, we have the gene specific ancestry contribution estimate $\hat{\mu}_{jk}$ $(k = 1, \ldots, 5)$ with variance $\hat{V}_{jk}$ of each population ancestry in SAC. To identify ancestries showing unusual gene specific ancestry contributions, we start by ranking different ancestors with a gene using their contribution scores. The ranked ancestor list is then used to compute a Pearson χ^2 score (Mazandu and Mulder, 2011*b*) or Pearson's cumulative test statistic for each ancestor subset, reflecting the tendency of ancestors in a particular set to occur towards the extremes of the list. The Pearson χ^2 score of a subset containing ancestors in the ranked list re-indexed from m to ℓ, $\chi^2_{\mathrm{P}}(m, \ell)$, is given by:

$$\chi^2_{\mathrm{P}}(m, \ell) = \sum_{k=m}^{\ell} \frac{(\hat{\mu}_{jk} - \mu_{m\ell})^2}{\mu_{m\ell}} \tag{3.4.3}$$

which is known to asymptotically approximate a χ^2-distribution with $\ell - m$ degree of freedom (dof), i.e., of variance $2 * (\ell - m)$ and where $\mu_{m\ell}$ is the expected value and represents aggregated score (ES) of the extracted subset and estimated as follows:

$$\mu_{m\ell} = \mathrm{ES}(\ell - m + 1) = \sum_{k=m}^{\ell} \hat{\mu}_{jk} \tag{3.4.4}$$

The ancestries showing unusual gene specific ancestry contributions are $\hat{\boldsymbol{s}}$ top ancestors with $\hat{\boldsymbol{s}}$ the index fold-change Pearson χ^2 score. Assuming that the ranked list contains n proteins, $\hat{\boldsymbol{s}}$ $(\hat{\boldsymbol{s}} \leq n)$ is the smallest index satisfying the following inequality:

$$\frac{\chi^2_{\mathrm{P}}(1, \hat{\boldsymbol{s}}) - r * \chi^2_{\mathrm{P}}(\hat{\boldsymbol{s}} + 1, n)}{\sqrt{1 + r^2}} > 0 \tag{3.4.5}$$

where

$$r = \sqrt{\frac{\hat{\boldsymbol{s}} - 1}{n - \hat{\boldsymbol{s}} - 1}} \tag{3.4.6}$$

which is simply the ratio of standard deviations of $\chi^2_{\mathrm{P}}(1, \hat{\boldsymbol{s}})$ and $\chi^2_{\mathrm{P}}(\hat{\boldsymbol{s}} + 1, n)$, and set to 1 if $\hat{\boldsymbol{s}}$=1.

The significance of the aggregated score, ES $(\hat{\boldsymbol{s}})$, of ancestries showing unusual gene specific ancestry contributions is assessed using sample randomization. We randomly select 1000 independent subsets (of same size $\hat{\boldsymbol{s}}$) of ancestry contributions and compute ES of each subset and then perform the Shapiro-Wilk test under the null hypothesis that the generated sample is drawn from a normal distribution. Following the acceptance or the rejection of the null hypothesis, we perform the T-test or the Wilcoxon to check whether identified ancestries showing unusual gene specific ancestry contributions is more than expected by chance.

3.5 Measuring proteins closeness at the functional level

Proteins perform an astonishing range of biological functions in an organism by collaborating in pathways and processes, and interacting with the cellular environment in some way to promote the cell's growth and function (Mazandu and Mulder, 2011*a*; Mulder *et al.*, 2014). This argues that the outcome of the disease or response to drugs requires concerted biological action of many genes involved in diverse processes or signalling pathways. Here, we use functional data from the Gene Ontology (GO) and the protein GO Annotation (GOA) to score putative proteins which are functionally similar, quantifying functional similarity between proteins based on their GO biological process annotations. Given two proteins p and q, the functional similarity between p and q, BMA (p, q), is computed using the Best Match Average model (Mazandu and Mulder, 2012*a*, 2013*b*, 2014*a*) as follows:

$$\mathrm{BMA}(p, q) = \frac{1}{2}\left(\frac{1}{n}\sum_{s \in T_p} \mathcal{S}(s, T_q) + \frac{1}{m}\sum_{s \in T_q} \mathcal{S}(s, T_p)\right) \tag{3.5.1}$$

where T_r is a set of process terms annotating a given protein r and $n = |T_p|$ and $m = |T_q|$ are the number of processes terms in these sets. $\mathcal{S}(s, T_r) = \max\{\mathcal{S}(s, t) : t \in T_r\}$ with $\mathcal{S}(s, t)$ is the semantic similarity score between process terms s and t computed using the GO-universal metric Mazandu and Mulder (2012*a*), and given by:

$$\mathcal{S}(s, t) = \frac{\mathrm{IC}(c)}{\max\{\mathrm{IC}(s), \mathrm{IC}(t)\}} \tag{3.5.2}$$

where $\mathrm{IC}(x)$ is the information content of the term x in the GO DAG and c is the most informative common ancestor between s and t. The IC value of the term is given by:

$$\mathrm{IC}\left(x\right) = -\ln\left(p\left(x\right)\right) \tag{3.5.3}$$

$p(x)$ is called the topological position characteristic of x, recursively obtained using its parents gathered in the set $\mathcal{P}_x = \{t : (t, x) \in \mathcal{L}_{GO}\}$ where $\mathcal{L}_{GO}$ expresses the set of links in the GO-DAG and $(t, x) \in \mathcal{L}_{GO}$ represents the link or association between a given parent t and its child x. This topological position characteristic, $p(x)$, is given by

$$p\left(x\right) = \begin{cases} 1 & \text{if } x \text{ is a root,} \\ \displaystyle\prod_{t \in \mathcal{P}_x} \frac{p\left(t\right)}{|\mathcal{C}_h(t)|} & \text{otherwise} \end{cases} \tag{3.5.4}$$

with $|\mathcal{C}_h(t)|$ the number of children with term t as parent.

3.6 Retrieving enriched processes and pathways of targets identified

We used the Gene Ontology (GO) and the protein GO Annotation (GOA) mapping provided by the UniProtKB-GOA project to reveal enriched processes in which a set of protein targets are involved. For elucidating statistically significant pathways implicated to the disease, we used the KEGG pathway dataset to extract all human pathways. For each GO term process and pathways, the p-value was calculated using its frequencies of occurrence in the reference dataset (human proteome) and target set, which composed of identified protein targets, and adjusted using the Bonferroni multiple testing correction. We used the hyper-geometric distribution, the p-value is the probability of observing at least s proteins from a target gene set of size n by chance, knowing that the reference dataset contains m annotated genes out of N genes. This is given by the following formula (Mazandu and Mulder, 2013*a*):

$$P\left[X \geq s\right] = 1 - \sum_{k=0}^{s-1} \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}} \tag{3.6.1}$$

where the random variable X is the number of genes annotated with the GO term under consideration within a given disease-associated gene subset. To account for relationships between GO terms in the GO structure, we used the concept of the GO term semantic similarity score (Mazandu and Mulder, 2013*a*) and the frequency of occurrence $f\left(t\right)$ of the target-associated process

t in a set of proteins P is given by:

$$f(t) = \sum_{q \in P} \delta_q(t) \tag{3.6.2}$$

where δ_q is the $q-$function indicator given by

$$\delta_q(t) = \begin{cases} 1 & \text{if } \mathcal{A}_q(t) > \varepsilon \\ 0 & \text{otherwise} \end{cases} \tag{3.6.3}$$

where $\varepsilon \geq 0$ is the agreement level or customized agreement at which a GO term is considered to be a possible annotation of the protein q and $\mathcal{A}_q(t) = \mathcal{S}(t, T_q)$, representing the semantic similarity degree at which a related term is considered to semantically reflect in the specification of the term t (Mazandu and Mulder, 2014*b*). The p-value of each process term was adjusted using the Bonferroni multiple testing correction.

Chapter 4

Results and discussion

To analyze susceptibility to tuberculosis (TB) at the systems level, we start by constructing intra- and inter-organism protein-protein functional interaction networks (human, MTB strain CDC1551 and human-MTB). In this chapter, we used computational methods described in the previous chapter to integrate different protein-protein interaction datasets in order to generate unified networks and map different targets/markers identified into these networks to assess the contribution of each target/marker to the system using network centrality measures. Moreover, we checked whether these targets are related by interacting in the map, belonging to the same cluster or being closely related based on GO processes annotating them. Since targets or markers, especially those associated with a specific phenotype, may differ between individuals or populations, but may be involved in the similar processes or work together in the same pathway, we also performed functional analyses of different targets, identifying enriched processes and pathways in which different markers are involved.

4.1 General topological structure of unified functional networks

In this section, we present general features and topological properties of different functional networks generated. It is worth mentioning that in different analyses, interactions with low confidence scores (between 0 and 0.3) were discarded unless they are predicted by at least two different sources. For the MTB strain CDC1551, we have predicted a total of 138429 functional interactions (edges) connecting 4170 proteins in the complete set of all 4201 proteins from the UniProt database. The human functional network is comprised of 2878644 functional interactions connecting 19274 proteins in the set of 20199 human reviewed proteins as read from the file retrieved from the UniProt database. Different predicted interactions were categorized in three classes, namely low,

medium and high confidence interactions and summarized in Table 4.1 and different general features of each network are shown in Table 4.2.

Table 4.1: **Predicted protein-protein functional interactions.** Functional interactions in different networks shown separately for each dataset per confidence range. '-' indicates that a source was not used because of lack of data for the organismes under consideration. 'Other' source is specifically related to human-MTB interactions extracted from interolog-DIP-known, interolog-DIP array and interolog-HPI-array (Rapanoel *et al.*, 2013).

Data source	Low			Medium			High		
	Human	MTB	Human-MTB	Human	MTB	Human-MTB	Human	MTB	Human-MTB
STRING	2591298	165668	-	729973	62780	-	283776	14340	-
Sequence data	517331	193	0	447354	26220	7260	1484883	40510	303
Interologs	-	-	0	-	-	0	-	-	92
DIP	0	-	-	0	-	-	61865	-	-
Intact	0	-	-	0	-	-	4714	-	-
Others	-	-	0	-	-	0	-	-	608
Combined score	2931698	159943	0	1104684	84133	7260	1773958	54286	552

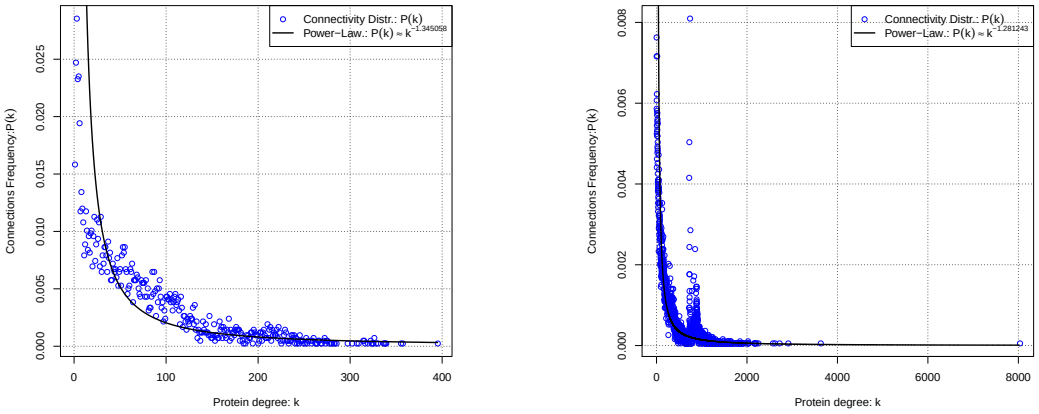
In MTB functional network, we identified 10 interactions with low confidence scores, but predicted by at least two different sources. The human and MTB functional networks were analyzed for general properties of the degree and path-length distributions and results are shown in Figures 4.1 and 4.2.

