GENETIC EPIDEMIOLOGY OF TUBERCULOSIS

Nooshin Ahmadipour

Department of Epidemiology and Biostatistics

McGill University, Montreal

December 2002

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements of the degree of Master of Science

Your file Votre référence
ISBN: 0-612-88142-3
Our file Notre référence
ISBN: 0-612-88142-3

TABLE OF CONTENTS

ABSTRACT

Background: Susceptibility to a complex disease such as tuberculosis generally involves interactions among several genes and environmental factors. Several association studies have been conducted to examine the association between candidate genes and tuberculosis. However, the genetic risk factors are not fully understood.

Objective: To examine the effect of several candidate genes, including natural resistance associated macrophage protein 1 (NRAMP1), vitamin D receptor (VDR), surfactant proteins (SFTPA1), and mannose-binding lectin (MBL), and also to assess the effect of several risk factors on their association with tuberculosis. The other objectives were to test for mode of inheritance and also to estimate the relative risks of disease for different genotypes.

Methods: A prospective case-parental control study was conducted. Ninety-five nuclear families were selected from an existing database of families with tuberculosis in Ethiopia. Each family consisted of one affected child and two parents. The primary outcome was transmission/nontransmission of alleles from parents to affected offspring.

Results: The transmission disequilibrium test showed that marker SFTPA1-294 was significantly associated with the outcome ($\chi^2 = 4.297$; $p = 0.038$). When other risk factors such as age, sex, ethnicity, certain symptoms or other genes were allowed to modify the transmission probabilities in a logistic regression model, several other markers were found to be significantly associated with the outcome.

Conclusions: Despite the limitations of this study, this thesis provided evidence for inheritance of susceptibility to tuberculosis in Hadiayan families in Ethiopia. To confirm the findings in this thesis, it would be useful to conduct similar research in populations with different ethnic origins, where genetic and environmental exposures can be examined and compared.

RÉSUMÉ

Contexte: La sensibilité à une maladie complexe comme la tuberculose provient généralement d'interactions entre différents gènes et facteurs environnementaux. Plusieurs études d'association ont été menées afin d'examiner les liens entre des gènes candidats et la tuberculose. Les facteurs de risque génétiques de cette maladie n'ont toutefois pas encore été entièrement élucidés.

Objectifs: Examiner l'effet de différents gènes candidats, notamment la résistance naturelle associée à la protéine 1 des macrophages (NRAMP1), au récepteur de la vitamine D (VDR), aux protéines de surfactant (SFTPA1) et à la lectine fixatrice du mannose (MBL), et évaluer l'effet de certains facteurs de risque sur l'association de ces gènes à la tuberculose. Établir le mode de transmission de cette maladie et estimer les risques d'atteinte relatifs de différents génotypes.

Méthode: Une étude prospective cas-parents témoins a été réalisée. Quatre-vingt-quinze familles nucléaires ont été sélectionnées à l'aide d'une base de données sur les familles atteintes de tuberculose en Éthiopie. Les familles choisies étaient constituées d'un enfant atteint et de deux parents. Le résultat principal était la transmission ou la nontransmission d'allèles des parents à l'enfant atteint.

Résultats: Le test de déséquilibre de transmission a révélé une association significative entre le marqueur SFTPA1-294 et le résultat (χ^2 = 4,297; p = 0,038). Lorsqu'on a permis à d'autres facteurs de risque comme l'âge, le sexe, l'origine ethnique, certains symptômes ou d'autres gènes de modifier les probabilités de transmission dans un modèle de régression logistique, on a observé une association significative entre plusieurs autres marqueurs et le résultat.

Conclusions: Malgré les limites de cette étude, les données obtenues soutiennent la thèse d'une transmission héréditaire de la sensibilité à la tuberculose chez les familles d'Hadiaya. Pour confirmer ces résultats, il faudra mener des recherches semblables sur des populations d'origines ethniques différentes chez lesquelles les facteurs génétiques et environnementaux peuvent être étudiés et comparés.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude and respect to my thesis supervisor, Dr. Celia Greenwood for her constant support, patience, and guidance during all stages of this thesis. I thank her for her statistical and genetic expertise. Her insightful suggestions and comments during the writing of this thesis are greatly appreciated.

I am grateful to my thesis co-supervisor, Dr. Theresa Gyorkos for her initial acceptance to take me as one of her graduate students. I also appreciate her support, time and advice throughout my thesis.

I also would like to thank Dr. Erwin Schurr, a member of my thesis committee, for giving me the permission to use his database, and for his comments.

I would like to thank all persons who participated in this study, without them none of this research would be possible. I also express my gratitude to Dr. Tewodros Eguale, the study physician in Ethiopia for his great job. I appreciate Mrs. Leah Simkin for her database design, programming, and guidance in using the database; Neil Malik for genotyping the samples at laboratory of Montreal General Hospital; and all other staff involved in AHRI project in Montreal and Ethiopia.

Finally, I would like to thank my husband for his constant support, understanding, and encouragement during the completion of this thesis.

Acknowledgement of financial support

I gratefully acknowledge the fellowship I received from Dr. Celia Greenwood and Dr. Erwin Schurr, while I was a student at the Research Institute of the Montreal General Hospital.

LIST OF ABBREVIATIONS

"A"	Adenine
AHRI	Armauer Hansen Research Institute (in Ethiopia)
AIDS	Acquired immunodeficiency syndrome
AM(s)	Alveolar macrophage(s)
ARI (ARTI)	Annual risk of infection (Annual risk of tuberculosis infection)
BCG	Bacille Calmette-Guérin
bp	base pairs
"C"	Cytosine
CDC	Centers for Disease Control and Prevention
CI	Confidence Interval
CMI	Cell-mediated immunity
DNA	Deoxyribonucleic acid
ds-DNA	Double stranded-DNA
DOTS	Directly observed therapy (short course)
DTH	Delayed type hypersensitivity
DZ	Dizygotic
EEA	Equal environmental assumption
EMB	Ethambutol
EPTB	Extrapulmonary tuberculosis
"G"	Guanine
G × E	Genotype-environment interaction
G × G	Gene-gene interaction
HIV	Human immune deficiency virus
IL	Interleukin
INH	Isoniazid
IUATLD	International Union Against Tuberculosis and Lung Disease
LD	Linkage disequilibrium
MBL (MBP)	Mannose-binding lectin (protein)
MDR-TB	Multi-drug resistant tuberculosis

M. tuberculosis	*Mycobacterium tuberculosis*
MZ	Monozygotic
NO	Nitrous oxide
Nramp1	Natural resistance-associated macrophage protein 1 (in mice)
NRAMP1	Natural resistance-associated macrophage protein 1 (in humans)
OR	Odds ratio
PPD	Purified Protein Derivative
PTB	Pulmonary tuberculosis
PZA	Pyrazinamide
RMP	Rifampin (Rifampicin)
RR	Risk ratio
SM	Streptomycin
SFTP	Surfactant proteins
"T"	Thymine
TB	Tuberculosis
TDT	Transmission disequilibrium test
TNF	Tumor necrosis factor
TST	Tuberculin Skin Test
VDR	Vitamin D Receptor
WHO	World Health Organization

LIST OF TABLES AND FIGURES

LIST OF APPENDICES

CHAPTER 1: INTRODUCTION

Tuberculosis (TB), caused by *Mycobacterium tuberculosis* (*M. tuberculosis)*, is a chronic infection, which usually affects the lungs, but any organ may be involved. Once infection is established, the disease may develop within a short period of time or may be delayed for months, years, or even an entire life (Moulding, 1999; Raviglione & O'Brien, 2001). Tuberculosis is an ancient infection that has affected humans since their early history (Evans, 1998). Records of epidemics of tuberculosis are documented from the 1600s, but their numbers have declined with social improvement in the 19th century, and continued to decline with the discovery of vaccine and effective therapy (Geiter, 2000). In fact, tuberculosis was almost ignored until only recently, when a worldwide resurgence became evident due to 1) Human Immunodeficiency Virus (HIV) pandemics; 2) multi-drug resistance TB; 3) poverty and homelessness; 4) neglect of control programs; 5) changing demography (increasing world population and changing age structure); and 6) immigration from high prevalence regions (Maher & Raviglione, 1999). Tuberculosis is the only disease declared by the World Health Organization (WHO) as a global emergency (Maher & Raviglione, 1999); WHO estimates that one person in the world becomes infected with TB every second, and that one third of the entire population of the world is infected, of whom 10% progress to clinically defined tuberculosis. Without prompt, effective action, the annual incident number of cases of TB globally may increase by 40% by 2020 (Raviglione & O'Brien, 2001).

Susceptibility to complex traits such as tuberculosis generally involves the contributions of many different genes, as well as environmental factors that can modify disease risk (Eaves & Sullivan, 2001). There is convincing evidence that host genes may play a major role in both the risk of acquiring infection and developing tuberculosis (Kwiatkowski, 2000). Several twin studies of TB (Kallmann & Reisner, 1943; Comstock, 1978; Jepson et al., 2001) have suggested that genetic factors participate in host susceptibility to tuberculosis. In addition, several investigations have indicated that certain racial or ethnic groups are more likely to develop tuberculosis (Stead et al., 1990; Moulding, 1999).

Early detection and prompt and appropriate treatment of TB cases are important tools to control this disease (Dye, 2000). The WHO strategy to control TB is the application of

directly observed treatment, short-course chemotherapy (DOTS). This strategy ensures adherence to treatment and reduces drug-resistance and relapse of the disease. However, this strategy is not effective for those already infected with drug-resistant strains of mycobacteria (Farmer, 2001). Vaccination and preventive chemotherapy are other strategies used to prevent and control tuberculosis. Nevertheless, the world now requires more effective vaccines and better therapeutic agents for all types of tuberculosis.

Developments in genetics have allowed a more systematic study of the interaction of genes and environment on the occurrence of infectious diseases. Several approaches, including linkage studies, association studies of candidate genes and animal models have enabled identification of host genes that regulate infectious diseases. The identification of host genes that contribute to susceptibility/resistance to tuberculosis will provide new keys to better understand the pathogenic mechanisms that result in disease (Abel & Dessein, 1997).

CHAPTER 2: BACKGROUND

2.1 TUBERCULOSIS

2.1.1 History

Tuberculosis appears to be as old as mankind. It has been speculated that the first human infections may have been caused by *Mycobacterium bovis* (*M. bovis*) acquired from domesticated animals and that *M. tuberculosis* may have evolved from *M. bovis* (Hass & Hass, 1995; Evans, 1998). Mummies and skeletal remains from the predynastic era (prior to 3000 B.C.) in Northern Africa, show evidence of pulmonary and spinal TB (Daniel et al., 1994). Similar proof of infection has been discovered in Germany, Peru, and Chile (Daniel et al., 1994; Evans, 1998).

Tuberculosis was probably a sporadic disease of humans in their early history. It has been suggested that tuberculosis became endemic in humans when stable social networks of several hundred people were established, about 10,000 years ago (Daniel et al., 1994). Crowding in European cities, and later the Industrial Revolution in Western Europe provided the necessary environmental conditions for an endemic disease to become epidemic in the early 1600s (Daniel et al., 1994; Stead & Bates, 1995). The epidemic grew and spread through Western Europe during the next two centuries. As Western Europeans moved about the globe, the epidemic of tuberculosis followed them to Eastern Europe, North America, Asia, and North Africa. However, tuberculosis came much later to many parts of the world, most notably sub-Saharan Africa, the Pacific Islands, and the highlands of New Guinea (Daniel et al., 1994; Stead & Bates, 1995). It is still unknown in some isolated tribes in the Amazon highlands.

During the 20th century, TB incidence rapidly decreased in industrialized countries. The invention of chemoprophylaxy (vaccine), effective chemotherapy, improvement in social and nutritional conditions, and changes in population size have commonly been assumed to have influenced the decline in incidence and mortality from tuberculosis. Furthermore, there is a current ongoing debate for the role of natural selection reducing TB susceptibility in populations: within a given population, resistant individuals may have survived to reproduce more often than susceptible individuals. However, it has been shown quantitatively that natural selection alone cannot account for the rapid decline in TB mortality over a period of 300 years (Stead & Bates, 1995; Lipsitch, & Sousa, 2002).

Before the discovery of the tubercle bacillus by Robert Koch in 1882, diseases such as tuberculosis and leprosy were widely believed to be inherited disorders (Cooke & Hill, 2001). Indo-European Aryans (about 1500 B.C.) believed that tuberculosis was caused by stress in people with an inherited susceptibility. They suggested that stress developed from fasting, sorrow, pregnancy, and chest wounds (Hass & Hass, 1995). Hippocrates (during the fifth century B.C.) considered tuberculosis as essentially non-contagious and associated with an inherited susceptibility (Hass & Hass, 1995). In the twentieth century, family studies have confirmed the heritability of susceptibility to several infectious diseases (Cooke & Hill, 2001), validating, in some sense, the intuition of Hippocrates. In 1949, Haldane hypothesized that infectious diseases had been an important force of natural selection for at least 5000 years (Lederberg, 1999); Haldane stated "It is much easier for a mouse to get a set of genes which enables it to resist *Bacillus typhimurium* than a set which enables it to resist cats."

2.1.2 Clinical features

1) Definition & etiology

Tuberculosis is caused by bacteria belonging to the *Mycobacterium tuberculosis* complex. The disease usually affects the lungs, although in up to one third of cases, extra-pulmonary sites are involved. Transmission usually occurs through the airborne spread of droplet nuclei produced by patients with infectious pulmonary TB (PTB).

The *M. tuberculosis* complex belongs to the family Mycobacteraceae and includes *M. tuberculosis*, *M. bovis*, *M. africanum*, *M. canettii*, and *M. microti* (Fitzpatrick & Braden, 2000). The majority of cases are caused by *M. tuberculosis*. In the past, transmission of infection with *M. bovis* through consumption of milk from infected cows was common, but this has been brought under control in developed countries. However, in developing countries, depending on pasteurization of milk, *M. bovis* may still be prevalent (Dutt & Stead, 1999). *M. africanum* has been isolated in West and Central Africa, and in Southeast England (Raviglione & O'Brien, 2001). *M. africanum* is highly transmissible and is probably spread by aerosol transmission. *M. microti* is a pathogen of rodents, and *M. canettii* is a newly described variant of *M. tuberculosis* that has been isolated from several patients who acquired infection in Africa. The nontuberculous mycobacteria

(such as *M. avium* complex), and *M. leprae* are also members of the family Mycobacteraceae (Raviglione & O'Brien, 2001).

2) Pathophysiology

M. tuberculosis is nearly always transmitted through the airborne spread of droplet nuclei produced by patients with pulmonary (Brahmer & Small, 1998; Raviglione & O'Brien, 2001) or laryngeal TB (Fitzpatrick & Braden, 2000) which are aerosolized by coughing, sneezing, singing, or speaking. When inhaled, the majority of bacilli are trapped in the upper airways and expelled by ciliated mucosal cells, however, droplets of 1 to 5 μm in diameter can reach the alveoli (Brahmer & Small, 1998). There, nonspecifically activated alveolar macrophages (AMs) ingest the bacilli. AMs secrete a number of cytokines: interleukin (IL) 1, IL-6 and tumor necrosis factor α (TNF-α) which contribute to the killing of mycobacteria, the formation of granulomas, and a number of systemic effects such as fever and weight loss. The bactericidal activity of the AMs and the number of bacilli and their rate of multiplication determine the final outcome (Brahmer & Small, 1998). If the bacillus is not destroyed, it multiplies and eventually kills the AM (Dannenberg, 1999). Nonactivated monocytes attracted from the blood-stream to the site by various chemotactic factors ingest the bacilli released from the lysed AMs. At the same time bacilli are transported to the regional lymph nodes, enter the bloodstream, and can establish sites of infection in other portions of the lungs and other organs (Brahmer & Small, 1998; Raviglione & O'Brien, 2001).

The initial stages of infection are usually symptomless (Fitzpatrick & Braden, 2000). However, adequate proliferation of organisms in either pulmonary or extrapulmonary sites may cause a clinically evident illness. Clinical illness soon following infection is classified as *primary TB*. Children up to the age of 4 years and immunosuppressed persons such as those with HIV infection are much more likely to develop TB soon after infection has occurred (Raviglione & O'Brien, 2001). In the vast majority of instances, however, a relative equilibrium is established between the host and the pathogen until specific immunity develops, usually within 4 to 8 weeks of the original infection (Brahmer & Small, 1998). This is marked by the conversion of the tuberculin skin test from negative to positive.

With the development of specific immunity and accumulation of large numbers of activated macrophages, granulomatous lesions are formed. Initially, a "tissue-damaging response" develops, which is the result of a delayed-type hypersensitivity (DTH) reaction to various bacillary antigens; it destroys nonactivated macrophages that contain multiplying bacilli and causes local solid necrosis in the center of the lesion (Raviglione & O'Brien, 2001). Although *M. tuberculosis* can survive, its growth is inhibited within this solid necrotic environment with low oxygen tension and low PH. At this point, some lesions may heal by fibrosis and calcification, while others undergo further evolution. The second host immune response, "macrophage-activating response," is a cell-mediated immunity (CMI) phenomenon (Brahmer & Small, 1998; Raviglione & O'Brien, 2001). In the majority of infected individuals, sensitized T-cells produce and release lymphokines on exposure to the antigens of the tubercle bacillus. These lymphokines attract and activate macrophages, which in turn become much more potent in killing and digesting the microorganism. These activated cells aggregate around the lesion's center and effectively neutralize tubercle bacilli without causing further tissue destruction (Dannenberg, 1999; Raviglione & O'Brien, 2001). In the central part of the lesion, the necrotic material resembles soft cheese (caseous necrosis). Even when healing occurs, viable bacilli may remain dormant within macrophages or in the necrotic material for years or even throughout the patient's lifetime. These healed lesions in the lung parenchyma and hilar lymph nodes may later undergo calcification. In a minority of cases, the CMI is weak, and *M. tuberculosis* growth can be inhibited only by intensified DTH reactions, which lead to further tissue destruction (Dannenberg, 1999; Raviglione & O'Brien, 2001). At the center of the lesion, caseous material liquifies. Bronchial walls as well as blood vessels are invaded and destroyed, and cavities are formed. The liquified caseous material, containing large numbers of bacilli, is drained through bronchi. Within the cavity, tubercle bacilli multiply well and spread into the airways and the environment through expectorated sputum.

Individuals who have a successful immune response usually harbor live tubercle bacilli in the parts of the body seeded by the early dissemination of the organism, so-called *latent TB infection* (Raviglione & O'Brien, 2001). In latent TB Infection, the dormant bacilli may persist for years before being reactivated to produce *secondary TB.*

The risk of developing active TB during the first 2 years after being infected is about 5 percent (Enarson & Murray, 1995). Another 5 percent of infected persons will develop active TB at a later time in their lives. It is believed that approximately 90 percent of individuals with latent infection will never develop TB (Raviglione & O'Brien, 2001). A recent *in vitro* study (Schoolnik, 2001) indicated that nitrous oxide (NO) in the human's activated macrophages might induce TB into a latent stage. Schoolnik suggested that NO, through a switch in carbon and energy metabolism of the organism, signals a change in the replication rate of *M. tuberculosis* in the host from a rapid replication state to a very slow or dormant state. Previously, Rockett and coworkers (1998) documented that there was plenty of evidence that NO produced in mice has an important role in host defence. However, they believed that human macrophages could not produce significant levels of NO. Hence, further research is needed to resolve the controversy to delineate the role of NO in humans.

Reinfection of a previously infected individual with new strain of *M. tuberculosis*, may increase the risk of development of disease. In persons infected with HIV, reinfection occurs more frequently (Small et al., 1993), however, there is bacteriologic and clinical evidence that reinfection also occurs in HIV negative individuals (Bates et al., 1976; du Plessis et al., 2001), but only rarely.

TB in children is non-infectious in 90 to 95% of cases and preadolescent cases play only a minor role in the spread of the disease (Donald & Beyers, 1998). The interval between infection and disease is usually short–weeks to months in children, while it is usually long–years to decades in adults (Starke, 1999). Children are more prone to extrapulmonary TB (EPTB), particularly TB meningitis and miliary TB, with the highest incidence rates in very young children (Donald & Beyers, 1998).

EPTB generally is a consequence of lymphohematogenous dissemination of *M. tuberculosis* during primary infection (Fitzpatrick & Braden, 2000). Persons with immunosuppressive illnesses are more likely to have EPTB. As a result of hematogenous dissemination in HIV-infected individuals, EPTB is seen more commonly today than in the past (Raviglione & O'Brien, 2001). The most common forms of EPTB are TB of lymph nodes (41%), pleural (21%), bone and joint (11%), genitourinary (7%), meningeal (5%), and peritoneal (4%) (Fitzpatrick & Braden, 2000). Other sites account for 11% of

cases. EPTB constitutes 15% to 20% of active cases and concomitant pulmonary and extra-pulmonary disease occurs in approximately 7% of cases.

Extensive hematogenous dissemination of tubercle bacilli may cause miliary TB, which produces a miliary pattern on chest radiographs or in biopsy specimens from organs (Fitzpatrick & Braden, 2000). Miliary TB most frequently occurs among high-risk groups, such as persons with HIV infection or other immunosuppressive illnesses, children younger than 12 months (Kondo et al., 2001) and persons who abuse alcohol (Fitzpatrick & Braden, 2000). Miliary TB accounts for about 0.2% of cases.

3) Diagnosis

The term "TB infection" refers to a positive tuberculin skin test (TST) with no evidence of active disease. The TST involves intradermal injection of 5 tuberculin units (usually 0.1 ml) of purified protein derivative (PPD) into the volar or dorsal area of the forearm. Induration is then assessed at 48 to 72 hours and interpreted based on risk groups. For instance, induration ≥15 mm is considered positive in persons with no risk factors (see Fitzpatrick and Braden, 2000, for tuberculin positivity criteria). Nevertheless, in some truly infected people, positive reaction may not be detected (anergy). This usually happens in: 1) severe forms of TB; 2) when the infection is very recent (less than 4 weeks); or 3) when the immune system is compromised (Raviglione & O'Brien, 2001). Of greater importance is the relatively poor specificity of the test. Persons infected with nontuberculous mycobacteria and those recently vaccinated with Bacille Calmette-Guérin (BCG) will often have a positive reaction. A positive skin test in a person vaccinated with BCG more than 5 years previously should be considered as caused by *M. tuberculosis* infection and not attributed to BCG vaccination (Anonymous, 2000). Nevertheless, it has been argued that 10 to 25 percent of individuals who received BCG vaccination between the age of 2 and 5 years will persistently have a positive tuberculin reaction even 20 to 25 years later (Canadian Immunization Guide, 2002)

"TB disease" refers to cases that have a positive acid-fast smear or culture for *M. tuberculosis* or radiographic and clinical presentation of TB. Although acid-fast examination is less sensitive than culture and it is not species specific, it is fast and inexpensive (Fitzpatrick & Braden, 2000). However, the "gold standard" for the

diagnosis of TB is isolation of the bacilli in a culture specimen. In EPTB, site-specific tissue or fluid samples or both are submitted for smear, culture, and histologic analysis.

Signs and symptoms of TB depend on the organs involved. In general, there are some systemic symptoms and signs that may be observed with any site involvement, such as fatigue, anorexia, weight loss, and persistent low-grade fever (Raviglione & O'Brien, 2001). Ten percent to 20% of persons with active TB, particularly persons with early disease and elderly persons, may have no symptoms (Fitzpatrick & Braden, 2000). Cough is the most prominent symptom of PTB, particularly for immuno-competent patients, although it may go unrecognized in patients with a history of chronic lung disease such as chronic bronchitis (Fitzpatrick & Braden, 2000).

4) Treatment

Standard treatment for TB caused by drug-sensitive organisms is a 6-month regimen divided into: 1) an initial or bactericidal phase for 2 months; and 2) a continuation or sterilizing phase for 4 months (Raviglione & O'Brien, 2001). The initial phase consists of first-line drugs including: isoniazid (INH), rifampin (RMP), pyrazinamide (PZA), and ethambutol (EMB) (or streptomycin (SM)), followed by INH and RMP in the continuation phase. Patients must be treated to completion to avoid development of resistant disease (Raviglione & O'Brien, 2001). For further details and other regimens see Raviglione & O'Brien, 2001. In very young children, ethambutol should be avoided, because it may be difficult to monitor for ocular toxicity (Patterson et al., 1999).

Without treatment, about one–fourth of the patients die within 2 years and 50% to 60% die within 5 years (Enarson & Murray, 1995). The highest death rates are observed among smear-positive patients. Twenty-five percent are spontaneously cured, and 25% remain smear positive and are a source of infection within the community. With appropriate chemotherapy, up to 90% are cured; the fatality rate is less than 10%; and only 2% remain smear positive (Enarson & Murray, 1995).

Strains of *M. tuberculosis* resistant to each anti-TB drug occur by spontaneous point mutations in the mycobacterial genome (Raviglione & O'Brien, 2001). A patient is said to have drug-resistant TB if the strain causing the disease is resistant to at least one of the first-line anti-TB drugs (Talenti & Iceman, 2000). Resistance is defined as "primary or

initial" when identified in a person who has not previously been treated, and "secondary or acquired" when failure of previous treatment due to an inappropriate regimen or non-adherance to treatment leads to resistance. Multidrug-resistant TB (MDR-TB) has been defined as resistance to at least INH and RMP. In MDR-TB involving resistance to 5 to 7 drugs, the cure rates are approximately 56% (Globe et al., 1993).

2.1.3 Epidemiology

1) Measurements of tuberculosis in the community

Before the introduction of chemotherapy and under stable conditions, the prevalence and the mortality rate of TB were good indicators of the size and trend of the TB problem in a community (Enarson & Murray, 1995). In the early part of the twentieth century, the tuberculin test, a third measure of the burden of TB, became available (Enarson & Murray, 1995; Maher & Raviglione, 1999). To estimate the incidence of the disease, the annual risk of TB infection (ARTI or ARI) is the most widely used method, which is based on tuberculin surveys that measure the prevalence of TB infection by age within a community. The ARI, however, has some limitations such as: 1) problems with standardization of tuberculin; 2) problems with standardization of the technique of administration, reading and interpretation of results; and 3) cross-reactions with BCG and other mycobacteria. An ARI of 1% corresponds approximately to an annual incidence of active cases of 100 per 100,000 population per year. When the ARI is above 1000, TB is considered as "epidemic;" above 100, the population can be defined as at "high risk" for TB; below 10, the population can be defined as at "low risk" for TB; below 1, entering the elimination phase of the program; and at 0.1, TB can be defined as eliminated (Enarson & Murray, 1995).

Case notification data may also provide useful information for obtaining incidence rates by age, gender, and risk groups (Maher & Raviglione, 1999). In the few countries where validity and completeness of the registration of cases are ensured, case notification closely approximates the true incidence of tuberculosis. However, in those countries where only a minority of the population has access to effective health service, case notifications often represent only a fraction of the true incident cases.

2) Current global estimate

The most recent published estimates of the global burden of TB are based on data available from a global surveillance and monitoring project conducted by WHO in 1997 (Dye et al., 1999). The main purpose of this project was to estimate the incidence and prevalence of TB, and mortality for 1997 (Table 2.1). To obtain these estimates three different methods were used based on i) cases notified to WHO; ii) ARI data derived from tuberculin surveys; and iii) data on prevalence of smear-positive pulmonary disease from prevalence surveys.

In 1997, nearly 8 million new cases of TB (136 per 100,000 population) worldwide were reported to WHO (Dye et al., 1999). About 3 million of them were cases of infectious pulmonary disease (smear positive), and the existing cases of TB totaled 16.2 million. Twenty-two countries in Asia and Africa accounted for about 80% of all incident TB cases. It was estimated that nearly 2 million deaths occurred from TB and the global fatality rate was 23 percent. Based on progression of the disease during the last few years, estimated by WHO (2001a), 10.2 million new cases are expected in 2005.

Table 2.1: Estimates of TB burden by WHO region in 1997 (adapted from Dye et al.,1999)

Data are given as rate per 100,000 population							
WHO Region	Population, thousands	Incidence	Prevalence	Infection prevalence, %	TB death rate	Case fatality, %	TB/ HIV*
Africa	611,604	259	384	35	88	34	1,194
The Americas	792,330	52	72	18	8	16	64
Eastern Mediterranean	475,415	129	258	29	30	23	23
Europe	870,386	51	73	15	7	14	10
Southeast Asia	1,458,274	202	524	44	48	24	162
Western Pacific	1,641,179	120	230	36	22	18	19
All Regions	5,849,188	136	277	32	32	23	183

* TB/HIV: Prevalence of TB/HIV coinfection

The AIDS epidemic has had a profound impact on TB over the last decade. The HIV pandemic has increased the number of TB cases in both developing and developed countries (Small & Selcer, 1999). The recent spread of the HIV epidemic in sub-Saharan Africa has been accompanied by a notable increase in the number (double or triple in a period of 10 years) of TB cases (Raviglione & O'Brien, 2001). In 1997, the prevalence of *M. tuberculosis* coinfected with HIV was 0.18% in the world and 8% of incident TB cases had HIV infection (Dye et al., 1999).

In addition, drug-resistant TB resulting predominantly from inappropriate use of currently available medications coupled with poor infection control measurement contribute to a worrying picture (Talenti & Iceman, 2000). Based on a worldwide survey (WHO/IUATLD, 2001) conducted between 1994 and 1996, the prevalence of resistance to at least one anti-TB drug among new cases ranged from 1.7% in Uruguay to 36.9% in Estonia. MDR-TB ranged from 0% in eight geographical settings to 14.1% in Estonia. Higher proportions of resistance to at least one drug were observed in foreign–born TB patients compared with non–foreign–born patients in Canada, Denmark, Finland, Germany, Iran, Netherlands, Sweden, England & Wales, and the United States. Encouragingly, France and the United States reported a significant downward trend in MDR-TB prevalence.

In 1990, it was estimated that 1.3 million cases of childhood TB occurred worldwide and that 450,000 children died of TB (Donald & Beyers, 1998). In a more recent study, the number of annual deaths from TB in children was estimated at 130,000 (Donald & Beyers, 1998). Data concerning TB infection in children are not available. On average, a reactive tuberculin test is observed in about 30% to 50% of all household contacts of a TB case (Starke, 1999). Although it is difficult to assess the global burden of TB in children, mainly due to difficulties in diagnosis, it is still feasible to estimate the infection rate in children. This could be a good estimator of future cases of tuberculosis.

3) Population trends

Age, gender, and marital status

The median age for TB has increased noticeably in developed countries where, as a result of a rapid decline in the risk of infection, the infected population segments have

become increasingly older (Haro, 1998). In contrast, in many developing countries, TB still peaks in young adults (IUATLD, 1998). The greatest risk of disease is during infancy and adolescence (Enarson & Murray, 1995). The lowest rates of TB, in any population, are usually observed in children between the ages of 5 and 14. In virtually all countries, the case rates are higher among males than among females (Holmes et al., 1998). Among women, the incidence peaks at 25 to 34 years of age. In this age group rates are usually higher among women than among men. Some studies indicate that women may have higher rates of progression to disease and a higher case fatality rate in their early reproductive ages (Holmes et al., 1998). The gender ratio for TB in children is about 1:1 (Starke, 1999). In a study in Denmark (Dutt & Stead, 1999), it was shown that the TB rate was higher in single, widowed, and divorced individuals compared with married individuals, and that the rate was always higher among men than among women.

In 1995, TB case rates in Canada were highest for those 65 years of age and older (Long et al., 1999). The case rates in the years 1987 to 1995 indicate that the rates are falling among older Canadians and increasing among younger Canadians (except those aged 25-34 years). Most cases under 5 years of age were in status Indians. In status Indians and foreign-born people, the highest rates were among those younger than 44, whereas in "other" Canadians, the highest rates were among those older than 44 years. Tuberculosis is more common among males. However, data from 1987 to 1995 show that the number of cases in males has decreased by 7.8%, while the number of cases in females has increased by 5.4%.

Race/Ethnicity

In some races (or subraces) such as blacks, Inuits and American Indians, the disease may have a more progressive course (Moulding, 1999). In studies of both nursing homes and prison populations, blacks were twice as likely to become infected than whites under similar exposure, and they also experienced more rapidly progressive disease (Stead et al., 1990).

In 1995, the overall risk of TB in Canada was 12 times higher for people born outside Canada and 25.8 times higher for status Indians than for "other" Canadians (Long et al., 1999).

Socioeconomic status, Malnutrition

Poverty has always been strongly associated with the incidence of TB (Enarson & Murray, 1995). Low socioeconomic indicators tend to result in crowded living conditions, resulting in a higher risk of transmission of infection and a generally higher incidence of disease. Poverty may also reduce access to health care services (Bergner & Yerby, 1968), thus prolonging the period of infectiousness of TB patients, and further increasing the risk of infection among case contacts. In addition, poverty is linked to malnutrition. Morbidity and mortality from bacterial respiratory infections are worsened by poor nutritional status (Berkowity, 1992). How exactly malnutrition favors the development of the disease is unknown. Enarson and Murray (1995) suggested that various components of the immune system, including T-lymphocyte function and CMI, are impaired.

Geographic distribution

The worldwide distribution of tuberculosis, according to TB case notification to WHO in 1997, was as follows: 39% in South-East Asian Region; 25% in the Western Pacific Region; 15% in the African Region; 3.5% in the Eastern Mediterranean Region, 10.5% in the European Region; and 7.5% in the American Region (Dye et al., 1999).

Based on data in 1995, over 75% of TB cases in Canada were observed in Ontario, Québec and British Columbia, where most new immigrants choose to live (Long et al., 1999). In these provinces, the highest numbers were reported from the large urban areas of Toronto, Montreal and Vancouver, possibly due to higher numbers of immigrants and inner city poors. In contrast, TB has almost been eliminated from the Atlantic provinces, where there were only 13 smear-positive cases of TB.

2.1.4 Risk factors

There are two distinct stages in the development of TB: infection and disease. Each has its own set of risk factors that must be considered separately. While the risk of acquiring infection appears to depend mostly on exogenous factors, risk of developing the disease after being infected seems to be more related to endogenous factors such as genetic factors (Raviglione & O'Brien, 2001). Moreover, a growing body of evidence

indicates that both the risk of acquiring infection and the risk of developing the disease may be partially determined by genetic factors (Kwiatkowski, 2000).

1) Infection stage (Table 2.2)

Most transmission of tubercle bacilli from one person to another is by the airborne route. The infectiousness of a TB case is determined by the type and extent of the disease (Enarson & Murray, 1995): *M. tuberculosis* is virtually always transmitted to other individuals from patients with pulmonary TB, as opposed to those with extra-pulmonary involvement; patients who are smear positive are more infectious than those who are smear negative. However, those who are smear negative but culture positive or those who are smear negative but with chest radiographs that show evidence of tuberculosis can infect others (Behr et al., 1999).

There are also other factors that determine the probability of transmission of *M. tuberculosis*. Environmental factors influence the viability of the bacilli in the droplet nuclei and the concentration of droplet nuclei in the air: sunlight and ultraviolet light destroy the organism; the bacilli do not survive in dry environments (Dannenberg, 1999). The duration of exposure and degree of ventilation of the environment also affect the likelihood of becoming infected with the organism (Enarson & Murray, 1995). In crowded places such as jails, homeless shelters, and nursing homes, one unrecognized TB case might endanger many people. Although household contacts are at the highest risk of infection, TB infection can also be acquired in more casual settings such as a bar (Tabet et al., 1994), a church (Mangura et al., 1998), or a classroom (Curtis et al., 1999). Some evidence shows that *M. tuberculosis* can be transmitted during long (i.e. more than 8 hours) flights from an infectious source (a passenger or crew member) to other passengers or crew members (WHO, 2001b).

Not every previously unexposed person acquires TB infection after inhaling the tubercle bacillus as only one fourth of close household contacts of smear-positive index cases were estimated to have been infected by their exposure (Enarson & Murray, 1995). This resistance to infection in previously uninfected individuals must depend largely on innate defense mechanisms. In his observations, Stead (1992) suggested that genetic factors play an important role in innate resistance to infection by *M. tuberculosis*. A vast

majority of genetic variants may regulate specific immunological, physiological and metabolic mediators (Kwiatkowski, 2000). Other factors that impair the immune system, such as drug abuse, alcoholism, and HIV infection, may increase the risk of contracting the infection (Reichman et al., 1979; Enarson & Murray, 1995). The summary of selected studies which have examined the relationship between different risk factors and risk of acquiring infection are described in tables 2.4 and 2.5.

Table 2.2: Factors determining the likelihood of transmitting TB infection

Characteristics of the case
- Presence of cough or other mechanism to produce droplet nuclei
- Type and extent of the disease: cases of PTB who are smear positive are more infectious
- Number of organisms in sputum, viability and virulence of organisms

Environmental factors
- Ventilation: the turnover of air in a specific space affects the concentration of organisms
- Crowded places such as jails, homeless shelters, and nursing homes: groups of individuals are exposed to TB case(s)
- Humidity: *M. tuberculosis* does not survive in a dry environment
- Exposure to sunlight/ ultraviolet light destroys the organism

Contact's characteristics
- Amount of time spent sharing air with case(s): household contacts are at highest risk
- Living in high-prevalence countries
- Aboriginal (Inuit or First Nation) background
- Travel to high prevalence countries
- Older age
- Drug abuse, alcoholism
- Health care occupation and other occupational contact with a high-prevalence group
- Innate defense mechanism

2) Disease stage (Table 2.3)

The host's immune system, particularly CMI, is the most important factor that determines whether a new or old infection will progress to disease (Enarson & Murray, 1995). Youmans (1979) stated concisely, "progression of tuberculosis can only take place in the presence of inadequate cellular immunity to infection."

The risk of disease in people who react to tuberculin and whose weight is less than 90% of ideal body weight (body mass index of less than 20) has twice the risk than

among those whose weight is in the ideal range, and 4 times more than those whose weight is more than 110% of the ideal (Canadian Lung Association, 1996).

It was estimated by CDC (1997) that over 30 million people worldwide were infected with HIV, and most of them reside in highly endemic areas for TB. The association between TB and HIV is not surprising given that the cellular immune system is the major target of HIV and the first host defense against TB (Small & Selcer, 1999). In individuals with latent TB who become infected with HIV, the chance of developing active TB is seven to 10 percent per year as opposed to a 10% lifetime risk for those with a normal immune system (American Lung Association, 2002). Approximately 40% of individuals with HIV infection who become exposed to *M. tuberculosis* will develop active TB, and more quickly, than do individuals with a normal immune system (Daley et al., 1992).

Table 2.3: Factors determining the likelihood of progression from infection to disease

Presence of inadequate immune response, especially cell-mediated immunity which is associated with:

- Genetic factors
- Certain tumors, particularly T-cell lymphomas
- Renal failure requiring dialysis
- Undernutrition
- Age
- High-dose corticosteroids
- Immunosuppressives
- Diabetes, especially the insulin dependent form
- Pulmonary silicosis
- Gastrectomy
- Pregnancy
- Recent tuberculin conversion
- And most notably, HIV infection

Tables 2.4 and 2.5 provide a summary of selected studies that have examined the correlation between environmental (Table 2.4) and genetic (Table 2.5) risk factors, and tuberculosis. These selected studies include one representative study for each gene and other risk factors, from different parts of the world, and mostly in recent years.

Table 2.4: Selected studies which have examined the relationship between different non-genetic risk factors and TB

Authors, Year, Place	*Patients*	*Design*	*Exposure*	*Risk estimate, (CI)*	*Result*
Cowie, 1994, South Africa	1,153 older gold miners with and without silicosis (TB free)	Longitudinal study (7 yrs)	Silicosis	OR = 2.8 (1.9 – 4.1)	There is a high risk of pulmonary and extra-pulmonary TB in men with silicosis
Mangura et al., 1998, U.S.A.	5 cases of TB 301 healthy persons All members of a church	Tuberculin test, DNA fingerprints in an outbreak of TB	Singing and location of ventilation outlet in the church	RR = 2.04 (1.17 – 3.56)	Tenors were more likely to be tuberculin reactors
Behr et al., 1999, U.S.A.	71 clusters of TB patients infected with matched fingerprint strains 43 clusters with positive smears 28 clusters with negative smears	A part of ongoing study of the molecular epidemiology of TB	Result of sputum examination (positive vs. negative)	Relative transmission rate: (negative/ positive) = 0.22 (0.16 – 0.32)	Although smear-positive individuals are the most infectious patients, patients with negative smears (culture positive) are also responsible for transmission of the bacilli
Davies et al., 1999, U.K.	All cases of TB	Retrospective study from 1853 to 1910	Social condition, based on earnings and population density per house	Annual decline in mortality rate = 1.71% (0.77 – 2.63)	Improvement of social conditions does not provide the total explanation, and natural selection, probably played a role
Cobelens et al., 2000, Netherlands	988 BCG-naïve immunocompetent travellers	Multicentre, prospective cohort study	Travel to high risk countries for TB (ARI ≥ 1%)	Overall incidence rate = 3.5 per 1000 person-months of travel (2.0 – 6.2)	The risk of TB infection in long- term travellers to high-risk countries is substantial
Barr et al., 2001, U.S.A.	3,343 cases of TB	Longitudinal study of TB incidence adjusted for AIDS, proportion of foreign born, and race/ethnicity	Neighborhood poverty	RR = 1.3 (1.30 – 1.36) per 10% increase in poverty	Neighborhood poverty was strongly associated with incidence
John et al., 2001, India	1251 cases of kidney transplant	Prospective study Analysis: Kaplan-Meier	Post-transplant cyclosporine therapy Vs. prednisolone plus azathioprine therapy	≤ 6 months of therapy RR = 2.5 (p = 0.031) ≤ 12 months of therapy RR = 1.9 (p = 0.043)	Cyclosporine therapy after kidney transplant was found to be associated with early post-treatment TB comparing with treatment by prednisolone plus azathioprine

Authors, Year, Place	*Patients*	*Design*	*Exposure*	*Risk estimate, (CI)*	*Result*
Czeizel et al., 2001, Hungary	22,865 newborns with congenital abnormality 38,151 newborns without congenital abnormality	Population-based case-control study	Oral anti-TB drug during the second and third months of gestation	OR = 0.6 (0.3 – 1.3)	Maternal exposure to oral anti-TB drugs during pregnancy did not show a detectable teratogenic risk to the fetus
Kondo et al., 2001, Tokyo	Children with TB 45 infants 12 mo or less 31 children 13 to 35 mo	Comparison study of clinical data	TB contact	RR = 6.0 p = 0.054	DTH and CMI to MTB among infants may be lower than children aged 13-35 mo
Saiman et al., 2001a, U.S.A.	Children 1-5 yr 96 cases of latent TB 192 controls without latent TB (TST = 0)	Case-control study matched on age and clinic	Foreign birth BCG immunization Foreign travel (study subjects or household members)	OR = 5.5 (2.79 – 12.44) OR = 7.8 (3.03 – 22.86) OR = 1.92 (0.92 – 2.83)	Foreign birth, BCG immunization, and foreign travel (study subjects or household members) were identified to be significant risk factors for a positive TST
Berggren Palme et al., 2001, Ethiopia	Cases: children with TB Controls: children without TB	One year hospital-based prospective case-control study	HIV, in children with TB	OR = 12.7 (2.9 – 55)	There is strong association between TB and HIV infection. In children with TB, risk of HIV infection is considerably higher
Shor-Posner et al., 2002, U.S.A.	259 HIV seropositive drug users followed for 2 years 12 cases of mycobacterial disease (9 TB and 3 other mycobacterial diseases) occurred, which were then compared with 32 controls matched on age, sex and HIV status	Case-control study following a two year double-blind, placebo-controlled clinical trial	Selenium level (≤135/>135 μg/l) BMI (<19/≥19 kg/m^2)	RR = 3.0 p = 0.015 RR = 2.0 p = 0.05 After control for CD4 count and anti-retroviral therapy	Selenium level probably has important effects on the pathogenesis of TB and other mycobacterial diseases

Table 2.5: Selected studies which have examined the relationship between different genetic risk factors and TB

Authors, Year, Place	***Patients***	***Design***	***Exposure***	***Risk estimate, (CI)***	***Result***
Stead et al., 1990, U.S.A.	21,010 white persons and 4388 black persons in 165 nursing homes All initially TST negative	Incidence of tuberculin conversion (TST≥12mm)	Tubercle bacilli	RR = 1.9 (1.7 – 2.1)	The risk of acquiring infection in black individuals is higher than white individuals under similar environment
Shaw et al., 1997, Northern Brazil	98 multicase TB pedigrees including 704 inviduals (205 nuclear families)	Linkage study, combined with segregation analysis	*NRAMP1* *TNF* (tumor necrosis factor)	General two-locus model $0.01 < p < 0.05$	TB susceptibility in Northern Brazil is under oligogenic control (neither is a major gene)
Bellamy et al., 1998a, West Africa	410 adults with smear-positive PTB 417 healthy controls	Case-ethnically matched-control study	Heterozygote for two *NRAMP1* polymorphisms in intron 4 and 3′ untranslated region of the gene Compared with those with the most common *NRAMP1* genotype	OR = 4.07 (1.86 – 9.12)	Genetic variation in *NRAMP1* affects susceptibility to TB in West Africans
Selvaraj et al., 1999, India	202 cases of PTB 109 controls without TB	Case-control study	Functional mutants of *MBP* (MBP 52, 54, and 57 wild and mutant alleles)	OR = 6.5 p = 0.008	Functional mutants of MBP are associated with PTB
Dubaniewicz et al., 2000, Poland	31 cases of PTB 58 healthy controls (volunteers)	Case-control study	HLA-DRB1*13 and HLA-DRB1*16	RR = 0.04 $p < 0.001$	Presence of HLA-DRB1*16 alleles may increase the risk of developing PTB, whereas HLA-DRB1*13 alleles may be resistant to TB
Wilkinson et al., 2000, U.K.	Asian population of Gujarati origin 126 cases of TB 116 healthy contacts	Hospital-based case-control study	Vitamin D deficiency Undetectable serum vit D Interaction of vit D deficiency and genotype *TT/Tt*	OR = 2.9 (1.3 – 6.5) OR = 9.9 (1.3 – 76.2) OR = 2.8 (1.2 – 6.5)	Vit D deficiency may contribute to the high occurrence of TB in this population. Polymorphisms in the *VDR* gene also contribute to susceptibility when considered in combination with vit D deficiency

Authors, Year, Place	*Patients*	*Design*	*Exposure*	*Risk estimate, (CI)*	*Result*
Lawn et al., 2000, Africa	105 subjects: 30 only PTB 20 only HIV+ 20 PTB and HIV+ 30 healthy controls	Prospective study Measurement of serum level of soluble CD14 (sCD14)	PTB and HIV	Ratio of levels, compared with healthy group: In only PTB = 3.1 ($p< 0.0001$) In only HIV+ = 4.1 ($p< 0.01$) In PTB and HIV+ = 8.7 ($p< 0.0001$)	The serum level of sCD14 which is a macrophage activation marker, is significantly higher in cases of TB and/or HIV than healthy population Also reduction of sCD14 after treatment was started did not show the same pattern observed with lymphocyte markers (sCD25)

2.1.5 Prevention and control

The global burden of tuberculosis will continue to increase without effective TB control programs. The commitment of governments is necessary to provide adequate funds to implement effective programs (Maher & Raviglione, 1999).

By far, the best way to prevent tuberculosis is the reduction of the source of infection through rapid diagnosis of infectious cases and appropriate treatment until cure (Raviglione & O'Brien, 2001). Prompt, appropriate drug treatment has the potential to reduce the TB burden by more than 50% in 10 years (Dye, 2000). The WHO strategy to control TB is directly observed treatment, short-course chemotherapy (DOTS) (Maher & Raviglione, 1999). One of the foundations of DOTS is the administration of standard short-course chemotherapy under direct observation of health care workers. This strategy ensures adherence to treatment and increases the cure rate. However, there are some factors that inhibit the expansion of DOTS: poor financial support for TB control, poor organization and management of health services, the HIV epidemic, and the rise of MDR-TB (Nunn et al., 2002).

In both developed and developing countries, a significant proportion of TB cases arises from the pool of persons with remote (acquired in the past) infection with *M. tuberculosis*. Therefore, treatment of infected persons has great importance in reducing the current TB morbidity (Raviglione & O'Brien, 2001). Preventive therapy or chemoprophylaxis is the administration of a chemotherapeutic agent(s) for a period of 6 to 12 months to prevent the development of active disease in infected individuals. It is assumed that administration of a chemoprophylactic agent results in sterilization of organisms from infected persons (Bates, 1995). A major limitation to this program is the requirement that an asymptomatic person take medication for a relatively long period of time (O'Brien, 1998).

A total of 19 randomized placebo-controlled trials, conducted in seven countries (both developed and developing), have investigated the effectiveness of preventive therapy (Comstock & Geiter, 1999). In these studies that involved more than 135,000 subjects at risk of TB, including contacts of active cases, tuberculin test converters, and institutionalized patients with mental disease, the decrease in the disease rate ranged

between 25% and 92%. Nevertheless, when analysis was restricted to persons who were apparently compliant with medication, the protective efficacy was approximately 90%.

An additional prevention strategy is BCG vaccination. BCG was derived from an attenuated strain of *M. bovis* and was first administered to humans in 1921 (Comstock & Geiter, 1999). Estimates of clinical efficacy of BCG in preventing pulmonary TB have varied from zero protection in the Southern United States and in South India, to approximately 80% in the United Kingdom (UK) (WHO, 2000). It seems possible that this variation is partly due to interactions between the vaccine and strains of mycobacteria in the environment (Brandt et al., 2002) that control the multiplication of the vaccine. In their study in Northern Malawi (the region with no protection for PTB), Brandt and colleagues demonstrated that prior exposure of laboratory mice to specific strains of mycobacteria affects the efficacy of BCG. In this area, two strains of *M. avium* complex were found to completely block BCG activity.

Although the overall efficacy of BCG in protecting against adult forms of the disease is questionable, data show that BCG is successful in protecting against TB meningitis and miliary TB (estimated 46-100% protection) in the first year of life (WHO, 2000). BCG vaccine is recommended for routine use at birth in all countries with high TB prevalences (WHO, 1995a). Babies born to mothers who develop active TB shortly before or shortly after delivery should be given prophylactic chemotherapy. BCG can then be given after chemotherapy ends, when BCG provides protection (WHO, 1998b). In individuals with depressed cellular immunity such as those with HIV infection, BCG vaccination is contraindicated (WHO, 1995b).

In countries where the prevalence of tuberculosis is low, BCG may be offered only to high-risk groups such as immigrants from high prevalence areas (WHO, 2000). It has been argued that discontinuation of BCG in low prevalence countries would facilitate the use of tuberculin testing for contact tracing, source identification and selection of individuals for preventive therapy (IUATLD, 1994). It seems a valid argument, but many years must pass after discontinuation of routine BCG vaccination to replace a vaccinated population with an unvaccinated population.

In addition, environmental control methods can be used to reduce the concentration of droplet nuclei in the air in high-risk areas. Environmental control refers to engineering

modalities for removal or disinfection of air containing *M. tuberculosis*. There are inexpensive methods such as maximizing natural ventilation and mechanical ventilation, as well as more costly methods such as ultraviolet germicidal irradiation (WHO, 1999).

The basics of control are summarized in two categories according to TB prevalence (Raviglione & O'Brien, 2001):

a) In low-prevalence countries with adequate resources, the greatest emphasis must be placed on: i) identification and treatment of active cases of TB to cure; ii) screening of high risk groups (such as HIV-seropositive individuals and immigrants from high prevalence countries); iii) preventive therapy for high risk persons with positive PPD; and iv) contact investigation. In the U.S., great attention is given to the transmission of TB in institutional settings such as hospitals, and for the homeless, particularly when associated with HIV infection.

b) In high prevalence countries, most of the efforts must be placed on stopping the transmission of tubercle bacilli in the community by: i) case detection, mainly via passive case-finding (e.g., microscopic examination of sputum for those with cough of more than 3 weeks); ii) administration of DOTS to all sputum smear positive patients, for at least the initial phase of treatment, establishment and maintenance of a system with regular drug supply; and iii) establishment and maintenance of a constant patient evaluation program.

Comstock (1978) suggested that a higher priority for preventive therapy and treatment is required for families where the patients have a positive family history of TB than for those with a negative family history. His argument rests on the assumption that genes contribute to familial susceptibility to tuberculosis.

2.2 GENETICS

2.2.1 Basic concepts of genetics

The basic concepts of genetics have been thoroughly reviewed by several researchers including Cummings, 1988; Ott, 1991; Elseth & Baumgardner, 1995a; and Mehlman & Botkin, 1998.

In summary, an individual's heredity is determined by deoxyribonucleic acid (DNA) which is packaged in 23 pairs of chromosomes in normal somatic (body) cells. The first 22 pairs, called autosomes, have the same structure in males and females, unlike the 23rd pair, which predicts an individual's gender and are not identical. These forty-six chromosomes contain the entire human genetic code (excluding mitrochondrial DNA which is not discussed in this thesis), called "the human genome."

A chromosome is a long stretch of double stranded-DNA (ds-DNA) that is tightly coiled and has a double helix configuration. Ds-DNA is present in the nucleus of virtually every cell in the body. The building blocks of DNA are the four nucleotides (called nucleotides) adenine, guanine, cytosine, and thymine. Combination of these 4 bases are arranged in a long string, and bonded to a complementary string. Each nucleotide is bonded to another nucleotide on the other strand. Together the two nucleotides make up a "base pair." Adenine (A) always pairs with thymine (T), and guanine (G) always pairs with cytosine (C). A DNA sequence is often described as an ordered list of bases (e.g., ACGTTG), which carries the genetic information. Certain sequences of DNA base pairs represent a code for synthesizing proteins. Specific combinations of three nucleotides create "codons," which are the fundamental units of this genetic code. Each codon can be translated into an amino acid. Since there are four different nucleotides, and a codon is made up of three nucleotides, there are $4^3 = 64$ different codons. However, these 64 codons specify only 20 different amino acids, the building blocks of proteins. Moreover, the majority of human DNA does not lead to protein formation. The function of the noncoding DNA in the genome, sometimes referred to as junk DNA, is largely unknown.

The physical site or location of a gene is called its "locus." At any particular locus, there can exist different variants of the DNA sequence; such variations are called "alleles." A "marker" locus is a locus that can be easily genotyped and whose location on a chromosome is known. Since individuals carry two copies of each chromosome, two

alleles may be present at each locus. The allele combination is called a genotype. If the two alleles in an individual are the same, the genotype is called "homozygous" and the individual with such a genotype is said to be a "homozygote." If the two alleles are different, the genotype is "heterozygous" and the individual with such a genotype is called a "heterozygote." The relative frequencies of the different alleles of a gene are called "allele frequencies." A gene is called often "polymorphic" (i.e., having many forms) when its most common allele has a population frequency of less than 95% (99% in some references).

A normal, functional gene can be altered by random errors in DNA replication, or by external exposure to radiation or chemical materials. Some of these alterations will lead to the production of proteins that cause illness or deformity. Such abnormal gene sequences are termed "mutations" when they occur at frequency of <1%. Since each individual has two copies of each chromosome, and therefore two copies of each gene, a mutation in only one copy may produce no ill effects. Conditions that require two abnormal copies of a gene to produce a clinical phenotype are termed "recessive" conditions, whereas conditions in which only one abnormal copy is sufficient to produce the phenotype are termed "dominant" conditions. A gene may interact with other genes and with the environment. Even when mutations in a single gene cause disease, other genes and environmental factors may play a significant role in modifying the severity of the disease.

Sexual reproduction is a process that is mediated by the formation and subsequent union of gametes. All inherited genetic information is contained in two cells, the sperm and the egg. These cells are produced by a special division process called "meiosis," a Greek word meaning to reduce. In meiosis, chromosome pairs are separated from each other and produce gametes or haploid cells, each with only 23 chromosomes. In a zygote, a cell formed by the union of two gametes, the chromosome number is restored to 46, the full complement of genetic information.