Table 4.2: **General network parameters.** Features of different functional networks in terms of number of proteins and functional interactions connecting them, as well members of connected components where possible.

Features	Value		
	Human	MTB	Human-MTB
Number of interactions	2878644	138429	7868
Number of proteins	19274	4170	2095
Number of human proteins	-	-	225
Number of CDC1551 proteins	-	-	103
Number of components	52	23	197

4.1.1 Fitting degree and path-length distribution

Degree distribution: The result in Figures 4.1 shows that these functional networks satisfy scale-free topology properties, i.e., the degree distribution of proteins approximates a power law function $\mathbb{P}(k) = k^{-\gamma}$, with the degree exponents $\gamma \sim 1.34506$, and 1.28124 for MTB and human, respectively. These degree exponent values were determined using the linear model: $\log(\mathbb{P}(k)) \sim -\gamma \log(k)$, linear in log, with p-values $< 2e-16$ for human and MTB functional networks, respectively, under the null hypothesis that $\gamma = 0$.



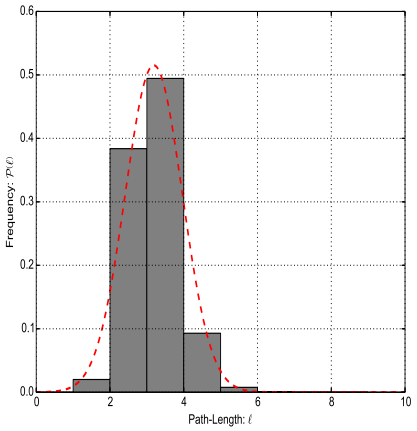
(a) MTB protein degree distribution

(b) Human protein degree distribution

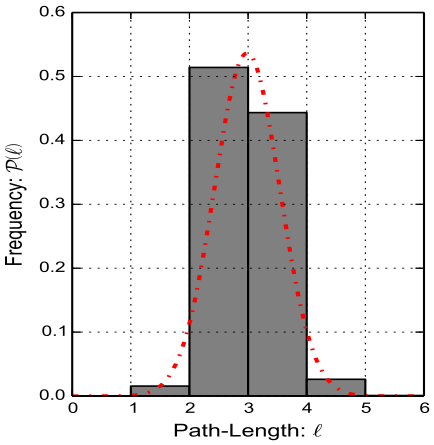
Figure 4.1: **Protein connectivity or degree distribution in MTB and human functional networks.** Circle mortar represents the frequency $\mathbb{P}(k)$ of observing a protein interacting with k partners in a functional network. The solid line plots the power-law function approximating the connectivity distribution.

This means that the observed data shows evidence that these power exponent values are significantly different from 0. This means that although some of the proteins would have many interacting partners, most of them would have few partners. The proteins that have many interacting partners are called "high degree" proteins or hubs or key proteins and probably ensure some basic chemical operations, such as energy transfer and redox reactions, essential for the survival of the organism (Albert *et al.*, 2000).

Path-length distribution: The average path lengths are approximately 3 (hops) for MTB and human functional networks, respectively, indicating that the spread of biological information in these systems is relatively fast. The two functional networks have a "small world" property, i.e, the transmission of biological information from a given protein to others is achieved through only a few steps. This provides an idea about the network navigability, indicating how fast the information can be spread in the system independently of the number of proteins (Albert *et al.*, 2000). Since MTB must survive in the host, this property may provide it with an evolutionary advantage in the sense that it would be able to efficiently respond to the perturbations in the environment and to quickly exhibit a qualitative change of behaviour in response to these perturbations. This is important, as MTB may be forced to react in response to host antibacterial immune mechanisms in order to promote its entry and replication into the host cell.



(a)MTB path-length distribution



(b) Human path-length distribution

Figure 4.2: **Path-length distribution in MTB and human functional networks.** Histogram plot represents the path-length distribution, i.e, frequency of occurrence of shortest path of length ℓ, $\ell = 1, 2, 3, \ldots$ and the dashed line plot is the normal distribution approximating the path length distribution.

4.1.2 Identification of network key proteins and clustering results

Revealing the hierarchical community structure of the networks, identifying key proteins and mapping disease associated genes or proteins onto unified functional networks can provide a better understanding of the disease pathogenesis and increased potential to fully characterize the susceptible genes. In the human system, 2878644 interactions out of a total of 5810340 predicted functional interactions in the unified network with scores ranging from medium to high confidence level were used. We identified 915 key proteins out of 19274 found in the human protein-protein functional network representing about 5% of interacting proteins in the human protein-protein functional interaction network. The clustering model described in section 3.2 splits this human networks into 52 clusters or densely connected communities, the biggest or giant cluster containing 4385 proteins, of which, 122 are key proteins. Among the 52 modules or clusters in the network, only 9 of the clusters contain key proteins and the distribution of these key proteins as shown in Table 4.3.

For the pathogen system, only 138429 out of 172108 protein-protein functional interactions (edges) connecting 4170 proteins have interaction scores ranging from medium to high confidence scores (scores greater than 0.3) or identified by at least two different sources, and used for further analyses. In this network,

Table 4.3: **Classification of human proteins in the functional network.** Distribution of proteins and key proteins in 9 different clusters.

Cluster number	Number of proteins	Number of key proteins
0	4385	122
1	1333	29
2	4233	323
3	2890	102
4	925	30
5	1941	9
6	987	60
7	1188	159
8	1288	81
Total	19170	915

we have identified 22 clusters or communities, the giant cluster contains a total of 907 proteins. Using the threshold value given for key proteins, there are 1095 key proteins in the MTB network.

4.2 Tuberculosis risk genes in different populations

A SNP-gene mapping file was downloaded from the International Haplotype Map Phase 3 website (Frazer *et al.*, 2007) at `http://hapmap.ncbi.nlm.nih.gov/`. SNPs were mapped to their corresponding genes in the human functional network based on the SNP-gene distance. Among these SNPs, 112 and 122 SNPs with moderate p-value ($\leq 5 \times 10^{-5}$) in SAC (Chimusa *et al.*, 2014) and Ghana-Gambia (Thye *et al.*, 2010) population, respectively, were retrieved from literature. There is no SNP shared between the admixed and homogeneous populations. From the human functional network, 18813 proteins were found in the SNP-gene mapping file, of which, 15 and 52 proteins have SNPs with moderate p-values and belong to SAC and Ghana-Gambia populations, respectively. No common gene is shared between the two populations.

4.2.1 Identification tuberculosis risk genes

To identify TB risk genes or candidate genes, we combine the effect of all SNPs that did not reach the intrinsic genome-wide significance threshold p-value of 5×10^{-8}, but with moderate p-values within a particular gene using the approach described in section 3.3. 6 proteins out of 15 in SAC population and 8 out 52 in Ghana-Gambia met the intrinsic genome-wide significance threshold p-value and are believed to be disease causing genes (disease candidate genes) or to make populations under consideration to be genetically susceptible to TB. These significant disease associated genes with their associated moderate SNPs and distances are shown in Table 4.4.

The 6 disease associated genes with significant combined p-value are A3KMH1, Q7Z5N4, Q6ZVL6, Q9NQ90, P19544, and Q8NEX9. These genes are involved in different processes and have different functions in the human system. The details about the proteins and their function can be found in the UniProt website (http://www.uniprot.org/). For instance, A3KMH1 has two GO-molecular functions; ATPase activity, which contributes the catalysis of the reaction: ATP + H2O = ADP + phosphate + 2 H+. It is an enzyme or a catalytic protein in ATP cycle through which energy passes during its transfer from catabolic to anabolic pathways. The other biochemcal activity of this gene is ATP binding, interacting selectively and non-covalently with ATP, adenosine 5'-triphosphate, a universally important coenzyme and enzyme regulator. This protein can be attacked and used by the pathogen to weaken the host system as it is important in energy release and consumption process.

Similarly, we get the set of disease associated genes in the homogeneous Ghana-Gambia population with their associated moderate SNPs and distances, see Table 4.4. The 8 significant disease associated genes are Q9BQI5, P32245, Q86Y38, A1KZ92, Q07020, Q5T5C0, Q8TEW8, and O60669. The details and functions of these proteins can also be found at the UniProt website (http://www.uniprot.org/). For example, the disease associated gene Q9BQI5 has 3 GO-molecular function and 6 GO-biological processes annotated to it.

4.2.2 Quantifying SAC ancestral contributions to TB susceptibility

We combined the SAC true locus-specific ancestry contribution retrieved from (Chimusa *et al.*, 2014) within each candidate gene using the model described in section 3.4 in order to elucidate the ancestor contributing to TB susceptibility in this admixed population. The South African Coloured (SAC) population is an admixed population group in South Africa which has a high incidence of TB. The population is a mixture of five populations; isiXhosa(YRI), Khomani SAN (KHS), European (CEU), Indian (GIH), and Chinese (CHS). By combining the local ancestry contribution of GWAS SNPs at gene level, we elucidated the ancestry which confers TB disease susceptibility to the SAC population.

We took the 6 SAC disease genes and combined their SNPs local ancestry contribution to the corresponding genes ancestry proportion of each ancestor using maximum likelihood estimation as shown in Table 4.5. Looking for the ancestry with high statistical significance in each disease gene, we found that African ancestor Khomani San (KHS) has high contribution in all the disease genes with p-values $< 2e - 16$. This means this ancestor contributes highly

Table 4.4: **Different disease associated genes identified.** Significant disease associated proteins of the admixed SAC (in the first part) and homogeneous Ghana-Gambia populations (in the second part) with their descriptions (name), cluster in which they are mapped (Cluster Ref), associated moderate SNPs and distances.