Chromosome behavior during the different stages of meiosis is shown in table 2.6. Meiosis consists of two nuclear divisions; each division is composed of 4 stages: prophase, metaphase, anaphase, and telophase. Meiosis I is a reduction division in which homologous chromosomes separate, whereas meiosis II is an equational division which

results in the separation of sister chromatids (the two daughter strands of a duplicated chromosome that are joined by a single centromere). In meiosis I, the number of chromosomes reduces from "2n" to "n", and number of chromatids reduces from "4n" to "2n". In meiosis II, the number of chromosomes remains "n", while the number of chromatids reduces from "2n" to "n". Meiosis produces genetically variable gametes through two mechanisms. This genetic variation is a crucial aspect of sexual reproduction. Firstly, the random alignment of the paternal and maternal homologues at metaphase I produces a random assortment of paternal and maternal chromosomes in the gametes. Secondly, additional variability is produced by "crossing-over" or "recombination". This process involves the physical exchange of chromosomal material between homologous chromosomes (non-sister chromatids) in prophase I (Table 2.6); crossing over redistributes this genetic information and produces new combinations of genes (Figure 2.1). Multiple cross-overs can take place within a single tetrad (four haploid products of a single meiotic cycle, see table 2.6). Therefore, human offspring are never genetically identical to either parent. On average, 1% recombination is equivalent to about a million base pairs of DNA and is defined as one centiMorgan (cM).

Table 2.6: Summary of Meiosis (Adapted from Cummings, 1988)

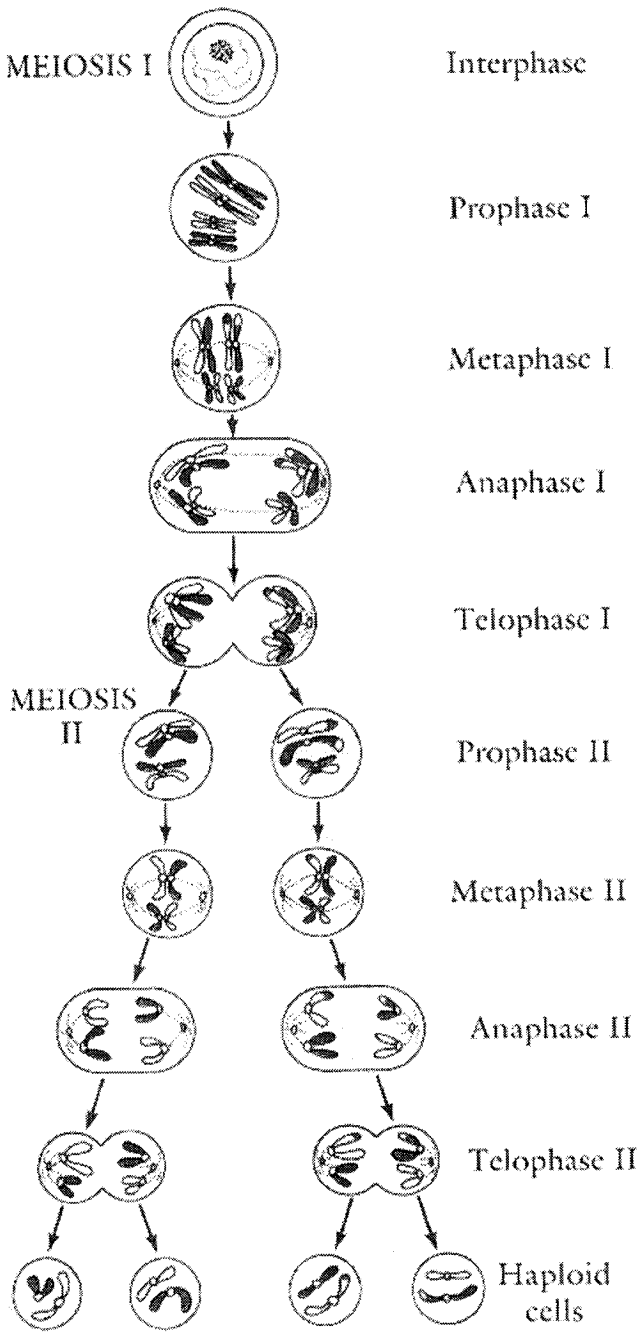
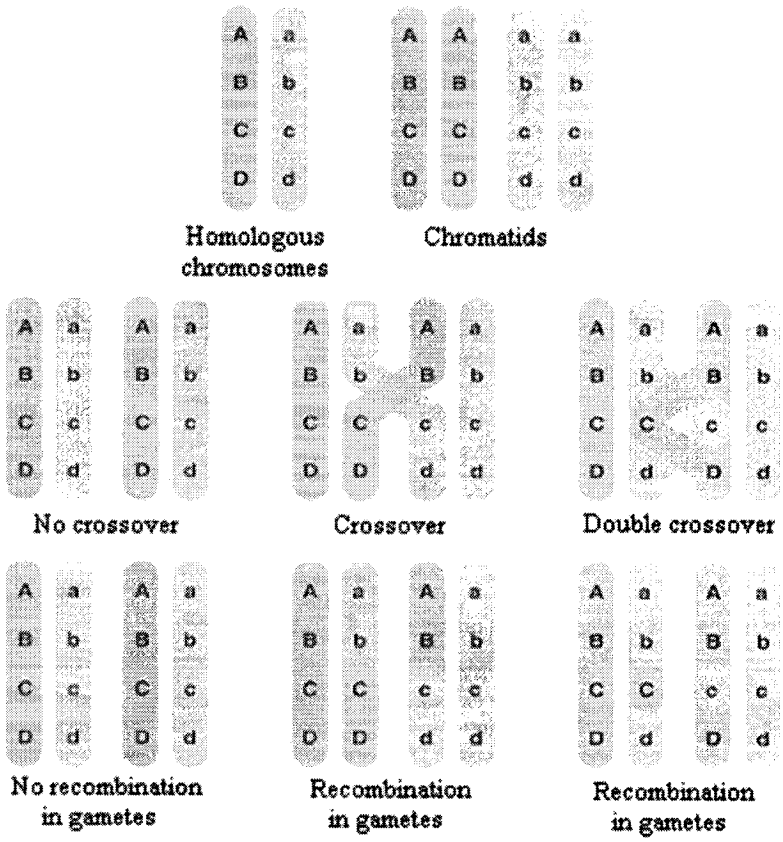
	Interphase: Chromosome replication takes place Prophase I: The duplicated chromosomes become visible; homologous chromosome (dyads) pairs (to form tetrads) and sister chromatids become visible; **recombination** occurs Metaphase I: Paired chromosomes align at equator of cell Anaphase I: Homologous chromosomes (or dyads) separate; members of each chromosome pair move to opposite poles Telophase I: Cytoplasm divides; two cells (each with haploid number of duplicated chromosomes) are produced Prophase II: Chromosomes recoil Metaphase II: Unpaired (haploid number of) chromosomes become aligned at equator of cell Anaphase II: Centromers split, sister chromatids separate Telophase II: haploid number of chromosomes enters each gamete; nuclear membrane reforms; cytoplasm divides; meiosis is complete

Figure 2.1: Crossing-over and recombination. Genetic recombination occurs through crossing-over and results in recombinant and non-recombinant chromosome segments in the gametes (from Raviglione & O'Brien , 2001)



2.2.2 Genetic epidemiology of tuberculosis

In the mid-1980s a new discipline called genetic epidemiology was established (Khoury et al., 1993a), focusing on the role of genetic factors and their interactions with environmental factors on the occurrence of human diseases. Remarkable advances in molecular biology have enhanced this field, and provided a new understanding of the etiology and pathogenetic mechanisms that are critical for disease development. Genetic epidemiologic studies apply both genetic (e.g., typing of genetic markers) and epidemiologic information (e.g., risk factors such as sex and age) to identify genes having a substantial influence on expression of human complex traits (Abel & Dessein, 1998). The strategies of genetic epidemiology include both population surveys (e.g., the study of the role of genetic factors in disease processes) and family studies (e.g., evaluation of familial clustering of disease).

The main goal of genetic epidemiology is to incorporate the role of genetic-environmental factors into the development of intervention and prevention strategies for human diseases (Khoury et al., 1993a).

A genetic basis for interindividual variability in human tuberculosis susceptibility is indicated by racial differences, twin, adoption, genetic linkage and association studies, and segregation analysis, which are discussed below.

1) Natural selection and racial differences

It is hypothesized that certain racial groups are more susceptible to tuberculosis than others. Among 41,000 tuberculin negative nursing home residents in Arkansas, white subjects were significantly more resistant (nearly twice) than black subjects to infection by *M. tuberculosis* (Stead et al., 1990). This natural resistance is reflected in the ability of the macrophages to restrict the growth of intracellular organisms (Crowle & Elkins, 1990). Macrophages from white individuals permit significantly less replication of tubercle bacilli compared to those from blacks. It has been suggested that susceptibility to *M. tuberculosis* infection is independent of factors that influence the progression to clinical disease (Stead et al., 1990). However, certain racial groups such as blacks, Inuit, and American Indians have also shown more rapid progressive disease, with a greater tendency to extra-pulmonary involvement (Moulding, 1999). In contrast, individuals of Caucasian and Mongolian descent experience more chronic disease, primarily affecting the lungs.

Infectious diseases are believed to have been an important force of natural selection in human history. In a retrospective study in England and Wales, Davies and his colleagues (1999) claimed "improving social conditions do not provide the total explanation for the decline in tuberculosis during Victorian times." They believed that through the force of natural selection, susceptibility to TB declined in successive generations. The theory of natural selection argues that a population's resistance to an infectious agent can change over several generations if the infection produces high mortality in the host population before or during the reproductive age. The argument assumes that those alive to reproduce are more resistant to the disease (Stead & Bates, 1995). In this way, the proportion of resistant individuals gradually increases over

generations. When Qu'Appelle Indians first became exposed to *M. tuberculosis* in Saskatchewan in 1890, nearly 10% of the population died annually from the disease. After 40 years, when over half the Indian families had been eliminated, the annual tuberculosis death rate fell to only 0.2% per year, presumably due to the strong selection pressure against TB susceptibility genes (Motulsky, 1960). However, administration of BCG also played an important role. In a recent study, Lipsitch and Sousa (2002) quantitatively evaluated the role of natural selection in mortality from tuberculosis. They demonstrated that although selective pressure plays a role in favor of human genes that confer protection against TB, it is unlikely that the increase in resistance explains the whole decline in TB mortality.

The impact of natural selection on susceptibility to infectious disease should be taken into account when different ethnic groups are compared. Stead (1992) suggested that there is a correlation between a person's resistance level to an infectious disease and the region of his or her ancestry. Ancestors of more resistant populations (e.g., Ashkenazi Jews) were from densely populated areas (e.g., Europe) and experienced high death rates from TB over several centuries, whereas the ancestors of more susceptible populations (e.g., blacks) were from areas once free of TB (e.g., sub-Saharan Africa).

In most infectious diseases such as tuberculosis, only some exposed individuals acquire infection or develop the disease (Enarson & Murray, 1995). This interindividual variability is partly determined by the host-parasite interaction, and involves the innate defense mechanisms of the host (Enarson & Murray, 1995; Marquet & Schurr, 2001). Identification of the important genes that contribute to tuberculosis susceptibility/ resistance will provide a better understanding of the pathogenesis of the disease and assist in developing new treatment strategies (Marquet & Schurr, 2001).

2) Twin studies

A well-conducted twin study has the potential to evaluate the relative contribution of genetic and environmental factors to a given illness. A twin study demonstrates how often the twin of an index case contracts the disease depending on whether the twin is monozygotic (MZ) or dizygotic (DZ). MZ or identical twins arise from the same ova and therefore share 100% of genes, whereas DZ or fraternal twins arise from two separate ova

and on average share only 50% of their genes. In fact, DZ twins are genetically like any other siblings in the family, and may be of the same or opposite sex. When twin pairs are both affected, they are said to be concordant. In other words, concordance is the probability that the pair of an affected twin becomes affected. By comparing concordance rates between MZ and DZ twins, one can estimate the role of genetic factors as well as environmental factors in the etiology of the disease (Khoury et al., 1993b). For instance, a concordance rate of 80% in MZ twins indicates a substantial role for genetic factors, while a concordance rate of 30% indicates a stronger role for environmental factors.

Although the study of twins has significantly helped in understanding complex traits, the validity of the results of many such studies has greatly been questioned. There is a framework consisting of 3 assumptions that can be used in the evaluation of twin studies (Bulik et al., 2000): 1) ascertainment of zygosity: misclassification of twins can bias the measurements and therefore invalidate the findings; 2) equal environmental assumption (EEA): the EEA assumes that MZ and DZ twins are equally exposed to environmental influences that are of etiologic relevance to the disease under study. If the EEA is not correct, the greater resemblance of MZ twins compared to DZ twins could be due to environmental factors; and 3) generalizability: that results obtained from twin studies are applicable to non-twins (singletons). Although there are differences between twins and nontwins (e.g., incidence of congenital malformations), empirical studies have generally shown that twins and singletons have similar risks.

Sampling strategy is another important issue in twin studies. Population-based random sampling methods may be the optimal approach. Then, affected twins can be identified through cross-linking with disease registries. Although these studies are expensive and need the resources of national registers, the bias associated with sampling is minimal (Hawkes, 1997).

Studies of twins have provided some convincing evidence of a hereditary susceptibility to tuberculosis. They have consistently indicated much higher concordance for disease among MZ than among DZ twins (Kallmann & Reisner, 1943; Comstock, 1978; Jepson et al., 2001). Kallmann and Reisner (1943) suggested that because both types of twins are likely to share the same environment, the explanation for a greater difference in disease rates for DZ twins than for MZ twins may be sought in heredity.

Nevertheless, it was argued by Comstock (1978) that the greater physical contact observed between MZ twin-pairs explains, in some degree, the higher concordance for TB of MZ compared with DZ twin pairs. Comstock (1978) also indicated that there was a higher TB concordance rate among monozygous co-twins when the index twin was less than 15 years of age at diagnosis, and also among those whose index case was reinfected with tuberculosis.

In the study conducted by Kallmann and Reisner (1943), the similarities and dissimilarities in the extent (e.g., mild forms) and outcome (e.g., subsequent arrest) of TB were compared in a group of MZ twins and a group of DZ twins. The ratio of similar behavior to dissimilar behavior in the MZ group was 8 to 1, while this ratio in the DZ group was only 2 to 1. In this nation-wide study, the twin index cases were collected from new TB admissions to hospitals and those visiting clinics. Then they were randomly selected for further analysis. The sampling strategy seems adequate. However, the diagnosis of zygosity was made based on personal examination and extended observation, which is prone to misclassification.

Most twin studies of tuberculosis have shown an approximately 35% concordance rate in MZ twins, and the remaining effect (65%) is attributed to environmental factors (Jepson, 1998). However, this result is obtained from published literature, and may not reflect the findings of studies that have not been published.

3) Adoption studies

Adoption studies are also applied to evaluate the influence of genetic factors and environmental factors on traits. The basic premise is that if a disease is influenced by genetic factors, the risk of that disease should be higher in biological relatives (parents, siblings, offspring) than in adopted relatives who live in the same environment. This study has its own complexities such as unavailability of adoption records in certain countries. Adopted children from other countries have been studied for prevalence of infection diseases, including tuberculosis in the United States. Although it has been demonstrated that there is an increased rate of latent TB infection among internationally adopted children, no further evaluations have been made of classic adoptions and tuberculosis (Saiman et al., 2001b).

4) Genetic Linkage studies, segregation analysis and association studies

There are several approaches to the identification of susceptibility genes that contribute to the acquisition of infectious diseases (Bellamy, 1998; Bellamy & Hill, 1998). Segregation analysis assesses whether there is evidence for genetic control of disease susceptibility, and family-based genetic linkage studies and association studies are two common strategies used to identify specific chromosomal regions or genes involved in a complex disease.

Segregation analysis is used to estimate the mode of inheritance of a given phenotype (Abel & Dessein, 1998). Familial aggregation of a phenotype (e.g., the same infection in multiple relatives) can be the result of genetic-environmental relationships but also cultural habits (Abel & Dessein, 1998). Complex segregation analysis primarily tests for the existence of a major gene effect on phenotype expression. Although the major gene may not be the only gene involved in the expression of the phenotype, its effect is assumed to be large. When there is evidence for a major gene, complex segregation analysis estimates the penetrances (probability of disease given the genotype) for the phenotype/genotype model. Several complex segregation analyses have been performed in infectious diseases such as tuberculosis, leprosy and malaria. A study of multicase TB pedigrees from Northern Brazil revealed that neither of the candidate genes *NRAMP1* and *TNFA* is a major gene in susceptibility to tuberculosis and that the disease is under oligogenic control (Shaw et al., 1997). In leprosy, it has been suggested that a recessive major gene may play a role (Abel et al., 1995). Also in human malaria, the intensity of infection was found to be controlled by a recessive major gene (Abel et al., 1992).

Linkage and association studies are complementary approaches for gene mapping. Familial linkage studies use data on anonymous highly polymorphic markers together with phenotype and pedigree structure information, to trace inheritance patterns through families. Co-inheritance of genetic markers and disease susceptibility suggests that a susceptibility gene is close, on the chromosome, to the genetic markers. Linkage studies can locate regions containing genes that exert major effects on the risk of developing the disease under study, and they necessarily require data either on multiple affected family members or on unaffected sibs (Spielman et al., 1993). Association studies also use data on anonymous markers, but the analysis uses unrelated individuals or small nuclear

families to identify particular alleles that are more (or less) frequent among the affected individuals than among the controls.

For association studies, two strategies can be used to select markers: 1) a candidate gene method that involves typing a few markers in small chromosomal regions containing genes assumed to be causally related to the disease under study; or 2) a genome-wide search, which involves a random search along the whole genome for chromosomal regions that could be involved in susceptibility to a given disease (Abel & Dessein, 1998). Candidate genes are genes that are likely to play a role in disease susceptibility, based on information from previous epidemiological, laboratory or gene function knowledge. Candidate genes can also be derived from experiments in animal models (commonly inbred strains of laboratory mice) of infectious diseases (Marquet & Schurr, 2001).

To date, most association studies have tested candidate genes, since genome-wide association studies require thousands of markers. In comparison, linkage studies can identify chromosomal regions in the absence of any information about candidate genes, using a few hundred markers. However, even after correcting for multiple testing, association studies can have greater power to detect susceptibility genes than linkage studies in genome-wide analysis, especially for genes with a small effect (Risch & Merikangas, 1996).

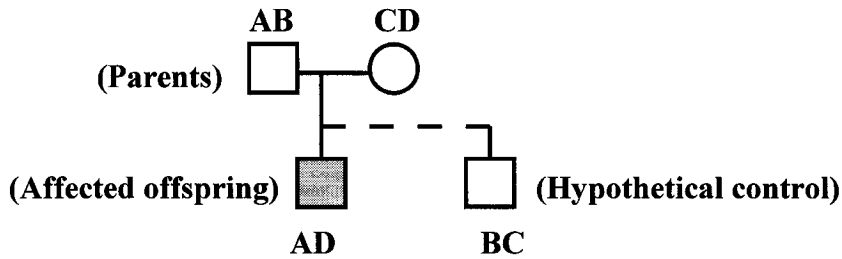
Among the classic epidemiologic methods, the case-control study has been the most widely applied classic design used to investigate potential associations between candidate genes and disease events (Khoury, 1998). The case-control design is particularly useful in genetic epidemiology for several reasons: i) case-control studies are less costly and easier to conduct than prospective cohort studies; ii) genes do not change with time, and are not affected by disease status; and iii) several candidate genes can be studied simultaneously (Khoury & Yang, 1998).

Classic case-control studies (population-based) compare allele frequencies in a set of unrelated individuals with the phenotype (e.g., disease) of interest with a set of healthy matched controls (Bellamy, 1998; Pericak-Vance, 1998). The control group should be matched with respect to ethnicity and other factors such as age. A particular marker allele is associated with the disease if the allele frequencies are different among affected

individuals compared to unaffected controls. This marker may be a genetic mutation or anonymous piece of DNA very close to a functional mutation (Flanders & Khoury, 1996).

Alternative study designs have been proposed more recently that examine allele frequencies within families. These methods use controls that have been selected from their families, siblings or "pseudosiblings" based on parental alleles. In the "case-parental control study," first proposed by Rubinstein and colleagues in 1981, parents of cases are used as a sort of control group (Flanders & Khoury, 1996). In this design, the frequency distributions of alleles transmitted from heterozygous parents to affected offspring are compared with the distribution of alleles not transmitted (controls). No "control" individual actually exists; however, the non-transmitted alleles can be considered as the genotype of a hypothetical sibling (or pseudosibling) of the case (Figure 2.2).

Figure 2.2: Hypothetical control in case-parental control design



Each method (population-based case-control and case-parental control) has its own advantages and limitations. The most important design issue in population-based case-control studies is the choice of an appropriate control group. These studies can produce false positive associations as a result of confounding due to population stratification (the mixture of individuals from heterogeneous genetic backgrounds) (Khoury & Flanders, 1996; Bellamy, 1998; Khoury, 1998). This can occur when cases and controls are poorly matched on ethnicity. Poor matching can occur when ethnic groups are defined in very crude terms, such as Caucasian, black or oriental, and hidden population stratification (maybe many "sub-ethnicities") cannot be completely controlled by ensuring that cases and controls belong to the same major ethnic group. In different ethnic populations, the evolutionary history of haplotypes (the arrangement of many alleles along a

chromosome) and linkage disequilibrium patterns can be significantly different (Cardon & Bell, 2001).

Linkage disequilibrium (LD) is defined as an association between two adjacent markers. In other words, two markers are said to be in LD when there is an excess or deficiency of certain haplotypes. When LD is present between markers, it reflects the recombination history in the population of that haplotype (Pericak-Vance, 1998; Cardon & Bell, 2001). For instance, individuals who inherited a variant or mutation on a chromosome from a common ancestor (some time in the past) should have also inherited DNA very close to the mutation from the same ancestor. The length of DNA that is identical by descent with the ancestor decays over generations, due to the process of crossing-over in meiosis that results in a rearrangement of markers. Hence, close to the mutation, descendants will often have the same alleles at markers.

Linkage disequilibrium between two genes or markers (say marker 1 with alleles A and a and marker 2 with alleles B and b) is often measured by $\Delta = q_{ab} - q_a\, q_b$. The population frequencies of alleles a and b are q_a and q_b, and q_{ab} is the population frequency of alleles a and b occurring together (on the same chromosome). Hence, Δ measures departures from independence, and is zero when there is no disequilibrium. The maximum possible value for Δ for a specific pair of markers depends on the allele frequencies q_a and q_b (Ewens & Spielman, 1997).

As previously stated, linkage disequilibrium around a mutation is at its maximum when the mutation occurs, and then decays with subsequent generations. However, this decay can be rather slow, depending on the recombination distance and the number of generations that have passed since the initial mutation (Pericak-Vance, 1998): markers closer to the mutation are exchanged more slowly than those further away from the mutation; mutations that occurred recently are more likely to show extensive LD that might extend over long distances among descendants. Moreover, patterns of LD can vary significantly within and between different populations due to different mutations, founding haplotypes, and recombination history. Admixture between populations can further complicate the picture of LD patterns. Therefore, in association studies, when cases and controls are from different ethnic groups, the frequencies of marker alleles may be different. When the risk of a disease is different in two ethnic groups, the genetic

factors that are also different in those groups will appear to be related to that disease (Wacholder et al., 2000). In a cross-sectional study in Native Americans of Pima and Papago tribes, a very strong negative association was found between the genetic marker Gm3;5;13;14 and type II or non-insulin-dependent diabetes mellitus (Knowler et al., 1988). Although individuals without this marker had a higher prevalence of diabetes than those with the marker, this marker was an index of white admixture. Since diabetes was less prevalent among Caucasians than among the Native Americans studied, having Caucasian ancestry was a protective factor for diabetes and the marker itself had no effect on diabetes mellitus. When the analysis controlled for admixture (i.e., stratification by degree of admixture), the association disappeared. In this example, both the genotype (Gm3;5,13,14 haplotype) and the phenotype (diabetes mellitus type II) were distributed differently in Native Americans than in those of mixed ethnicity.

In a cancer study in the United States, Wacholder and colleagues (2000) quantified the bias from population stratification in a case-control study. They showed that ignoring ethnicity in a study of common genetic variants and cancer in multiethnic non-Hispanic European-Americans leads to a very small bias (≤1%). They suggested that a case-control study of common genetic polymorphisms and cancer that is properly designed, conducted, and analyzed is unlikely to be subject to substantial bias. Nevertheless, their study examined only one situation, and they have suggested that further studies are needed to estimate the effect of population stratification within other populations.

Population-based case-control studies have also some advantages (Khoury & Yang, 1998). They are easier to conduct, especially for adult-onset diseases where parents may not be available; they have similar if not higher statistical power than family-based case-control studies; and they can also assess gene-gene and gene-environment interactions in measurement of the relative risk of disease.

A limitation to the family-based design is confounding due to "comorbidity" (Robins et al., 2001). This occurs when the gene under study is in linkage disequilibrium with a comorbid phenotype (e.g., AIDS), but is not linked to the main phenotype (e.g., tuberculosis). In complex diseases, comorbidity or association of two disorders occurs commonly. For instance, in 1997, the worldwide incidence rate of TB among individuals with HIV infection was estimated as 183 per 100,000 but it was 136 per 100,000 in those

without HIV infection, because infection with HIV increases the number of TB cases. Therefore, in the presence of comorbidity between the target disease and a second disease that may be in linkage disequilibrium with the gene, the case-parental control test may be an invalid test of linkage between the gene and the disease of interest. Robins and associates (2001) suggested that a population-based study, where a random sample of individuals with the target disease are recruited, would reduce the bias due to comorbidity.

Nevertheless, the case-parental control design has the main advantage that the control genotype comes from the same ethnic population as the case genotype. This matched analysis eliminates the possibility that case-control differences in allele frequencies are due to unrecognised genetic backgrounds (Flanders & Khoury, 1996). It has also been shown that using parents makes the study more powerful for detecting gene-environment interactions (Witte et al., 1999). Furthermore, selection of parents can make the design more cost-efficient, since they are from the same family (and ethnic group) as cases, which consequently eliminates the problem of selecting appropriate controls from the general population. Finally, parents are more motivated to participate in the study than unrelated individuals from general population. However, in diseases of adult onset such as diabetes mellitus, or tuberculosis in the elderly, parents are often unavailable, and consequently many potential cases will be excluded from study (Witte et al., 1999).

There is an alternative method to test for marker-disease association. The "sib TDT" or "S-TDT," compares the frequency of alleles in affected and unaffected sibs. This method is useful for diseases of late onset or when one or both parents are unavailable. In this method, at least one affected and one unaffected sib is required, hence in families where all children are affected S-TDT cannot be used. However, there are procedures that combine the data from both types of families (offspring-parents and siblings) into one overall test (Spielman & Ewens, 1998). Disadvantages of this method will be briefly discussed in the methods (4.2.4).

4.1) Candidate gene studies

Several candidate genes have been identified to influence susceptibility/resistance to tuberculosis. The following candidate genes have been genotyped in the data used in this thesis; some of the evidence for these candidate genes will be discussed.

4.1.1) Natural resistance-associated macrophage protein 1 (*Nramp1*, *NRAMP1*)

One well-known susceptibility gene for tuberculosis was initially identified in the mouse: the natural resistance-associated macrophage protein 1 (*Nramp1*) gene. This gene, located on mouse chromosome 1, controls innate (non-immune) resistance to infection with *M. bovis* (BCG) (Gros et al., 1981). Innate defense to infection with several intracellular parasites such as *Mycobacterium*, *Leishmania*, and *Salmonella* has also been shown to be under the control of a single G169D amino acid substitution in the *Nramp1* protein in inbred mouse strains (Vidal et al., 1993; 1995). This observation led to the hypothesis that polymorphisms in the human homologue may similarly be associated with *M. tuberculosis* susceptibility.