ID	Name	Gene name	Cluster Ref	Total SNPs	Closest SNPs	SNPs	Distance	P-value	Combined P-value
A3KMH1	von Willebrand factor A domain-containing protein 8	VWA8	7	306	219	rs9525555 rs1900442	0 0	5.72115e-06 3.00952e-06	1.46047e-10
Q7Z5N4	Protein sidekick-1	SDK1	6	989	746	rs12701526 rs1542222	4313 5356	8.10041e-06 2.48252e-06	1.71852e-10
Q6ZVL6	UPF0606 protein KIAA1549L	KIAA1549L	3	165	105	rs4755689 rs12273774 rs6184655	18621 22713 47656	3.24968e-07 2.92121e-06 6.41033e-06	5.81170e-16
Q9NQ90	Anoctamin-2	ANO2	4	351	315	rs12426185	0	4.44541e-08	4.44541e-08
P19544	Wilms tumor protein	WT1	8	135	42	rs11031731 rs10767930 rs7924866 rs11031714 rs7928458	43891 47636 58969 67298 81257	2.6962e-06 7.2745e-06 1.67123e-06 1.00134e-06 3.02134e-06	9.77582e-25
Q8NEX9	Short-chain dehydrogenase/reductase family 9C member 7	SDR9C7	0	63	18	rs1872436 rs719750 rs7970701 rs840160	0 0 0 0	1.05439e-10 3.74505e-09 3.74505e-09 3.74505e-09	1.12956e-32
Q9BQI5	SH3-containing GRB2-like protein 3-interacting protein 1	SGIP1	1	311	202	rs10493412 rs12141816 rs17436311	0 0 0	5.58e-05 5.73e-05 5.17e-05	1.04617e-11
P32245	Melanocortin receptor 4	MC4R	2	258	3	rs7236588 rs1943238 rs1943240	70882 75228 76434	1.36e-05 8.7e-05 8.39e-06	7.48991e-13
Q86Y38	Xylosyltransferase 1	XYLT1	5	556	227	rs7197476 rs7190310 rs1542421	0 0 0	4.46e-05 7.61e-05 5.43e-05	1.16627e-11
A1KZ92	Peroxidasin-like protein	PXDNL	6	415	325	rs1837087 rs7844531 rs6986651 rs4873534	5674 12506 13020 43923	1.52e-05 8.45e-06 8.45e-06 1.17e-05	9.28385e-18
Q07020	60S ribosomal protein L18	RPL18	0	191	3	rs1353690 rs904263	250064 256730	1.83e-05 5.43e-05	7.75686e-09
Q5T5C0	Syntaxin-binding protein 5	STXBP5	7	153	54	rs9373523 rs4896905	0 0	1.23e-06 3.22e-05	3.70374e-10
Q8TEW8	Partitioning defective 3 homolog B	PARD3B	2	813	646	rs2335704 rs2335705 rs13390761	0 0 0	4.63e-06 4.59e-06 2.73e-05	9.28385e-18
O60669	Monocarboxylate transporter 2	SLC16A7	5	445	99	rs2175950 rs11173067 rs10877268	69438 166303 307062	1.79e-05 3.32e-05 8.54e-05	3.46121e-12

to TB susceptibility in the population. This fact agrees with other results that the African ancestry of the SAC population has higher risk of TB infection (Daya *et al.*, 2014*a*). It is worth mentioning that the fact that the genetic

Table 4.5: **Gene level ancestry contribution in SAC disease associated genes.** Combining ancestry specific TB risk at gene level in the SAC population to predict ancestries conferring disease risk to this admixed population.

Gene	KHS	GIH	CHS	YRI	CEU
A3KMH1	0.49658	0.00110	0.00034	0.00170	0.0
Q8NEX9	0.73704	0.01841	0.12653	0.00699	0.0
Q7Z5N4	0.98908	0.00614	0.00477	0.0	0.0
Q6ZVL6	0.93727	0.01569	0.33856	0.10504	0.00182
Q9NQ90	0.93315	0.01296	0.00068	0.01568	0.00477
P19544	0.76862	0.02032	0.10463	0.03492	0.00436

contributions of different ancestries with a gene do not sum to 1 provides an indication that the genetic component does not fully explain susceptibility to the disease. This means that the environment factors, including life style, may also contribute to TB susceptibility in this population.

4.2.3 Mapping different candidate genes onto functional networks

As indicated previously, the two population groups under consideration in this study, namely the admixed SAC and the homogeneous Ghana-Gambia have no disease associated gene in common. Nevertheless it is possible that these genes interact and influence each other in the same sub-network to produce similar effect in these populations. Thus, examining the effects of these genes beyond single gene approach can shed light to the full characterization of candidate genes and the genetic structure of TB in these populations. In this section, we use human, MTB and human-MTB protein-protein functional networks as an integrated system to check how identified candidate genes work together in the same module or whether they interact directly and indirectly with the pathogen.

We explored how candidate genes identified in different populations interact using the Protein Interaction Visualizer (PINV) tool (Salazar *et al.*, 2014) at `http://biosual.cbio.uct.ac.za/pinViewer.html?core=HumanColl` (accessed on March 01 2016) by searching candidate genes for each population as the protein queries. Figures 4.3 and 4.4 show the general interaction structure of these proteins.

4.2.3.1 Graph-based relationships between candidate genes

Figure 4.3 shows the sub-network containing the SAC disease genes, which potentially represents the SAC sub-network of interacting genes underlying ethnic differences in disease risk. The disease gene Q725N4 has the biggest degree of 899 interacting proteins and Q62VL6 has the smallest, 15 proteins. Figure 4.4 shows a potential Ghana-Gambia sub-network of interacting genes. From all A1KZ92 has the highest degree with 1087 interacting partners. It is worth mentioning that one disease associated protein in the homogeneous population, namely P32245 is a key protein, which is expected to be evolutionary more conserved (Kaçar and Gaucher, 2013) than other disease candidate proteins in the population. This happens because this gene is likely to experience stronger selective constraints than other candidate genes in the network due to its importance in the functioning of the system (Albert *et al.*, 2000).

Among 8 and 6 disease associated genes for the two populations, namely Ghana-Gambia and SAC populations, respectively, 3 gene pairs interact in common clusters: Q07020 and Q8NEX9, A1KZ92 and Q7Z5N4, and Q5T5C0 and A3KMH1. For the remaining genes in the set of candidate genes in the two population groups each gene is singularly located in its cluster. The difference in disease susceptibility between the two populations may come from these

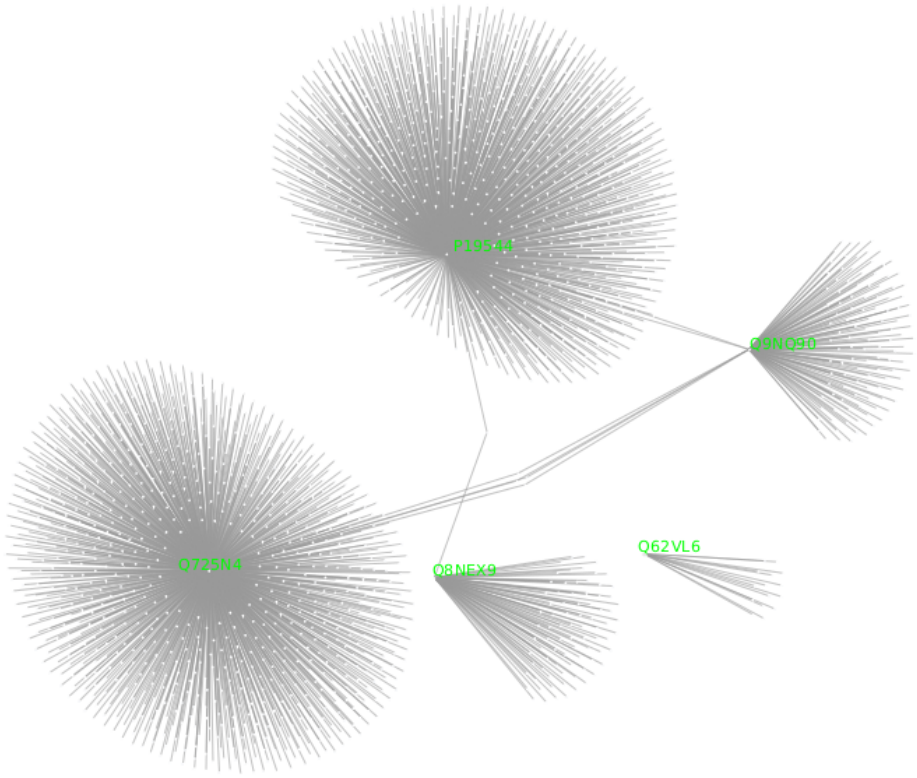


Figure 4.3: **SAC disease genes mapped on to the human functional network:** The sub-network containing all identified SAC disease genes in green and showing how these genes are connected.

disease associated proteins. Clusters containing disease genes are statistically enriched with 1039 main biological processes and some of them are shown in Table 4.6. Several processes are related to immune system and indicate that these candidate genes may contribute to down-regulate various biological processes, such as tissue repair and immune responses, among many others, and may have an effect in progression for infection to disease. This provides an indication that these sub-networks are potentially associated with disease risk in the population under consideration.

4.2.3.2 Prediction of human-MTB interactions influencing susceptibility to tuberculosis

The infection outcome depends on the interplay between the timing and concentration of various signalling molecules and this is critical for the fate of the host-pathogen interaction. As consequence, the failure of this process may yield increased susceptibility to the pathogen, which possibly negotiates its entry into host cell by activating processes to thwart host defence mechanisms. In this section, we identify potential MTB proteins which interact with candidate proteins, in which case, these candidates genes may

Table 4.6: **Some statistically enriched biological processes in which non common clusters containing disease candidate genes are involved.** For each process identified level of the term in the GO DAG description, p-value and corrected p-value following Bonferroni multiple testing correction are provided.

Term ID	Level	Process Name	P-value	Corrected P-value
GO:0061485	11	memory T cell proliferation	5.93764e-06	0.02774
GO:0001865	12	NK T cell differentiation	5.56822e-09	2.60147e-05
GO:0050798	11	activated T cell proliferation	1.58631e-06	0.00741
GO:0045065	11	cytotoxic T cell differentiation	1.65457e-08	7.73017e-05
GO:0043382	10	positive regulation of memory T cell differentiation	6.22663e-06	0.02909
GO:0072369	12	regulation of lipid transport by positive regulation of transcription from RNA polymerase II promoter	9.09354e-11	4.24850e-07
GO:0034389	4	lipid particle organization	6.21640e-11	2.90430e-07
GO:1903955	10	positive regulation of protein targeting to mitochondrion, saturated fatty acid	2.73654e-07	0.00128
GO:0002291	11	T cell activation via T cell receptor contact with antigen bound to MHC molecule on antigen presenting cell	1.34764e-08	6.29621e-05
GO:0046597	8	negative regulation of viral entry into host cell	3.57554e-08	0.00017
GO:0046007	10	negative regulation of activated T cell proliferation	8.38265e-06	0.03916
GO:0042492	11	gamma-delta T cell differentiation	4.27773e-08	0.00019
GO:0038168	11	epidermal growth factor receptor signaling pathway via I-kappaB kinase/NF-kappaB cascade	8.91922e-11	4.16706e-07
GO:0035457	6	cellular response to interferon-alpha	3.12971e-07	0.00146
GO:0032608	5	interferon-beta production	1.33742e-09	6.24844e-06
GO:0034116	7	positive regulation of heterotypic cell-cell adhesion	3.18167e-07	0.00148
GO:1900165	10	negative regulation of interleukin-6 secretion	1.11220e-08	5.19620e-05
GO:0050718	11	positive regulation of interleukin-1 beta secretion	2.30484e-07	0.00107
GO:0035963	6	cellular response to interleukin-13	2.73974e-07	0.00128
GO:0032618	4	interleukin-15 production	3.28122e-08	0.00015
GO:0032632	5	interleukin-3 production	3.28122e-08	0.00015
GO:0071351	6	cellular response to interleukin-18	2.73974e-07	0.00128
GO:2000778	10	positive regulation of interleukin-6 secretion	3.89489e-11	1.81969e-07
GO:0045416	8	positive regulation of interleukin-8 biosynthetic process	1.03925e-05	0.04855
GO:0036016	6	cellular response to interleukin-3	1.08243e-07	0.00050
GO:0071104	5	response to interleukin-9	5.93932e-07	0.00277
GO:0071105	5	response to interleukin-11	1.21508e-06	0.00567
GO:0030890	9	positive regulation of B cell proliferation	8.30169e-12	3.87855e-08
GO:0002326	11	B cell lineage commitment	3.85891e-12	1.80288e-08
GO:0002331	13	pre-B cell allelic exclusion	7.36557e-11	3.44119e-07
GO:0001922	8	B-1 B cell homeostasis	7.49510e-08	0.00035
GO:0002337	13	B-1a B cell differentiation	6.81722e-11	3.18500e-07
GO:0002352	12	B cell negative selection	4.66575e-11	2.17984e-07
GO:0002358	8	B cell homeostatic proliferation	0.0	0.0
GO:0090340	9	positive regulation of secretion of lysosomal enzymes	3.08500e-10	1.44131e-06
GO:0045429	8	positive regulation of nitric oxide biosynthetic process	8.62045990146e-11	4.02747886596e-07
GO:0051000	8	positive regulation of nitric-oxide synthase activity	6.78463784265e-07	0.00316978280009
GO:0051001	8	negative regulation of nitric-oxide synthase activity	1.03924840439e-05	0.048553685453

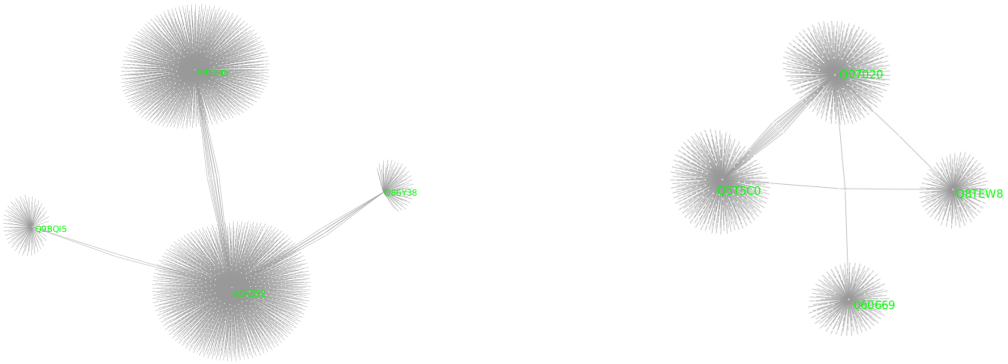


Figure 4.4: **Ghana-Gambia disease genes mapped on to the network:** The sub-network containing all Ghana-Gambia significant disease genes in green and showing how these genes are connected.

potentially contribute to the various stages of the pathogen life cycle within the host. These candidate genes may possibly help the pathogen to achieve its goal, preventing host protective immunity against the pathogen and likely to cause disease.

In the human-MTB functional network, the total number of human proteins interacting with MTB is 255 and that of MTB proteins interacting with human is 103.

From the list of candidate disease genes, the SAC disease associated genes A3KMH1, Q9NQ90, and Q8NEX9 and their direct neighbours interact with the pathogen. On another side, there are 349 MTB proteins interacting with SAC disease genes or their direct interacting neighbours. Table 4.7 shows statistically significant or enriched biological processes in which these MTB proteins are involved. For homogeneous group, the homogeneous disease

Table 4.7: **Some statistically enriched biological processes in which MTB proteins interacting with SAC disease genes or its partners are involved.** For each process identified level of the term in the GO DAG description, p-value and corrected p-value following Bonferroni multiple testing correction are provided.

Term ID	Level	Process Name	P-value	Corrected P-value
GO:0019367	11	fatty acid elongation	1.45863e-10	1.38570e-08
GO:0019605	10	butyrate metabolic process	1.15010e-08	1.09260e-06
GO:0097089	9	methyl-branched fatty acid metabolic process	1.15010e-08	1.09260e-06
GO:0001676	9	long-chain fatty acid metabolic process	7.38368e-09	7.01449e-07

associated genes with their direct neighbours interacting with the MTB proteins are P32245, Q86Y38, A1KZ92, Q07020, Q5T5C0, Q8TEW8, and O60669. On the other hand, the number of MTB proteins interacting with homogeneous disease genes or its interacting partner is 18. In this case, we searched for processes of clusters in which the proteins belong. These 18 interacting MTB proteins belong to six clusters involved in main significant biological processes shown in Table 4.8.

Table 4.8: **Some statistically enriched biological processes in which MTB proteins interacting with homogeneous disease genes or its partners are involved.** For each process identified level of the term in the GO DAG description, p-value and corrected p-value following Bonferroni multiple testing correction are provided.

Term ID	Level	Process Name	P-value	Corrected P-value
GO:0070929	9	trans-translation	2.61956e-11	9.32563e-09
GO:0006415	8	translational termination	4.61710e-11	1.64368e-08
GO:0001676	9	long-chain fatty acid metabolic process	1.15763e-05	0.00412
GO:0006413	8	translational initiation	4.61710e-11	1.64368e-08
GO:0019605	10	butyrate metabolic process	2.44826e-05	0.00871
GO:0019679	11	propionate metabolic process, methylcitrate cycle	1.48611e-05	0.00529
GO:0097089	9	methyl-branched fatty acid metabolic process	2.44826e-05	0.00871
GO:0019367	11	fatty acid elongation, saturated fatty acid	6.63498e-05	0.02362

The results obtained indicate that MTB proteins interacting with disease associated genes are involved in fatty acid related processes. Indeed, the number of MTB proteins which are involved in the fatty acid related processes and interact with disease associated genes in the two population groups, SAC and homogeneous Ghana-Gambia, was more than expected by chance with p-values 6.72517e-12 and 5.85126e-05 respectively. Fatty acids are very important nutrients for the survival of the MTB cells in the host environment (Kinsella *et al.*, 2003). MTB has approximately 250 genes involved in fatty acid metabolism, a much higher proportion than in any other organism. Biosynthesis of fatty acids is important because its mycolic acids, produced from elongated fatty acids, plays an essential role in the formation of complex lipids layer on the cell membrane which help it to survive inside the host, contributing to its virulence and pathogenesis. This membrane layer have been targeted by many of the drugs used to treat MTB infection (Kinsella *et al.*, 2003). MTB uses the envelop of complex lipids on its cell membrane to protect itself in the host environment.

4.2.4 Retrieving potential enriched processes and pathways of candidate genes

As pointed out previously, proteins can also interact with one another in different pathways and processes, which are responsible for the biological dynamics of a system (Mulder *et al.*, 2014), in which case, these proteins influence each other either through enhancement or through hindrance. This indicates that risk-associated genes may be different between populations or individuals, but may influence each other in the same pathway or contribute to the same process. Thus, identifying and understanding pathways or processes in which candidate genes are involved can provide a deeper insight into the underlying genetic mechanisms of TB pathogenesis and may contribute to the development of TB therapeutics. Here, we identified statistically significant processes and pathways in which candidate genes in different populations under consideration are involved.

Reading human-GO annotations to retrieve biological process terms associated with the disease associated genes, the total number of biological processes that involve the disease associated proteins in SAC population is 61 with 44 filtered processes. Among these, 7 processes are found to be statistically significant. Table 4.9 presents the 7 significant biological processes in which the SAC disease associated proteins are involved. For the Ghana-Gambia homogeneous population, 46 processes were identified with 35 filtered processes, no enriched process was detected for the Ghana-Gambia homogeneous population.

Table 4.9: **Some statistically enriched biological processes in which SAC disease associated proteins are involved.** For each process identified level of the term in the GO DAG description, p-value and corrected p-value following Bonferroni multiple testing correction are provided.

Term ID	Level	Process Name	P-value	Corrected P-value
GO:0072284	11	metanephric S-shaped body morphogenesis	0.00093	0.04116
GO:0035802	9	adrenal cortex formation	0.00047	0.02058
GO:2000195	6	negative regulation of female gonad development	0.00062	0.02744
GO:2001076	8	positive regulation of metanephric ureteric bud development	0.00031	0.01372
GO:0072302	11	negative regulation of metanephric glomerular mesangial cell proliferation	0.00016	0.00686
GO:0072166	11	posterior mesonephric tubule development	0.00078	0.03430
GO:0060923	10	cardiac muscle cell fate commitment	0.00093	0.04116

We performed a pathway enrichment analysis using set of human pathways

collected from the KEGG pathway (Kanehisa and Goto, 2000). No enriched pathway was identified for the SAC population and for the Ghana-Gambia population, 2 enriched pathways were detected: *Glycosaminoglycan biosynthesis-chondroitin sulfate/dermatan sulfate* (KEGG ID:hsa00532) and *Glycosaminoglycan biosynthesis-heparan sulfate/heparin* (KEGG ID:hsa00534), with adjusted p-values of 0.000245. These pathways are shown in Figures 4.5 and 4.6. These pathways are likely to be triggered to enhance