The human *NRAMP1*, located on chromosome region 2q35, has a genomic length of 14kb (Cellier et al., 1994; Marquet et al., 2000). The mouse mutation has not been seen in humans. It has been suggested that in human *NRAMP1*, substitutions of glycin are not tolerated, and the hypothesis is that other mutation(s) are causing susceptibility to TB (Cellier et al., 1994). Although both *NRAMP1* and *Nramp1* are expressed in macrophages, there are differences in tissue and functional expression of these genes between two species (Schneemann et al., 1993; Vidal et al., 1993; Cellier et al., 1994, 1997). The highest expression of *NRAMP1* is observed in peripheral blood leukocytes, followed by lung and spleen, and it may affect the susceptibility/resistance to the organism by altering the macrophage activation. In contrast, *Nramp1* in mice regulates bacterial growth in the spleen and liver.

In the human *NRAMP1* gene, a total of 11 polymorphisms have been identified (Buu et al., 2000). In a case-control study in West Africa, it was found that tuberculosis risk was associated with specific *NRAMP1* alleles (Bellamy et al., 1998a). Association with *NRAMP1* alleles and tuberculosis was also obtained in a newly studied population in West Africa, using case-parental control design (Cervino et al., 2000). In a linkage study

in Vietnamese families (Abel et al., 1995) it was demonstrated that the *NRAMP1* gene was linked to leprosy susceptibility. Greenwood and colleagues (2000) studied the *NRAMP1* gene-environment interaction in a tuberculosis outbreak in an extended Canadian Aboriginal family. They demonstrated that *NRAMP1* was linked to susceptibility to tuberculosis, but linkage was seen only when age together with vaccination status was taken into account. This suggests that genetic effects may remain undetected if the gene–and clinico–epidemiological information are not appropriately accounted for in the analysis (Gros & Schurr, 2002).

4.1.2) Vitamin D Receptor (*VDR*)

Susceptibility to tuberculosis may be increased by deficiencies in vitamin D (25-hydroxycholecalciferol) (Davies et al., 1985). In two prospective studies of patients with tuberculosis (Davies et al., 1985, 1987), the serum level of 25-hydroxycholecalciferol in patients with TB was between 35% and 66% of healthy matched control values, but rapidly rose to equal control values as soon as treatment ended. Davies et al (1985, 1987) also postulated that low serum vitamin D may be a consequence of TB and anti-TB therapy.

An early summer peak is observed in TB rates in the UK, which suggests that low concentrations of vitamin D following winter may contribute to reactivation of latent infection (Douglas et al., 1996). Vegetarian diet and the associated vitamin deficiency may be a risk factor for tuberculosis in some Asian immigrants in London (Finch et al., 1991).

The active metabolite of vitamin D, 1,25-dihydroxyvitamin D (1,25-D3), interacts with the vitamin D receptor, which is present on human monocytes, and helps the monocytes to control proliferation of *M. tuberculosis* (Rook et al., 1986; Rockett et al., 1998). This effect may be influenced by polymorphisms at three sites in the *VDR* gene, which is located on chromosome region 12q12-q14 (Labuda et al., 1991). At the *VDR* locus, there are numerous *VDR* alleles, including (*Aa, Bb, Ff, Tt*). Homozygotes (genotype *tt*) for a variant of *VDR* were significantly underrepresented among TB patients, which suggests that this genetic variant may confer a higher level of resistance to tuberculosis (Bellamy et al., 1999).

In a more recent study, Wilkinson and coworkers (2000) investigated the interaction between serum 25-hydroxycholecalciferol concentrations and VDR genotype on susceptibility to TB in a group of Asians living in the UK with a high rate of tuberculosis. In this hospital-based case-control study, they found an association between 25-hydroxycholecalciferol deficiency (≤10 nmol/L) and active disease, and even a higher risk for TB among those with undetectable serum 25-hydroxycholecalciferol (<7 nmol/L). The combination of genotype *TT/Tt* and 25-hydroxycholecalciferol deficiency was associated with disease, and the presence of genotype *ff* was strongly associated with disease.

4.1.3) Mannose-binding lectin or protein (*MBL* or *MBP*)

Mannose binding protein, located on chromosome region 10q11.2-q21, is one of the most important components of the innate immune system (Turner, 1996; Turner et al., 2000). MBL is considered as an "ante-antibody," and it has been demonstrated (in vitro) that MBL binds *M. tuberculosis* and acts as an opsonin in the phase of innate immunity (Hoal-Van Helden et al., 1999). In a recent study (Selvaraj et al., 2000), *MBP* was presented as a gene which influences the cell mediated immune response in patients with PTB.

Deficiency of MBL originates from three known structural gene mutations in exon 1 of the *MBL* gene and is associated with increased risk of tuberculosis and other infections such as malaria (Summerfield et al., 1995; Selvaraj et al., 1999). These mutations have been described in codon 52 (variant D), 54 (variant B) and 57 (variant C). The highest frequency of variant B has been found in native American populations and Eurasians whereas the C allele is frequent in most sub-Saharan African populations (Lipscombe et al., 1992). In a case-control study conducted in Africa, low levels of MBL were found to be associated with increased risk of tuberculosis and HIV infection (Garred et al., 1997). In contrast, MBL deficiency was not a significant risk factor for pulmonary TB nor for malaria in a retrospective study in Gambia (Bellamy et al., 1998b). In a population-based case-control study of an epidemic of tuberculosis in South Africa, the MBP B allele (G54D) was even found to be protective against TB meningitis (Hoal-Van Helden et al., 1999).

4.1.4) Surfactant proteins (*SFTP*)

Pulmonary surfactant is a lipoprotein complex synthesized and secreted by alveolar cells that reduces the surface tension of the alveoli, and appears to play important roles in the innate host defense (Hoover & Floros, 1998; Ferguson & Schlesinger, 2000). Pulmonary surfactant consists of four proteins, including surfactant protein (SP) A, B, C and D. SP-B (Floros et al., 2000) and SP-C (Conkright et al., 2002) are essential for normal lung function. SP-B probably plays an indirect role in TB infection by further compromising lung function in the presence of *M. tuberculosis*. SP-A and SP-D gene loci have been physically mapped on chromosome region 10q22-q23 (Hoover & Floros, 1998). Polymorphic marker loci have been characterized for *SP-A* (consists of *SP-A1* and *SP-A2)* and *SP-D* genes (Floros et al., 2000). SP-A and SP-D bind *M. tuberculosis* and modulate phagocytosis by AMs. It has been demonstrated (Ferguson et al., 2002) that SP-A increases the phagocytosis of *M. tuberculosis* due to direct interaction with AMs, and is likely to be harmful to the host. In contrast, SP-D reduces the phagocytosis of *M. tuberculosis* by AMs due to an interaction between the organism and the protein, and reduces the intracellular growth of *M. tuberculosis*.

A population-based case-control study conducted in Japan (Kondo et al. 1998), revealed that the SP-D serum level is a useful indicator of disease activity in pulmonary tuberculosis. It was found that higher levels of SP-D protein (>134.6 ng/ml) are significantly associated with the number of tubercle bacilli in the sputum. The results also indicated that higher levels of SP-D were observed in patients with cavity formation than in those without cavity formation. In another population-based case-control study in Mexican families, Floros and associates (2000) found that some of the SP marker alleles are associated with an increased risk for tuberculosis (i.e., *SP-A1* ($6A^4$), *SP-A2* ($1A^3$), and SP-B (*B1013_A*)).

2.3 SUMMARY

This chapter has provided the biological rationale for each of the candidate genes studied in the thesis, in relation to the risk for tuberculosis, in different populations. For each gene, several studies using different designs, such as linkage studies, animal models and case-control studies, have been used to investigate the relationship between the

candidate genes and tuberculosis. However, except for *NRAMP1*, no family-based studies have been done to examine the effect of these genes on the risk of developing tuberculosis. Moreover, to validate candidate genes reported by others, and to examine associations in different populations to see if associations are consistent across populations, further studies are required. This study aimed to examine how these candidate genes may influence the risk of the disease by conducting a case-parental control study in Ethiopian families.

CHAPTER 3: OBJECTIVES

The main objectives of this study were to 1) examine whether alleles at several candidate genes, including natural resistance associated macrophage protein 1 (*NRAMP1*), vitamin D receptor (*VDR*), surfactant proteins (*SFTPA1*), and mannose-binding lectin (*MBL*) alter the risk of developing tuberculosis. (The transmission disequilibrium test (TDT) was used to test for association between these markers and the phenotype of tuberculosis); and 2) assess how other factors such as age–of–onset, gender, ethnicity, and certain symptoms of the disease and other genes may influence the association between these marker alleles and TB.

The secondary objectives were to i) test for dominant or recessive inheritance; and ii) to estimate the relative risks of disease for individuals carrying either zero or one high risk allele versus homozygotes with two high risk alleles.

CHAPTER 4: MATERIALS AND METHODS

This research involves the analysis of a portion of an existing database containing family data from Ethiopia that has been collected by Dr. Erwin Schurr (McGill University) and his colleagues in order to investigate genetic susceptibility to tuberculosis. In this chapter, some aspects of the design of the original study of Dr. Schurr will be presented, separately from the description of families selected for analysis in this thesis.

4.1 DESIGN OF THE ORIGINAL PROJECT (AHRI STUDY)

4.1.1 Study population

The study population originates from Hosanna, a rural area in Ethiopia. Hosanna is located 250 km Southeast of Addis Ababa, and has a population of approximately 40,000. Recruitment was started in August 1997 and is ongoing. Individuals who were diagnosed with tuberculosis at the Hosanna hospital or affiliated health centers were asked if their parents or family members would be willing to participate in the study. Hence, these cases (probands) are self-selected. In this study, three kinds of families were enrolled: i) families for which one or both parents of an affected child were available and where either none or one parent was affected by TB; ii) families where one parent, and at least one affected child, and one or more healthy sib(s) were available; and iii) families where both parents were missing, but two or more healthy sibs were available, in addition to at least one affected child. For the study of multiplex families or families with more than one affected offspring it was decided to also include historical cases (siblings of affected child who had had TB in the past). The recruitment was based on inclusion and exclusion criteria that are described below.

4.1.2 Inclusion and exclusion criteria

The inclusion criterion for a case was the detection of acid-fast bacilli in sputum and/ or successful cultivation of *M. tuberculosis* isolates from sputum samples.

Subjects were excluded from the study if:

1) The families did not consent to participate.
2) The families did not reside close enough to the Hosanna hospital for visits.
3) Someone in the family was (clinically) suspected to be positive for AIDS.

4) Very young children (<7 years old) for ethical reasons.

4.1.3 Subject enrollment and sample acquisition

Subjects were identified from records of the TB unit at Hosanna hospital. TB cases and their families were contacted by local health care workers; the nature of the study was explained to patients and their families; and they were invited to enroll in the study. For those families who agreed to participate, clinical data of patients were forwarded to the study physician to review the records. During the interview the nature of the study was explained by the study physician to all family members again and written (signature or finger print) consent to participate in the study was obtained. The physician was blinded to the genetic data because clinical diagnosis was made prior to genotyping analysis.

Families who gave their consent were examined by the study physician who also filled out a one-page questionnaire for each person. To confirm case status, blood and sputum samples were obtained from cases and their families by a team of 3 people consisting of the study physician, a nurse, and a skilled phlebotomist/technician. All specimens were transported to Armauer Hansen Research Institute (AHRI) located in Addis Ababa on the same day and further processing was done within 12 hours.

At AHRI, blood was separated into mononuclear cells, polymorphonuclear cells, and plasma. DNA, extracted from polymorphonuclear cells, was aliquoted and half of the total amount was shipped to Montreal. Genetic marker typing was performed in the laboratory of Dr. Schurr at The Montreal General Hospital.

4.1.4 Rationale for choice of Hossana, Ethiopia

1) Hosanna has a large number of TB patients: Annually, approximately 3000 new cases of TB are registered by the Hosanna hospital and its primary health care clinics (the catchment population is approximately 3 million).
2) There is a low rate of HIV/AIDS: The prevalence of HIV-positive persons has been estimated at less than 0.1% (Schurr, personal communication).
3) The staff is well trained.
4) Record keeping is reliable.

5) Most individuals come from one ethnic group: approximately 80% of the study population belongs to the Hadiya ethnic group. Minority ethnic groups in the area include Gurage, Kembata, and Amhara. In addition, inter-group marriage is not common in this region.

4.1.5 Data acquisition form

A one–page questionnaire was filled out for each individual at Hosanna hospital. This form contains all personal information including demographics (age, gender, ethnic group), family ID, individual ID, hospital number, and case status. In this questionnaire, clinical data were also recorded, which included general health condition, weight, recent weight loss, symptoms of TB, BCG scars, history of previous treatment for TB, clinical data, and result of sputum examination. A copy of this form is included in appendix I.

4.1.6 Ethical considerations

Documented informed consent to participate in the study was obtained from each subject (patient and parents) after the nature of study was explained by the physician.

Confidentiality was ensured through the denominalization of the data, a method that presents information in such a way that the identity of the subject cannot be ascertained from the collected information.

Ethical approval for the primary study was obtained from The Federal Democratic Republic of Ethiopia (Ethiopian Science and Technology Commission), and The Research Ethics Committee at the Montreal General Hospital, Montreal, Quebec, Canada. Copies of the study approval are included in appendix I.

4.2 DESIGN OF THIS RESEARCH

4.2.1 Study design

In this research, the prospective case-parental control design was used to address the questions under study. In genetic epidemiology, when data from affected offspring and their parents are available, the case-parental control study is a powerful design for the study of candidate genes that is not susceptible to bias due to ethnic diversity (Sham, 1996).

4.2.2 Study population

The study subjects for this thesis were selected from the pool of individuals enrolled in the original project. Since it is an ongoing project, it was decided to select families recruited from August 1997 to October 2001. For this research, only families with exactly one affected offspring and both parents were included. A total of 95 nuclear families (n = 285) with one affected offspring (n = 95) and two parents (n = 190) were identified from the original database (n = 620).

4.2.3 Inclusion and exclusion criteria

From all families recruited in the primary project only those with only one affected child and two parents were included.

All other families such as those with multiple affected offspring, and families with one or no parent (with or without sibships) were excluded for the purpose of this thesis.

4.2.4 Rationale for the choice of families with only one affected child and two parents

Multiplex families (families with more than one affected offspring) were excluded from this research because:

1) Multiplex sibships make the contingency statistic (TDT) invalid as a test of association, unless only one affected child from each family is used in the analysis (Ewens & Spielman, 1995).
2) Multiplex families may be different from families with only one affected child (simplex families), due to increased genetic or environmental susceptibility. Restricting to simplex families may make the sample more homogenous (personal communication, Erwin Schurr).
3) Historical cases, those who had been diagnosed in the past by other physicians and possibly by different diagnostic strategies (e.g., only based on clinical findings), were only found in multiplex families, therefore, by excluding the multiplex families, historical cases were automatically eliminated from the study.

Families where one or both parents were missing, were also excluded because:

1) Bias can be introduced when one or both parents are missing and sibs genotypes are used to deduce the missing parental alleles. Although simple tests of association can be performed when one or both parents are missing, a bias can be introduced if the analysis is not carefully controlled.
2) Different analytical methods should be applied that are beyond the scope of this thesis.

4.2.5 Ethical considerations

Ethics approval for this study was obtained from The Research Ethics Committee at The Montreal General Hospital, Montreal, Quebec, Canada. A copy of this study approval is also included in appendix I.

4.2.6 Measurement

In traditional case-control studies we can define "exposure" and "outcome" variables: in such designs the "outcome" under study is usually disease status and the "exposures" are potential etiologic modifiable or non-modifiable characteristics, risk factors or determinants of the disease in the study population.

The case-parental control design, selected for this research, cannot really be fitted into an "outcome" and "exposure" paradigm. In fact, the "outcome" of interest is transmission /nontransmission of particular marker alleles at candidate genes from parents to affected offspring, and the "exposures" are candidate gene variants or markers. In other words, exposure and outcome are not distinct from one another. We study, for instance, if allele "1" was transmitted more frequently than expected to the affected offspring. "Exposure" can be considered to be the candidate gene under study, and "outcome" can be considered to be which allele was transmitted. Hence, there are new "outcome" variables for every candidate gene examined.

The candidate genes examined were: 1) natural resistance-associated macrophage protein 1 (*NRAMP1*); 2) mannose-binding protein (*MBP*); 3) vitamin D receptor (*VDR*); and 4) surfactant proteins A1 (*SFTPA1*). Table 4.1 describes the locus → polymorphism or marker → allele staging. The number of missing values varies for different alleles at different markers. These missing values can either be due to absence of samples,

technical problems in genotyping or because the samples had not yet been genotyped. All markers except *NRAMP1-5'(GT)n* were biallelic. Six individuals were identified to have allele "3" at marker allele *NRAMP1-5'(GT)n*: 2 cases and two parents with genotype "1,3", and 2 parents with genotype "2,3". These participants' genotypes were not included in the analysis for this marker locus as they only made up a very small proportion of the total genotypes, and analysis of biallelic markers is more straight-forward.

Table 4.1: List of marker alleles (adapted from AHRI website)

Locus	Marker	Allele	Description	Locus	Marker	Allele	Description
NRAMP1	NRAMP1-5'(GT)n	1	133bp	SFTPA1	SFTPA1-170	1	C
NRAMP1	NRAMP1-5'(GT)n	2	135bp	SFTPA1	SFTPA1-170	2	T
NRAMP1	NRAMP1-5'(GT)n	3	137bp	SFTPA1	SFTPA1-256	1	C
NRAMP1	NRAMP1-274C/T	1	T	SFTPA1	SFTPA1-256	2	G
NRAMP1	NRAMP1-274C/T	2	C	SFTPA1	SFTPA1-294	1	A
NRAMP1	NRAMP1-469+14G/C	1	G	SFTPA1	SFTPA1-294	2	G
NRAMP1	NRAMP1-469+14G/C	2	C	SFTPA1	SFTPA1-507	1	A
NRAMP1	NRAMP1-3'UTR	1	+CAAA	SFTPA1	SFTPA1-507	2	G
NRAMP1	NRAMP1-3'UTR	2	-CAAA	SFTPA1	SFTPA1-763	1	C
MBL	MBL-G54D	1	C	SFTPA1	SFTPA1-763	2	T
MBL	MBL-G54D	2	T	VDR	VDR-117	1	T
MBL	MBL-G57Q	1	G	VDR	VDR-117	2	C
MBL	MBL-G57Q	2	A	VDR	VDR-1056	1	C
MBL	MBL-Cod52	1	C	VDR	VDR-1056	2	T
MBL	MBL-Cod52	2	T				

Locus: The candidate gene
Marker: The name of the polymorphism studied in the gene
Allele: Assigned labels for different alleles
Description: Exact description of each allele
bp: base pairs, T: thymine, C: cytosine, A: adenine, G: guanine

In this research, the effects of individual characteristics, signs and symptoms of tuberculosis, and laboratory findings on the transmission pattern were investigated. There were no missing values in these variables (symptoms apply only to affected offspring). The following variables were examined:

1) Age at diagnosis was measured in years as a continuous variable. However, there is a possibility for inaccuracy in the estimation of age, because the age was based on the participants' responses, and many individuals were unsure of their age. There were no documents such as a birth certificate to verify the exact age. When possible, the study physician attempted to improve estimates of age by asking about memories of major events. In general, younger individuals (under 30 years old) were more likely to know their age. Age was analyzed both as a categorical and a continuous variable. Age was grouped into 5 categories: under 15; 15-19; 20-24; 25-29; 30 and over.
2) Gender was a dichotomous variable, female or male.
3) Ethnicity was divided into 2 major groups: Hadiya and others.
4) General condition (subjective) was categorized by the study physician into 4 groups: good, bad, cachectic and bed-ridden.
5) Weight at interview was measured in kilograms (kg).
6) Weight loss more than 10 kg (in the past 6 months) was estimated by the patients (or their parents for younger children). This variable was dichotomized to yes or no.
7) BCG scar was evaluated by the study physician and was dichotomized to yes if present, and no, if absent.
8) Cough, hemoptysis, cough and sputum, and night sweats were dichotomized to yes or no. Duration of these variables was measured in weeks (continuous).
9) Previous treatment for TB was measured as a dichotomous variable: yes or no.
10) Acid-fast bacilli were graded from 0 to 6+. In this study, this variable was dichotomized to positive (1+ to 6+) or negative (0). An intensity measurement could not be used because samples were graded at either Hosanna or AHRI and unfortunately the two participating laboratories had used different scales.
11) Culture was dichotomized to positive or negative.

12) Extrapulmonary TB was dichotomized to yes or no.

4.3 STATISTICAL ANALYSIS

There were three main types of analysis performed on these data. Firstly, a matched pair analysis or McNemar test, using the TDT computer package (Spielman & Ewens, 1999), was performed to test for linkage disequilibrium between marker alleles and tuberculosis (Spielman et al., 1993; Ewens & Spielman, 1995). Then, a genetics computer package named "gassoc" (Schaid & Rowland, 2001) was used to i) estimate relative risks for genotypes (1,2 or 2,2) versus genotype (1,1); and ii) test for dominant and recessive inheritance (Schaid & Sommer, 1993, 1994). Finally, a logistic regression adaptation was used to test for gene-environment interactions and gene-gene interactions (Sham & Curtis, 1995; Eaves & Sullivan, 2001), using SAS (SAS Language, 1990). Each approach is described in detail below.

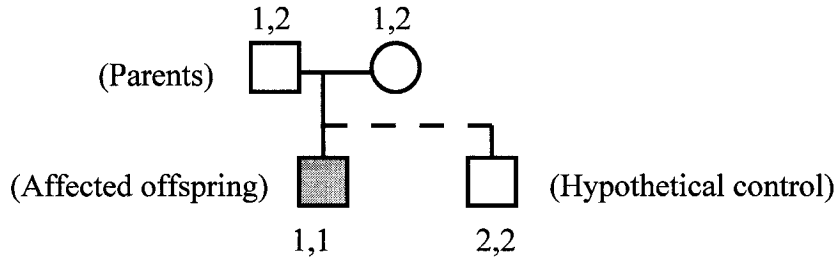
4.3.1 Transmission disequilibrium test

The transmission disequilibrium test (TDT) was first introduced by Spielman et al. (1993) as a test for linkage and association between a biallelic genetic marker and a complex disease, in family-based case-control studies. The TDT is also frequently employed to confirm disease-marker associations detected in population-based case-control studies. Given a reasonable numbers of families (see section 5.7), this test has enough statistical power to identify the associations between markers and disease. Although the TDT has most power when the true mode of inheritance is additive (see section 6.1), it still is a valid test statistic for other modes of inheritance (i.e., dominant or recessive) (Ewens & Spielman, 1995).

The TDT uses data from families in which marker genotypes are known for the parents and the affected offspring, but only families where at least one parent is heterozygous for the marker alleles under study are considered. In a TDT analysis, one compares the number of times that heterozygous parents transmit a putative, high-risk marker allele to affected offspring, with the number of times that they transmit the alternative marker allele. If this is different from 50% (what is expected under Mendel's laws of inheritance), then the high-risk allele may be a disease susceptibility allele itself or may be in linkage disequilibrium with one (Spielman et al., 1993). The following

figure (4.1) shows an example of a family with both heterozygous parents (1,2). In this example, allele "1" from both parents is transmitted to the affected child, therefore, this child carries genotype "1,1", and the nontransmitted alleles are allele "2" from both parents or "2,2". In this family, the transmitted and nontransmitted alleles from the heterozygous father are counted independently from the heterozygous mother.

Figure 4.1: Transmitted and nontransmitted alleles from heterozygous parents to an affected child



All possible genotypes for an index case, its hypothetical control, and the parents, are shown in table 4.2 for a biallelic marker. The first column and the first row (bold) illustrate the transmitted, and nontransmitted alleles, respectively. The body of the table shows the parents' genotypes (father's genotype/mother's genotype). The TDT only analyses families with heterozygous parents (underlined).

Table 4.2: Transmission of alleles from heterozygous parents to affected offspring (Adapted from Sun et al., 1998)

Transmitted alleles (case)	Non-transmitted alleles (hypothetical control)			
	1,1	**1,2**	**2,1**	**2,2**
1,1	1,1/1,1	1,1/<u>1,2</u>	<u>1,2</u>/1,1	<u>1,2</u>/<u>1,2</u>
1,2	1,1/<u>2,1</u>	1,1/2,2	<u>1,2</u>/<u>2,1</u>	<u>1,2</u>/2,2
2,1	<u>2,1</u>/1,1	<u>2,1</u>/<u>1,2</u>	2,2/1,1	2,2/<u>1,2</u>
2,2	<u>2,1</u>/<u>2,1</u>	<u>2,1</u>/2,2	2,2/<u>2,1</u>	2,2/2,2

In any pair of alleles, the first comes from the father and the second from the mother

Suppose we have a sample of n families with a single affected child. The 2n parents can be described in terms of transmitted and nontransmitted alleles of a biallelic marker

(i.e., alleles 1 and 2) to the affected child in a 2×2 table, indicated by *a*, *b*, *c*, and *d* (Table 4.3). The important elements of the table are *b* and *c* or discordant alleles, which reflect the data from heterozygous "1,2" parents. The appropriate matched-pair analysis is the TDT or McNemar test: $(b\text{-}c)^2 / (b+c)$, which is approximately distributed as a χ^2 statistic with one degree of freedom (Spielman et al, 1993). In this analysis, *b* is denoted as the number of affected children who have received allele "1" and not received allele "2" from their heterozygous parent and *c* those who have received allele "2" and have not received allele "1". When cell count *b* is larger than cell count *c*, allele "1" is more likely to have been transferred to the cases than allele "2", which implies an association between allele "1" and the disease. The information from homozygous parents (1,1 or 2,2) is excluded from the analysis, because they are not informative. The labels "1" and "2" are arbitrary names for different allelic variants at each marker locus. If both parents of an index case are heterozygous, they contribute independently to the table. Statistically, their transmissions are independent under the null hypothesis.

Table 4.3: Transmitted and nontransmitted marker alleles "1" and "2" among 2n parents of n affected children (from Spielman et al., 1993)

	Nontransmitted allele (control)		
Transmitted allele (case)	1	2	Total
1	*a*	*b*	*a* + *b*
2	*c*	*d*	*c* + *d*
Total	*a* + *c*	*b* + *d*	*2n*

$$TDT = (b - c)^2 / (b + c) \approx \chi^2 \ (m\text{-}1 \text{ or } 1 \text{ df})$$

m: number of marker alleles

Let θ be the recombination fraction between the marker and the susceptibility gene, and let δ be the measure of disequilibrium between the marker and the disease gene. The null hypothesis is $\delta (1 - 2\theta) = 0$, which would be zero if $\delta = 0$ (no linkage disequilibrium or no association) or $\theta = ½$ (no linkage). The parameter θ represents the probability of recombination occurring during meiosis. That is, for two genes A and B, θ measures the probability that gene A came from the mother and gene B from the father, or vice versa.

For a candidate gene study, in fact, we are assuming that linkage is present since the candidate gene is at the marker (i.e., $\theta = 0$). Hence, we are testing that $\delta = 0$ or that there is no association or LD. The compound null hypothesis is important, because if linkage is absent ($\theta = ½$), the test statistic χ^2 will be expected to be zero; which implies that association cannot be detected in the absence of linkage (and vice versa). As theta goes farther away from ½ towards zero, the value of χ^2 increases, and the association will be more easily detected.

4.3.2 Relative risk estimation and mode of inheritance

The previous analysis treated transmissions from parents independently. An alternative analysis method to study the association of a candidate gene with a disease using cases and parental data was presented by Schaid and Sommer (1993, 1994). This approach allows:

i) Testing dominant or recessive inheritance, which implies dependence between parents. If a mode of inheritance (e.g., recessive) is assumed, then, this approach can test for association based on that assumption.

ii) Estimating the two relative risks (RR): 1) the RR of disease for heterozygotes with one susceptible allele (1,2) versus homozygotes with two susceptible alleles (1,1); and 2) the RR for homozygotes without susceptible alleles (2,2) versus homozygotes with two susceptible alleles (1,1).

In this analysis, a likelihood method is used which is conditional on parental genotype. For this description, allele "1" is considered as the high risk allele, and allele "2" as the normal allele. Schaid and Sommer (1993) defined the following notations: p: the population frequency of allele "1"; q: the population frequency of allele "2"; and D: the event that an individual has the disease. Then, the probabilities of disease, conditional on the genotypes at a given candidate gene can be defined as $f_2 = P(D|\ 1,1)$, $f_1 = P(D|\ 1,2)$, $f_0 = P(D|\ 2,2)$. Here, the subscript "j" of f_j indicates the number of copies of allele "1" in the cases. The two RR parameters discussed earlier are defined as $\psi_1 = f_1 / f_2$ and $\psi_2 = f_0 / f_2$ and measure the relative decrease in disease probabilities for heterozygous "1,2" cases and homozygous "2,2", compared to the probability for homozygous "1,1" cases,

respectively. These RRs measure the relative changes when only one candidate gene (single-locus) is considered at a time. Schaid and Sommer (1993) argued that for complex diseases, the value of the tests obtained from each locus represents an average association over all multilocus genotypes that are associated with the disease and have not been measured. Under the null hypothesis of no association between the candidate gene and the disease, all conditional disease probabilities have equal quantity ($f_1 = f_2 = f_0$), which implies that $\psi_1 = \psi_2 = 1$. [In order to use consistent allele labelling throughout the thesis, the parameter definitions given here are slightly modified from the ones in Schaid & Sommer (1993)].

To understand the test statistics, a detailed explanation follows. For a given locus with two alleles there are three possible genotypes for each parent: "1,1", "1,2", and "2,2". Table 4.4 describes the informative parental mating types, case and non-transmitted parental genes (one allele from the father and one allele from the mother), and the genotype probabilities for cases. The counts for each transmission type, conditional on mating type, are denoted x_j. Schaid and Sommer's (1993) likelihood approach led to the probabilities in the last column, so that the transmission probabilities can be expressed as a function of the relative risks.

Table 4.4: Classification of affected offspring according to their genotype and their parents' genotypes at the candidate gene locus (adapted from Schaid & Sommer, 1993)

(Informative) parental mating type	Notation	Case genotype	Non-transmitted parental genes	Probability of case genotype
a) 1,1 x 1,2	x_{a2} x_{a1}	1,1 1,2	1,2 1,1	$\psi_1 / (\psi_1 + 1)$ $1 / (\psi_1 + 1)$
b) 1,2 x 1,2	x_{b2} x_{b1} x_{b0}	1,1 1,2 2,2	2,2 1,2 1,1	$1 / (\psi_2 + 2\psi_1 + 1)$ $2\psi_1 / (\psi_2 + 2\psi_1 + 1)$ $\psi_2 / (\psi_2 + 2\psi_1 + 1)$
c) 1,2 x 2,2	x_{c1} x_{c0}	1,2 2,2	2,2 1,2	$\psi_1 / (\psi_1 + \psi_2)$ $\psi_2 / (\psi_1 + \psi_2)$

The numbers (0,1,2) in subscripts describe the number of susceptible alleles "1" within possible parental mating. Non-informative mating types (both parents homozygous) have been excluded.

Tests for association can then be based on testing whether ψ_1 and ψ_2 are 1.0. In addition, since the two relative risk (RR) parameters can be estimated, this likelihood method can also test for i) dominant effect of the mutant allele or allele "1" where $f_1 = f_2$ (H_D: $\psi_1 = 1$); and ii) recessive model (H_R: $\psi_1 = \psi_2$) where individuals with one copy of allele "1" (heterozygotes 1,2) are not at higher risk. Therefore, we can test whether a dominant or recessive model best fits our data at different loci (Schaid & Sommer, 1993, 1994). A third hypothesis tests an allele-dosage model log (R_i) = β_i, where i represents the count of allele "1" and can be 0, 1, or 2 (Schaid & Sommer, 1994). In the allele-dosage model it is assumed that the dose of the allele "1" has a linear effect on the log RR. This model suggests that $R_2 = (R_1)^2$ when $R_1 < R_2$. In fact, this is equivalent to the standard TDT of section 4.3.1 (Schaid & Sommer, 1994). Testing for dominant and recessive alternatives can be performed with the following equations, which are expected to be χ^2 statistics:

$$\text{DOM} = [(x_{b2} + x_{b1} - 3n_4 / 4) + (x_{c1} - n_5 / 2)]^2 / (3n_4 / 16 + n_5/4)$$
$$\text{REC} = [(x_{a2} - n_2 / 2) + (x_{b2} - n_4 / 4)]^2 / (n_2 / 4 + 3n_4 / 16)$$

In the dominant model, cases from mating type "a" (1,1 x 1,2) are not included in the equation since they are not informative under a dominant model. Homozygotes and heterozygotes are also grouped together ($x_{b2} + x_{b1}$), because it is a dominant model. In the recessive model, cases from mating type "c" (1,2 x 2,2) are excluded from the equation since the offspring carry at most one susceptibility allele and are, hence, not informative under a recessive model. Under the null hypothesis, only the numbers of cases that are homozygous are compared with their expected values. For the allele dosage model, TDT statistics are the most powerful χ^2 statistics. Schaid and Sommer (1993) indicated that in general, smaller sample sizes are required to detect genes with a recessive mode of inheritance.

4.3.3 Logistic regression analysis

A bivariate analysis was first conducted to evaluate the correlation between variables under study. Highly correlated covariates were assessed independently in the further analysis described below.

A logistic regression method, proposed by Sham and Curtis (1995) was used for the analysis of the families. This approach provides tests for both: i) the main effects of marker alleles on genetic risk; and ii) genotype-environment interaction (G × E). If the risk of disease associated with a particular gene is modified by environmental exposures, then models that estimate G × E may detect associations that cannot be seen in univariate analysis. In several studies age, gender, or one or more associated clinical outcomes have been considered as the "environment" that interacts with genotype (Eaves & Sullivan, 2001). Such interactions between covariates and the effect of marker alleles can be modeled in a basic logistic regression model that can be implemented in standard statistical software such as SAS (SAS Language, 1990).

Like the standard TDT analysis (section 4.3.1), the outcome variable (Y) in this analysis is transmission/nontransmission of alleles from heterozygous parents to affected offspring where Y = 1 if allele "1" (the putative, high risk allele) is transmitted (or allele "2" is not transmitted), and Y = 0 if allele "1" is not transmitted (or allele "2" is transmitted). In a logistic model without covariates and with only the intercept, testing that the intercept is zero is equivalent to the TDT test,

$$\text{Log}\,[P(y = 1) / P(y = 0)] = \alpha.$$

If $\alpha = 0$, then $P(y = 1) = P(y = 0)$, and hence the probability of transmission equals the probability of nontransmission.

In this analysis, unlike traditional case-control studies, a control (genotype) is constructed and does not really exist as a real person, and hence has no phenotype. Therefore, the covariates used in the analysis apply only to the affected offspring. The phenotypic status (covariate) of the offspring is denoted by X where X = 1 if the environmental factor (e.g., hemoptysis) is present, and X = 0 if absent. Here, the model

$$\text{Log}\,[P(y = 1) / P(y = 0)] = \alpha + \beta x$$

can be used to test for $\beta = 0$, which examines whether the covariate modifies the transmission pattern. An equivalent model,

$$\text{Log}\,[P(y = 1) / P(y = 0)] = \beta_1 I\,(x = 1) + \beta_2 I\,(x = 0)$$

specifically performs the TDT in subgroups defined by the covariate. Covariates for the affected offspring were coded in the usual way.

The logistic regressions (one for each marker) were run using SAS (SAS Language, 1990), with and without covariates (the code for this program was written by Dr. Celia Greenwood). We also looked at interactions between markers where one marker served as the outcome variable "Y", and the second marker as a covariate. In this case, the covariate "X" contained the transmission information for a particular allele. For example, suppose that "X" refers to a marker in *NRAMP1*. To examine its effect on a transmission pattern at another unlinked marker in a different gene, the covariate was coded as:

X = 1 if allele "1" was transmitted from heterozygous parent to affected child;

X = -1 if allele "2" was transmitted from a heterozygous parent; and

X = 0 if otherwise (i.e., homozygous parent).

For all analyses, including TDT univariate and multivariate analyses, a family with 2 heterozygous parents at a particular marker contributes 2 data points to the analysis. Covariates for the gene-gene interaction (G × G) analysis were coded to correspond to the same parental transmission. When a parent was homozygous at marker "Y", the transmission was excluded from analysis. However, a homozygous parent at "X" was retained in the analysis for comparability when $\beta = 0$.

After identifying demographic and environmental variables that appeared to influence transmission distortion at a single marker, gene-gene interactions were examined. Only (demographic or environmental) covariates identified in the single locus analysis were included in the G × G models. Subgroup analyses examined the transmission distortion at one gene in subgroups defined by transmission patterns at a second gene.

In all steps of logistic analysis discussed above, likelihood ratio chi-squares were used to test for significance.

CHAPTER 5: RESULTS

5.1 BASELINE CHARACTERISTICS OF STUDY POPULATION

The study population consisted of 95 independent nuclear families recruited from August 1997 to October 2001. Each family consisted of one offspring diagnosed with pulmonary tuberculosis and his or her parents. The information used here in the analysis was obtained at the recruitment interview.

Tables 5.1 and 5.2 illustrate the distribution of demographics, clinical data, and laboratory results for the (tuberculous) offspring. Approximately 54% of these children were male and 76% were from the primary ethnic group Hadiya. The mean age was 18 years (range 7–35); those between ages 15 to 19 years made up 40% of all cases and 22.1% were between ages 7 to 15. All offspring had pulmonary tuberculosis with no extrapulmonary involvement.

Almost 79% of the cases were in fairly good condition, although all 95 cases complained from night sweats, cough and sputum. The duration of these symptoms ranged from 2-156 weeks, with a mean of about 21 weeks. Hemoptysis was present in 48.4% of cases (mean duration 2.7 weeks, range 0–24). The offspring's weight ranged from 13.7 to 66.5 kg (mean 40.3). Weight loss was documented for all cases. Weight loss of more than 10 kg was reported in 25.3% of the cases. Only 2.1% had BCG scars showing evidence of vaccination, and no one had been previously treated for tuberculosis. All case subjects had a positive sputum smear examination, although only 62% were culture positive.

The age and ethnic distribution of the parents are shown in tables 5.3 and 5.4. Like the offspring, the majority (79%) of parents were Hadiyan; not surprisingly, about 98% were 30 years and older. The mean age for parents was about 45 years, ranging from 25 to 90. Note that the demographic information about the parents is not used in the subsequent covariate analysis.

Table 5.1: Frequency of selected variables among affected offspring (N = 95)

Variable	Coding	Final coding	Frequency	Percent
Gender	1 = Male 2 = Female	1 = Male 0 = Female	51 44	53.7 46.3
Generation	1 = G_1 2 = G_2	1 = G_1 0 = G_2	0 95	0.0 100.0
Ethnicity	1 = Hadiya 2 = Other	1 = Hadiya 0 = Other[a]	72 23	75.8 24.2
Age-group (years)	1 = (7–15) 2 = (15–19) 3 = (20–24) 4 = (25–29) 5 = (≥30)	1 = (15–19)[b] 0 = Other ; and age groups as dummy variables with 15–19 age group as reference	21 38 22 8 6	22.1 40.0 23.2 8.4 6.3
Extrapulmonary TB	1 = Yes 2 = No	1 = Yes 0 = No	0 95	0.0 100.0
General condition	1 = Good 2 = Bad 3 = Cachectic 4 = Bed-Ridden	1 = Good 0 = Other	75 10 9 1	78.9 10.5 9.5 1.1
Cough	1 = Yes 2 = No	1 = Yes 0 = No	95 0	100.0 0.0
Cough-sputum	1 = Yes 2 = No	1 = Yes 0 = No	95 0	100.0 0.0
Hemoptysis	1 = Yes 2 = No	1 = Yes 0 = No	46 49	48.4 51.6
Night sweats	1 = Yes 2 = No	1 = Yes 0 = No	95 0	100.0 0.0
Wt-loss>10 kg	1 = Yes 2 = No	1 = Yes 0 = No	24 71	25.3 74.7
Acid fast bacilli (AFB)	1 = Pos 2 = Neg	1 = Pos 0 = Neg	95 0	100.0 0.0
Culture	1 = Pos 2 = Neg	1 = Pos 0 = Neg	59 36	62.1 37.9
BCG scar	1 = Yes 2 = No	1 = Yes 0 = No	2 93	2.1 97.9
Previous treatment	1 = Yes 2 = No	1 = Yes 0 = No	0 95	0.0 100.0

a: includes Amhara, Gurage, Kembata, Silti, and those whose parents were from different ethnic groups (father's/mother's ethnic group) including Amhara/Hadiya, Hadiya/Amhara, Hadiya/Kembata, Kembata /Hadiya, Hadiya/Gurage, Hadiya/Gurage-Silti, Gurage-Silti, Silti/Hadiya, and Gurage-Sebata

b: this age group was selected as the reference group, because it included the largest number of cases

Table 5.2: Distribution of selected variables among affected offspring (N = 95)

Variable	Mean	SD	Min	Max
Age (yrs)	18.05	6.02	7.00	35.00
Weight (kg)	40.32	12.14	13.70	66.50
Cough duration (wks)	21.18	23.71	2.00	156.00
Hemoptysis duration (wks)	2.75	5.11	0.00	24.00
Cough-sputum duration (wks)	20.88	23.80	2.00	156.00
Night sweats duration (wks)	21.07	23.72	2.00	156.00

Table 5.3: Frequency of demographic variables among parents (N = 190)

Variable	Coding	Frequency	Percent
Gender	1 = Male	95	50.0
	2 = Female	95	50.0
Generation	1 = G_1	190	100.0
	2 = G_2	0	0.0
Ethnicity	1 = Hadiya	150	79.0
	2 = Other	40	21.0
Age-group (years)	1 = (<15)	0	0.0
	2 = (15–19)	0	0.0
	3 = (20–24)	0	0.0
	4 = (25–29)	4	2.1
	5 = (≥30)	186	97.9

Table 5.4: Distribution of age among parents (N = 190)

Variable	Mean		SD		Min		Max	
	Mother	Father	Mother	Father	Mother	Father	Mother	Father
Age (yrs)	44.90		10.49		25.00		90.00	
	40.23	49.57	7.79	10.80	25.00	28.00	65.00	90.00

5.2 GENE FREQUENCIES AND HETEROZYGOSITY

The frequencies of marker alleles at candidate genes were measured among both affected children and their parents. Tables 5.5 through 5.8 describe the frequencies of genes and heterozygosity at the 4 loci and 13 markers for the affected offspring and their parents. In the Methods chapter, table 4.1 illustrated which polymorphisms correspond to the allele labels "1", "2", and "3".

Table 5.5 describes 5 genotypes identified at marker NRAMP1-5′(GT)n, out of the possible six genotypes for a marker with 3 alleles. Only this marker had more than 2 alleles and only 6 individuals carried allele "3". The highest number of missing genotypes in both offspring and their parents, on average, was observed at locus *SFTPA1* (28%); whereas for the rest of the loci including *NRAMP1*, *MBL* and *VDR*, about 13% of genotypes were missing for all study subjects.

The frequency of allele "1" and allele "2" was measured at each marker for the parents. The frequency of allele "1" at NRAMP1-5′(GT)n was 0.69 ([(2×78)+(1×69)]/ 2×163), and 0.31 for allele "2". The frequency of allele "1" was as follows (at each of the markers): 0.28 at marker NRAMP1-274C/t; 0.74 at NRAMP1-469+14G/c; 0.50 at NRAMP1-3′UTR; 0.90 at MBL-G54D; 0.87 at MBL-G57Q; 0.81 at VDR-117; 0.56 at VDR-1056; 0.93 at SFTPA1-170; 0.75 at SFTPA1-294; 0.30 at SFTPA1-256; 0.11 at SFTPA1-763; and 0.94 at SFTPA1-507. The allele frequencies for the affected children were approximately the same as their parents at each marker. The average frequency of the rarer allele was 0.23 with a range from 0.06 to 0.50.

The heterozygosity measures the genetic diversity or extent of genetic variation in a population for a marker. A value close to zero shows that the heterozygosity is low and the marker is not informative, and the maximum heterozygosity for a biallelic marker is 0.5. These values are calculated using the following equation and listed in the last columns of tables 5.5–5.8. In this equation "q" stands for allele frequency; "k" is a given marker; and "*i*" is the number of alleles (Elseth & Baumgardner, 1995b):

$$H_k = 1 - \sum_{i=1}^{n} q_{ik}^2$$

In table 5.8, for instance, the heterozygosity for parents at SFTPA1-170 was 0.13 (1-(0.93^2 + 0.07^2)), where 0.93 is the frequency distribution of allele "1" and 0.07 is the

frequency distribution of allele "2". The average heterozygosity, locus by locus, was calculated by the mean of the marker-specific values. Two loci had high average heterozygosities (0.43 for *NRAMP1*, 0.40 for *VDR*), but markers at *MBL* and *SFTPA1* had low average heterozygosity (0.20 and 025, respectively). Hence, associations with the latter 2 markers may be harder to identify since few parents will be heterozygous. The average heterozygosity of all loci in this study was 0.32 (4.16/13) in the parents.

Table 5.5: Frequencies of genes and heterozygosity at *NRAMP1* locus among offspring and their parents

NRAMP1-5′(GT)N							
Status	Genotype						
	1,1	1,2	2,2	1,3[a]	2,3[b]	# Missing	Heterozygosity
Case	39	36	6	2	0	12	0.42
Parents	78	69	16	2	2	23	0.43
NRAMP1-274C/t							
Status	Genotype						
	1,1	1,2	2,2	# Missing	Heterozygosity		
Case	6	39	38	12	0.43		
Parents	16	63	89	22	0.40		
NRAMP1-469+14G/c							
Status	Genotype						
	1,1	1,2	2,2	# Missing	Heterozygosity		
Case	43	37	3	12	0.38		
Parents	94	60	13	23	0.38		
NRAMP1-3′UTR							
Status	Genotype						
	1,1	1,2	2,2	# Missing	Heterozygosity		
Case	17	47	19	12	0.50		
Parents	38	90	39	23	0.50		

a and b columns were not included in the calculations of allele frequency and heterozygosity

Table 5.6: Frequencies of genes and heterozygosity at *MBL* genes among offspring and their parents

MBL-G54D					
Status	Genotype				
	1,1	1,2	2,2	# Missing	Heterozygosity
Case	71	11	1	12	0.12
Parents	136	30	1	23	0.18
MBL-G57Q					
Status	Genotype				
	1,1	1,2	2,2	# Missing	Heterozygosity
Case	63	18	2	12	0.23
Parents	129	36	3	22	0.22

Table 5.7: Frequencies of genes and heterozygosity at *VDR* genes among offspring and their parents

VDR-117					
Status	Genotype				
	1,1	1,2	2,2	# Missing	Heterozygosity
Case	58	23	2	12	0.27
Parents	108	57	3	22	0.31
VDR-1056					
Status	Genotype				
	1,1	1,2	2,2	# Missing	Heterozygosity
Case	28	40	15	12	0.49
Parents	60	68	39	23	0.49

Table 5.8: Frequencies of genes and heterozygosity at *SFTPA1* genes among offspring and their parents

SFTPA1-170					
Status	Genotype				
	1,1	1,2	2,2	# Missing	Heterozygosity
Case	64	9	0	22	0.11
Parents	129	20	0	41	0.13
SFTPA1-294					
Status	Genotype				
	1,1	1,2	2,2	# Missing	Heterozygosity
Case	49	30	3	13	0.62
Parents	95	61	11	23	0.37
SFTPA1-256					
Status	Genotype				
	1,1	1,2	2,2	# Missing	Heterozygosity
Case	6	30	43	16	0.39
Parents	12	72	74	32	0.42
SFTPA1-763					
Status	Genotype				
	1,1	1,2	2,2	# Missing	Heterozygosity
Case	1	9	48	37	0.16
Parents	3	19	94	74	0.20
SFTPA1-507					
Status	Genotype				
	1,1	1,2	2,2	# Missing	Heterozygosity
Case	45	6	0	44	0.11
Parents	92	12	0	86	0.11

5.3 CORRELATION BETWEEN COVARIATES

The correlation coefficients between covariates in the affected children were assessed in a bivariate analysis. Cough-sputum duration and night sweats duration were the only variables found to be highly correlated (r = 0.998). It is likely that these two variables are both measuring a similar phenotype–illness duration. Some other covariates showed significant correlations that are listed in table 5.9. Smaller correlations were observed between duration of cough-sputum and night sweats and duration of hemoptysis. Older children were heavier and also lost more weight than younger children and children with lower weights. More positive cultures were observed in older children who also lost more weight. Highly correlated variables were not included together (assessed independently) in the subsequent analysis (multivariate logistic regression) due to problems of collinearity.

Table 5.9: Correlation coefficients (Pearson) between variables

Variables	(p-value) Correlation coefficient
Age by weight	(p< 0.0001) 0.660
Age by weight loss > 10 kg	(p< 0.0001) 0.411
Age by culture positive	(p = 0.034) 0.217
Age by cough-sputum duration	-0.052
Age by night sweats duration	-0.058
Age by hemoptysis duration	-0.138
Age by sex	-0.036
Weight by weight loss > 10 kg	(p = 0.018) 0.241
Weight loss > 10 kg by cough-sputum duration	-0.074
Weight loss > 10 kg by culture positive	(p = 0.046) 0.204
Weight loss > 10 kg by ethnicity	-0.180
Cough-sputum duration by culture positive	0.051
Cough-sputum duration by hemoptysis duration	(p = 0.0003) 0.360
Cough-sputum duration by night sweats duration	(p< 0.0001) 0.998
Night sweats duration by hemoptysis duration	(p = 0.0004) 0.358
Night sweats duration by weight loss > 10 kg	-0.079

5.4 TRANSMISSION DISEQUILIBRIUM TEST (TDT)

The transmission disequilibrium test was performed by the computer program TDT (Spielman & Ewens, 1999). This program only analyzes the data from families where one or both parents are heterozygous for a given marker allele and excludes the rest of the families (see Methods). A family with two heterozygous parents contributes two observations to the analysis. At marker NRAMP1-5′(GT)n, any individual carrying the third allele (allele 3) was excluded from the analysis, because the number of people with this allele was too small to detect any association. Moreover, the classic TDT uses only biallelic markers.

Table 5.10 illustrates the TDT results (TDT outputs are included in the appendix II). Among markers tested, only SFTPA1-294 was significantly associated with the outcome ($\chi^2 = 4.413$; $p = 0.035$). There were 58 heterozygous parents of whom 37 transmitted allele "1" and 21 transmitted allele "2" to their offspring. Therefore, approximately 64% (%[37/37+21]) of the offspring received the "1" allele. Hence, this distortion implies an association between the surfactant polymorphism "294", or a variant in linkage disequilibrium with this marker. A nearby polymorphism SFTPA1-763, showed a similar transmission distortion of 68%, although the test was not statistically significant. Furthermore, the marker with the largest number of heterozygous parents (NRAMP1-3′UTR) (hence the most power to detect an association) showed no association at all. These results must be interpreted cautiously since 13 markers were tested for association. One result significant at $p < 0.05$ could easily occur by chance alone.

The classic TDT only tests for association between a single marker and disease. Effects of other markers or environmental factors have been examined by other statistical tools that will follow.

Table 5.10: Transmission disequilibrium test (TDT), only if parent is heterozygous

Marker allele	χ^2	p-value	Ctn	Cnt	# Observations
NRAMP1-5′(GT)n-A1 NRAMP1-5′(GT)n-A2	0.138	0.709	34 31	31 34	65
NRAMP1-274C/T-A1 NRAMP1-274C/T-A2	1.066	0.301	34 26	26 34	60
NRAMP1-469+14G/C-A1 NRAMP1-469+14G/C-A2	0.438	0.507	26 31	31 26	57
NRAMP1-3′UTR-A1 NRAMP1-3′UTR-A2	0.000	1.000	43 43	43 43	86
MBL-G54D-A1 MBL-G54D-A2	1.689	0.193	18 11	11 18	29
MBL-G57Q-A1 MBL-G57Q-A2	0.028	0.865	17 18	18 17	35
SFTPA1-170-A1 SFTPA1-170-A2	0.222	0.637	10 8	8 10	18
SFTPA1-256-A1 SFTPA1-256-A2	1.246	0.264	28 37	37 28	65
SFTPA1-294-A1 SFTPA1-294-A2	4.413	0.035	37 21	21 37	58
SFTPA1-507-A1 SFTPA1-507-A2	0.111	0.738	4 5	5 4	9
SFTPA1-763-A1 SFTPA1-763-A2	2.578	0.108	6 13	13 6	19
VDR-117-A1 VDR-117-A2	1.851	0.173	32 22	22 32	54
VDR-1056-A1 VDR-1056-A2	0.062	0.802	31 33	33 31	64

Ctn: Count of allele transmitted
Cnt: Count of allele non-transmitted
Observations: Reflects the number of heterozygous parents or in other words the number of times transmission/nontransmission of alleles from a heterozygous parents to a child has been assessed
A1: First allele
A2: Second allele

5.5 RELATIVE RISKS AND HEREDITARY MODE OF TRANSMISSION

Relative risks (RR) for homozygotes "2,2" and heterozygotes "1,2" versus homozygotes "1,1" were estimated using the genetic statistical program "gassoc" (Schaid & Rowland, 2001). This program considers genotype "1,1" as the comparison genotype and then measures the relative risks of other two genotypes "1,2" and "2,2" compared to

"1,1." These RRs help estimate the strength of the associations. We also tested for mode of inheritance.

Two marker alleles SFTPA1-294 and SFTPA1-763 showed significant associations. At marker SFTPA1-294, the results showed that genotype "2,2" is a low risk genotype when compared to genotype "1,1" (ß = -2.131; RR = 0.118; p = 0.047). Genotypes "1,2" and "1,1" had about the same risk with slightly lower (and nonsignificant) risk for "1,2" (ß = -0.356; RR = 0.7; p = 0.251). The results of the mode of inheritance test indicated a significant association under a recessive pattern (p = 0.043). This implies that offspring who received both susceptible alleles (1,1) at this marker are at higher risk for tuberculosis than those with other genotypes (1,2 or 2,2).

At marker SFTPA1-763 only one individual carried genotype "1,1", the gassoc program detected some evidence of association (p = 0.054), where genotype "2,2" was higher risk than "1,2" (β = 1.115 [0.189 - (-0.926)], RR = 3.049 [exp $^{1.115}$]). Although the p-value was obtained under a dominant mode of inheritance, the mode cannot be estimated well with only one case carrying "1,1." The outputs are included in appendix III.

5.6 LOGISTIC REGRESSION ANALYSIS

An adapted logistic regression was used to analyze the transmission data from the ninety-five nuclear families with covariates (see Methods). Both phenotypic and genotypic data from children, together with genotypic data from their parents were used. Transmitted alleles from heterozygous parents were precalculated and stored, and used as the binary outcome variable (e.g., if allele "1" was transmitted, Y = 1; if allele "2" was transmitted, Y = 0). Families with two heterozygous parents contributed 2 observations to the program. The number of missing values was different for each marker. Analysis included all families with genotype data at the marker being studied, even if data were missing at other markers. Therefore the number of observations varied across markers according to 1) number of heterozygous parents; and 2) number of missing values.

First, the logistic analysis was performed at each marker with only an intercept. The obtained χ^2 values were the same as in the TDT with slight differences at some markers (Tables 5.11–5.16). The 2 tests are asymptotically equivalent; this demonstration

confirmed that the precoding of the transmissions was correct. As before, marker SFTPA1-294 was the only marker found to be significantly associated with transmission of allele "1". The likelihood ratio chi-square (LR χ^2) for this marker was 4.297 (p = 0.038).