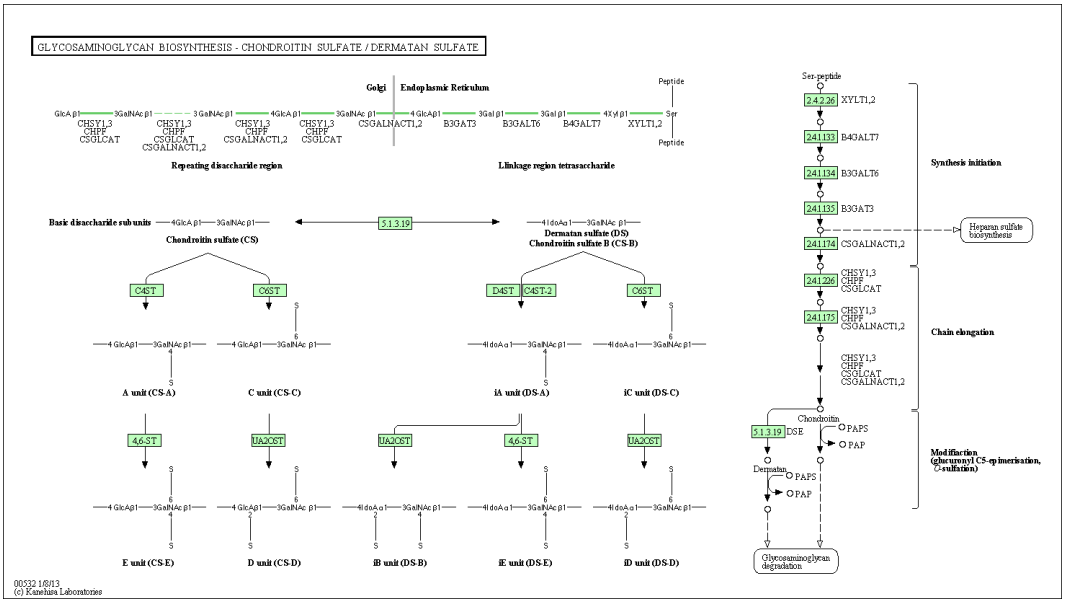


Figure 4.5: ***Glycosaminoglycan biosynthesis-chondroitin sulfate/dermatan sulfate* (KEGG ID:hsa00532).** The KEGG map as retrieved from the KEGG website (http://www.genome.jp/kegg-bin/show_pathway?hsa00532).

the first line of contact between pathogen and host cell, which is essential to a pathogen's invasive potential, and activate biological interactions influencing many physiological and pathological processes, including cell-to-cell communication, adhesion and the immune response (Kamhi *et al.*, 2013). These interactions, GAGs modulate various biological processes, such as cell adhesion, proliferation and migration, tissue repair, coagulation, and immune responses, among many others. Glycosaminoglycan–pathogen interactions affect most, if not all, the key steps of microbial pathogenesis, including host cell invasion, cell–cell transmission and evasion of host defense mechanisms (Aquino *et al.*, 2010). These host glycosaminoglycan pathways mediates the entry of pathogens into human cells, and it has been observed that in *Drosophila fly*, this pathway promote pathogen virulence and to confer susceptibility to infection (Aquino *et al.*, 2010). These observations

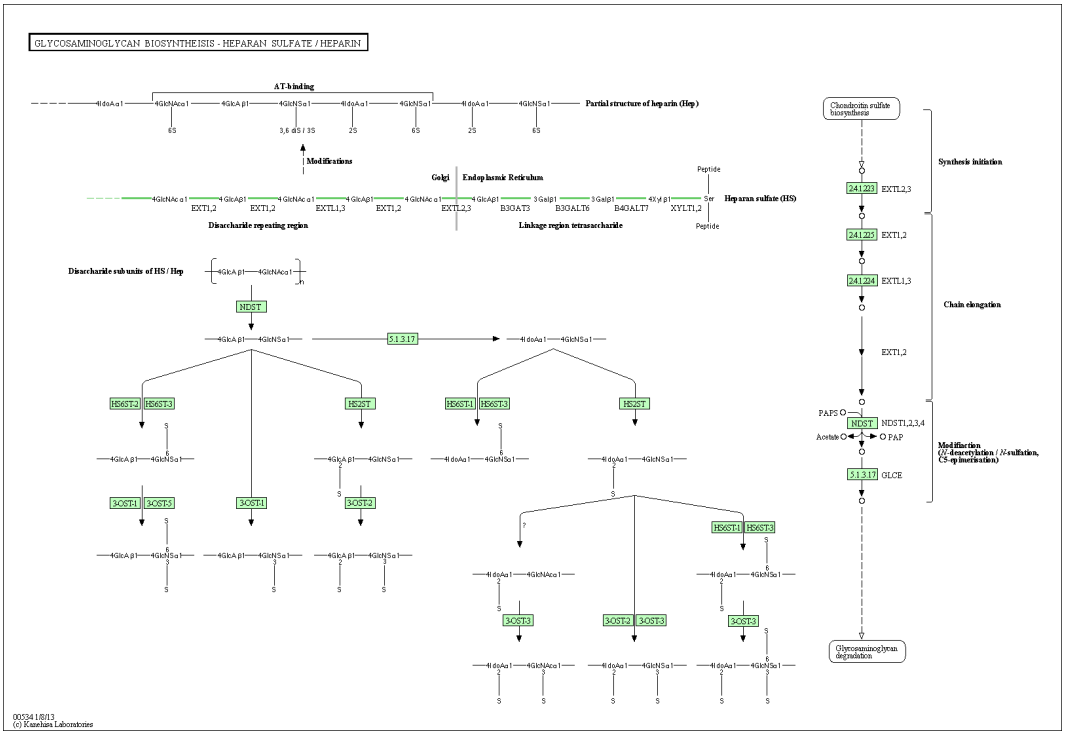


Figure 4.6: ***Glycosaminoglycan biosynthesis-heparan sulfate/heparin* (KEGG ID:hsa00534).** The KEGG map as retrieved from the KEGG website (http://www.genome.jp/kegg-bin/show_pathway?hsa00534).

indicate that glycosaminoglycan–pathogen interactions serve diverse functions that affect the pathogenesis of infectious diseases.

As pointed out previously, the two population groups, the admixed coloured population and the homogeneous Ghana-Gambia population do not share disease associated genes, however, we examine how close are these candidate genes at the functional level based on the GO biological process terms using the similarity measure discussed in section 3.5. So, we compute functional similarity scores between protein pairs using the GO-Universal metric (Mazandu and Mulder, 2014*b*) and cluster these proteins using distance (dissimilarity) scores derived from functional similarity scores. Figure 4.7 shows how close are these genes using the hierarchical clustering map or dendrogram. For example the two proteins SDR9C7 and PXDNL in figure 4.7 clustered together and the distance between the two is approximately 0.35. so, the functional similarity score between the two proteins is $\mathcal{S}(SDR9C7, PXDNL) = 0.65$. The two proteins are functionally most similar compared to the other proteins. Similarly, the proteins ANO2 (from SAC) and SLC16A7 (from Ghana-Gambia) are clustered with similarity score of 0.58 approximately, indicating that theses proteins are more similar functionally to each other than they are similar

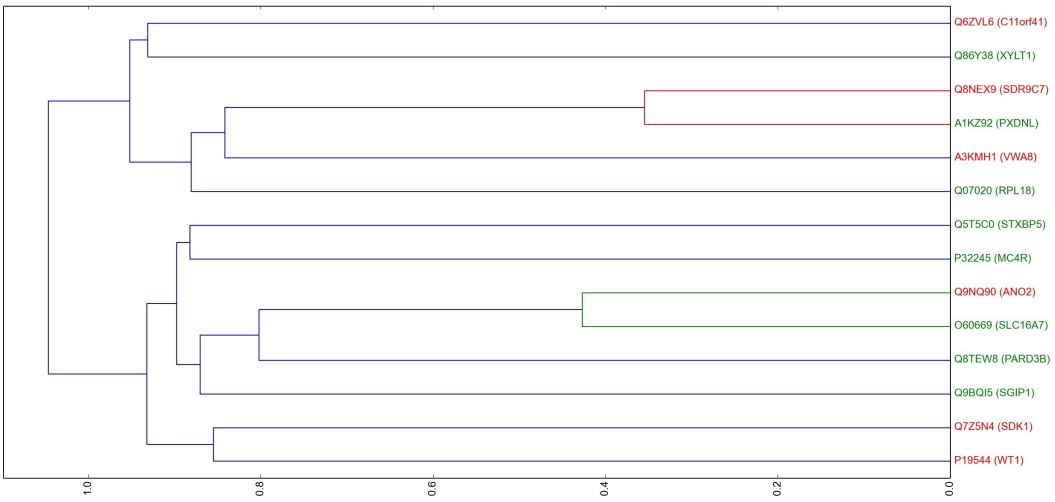


Figure 4.7: **Hierarchical clustering map of disease genes.** Horizontal axis shows the distance or dissimilarity score between a pair of proteins or clusters in the set of disease associated proteins. The proteins in green are those from the homogeneous Ghana-Gambia population and the red ones are the disease genes of the admixed SAC population. The hierarchical clustering map shows how similar or dissimilar are gene or protein pairs at functional level and shows their functional cluster group.

to other proteins. But the distance d($Q62VL6, Q86Y38$) is approximately 0.9 and their functional similarity is approximately 0.1. Therefore the two proteins are functionally very different. From Figure 4.7 we generally have two main clusters and we see that most of the disease genes are dissimilar from each other and dispersed. This indicates that these proteins have variety of functional purposes in the system. Even the proteins within the same population are functionally dissimilar.

4.3 Disease candidate genes and drug targets

Front-line TB therapies last at least six months using the initial combination of isoniazid, rifampin, pyrazinamide, and ethambutal (Mulder *et al.*, 2013), a control strategy implemented in response to the global TB epidemic, known as Directly Observed Therapy, Short-course (DOTS). It is worth mentioning that these anti-TB compounds were introduced between 1952 and 1963 (Mazandu and Mulder, 2012*b*): isoniazid (1952), pyrazinamide (1954), cycloserine (1955), ethambutol (1962), and rifampin (1963), indicating that TB is still being treated with a decades-old drug regimen. Adverse drug reaction contributes to non-compliance with this long duration of TB treatment, leading to the development of resistance, the main failure in TB treatment. Furthermore,

using drugs or chemical agents/antibiotics to treat an infectious disease has selective action on infective or pathogenic organisms and also has adverse side effects in humans, in which case the drug used can be either ineffective or even toxic. 59% of drugs causing adverse side effects are metabolized by polymorphic enzymes(Huang *et al.*, 2006), indicating that genetic component is one of the critical factors that contribute to TB treatment outcomes. In this section, we investigate potential role of genetic factors in TB therapy outcome in relation to disease associated genes in admixed (SAC) and homogeneous (Ghana-Gambia) populations. Certain genetic variants or SNPs are associated with significant changes in drug efficacy, drug disposition (Roden *et al.*, 2006; Eichelbaum *et al.*, 2006; Ma and Lu, 2011) as they may directly impact drug metabolizing enzymes, drug targets and drug receptors (Ramos *et al.*, 2012). Thus, we map SNPs to drug targets and drug targets to the unified functional network in order to examine relationships between front-line drug targets and disease associated genes and to predict the interaction between human and bacterial pathogen potentially influencing drug responses.