In the next step of the analysis, covariates were added to the models to evaluate the G x E effects (see Methods). If certain environmental or demographic characteristics interact with genotypes to modify the disease risk, then covariates will appear to affect the transmission probabilities. This analysis identified some additional markers as being associated with the disease when covariates were included in the models. At locus *NRAMP1*, a significant G x E was observed in the model with marker NRAMP1-5′(GT)n. Variables weight, age group, and culture positive (one at a time) showed associations with $p < 0.05$ (Table 5.11). The smallest overall p-value was observed when both age group and culture positive were considered. One 10 kilogram increase in weight is predicted to increase the probability of transmitting allele "1" by a factor of $e^{0.049} = 1.05$ (LR $\chi^2 = 4.91$, p = 0.026). Transmission distortion also appeared to differ by age. Age groups 3 (20–24 yr) and 4 (25–29 yr) showed excess transmission of allele "1", and were significantly different from the comparison age group 15–19 (p = 0.024 for ages 20–24; p = 0.057 for ages 25–29). There was a suggestion that culture status influenced transmission probabilities (p = 0.066) when considered alone. However, in a model also containing the age groups, the evidence that these 2 covariates had an effect was quite strong (LR $\chi^2 = 16.245$, p = 0.006). This model shows a strong bias towards allele "1" among culture positive cases between 20 and 29 years of age. Conversely, culture negative cases over 30 or under 15 show excess transmission of allele "2".

At locus *SFTPA1*, models with markers SFTPA1-170, SFTPA1-256, SFTPA1-763, and SFTPA1-507 all showed some associations with $p < 0.05$ when interactions with covariates were examined. In the model with SFTPA1-170, variables weight and weight loss > 10 kg, alone and together affected transmission distortion (Table 5.12). In this model, lower weight was associated with the transmission of allele "1". Weight loss of more than 10 kg was also associated with allele transmission, where 86% of offspring received allele "1" when weight had been lost (see appendix IV for calculations). The chi-square for weight and weight loss >10 kg were 3.314 (p = 0.068), and 3.545 (p =

0.059), respectively. The effect of each of these variables became bigger and more significant when included together in the model. However, it should be noted that there were only 18 heterozygous parents at SFTPA1-170. Hence, these relationships with covariates should be very cautiously interpreted.

In another model, interactions between SFTPA1-256 and covariates were examined (Table 5.13). Night sweats duration, and age group 2 (vs. other age groups), and cough-sputum duration interacted with allele transmission. In age group 15–19, 73% ($e^{-0.15+1.15}$ /1+ $e^{-0.15+1.15}$) of cases received allele "2" from their heterozygous parents. Longer cough-sputum duration ($p = 0.036$) and night sweats duration ($p = 0.030$) were also associated with more transmission of allele "2". Models containing age group 2 and duration of cough ($p = 0.005$); and age group 2 and duration of night sweats ($p = 0.003$) showed strong evidence for these altering transmission distortions. As was explained in the bivariate analysis earlier, night sweats duration and cough-sputum duration are highly correlated and hence the models containing these variables are very similar.

The third marker at locus *SFTPA1*, SFTPA1-763, showed borderline significant evidence of association when age (continuous) was included in a gene-environment interaction (Table 5.14). The model shows that a 1 year increase in age is associated with an increased risk of transmission of allele "2" by 1.18 ($e^{0.166}$). Although ethnicity (Hadiya vs. other) alone had a small effect ($\chi^2 = 3.041$; $p = 0.081$), when added to the model with age, the overall model became more significant (LR $\chi^2 = 7.026$, $p = 0.029$). There were only 19 heterozygous parents at SFTPA1-763. Therefore, these relationships with variables should be cautiously interpreted.

None of the covariates had significant effects on SFTPA1-249, the only marker that showed a significant association in the model with only an intercept.

At the third locus under study, *VDR*, both markers showed some significant associations when weight or weight loss were included in the models (Tables 5.15 and 5.16). For VDR-1056, transmission of allele "2" was greater at lower weights ($\chi^2 = 3.648$, $p = 0.056$). For VDR-117, 91% of those reporting a large weight loss received allele "1" compared to 49% of those without the weight loss ($\chi^2 = 4.266$, $p = 0.038$). No gene environment interactions were identified with any markers at MBL.

Finally, all possible interactions between unlinked markers that showed important effects previously were examined. The interaction between VDR-117 and SFTPA1-294 was just significant ($\chi^2 = 3.945$, $p = 0.047$; LR $\chi^2 = 4.603$, $p = 0.031$) (Table 5.17). When hemoptysis was also included, the association became stronger ($\chi^2 = 4.691$, $p = 0.030$; LR $\chi^2 = 8.033$, $p = 0.018$). In this model, transmission at VDR-117 was considered as the outcome variable (Y) and transmission at SFTPA1-294 was the predictor variable (X). As it was discussed in the Methods, parents homozygous at "X" were included in this analysis, but the covariate was coded "0". The results of table 5.17 assume that the predictor variable has a linear effect (i.e., bias in transmission at the outcome variable change linearly with the predictor variable going from allele "2" transmission (-1) to homozygous (0) to allele "1" (1)).

The last 3 columns of table 5.17 break down the transmission distortion at VDR-117 by the SFTPA1-294 alleles received. Distortion towards allele "1" at VDR-117 is strongest when allele "1" is transmitted at SFTPA1-294 from a heterozygous parent ($p = 0.079$) (80% received allele "1" [$e^{1.38}/1+ e^{1.38}$]).

Furthermore, excess transmissions of allele "1" at VDR-117 were seen more often with transmission of allele "1" at NRAMP1-5′(GT)n ($p = 0.091$) (Table 5.18). Addition of the weight loss > 10 kg covariate, however, weakened the gene-gene interaction. Although the p-value for the covariate itself did not show any significance, the overall model became significant ($p = 0.012$), reflecting the previously shown effect of weight loss. Examination of subgroup estimates showed that the VDR-117 distortion was much larger when allele "1" was transmitted at NRAMP1-5′(GT)n from a heterozygous parent (79% of cases received allele "1" at VDR-117 when NRAMP1-5′(GT)n allele "1" was transmitted, and 45% of cases received allele "1" at VDR-117 when NRAMP1-5′(GT)n allele "2" was transmitted).

Many different regression models were fit to these data, and the number of observations was small (very small for some markers). Hence, all these results could be chance findings and should be cautiously interpreted.

Goodness of fit was checked by Hosmer-Lemeshow for the models discussed above. The p-values were not significant, but there was no problem with the fit, all the observed and expected values were close to each other.

Table 5.11: Logistic regression results for marker allele "NRAMP1-5′(GT)n"

Marker allele: NRAMP1-5′(GT)n; TDT $\chi^2 = 0.138$, $p = 0.709$						
Covariate	Estimate	SE	χ^2	p	Likelihood ratio χ^2	p
Only intercept	-0.092	0.248	0.138	0.709		
Intercept	1.969	1.020	3.724	0.536		
Weight (kg)	-0.049	0.023	4.371	0.036	4.910	0.026
Intercept	0.405	0.408	0.986	0.241		
Age-group 1	0.287	0.736	0.152	0.695		
Age-group 3	-1.584	0.702	5.084	0.024	12.253	0.015
Age-group 4	-2.196	1.154	3.620	0.057		
Age-group 5	0.693	1.224	0.320	0.571		
Intercept	0.619	0.468	1.743	0.186		
Culture positive	-1.024	0.558	3.359	0.066	3.501	0.061
Intercept	1.353	0.668	4.103	0.042		
Age-group 1	0.035	0.774	0.002	0.964		
Age-group 3	-1.812	0.752	5.799	0.016		
Age-group 4	-2.141	1.180	3.293	0.069	16.245	0.006
Age-group 5	1.001	1.236	0.656	0.417		
Culture positive	-1.256	0.650	3.730	0.053		

Table 5.12: Logistic regression results for marker allele "SFTPA1-170"

Marker allele: SFTPA1-170; TDT $\chi^2 = 0.222$, $p = 0.637$						
Covariate	Estimate	SE	χ^2	p	Likelihood ratio χ^2	p
Only intercept	-0.223	0.474	0.221	0.633		
Intercept	-5.667	3.144	3.248	0.071		
Weight (kg)	0.124	0.068	3.314	0.068	6.225	0.012
Intercept	0.559	0.626	0.797	0.371		
Weight loss>10 kg	-2.351	1.248	3.545	0.059	4.568	0.032
Intercept	-9.069	5.792	2.451	0.117		
Weight	0.251	0.144	3.040	0.081		
Weight loss>10 kg	-5.035	2.421	4.325	0.037	15.465	0.0004

Table 5.13: Logistic regression results for marker allele "SFTPA1-256"

Marker allele: SFTPA1-256; TDT χ^2 = 1.246, p = 0.264						
Covariate	Estimate	SE	χ^2	p	Likelihood ratio χ^2	p
Only intercept	0.278	0.250	1.238	0.265		
Intercept Age-group 2	-0.154 1.152	0.321 0.546	0.230 4.448	0631 0.034	4.734	0.029
Intercept Night sweats duration (wk)	-0.712 0.052	0.492 0.024	2.094 4.671	0.147 0.030	6.969	0.008
Intercept Age-group 2 Night sweats duration	-1.192 1.155 0.055	0.578 0.577 0.026	4.253 3.995 4.616	0.039 0.045 0.031	11.217	0.003
Intercept Cough sputum duration (wk)	-0.653 0.049	0.482 0.023	1.836 4.369	0.175 0.036	6.480	0.010
Intercept Age-group 2 Cough sputum duration	-1.105 1.129 0.051	0.561 0.573 0.025	3.874 3.876 4.227	0.049 0.049 0.039	10.588	0.005

Table 5.14: Logistic regression results for marker allele "SFTPA1-763"

Marker allele: SFTPA1-763; TDT χ^2 = 2.578, p = 0.108						
Covariate	Estimate	SE	χ^2	p	Likelihood ratio χ^2	p
Only intercept	0.773	0.493	2.454	0.117		
Intercept Age (continuous)	-2.151 0.166	1.689 0.096	1.622 2.983	0.202 0.084	3.766	0.052
Intercept Hadiya	-0.287 1.897	0.763 1.087	0.141 3.041	0.706 0.081	3.324	0.068
Intercept Age Hadiya	-3.446 0.169 2.133	2.058 0.100 1.262	2.802 2.854 2.854	0.094 0.091 0.091	7.026	0.029

Table 5.15: Logistic regression results for marker allele "VDR-1056"

Marker allele: VDR-1056; TDT $\chi^2 = 0.062$, $p = 0.802$						
Covariate	Estimate	SE	χ^2	p	Likelihood ratio χ^2	p
Only intercept	0.062	0.250	0.062	0.802		
Intercept	1.743	0.920	3.586	0.058		
Weight (kg)	-0.041	0.021	3.648	0.056	3.942	0.047

Table 5.16: Logistic regression results for marker allele "VDR-117"

Marker allele: VDR-117; TDT $\chi^2 = 1.851$, $p = 0.173$						
Covariate	Estimate	SE	χ^2	p	Likelihood ratio χ^2	p
Only intercept	-0.374	0.277	1.830	0.176		
Intercept	-0.046	0.305	0.023	0.878		
Weight loss>10 kg	-2.256	1.092	4.266	0.038	6.707	0.009

Table 5.17: Gene-gene interaction (VDR-117 X SFTPA1-294)

Marker allele interaction	Y= VDR-117 X= SFTPA1-294	Including Hemoptysis	SFTPA1 allele 1	SFTPA1 allele 2	SFTPA1 homozygous
Likelihood ratio χ^2	4.603	8.033	6.52		
p	0.031	0.018	0.089		
Estimate (X)	-1.148	-1.322	-1.386	0.916	-0.383
SE	0.578	0.610	0.790	0.836	0.335
χ^2	3.945	4.691	3.074	1.199	1.308
p	0.047	0.030	0.079	0.273	0.253

Columns 2 and 3 show the effect of allele transmissions at SFTPA1on allele transmissions at VDR, assuming a linear relationship; where transmissions from homozygous parents are coded as 0, transmissions of allele 1 from heterozygous parents are coded as 1, and transmissions of allele 2 from heterozygous parents are coded as -1
The last 3 columns assess the transmission distortion at VDR for the three SFTPA1 subcategories of individuals: transmission of allele 1 from parents heterozygous at SFTPA1, transmission of allele 2 from heterozygous parents, and transmission from homozygous parents

Table 5.18: Gene-gene interaction (VDR-117 X NRAMP1-5′(GT)n)

Marker allele interaction	Y= VDR-117 X= NRAMP1-5′(GT)n	Including weight loss> 10 kg	NRAMP1 allele 1	NRAMP1 allele 2	NRAMP1 homozygous
Likelihood ratio χ^2	3.031	8.740	5.26		
p	0.081	0.012	0.154		
Estimate (X)	-0.733	-0.6306	-1.299	0.182	-0.208
SE	0.434	0.4526	0.651	0.606	0.373
χ^2	2.844	1.9413	3.979	0.091	0.309
p	0.091	0.163	0.046	0.763	0.578

Columns 2 and 3 show the effect of allele transmissions at NRAMP1on allele transmissions at VDR, assuming a linear relationship; where transmissions from homozygous parents are coded as 0, transmissions of allele 1 from heterozygous parents are coded as 1, and transmissions of allele 2 from heterozygous parents are coded as -1
The last 3 columns assess the transmission distortion at VDR for the three NRAMP1 subcategories of individuals as described above

CHAPTER 6: DISCUSSION

The case-parental control design was used to investigate the role of 4 candidate genes (13 markers) on the risk of developing tuberculosis. In Ethiopia, it is very difficult to enroll large families for a study. Furthermore, the effects of any of the candidate genes for tuberculosis are expected to be relatively small. Therefore, a linkage study would have been inappropriate, and an association study was the most appropriate design to test the effects of the candidate genes.

In this family-based study, genotypic and phenotypic data were analyzed for 95 nuclear families (with one affected child and two parents), selected from the AHRI database with various family structures. Analyses were undertaken independently at each marker. There were two main concerns in this study, including i) missing values for marker alleles, which caused a shrinkage in the study population, since families with missing value(s) were eliminated from the analysis; and ii) lack of heterozygosity (at some markers) in the parents, since establishment of an association is more difficult when fewer parents are heterozygous.

Regarding covariates, the small sample size in most age groups, and the method of grouping of age, when used as a categorical variable, were other concerns in this study. Only six affected offspring were 30 years and older. Therefore, age was grouped into 5-year intervals, with the top category consisting of "30 and older." Different forms for the age covariate were used in the final models including: a) age group 2 (15–19 yr) with highest number of subjects versus others; b) 4 age categories (dummies) with age group 2 as the reference age group (table 5.1); and c) age as a continuous variable. Due to the probably inaccuracies in age, the latter form may lead to parameter estimates, for the effect of age, that are the most susceptible to error.

6.1 TRANSMISSION DISEQUILIBRIUM TEST

Since only the genotyping data of heterozygous parents are analyzed, the number of heterozygous parents is critical for the TDT and the multivariate analyses. Hence, markers with few heterozygous parents will have low power to detect associations.

The primary finding of this study was the significant association between marker SFTPA1-294 and transmission of allele "1" (or transmission of allele with an "adenine"

substitution, in terms of polymorphism). Sixty-four percent of affected offspring in this study were shown to have received allele "1" at this marker, which implies an association between this allele and tuberculosis. However, the association with this surfactant marker, if real, does not necessarily imply that the SFTPA1-294 marker increases the disease risk. Due to LD, a functional variant may be nearby within *SFTPA1*, or nearby but in another gene or in a regulatory region. Once this association is validated in another data set, search for a functional variant may be undertaken.

The power to detect an association will increase with the number of heterozygous parents. For example, there were only 19 heterozygous parents at SFTPA1-763; although the TDT is not significant, 68% of offspring received allele "2," similar to the distortion observed at SFTPA1-294. If additional families heterozygous at SFTPA1-763 were recruited, this marker might also demonstrate an important association, and the evidence for a role for surfactant protein in tuberculosis susceptibility would be strengthened, since results from several linked markers are stronger than one isolated marker. It would be interesting in the future to examine risks associated with 2-marker haplotypes.

6.2 RELATIVE RISKS AND MODE OF TRANSMISSION

At marker SFTPA1-294, there was evidence to support a recessive mode of inheritance. In other words, patients must receive both susceptible alleles (allele 1) to be at high risk. Therefore, at this marker, genotype "1,1" should be considered as high risk and the other 2 genotypes (2,2 and 1,2) as low risk genotypes for TB.

At the nearby marker SFTPA1-763, the RRs measured at this marker were not significantly different from 1.0. However, this was probably due to the fact that only one case carried the "1,1" genotype, the genotype that was automatically allocated to be the reference genotype by the program (gassoc). Nevertheless, it was shown that the "1,2" genotype was at lower risk than the "2,2" genotype.

The TDT assumes a log additive model, where the relative risk for individuals homozygous for the susceptibility allele is the square of the relative risk for one copy of the high-risk allele. This model can be thought of as intermediate between recessive and dominant models. The TDT therefore has one fewer parameter than the more general models that estimate the separate relative risks R_1 and R_2, and can be expected to be a

more powerful test on that basis. However, it is useful to be able to estimate the mode of inheritance, since future investigations of the role of a candidate gene may be helped by knowing the possible mode.

6.3 LOGISTIC REGRESSION ANALYSIS

Several covariates proved to be statistically significant modifiers ($p< 0.05$) of the effect of the genetic markers on the risk of tuberculosis. For each marker, approximately 20 models including all possible combinations of covariates were tested. It was expected to observe an important effect for gender and age. Although important effects were found for age, no effects were seen for gender. This study showed that the genetic effects are different in younger individuals than in older individuals, but it is similar in the 2 genders. In children, TB rates are normally the same in males and females. Perhaps the factors that lead to differences in TB rates by gender among young adults were not seen in this study since most of the affected offspring were young.

The models that showed significant associations when certain variables were allowed to modify the effects are discussed below. Overall, numerous regression models were fit to the data. Therefore, all these associations with covariates could be chance findings and should be cautiously interpreted and validated in independent data sets. In each case, the genes could be influencing susceptibility to disease, or progression of disease to a clinically serious state such that the affected offspring came to the clinic for treatment.

1) The models with marker NRAMP1-5′(GT)n and covariates weight, age group and culture positive (one at a time). The results showed that there was a larger bias towards transmission of allele "1" (a 133 bp allele) among cases between 20–29 years (age groups 3 and 4), heavier cases and culture positive cases. Hence, it could be by hypothesized that the *NRAMP1* gene is playing a role in TB susceptibility or progression among older and heavier individuals. However, interpretation of such results is difficult. This association could be due to unknown factors such as hormonal, emotional or other factors that influence the immune system at these years of age.

2) The model with marker SFTPA1-170 and variables weight and weight loss >10 kg. It appears that individuals who lost more than 10 kg weight, and as a result have low

weights, are those receiving allele "1." Hence, this implies that the effect of SFTPA1-170 on TB is more important among those with low weight. However, this result must be interpreted cautiously as the sample size is extremely small (only 18 parents were heterozygous at this marker). From another point of view, the wide range of both weight (14–67 kg) and age (7–35 yr) in the offspring makes the interpretation very difficult. A continuous measurement of weight loss could probably provide more information for this assessment. Furthermore, since age and weight are somewhat confounded, a larger sample size would be required to study the effect of weight independent of age.

3) The model with marker SFTPA1-256 and duration of night sweats, or duration of cough-sputum and age between 15–19 years (vs. other ages). The effect of these covariates became larger and more significant when age group 2 (15–19 yr) was included together with either cough-sputum duration or night sweats duration. Again, it is difficult to interpret the results. It can be suggested that those who received allele "2" at this marker are more likely to develop tuberculosis between ages 15–19 years and that night sweats and productive cough occur earlier during the course of the disease. It is also possible that patients of this age group paid less attention to their disease, and consequently visited the clinic visited when the disease was more advanced. Individuals with longer disease duration may also be those who have lost more weight; hence this model may be detecting the same phenomenon seen at SFTPA1-170.

4) The model with age (continuous) and ethnicity at marker SFTPA1-763. The overall model showed more significance when both variables were considered. It appears that excess transmission of allele "2" to the affected children increases with age, and the distortion is particularly strong among Hadiyan. Only 19 parents were heterozygous at this marker, so results must be interpreted cautiously. The same hypothetical interpretation can be used for this model: Hadiyan children who received the allele "2" from their parents at marker SFTPA1-763 tend to develop tuberculosis at an older age.

5) The model with marker VDR-1056 and weight. At this marker, distortion of allele "2" transmission was greater at lower weight. This result could imply that the effect of the Vitamin D receptor on susceptibility is greater among younger or malnourished

individuals, or that the Vitamin D receptor influences disease progression. However, since age and weight are confounded, their effects would need to be studied in larger samples, together with an assessment of whether sun exposure varies with age or weight.

6) The model with marker VDR-117 and weight loss > 10 kg. At this marker, a larger bias was observed in the transmission of allele "1" among offspring reporting a large weight loss. This result may indicate that those who received allele "1" at the VDR-117 marker lose more weight during the illness. However, as was explained in other models, this interpretation is questionable since age and weight are confounded.

7) Finally, a strong transmission distortion towards allele "1" was found when interactions between markers VDR-117 and SFTPA1-294; and VDR-117 and NRAMP1-5′(GT)n were examined. There is significant transmission distortion at VDR-117 for offspring who received allele "1" at NRAMP1-5′(GT)n, but only negligible transmission distortion for those who received allele "2." Hemoptysis seems to be the prominent symptom in those who received the "1" allele at SFTPA1-294. Weight loss > 10 kg is also more frequently seen in those who received the "1" allele from their parents at NRAMP1-5′(GT)n.

The covariate effects discussed above may vary among case-patients as a result of interactions between an individual's genes and environment (including the myco-bacteria's genes). Such variation could lead to a more severe, longer duration of symptoms, development of the disease early at childhood or later in life, or milder forms of TB with different characteristics of the disease. Awareness of the role of nutrition status, living arrangements, humidity and other environmental factors that may affect the acquisition of the infection and also the interval between infection and developing the disease is also important. Although these factors have not been measured in our study population family–by–family, investigators who collected the data have estimated a low socioeconomic status for all families recruited in the study. Therefore, it is presumed that all families have the same low standard of living, which eliminates the role of these factors in differences observed in the phenotypes.

6.4 ASSESSMENT OF STUDY POWER

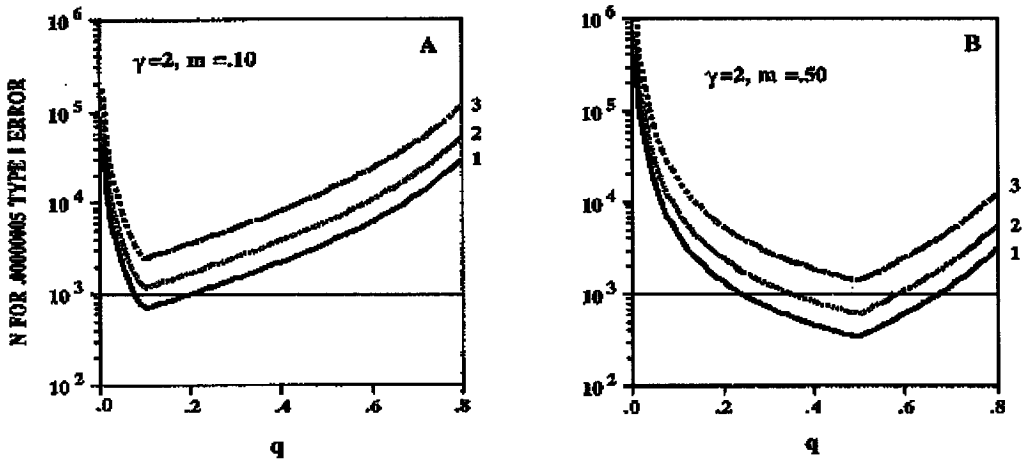
The power of the TDT depends on i) the strength of the linkage disequilibrium (Δ) between the analyzed marker and the disease locus; and ii) the population frequencies of both the associated marker allele (m), and the disease susceptibility allele (q) (Risch & Merikangas, 1996; Abel & Müller-Myhsok 1998). Figures 6.1 and 6.2 illustrate variation in the required number of families (N) by q and m. The parameter "γ" measures the effect of the susceptibility gene. That is, if allele "1" is the high risk allele, "γ" is the relative risk for subjects with genotype "1,2" relative to genotype "1,1". A multiplicative model is assumed so that the relative risk for "1,1" subjects is γ^2. (In section 4.3.2 in the Methods, the same parameter was called R: these power calculations are performed assuming $R_2 = R_1^2$). In figure 6.1, using a genomewide-screening strategy, N is estimated for a type I error (α) of 5×10^{-8}, and for $\gamma = 2$ and different values for m (0.10 and 0.50), with various linkage disequilibrium strengths. In figure 6.2, where a candidate-gene strategy is used, N is estimated for a type I error of 5×10^{-4}, for $\gamma = 2$ and $\gamma = 4$ and $m = 0.50$. In both strategies, a power of (1 -β) 80% is considered.

In figure 6.1, for $\gamma = 2$, the required sample size is significantly greater than 1,000 families, except for the optimal situation ($m = q$ and $\delta = \delta_{max}$), when the smallest sample size is needed. With the candidate-gene approach, in figure 6.2, a smaller sample size is needed since many fewer tests are performed: the chosen α level corresponds to investigating 10 biallelic candidate genes. Hence, the type I error for each test is 0.0005, for an overall type I error of 0.05. Figure 6.2 demonstrates that a sample size near 100 families is required to detect susceptibility alleles with a genotype relative risk of 4.0, providing that the linkage disequilibrium, δ, is at least 75% of its possible maximum, and that the marker allele and the susceptibility allele have frequencies near 0.5. If the genotype relative risk is only 2.0, then at least 200 families would be required to detect the gene, under ideal conditions with maximum linkage disequilibrium.

In this thesis, we have investigated 13 candidate genes using 95 families. Hence, the sample size and the number of candidate genes examined is similar to the values in figure 6.2. Therefore, this study should have adequate power to detect a gene with a genotype relative risk of 4.0, and may be able to detect genes with smaller effects if the conditions

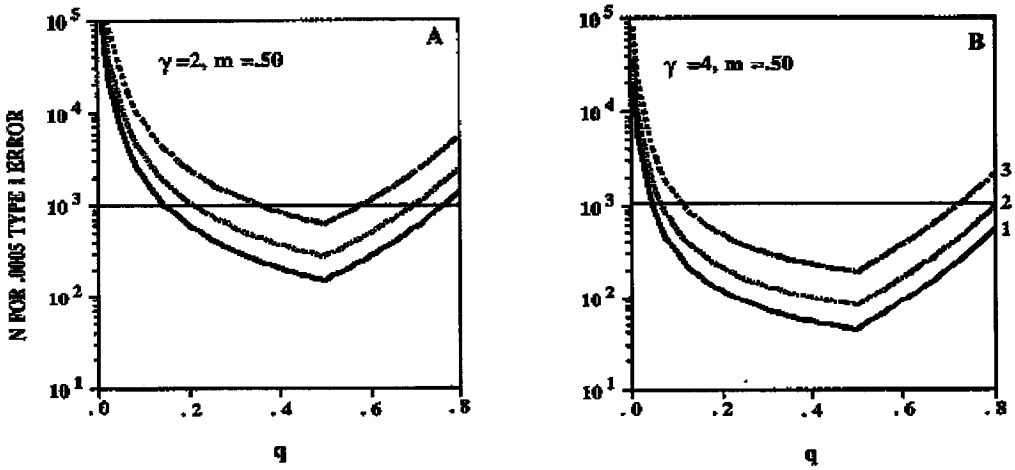
are right (i.e., if the associated marker allele is close to the frequency of the susceptibility allele, and the polymorphism used as a marker is in almost complete linkage disequilibrium with the susceptibility variant).

Figure 6.1: N (log scale) required for detection of association, for a type I error of 5×10^{-8} (genomewide-screening strategy) and a power of 80%, according to q, for $\gamma = 2$ (From Abel & Müller-Myhsok, 1998)



In curve 1, the maximum LD is assumed, in curve 2; $\delta = 0.75\ \delta_{max}$ is assumed; and in curve 3, $\delta = 0.50\ \delta_{max}$ is assumed

Figure 6.2: N (log scale) required for detection of association, for a type I error of 5×10^{-4} (candidate-gene strategy) and a power of 80%, according to q, for $\gamma = 2$ (A) and $\gamma = 4$ (B) (From Abel & Müller-Myhsok, 1998)



In curve 1, the maximum LD is assumed; in curve 2, $\delta = 0.75\ \delta_{max}$ is assumed; and in curve 3, $\delta = 0.50\ \delta_{max}$ is assumed

These power estimates apply only to the classic TDT results. However, in the logistic regression models with covariates, many more tests of association were performed, and so the curves in figure 6.2 are no longer appropriate. Many more families would be necessary to have good power to detect gene-environment interactions.

6.5 STRENGTHS AND LIMITATIONS OF THE AHRI AND THIS STUDY

The results obtained in this study need to be interpreted in light of its strengths and limitations. One important strength of the AHRI study was the choice of Hosanna which has a high incidence rate for tuberculosis and a low prevalence for HIV-positivity. Different genes may influence susceptibility among HIV-positive individuals, since the risk of TB is much larger and the immune system is already compromised.

Limitations to this study include those common to traditional case-control studies, including errors in recall and misclassification of the phenotypes and genotypes. There are two potential sources of measurement error in the AHRI study. The first relates to the way that demographic and clinical data have been assessed. Information on covariates such as age, weight loss, presence and duration of symptoms including cough, hemoptysis and night sweats were all self-reported. Study cases may not recall the exact duration or severity of their symptoms. A small weight loss may appear very large to some subjects but not to the others. Although the study physician attempted to more accurately estimate the age of the participants, there remains a high possibility for inaccuracy in the estimation of age, particularly in older subjects. This can seriously influence the interpretation of the results when effects of these variables are being assessed. The second source of misclassification bias is related to genetic marker typing. This analysis was subject to measurement error. Phenotypic data were only analyzed for the affected children, so differential or non-differential misclassification of these covariates was not problematic in this study. Genotyping errors could lead to exclusion of families due to apparent non-Mendelian inheritance, or to incorrect inference about allele transmission. However, it is unlikely that magnitude of errors depends on whether the person is an offspring or a parent. Therefore, errors in genotype data, although considered to be nondifferential, may have biased the true estimates towards the null.

In historical studies, where recorded data are generally used that have usually been created by different examiners and for different purposes, the possibility of information bias is relatively high. In the AHRI study, the data were collected prospectively and by only one knowledgeable physician who was also responsible for the physical examination of all study subjects. Clinical criteria for a given disease may vary over time and between physicians. Although the information obtained by study subjects might have led to information bias, this was minimized because only a single physician was involved in collecting this information. Moreover, the questionnaire designed for the AHRI study contained simple questions concerning symptoms related to TB that were not difficult or confusing for study participants to answer.

In the AHRI database, the data related to demographic characteristics, clinical and laboratory information were complete. However, there were many missing values for genotypes at different loci which may have affected the power of this study by eliminating the families with missing values from the analysis. Another important concern is about the validity of the results of the microscopic examination of sputum (the main criterion for an individual to be considered as a case), which was done either at Hosanna hospital or AHRI. No inter or intra-observer reliability test was performed at these centers.

Another weakness of the AHRI study is the lack of external validity (the extent to which the associations observed in a study are true in a larger, external population). However, even if the results of this study cannot be applied to the entire population in Africa or other parts of the world, it is possible to define a certain ethnic group (Hadiya) to which the results can be generalized with some degree of confidence. In addition, this apparent limitation, in fact, was a major advantage of this study: if different associations occur in different ethnic groups, the TDT in a mixed population may not detect an association due to the heterogeneity. In this study, since 80% are Hadiyan, heterogeneity should not have been a problem.

One more limitation to the AHRI study is the restriction of recruitment to families living close to Hosanna hospital. Although this makes the study more cost-efficient, it can distort the estimates of effect due to selection of families in specific areas (selection bias). Families living in more distant areas or those who refused to participate may be

different in some characteristics such as ethnicity, genetic components and clinical patterns of tuberculosis.

In regard to the classic TDT, a serious limitation is when genotype information is lost due to missing parent(s), which creates a loss of statistical power to detect a significant association. In this study, many families were eliminated only because parent(s) were not available. So, a relatively small sample size and limited power was a limitation to this study. Although the complete data set (AHRI study) includes some families without both parents, but with more than one child, implementation of the sib-TDT or other methods for testing for associations in general pedigrees was beyond the scope of this thesis.

CHAPTER 7: CONCLUSION

This study sought to examine 1) whether particular candidate genes influence the susceptibility/resistance to tuberculosis; and 2) whether certain individual characteristics modify the association between candidate genes and tuberculosis.

Despite the limitations mentioned, using a case-parental control design, this thesis provided evidence for inheritance of susceptibility to tuberculosis. In addition, the results of this analysis revealed significant interactions between different covariates (e.g., age, ethnicity, and clinical symptoms of tuberculosis) and genetic components to modify the risk of tuberculosis. The findings of this study are as a minimum generalizable to Hadiyan families in Ethiopia.

To confirm the findings in this thesis, it would be useful to conduct similar research (and with larger sample sizes) in numerous regions of the world where genetic and environmental exposures in populations with different ethnic origins can be examined and compared. Despite the fact that replication of these findings in other populations will increase the credibility of the findings, lack of replication does not necessarily refute the results as it can simply occur due to heterogeneity (Borecki & Suarez, 2001). Even when the associations are confirmed, the functional variants are still unknown. Much more work is required to find the exact DNA changes that influence disease susceptibility and resistance, to figure out what these sequence changes do and what they interact with, and to assess the risk ratios and attributable risks in populations around the world.

Studies such as these could also provide a surveillance system for population genotypes, such as those for incidence and prevalence of different diseases in different regions of the world. This can facilitate ways to inform prevention or treatment programs that consider genetic risk factors for tuberculosis.

REFERENCES

Abel L, Cot M, Mulder L, Carnevale P, Feingold J. Segregation analysis detects a major gene controlling blood infection levels in human malaria. *Am J Hum Genet* 1992; 50 (6):1308-17.

Abel L, Vu DL, Oberti J, Nguyen VT, Van VC, Guilloud-Bataille M, Schurr E, Lagrange PH. Complex segregation analysis of leprosy in southern Vietnam. *Genet Epidemiol* 1995; 12(1):63-82.

Abel L, Dessein AJ. Genetic epidemiology of infectious diseases in humans: design of population-based studies. *Emerg Infect Dis* 1998; 4(4):593-603.

Abel L, Müller-Myhsok B. Maximum-likelihood expression of the tranmission/ disequilibrium test and power calculation. *Am J Hum Genet* 1998; 63:664-667.

American Lung Association. American Lung Association fact sheet: TB and HIV, 2002. http://www.lungusa.org/diseases/tbhivfac.html.

Anonymous. Diagnostic standards and classification of TB in adults and children. *Am J Respir Crit Care Med* 2000; 161(4 Pt 1):1376-1395.

Barr RG, Diez-Roux AV, Knirsch CA, Pablos-Mendez A. Neighborhood poverty and the resurgence of TB in New York City, 1984-1992. *Am J Public Health* 2001; 91(9): 1487-1493.

Bates JH, Stead WW, Rado TA. Phage type of tubercle bacilli isolated from two or more sites of organ involvement. *Am Rev Respir Dis* 1976; 114:353-358.

Bates JH. The tuberculin skin test and preventive treatment for TB. In: Tuberculosis (Eds: Rom WN, Garay S), New York: Brown and Company 1995:865-871.

Behr MA, Warren SA, Salamon H, Hopewell PC, Ponce de Leon A, Daley CL, Small PM. Transmission of *M. tuberculosis* from patients smear-negative for acid-fast bacilli. *Lancet* 1999; 353(9151):444-449.

Bellamy R. Genetics and pulmonary medicine. 3. Genetic susceptibility to TB in human populations. *Thorax* 1998; 53(7):588-593.

Bellamy R, Hill AVS. Genetic susceptibility to mycobacteria and other infectious pathogens in humans. *Curr Opinion Immunol* 1998; 10:483-487.

Bellamy R, Ruwende C, Corrah T, McAdam KP, Whittle HC, Hill AV. Variations in the *NRAMP1* gene and susceptibility to TB in West Africans. *N Engl J Med* 1998a; 338 (10):640-644.

Bellamy R, Ruwende C, McAdam KP, Thursz M, Sumiya M, Summerfield J, Gilbert SC,

Corrah T, Kwiatkowski D, Whittle HC, Hill AV. Mannose binding protein deficiency is not associated with malaria, hepatitis B carriage nor TB in Africans. *Q J Med* 1998 b; 91(1):13-18.

Bellamy R, Ruwende C, Corrah T, McAdam KP, Thursz M, Whittle HC, Hill AV. TB and chronic hepatitis B virus infection in Africans and variation in the vitamin D receptor gene. *J Infect Dis* 1999; 179(3):721-724.

Berggren Palme I, Gudetta B, Degefu H, Muhe L, Bruchfeld J, Giesecke J. A controlled estimate of the risk of HIV infection in Ethiopian children with TB. *Epidemiol Infect* 2001; 127(3):517-525.

Bergner L, Yerby AS. Low income and barriers to use of health services. *N Engl J Med* 1968; 278:541-546.

Berkowity FE. Infections in children with severe protein-energy malnutrition. *Pediatr Infect Dis J* 1992; 11:750-759.

Borecki IB, Suarez BK. Linkage and association: Basic concepts. In: Genetic dissection of complex traits (Eds: Bao DC, Province MA), St. Louis: Academic Press 2001:45-63.

Brahmer JR, Small PM. Tuberculosis and nontuberculous mycobacterial infection. In: Textbook of Internal Medicine, fifth edition (Ed: Stein JH), St. Louis: A Times Mirror Company 1998:1625-1642.

Brandt L, Feino CJ, Weinreich OA, Chilima B, Hirsch P, Appelberg R, Andersen P. Failure of the *M. bovis* BCG vaccine: some species of environmental mycobacteria block multiplication of BCG and induction of protective immunity to TB. *Infect Immun* 2002; 70(2):672-678.

Bulik CM, Sullivan PF, Wade TD, Kendler KS. Twin studies of eating disorders. *Int J Eat Disord* 2000; 27(1):1-20.

Buu N, Sanchez F, Schurr E. The Bcg host-resistance gene. *Clin Infect Dis* 2000; 31 (Suppl 3):S81-85.

Canadian Lung Association. Canadian TB standards, 4th edition. Ottawa: Canadian Lung Association 1996.

Canadian Immunization Guide. Sixth edition. Ottawa: Canadian Medical Association 2002:71-76.

Cardon LR, Bell JI. Association study designs for complex diseases. *Nat Rev Genet* 2001; 2(2):91-99.

CDC (Centers for Disease Control and Prevention). TB and acquired immunodeficiency syndrome. New York. *MMWR* 1997; 36:785-797.

Cellier M, Govoni G, Vidal S, Kwan T, Groulx N, Liu J, Sanchez F, Skamene E, Schurr E, Gros P. Human natural resistance-associated macrophage protein: cDNA cloning, chromosomal mapping, genomic organization, and tissue-specific expression. *J Exp Med* 1994; 180(5):1741-1752.

Cellier M, Shustik C, Dalton W, Rich E, Hu J, Malo D, Schurr E, Gros P. Expression of the human *NRAMP1* gene in professional primary phagocytes: studies in blood cells and in HL-60 promyelocytic leukemia. *J Leukoc Biol* 1997; 61(1):96-105.

Cervino AC, Lakiss S, Sow O, Hill AV. Allelic association between the *NRAMP1* gene and susceptibility to tuberculosis in Guinea-Conakry. *Ann Hum Genet* 2000; 64(Pt 6):507-512.

Cobelens FGJ, Deutekom HV, Draayer-Jansen IWE, Schepp-Beelen ACHM, Gerven PJHJV, Kessel RPMV, Mensen MEA. Risk of infection with *M. tuberculosis* in travellers to areas of high TB endemicity. *Lancet* 2000; 356:461-465.

Comstock GW. Tuberculosis in twins: A re-analysis of the prophit survey. *Am Rev Resp Dis* 1978; 117:621-624.

Comstock GW, Geiter LJ. Prophylaxis. In: Tuberculosis and nontuberculous mycobacterial infections, fourth edition (Ed: Schlossberg D), Philadelphia: W.B. Saunders Company 1999:98-103.

Conkright JJ, Na CL, Weaver TE. Overexpression of surfactant protein-C mature peptide causes neonatal lethality in transgenic mice. *Am J Respir Cell Mol Biol* 2002; 26(1): 85-90.

Cooke GS, Hill AVS. Genetics of susceptibility to human infectious disease. *Nature Rev Genet* 2001; 2(12):967-977.

Cowie RL. The epidemiology of TB in gold miners with silicosis. *Am J Respir Crit Care Med* 1994; 150(5 Pt 1):1460-1462.

Crowle AJ, Elkins N. Relative permissiveness of macrophages from black and white people for virulent tubercle bacilli. *Infect Immunol* 1990; 58:632-638.

Cummings MR. Chromosomes, Mitosis, and Meiosis. In: Human Heredity, principles and issues (Ed: Cummings MR), St. Paul (U.S.A.): West Publishing Company 1988:11-39.

Curtis AB, Ridzon R, Vogel R, McDonough S, Hargreaves J, Ferry J, Valway S, Onorato IM. Extensive transmission of *M. tuberculosis* from a child. *N Engl J Med*

1999; 341(20):1491-1495.

Czeizel AE, Rockenbauer M, Olsen J, Sorensen HT. A population-based case-control study of the safety of oral anti-TB drug treatment during pregnancy. *Int J Tuberc Lung Dis* 2001; 5(6):564-568.

Daley CL, Small PM, Schecter GF, Schoolnik GK, McAdam RA, Jacobs WR, Hopewell PC. An outbreak of TB with accelerated progression among persons infected with the human immunodeficiency virus. An analysis using restriction-fragment-length polymorphisms. *N Engl J Med* 1992; 326:231-235.

Daniel TM, Bates JH, Downes KA. History of TB. In: Tuberculosis: pathogenesis, protection, control, (Ed: Bloom BR), Washington DC: American Society for Microbiology 1994:13-24.

Dannenberg AM. Pathophysiology of TB. In : Tuberculosis and nontuberculous mycobacterial infections, fourth edition (Ed: Schlossberg D), Philadelphia: W.B. Saunders Company 1999:17-28.

Davies PD, Brown RC, Woodhead JS. Serum concentrations of vitamin D metabolites in untreated TB. *Thorax* 1985; 40(3):187-190.

Davies PD, Brown RC, Church HA, Woodhead JS. The effect of anti-TB chemotherapy on vitamin D and calcium metabolism. *Tubercle* 1987; 68(4):261-266.

Davies RP, Tocque K, Bellis MA, Rimmington T, Davies PD. Historical decline in TB in England and Wales: improving social conditions or natural selection? *Int J Tuberc Lung Dis* 1999; 3:1051-1054.

Donald PR, Beyers N. TB in childhood. In: Clinical TB, second edition (Ed: Davies PDO), London: Chapman & Hall 1998:205-222.

Douglas AS, Strachan DP, Maxwell JD. Seasonality of TB: The reverse of other respiratory diseases in the UK. *Thorax* 1996; 51(9):944-946.

Dubaniewicz A, Lewko B, Moszkowska G, Zamorska B, Stepinski J. Molecular subtypes of the HLA-DR antigens in pulmonary TB. *Int J Infect Dis* 2000; 4(3):129-133.

Du Plessis DG, Warren R, Richardson M, Joubert JJ, Van Helden PD. Demonstration of reinfection and reactivation in HIV-negative autopsied cases of secondary TB: multi-lesional genotyping of *M. tuberculosis* utilizing IS 6110 and other repetitive element-based DNA fingerprinting. *Tuberculosis* 2001; 81(3):211-220.

Dutt AK, Stead WW. Epidemiology and host factors. In: Tuberculosis and non-tuberculous mycobacterial infections, fourth edition (Ed: Schlossberg D), Philadelphia: W.B. Saunders Company 1999:3-18.

Dye C, Scheele S, Dolin P, Pathania V, Raviglione MC. Global Burden of TB. Estimated incidence, prevalence, and mortality by country. *JAMA* 1999; 282(7):677-686.

Dye C. Tuberculosis 2000-2010: Control, but not elimination. *Int J Tuberc Lung Dis* 2000; 4(12 Suppl 2):S146-152.

Eaves LJ, Sullivan P. Genotype-environment interaction in transmission disequilibrium tests. In: Genetic dissection of complex traits (Eds: Rao DC, Province MA), 2001: 223-240.

Elseth GD, Baumgardner KD. Genetics: Early history and cytological foundations. In: Principles of modern genetics (Eds: Elseth GD, Baumgardner KD), St. Paul (U.S.A.): West publishing company 1995a:1-19.

Elseth GD, Baumgardner KD. Molecular evolution. In: Principles of modern genetics (Eds: Elseth GD, Baumgardner KD), St. Paul (U.S.A.): West publishing company 1995b:639-657.

Enarson DA, Murray JF. Global epidemiology of TB. In: Tuberculosis (Eds: Rom WN, Garay S), New York: Brown and Company 1995:57-75.

Evans CC. Historical background. In: Clinical TB, second edition (Ed: Davies PDO), London: Chapman & Hall 1998:3-19.

Ewens WJ, Spielman RS. The transmission/disequilibrium test: history, subdivision, and admixture. *Am J Hum Genet* 1995; 57:455-464.

Ewens WJ, Spielman RS. Disease association and the transmission/disequilibrium test (TST). *Curr Prot Hum Genet* 1997:1.12.1-1.12.13.

Farmer P. DOTS and DOTS-plus: not the only answer. *Ann N Y Acad Sci* 2001; 953:165-184.

Ferguson JS, Schlesinger LS. Pulmonary surfactant in innate immunity and the pathogenesis of TB. *Tuberc Lung Dis* 2000; 80(4-5):173-184.

Ferguson JS, Voelker DR, Ufnar JA, Dawson AJ, Schlesinger LS. Surfactant protein D inhibition of human macrophage uptake of *M. tuberculosis* is independent of bacterial agglutination. *J Immunol* 2002; 168(3):1309-1314.

Finch PJ, Millard FJ, Maxwell JD. Risk of TB in immigrant Asians: culturally acquired immunodeficiency? *Thorax* 1991; 46(1):1-5.

Fitzpatrick LK, Braden C. Tuberculosis. In: Kelley's textbook of internal Medicine, fourth edition (Ed: Humes HD), Philadelphia: Lippincott Williams & Wilkins 2000: 2055-2065.

Flanders D, Khoury MJ. Analysis of case-parental control studies: Method for study of associations between disease and genetic markers. *Am J Epidemiol* 1996; 144(7):696-703.

Floros J, Lin HM, Garcia A, Salazar MA, Guo X, DiAngelo S, Montano M, Luo J, Pardo A, Selman M. Surfactant protein genetic marker alleles identify a subgroup of TB in a Mexican population. *J Infect Dis* 2000; 182(5):1473-1478.

Garred P, Richter C, Andersen AB, Madsen HO, Mtoni I, Svejgaard A, Shao J. Mannan-binding lectin in the sub-Saharan HIV and TB epidemics. *Scand J Immunol* 1997; 46 (2):204-208.

Geiter L. Ending neglect. The elimination of tuberculosis in the United States (Ed: Geiter L), Washington, DC: National Academy Press 2000:1-292.

Globe M, Iseman MD, Madsen LA. Treatment of 171 patients with pulmonary TB resistance to isoniazid and rifampin. *N Engl J Med* 1993; 328:527-532.

Greenwood CMT, Fujiwara TM, Boothroyd LJ, Miller MA, Frappier D, Fanning EF, Schurr E, Morgan K. Linkage of TB to chromosome 2q35 loci, including *NRAMP1*, in a large aboriginal Canadian family. *Am J Hum Genet* 2000; 67:405-416.

Gros P, Skamene E, Forget A. Genetic control of natural resistance to *M. bovis* (BCG) in mice. *J Immunol* 1981; 127(6):2417-2421.

Gros P, Schurr E. Immunogenetics of host response to bacteria in mice. In: Immunology of infectious disease (Eds: Kaufmann SHE, Ahmad R), Washington DC: ASM Press 2002.

Haro AS. Tuberculosis in Finland. Past-present-future. TB and respiratory diseases Yearbook 1998; 18:1-109.

Hass F, Hass SS. The origin of *M. tuberculosis* and the notion of its contagiousness. In: Tuberculosis (Eds: Rom WN, Garay S), New York: Brown and Company 1995:3-19.

Hawkes CH. Twin studies in medicine-what do they tell us? *Q J Med* 1997; 90:311-321.

Hoal-Van Helden EG, Epstein J, Victor TC, Hon D, Lewis LA, Beyers N, Zurakowski D, Ezekowitz AB, Van Helden PD. Mannose-binding protein B allele confers protection against TB meningitis. *Pediatr Res* 1999; 45(4 Pt 1):459-464.

Holmes CB, Hausler H, Nunn P. A review of sex differences in the epidemiology of TB. *Int J Tuberc Lung Dis* 1998; 2:96-104.

Hoover RR, Floros J. Organization of the human SP-A and SP-D loci at 10q22-q23. Physical and radiation hybrid mapping reveal gene order and orientation. *Am J Respir*

Cell Mol Biol 1998; 18(3):353-362.

IUATLD (International Union Against TB and Lung Disease). Criteria for discontinuation of vaccination programmes using BCG in countries with a low prevalence of TB. *Tuberc Lung Dis* 1994; 75:179-181.

IUATLD. Program National de Lutte contre la Tuberculose de la République du Sénègal. Rapport No. 24 de l'UICTMR. Paris: IUATLD, 1998.

Jameson JL, Kopp P. Principles of human genetics. In: Harrison's, Principles of internal medicine, fifteenth edition (Eds: Braunwald E, Fauci AS, Kasper D, Hauser S, Longo DL, Jameson JL), United States: McGraw Companies 2001:375-396.

Jepson A. Twin studies for the analysis of heritability of infectious diseases. *Bull Inst Pasteur* 1998; 96:71-81.

Jepson A, Fowler A, Banya W, Singh M, Bennett S, Whittle H, Hill AV. Genetic regulation of acquired immune responses to antigens of *M. tuberculosis*: A study of twins in West Africa. *Infect Immun* 2001; 69(6):3989-3994.

John GT, Shankar V, Abraham AM, Mukundan U, Thomas PP, Jacob CK. Risk factors for post-transplant TB. *Kidney Int* 2001; 60(3):1148-1153.

Kallmann FJ, Reisner D. Twin studies on the significance of genetic factors in TB. *Am Rev Tuberc* 1943; 147:549-571.

Khoury MJ, Beaty TH, Cohen BH. Scope and strategies of genetic epidemiology. In: Fundamentals of genetic epidemiology. New York: Oxford University Press 1993a: 1-25.

Khoury MJ, Beaty TH, Cohen BH. Epidemiologic approaches to familial aggregation. In: Fundamentals of genetic epidemiology. New York: Oxford University Press 1993 b:164-199.

Khoury MJ, Flanders WD. Nontraditional epidemiologic approaches in the analysis of gene-environment interaction: case-control studies with no controls! *Am J Epidemiol* 1996; 144(3):207-213.

Khoury MJ. Genetic Epidemiology. In: Modern epidemiology, second edition (Ed: Rothman K, Greenland S), Philadelphia: Lippincott-Raven 1998:609-621.

Khoury MJ, Yang Q. The future of genetic studies of complex human diseases: An epidemiologic perspective. *Epidemiol* 1998; 9(3):350-354.

Knowler WC, Williams RC, Pettit DJ, Steinberg AG. Gm3;5,13,14 and type 2 diabetes mellitus: an association in American Indians with genetic admixture. *Am J Hum*

Genet 1988; 43(4):520-526.

Kondo A, Oketani N, Maruyama M, Taguchi Y, Yamaguchi Y, Miyao H, Mashima I, Oono M, Wada K, Tsuchiya T, Takahashi H, Abe S. Significance of serum surfactant protein-D (SP-D) level in patients with pulmonary TB. *Kekkaku* 1998; 73(10):585-590.

Kondo S, Ito M, Kageyama S. Infants 12 months-old or less as a high risk group in TB-comparison of clinical data with those in children aged one to two years. *Kekkaku* 2001; 76(5):407-411.

Kwiatkowski D. Genetic dissection of the molecular pathogenesis of severe infection. *Intensive Care Med* 2000; 26(Suppl 1):S89-97.

Labuda M, Ross MV, Fujiwara TM, Morgan K, Ledbetter D, Hughes MR, Glorieux F. Two hereditary defects related to vitamin D metabolism map to same region of human chromosome 12q. *Cytogenet Cell Genet* 1991; 58:1978.

Lawn SD, Labeta MO, Arias M, Acheampong JW, Griffin GE. Elevated serum concentrations of soluble CD14 in HIV- and HIV+ patients with TB in Africa: Prolonged elevation during anti-TB treatment. *Clin Exp Immunol* 2000; 120(3):483-487.

Lederberg J. JBS. Haldane (1949) on infectious disease and evolution. *Genetics* 1999; 153(1):1-3.

Lipscombe RJ, Sumiya M, Hill AV, Lau YL, Levinsky RJ, Summerfield JA, Turner MW. High frequencies in African and non-African populations of independent mutations in the mannose binding protein gene. *Hum Mol Genet* 1992; 1:709–715.

Lipsitch M, Sousa AO. Historical intensity of natural selection for resistance to tuberculosis. *Genetics* 2002; 161:1599-1607.

Long R, Nijoo H, Hershfield E. Tuberculosis: Epidemiology of the disease in Canada. *Can Med Assoc J* 1999; 160:1185-1190.

Maher D, Raviglione MC. The global epidemic of TB: A WHO perspective. In: Tuberculosis and nontubercolous mycobacterial infection , fourth edition (Ed: Schlossberg D), Philadelphia: W.B. Saunders Company 1999:104-115.

Mangura BT, Napolitano EC, Passannante MR, McDonald RJ, Reichman LB. *M. tuberculosis* miniepidemic in a church gospel choir. *Chest* 1998; 113:234-237.

Mehlman MJ, Botkin J. The human genome project. In: Access to the genome, the challenge to equality (Eds: Mehlman MJ, Botkin J), Georgetown University Press, Washington, D.C. 1998:7-19.

Motulsky AG. Metabolic polymorphisms and the role of infectious diseases in human evolution. *Hum Biol* 1960; 32:28-62.

Moulding T. Pathogenesis, pathophysiology, and immunology: Clinical orientations. In: Tuberculosis and nontubercolous mycobacterial infection , fourth edition (Ed: Schlossberg D), Philadelphia: W.B. Saunders Company 1999:48-56.

Nunn P, Harries A, Godfrey-Faussett P, Gupta R, Maher D, Raviglione M. The research agenda for improving health policy, systems performance, and service delivery for tuberculosis control: A WHO perspective. *Bull World Health Organ* 2002; 80(6):471-476.

Marquet S, Lepage P, Hudson TJ, Musser JM, Schurr E. Complete nucleotide sequence and genomic structure of the human NRAMP1 gene region on chromosome region 2q35. *Mamm Genome* 2000; 11(9):755-762.

O'Brien RJ. Preventive therapy. In: Clinical TB, second edition (Ed: Davies PDO), London: Chapman & Hall 1998:397-416.

Ott J. Basic genetics and cytogenetics. In: Analysis of human genetic linkage. Revised edition (Ed: Ott J), Baltomore and London: The John Hopkins University Press 1991:1-38.

Patterson PE, Kimerling ME, Bailey WC, Dunlap NE. Chemotherapy of TB. In: Tuberculosis and nontuberculous mycobacterial infections, fourth edition (Ed: Schlossberg D), Philadelphia: W.B. Saunders Company 1999:71-82.

Pericak-Vance MA. Linkage disequilibrium and allelic association. In: Approaches to gene mapping in complex human diseases (Eds: Haines JL, Pericak-Vance MA), New York: J. Wiley & Sons 1998:323-333.