4.3.1 GO-based functional relationship between drug targets

We used GO processes to examine functional similarity between drug targets using GO-universal metric based on enriched process terms in which these targets are involved. Thus, we retrieved protein biological processes most pertinent to each target set, the set of drug target proteins. We obtained 53 filtered processes associated with isoniazid drug target proteins and 12 are significant biological processes that are performed by isoniazid drug targets as show in Table 4.10. There are 72 filtered biological processes associated with rifampin target proteins and 11 are significant processes shown in Table 4.11. There are 55 filtered biological processes associated with pyrazinamide target proteins and 16 are significant processes performed by pyrazinamide drug targets (see Table 4.12). There are no processes found which are associated with ethambutol drug target, which has only one putative targets, namely P11473, extracted from the Guide to Pharmacology database. These target sets share some proteins in common and contribute in several common biological processes, including *exogenous drug catabolic process* (GO:0042738) and *toxin biosynthetic process* (GO:0009403).

We plot the hierarchical clustering map for the set of drug target proteins that shows their pairwise functional similarity scores, assessing how close are these targets to each other. The clustering results are shown in figure 4.8. The vertical axis is the set of drug target proteins and the horizontal axis represents the distance or dissimilarity between a pair of proteins or clusters. In Figure 4.8 the target proteins in isoniazid are in red, proteins in rifampin

Table 4.10: **Some statistically enriched biological processes in which isoniazid drug target proteins are involved.** For each process identified level of the term in the GO DAG description, p-value and corrected p-value following Bonferroni multiple testing correction are provided.

Term ID	Term Level	Process Name	P-value	Corrected P-value
GO:0019373	11	epoxygenase P450 pathway	3.01841e-11	1.53938e-09
GO:0006805	5	xenobiotic metabolic process	0.0	0.0
GO:0006706	5	steroid catabolic process	1.97002e-05	0.00100
GO:0016098	7	monoterpenoid metabolic process	0.0	0.0
GO:0097267	11	omega-hydroxylase P450 pathway	7.45685e-11	3.80299e-09
GO:0070989	4	oxidative demethylation	0.0	0.0
GO:0071615	4	oxidative deethylation	5.76946e-10	0.04116
GO:0090350	7	negative regulation of cellular organofluorine metabolic process	0.00026	0.01325
GO:0009822	5	alkaloid catabolic process	1.82343e-07	9.29954e-06
GO:0046226	6	coumarin catabolic process	0.00052	0.02651
GO:0009403	5	toxin biosynthetic process	0.00026	0.01326
GO:0042738	5	exogenous drug catabolic process	3.87484e-11	1.97617e-09

Table 4.11: **Some statistically enriched biological processes in which rifampin drug target proteins are involved.** For each process identified level of the term in the GO DAG description, p-value and corrected p-value following Bonferroni multiple testing correction are provided.

Term ID	Term Level	Process Name	P-value	Corrected P-value
GO:0019373	11	epoxygenase P450 pathway	0.0	0.0
GO:0006706	5	steroid catabolic process	4.58716e-05	0.00330
GO:0070980	7	biphenyl catabolic process	0.00039	0.02807
GO:0016098	7	monoterpenoid metabolic process	7.92217e-11	5.70396e-09
GO:0097267	11	omega-hydroxylase P450 pathway	4.69371e-11	3.37947e-09
GO:0070989	4	oxidative demethylation	0.0	0.0
GO:0071615	4	oxidative deethylation	2.50821e-11	1.80591e-09
GO:0006789	7	bilirubin conjugation	0.00039	0.02807
GO:0009822	5	alkaloid catabolic process	4.25558e-07	3.06401e-05
GO:0009403	5	toxin biosynthetic process	0.00038	0.02807
GO:0042738	5	exogenous drug catabolic process	0.0	0.0

Table 4.12: **Some statistically enriched biological processes in which pyrazinamide drug target proteins are involved.** For each process identified level of the term in the GO DAG description, p-value and corrected p-value following Bonferroni multiple testing correction are provided.

Term ID	Term Level	Process Name	P-value	Corrected P-value
GO:2001213	8	negative regulation of vasculogenesis	0.00013	0.00714
GO:0042816	6	vitamin B6 metabolic process	0.00065	0.03573
GO:0009822	5	alkaloid catabolic process	0.00039	0.02144
GO:0006805	5	xenobiotic metabolic process	0.00033	0.01819
GO:1900747	7	negative regulation of vascular endothelial growth factor signaling pathway	0.00052	0.02859
GO:0036378	8	calcitriol biosynthetic process from calciol	0.00052	0.02859
GO:0016098	7	monoterpenoid metabolic process	0.00025	0.01392
GO:0060058	10	positive regulation of apoptotic process involved in mammary gland involution	0.00052	0.02859
GO:0070989	4	oxidative demethylation	1.41804e-06	7.79922e-05
GO:0045602	7	negative regulation of endothelial cell differentiation	0.00078	0.04288
GO:0006706	5	steroid catabolic process	4.38698e-06	0.00024
GO:0071615	4	oxidative deethylation	1.76591e-05	0.00097
GO:0009115	9	xanthine catabolic process	0.00026	0.01429
GO:0007595	8	lactation	1.21781e-05	0.00067
GO:0009403	5	toxin biosynthetic process	0.00013	0.00714
GO:0042738	5	exogenous drug catabolic process	1.33501e-05	0.00073

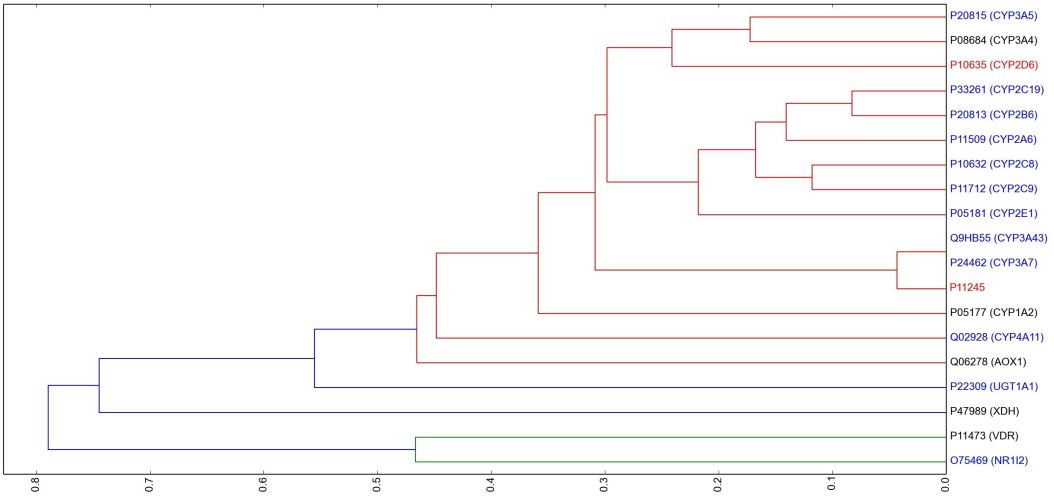


Figure 4.8: **Hierarchical clustering map for drug target proteins.** Horizontal axis shows the distance or dissimilarity score between a pair of proteins or clusters in the set of drug target proteins. The target proteins in isoniazid are in red, proteins in rifampin are in blue, proteins in pyrazinamide are in black, and proteins in ethambutol are in green. But notice that the drugs have common protein targets and those common targets take only one of the colors.

are in blue, proteins in pyrazinamide are in black, and proteins in ethambutol are in green. But notice that the drugs have common protein targets and those common targets take only one of the colors. The protein P11473(VDR) is common target for all the four drugs though it appears black in the dendrogram. Similarly, P05177 (CYP1A2) and P08684(CYP3A4) are common targets for the three drugs isoniazid, rifampin, and pyrazinamide. P05181 (CYP2E1), P11712 (CYP2C9), P11509 (CYP2A6), P10632 (CYP2C8) are common targets for isoniazid and rifampin.

From Figure 4.8, we observe that rifampin protein targets Q9HB55 and P24462 are the same functionally since the distance $\mathrm{d}(Q9HB55, P24462) = 0$. These two proteins and isoniazid target protein P11245 are in the same cluster with dissimilarity score $\mathrm{d}(P11245, P24462) = \mathrm{d}(P11245, Q9HB55) \approx 0.05$. These proteins are the most similar compared to the other proteins. The second most similar proteins are P33261 and P20813 which are rifampin target proteins. The common target protein for all drugs, P11473 (VDR) is in the same cluster (similar) with the protein O75469 (NR1I2) of rifampin with similarity score of approximately 0.55 and it is in the same cluster with the other proteins with similarity score of 0.20.

In most cases, functional similarity values are greater than 0.5 and 15 out of 19 proteins belong to the same cluster. All protein targets of rifampin and isoniazid except P22309 are in this big cluster and P05177 is the only target of pyrazinamide which is also target for isoniazid and rifampin is in this cluster. In general, using drug combination therapies is often necessary to ensure that the disease is contained (Winter *et al.*, 2012), but interactions between these drugs can produce either antagonistic, indifferent or synergistic effects (Maity *et al.*, 2014). In the context of the TB front-line therapy, the fact that drug targets share high functional similarity scores provides an indication that interactions between these drugs are likely to produce synergistic or indifferent effects. Thus, further pharmaco-kinetic or dynamic analysis of potential drug-drug interactions are essential to determine drug response effects.

4.3.2 Drug targets vs disease risk genes and MTB system

In Table 4.13, we have TB front-line drugs protein targets with the number of associated SNPs and closest distance SNPs ($d = 0$). These SNPs can be responsible for drug metabolism and drug response. They are mostly found in drug metabolising enzymes and they are responsible for variation in drug response (Alwi, 2005). The front line target proteins or the pharmacogenes we have considered are important in drug metabolism to have effective

output from the drug taken. From the result in Table 4.13 none of the drug targets is found to be key proteins in the human functional network. Interestingly, these drug targets share common clusters with disease associated genes especially in the case of homogeneous population. Regarding interaction with MTB proteins, only pyrazinamide target proteins Q06278 and P47989 interact with MTB proteins. This provides an indication that these targets are likely to treat the underlying cause of disease, but further analyses are still needed to examine the effectiveness of these targets.

Considering MTB proteins front-line drug targets; rifampin: P9WGY8, isoniazid: P9WGR0, pyrazinamide: Q7D6Z3, and ethambutol: P9WNL4, P9WNL8, P9WNL6. These target proteins are key proteins in MTB network and they interact with human proteins except isoniazid target protein P9WGR0. This may cause harmful side-effects and contributes to drug inefficiency (Mulder *et al.*, 2013).

Table 4.13: **Relationship between TB front-line drug targets and disease associated genes.** Mapping different SNPs to their corresponding human targets elucidating targets which are key proteins in the functional networks and identifying those interacting with the other organism proteins and those located in the same cluster with disease associated genes. '1' indicates that a target under consideration is a key protein/shared a common clusters with disease associated genes/ interacts with the other organism (human or pathogen), and '0' indicates otherwise.

Target ID	Organism	Key protein	Interacting with MTB/Human protein	Total Number of SNPs	Closest SNP	common cluster as admixed disease genes	common cluster as Homogeneous disease genes
P11712	human	0	0	107	90	0	1
P05177	human	0	0	101	91	0	1
P10632	human	0	0	120	90	0	1
P08684	human	0	0	54	48	0	1
P20813	human	0	0	69	46	0	1
P22309	human	0	0	31	31	0	1
P33261	human	0	0	109	70	0	1
P11509	human	0	0	24	11	0	1
P05181	human	0	0	88	33	0	1
Q9HB55	human	0	0	38	30	1	1
P20815	human	0	0	42	31	0	1
P24462	human	0	0	49	39	0	1
Q02928	human	0	0	46	27	1	1
O75469	human	0	0	85	74	0	1
P11473	human	0	0	94	65	0	1
P10635	human	0	0	25	9	0	1
P11245	human	0	0	0	0	0	1
P47989	human	0	1	148	92	0	1
Q06278	human	0	1	170	109	1	1
P9WGY8	MTB	1	1	-	-	–	-
P9WGR0	MTB	0	0	-	-	–	-
Q7D6Z3	MTB	1	1	-	-	–	-
P9WNL4	MTB	1	0	-	-	–	-
P9WNL8	MTB	1	0	-	-	–	-
P9WNL6	MTB	1	0	-	-	–	-

Chapter 5

Conclusion

Currently, tuberculosis (TB) is a major global health challenge especially in Sub-Saharan Africa. Significant progress is being made to stop TB, but the emergence of drug resistance MTB strains and its synergy with HIV make the implementation of control measurement difficult. TB latently infected approximately one third of the world population and the admixed South African Coloured (SAC) population which is located on the western cape of South Africa has a high incidence of TB and is ideally suited to the discovery of TB susceptibility genetic variants and their probable ethnic origins. In this project, we used an integrated framework model to analyse susceptibility to tuberculosis in homogeneous and admixed population. we analysed human genetic susceptibility in relation to the MTB system. We identified potential genes, biological processes, and pathways involved in TB susceptibility, and the ancestry showing association with TB susceptibility in the admixed SAC population. Furthermore, we have examined the relationship between different disease associated genes and front-line drug targets.

Using a graph-based model on human and MTB protein-protein functional interactions, we have analyzed topological properties of the networks, including small-world and scale-free properties of networks and characterized key proteins in the network using centrality measures. Combining association signals of GWAS and integrating with the human protein-protein functional interaction network, we identified 6 disease associated genes for the admixed SAC population and 8 disease associated genes for the homogeneous Ghana-Gambia population. We identified clusters or sub-graphs in the network containing these disease associated genes. We elucidated a total of 9 disease associated clusters where 3 of the clusters are common that contain disease genes from both population groups and 6 clusters are non-common. These genes can be responsible for the difference in genetic susceptibility to TB between the two population groups.

Gene Ontology (GO) term and pathway enrichment analyses were performed

targeting disease genes and their clusters, and TB first line drug targets to elucidate potential biological processes in which these genes contribute. Moreover, for the admixed SAC population, we combined local ancestry contribution of SNPs at gene level for the identified disease associated genes. We observed that the African Khomani San ancestry of the SAC population is overrepresented in all the disease genes, thus contributing to TB risk in this population and this observation agrees with previous findings. A disease is often a result of several interactions between proteins or genes affecting multiple biological processes and pathways. This suggest that the identification of disease associated genes and their underling biological processes and pathways can advance our understanding of TB pathogenesis and may be useful for the development of novel drugs based on target proteins and processes involved in the disease.

List of References

Albert, R., Jeong, H. and Barabási, A.-L. (2000). Error and attack tolerance of complex networks. *Nature*, vol. 406, no. 6794, pp. 378–382.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, vol. 215, no. 3, pp. 403–410.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997). Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Research*, vol. 25, no. 17, pp. 3389–3402.

Alwi, Z.B. (2005). The use of snps in pharmacogenomics studies. *The Malaysian Journal of Medical Sciences (MJMS)*, vol. 12, no. 2, p. 4.

Aquino, R.S., Lee, E.S. and Park, P.W. (2010). Diverse functions of glycosaminoglycans in infectious diseases. *Progress in Molecular Biology and Translational Science*, vol. 93, pp. 373–394.

Arisoa, R.H. (2012). Network generation and analysis for the investigation of host-pathogen interactions in tuberculosis holifidy arisoa rapanoel. *MSc thesis, University of Cape Town Press.*

Barabasi, A.-L. and Oltvai, Z.N. (2004). Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, vol. 5, no. 2, pp. 101–113.

Bastien, O. and Maréchal, E. (2008). Evolution of biological sequences implies an extreme value distribution of type i for both global and local pairwise alignment scores. *BMC Bioinformatics*, vol. 9, no. 1, p. 332.

Bastien, O., Ortet, P., Roy, S. and Maréchal, E. (2005). A configuration space of homologous proteins conserving mutual information and allowing a phylogeny inference based on pair-wise z-score probabilities. *BMC Bioinformatics*, vol. 6, no. 1, p. 1.

Begum, F., Ghosh, D., Tseng, G.C. and Feingold, E. (2012). Comprehensive literature review and statistical considerations for gwas meta-analysis. *Nucleic acids research*, p. gkr1255.

Bellamy, R. (1998). Genetic susceptibility to tuberculosis in human populations. *Thorax*, vol. 53, no. 7, pp. 588–593.

Berrington, W.R. and Hawn, T.R. (2007). Mycobacterium tuberculosis, macrophages, and the innate immune response: does common variation matter? *Immunological Reviews*, vol. 219, no. 1, pp. 167–186.

Blondel, V., Guillaume, J., Lambiotte, R. and Lefebvre, E. (2008). Fast unfolding of community hierarchies in large networks, 1–6. *arXiv preprint arXiv:0803.0476*.

Bonacich, P. (2007). Some unique properties of eigenvector centrality. *Social Networks*, vol. 29, no. 4, pp. 555–564.

Botstein, D., Cherry, J., Ashburner, M., Ball, C., Blake, J., Butler, H., Davis, A., Dolinski, K., Dwight, S., Eppig, J. *et al.* (2000). Gene ontology: tool for the unification of biology. *Nat Genet*, vol. 25, no. 1, pp. 25–29.

Chatr-Aryamontri, A., Breitkreutz, B.-J., Heinicke, S., Boucher, L., Winter, A., Stark, C., Nixon, J., Ramage, L., Kolas, N., O'Donnell, L. *et al.* (2013). The biogrid interaction database: 2013 update. *Nucleic Acids Research*, vol. 41, no. 1, pp. D816–D823.

Chimusa, E.R., Daya, M., Möller, M., Ramesar, R., Henn, B.M., Van Helden, P.D., Mulder, N.J. and Hoal, E.G. (2013). Determining ancestry proportions in complex admixture scenarios in south africa using a novel proxy ancestry selection method. *PloS ONE*, vol. 8, no. 9, p. e73971.

Chimusa, E.R., Mbiyavanga, M., Mazandu, G.K. and Mulder, N.J. (2015). AncGWAS: a post genome-wide association study method for interaction, pathway, and ancestry analysis in homogeneous and admixed populations. *Bioinformatics*, vol. 32, no. 4, pp. 549–556.

Chimusa, E.R., Zaitlen, N., Daya, M., Möller, M., van Helden, P.D., Mulder, N.J., Price, A.L. and Hoal, E.G. (2014). Genome-wide association study of ancestry-specific tb risk in the south african coloured population. *Human Molecular Genetics*, vol. 23, no. 3, pp. 796–809.

Cousins, D.V., Bastida, R., Cataldi, A., Quse, V., Redrobe, S., Dow, S., Duignan, P., Murray, A., Dupont, C., Ahmed, N. *et al.* (2003). Tuberculosis in seals caused by a novel member of the mycobacterium tuberculosis complex: Mycobacterium pinnipedii sp. nov. *International Journal of Systematic and Evolutionary Microbiology*, vol. 53, no. 5, pp. 1305–1314.

Crombie, I.K. and Davies, H.T. (2009). What is meta-analysis. *What is*, pp. 1–8.

Daya, M., van der Merwe, L., Galal, U., Möller, M., Salie, M., Chimusa, E., Galanter, J., van Helden, P., Henn, B., Gignoux, C. *et al.* (2012). A panel of ancestry informative markers for the complex five-way admixed south african coloured population. *PloS ONE*, vol. 8, no. 12, pp. e82224–e82224.

Daya, M., van der Merwe, L., Gignoux, C.R., Van Helden, P.D., Möller, M. and Hoal, E.G. (2014*a*). Using multi-way admixture mapping to elucidate TB susceptibility in the south african coloured population. *BMC Genomics*, vol. 15, no. 1, p. 1021.

Daya, M., van der Merwe, L., van Helden, P.D., Möller, M. and Hoal, E.G. (2014*b*). The role of ancestry in TB susceptibility of an admixed south african population. *Tuberculosis*, vol. 94, no. 4, pp. 413–420.

Daya, M., van der Merwe, L., van Helden, P.D., Möller, M. and Hoal, E.G. (2015). Investigating the role of gene-gene interactions in TB susceptibility. *PloS ONE*, vol. 10, no. 4.

Dubos, R.J. and Dubos, J. (1952). *The white plague: tuberculosis, man, and society*. Rutgers University Press.

Eichelbaum, M., Ingelman-Sundberg, M. and Evans, W.E. (2006). Pharmacogenomics and individualized drug therapy. *Annu. Rev. Med.*, vol. 57, pp. 119–137.

Fleischmann, R., Alland, D., Eisen, J., Carpenter, L., White, O., Peterson, J., DeBoy, R., Dodson, R., Gwinn, M., Haft, D. *et al.* (2002). Whole-genome comparison of mycobacterium tuberculosis clinical and laboratory strains. *Journal of Bacteriology*, vol. 184, no. 19, pp. 5479–5490.

Frazer, K.A., Ballinger, D.G., Cox, D.R., Hinds, D.A., Stuve, L.L., Gibbs, R.A., Belmont, J.W., Boudreau, A., Hardenbol, P., Leal, S.M. *et al.* (2007). A second generation human haplotype map of over 3.1 million snps. *Nature*, vol. 449, no. 7164, pp. 851–861.

Gene Ontology Consortium (2010). The gene ontology in 2010: extensions and refinements. *Nucleic Acids Research*, vol. 38, no. suppl 1, pp. D331–D335.

Han, B. and Eskin, E. (2011). Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. *The American Journal of Human Genetics*, vol. 88, no. 5, pp. 586–598.

Huang, S.-M., Goodsaid, F., Rahman, A., Frueh, F. and Lesko, L.J. (2006). Application of pharmacogenomics in clinical pharmacology. *Toxicology Mechanisms and Methods*, vol. 16, no. 2-3, pp. 89–99.