Raviglione MC, O'Brien RJ. Tuberculosis. In: Harrison's, Principles of internal medicine, fifteenth edition (Eds: Braunwald E, Fauci AS, Kasper D, Hauser S, Longo DL, Jameson JL), United States: McGraw Companies 2001:1024-1034.

Reichman LB, Felton CP, Edsall JR. Drug dependence, a possible new factor for TB disease. *Arch Intern Med* 1979; 139:337-339.

Risch N, Merikangas K. The future of genetic studies of complex human diseases. *Science* 1996 ;273(5281):1516-1517.

Robins JM, Smoller JW, Lunetta KL. On the validity of the TDT test in the presence of comorbidity and ascertainment bias. *Genet Epidemiol* 2001; 21:326-336.

Rockett KA, Brookes R, Udalova I, Vidal V, Hill AV, Kwiatkowski D. 1,25-dihydroxy-vitamin D3 induces nitric oxide synthase and suppresses growth of *M. tuberculosis* in

a human macrophage-like cell line. *Infect Immunol* 1998; 66(11):5314-5321.

Rook GA, Steele J, Fraher L, Barker S, Karmali R, O'Riordan J, Stanford J. Vitamin D_3, gamma interferon, and control of proliferation of *M. tuberculosis* by human monocytes. *Immunology* 1986; 57(1):159-163.

Saiman L, San Gabriel P, Schulte J, Vargas MP, Kenyon T, Ontario I. Risk factors for latent TB infection among children in New York City. *Pediatrics* 2001a; 107(5):999-1003.

Saiman L, Aronson J, Zhou J, Gomez-Duarte C, Gabriel PS, Alonso M, Maloney S, Schulte J. Prevalence of infectious diseases among internationally adopted children. *Pediatrics* 2001b; 108(3):608-612.

SAS Language. Reference, Version 6.0, first edition Cary, NC: SAS Institute Inc. 1990.

Schaid DJ, Sommer SS. Genotype relative risks: methods for design and analysis of candidate-gene association studies. *Am J Hum Genet* 1993; 53(5):1114-1126.

Schaid DJ, Sommer SS. Comparison of statistics for candidate-gene association studies using cases and parents. *Am J Hum Genet* 1994; 55(2):402-409.

Schaid DJ, Rowland CM. Genetic ASSOCiation analysis software for cases and parent, Version 1.06, 2001. http://www.mayo.edu/statgen/.

Schneemann M, Schoedon G, Hofer S, Blau N, Guerrero L, Schaffner A. Nitric oxide synthase is not a constituent of the antimicrobial armature of human mononuclear phagocytes. *J Infect Dis* 1993; 167(6):1358-1363.

Schoolnik G. Tuberculosis metabolism. *BioMedNet* Conference Reporter from American Society for Microbiology and The Institute for Genomic Research-Microbial Genomes, 2001. http://news.bmn.com/conferences/list/view?bn_id=3120.

Selvaraj P, Narayanan PR, Reetha AM. Association of functional mutant homozygotes of the mannose binding protein gene with susceptibility to pulmonary TB in India. *Tuberc Lung Dis* 1999; 79(4):221-227.

Selvaraj P, Kurian SM, Uma H, Reetha AM, Narayanan PR. Influence of non-MHC genes on lymphocyte response to *M. tuberculosis* antigens and tuberculin reactive status in pulmonary TB. *Indian J Med Res* 2000; 112:86-92.

Sham P. Genetic epidemiology. *Br Med Bull* 1996; 52(2):408-433.

Sham P, Curtis D. An extended transmission/disequilibrium test (TDT) for multiallele marker loci. *Ann Hum Genet* 1995; 59:323-336.

Shaw MA, Collins A, Peacock CS, Miller EN, Black GF, Sibthorpe D, Lins-Lainson Z, Shaw JJ, Ramos F, Silveira F, Blackwell JM. Evidence that genetic susceptibility to *M. tuberculosis* in a Brazilian population is under oligogenic control: linkage study of the candidate genes *NRAMP1* and *TNFA*. *Tuberc Lung Dis* 1997; 78(1):35-45.

Shor-Posner G, Miguez MJ, Pineda LM, Rodriguez A, Ruiz P, Castillo G, Burbano X, Lecusay R, Baum M. Impact of selenium status on the pathogenesis of mycobacterial disease in HIV-1-infected drug users during the era of highly active antiretroviral therapy. *J Acquir Immune Defic Syndr* 2002; 29(2):169-173.

Small PM, Shafer RW, Hopewell PC. Exogenous reinfection with MDR-*M. tuberculosis* in patients with advanced HIV infection. *N Engl J Med* 1993; 328:1137-1144.

Small PM, Selcer UM. Human immunodeficiency virus and TB. In: Tuberculosis and nontubercolous mycobacterial infection , fourth edition (Ed: Schlossberg D), Philadelphia: W.B. Saunders Company 1999:329-338.

Spielman RS, McGinnis RE, Ewens WJ. Transmission test for linkage disequilibrium: The insulin gene region and insulin-dependent diabetic mellitus. *Am J Hum Genet* 1993; 52:506-516.

Spielman RS, Ewens WJ. A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. *Am J Hum Genet* 1998; 62(2):450-458.

Spielman RS, Ewens WJ. Transmission disequilibrium test and sib transmission disequilibrium test, Version 1.1, 1999, http://spielman07.med.upenn.edu/TDT.htm

Starke JR. Tuberculosis in Infants and Children. In: Tuberculosis and nontuberculous mycobacterial infections, fourth edition (Ed: Schlossberg D), Philadelphia: W.B. Saunders Company 1999:303-324.

Stead WW, Senner JW, Reddick WT. Racial differences in susceptibility to infection by *M. tuberculosis*. *N Engl J Med* 1990; 322:422-427.

Stead WW. Genetics and resistance to TB: Could resistance be enhanced by genetic engineering? *Ann Intern Med* 1992; 116(11):937-941.

Stead WW, Bates JH. Geographic and evolutionary epidemiology of TB. In: Tuberculosis (Eds: Rom WN, Garay S), New York: Brown and Company 1995:77-83.

Summerfield JA., Ryder S, Sumiya M, Thursz M, Gorchein A, Monteil MA, Turner MW. Mannose-binding protein gene-mutations associated with unusual and severe infections in adults. *Lancet* 1995; 345:886-889.

Sun F, Flanders WD, Yang Q, Khoury MJ. A new method for estimating the risk ratio in studies using case-parental control design. *Am J Epidemiol* 1998; 148(9):902-909.

Tabet SR, Goldbaum GM, Hooton TM, Eisenach KD, Nolan CM. Restriction fragment length polymorphism analysis detecting a community-based TB outbreak among persons infected with HIV. *J Infect Dis* 1994; 169(1):189-192.

Talenti A, Iceman M. Drug-Resistant TB. *Drugs* 2000; 59(2):171-179.

Turner MW. Mannose-binding lectin: the pluripotent molecule of the innate immune system. *Immunol Today* 1996; 17(11):532-540.

Turner MW, Dinan L, Heatley S, Jack DL, Boettcher B, Lester S, McCluskey J, Roberton D. Restricted polymorphism of the mannose-binding lectin gene of indigenous Australians. *Hum Mol Genet* 2000; 9(10):1481-1486.

Vidal SM, Malo D, Vogan K, Skamene E, Gros P. Natural resistance to infection with intracellular parasites: isolation of a candidate for Bcg. *Cell* 1993; 73(3):469-485.

Vidal SM, Tremblay ML, Govoni G, Gauthier S, Sebastiani G, Malo D, Skamene E, Olivier M, Jothy S, Gros P. The Ity/Lsh/Bcg locus: Natural resistance to infection with intracellular parasites is abrogated by disruption of the Nramp1 gene. *J Exp Med* 1995; 182(3):655-666.

Wacholder S, Rothman N, Caporaso N. Population stratification in epidemiologic studies of common genetic variants and cancer: quantification of bias. *J Natl Cancer Inst* 2000; 92(14):1151-1158.

WHO (World Health Organization). Global TB programme and global programme for vaccines. Statement on BCG revaccination for the prevention of TB. *Wkly Epidemiol Rec* 1995a; 70:229-231.

WHO. Global Programme for Vaccines and Immunization. Immunization policy. Geneva: WHO, 1995b:1-51.

WHO. Breastfeeding and Maternal TB. A statement prepared jointly by CHD, GTB, GPV and RHT (WHO). Update 1998b, No 23.

WHO. Guidelines for the prevention of TB in health care facilities in resource-limited settings, 1999.

WHO. TB and BCG 2000. http://www.who.int/vaccines/intermediate/TB.htm.

WHO. Global TB Control. *WHO Report*. Geneva, Switzerland, WHO 2001a.

WHO. TB and Air Travel-Executive Summary, Global TB Programme. Guideline for prevention and control. WHO 2001b.

WHO/IUATLD. The WHO/IUATLD Global project of Anti-TB drug resistance

surveillance 2001. Report No 2- Prevalence and trends.

Wilkinson RJ, Llewelyn M, Toossi Z, Patel P, Pasvol G, Lalvani A, Wright D, Latif M, Davidson RN. Influence of vitamin D deficiency and vitamin D receptor Polymorphisms on TB among Gujarati Asians in west London: a case-control study. *Lancet* 2000; 355(9204):618-621.

Witte JS, Gauderman WJ, Thomas DC. Asymptotic bias and efficiency in case-control studies of candidate genes and gene-environment interactions: basic family designs. *Am J Epidemiol* 1999; 149(8):693-705.

Youmans GP. Tuberculosis. Philadelphia W.B. Saunders Company 1979:323.

APPENDIX I

Centre universitaire de santé McGill
McGill University Health Centre

December 6, 2001

Dr. Nooshin Ahmadipour
1745 Cedar Avenue #514
Montreal, Quebec
H3G 1A7

Dear Dr. Ahmadipour:

Thank you for your letter requesting ethics approval to conduct research for your Master's thesis under the supervision of Dr. Theresa Gyorkos and Dr. Celia Greenwood in relation to Dr. Erwin Schurr's study entitled "Genetic Epidemiology of Tuberculosis".

It is noted that you wish to test the polymorphisms in candidate genes that are non-randomly associated with susceptibility to tuberculosis. You also plan to investigate the strength of genetic effect associated with different genes, and how this association varies with risk factors such as age of onset and severity of the disease, and also to identify the role of different genes in progression of disease to its later stages.

I am pleased to inform you that it is deemed appropriate, and I hereby grant you approval to conduct your Master's thesis in relation to the information already collected for Dr. Erwin Schurr's study. Please provide us with a letter of permission from Dr. Schurr allowing you permission to have access to this information.

Good luck with your Master's thesis.

Sincerely,

Denis Cournoyer,
Chairman
Research Ethics Committee
MUHC-Montreal General Hospital

Cc: Dr. Theresa Gyorkos
Dr. Celia Greenwood
Dr. Erwin Schurr

1650 Cedar, Montréal (Québec) Canada, H3G 1A4 (514) 937-6011

Centre universitaire de santé McGill
McGill University Health Centre

December 2, 1998

Dr. Erwin Schurr
Associate Professor
Medicine and Human Genetics
McGill University

RE: REC. 98-040 entitled "Genetic Epidemiology of Tuberculosi s."

Dear Dr. Schurr:

We are pleased to inform you that your modifications complying to the recommendations made by the Committee on November 24, 1998, we e found ethically acceptable. The Research Ethics Committee hereby grants y ou approval, for the revised consent forms on December 2, 1998, valid until November ', 1999.

Please be reminded that our policies require the following information:

- **date of activation and completion of the study**
- **notification of any change to the protocol**
- **notification of any adverse events, drug reactions, or pr c blems occurring during the course of the study**
- **an annual report (a short questionnaire will be sent to yo u in approximately 10 months from approval date)**
- **a reprint of article arising from the study**
- **a final report upon completion of the study**

Please be reminded that your study has not been approved for recruitm ent of minors or adults not competent to consent and that consent forms shou ld disclose all reasonably foreseeable risks no matter how rare or minimal. Please note that patient recruitment and study procedures should follow recommendatio ns found in the minutes (November 24, 1998) of the Research Ethics Committee m eeting.

We trust that this meets with your complete satisfaction.

Sincerely,

Dr. Denis Cournoyer, Chairman
MGH Research Ethics Committee

Le Centre universitaire de santé McGill (CUSM) comprend l'Hôpital de Montréal pour Enfants, l'Hôpital général de Montréal, l'Hôpital neurologique de Montréal et l'Hôpital Royal Victoria. Le CUSM est affilié à la Faculté de médecine de l'Université McGill.
The McGill University Health Centre (MUHC) consists of The Montreal Children's Hospital, The Montreal General Hospital, The Montreal Neurological Hospital, and The Royal Victoria Hospital. The MUHC is affiliated with the McGill University Faculty of Medicine.

HÔPITAL GÉNÉRAL DE MONTRÉAL
1650, av. Cedar, Montréal (Québec) H3G 1A4, Tél.: (514) 937-6011

31 01 008703

Centre universitaire de santé McGill
McGill University Health Centre

February 15, 2001

Dr. Erwin Schurr
Research Institute
Montreal General Hospital

RE: REC. 98-040 entitled "Genetic Epidemiology of Tuberculosis."

Dear Dr. Schurr:

Thank you for your follow-up for renewal of the above-mentioned study. Please note that the above mentioned protocol continues to meet the approval of the Montreal General Hospital Research Ethics Committee. We are pleased to inform you that re-approval is hereby granted on February 15, 2001, to continue your study at the Montreal General Hospital. You must seek renewal in January 2002.

Please be reminded that our policies require the following information:

- **date of activation (if not already provided) and completion of the study**
- **notification of any change to the protocol**
- **notification of any adverse events, drug reactions, or problems occurring during the course of the study**
- **an annual report (a short questionnaire will be sent to you in approximately 10 months from re-approval date)**
- **a reprint of article arising from the study**
- **a final report upon completion of the study**

Also, any condition that applied to the initial approval of your study continues to apply to this renewed approval.

Sincerely,

Denis Cournoyer,
Chairman
Research Ethics Committee
MUHC-MGH

በኢትዮጵያ ፌደራላዊ ዲሞክራሲያዊ ሪፑብሊክ
የኢትዮጵያ ሳይንስና ቴክኖሎጂ ኮሚሽን
The Federal Democratic Republic of Ethiopia
Ethiopian Science and Technology Commission

ቁጥር Ref No. RDHE/57-26/89

ቀን Date 11 MAR 1994

Prof.Sven Britton
Director of AHRI
Addis Ababa

Re:- Project proposal " Genetic Epidemiology of Tuberculosis"

We have received an Institutional letter of consent from the Southern Nation, Nationalities and Peoples Regional Health Bureau regarding the aforementioned project.

It is, therefore, our pleasure to inform you that the project is ethically approved for implementation.

With regards,

Sincerely youre,

YEMANE TEKLAI (Dr.)
Head, Health Department

and re

Received 199
Copy to

You may contact:
P. O. Box Addis Ababa, Ethiopia
Tel. 13 44
Telex 21568

GENETIC EPIDEMIOLOGY OF TUBERCULOSIS
DATA ACQUISITION FORM
HOSSANA HOSPITAL

DATE

NAME SEX (F/M) AGE (Yrs)

STUDY CODE ETHNICITY FAMILY ID

Is the individual a Case (C) Family (F) Population (P)

If Family (F)	
Study code of the index case	HOS
Relationship to the index case	

If Case (C)	
HOSSANA HOSPITAL#	
TB UNIT #	
If not the index case,	
Study code of the index case	HOS
Relationship to the index case	

CLINICAL EXAMINATION

General condition: GOOD | BAD | BED RIDDEN | CACHECTIC

WEIGHT (Kg)

	Y/N
Weight loss in past 6 months	
Loss less than 10 kg	
Loss more than 10 kg	
Do you suspect EPTB	

SYMPTOM	Y/N	DURATION (Wks)
Cough		
Hemaptysis		
Cough + Sputum		
Night Sweats		
Fever		

Does the individual have BCG vaccination scar Y | N | NOT SURE

If not a case, has he/she been treated for TB before Y | N | NOT SURE

SPUTUM RESULTS

Was sputum done at the clinic (Y/N) Clinic lab.#

Date done Result (POS/NEG) Ridley's scale grading

Was sputum collected to AHRI (Y/N) date collected

Was blood collected to AHRI (Y/N) date collected

Was the individual enrolled in the vaccine study (Y/N) Vaccine code

REMARKS _____

_____

AHRI #

APPENDIX II

Data from file: TDT-NRAMP1.txt

TDT-STDT Program 1.1

TDT with one missing parent allowed

Locus: 1 (NRAMP1-5′(GT)n)

	TDT			S-TDT				Combined Scores			
Allele	b	c	Chi-Sq	Y	Mean(A)	Var(V)	z'	W	Mean(A)	Var(V)	z'
1	34	31	0.138	0	0.000	0.000	N/A	34	32.500	16.250	0.248
2	31	34	0.138	0	0.000	0.000	N/A	31	32.500	16.250	0.248

Locus: 2 (NRAMP1-274C/T)

	TDT			S-TDT				Combined Scores			
Allele	b	c	Chi-Sq	Y	Mean(A)	Var(V)	z'	W	Mean(A)	Var(V)	z'
1	34	26	1.067	0	0.000	0.000	N/A	34	30.000	15.000	0.904
2	26	34	1.067	0	0.000	0.000	N/A	26	30.000	15.000	0.904

Locus: 3 (NRAMP1-469+14G/C)

	TDT			S-TDT				Combined Scores			
Allele	b	c	Chi-Sq	Y	Mean(A)	Var(V)	z'	W	Mean(A)	Var(V)	z'
1	26	31	0.439	0	0.000	0.000	N/A	26	28.500	14.250	0.530
2	31	26	0.439	0	0.000	0.000	N/A	31	28.500	14.250	0.530

Locus: 4 (NRAMP1-3′UTR)

	TDT			S-TDT				Combined Scores			
Allele	b	c	Chi-Sq	Y	Mean(A)	Var(V)	z'	W	Mean(A)	Var(V)	z'
1	43	43	0.000	0	0.000	0.000	N/A	43	43.000	21.500	-0.108
2	43	43	0.000	0	0.000	0.000	N/A	43	43.000	21.500	-0.108

Transmissions to unaffected sibs in families where TDT is used.

Locus: 1

Allele	b	c
1	0	0
2	0	0

Locus: 2

Allele	b	c
1	0	0
2	0	0

Locus: 3

Allele	b	c
1	0	0
2	0	0

Locus: 4

Allele	b	c
1	0	0
2	0	0

Data from file: TDT-VDR.txt

TDT-STDT Program 1.1

TDT with one missing parent allowed

Locus: 1 (VDR-117)

		TDT		S-TDT				Combined Scores			
Allele	b	c	Chi-Sq	Y	Mean(A)	Var(V)	z'	W	Mean(A)	Var(V)	z'
1	32	22	1.852	0	0.000	0.000	N/A	32	27.000	13.500	1.225
2	22	32	1.852	0	0.000	0.000	N/A	22	27.000	13.500	1.225

Locus: 2 (VDR-1056)

		TDT		S-TDT				Combined Scores			
Allele	b	c	Chi-Sq	Y	Mean(A)	Var(V)	z'	W	Mean(A)	Var(V)	z'
1	31	33	0.062	0	0.000	0.000	N/A	31	32.000	16.000	0.125
2	33	31	0.062	0	0.000	0.000	N/A	33	32.000	16.000	0.125

Transmissions to unaffected sibs in families where TDT is used.

Locus: 1

Allele	b	c
1	0	0
2	0	0

Locus: 2

Allele	b	c
1	0	0
2	0	0

Data from file: TDT-MBL.txt

TDT-STDT Program 1.1

TDT with one missing parent allowed

Locus: 1 (MBL-G54D)

	TDT			S-TDT				Combined Scores			
Allele	b	c	Chi-Sq	Y	Mean(A)	Var(V)	z'	W	Mean(A)	Var(V)	z'
1	18	11	1.690	0	0.000	0.000	N/A	18	14.500	7.250	1.114
2	11	18	1.690	0	0.000	0.000	N/A	11	14.500	7.250	1.114

Locus: 2 (MBL-G57Q)

	TDT			S-TDT				Combined Scores			
Allele	b	c	Chi-Sq	Y	Mean(A)	Var(V)	z'	W	Mean(A)	Var(V)	z'
1	17	18	0.029	0	0.000	0.000	N/A	17	17.500	8.750	0.000
2	18	17	0.029	0	0.000	0.000	N/A	18	17.500	8.750	0.000

Transmissions to unaffected sibs in families where TDT is used.

Locus: 1

Allele	b	c
1	0	0
2	0	0

Locus: 2

Allele	b	c
1	0	0
2	0	0

Data from file: TDT-SFTPA1.txt

TDT-STDT Program 1.1

TDT with one missing parent allowed

Locus: 1 (SFTPA1-170)

		TDT		S-TDT				Combined Scores			
Allele	b	c	Chi-Sq	Y	Mean(A)	Var(V)	z'	W	Mean(A)	Var(V)	z'
1	10	8	0.222	0	0.000	0.000	N/A	10	9.000	4.500	0.236
2	8	10	0.222	0	0.000	0.000	N/A	8	9.000	4.500	0.236

Locus: 2 (SFTPA1-256)

		TDT		S-TDT				Combined Scores			
Allele	b	c	Chi-Sq	Y	Mean(A)	Var(V)	z'	W	Mean(A)	Var(V)	z'
2	37	28	1.246	0	0.000	0.000	N/A	37	32.500	16.250	0.992
1	28	37	1.246	0	0.000	0.000	N/A	28	32.500	16.250	0.992

Locus: 3 (SFTPA1-294)

		TDT		S-TDT				Combined Scores			
Allele	b	c	Chi-Sq	Y	Mean(A)	Var(V)	z'	W	Mean(A)	Var(V)	z'
1	37	21	4.414	0	0.000	0.000	N/A	37	29.000	14.500	1.970
2	21	37	4.414	0	0.000	0.000	N/A	21	29.000	14.500	1.970

Locus: 4 (SFTPA1-507)

		TDT		S-TDT				Combined Scores			
Allele	b	c	Chi-Sq	Y	Mean(A)	Var(V)	z'	W	Mean(A)	Var(V)	z'
1	4	5	0.111	0	0.000	0.000	N/A	4	4.500	2.250	0.000
2	5	4	0.111	0	0.000	0.000	N/A	5	4.500	2.250	0.000

Locus: 5 (SFTPA1-763)

		TDT		S-TDT				Combined Scores			
Allele	b	c	Chi-Sq	Y	Mean(A)	Var(V)	z'	W	Mean(A)	Var(V)	z'
2	13	6	2.579	0	0.000	0.000	N/A	13	9.500	4.750	1.376
1	6	13	2.579	0	0.000	0.000	N/A	6	9.500	4.750	1.376

Transmissions to unaffected sibs in families where TDT is used.

Locus: 1

Allele	b	c
1	0	0
2	0	0

Locus: 2

Allele	b	c
2	0	0
1	0	0

Locus: 3

Allele	b	c
1	0	0
2	0	0

Locus: 4

Allele	b	c
1	0	0
2	0	0

Locus: 5

Allele	b	c
2	0	0
1	0	0

APPENDIX III

The program "gassoc" in appendix III, compares genotypes "1,2" and "2,2" to genotype "1,1." Throughout the Methods section in this thesis, allele "1" has been considered to be the high risk allele. For this reason, the definitions of the two relative risk parameters ψ_1 and ψ_2 are altered from the definitions used in Schaid and Sommer 1994.

```
**********************************************
*  gassoc Version 1.06
**********************************************
Pedigree 10, Person 138797: genotype inconsistent with parents
Pedigree 116, Person 16599: genotype inconsistent with parents
=====================================================================

ANALYSIS FOR Marker - SFTPA1-294, Locus =  4
=======================
Summary Info:

  # of valid lines in input file: 285
  # of affected cases: 106
  # of affected cases used in analysis:  77
  # of affected cases not used: 29
    # not used due to missing parent or missing parent alleles: 14
    # not used due to case missing alleles: 13
    # not used due to inconsistent parent/case alleles: 2


GRR coding scheme
Conditional Logistic:
 Final estimates of Beta:

                          Rel. Risk
Genetype      Beta        exp(Beta)     SE(Beta)      Z           P(2-sided)
2|X           -0.3567     0.7000        0.3112        -1.1462     0.25170524
2|2           -2.1317     0.1186        1.0740        -1.9848     0.04716295

 LR Statistic:        6.5805,   df=2,          p=0.037244387


 Covariance/Correlation Matrix (*=Corr(Bi,Bj)):
       0.0968       0.2183*
       0.0730       1.1534

TDT coding scheme
Conditional Logistic:
 Final estimates of Beta:

                      Rel. Risk
 Allele  Beta         exp(Beta)     SE(Beta)      Z           P(2-sided)
 2       -0.5664      0.5676        0.2732        -2.0730     0.03817467


 LR Statistic:        4.4716,  df=1,  p=0.034463681


 Covariance/Correlation Matrix (*=Corr(Bi,Bj)):
       0.0746


Score Statistics:

              Score     df  P-value          Sim P-value(Simulations=20)
GTDT:         4.4138    1   0.035649489      0.050000000
GDOM:         2.0250    1   0.154728924      0.200000000
GREC:         4.0833    1   0.043308143      0.000000000


Note: Seeds used for random# generation were 1000, 1200, 1400
**********************************************
```

* gassoc Version 1.06
****
Pedigree 84, Person 116298: genotype inconsistent with parents
Pedigree 116, Person 16599: genotype inconsistent with parents
==========

ANALYSIS FOR Marker - SFTPA1-763, Locus = 6
==========
Summary Info:

# of valid lines in input file: 285

# of affected cases: 106
# of affected cases used in analysis: 47
# of affected cases not used: 59
not used due to missing parent or missing parent alleles: 20
not used due to case missing alleles: 37
not used due to inconsistent parent/case alleles: 2

GRR coding scheme
Conditional Logistic:
Final estimates of Beta:

Genetype	Beta	Rel. Risk exp(Beta)	SE(Beta)	Z	P(2-sided)
2|X	-0.9266	0.3959	1.2752	-0.7267	0.46742082
2|2	0.1892	1.2082	1.2467	0.1517	0.87939115

LR Statistic: 4.2094, df=2, p=0.121880266

Covariance/Correlation Matrix (*=Corr(Bi,Bj)):
1.6261 0.8916*
1.4174 1.5541

TDT coding scheme
Conditional Logistic:
Final estimates of Beta:

Allele	Beta	Rel. Risk exp(Beta)	SE(Beta)	Z	P(2-sided)
1	-0.7730	0.4616	0.4935	-1.5661	0.11732104

LR Statistic: 2.6407, df=1, p=0.104157094

Covariance/Correlation Matrix (*=Corr(Bi,Bj)):
0.2436

Score Statistics:

	Score	df	P-value	Sim P-value(Simulations=20)
GTDT:	2.5789	1	0.108293656	0.250000000
GDOM:	3.6885	1	0.054788060	0.050000000
GREC:	0.1111	1	0.738882680	1.000000000

Note: Seeds used for random# generation were 1000, 2000, 3000

APPENDIX IV

Calculation of the proportion of transmissions: Sample calculations are presented here to show how the proportion of transmissions that received allele "1" was calculated from the logistic models. For table 5.12, the model predicting SFTPA1-170 with the binary covariate of weight loss >10 kg (WL10), calculations go as follows:

The estimate for the intercept = 0.559
The estimate for WL10 = -2.351

The logistic model fit by SAS PROC LOGISTIC is:

$$\log p\,(y = 0)/\, p(\, y = 1) = \alpha + \beta x = 0.559 - 2.351$$

Therefore $p\,(y = 0) = p(\text{alleles transmitted}) = e^{\alpha + \beta x}/\, 1 + e^{\alpha + \beta x}$

When WL10 = 0, $p(y = 0) = e^{0.559}/\, 1 + e^{0.559} = 0.63$
When WL10 = 1, $p(y = 0) = e^{-1.79}/\, 1 + e^{-1.79} = 0.14$
Therefore, $p(y = 1) = 1 - 0.14 = 0.86$