Huntley, R.P., Sawford, T., Mutowo-Meullenet, P., Shypitsyna, A., Bonilla, C., Martin, M.J. and O'Donovan, C. (2015). The goa database: gene ontology annotation updates for 2015. *Nucleic Acids Research*, vol. 43, no. D1, pp. D1057–D1063.

Jia, P., Zheng, S., Long, J., Zheng, W. and Zhao, Z. (2011). dmgwas: dense module searching for genome-wide association studies in protein–protein interaction networks. *Bioinformatics*, vol. 27, no. 1, pp. 95–102.

Kaçar, B. and Gaucher, E.A. (2013). Experimental evolution of protein–protein interaction networks. *Biochemical Journal*, vol. 453, no. 3, pp. 311–319.

Kamhi, E., Joo, E.J., Dordick, J.S. and Linhardt, R.J. (2013). Glycosaminoglycans in infectious disease. *Biological Reviews*, vol. 88, no. 4, pp. 928–943.

Kanehisa, M. and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, vol. 28, no. 1, pp. 27–30.

Keshavjee, S. and Farmer, P.E. (2012). Tuberculosis, drug resistance, and the history of modern medicine. *New England Journal of Medicine*, vol. 367, no. 10, pp. 931–936.

Kinsella, R.J., Fitzpatrick, D.A., Creevey, C.J. and McInerney, J.O. (2003). Fatty acid biosynthesis in mycobacterium tuberculosis: lateral gene transfer, adaptive evolution, and gene duplication. *Proceedings of the National Academy of Sciences*, vol. 100, no. 18, pp. 10320–10325.

Kubica, G.P., Kim, T.H. and Dunbar, F.P. (1972). Designation of strain h37rv as the neotype of mycobacterium tuberculosis. *International Journal of Systematic Bacteriology*, vol. 22, no. 2, pp. 99–106.

Kumar, R. and Nanduri, B. (2010). HPIDB-a unified resource for host-pathogen interactions. *BMC Bioinformatics*, vol. 11, no. Suppl 6, pp. 1–6.

Lienhardt, C., Fielding, K., Sillah, J., Bah, B., Gustafson, P., Warndorff, D., Palayew, M., Lisse, I., Donkor, S., Diallo, S. *et al.* (2005). Investigation of the risk factors for tuberculosis: a case–control study in three countries in West Africa. *International Journal of Epidemiology*, vol. 34, no. 4, pp. 914–923.

Ma, Q. and Lu, A.Y. (2011). Pharmacogenetics, pharmacogenomics, and individualized medicine. *Pharmacological reviews*, vol. 63, no. 2, pp. 437–459.

Maity, S.N., Chintaparthi, M.R., Hima Bindu, M., Kanta, R. and Kapur, I. (2014). In vitro antimicrobial activity of cefsulodin and kanamycin in combinations. *International Journal of Research in Medical Sciences*, vol. 2, no. 2, pp. 677–680.

Maliarik, M.J. and Iannuzzi, M.C. (2003). Host genetic factors in resistance and susceptibility to tuberculosis infection and disease. *Semin Respir Crit Care Med*, vol. 24, no. 2, pp. 223–228.

Mazandu, G.K. (2010). Data integration for the analysis of uncharacterized proteins in mycobacterium tuberculosis. *PhD thesis, University of Cape Town Press.*

Mazandu, G.K., Chimusa, E.R., Mbiyavanga, M. and Mulder, N.J. (2015). A-DaGO-FUN: an adaptable gene ontology semantic similarity-based functional analysis tool. *Bioinformatics*, vol. 32, no. 3, pp. 477–479.

Mazandu, G.K. and Mulder, N.J. (2011*a*). Generation and analysis of large-scale data-driven mycobacterium tuberculosis functional networks for drug target identification. *Advances in Bioinformatics*, vol. 2011.

Mazandu, G.K. and Mulder, N.J. (2011*b*). Scoring protein relationships in functional interaction networks predicted from sequence data. *PLoS ONE*, vol. 6, no. 4, p. e18607.

Mazandu, G.K. and Mulder, N.J. (2012*a*). A topology-based metric for measuring term similarity in the gene ontology. *Advances in Bioinformatics*, vol. 2012.

Mazandu, G.K. and Mulder, N.J. (2012*b*). Using the underlying biological organization of the mycobacterium tuberculosis functional network for protein function prediction. *Infection, Genetics and Evolution*, vol. 12, no. 5, pp. 922–932.

Mazandu, G.K. and Mulder, N.J. (2013*a*). DaGO-FUN: tool for gene ontology-based functional analysis using term information content measures. *BMC Bioinformatics*, vol. 14, no. 1, p. 284.

Mazandu, G.K. and Mulder, N.J. (2013*b*). Information content-based gene ontology semantic similarity approaches: toward a unified framework theory. *BioMed Research International*, vol. 2013.

Mazandu, G.K. and Mulder, N.J. (2014*a*). Information content-based gene ontology functional similarity measures: which one to use for a given biological data type? *PloS ONE*, vol. 9, no. 12, p. e113859.

Mazandu, G.K. and Mulder, N.J. (2014*b*). The use of semantic similarity measures for optimally integrating heterogeneous gene ontology data from large scale annotation pipelines. *Front. Genet*, vol. 5, no. 264, pp. 10–3389.

Mulder, N.J., Akinola, R.O., Mazandu, G.K. and Rapanoel, H. (2014). Using biological networks to improve our understanding of infectious diseases. *Computational and Structural Biotechnology Journal*, vol. 11, no. 18, pp. 1–10.

Mulder, N.J., Mazandu, G.K. and Rapano, H.A. (2013). Using host-pathogen functional interactions for filtering potential drug targets in mycobacterium tuberculosis. *Mycobacterial Diseases*, vol. 3, p. 126.

Parish, T. and Brown, A. (2009). *Mycobacterium: genomics and molecular biology*. Horizon Scientific Press.

Qu, H.-Q., Li, Q., McCormick, J.B. and Fisher-Hoch, S.P. (2011). What did we learn from the genome-wide association study for tuberculosis susceptibility? *Journal of Medical Genetics*, pp. jmg–2010.

Ramachandran, G. and Swaminathan, S. (2012). Role of pharmacogenomics in the treatment of tuberculosis: a review. *Pharmacogenomics and Personalized Medicine*, vol. 5, pp. 89–98.

Ramos, E., Callier, S.L. and Rotimi, C.N. (2012). Why personalized medicine will fail if we stay the course. *Personalized Medicine*, vol. 9, no. 8, pp. 839–847.

Rapanoel, H.A., Mazandu, G.K. and Mulder, N.J. (2013). Predicting and analyzing interactions between mycobacterium tuberculosis and its human host. *PloS ONE*, vol. 8, no. 7, p. e67472.

Roden, D.M., Altman, R.B., Benowitz, N.L., Flockhart, D.A., Giacomini, K.M., Johnson, J.A., Krauss, R.M., McLeod, H.L., Ratain, M.J., Relling, M.V. *et al.* (2006). Pharmacogenomics: challenges and opportunities. *Annals of Internal Medicine*, vol. 145, no. 10, pp. 749–757.

Salazar, G.A., Meintjes, A., Mazandu, G.K., Rapanoël, H.A., Akinola, R.O. and Mulder, N.J. (2014). A web-based protein interaction network visualizer. *BMC Bioinformatics*, vol. 15, no. 1, p. 1.

Salwinski, L., Miller, C.S., Smith, A.J., Pettit, F.K., Bowie, J.U. and Eisenberg, D. (2004). The database of interacting proteins: 2004 update. *Nucleic acids research*, vol. 32, no. suppl 1, pp. D449–D451.

Schurr, E. (2011). The contribution of host genetics to tuberculosis pathogenesis. *Kekkaku*, vol. 86, no. 1, pp. 17–28.

Snyder, E., Kampanya, N., Lu, J., Nordberg, E.K., Karur, H., Shukla, M., Soneja, J., Tian, Y., Xue, T., Yoo, H. *et al.* (2007). PATRIC: the VBI pathosystems resource integration center. *Nucleic Acids Research*, vol. 35, no. suppl 1, pp. D401–D406.

Stark, C., Breitkreutz, B.-J., Reguly, T., Boucher, L., Breitkreutz, A. and Tyers, M. (2006). Biogrid: a general repository for interaction datasets. *Nucleic Acids Research*, vol. 34, no. suppl 1, pp. D535–D539.

Stead, W.W., Senner, J.W., Reddick, W.T. and Lofgren, J.P. (1990). Racial differences in susceptibility to infection by mycobacterium tuberculosis. *New England Journal of Medicine*, vol. 322, no. 7, pp. 422–427.

Steinhaeuser, K. and Chawla, N.V. (2010). Identifying and evaluating community structure in complex networks. *Pattern Recognition Letters*, vol. 31, no. 5, pp. 413–421.

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K.P. *et al.* (2015). String v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Research*, vol. 43, no. Database issue, p. D447.

Thye, T., Vannberg, F.O., Wong, S.H., Owusu-Dabo, E., Osei, I., Gyapong, J., Sirugo, G., Sisay-Joof, F., Enimil, A., Chinbuah, M.A. *et al.* (2010). Genome-wide association analyses identifies a susceptibility locus for tuberculosis on chromosome 18q11. 2. *Nature Genetics*, vol. 42, no. 9, pp. 739–741.

UniProt Consortium (2015). UniProt: a hub for protein information. *Nucleic Acids Research*, vol. 43, no. Database issue, pp. D204–D212.

Wagner, A. (2003). How the global structure of protein interaction networks evolves. *Proceedings of the Royal Society of London B: Biological Sciences*, vol. 270, no. 1514, pp. 457–466.

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L. *et al.* (2014). The NHGRI GWAS catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research*, vol. 42, no. D1, pp. D1001–D1006.

Winter, G.E., Rix, U., Carlson, S.M., Gleixner, K.V., Grebien, F., Gridling, M., Müller, A.C., Breitwieser, F.P., Bilban, M., Colinge, J. *et al.* (2012). Systems-pharmacology dissection of a drug synergy in imatinib-resistant CML. *Nature Chemical Biology*, vol. 8, no. 11, pp. 905–912.

Yellaboina, S., Goyal, K. and Mande, S.C. (2007). Inferring genome-wide functional linkages in e. coli by combining improved genome context methods: comparison with high-throughput experimental data. *Genome research*, vol. 17, no. 4, pp. 527–535.

Zhang, A. (2009). *Protein interaction networks: Computational analysis*. Cambridge University Press.

Zhang, Q., Long, Q. and Ott, J. (2014). Apriorigwas, a new pattern mining strategy for detecting genetic variants associated with disease through interaction effects. *PLoS Comput Biol*, vol. 10, no. 6, p. e1003627.